# Multi-year maintenance optimisation for paved public roads - segment based modelling and price-directive decomposition 

Per-Åke Andersson

Department of Mathematics
Linköping 2007

Multi-year maintenance optimisation for paved public roads -- segment based modelling and price-directive decomposition
(C) 2007 Per-Åke Andersson

Matematiska institutionen
Linköpings universitet SE-581 83 Linköping, Sweden
peand@mai.liu.se
Linköping studies in science and technology. Dissertations, No. 1107 ISBN: 978-91-85831-73-9
ISSN 0345-7524

## Acknowledgements

This project has been administered by CDU, KTH (Centre for research and education in operation and maintenance of infrastructure, Royal institute of technology) and performed at MAI, LiU (Department of mathematics, Linköping university), with financial support from VV (Swedish road administration), Vinnova, VTI (Swedish national road and transport research institute), Örebro university and my wife Lena.

I thank

- professor em. Sven Erlander, MAI, for giving me a $1^{\text {st }}$ chance and for introducing me into optimisation and traffic, professor em. Gunnar Aronsson, MAI, for a $2^{\text {nd }}$ and my main supervisor professor em. Per Olov Lindberg for a $3^{\text {rd }}$ chance and for his patience and the constructive-minded reading,
- my other supervisors Torbjörn Larsson, MAI, for his wit at meetings, and Lars-Göran Wågberg, VTI,
- the other members of the project consultative group Hans Cedermark / Håkan Westerlund, CDU, for their commitment, Jaro Potucek / Arne Johansson, VV - and Carl-Gösta Enocksson and Johan Lang, VV, both also crucial members of the modelling and validation group,
- Jörgen Blomvall, MAI, for making my parallelisation runs on the Penta PC-cluster possible,
- NSC for granting me 48000 CPU-hours in all on the Monolith PC-cluster,
- all personnel at MAI, especially the optimisation group, for their kindness and servicemindedness, and Tommy Elfving, MAI, for his encouragement.

Linköping May, 2007
Per-Åke Andersson

## Abstract

The thesis deals with the generation of cost efficient maintenance plans for paved roads, based on database information about the current surface conditions and functional models for costs and state changes, partly developed in cooperation with Vägverket (VV, Swedish Road Administration). The intended use is in a stage of budgeting and planning, before concrete project information is available. Unlike the up to now used models, individual maintenance plans can be formulated for each segment (a homogeneous road section as to the current pavement state and paving history), in continuous state and works spaces. By using Lagrangean relaxation optimisation techniques, the special benefit/-cost-ratio constraints that VV puts on each maintenance project can be naturally mastered by dual prices for the budget constraints. The number of segments competing for budget resources is usually large. Data from VV Vägdatabank (SRA Road Database) in county Värmland were used, comprising around 9000 road segments. Due to the large data amount the implemented programs rely on parallel computation. During the thesis work, access to the PC-cluster Monolith at NSC was granted. In order to reduce optimisation run times, model \& method development was needed. By aggregating the road segments into road classes, good initial values of the dual prices were achieved. By adding new state dimensions, the use of the Markov property could be motivated. By developing a special residual value routine, the explicitly considered time period could be reduced. At solving the dual subproblem special attention was paid to the discretization effects in the dynamic programming approach. One type of study is on a sub-network, e.g. a road. Validation studies were performed on road 63 in Värmland - with promising but not satisfactory results (see below). A special model for co-ordinated maintenance considers the fine-tuned cost effects of simultaneous maintenance of contiguous road segments. The other main type of study is for a whole network. Several method types have been applied, both for solving the relaxed optimisation problems and for generating maintenance plans that fit to the budgets. For a decent discretization, the run time for the whole Värmland network is less than 80 CPU-hrs.A posterior primal heuristics reduces the demands for parallel processing to a small PC-cluster. The thesis further studies the effects of redistributing budget means, as well as turning to a transparent stochastic model - both showing modest deviations from the basic model.

Optimisation results for Värmland indicate budget levels around $40 \%$ of the actual Värmland budget as sufficient. However, important cost triggers are missing in this first model round, e.g., certain functional performance (safety), all environmental performance (noise etc.) and structural performance (e.g. bearing capacity, only modelled by an age measure). For increased credibility of PMS in general and optimisation in particular, the discrepancies should be further analysed and lead to improvements as to condition monitoring, state effect \& cost modelling and mathematical modelling \& implementation.

## Sammanfattning

I avhandlingen studeras hur kostnadseffektiva underhålls- (uh-)planer för belagd väg kan genereras, på basis av information om aktuellt vägytetillstånd och funktionella modeller för kostnads- och tillståndsförändringar, delvis utvecklade i samarbete med svenska Vägverket (VV). Tilltänkt användning är på strategisk och programnivå, innan mer detaljerad objektinformation finns att tillgå. Till skillnad från hittills använda modeller, så genereras individuella uh-planer för varje vägsegment (en homogen vägsträcka vad gäller aktuellt beläggningstillstånd och beläggningshistorik), i kontinuerliga tillstånds- och åtgärdsrum. Genom användning av Lagrangerelaxerande optimeringsteknik, så kan de speciella nytto/kostnads-kvot-villkor som VV ålägger varje uh-objekt naturligen hanteras med dualpriser för budgetvillkoren. Antalet vägsegment som konkurrerar om budgetmedlen är vanligtvis stort. Data från VV:s Vägdatabank för Värmland har använts, omfattande ca 9000 vägsegment. Genom den stora datamängden har datorprogrammen implementerats för parallellbearbetning. Under avhandlingsarbetet har projektet beviljats tillgång till Monolith PCklustret vid NSC. För att kunna reducera optimeringskörtiderna har modell- och metodutveckling varit nödvändig. Genom att aggregera vägsegmenten till vägklasser har goda startvärden på dualpriserna erhållits. Genom utvecklingen av en speciell restvärdesrutin har den explicit behandlade tidsperioden kunnat reduceras. Vid lösandet av det duala subproblemet har speciell uppmärksamhet ägnats åt de diskretiseringseffekter som uppstår i metoden dynamisk programmering. En typ av tillämpning avser ett delvägnät, exempelvis en väg. Valideringsstudier har genomförts på väg 63 i Värmland - med lovande men inte tillfredsställande resultat (se nedan). En speciell modell för samordnat uh beaktar stordriftsfördelarna vid samtidig åtgärd på en hel vägsträcka. Den andra huvudtypen av studier gäller ett helt nätverk. Flera metodtyper har tillämpats, både för att lösa de relaxerade optimeringsproblemen och för att generera uhplaner som uppfyller budgetvillkoren. För en anständig diskretisering är körtiderna för hela Värmland mindre än 80 CPU-timmar. Genom en a posteriori primal heuristik reduceras kraven på parallellbearbetning till ett litet PC-kluster. Avhandlingen studerar vidare effekterna av omfördelade budgetmedel samt en övergång till en transparent, stokastisk modell - vilka båda visar små avvikelser från basmodellen.

Optimeringsresultaten för Värmland indikerar att budgetnivåer på ca $40 \%$ av Värmlands verkliga uh-budget är tillräckliga. Dock saknas viktiga kostnadsdrivande faktorer i denna första modellomgång, exempelvis vissa funktionella prestanda (säkerhet), all miljöpåverkande prestanda (buller etc.) och strukturell prestanda (ex.vis bärighet, som enbart modelleras via ett åldersmått). För ökad tilltro till PMS i allmänhet och optimering i synnerhet, bör avvikelserna analyseras ytterligare och leda till förbättringar vad gäller tillståndsmätning, tillståndseffekt- \& kostnadsmodellering samt matematisk modellering \& implementering.

## Contents

1 Introduction
1.1 Maintenance problem ..... 1
1.2 Optimisation ..... 3
1.2.1 Simplified optimisation model ..... 4
1.2.2 Lagrangean relaxation ..... 4
1.2.3 Dual optimisation by subgradient methods ..... 6
1.2.4 Dual optimisation by the Dantzig-Wolfe method ..... 7
1.2.5 Subproblem solving by dynamic programming ..... 9
1.3 Program system ..... 11
1.4 Earlier studies ..... 15
1.4.1 Systems based on Markov decision processes (MDPs) ..... 15
1.4.2 The Arizona model ..... 16
1.4.3 The NOS-based PM-systems in Kansas, Alaska and Finland ..... 17
1.4.4 Swedish PMS-based optimisation ..... 18
1.4.5 PMS-systems in Denmark and Norway ..... 18
1.4.6 Other optimisation techniques ..... 18
1.4.7 Decision support systems and integration ..... 19
1.4.8 Survey articles and implementation experience ..... 20
1.5 Study aim and outline ..... 21
1.5.1 Aim of study ..... 21
1.5.2 Outline of the thesis ..... 21
2 Initial study
2.1 Introduction ..... 23
2.2 Model ..... 24
2.2.1 Problem description ..... 24
2.2.2 Mathematical formulation ..... 25
2.3 Method ..... 26
2.3.1 Main procedure ..... 26
2.3.2 Subproblem characteristics ..... 26
2.3.3 Start routine ..... 28
2.3.4 Primal Feasibility Heuristics ..... 29
2.4 Application: PMS- and HIPS-based data ..... 29
2.4.1 Data conversion to OPM ..... 30
2.4.2 Discretization and interpolation experiments ..... 31
2.4.3 Some results ..... 33
2.5 Outlook ..... 36
3 Main study: Problem description
3.1 Road data ..... 37
3.1.1 Road segments ..... 37
3.1.2 Road classes ..... 38
3.1.3 Road constants and pavement data ..... 38
3.2 Pavement state ..... 39
3.2.1 State variables ..... 39
3.2.2 State limits ..... 40
3.2.3 IRI-value limits ..... 40
3.2.4 Rut depth limits ..... 40
3.2.5 Age limits ..... 40
3.3 Maintenance ..... 41
3.3.1 Works types ..... 41
3.3.2 Works extent: Layer thickness ..... 41
3.3.3 Restrictions on layer thickness ..... 42
3.4 Traffic effects and costs ..... 42
3.4.1 Travel time cost ..... 42
3.4.2 Vehicle operating cost ..... 42
3.5 Maintenance effects (state transitions) ..... 43
3.5.1 IRI-value immediately after a major maintenance operation, IRI $_{\text {after }}$ ..... 43
3.5.2 IRI-deterioration rate after a major maintenance operation, $\Delta I R I_{\text {after }}$ ..... 43
3.5.3 Rut depth immediately after a major maintenance operation, $R D_{\text {after }}$ ..... 44
3.5.4 Deterioration rate of rut depth after a major maint. operation, $\triangle R D_{\text {after }}$ ..... 44
3.5.5 Age immediately after a major maintenance operation, Age after ..... 44
3.5.6 Updated IRI-value, rut depth and age after one year of routine maint. ..... 45
3.6 Maintenance costs ..... 45
3.6.1 Fixed major maintenance costs ..... 46
3.6.2 Variable major maintenance costs ..... 46
3.6.3 Cost of routine maintenance ..... 47
3.7 General data ..... 47
3.7.1 Interest rate ..... 47
3.7.2 Capital scarcity factor ..... 47
3.8 Stochastic submodels ..... 48
3.8.1 Interpolation of state limits ..... 48
3.8.2 Stochastic state transitions ..... 50
3.9 Discussion ..... 52
3.9.1 Limitation ..... 52
3.9.2 Homogeneous segments ..... 53
3.9.3 Robustness ..... 53
3.9.4 Data corrections ..... 54
4 Main study: Basic model and subnet results ..... 56
4.1.1 Return rate restrictions ..... 56
4.1.2 Segment oriented problem ..... 57
4.1.3 Road class oriented problem ..... 58
4.2 Method ..... 60
4.2.1 Dual optimisation ..... 60
4.2.2 Dual subproblem solving ..... 62
4.2.3 Primal heuristics ..... 64
4.2.4 Optimisation of works extent ..... 68
4.2.5 Residual values ..... 70
4.3 Implementation ..... 71
4.3.1 Running on a PC-cluster ..... 71
4.3.2 Computer memory ..... 71
4.3.3 Absolute and relative state limits ..... 72
4.3.4 Discretization errors ..... 72
4.3.5 Reservations as to traffic evolution ..... 76
4.4 Case study: Värmland ..... 77
4.4.1 General results ..... 77
4.4.2 Budget runs ..... 78
4.5 Case study: road 63 ..... 79
4.5.1 Run strategy ..... 79
4.5.2 Results ..... 79
5 Residual values
5.1 Background ..... 85
5.2 Models ..... 86
5.2.1 Residual values model ..... 86
5.2.2 Length distributions model ..... 89
5.2.3 Markov chains ..... 93
5.3 Method ..... 94
5.3.1 Dual optimisation ..... 94
5.3.2 Value iteration ..... 98
5.3.3 Policy iteration ..... 98
5.3.4 LP iteration ..... 99
5.3.5 Newton iteration ..... 102
5.3.6 Length distributions ..... 104
5.4 Results ..... 105
6 Co-ordinated maintenance
6.1 Mathematical formulation ..... 111
6.1.1 Indices and data ..... 112
6.1.2 Model ..... 113
6.2 Method ..... 114
6.2.1 Dual optimisation ..... 114
6.2.2 Primal heuristics ..... 116
6.3 Case study: road 63 ..... 121
6.3.1 Run strategy ..... 121
6.3.2 Results ..... 122
6.4 Discussion ..... 126
6.5 An alternative dual price updating method ..... 126
7 Stochastic model
7.1 Introduction ..... 131
7.2 Model ..... 132
7.2.1 Stochastics ..... 132
7.2.2 Truncation of distributions ..... 133
7.2.3 Problem formulation ..... 133
7.3 Method ..... 134
7.3.1 Computation of stochastic elements ..... 134
7.4 Results ..... 138
7.5 Discussion ..... 139
8 Redistributable budget
8.1 Model ..... 141
8.2 Method ..... 142
8.2.1 Dual optimisation ..... 142
8.2.2 Primal heuristics ..... 149
8.3 Results ..... 155
9 Methods for road network application
9.1 Parallelisation ..... 161
9.2 Subgradient method with full and partial updating ..... 163
9.2.1 Input data precision, discretization errors and run strategy ..... 163
9.2.2 Full and partial updating ..... 165
9.2.3 Camerini-Fratta-Maffioli (CFM) modification ..... 168
9.3 Interpolation improvements ..... 169
9.3.1 Multi-linear-quadratic interpolation ..... 169
9.3.2 Double grid interpolation ..... 173
9.4 Dantzig-Wolfe decomposition ..... 183
9.4.1 Master problem ..... 183
9.4.2 Implicit simplex pivoting ..... 188
9.4.3 Results ..... 195
9.5 Primal heuristics ..... 196
9.5.1 Problem ..... 197
9.5.2 Method ..... 198
9.5.3 Results ..... 201
9.6 Conclusions ..... 202
10 References ..... 203
A1 Appendix 1
A1.1 Road user costs ..... 211
A1.1.1 Travel time costs ..... 211
10.1 Vehicle operating costs ..... 211
A1.1.2 Fuel consumption ..... 212
A1.1.3 Tyre consumption ..... 212
A1.1.4 Parts consumption ..... 213
A1.1.5 Labour hours ..... 213
A1.1.6 Capital costs ..... 214

## 1 Introduction

### 1.1 Maintenance problem

What pavement properties are important: For a road user (traveller and vehicle)? For the road administrator (agency)? In [Ihs and Magnusson (2000)] a number of quantities are identified, e.g. different kinds of surface unevenness, friction and surface texture, bearing capacity. The most important descriptor of the pavement state is longitudinal unevenness. The influence of its different "wave lengths" are weighed into a collective single value of the state variable IRI (International Roughness Index, expressed in $\mathrm{mm} / \mathrm{m}$ ). Its strange unit is because IRI measures a standardised passenger's vertical movement (mm) in a standardised chair in a standardised vehicle, driving a length unit $(\mathrm{m})$ at $80 \mathrm{~km} / \mathrm{h}(=\mathrm{kph})$, see e.g. [Öberg (2001)]. The IRI-value is a 20 m average and corrected for unwanted influences from a hilly length profile of the road. Since its expected time evolution - especially the annual change, the degradation rate - can be pretty well predicted from historical data, see e.g. [Lang (2007)], it is important to regularly collect such information - and it makes studies like ours meaningful.

In Sweden the paved roads are dominated by flexible pavement, i.e. layers of bitumen based asphalt products. The road surface is degraded by time, due to, e.g., climate, pavement age, traffic load and tyre studs. To keep the road standard intact, every year some kind of maintenance is needed. The default option, routine (=minor) maintenance, e.g. pothole repair and minor crack sealing, eventually becomes insufficient, since the degradation will continue at increasing costs. When is major maintenance, i.e. resurfacing, cost-efficient? What kind of paving is the most cost-efficient? Where should the limited budget resources be spent?

For an analysis we have to consider the effects of different maintenance works on a number of state variables characterising the pavement conditions. Ideally all such parameter values are easily available. In practice the measuring of, e.g., bearing capacity needs more personnel and time. In Sweden, Vägverket (VV, Swedish Road Administration) is responsible for the State roads. The automatic Laser-RST VV-measurement programme, see [Forsberg and Göransson (2000)], regularly scanning 150000 road segments of homogeneous pavement conditions, is presently confined to texture and unevenness along and across the road surface. By assuming that historical RST-data are relevant for the future development, we can apply effect models describing what pavement state will result from different maintenance works.

According to internationally established models, see e.g. [HDM-4 (2000)], the pavement state determines different traffic effects on vehicles and road users. By applying official governmental exchange rates, see [Effektsamband 2000 (2001a)], between e.g. travel time and society cost (in "society-SEK", adjusted to 120 SEK/h for passenger cars and 150 SEK/h for lorries), we can summarise the various traffic effects as a traffic cost (= road user cost, in societySEK), i.e. the additional cost that pavement imperfections will cause, in comparison to an ideal state.

Since the maintenance costs (= agency costs) vary between the paving projects, we must rely on models of the cost values for different pavement states and different works types and
extents. The overall problem is how to balance the traffic cost and the maintenance cost (in road agency-SEK), as Fig 1.1 illustrates (cf. [Lindberg et al (2001)]). The two traffic cost curves are for two different road segments or road classes.


Figure 1.1 Traffic and maintenance costs for different average pavement states.
This balance, defining the optimal average state, is unique for each road segment, since the cost curves will look different. This follows from the varying road characteristics, both static, e.g. traffic and width, and dynamic, e.g. state. For a final choice in Fig 1.1, a translation between the cost scales is needed. Such a translation is better illustrated in a concrete choice situation, see Fig 1.2. A road segment is in a given state at the start of year $t$. Should major maintenance be performed in year $t$, initially improving the state, or be postponed? The advantage of a major operation (index 1 ) in yr $t$ over routine maintenance ( 0 ) is a reduced immediate traffic cost $\left(f_{1}<f_{0}\right)$. The disadvantage is an increased immediate maintenance cost ( $g_{1}>g_{0}$ ). However, it would be unfair to let $f_{0}, f_{1}$ only stand for the $1-\mathrm{yr}$ traffic costs, as measured from the middle of year $t$ (where the major maintenance is assumed to be performed), since the costs will differ also thereafter. On the other hand, just letting $f_{0}, f_{1}$ summarise all the discounted future traffic costs would still be unfair since, if optimal works options are chosen every year, a major maintenance may be performed the next year $t+1$, as the upper dotted curve in Fig 1.2 illustrates - implying low future traffic costs "for nothing". Therefore also the differences between the future maintenance costs should be considered, letting $g_{0}, g_{1}$ summarise all the discounted future maintenance costs. We arrive at a comparison between the traffic and maintenance costs, by the net present values (NPVs) of the summed up differences of discounted costs, between major and minor maintenance. Using a 1 -yr discount factor $d$ and assuming that $f_{0}, f_{1}$ are NPVs at the start of year $t+1$ and $g_{0}, g_{1}$ half a year earlier, a decisive quantity is the (incremental) benefit/cost ratio (BCR), see e.g. [HDM-4 (2000), vol. 1, pp G1-20, G1-24],

$$
\begin{equation*}
\frac{\sqrt{d} \cdot\left(f_{0}-f_{1}\right)}{g_{1}-g_{0}} \tag{1.1}
\end{equation*}
$$

For the immediate major maintenance to be chosen, the ratio must meet the governmental
BCR-lower bound, presently $v_{\mathrm{BCR}}=1.2439$, i.e. the discounted future traffic cost savings
must exceed the immediate extra (tax financed) maintenance costs by around $\frac{1}{4}$. In practice the current budget situation may put further restrictions on the ratio. Such a lower bound may serve as the wanted exchange rate between agency-SEK and society-SEK in Fig 1.1. The best major maintenance option is the one with the greatest benefit/cost ratio. Comparing two identical road segments, one with twice the traffic load of the other, one and the same works option will produce a doubled traffic cost and benefit/cost ratio for the former segment (cf. the dotted curve in Fig 1.1). For this reason more agency-SEK will in general be spent on segments with heavy traffic - moving the balance in Fig 1.1 to the left. Extremely bad pavement conditions on segments with sparse traffic may be prevented by the introduction of state restrictions.


Figure 1.2 State evolutions for two works options: immediate and postponed major maintenance. Traffic costs $f$, maintenance costs $g$.

However, we are discontent with (1.1) as well, since it does not focus on the 1 -yr postponing decision alternative and does not account for the expected future budget scarcity. Therefore we will apply a slightly different BCR-measure below (cf. Ch 4).

### 1.2 Optimisation

In Sec 1.2.1 we present a simple model, of the same principal structure as the models we will study in the following chapters, and list some problem properties. In the following subsections we give an introductory exposition of the main methods that we later will apply: Sec 1.2.2 includes a reformulation of the original optimisation problem as a Lagrangean dual problem and Secs 1.2.3-1.2.4 are devoted to two methods, the subgradient and Dantzig-Wolfe techniques, for solving the dual problem. In Sec 1.2 .5 we introduce our main method, dynamic programming, for solving the subproblems that arise, linked to the Lagrangean dual.

### 1.2 Optimisation

### 1.2.1 Simplified optimisation model

In order to illustrate various method principles, we will refer to the simplified optimisation problem

$$
\begin{align*}
& \underset{\mathbf{x}_{1}, \mathbf{x}_{2}}{\operatorname{minimise}} F\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right):=F_{1}\left(\mathbf{x}_{1}\right)+F_{2}\left(\mathbf{x}_{2}\right)  \tag{1.2a}\\
& \text { subject to }\left\{\begin{array}{cc}
G_{11}\left(\mathbf{x}_{1}\right)+G_{12}\left(\mathbf{x}_{2}\right) \leq b_{1} \\
G_{21}\left(\mathbf{x}_{1}\right)+G_{22}\left(\mathbf{x}_{2}\right) \leq b_{2} \\
\mathbf{x}_{j} \in \mathcal{X}_{j} & j=1,2
\end{array}\right. \tag{1.2b}
\end{align*}
$$

Here $\mathbf{x}_{1}, \mathbf{x}_{2}$ are variable vectors, in general mixed integer, i.e. including both discrete (integer) variables and continuous (real) variables as vector components. Introducing vector notations $\mathbf{b}=\left(b_{t}\right)_{t=1}^{2}$ and $\mathbf{G}=\left(G_{t}\right)_{t=1}^{2}, G_{t}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right):=G_{t 1}\left(\mathbf{x}_{1}\right)+G_{t 2}\left(\mathbf{x}_{2}\right) \quad t=1,2$, our aim is to find $\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)$-values such that the objective function $F$ is brought to the minimum, whereas $\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)$ have to satisfy all the vector-valued constraints $\mathbf{G}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right) \leq \mathbf{b}$, and $\mathbf{x}_{j} \in \mathcal{X}_{j}$ for $j=1,2$. In this thesis $F$ often stands for total traffic cost and (1.2b) - (1.2c) for $T$ budget constraints, one per year, with $\mathbf{b}$ denoting the maintenance budgets and $\mathbf{G}$ the total maintenance costs. The variables $\mathbf{x}=\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)$, generalised to $\mathbf{x}=\left(\mathbf{x}_{1}, \mathbf{x}_{2}, \ldots, \mathbf{x}_{J}\right)=\left(\mathbf{x}_{j}\right)_{j=1}^{J}$, often refer to pavement state and maintenance works at different times, for $J$ different (categories of) road segments. The short-form $\mathbf{G}(\mathbf{x}) \leq \mathbf{b}$ is formally independent of $T, J$.

Any optimisation problem, even where the constraint relations are equalities ( $=$ ) or inequalities $(\geq)$ or the purpose is maximisation, can be transferred into the form (1.1). The problem (1.2) has some structure, insofar as the constraints $\mathbf{G}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right) \leq \mathbf{b}$ involve all the variables, whereas $\mathbf{x}_{j} \in \mathcal{X}_{j}$ for $j=1,2$ concern one variable vector each. Moreover, both the objective $F$ and the coupling-constraint functions $G_{t}$ in (1.2b) - (1.2c) are additive in ( $\mathbf{x}_{1}, \mathbf{x}_{2}$ ). In this thesis the sets $\left(X_{j}\right)_{j}$ may refer to given state bounds and rules for the transition between states as a consequence of different maintenance works.

The optimisation problem (1.2) would be convex, if the objective $F$ is a convex function of $\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)$ and if the set of feasible solutions, i.e. the ( $\mathbf{x}_{1}, \mathbf{x}_{2}$ ) -region where all constraints (1.2b) - (1.2d) are satisfied, is a convex set. For convex problems the task of finding an optimal solution is relatively simple. As soon as any discrete variable can take two different values in the feasibility set the problem becomes non-convex.

### 1.2.2 Lagrangean relaxation

Our maintenance optimisation problem, as we will formulate it, is a mixed integer nonlinear problem with millions of variables, non-convex due to the maintenance cost functions $\mathbf{G}$ in (1.2) as well. In order to solve such complex problems, it is necessary to find and use the problem structure. The recurring method principle we will apply in different shapes is Lagrangean relaxation. The aim is to transfer a set of "difficult" constraints, e.g. (1.2b) (1.2c), from absolute bounds $\mathbf{b}$ into a milder, relaxed form, where constraint $t$ violations
$G_{t}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)>b_{t}$ are penalised instead of forbidden and the resulting optimisation problem is easier to solve, even repeatedly. For convex problems it can be shown, see e.g. [Minoux (1986), p 204], that the relaxed (dual) form is equivalent to the original (primal) formulation for certain penalty values, the shadow prices $\mathbf{v}$, in the following sense: a dual optimum exists, directly linked to a primally feasible solution - the optimum of the original problem.

For problem (1.2) a natural simplification is to relax the coupling constraints (1.2b) - (1.2c), by introducing one dual variable $v_{t}$ for each constraint component $t, \mathbf{v}=\left(v_{t}\right)_{t}$, and get

Dual: $\underset{\boldsymbol{v} \geq \mathbf{0}}{\operatorname{maximise}} \varphi(\mathbf{v})$
Dual subproblem: $\varphi(\mathbf{v}):=\min _{\mathbf{x}_{1}, \mathbf{x}_{2}}\left(F\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)+v_{1} \cdot\left[G_{1}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)-b_{1}\right]+v_{2} \cdot\left[G_{2}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)-b_{2}\right]\right) \equiv$ $\equiv \min _{\mathbf{x}}\left(F(\mathbf{x})+\mathbf{v}^{T}[\mathbf{G}(\mathbf{x})-\mathbf{b}]\right)$
subject to $\mathbf{x}_{j} \in \mathcal{X}_{j} \quad j=1,2$.
Here the upper index ${ }^{T}$ denotes transposing and $\mathbf{v}^{T} \mathbf{G}$ is a scalar product. Irrespective of the primal structure, the Lagrangean dual (1.3) is (equivalent to) a convex problem, see e.g. [ibid, p203], i.e. relatively simple to solve. Since both $F$ and $\mathbf{G}$ are additive, the dual subproblem (1.4) is separable into one subproblem for each variable vector $\mathbf{x}_{j}$, i.e. for a fixed $\mathbf{v}$ we can (easier) solve one subproblem for each $\mathbf{x}_{j}-$

$$
\begin{align*}
\text { Dual } j \text {-subproblem: } & \varphi_{j}(\mathbf{v}):=\min _{\mathbf{x}_{j}}\left(F_{j}\left(\mathbf{x}_{j}\right)+v_{1} \cdot G_{1 j}\left(\mathbf{x}_{j}\right)+v_{2} \cdot G_{2 j}\left(\mathbf{x}_{j}\right)\right)  \tag{1.5}\\
& \text { subject to } \mathbf{x}_{j} \in \mathcal{X}_{j}
\end{align*}
$$

- and then sum up the two contributions to the dual objective value

$$
\varphi(\mathbf{v})=\varphi_{1}(\mathbf{v})+\varphi_{2}(\mathbf{v})-\mathbf{v}^{T} \mathbf{b}
$$

This kind of simplification motivates the Lagrangean relaxation as an adequate method here.
In the dual subproblem (1.4), $\left(v_{t}\right)_{t}$ act as weights for the constraint functions $\left(G_{t j}\right)_{t, j}$ in balance with the original objective $F$, i.e. they take the role of exchange rates between the maintenance costs (in agency-SEK) and the traffic costs (in society-SEK). The shadow price $v_{t}$ is a capital scarcity factor: the higher shadow price, the higher benefit/cost ratio (1.1) is needed for a (major) maintenance project to become realised, and in fact $v_{t}$ normally corresponds to the BCR for the "last accepted" project, i.e. the lowest used ratio. (Therefore $\left(v_{t}\right)_{t}$ are our means for a modification of (1.1) - anticipated at the end of Sec 1.1.)

The advantages of relaxation are easier problems to solve. The disadvantage is a necessity to solve the relaxed problem for a sequence of trial $\mathbf{v}$-values, ideally until the optimum $\mathbf{v}=\mathbf{v}^{*}$ for (1.3) is found. Thus in iteration $i=0,1, \ldots$ we will solve a dual subproblem for fixed $\mathbf{v}=\mathbf{v}^{(i)}$, leading to the dual objective value $\varphi\left(\mathbf{v}^{(i)}\right)$ in (1.3). In every iteration $i$ we may also use the subproblem solution for the generation of a primally feasible solution, corresponding to the primal objective value $F=F_{\text {prim }}^{(i)}$ in (1.2). For non-convex primal problems like ours, a duality gap may exist, i.e. a true difference between the optimal primal and dual objective

### 1.2 Optimisation

values. But this gap is always nonnegative, see e.g. [ibid, p 203], meaning that by solving the dual subproblem repeatedly, the optimal primal objective value $F^{*}$ fulfils

$$
\max _{i} \varphi\left(\mathbf{v}^{(i)}\right) \leq F^{*} \leq \min _{i} F_{\text {prim }}^{(i)}
$$

i.e. at every iteration we can present an error estimate of the unknown $F^{*}$, as trapped between the best primal and dual values found so far (cf. the figure on the thesis cover). Since this also means that $F^{*}$ is not directly linked to $\mathbf{v}^{*}$, we may even generate a primal optimum during the iterative process, before $\mathbf{v}^{*}$ is reached.

### 1.2.3 Dual optimisation by subgradient methods

Letting $\mathbf{v} \geq \mathbf{0}$ vary in the dual (1.3), to simplify the presentation let us assume that the dual subproblem (1.4) has just three different optimal solutions $\mathbf{x}=\mathbf{x}^{(i)}$, for $i=1,2,3$. Thus as $\mathbf{v}$ varies, the function values $F\left(\mathbf{x}^{(i)}\right), \mathbf{G}\left(\mathbf{x}^{(i)}\right)$ are constant in $\mathbf{v}$-regions $\mathcal{V}_{i}$ for $i=1,2,3$ - see Fig 1.3. The dual objective turns into

$$
\varphi=\varphi(\mathbf{v})=F\left(\mathbf{x}^{(i)}\right)+\mathbf{v}^{T}\left[\mathbf{G}\left(\mathbf{x}^{(i)}\right)-\mathbf{b}\right] \text { for } \mathbf{v} \in \mathcal{V}_{i},
$$

i.e. is fully characterised by three planes in the 3D $(\mathbf{v}, \varphi)$-space. The dotted lines in Fig 1.3 represent level curves of constant $\varphi$-values, and we may visualise the $3^{\text {rd }} \varphi$-dimension: the function $\varphi=\varphi(\mathbf{v})$ is an irregular tetrahedron with its top at $\mathbf{v}^{*}$. In the following chapters the function surfaces $\varphi=\varphi(\boldsymbol{v})$ will have many facets, not always planar.


Figure 1.3 Illustrative example of subgradients and affine majorants for a dual.

Consider the point $\mathbf{v}_{\mathrm{A}}$ in the interior of $\mathcal{V}_{1}$. Here the $\varphi$-gradient $\nabla \varphi$ is well-defined, with the partial $\varphi$-derivatives as components:

$$
\frac{\partial \varphi}{\partial v_{t}}=G_{t}\left(\mathbf{x}^{(1)}\right)-b_{t} \quad t=1,2 \quad \Leftrightarrow \quad \nabla \varphi\left(\mathbf{v}_{\mathrm{A}}\right)=\mathbf{G}\left(\mathbf{x}^{(1)}\right)-\mathbf{b}
$$

pointing orthogonally to the level line in the direction of steepest ascent, as the arrow in Fig 1.3 illustrates. By stepping along $\nabla \varphi\left(\mathbf{v}_{\mathrm{A}}\right)$ from $\mathbf{v}_{\mathrm{A}}$ we expect the $\varphi$-value to change (initially) as if we move on the tetrahedron plane valid in $\mathcal{V}_{1}$ :

$$
\begin{equation*}
\varphi(\mathbf{v})=\varphi\left(\mathbf{v}_{\mathrm{A}}\right)+\nabla \varphi\left(\mathbf{v}_{\mathrm{A}}\right)^{T}\left(\mathbf{v}-\mathbf{v}_{\mathrm{A}}\right) \tag{1.6}
\end{equation*}
$$

However, at $\mathbf{v}_{\mathrm{B}}$ on the border between $\mathcal{V}_{1}$ and $\mathcal{V}_{2}$ the gradient is undefined: if we approach $\boldsymbol{v}_{\mathrm{B}}$ from the left (passing $\left.\mathbf{v} \in \mathcal{V}_{1}\right), \nabla \varphi(\mathbf{v})$ is the same as before but from the right $\left(\mathbf{v} \in \mathcal{V}_{2}\right)$ another level line and gradient direction are valid. In such a case the gradient concept is extended to subgradients $\gamma$, see e.g. [Minoux (1986), p 14], defined by:

$$
\begin{equation*}
\varphi(\mathbf{v}) \leq \varphi\left(\mathbf{v}_{\mathrm{B}}\right)+\boldsymbol{\gamma}^{T}\left(\mathbf{v}-\mathbf{v}_{\mathrm{B}}\right) \quad \forall \mathbf{v} \geq \mathbf{0} \tag{1.7}
\end{equation*}
$$

(where $\forall$ stands for "for every"), i.e. a linearization according to $\gamma$ from $v_{\mathrm{B}}$ to $\boldsymbol{v}$ is nowhere lower than the correct $\varphi$-value. Since the tetrahedron surface $\varphi=\varphi(\mathbf{v})$ for any $\mathbf{v}$ is the minimum of the three surface planes, each defined analogously to (1.6), we realise that both arrow directions at $\mathbf{v}_{\mathrm{B}}$ are subgradients, as well as all directions in between them, the "sector arc" in Fig 1.3, i.e. the convex combinations of the extreme (arrow) directions. Moreover, the precise meaning of a "convex" dual problem is that $\varphi=\varphi(\mathbf{v})$ is a concave function. At $\mathbf{v}_{\mathrm{A}}-$ where the function is differentiable - it follows the definition (1.7) that the gradient $\gamma=\nabla \varphi\left(\mathbf{v}_{\mathrm{A}}\right)$ is the only subgradient direction.

From Fig 1.3 we can see that all the subgradients point more or less towards the tetrahedron top, cf. [ibid, p 17]. A subgradient method, see e.g. [ibid, p 109], means that the next dual iteration point $\mathbf{v}^{(i+1)}$ is chosen by taking a step along any subgradient direction from the current dual iteration point $\mathbf{v}^{(i)}$. If the step lengths are chosen with some care, e.g. according to [Polyak (1966)], cf. [Minoux (1986), p 110], the subgradient method is guaranteed to converge asymptotically, which for a compact $\mathbf{v}$-domain can be written $\varphi\left(\mathbf{v}^{(i)}\right) \rightarrow \max _{\mathbf{v} \geq \mathbf{0}} \varphi(\mathbf{v}) \equiv \varphi\left(v^{*}\right)$ as $i \uparrow \infty$, i.e. it approaches the top value in Fig 1.3.

### 1.2.4 Dual optimisation by the Dantzig-Wolfe method

The "pure" subgradient method introduced in Sec 1.2 .3 builds on the local problem properties $(F, \mathbf{G})$ at the current dual iterate $\mathbf{v}^{(i)}$. In the "pure" Dantzig-Wolfe approach, see [Dantzig (1963), p 448], the corresponding information from all earlier iterations is used. For each dual iteration $i$ a subproblem optimum $\mathbf{x}^{(i)}$ (observe the new meaning of the notation) and the corresponding costs $F\left(\mathbf{x}^{(i)}\right), \mathbf{G}\left(\mathbf{x}^{(i)}\right)$ define a subgradient $\gamma:=\mathbf{G}\left(\mathbf{x}^{(i)}\right)-\mathbf{b}$ and a corresponding affine majorant $\hat{\varphi}$, defined by the linearization

$$
\begin{align*}
& \hat{\varphi}\left(\mathbf{v} ; \mathbf{v}^{(i)}\right):=\varphi\left(\mathbf{v}^{(i)}\right)+\boldsymbol{\gamma}^{T}\left(\mathbf{v}-\mathbf{v}^{(i)}\right)= \\
& =\left(F\left(\mathbf{x}^{(i)}\right)+\mathbf{v}^{(i)^{T}}\left(\mathbf{G}\left(\mathbf{x}^{(i)}\right)-\mathbf{b}\right)\right)+\gamma^{T}\left(\mathbf{v}-\mathbf{v}^{(i)}\right)=  \tag{1.8}\\
& =F\left(\mathbf{x}^{(i)}\right)+\mathbf{v}^{T}\left(\mathbf{G}\left(\mathbf{x}^{(i)}\right)-\mathbf{b}\right)
\end{align*}
$$

### 1.2 Optimisation

and satisfying (cf. (1.7))

$$
\varphi(\mathbf{v}) \leq \hat{\varphi}\left(\mathbf{v} ; \mathbf{v}^{(i)}\right) \quad \forall \mathbf{v} \geq \mathbf{0} .
$$

Since every iteration $i$ generates an affine majorant, the least upper bound $\min _{i} \hat{\varphi}\left(\mathbf{v} ; \mathbf{v}^{(i)}\right)$ at $\mathbf{v}$ is decisive for our linearization of the surface $\varphi=\varphi(\mathbf{v})$. Due to the additive cost functions in (1.2) the corresponding majorant property holds in each $j$-subproblem (1.5), which means that the iteration providing the least upper bound can be different for different $j$ 's, written $i=i_{j}$. This means a tighter upper bound than by forcing all $i_{j}$ 's to be equal. For our simplified problem (1.2) and the corresponding Lagrangean dual (1.3) we get the least upper bound at $\mathbf{v}$ as

$$
\begin{align*}
& \varphi(\mathbf{v}) \leq \min _{i_{1}, i_{2}}\left[F_{1}\left(\mathbf{x}_{1}{ }^{\left(i_{1}\right)}\right)+F_{2}\left(\mathbf{x}_{2}{ }^{\left(i_{2}\right)}\right)+\mathbf{v}^{T}\left(\mathbf{G}_{1}\left(\mathbf{x}_{1}{ }^{\left(i_{1}\right)}\right)+\mathbf{G}_{2}\left(\mathbf{x}_{2}{ }^{\left(i_{2}\right)}\right)-\mathbf{b}\right)\right]= \\
& =\min _{i_{1}}\left[F_{1}\left(\mathbf{x}_{1}{ }^{\left(i_{1}\right)}\right)+\mathbf{v}^{T} \mathbf{G}_{1}\left(\mathbf{x}_{1}{ }^{\left(i_{1}\right)}\right)\right]+\min _{i_{2}}\left[F_{2}\left(\mathbf{x}_{2}{ }^{\left(i_{2}\right)}\right)+\mathbf{v}^{T} \mathbf{G}_{2}\left(\mathbf{x}_{2}{ }^{\left(i_{2}\right)}\right)\right]-\mathbf{v}^{T} \mathbf{b} . \tag{1.9}
\end{align*}
$$

Fig 1.3 may illustrate this linearization as well, after a reinterpretation where we assume that the true function surface $\varphi=\varphi(\boldsymbol{v})$ is unknown. Instead let the level curves in Fig 1.3 represent the three decisive majorants (supporting planes) that can be formed from (1.9) after three dual iterations. The Dantzig-Wolfe (D-W) method now chooses the next dual iterate at the top $\mathbf{v}^{*}$, according to this linearization. New information is added for each iteration, in non-degenerate cases cutting off the top in the previous linearization. Eventually a detailed facet-like structure of supporting planes will cover the true function surface $\varphi=\varphi(\mathbf{v})$. Unlike the more robust subgradient technique, the convergence of the $\mathrm{D}-\mathrm{W}$ method depends on the primal problem structure. For convex problems asymptotic convergence is guaranteed, for a compact (closed and bounded) $\boldsymbol{v}$-domain written $\varphi\left(\mathbf{v}^{(i)}\right) \rightarrow \max _{\mathbf{v} \geq \mathbf{0}} \varphi(\mathbf{v})$ as $i \uparrow \infty$, see [ibid, p 477]. For nonconvex problems, like the ones in the thesis, no general theorem applies. Moreover, the addition of new information for each iteration and the many possible iteration combinations (here of $\left(i_{1}, i_{2}\right)$ ) lead to a complex master problem for the identification of the top:[]

```
maximise \varphi
    v,\varphi
```

subject to $\left\{\begin{array}{c}\varphi \leq\left[F_{1}\left(\mathbf{x}_{1}{ }^{\left(i_{1}\right)}\right)+\mathbf{v}^{T} \mathbf{G}_{1}\left(\mathbf{x}_{1}{ }^{\left(i_{1}\right)}\right)\right]+\left[F_{2}\left(\mathbf{x}_{2}{ }^{\left(i_{2}\right)}\right)+\mathbf{v}^{T} \mathbf{G}_{2}\left(\mathbf{x}_{2}{ }^{\left(i_{2}\right)}\right)\right]-\mathbf{v}^{T} \mathbf{b} \quad \forall i_{1}, i_{2} \\ \mathbf{v} \geq \mathbf{0}\end{array}\right.$

Since the master problem is a linear programming (LP) problem, it can be solved in a finite number of LP-iterations by the simplex method, see [ibid, p 120]. We will surround the current iterate $\mathbf{v}^{(i)}$ with some kind of box, confining the feasible LP-solutions to a compact set. The LP-optimum is the wanted top $\mathbf{v}^{*}$ (in the box) for the current dual iteration. The next dual iteration $i+1$ means a resolving of the dual subproblem, now for $\mathbf{v}^{(i+1)}=\mathbf{v}^{*}$. Normally the subproblem optimum means that a new supporting (hyper-)plane (=affine majorant) is generated - a cut to be added to the master constraints. But if a dual iteration leads to the very same subproblem optimum as the previous iteration did, stagnation occurs - and a dual optimum has been reached, see [ibid, p 475].

### 1.2.5 Subproblem solving by dynamic programming

Consider the dual $j$-subproblem (1.5). We will reformulate this as (1.10) below. Let the (road class or segment) index $j$ be implicit. We interpret the objective term $F(\mathbf{x})$ in (1.5) as the total traffic cost (normally discounted), summing up contributions $f_{t}\left(x_{t}\right)$ from different years $t=1,2$, for the pavement states $x_{t}$ at the start of year $t$, and including the residual value $\varphi_{3}\left(x_{3}\right)$, comprising the optimal future costs from state $x_{3}$ at the end of the explicitly considered 2-yr period. (A variant of the dynamic programming method - to be introduced here - is utilised also for the determination of residual values, as sketched in Sec 1.5.2 below and further studied in Ch 5.) Moreover, in (1.5) we interpret $G_{t}(\mathbf{x})$ as the maintenance cost in year $t$, depending on the current state $x_{t}$ as well as the works $u_{t}$ performed in year $t$, therefore rewritten as $g_{t}\left(x_{t}, u_{t}\right)$ in (1.10). The constraints $\mathbf{x} \in \mathcal{X}$ in (1.5) stand for

- the transitions between consecutive states, with $x_{t+1}$ dependent on the previous $x_{t}$ as well as the works $u_{t}$, therefore rewritten as $x_{t+1}=h\left(x_{t}, u_{t}\right)$,
- works restrictions due to the current state, $u_{t} \in \mathcal{U}\left(x_{t}\right)$,
- a given initial state $x_{1}=x_{1}^{0}$.

In the residual value notation $\varphi_{3}\left(x_{3}\right)$ the (optimal) Lagrangean multipliers $\mathbf{v}$ for the succeeding time periods are permanently fixed and implicit. By expanding this notation to the explicitly considered time periods in the Lagrangean dual, where $\mathbf{v}$ varies, we let $\varphi_{t}\left(x_{t} ; \mathbf{v}\right)$ denote the minimum cost in state $x_{t}$ from year $t$ and onwards, for occasionally fixed dual prices $\mathbf{v}$. The subproblem turns into

$$
\begin{align*}
& \varphi_{1}\left(x_{1}, \mathbf{v}\right):=\min _{\left(x_{t}\right)_{t>1},\left(u_{t}\right)_{t}}\left[f_{1}\left(x_{1}\right)+f_{2}\left(x_{2}\right)+\varphi_{3}\left(x_{3}\right)+v_{1} \cdot g_{1}\left(x_{1}, u_{1}\right)+v_{2} \cdot g_{2}\left(x_{2}, u_{2}\right)\right]  \tag{1.10}\\
& \text { subject to }\left\{\begin{array}{c}
x_{2}=h\left(x_{1}, u_{1}\right) \\
x_{3}=h\left(x_{2}, u_{2}\right) \\
u_{t} \in \mathcal{U}\left(x_{t}\right) \quad t=1,2 \\
x_{1}=x_{1}^{0} .
\end{array}\right.
\end{align*}
$$

The quantities are illustrated in Fig 1.4. The black dots denote chosen states. One minor maintenance option, leading to progressive deterioration (upwards), and one or more major maintenance options are distinguished. This is a simplified picture - in practice the degradation rates vary with the works extent.


Figure 1.4 Computational quantities in a $2-\mathrm{yr}$ dual ( $j$-)subproblem solved by DP.
A naïve approach for solving the subproblem (1.10) would be to calculate the cost for each possible works combination $\left(u_{1}, u_{2}\right)$. As a general method, with many works options, the number of paths starting at $x_{1}=x_{1}^{0}$ increases exponentially with the number of considered years. The dynamic programming (here DynP; to distinguish from DP $=$ Decision Process below) approach is to consider one year at a time, in a backward iteration procedure, from the start of the last year, here $t=2$. From a fixed state $x_{2}$ we consider each works option $u_{2}$ in yr 2. Its future cost is $f_{2}\left(x_{2}\right)+v_{2} \cdot g_{2}\left(x_{2}, u_{2}\right)+\varphi_{3}\left(h\left(x_{2}, u_{2}\right)\right)$. The lowest future cost is

$$
\begin{equation*}
\varphi_{2}\left(x_{2} ; \boldsymbol{v}\right)=\min _{u_{2} \in \mathcal{U}\left(x_{2}\right)}\left[f_{2}\left(x_{2}\right)+v_{2} \cdot g_{2}\left(x_{2}, u_{2}\right)+\varphi_{3}\left(h\left(x_{2}, u_{2}\right)\right)\right] . \tag{1.11}
\end{equation*}
$$

With the lowest cost $\varphi_{2}\left(x_{2} ; \mathbf{v}\right)$ determined for each state $x_{2}$, we now consider yr 1 in the same way, getting

$$
\varphi_{1}\left(x_{1} ; \boldsymbol{v}\right)=\min _{u_{1} \in \mathcal{U}\left(x_{1}\right)}\left[f_{1}\left(x_{1}\right)+v_{1} \cdot g_{1}\left(x_{1}, u_{1}\right)+\varphi_{2}\left(h\left(x_{1}, u_{1}\right) ; \boldsymbol{v}\right)\right]
$$

In practice we only have to consider the given state $x_{1}=x_{1}^{0}$ at the start of yr 1 .
Is this all? In theory yes, in practice no. The analysis presumes that all possible states $x_{2}$ are handled in (1.11) and that the residual values $\varphi_{3}\left(x_{3}\right)$ are given for all possible $x_{3}$-values. Since the pavement states in many respects are described by continuous variables and the cost and state functions are complex, the task is in practice impossible. Instead we have to discretize the state space, calculating $\left(\varphi_{t}\left(x_{t} ; \boldsymbol{v}\right)\right)_{t}$ for a finite nodal set $\mathcal{N}$ of states $x_{t}$ and rely on interpolation between neighbouring nodal states, as soon as $h\left(x_{t}, u_{t}\right)$ leads to a non-nodal state $x_{t+1}$. Letting $\bar{\varphi}_{t+1}\left(h\left(x_{t}, u_{t}\right) ; \mathbf{v}\right)$ denote such an interpolation result, the discretized DynP backward procedure for $T$ years means that we for $t=T, T-1, \ldots, 1$ determine

$$
\begin{equation*}
\varphi_{t}\left(x_{t} ; \boldsymbol{v}\right)=\min _{u_{t} \in \mathcal{U}\left(x_{t}\right)}\left[f_{t}\left(x_{t}\right)+v_{t} \cdot g_{t}\left(x_{t}, u_{t}\right)+\bar{\varphi}_{t+1}\left(h\left(x_{t}, u_{t}\right) ; \mathbf{v}\right)\right] \quad \forall x_{t} \in \mathcal{N} . \tag{1.12}
\end{equation*}
$$

To get more precise information about the optimal maintenance plan we also perform a DynP forward iteration procedure. The optimal path is identified by stepping one year $t$ at a time
from the given initial state. For established optimal states and works $\left(x_{\tau}, u_{\tau-1}^{*}\right)_{\tau \leq t}$ up to the start of year $t$, (1.12) is applied to the one (even non-nodal) reached state $x_{t}$, in order to find the minimising works $u_{t}^{*}$ in year $t$. By using $x_{t+1}=h\left(x_{t}, u_{t}^{*}\right)$, the next step on the optimal path is determined. The procedure is illustrated by Fig 1.4, read from the left to the right.

The 1 yr-stepping DynP procedures are applicable, if the costs depend on states and works as in (1.12). For a given state $x_{t}$ at the start of year $t$ the future costs must not depend on the previous states and maintenance decisions, i.e. for the future costs the path up to $x_{t}$ is irrelevant. For general stochastic processes this is the Markov property, see e.g. [Minoux (1986), p 410]. For instance, a state characterisation only by the current IRI-value (cf. Sec 1.1) is insufficient since, according to VV regression models, the degradation rate (cf. Sec 1.1) is valid for the whole time period between two consecutive major maintenance operations: the future evolution depends on the IRI-history. Such troubles can be solved by the introduction of the current degradation rate as part of the state description. The DynP method was formulated by [Bellman (1957)] and [Howard (1960)], see e.g. [Minoux (1986), p 381]. Using DynP the computational work increases linearly with the number of years $T$.

### 1.3 Program system

The study character as, apart from being a LiU-project (Linköping university), being a CDUproject (Centre for research and education in operation and maintenance of infrastructure) means special commitments to CDU and the financiers. The full aim is formulated in Sec 1.5.1 below. The stand alone program system that we have developed in the project, further described in the following chapters, is called OPM (= Optimisation for Pavement Management). The package consists of four linked main $\mathrm{C}++$ programs and a number of Matlab® statistics plot routines, all exchanging information via binary and text data files. The program code for all the different studies in the thesis, except the initial study (Ch 2), has been developed as one integrated system, with the aim to accomplish flexible model and program building blocks for future use.

During a program run multidimensional hierarchical structures are built dynamically, admitting changeable numbers of explicitly considered years, segments, states, works types, etc. The three main programs that contain iterative methods for solving a dual problem - the residual value routine RestOPM, the road class oriented StartOPM and the segment oriented OPM - are intended for parallelisation runs on a user controlled number of processors, even a single PC may do. The recommendable PC-cluster size would depend on the processor capacities as well as the modelling ambitions: both the (primary) memory needs and the total CPU-times increase essentially linearly with the number of discrete pavement states that are distinguished in the DynP backward iteration procedure (cf. Sec 1.2.5). The size of the data structures can also be varied by the chosen degree of "semi-manufacturing" of static data (pre-calculated and stored or recalculated each time needed) and by the choice between simultaneous access to the whole data structures in primary computer memory or via intermediate binary files.

### 1.3 Program system

In its present form the OPM system is a R\&D-product, not ready for production runs. Before that, the large set of options should be reduced to those of special interest. Moreover, both the input of data and the running of the whole program system should be carefully documented and equipped with supporting menu oriented routines. Although developed in dialogue with the Swedish VV an international interest cannot be excluded, for integration in existing PMSs.

We present the structure as flowcharts, without further comments. (In the file names the star notations ** are for processor and/or run identification.)



### 1.3 Program system



### 1.4 Earlier studies

Surveys of pavement management (PM) in general, and budget related questions in particular, are found in [Paterson (1987)], [Haas et al (1994), esp. ch 19: Priority programming], [Wallman et al (1995)], [Haas (1997)], [Robinson et al (1998), esp. ch 7: Prioritisation] and [HDM-4 (2000), esp. vol. 1, part G: Economic analysis]. The last reference is to the World Bank project "Highway Development and Management", primarily designed for the developing countries, but since the models are general and flexible, the latest development led by the University of Birmingham and extended to models for cold climate and safety, this project is of universal interest. In App 1 below we document the Swedish adjustments of the traffic cost models. For maintenance policies in general, the survey by [Wang (2002)] includes many references. In our project, literature summaries are given by [Lindberg et al (1997), (2001)]. In the following we go through PMS related articles found up to 2003 with focus on the use of optimisation, partly grouped by the models and methods being used.

### 1.4.1 Systems based on Markov decision processes (MDPs)

The Markov property was mentioned in Sec 1.2.5. For a road segment, being a (finite, homogeneous) Markov chain (MC) means that the number of pavement states is finite and that the state may change at given discrete time points, say once a year, according to static probabilities, irrespective of earlier state history. For a Markov decision process the MC time evolution is (partly) controlled by decisions (of works) that either can be taken freely and independently or are subject to, e.g., common budget constraints. Since the theory and implementation results for MDPs are extensive, see e.g. [Ross (1970)], [White (1985), (1993a), (1993b)] and [Carnahan (1988)], it is not surprising that most pavement optimisation systems (described in literature) build on MDP assumptions. If a finite set of pavement states and works options are distinguished, MDP problems can be reformulated as LP-problems, originally by [d'Épenoux (1960), (1963)] , [Manne (1960)] and [de Ghellinck (1960)] - a topic that we will return to in Ch 5 . Our segment oriented model in Ch 4 below handles a continuum of states and works.

In [Nazareth (2000)] an extension to stochastic budgets is mentioned, suggested to be solved with stochastic dynamic programming. Representing another kind of extension, by admitting random time steps between the state changes [Nesbitt et al (1993)] and [Ravirala and Grivas (1996)] use a semi-Markov approach, letting the time that a segment spends in a pavement state be stochastic. Of these two, the former model is applied to flexible, heavy-duty pavements in Manitoba, the latter to New York State data, but both lack budget restrictions. Also [Butt et al (1994)], [Carnahan et al (1987)] and [Smadi and Maze (1994)] avoid the budget restriction difficulties, by using dynamic programming for minimising the total maintenance cost. (Butt's et al handling of the budget constraints is by heuristics, ranking projects according to their benefit/cost ratios.)
[Ben-Akiva et al (1993)], [Madanat (1993b)] and [Madanat and Ben-Akiva (1994)] take a factor analytical approach, by using "latent MDPs". The idea is that, apart from degradation, also inspections suffer from uncertainty, manifested as measurements with errors (whereas the
true state is latent), and this is a way to handle such. We think of a possible use e.g. for the valuing of information by two different measurement techniques, one simplified and subject to greater errors than the other. [Smilowitz and Madanat (2000)] extend the latent MDP approach to the network decision level, expressing the budget constraints by enclosing the expected maintenance cost between two budget levels, and bounding the fractions (of segments) for each state and time. Models for both finite and infinite horizons are formulated. The sum of agency and traffic costs is minimised by LP.
[Chua (1996)] presents another extension, letting traffic and mechanistic, stochastic pavement effect submodels, partly based on AASHTO test results (cf. [AASHTO (1993)]), determine the life time of a pavement. By applying stochastic DynP to one road segment on a finite planning period, minimising a weighted sum of maintenance and road user costs (without budget constraints), the author admits dynamic state variable values and time evolutions of the traffic volume for two vehicle types. Also [Li et al (1997b)] consider dynamic (non-homogeneous) MDPs, permitting evolutions of traffic and environmental effects.

### 1.4.2 The Arizona model

When the Arizona model based PM-system was introduced in Arizona in 1980-81, it is reported, see [Golabi et al (1982)] and [Kulkarni (1984)], to have saved almost $1 / 3$ or 14 million USD of the road preservation budget for the State of Arizona. The reason would be a more unified handling of all road maintenance projects. The implemented models are a longterm and a short-term model. Both build on a subdivision into a finite set of pavement states and works options. The long-term model is of Markov chain type, finding stationary state probabilities minimising the total maintenance cost, for given transition probabilities and expected maintenance costs per state and works option. Instead of explicit traffic costs, the LPformulation includes lower bound constraints for all "acceptable"-state probabilities and upper bounds for all "unacceptable"-state probabilities. The resulting steady-state probabilities and optimal (lowest possible) total maintenance cost are taken as inputs to the short-term model. From a given state distribution, the total discounted maintenance cost is minimised for an explicit time period, letting the final distribution deviate at most a given percentage from the stationary probabilities and total cost. After five years use of the models, the authors report a turning from corrective to preventive works, of a concomitant smaller extent but slightly more frequent. Moreover, the model chooses thinner pavement layers than was used before. Both of these experiences are of special interest to us - cf. our case studies in Secs 4.4 and 6.3 below.

In the reported Arizona study, the entire road network was subdivided into nine categories, handled by separate models. For the reported implementation, aggregating 7400 1-milesegments, the category based results for all segments in a common pavement state mean that the relative use of different (discrete) works options should obey the determined optimal percentages. The interpretation of these percentages meant no problems, when putting the results into practice. A more serious concern was that the results could mean that for a 3-mile road section one works type was suggested for the first and third mile, and another for the second. We will handle such objections in Ch 6 below.

The [Wang and Zaniewski (1996)] hindsight describes the experiences of the Arizona and the related NOS models (in Sec 1.4.3). The entire Arizonian network is now subdivided into 15 road categories, and 6 works options are available. A $10-\mathrm{yr}$ planning horizon is used in the reported runs. We will use 40 yrs in the segment oriented runs below, and 80 yrs in the road class oriented - with our subdivision into 29 road classes the model most resembling the NOS model. The authors realise that the steady state is to be viewed as an ideal scenario and has never occurred in Arizona, due to fluctuations in budgeting and pavement behaviour. We may add non-homogeneous traffic evolution, since changed traffic volumes will shift the optimum, as we argued at Fig 1.1.

### 1.4.3 The NOS-based PM-systems in Kansas, Alaska and Finland

The further development of the original Arizona model has resulted in the more general Network Optimisation System (NOS). Whereas the Arizona model answered the question (1) "What are the minimum budget requirements necessary to maintain prescribed performance standards?", in [Alviti et al (1994), (1996)] also the reverse question (2) "What maximum performance standards can be maintained for a fixed budget?" is answered. The latter question is much harder. The authors have chosen price-directive Dantzig-Wolfe decomposition and solve the dual subproblem - separable into one subproblem per segment category (for $23-69$ categories) - by LP. This is similar to the methodology that we will use - but we will model each segment individually and use a more general method tool box, including DynP. In NOS it is possible to choose between (maintenance) cost-constrained benefit maximisation and benefit-constrained cost minimisation. In the total cost objective, the benefits for the different segment categories are weighted by road area. This seems strange to us, since (in the first place, and together with length) traffic volume is responsible for total benefit, whereas area is decisive for maintenance costs. NOS has been implemented in, e.g., Kansas and Alaska. In the Alaskan use, the differences implied by the questions (1) - (2) are summarised. Whereas the original Arizona model (1) results in the minimum budget 40 million USD for the average benefit level 0.82 (given from historical benefits per category, 69 categories), the benefit maximising NOS-run (2) for the fixed annual budget 40 million USD (and optimisation between categories) achieves benefit level 0.86 . Reversely, to satisfy this benefit level by answering question (1) would require a 48.2 million USD budget (but we cannot see how the benefits per category were chosen). In Alaska about 90 percent of the projects recommended by NOS during a $5-\mathrm{yr}$ period were selected for implementation.

Also the Finnish PMS Highway Investment Programming System (HIPS) is of NOS-type. According to [Olsonen (1988)], optimisation is performed for each segment category separately, like the original Arizona model. The total road user and maintenance cost is minimised - but we cannot find the principles for weighing between the two (cf. Fig 1.1). In [Sikow et al (1994)] results from a Lapland study are summarised. The effects of varying the annual budgets (to be specified for each segment category) are analysed. A significant conclusion is that road maintenance (and its split between summer and winter maintenance) and highway investments should not be treated separately, but all become components of an overall "road keeping" optimisation. In [Thompson et al (1989)] and [Äijö et al (1990)] more
technical details are given. It discusses the possibility to use the slope of the road-user cost vs. maintenance cost curve (cf. (1.1)) in the long-term model for weighing between the two costs, in scarce budget situations, and for weighing between the different segment categories. The authors reveal that the weight factors are user specified. The allocation of budget means from the nationwide network to 13 highway districts is implied by the optimal long-term and shortterm length distributions per state and works option, together with the number of kilometres per segment category and district.

### 1.4.4 Swedish PMS-based optimisation

The above mentioned Finnish HIPS package has been applied to Swedish data, see [Äijö (1995)], [Virtala (1996)] and [Lang (1996)], using the total traffic + maintenance cost as objective for minimisation. We have not found the weigh principles that were used. The effects of budget variations are described. These seem to be performed by BCR-ranking (cf. (1.1)), similar to that of [Butt et al (1994)], with $B C R=1$ as target value for an "optimal" budget. (Cf. Fig 1.1, where $B C R=1$ means that the contributions from the two solid cost curves can be summed up to the total cost.) The long-term optimal budget was 1242 million SEK, when summarised over 12 sub-networks. For an 8 -yr short-term time period, the use of this budget level would mean an improvement of the average state, in comparison to the current conditions in 1996, although the total cost would decrease for short-term budget levels up to around 1750 million SEK. [Lang (1999), (2002)] are documentations of the HIPS-input data changes that were made during 1995 - 2002. Subdivisions of the Swedish road network into 6 traffic classes and 3 climate zones were performed. In HIPS five works types are distinguished. VV utilises HIPS on a strategic management level, whereas prioritisation is used otherwise. In all prioritysation VV calculates the additional traffic costs, in comparison to the cost of an ideal pavement state, and uses it together with the full maintenance cost, see [Vägverket (1997)] and [Lang (1997)].

### 1.4.5 PMS-systems in Denmark and Norway

The Danish PMS, BELMAN, is presented by [Jansen and Schmidt (1994)]. Whereas the important input data and functions are described in detail, the proceeding lacks information about the optimization models that generate maintenance plans within budget constraints. The Norwegian PMS, see [Haugødegård et al (1994)], includes network optimisation, minimising agency + traffic cost subject to budget constraints. The road standard, in terms of longitudinal and transversal unevenness, is not admitted to exceed given road standards, different for different road types and traffic classes. No optimisation details are revealed.

### 1.4.6 Other optimisation techniques

[Flintsch et al (1998)] supplement the budget recommendations from the NOS-model with a ranking ("rate") formula for each candidate project, based on current state and maintenance costs, and the final choice of projects is made according to the rates, up to the recommended budget. In [Artman et al (1983)] an optimisation model is formulated for maintenance of
airfield pavements, with similar structure as for road maintenance. For a finite set of works options, the problem is expressed in terms of $0 / 1$-variables. The resulting integer
programming (IP) problem is solved with a heuristics of gradient-search type, invented by [Toyoda (1978)] and providing good but not necessarily optimal maintenance plans. The same technique is reported to have been utilised for highway maintenance optimisation by [Philips and Lytton (1977)]. In [Li et al (1997a), (1998)] a similar IP problem is formulated and solved to optimality - for a network consisting of 18 and 5 segments, respectively. The state description is 1D and the objective minimises a sum of benefit/cost ratios.

For continuous time, continuous state space, infinite planning horizon and deterministic time evolutions of pavement states, [Li and Madanat (2002)] find the optimal steady-state frequency of major works for one segment. The model is an optimal control minimisation of a functional describing total agency and traffic costs, subject to a $1^{\text {st }}$ order ordinary differential equation (and no budget restrictions). Before that, [Tsunokawa and Schofer (1994)] formulated and solved such a problem, comparing different integration methods and distinguishing transition time and steady state. The latter means "periodic maintenance", manifested through sawtooth curves describing the state evolutions.
[Mamlouk et al (2000)] uses the official standard AASHTO, see [AASHTO (1993)], for design modelling, and mechanistic models for the pavement degradation, in terms of equivalent single axle load. Subject to restrictions on the terminal state variable values, nonlinear DynP is applied for minimising the weighted sum of agency and traffic costs. The PC-based program is intended for the project decision level.

Examples of genetic algorithms (GAs) are [Fwa et al (1994)] and [Chan et al (2001)], both applied to a problem with a lot of resource constraints, but lacking the typical time structure of road maintenance. The former paper compares GA with integer programming and conclude that GA is a real, PC-based alternative. The latter article presents a new method, the prioritised resource allocation method (PRAM), and compare its performance with two established methods. PRAM outperforms the other two. The original GA-formulation, see [Holland (1975)], had no special means for handling constraints.

### 1.4.7 Decision support systems and integration

[Worm and van Harten (1996)] apply an OR-view to the maintenance planning problem, using different optimisation models and methods on four different decision levels, including MDP theory, DynP and shortest route calculation. The last mentioned technique is applied on the $3^{\text {rd }}$ level, where single-segment projects for a road are joined - a coordination facility that we will integrate in Ch 6 below. The common objective on all decision levels is maintenance cost minimisation. In [Davis and Carnahan (1987)] the MDP-based DynP-optimisation is supplied with a Monte Carlo simulation tool, generating cost and state statistics for the optimal works policies. The MDP-model includes constraints that forbid violations of given state bounds, by more than given probabilities. Also [Feighan et al (1988)] speak for the simulation possibilities. [Hanna et al (1993)] describe an expert system, for engineer support on the
project decision level. Based on input data, the system produces a list of recommendable works options and their consequences. [Ritchie (1987)] and [Ritchie et al (1987)] describe another expert system and its possible integration with network optimisation. [Wang et al (1994)] write about a system in Arizona, where the NOS-model in Sec 1.4.3 is supplemented with an expert system for the translation into concrete maintenance projects. The role of NOS becomes to produce budgets per road category (=segment category). The article by [Sundin and BrabanLedoux (2001)] is a status report on the use and findings of artificial intelligence, including artificial neural networks (ANNs), fuzzy logic genetic algorithms and expert systems, in PMS decision support systems. Apart from providing a support for road inspectors (not reported here), in [Yang et al (2003)] ANN models are used for the generation of composite pavement state indices: pavement condition rating.
[Madanat (1993a)] analyses the potentials of incorporating inspection decisions in PMS. The optimisation model includes choices of inspection or not per year. If the forecasting models are accurate, the author foresees that such an integration can lead to substantial cost savings. [Grivas and Schultz (1994)] formulate an optimisation model for the integration of neighbouring pavement and bridge projects. This is also part of an integration in [Chasey (1995)], who also recommends an integration with highway investments, with decision support from simulation models.

### 1.4.8 Survey articles and implementation experience

The role for operations research (OR) in general maintenance planning is analysed by [Dekker (1995), (1996)]. Various OR implementations and a lot of references are listed (including somewhat road maintenance). The paper [Hagquist (1994)] includes figures showing that substantial benefit increases are possible, if optimisation software is being used, especially for scarce budgets. In [Lytton (1994)] the available optimisation methods are described, plus some case studies, and the difference between optimisation and ranking is clarified. [Thompson (1994)] discusses the role of optimisation as a means for accomplishing consistent pavement policies and stable funding - and agency credibility.

The experiences from the implementation of the original Arizona model are analysed in [Hudson and Haas (1994)]. Apart from the results reported in Secs 1.4.2-1.4.3, the authors mention a general increase of awareness of all the factors that are involved in pavement management. The Alaskan implementation is treated in [Johnson et al (1994)], and the importance of adequate organisational changes as well as involvement of both technicians and management is emphasised. In [Tarricone (1993)] the difficulties in collecting and storing relevant input data are addressed. The use of PMS and HIPS-optimisation in Finland as a tool for performance contracting and bid evaluations, as well as product pricing, is described in [Tapio and Piirainen (1994)], whereas the conference contribution from [Knudsen and Simonsen (1994)] present the output data from the Danish BELMAN system, and how they are differently used by decision makers on the network and project management levels.

### 1.5 Study aim and outline

### 1.5.1 Aim of study

Apart from meeting scientific quality demands, our purpose is to accomplish a tool that can be integrated into a PM (Pavement Management) decision support system in general, for VV in particular. The implementation should comply with existing VV standards as to computer operating systems and programming languages. The model system should be modularised and flexible as to model types, e.g. admit both discrete and continuous state descriptions, and data input, and should be formulated in dialogue with VV. The intended use is in a stage of budgeting and planning, before concrete project information is available. The implementation should be capable of handling regional and even national road networks. On the network level we think of budget scenario studies, as well as budget allocation. Moreover, an intended use is on the road level, for the identification of candidate projects and state supervision. The study should include some kind of validation.

### 1.5.2 Outline of the thesis

In Ch 2 an initial study is performed, based on principally the same kind of data as the existing NOS-model does, see [Lang (1999), (2002)]. In the huge data sets we encounter several inconsistencies and ad hoc assumptions.

From the weaknesses we identify in the initial study a new, more transparent input model is formulated in Ch 3 together with VV - a thorough revision of both data and functions. The setup program for an optimisation run, joining the input data and functions into streamlined road characteristics, is DataOPM in the program system, cf. Sec 1.3.

In Ch 4 the basic optimisation models are formulated, both a road class oriented and a segment oriented model. Whereas the road class model is intended for the network level, both for scenario studies and as a general start routine for the optimisation, StartOPM in Sec 1.3, the segment oriented model is the main routine, OPM in Sec 1.3, also on the road level. Whereas the road class oriented model treats the segments simply by their road lengths, the segment orientted model considers their individual characteristics. Both the pavement state and works option spaces (of possible variable values) are continuous. The tricky governmental BCRlower bounds are complied with, as dual constraints. Here we make use of the optimisation concepts and methods in Sec 1.2. The validation of a special study of road 63 in Värmland is reported.

Ch 5 is devoted to our residual values routine, RestOPM in Sec 1.3. The solution method DynP, cf. Sec 1.2.5, is applied in a special form, only using a 1 -yr explicit time period. See Fig 1.5 b for an illustration. "Special" means that the values $\varphi$, which are determined per nodal state (unfilled circles), shall coincide at the start and end of the modelled year. This is our characterisation of a stationary process, as to states and costs. The residual values are transferred to StartOPM, as Fig 1.5a illustrates. The initial segment states are fully considered (filled circles.) The different numbers of works options per state in Fig 1.5 reflect state
restrictions: on the highest node level (a worst acceptable state), minor maintenance is forbidden, whereas on the lowest (ideal) level no major maintenance is conceivable. For the translation to OPM, further individualisation is performed for a transfer period $(T, 2 T)$, providing residual values for the time period $(0, T)$ of $T$ years, which is explicitly considered in the segment based analysis.


Figure 1.5 Illustration of DP-use in road class oriented routine (a) and residual values routine (b).

Ch 6 describes a further development of the segment oriented base model for application on the road level, fine-tuning the cost description of the coordinated projects that consist of simultaneous maintenance of a varying number of neighbouring road segments. For a road section of $N$ contiguous segments and a choice between two works types - minor or major maintenance - for each segment and year, the number of choice combinations becomes $2^{N}$ for each year. Such cost considerations cannot be taken on the road class level, where most articles in Sec 1.4 are found. A validation on road 63 is reported.

Ch 7 treats a stochastic, i.e. a more realistic, variant of the base model. By assuming independency between the random events that determine the state evolutions of different segments or road class states and assuming that the law of large numbers, see e.g. [Devore (2004)], is applicable, the model becomes analytical. The approach may be extended to uniform impacts of weather/climate variations and/or coordinated maintenance, without any tedious Monte Carlo simulation.

Ch 8 is devoted to the effects of budget redistribution between the years, providing more flexibility in times of scarce budgets. Different deposit and lending rates are distinguished.

In Ch 9 our focus is on methods for the network level, investigating different variants of the models and methods sketched in Sec 1.2. The aim is to reduce CPU-times and the number of parallel processors.

## 2 Initial study

The multiyear maintenance-planning problem concerns when to maintain which paved public road segments by what rehabilitation actions. We present an optimisation model, minimising the overall road user cost subject to annual budget constraints, and apply it to a regional network in Sweden, comprising almost 10,000 road segments. The solution approach we propose involves Lagrangean relaxation, with segment-specific dynamic programming subproblems, subgradient-based updating of the dual multipliers, warm-start, and a primal feasibility heuristic. The first results show that our approach is computationally feasible on the regional planning level.

### 2.1 Introduction

Vägverket (VV, Swedish Road Administration) spends around MEUR 300 each year on the maintenance of the paved state road network, and even small percentage cost reductions mean great savings. The default option is routine maintenance (minor patching of pot holes and cracks, etc.), which becomes more expensive as the road surface deteriorates. The condition of around 150,000 homogeneous road segments is measured regularly by VV , and the future deterioration is forecasted. Many road administrations around the world store huge data sets, and use Pavement Management Systems (PMSs) to handle such information. VV registers both road constants and time-dependent data for each segment in its PC-based PMS, see [Lang and Dahlgren (2001)], [Lang and Potucek (2001)].

Both on a strategic (national) and a programme (regional) planning level, optimisation might be useful for scenario studies and as decision support for budget allocation or identification of candidate projects. As a research project but intended for later integration with VV-PMS, we have developed and implemented a method, OPM (Optimisation for PM), based on Lagrangean relaxation, and so far applied it to the county of Värmland in Sweden, comprising almost 10,000 segments. Its prospective success among PMS users further depends to a large extent on the quality of input data. We extract road segment data from the official VV-PMS database, whereas cost and deterioration models are taken from an experimental database (VV-HIPS).

As for the optimisation problem, annual budget constraints are the only constraints coupling the road segments, and Lagrangean relaxation (cf. Sec 1.2.2) then results in one separate multiyear subproblem for each segment. The dual prices (Lagrangean multipliers) can be interpreted as exchange rates between agency-EUR and society-EUR (also cf. Sec 1.1) for the years of the planning period, and provide every subproblem with a combined road user and maintenance cost objective.

The published large-scale studies on the segment level are few. [Lindberg et al (1997)] suggest the use of Lagrangean relaxation and include an international overview of optimisation in PMSs. As for a state of the art of optimisation for maintenance decision making in general, see [Dekker and Scarf (1998)]. The literature is more extensive on an aggregate - road class level, where groups of segments of, e.g., similar traffic load are distinguished. [Alviti et al
(1996)] describe a road class based Markovian model, cf. Sec 1.4.3, a development of the "Arizona model" [Golabi et al (1982)]) and a solution technique, NOS, based on DantzigWolfe decomposition, for our kind of problem. (The mentioned VV-HIPS is of NOS-type.) Their optimal decisions are the proportions of road lengths to maintain by different actions, presented per road class and cell, each cell representing a range of (undistinguishable) pavement conditions. The subdivision into cells corresponds to a discretization of a multidimensional state space. Our approach, on the contrary, makes it possible to model and get maintenance plans for individual road segments, in a state space of continuously varying pavement conditions.

### 2.2 Model

### 2.2.1 Problem description

The pavement condition of a road segment can be characterised by a number of state variables. Apart from Pavement Age, two main statistics of profilometer data are monitored (in Sweden): International Roughness Index (IRI) for longitudinal unevenness and Rut Depth for transversal unevenness. Other candidate state variables, used by e.g. [Alviti et al (1996)], are measures of bearing capacity and cracking. Due to varying winter conditions and other disturbances, deterioration of the pavement condition may be viewed as a stochastic process. In a Markovian type of model, e.g. [ibid.], both the time evolution and all costs are determined by the current combination of the state variable values, i.e. the pavement state, together with a number of local road characteristics, e.g., width and traffic - the road constants - and state transition probabilities. By applying the common approach of planning by the year, we will model a discrete time - as with Markov chains. However, the estimates of the transition probabilities are currently considered unreliable in Sweden. Therefore our aim, to consider the state transitions as deterministic by modelling only the average behaviour, does not involve any loss of information in practice.

For planning purposes the road network is partitioned into a number of classes, according to road type (motorway, national " $2+1$ lanes" road, etc), annual average daily traffic (AADT) interval, climate zone, and/or speed limit. Every road class has specific pavement attributes in some sense, e.g. maximum allowable IRI and Rut Depth values. Due to budget scarcity such limits are violated in practice, although this is unwanted.

The individual road segments are considered homogeneous as to the pavement conditions. In reality the subdivision of the roads into segments of varying lengths mainly identifies sections sharing pavement history, and high standard deviations of IRI and Rut Depth may be registered, especially for long segments. On the other hand, using short segments and disregarding the potential coordination profit from the simultaneous maintenance of nearby segments could lead to cost overestimates. We ignore the effects of coordinated maintenance, assuming that the segments are acceptably long.

In practice, a lot of maintenance actions are available, as to different pavement materials, methods and extent, see e.g. [Haas (1997), p 82]. On the regional or national planning level,
without detailed project/preparation information about the local pavement conditions, it is more relevant to choose between types of maintenance actions - e.g. patching, sealing, resurfacing and reconstruction of flexible pavements - and to use the characteristics of these types in the model. In Sweden all routine maintenance is financed from one budget - for operations - and the rehabilitation actions (collective for all actions except routine maintenance) from another, but we consider an overall annual budget for all the maintenance costs $C_{M}$, thus also determining an optimal budget split.

The traffic cost $C_{T}$ measures all the road user sacrifices attributable to the pavement conditions, e.g. vehicle operating costs and time costs, as valued by society. In traditional maintenance planning by net present value (NPV)/cost ratios, see e.g. [Robinson et al (1998), p 116], a rehabilitation action is preferred to routine maintenance if its cost (PV) difference satisfies (cf. (1.1))

$$
\Delta C_{T}+v_{B C R} \cdot \Delta C_{M} \leq 0
$$

where $v_{B C R}$ is a "tax factor" used to exchange agency-EUR for society-EUR. Since we expect results showing budget scarcity, we ignore this type of constraint. It is easily checked afterwards, since the Lagrangean multipliers used below will act like $v_{B C R}$.

### 2.2.2 Mathematical formulation

In a $D$-dimensional state space the pavement state $\mathbf{x}=\left(x_{d}\right)_{d=1}^{D}$ is a point defined by the values of the state-variables $x_{d} \in \mathfrak{R}$. The higher $x_{d}$-value the worse is the state. Every road segment $s$ out of $S$ given belongs to a unique road class $r=r(s)$. A planning period of $T$ years is considered. The state of segment $s$ at time $t$ is denoted $\mathbf{x}_{s t}$, and the initial state $\mathbf{x}_{s 0}=\mathbf{a}_{s}$ is given. In road class $r$ a discrete set $\mathcal{U}_{r}(\mathbf{x})$ of feasible maintenance actions $u$ is available. Deterministic state-transitions are assumed: If action $u_{s t}$ is applied to segment $s$ at time $t$ the resulting state is $\mathbf{x}_{s, t+1}=h_{r(s)}\left(\mathbf{x}_{s t}, u_{s t}\right)$. The maintenance costs $c_{s}\left(\mathbf{x}_{s t}, u_{s t}\right)$ are restricted by overall annual budgets $\left(b_{t}\right)_{t=0}^{T-1}$. The traffic costs $f_{s}\left(\mathbf{x}_{s t}\right)$ are convex functions of $\mathbf{x}_{s t}$. The objective is to minimise the total future traffic cost, discounted to time $t=0$ by an annual discount factor $d$ (not to mistake for state dimension), and including a residual cost contribution $\varphi_{S T} \bullet\left(\mathbf{x}_{s T}\right)$ for the time after $t=T$. The optimisation model is:

$$
\begin{align*}
& \underset{\left(\mathbf{x}_{s t}\right)_{s, t},\left(u_{s t}\right)_{s, t}}{\operatorname{minimise}} \sum_{s}\left(\sum_{t=0}^{T-1} d^{t} \cdot f_{s}\left(\mathbf{x}_{s t}\right)+d^{T} \cdot \varphi_{s T} \bullet\left(\mathbf{x}_{s T}\right)\right) \\
& \text { subject to }\left\{\begin{array}{cc}
\sum_{s} c_{s}\left(\mathbf{x}_{s t}, u_{s t}\right) \leq b_{t} & \forall t \\
\mathbf{x}_{s, t+1}=\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, u_{s t}\right) & \forall s, t \\
\mathbf{x}_{s 0}=\mathbf{a}_{s} & \forall s \\
u_{s t} \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s t}\right) & \forall s, t .
\end{array}\right. \tag{2.1a}
\end{align*}
$$

This formulation admits that the maintenance set $\mathcal{U}_{r}(\mathbf{x})$ is restricted (i.e. routine maintenance is excluded) above certain $x_{d}$-levels defining the acceptance limits of really poor road con-

### 2.2 Model

ditions. As an alternative we use penalties added to the traffic costs for the violation of such acceptance limits. There are $S \cdot D \cdot(T+1)$ state-variables $x_{s t d}$ and $S \cdot T$ decision- (or control-) variables $u_{s t}$. The problem is non-convex due to the discrete nature of $\mathcal{U}_{r}(\mathbf{x})$ and possibly by the form of the maintenance costs $c_{s}$ in (2.1a).

### 2.3 Method

### 2.3.1 Main procedure

Since the budget constraints (2.1a) are the only constraints that link the road segments, we apply Lagrangean relaxation, introducing Lagrangean multipliers $\left(v_{t}\right)_{t=0}^{T-1}$, and receive a separable dual. Subproblem $s$ is written:

$$
\begin{align*}
& \Phi_{s}(\mathbf{y}):=\min _{\left(\mathbf{x}_{s t}\right)_{t}\left(u_{s t}\right)_{t}} \sum_{t=0}^{T-1} d^{t} \cdot\left(f_{s}\left(\mathbf{x}_{s t}\right)+\sqrt{d} \cdot v_{t} \cdot c_{s}\left(\mathbf{x}_{s t}, u_{s t}\right)\right)+d^{T} \cdot \varphi_{s T \bullet}\left(\mathbf{x}_{s T}\right)  \tag{2.2a}\\
& \text { subject to }\left\{\begin{array}{cl}
\mathbf{x}_{s, t+1}=\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, u_{s t}\right) & \forall t \\
\mathbf{x}_{s 0}=\mathbf{a}_{s} & \\
u_{s t} \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s t}\right) & \forall t .
\end{array}\right. \tag{2.2b}
\end{align*}
$$

Here the state transitions are assumed to occur in the middle of each year, and the $\sqrt{d}$-factor reflects that the maintenance cost (in current prices) is discounted for half a year. Hence $v_{t}$ expresses a rate of exchange from agency-EUR to society-EUR in the middle of year $t$. In order to solve the dual,

$$
\begin{equation*}
\underset{\mathbf{v} \geq \mathbf{0}}{\operatorname{maximise}} \sum_{s} \Phi_{s}(\mathbf{v})-\sqrt{d} \cdot \sum_{t=0}^{T-1} d^{t} b_{t} v_{t}, \tag{2.3}
\end{equation*}
$$

we use subgradient techniques with Polyak steps, e.g. [Minoux (1986), p 112], and primal target costs generated by heuristics. In each iteration, with $\mathbf{v}$ fixed, the segment-specific Lagrangean subproblems (2.2) are shortest-path problems, which we solve by dynamic programming (DynP, cf. Sec 1.2.5) - using a continuous state space but interpolating costs and succeeding states between grid points (where all input data concerning states are found). If $L$ grid levels are used per state space dimension, then $S \cdot L^{D} \cdot T$ discrete decision-variables are needed.

### 2.3.2 Subproblem characteristics

Due to the discrete nature of $\mathcal{U}_{r}$, we will use $m$ to index the maintenance actions $u$. In order to apply DynP, we introduce $n$ as a grid point (node) index and let $\varphi_{s t m}\left(\widetilde{\mathbf{x}}_{r n}\right)$ denote the optimal future cost for segment $s$ when action $m$ is used in grid state $\widetilde{\mathbf{x}}_{r n}, r=r(s)$, at time $t$, for given multiplier values $\mathbf{v}=\left(v_{t}\right)_{t=0}^{T-1}$ (suppressed). In general, the application of a maintenance action $m$ to a road segment $s$ in grid state $\widetilde{\mathbf{x}}_{r n}$ in (the beginning of) year $t$ will result in a non-grid state $\mathbf{x}=\xi_{r n m}$ at the end of the year. The succeeding optimal $\operatorname{cost} \bar{\varphi}_{s, t+1, m}(\mathbf{x})$ is computed as a
weighted average cost of the $2^{D}$ grid points $n^{\prime} \in \mathcal{N}_{r}(\mathbf{x})$, the set of neighbouring grid points to $\mathbf{x}$. We will use this "bar notation" for such averages in general. The weights $w_{r n^{\prime}}(\mathbf{x})$ are chosen (in the natural way) to guarantee consistent state variable values, i.e.

$$
\begin{equation*}
\mathbf{x}=\sum_{n^{\prime} \in \mathcal{N}_{r}(\mathbf{x})} w_{r n^{\prime}}(\mathbf{x}) \cdot \widetilde{\mathbf{x}}_{r n^{\prime}} \tag{2.4}
\end{equation*}
$$

The interpolated cost becomes

$$
\begin{equation*}
\bar{\varphi}_{s, t+1, m}(\mathbf{x})=\sum_{n^{\prime} \in \mathcal{N}_{r}(\mathbf{x})} w_{r n^{\prime}}(\mathbf{x}) \cdot \varphi_{s, t+1, m}\left(\widetilde{\mathbf{x}}_{r n^{\prime}}\right) . \tag{2.5}
\end{equation*}
$$

Below the highest grid levels this amounts to linear interpolation, with weights corresponding to transition probabilities in an expected cost evaluation, see e.g. [Bertsekas (1995), p 282]. If violations of acceptance limits are penalised instead of forbidden, we apply linear extrapolation as well. As the notation $\varphi_{s t m}\left(\widetilde{\mathbf{x}}_{r n}\right)$ indicates, we compute optimal future costs for each $s, t, n$ and $m$, i.e. decision-costs, in order to reduce the discretization error and to facilitate the choice of optimal action. By separating the decisions $m$ we avoid the systematic underestimation of cost that would result (at least in state space regions where the $c_{S}(\bullet, u)$ functions are concave) if we stored only the lowest cost (among $\left.m \in \mathcal{U}_{r}\left(\widetilde{\mathbf{x}}_{r n}\right)\right)$ at each grid point $\widetilde{\mathbf{x}}_{r n}-$ see Fig 2.1 for an illustration. If the actions are few this may well outweigh the drawbacks of greater memory and computation needs. Letting $\boldsymbol{\xi}_{r(s), n m}=\overline{\mathbf{h}}_{r(s)}\left(\widetilde{\mathbf{x}}_{r(s), n}, m\right)$, the DynP backward recursion for $\varphi_{s t m}, t=T-1, \ldots, 0$, becomes

$$
\begin{align*}
\varphi_{s t m}\left(\widetilde{\mathbf{x}}_{r(s), n}\right)= & f_{s}\left(\widetilde{\mathbf{x}}_{r(s), n}\right)+v_{t} \cdot \sqrt{d \cdot} c_{s}\left(\widetilde{\mathbf{x}}_{r(s), n}, m\right) \\
& +d \cdot \min _{m^{\prime} \in \mathcal{U}_{r(s)}\left(\xi_{r(s), n m}\right)}\left[\bar{\varphi}_{s, t+1, m^{\prime}}\left(\xi_{r(s), n m}\right)\right] \quad \forall s, n, m . \tag{2.6}
\end{align*}
$$

Figure 2.1 Interpolation of costs between two grid levels $A, B$ in the presence of two actions $m_{l}$ and $m_{2}$.


Compared to ordinary DynP, the decison-cost interpolation means that the minimisation is over the actions of the succeeding instead of the current year, and the computational work involves
$\left|\mathcal{U}_{r}\right|$ decision-cost interpolations instead of one linear. For each segment $s$, the DynP forward recursion identifies the shortest-path from $\mathbf{x}_{s 0}=\mathbf{a}_{s}$, through

$$
\left\{\begin{array}{l}
u_{s t}^{*} \in \underset{\arg \min }{\arg } \bar{\varphi}_{s t m}\left(\mathbf{x}_{s t}\right)  \tag{2.7}\\
\mathbf{x}_{s, t+1}=\overline{\mathbf{h}}_{r(s)}\left(\mathbf{x}_{s t}, u_{s t}^{*}\right)
\end{array} \quad t=0, \ldots, T-1\right.
$$

This means another $\left|\mathcal{U}_{r}\right|$-fold increase of the number of cost interpolations for the decisioncost method. In our implementation we assume that the cost and transition data are given for a basic grid only (of least acceptable node density), and for every non-grid state $\mathbf{x}$ the function values $f_{S}(\mathbf{x}), c_{s}(\mathbf{x}, m)$ and $\mathbf{h}_{r}(\mathbf{x}, m)$ of the optimisation model have to be approximated by the interpolated $\bar{f}_{S}(\mathbf{x}), \bar{c}_{S}(\mathbf{x}, m)$ and $\overline{\mathbf{h}}_{r}(\mathbf{x}, m)$, respectively.

### 2.3.3 Start routine

Since we anticipate a large number of road segments to handle, the number of dual iterations is crucial, and so is the time horizon $T$ : The more appropriate residual values $\varphi_{s T} \bullet\left(\mathbf{x}_{s T}\right)$, the shorter time horizon can be used. For these reasons, we use a separate, simplified start routine to determine good initial multiplier values $\mathbf{v}$ for the main routine, and to estimate (realistic) residual values at the end of year $T$. In the start routine we solve an aggregate optimisation problem, analogous to (2.1), by collecting the road segments into their respective road classes and determining the optimal road length distribution over the grid points for each road class and year in a longer planning period (we use $2 T$ ). In this aggregate problem each road segment contributes its length only, since we disregard the segment differences as to road width and traffic load and use the road class averages. Contrary to the DynP forward recursions (2.7) of the main routine, the start routine is grid point based: If the state transition from $\widetilde{\mathbf{x}}_{r n}$ leads to a non-grid state $\mathbf{x}=h_{r}\left(\widetilde{\mathbf{x}}_{r n}, u\right)$, the corresponding road length is distributed among the neighbouring grid points $n^{\prime} \in \mathcal{N}_{r}(\mathbf{x})$, according to the weights $w_{r n^{\prime}}(\mathbf{x})$ from Sec 2.3.2, now used like probabilities in a stochastic process. This guarantees consistency: the inverse operation - linear interpolation - would satisfy (2.4). Initially the segment $s$ length is distributed among the neighbouring grid points of the initial state $\mathbf{a}_{s}$. The DynP backward recursions are made for a unit-length average-segment per class, and since the net state transitions are between grid points and no interpolation is involved, there is no need to differentiate between maintenance actions $m$ in the optimal future cost, i.e. $\varphi_{r t}\left(\widetilde{\mathbf{x}}_{r n}\right)$ replaces $\varphi_{s t m}\left(\widetilde{\mathbf{x}}_{r n}\right)$ in (2.6). Instead of one subproblem per segment as in the main routine, and one active segment state per year in the DynP forward recursions, we get one subproblem per road class in the start routine, but have to handle all states (of positive distributed length) for every year.

The road class and grid based computing, and the state transition distributions that can be interpreted as probabilities, make our start routine directly comparable to NOS. Notable differences are that we use the exact initial segment states, distributed, as input and also admit extrapolation (with some weights negative). Moreover, we let (dual) prices control the time
evolution, rather than forcing the length distribution to a pre-calculated steady-state distribution in some 10 years.

### 2.3.4 Primal Feasibility Heuristics

In order to get (good) feasible solutions (to (2.1)), we execute a primal feasibility heuristics every $I_{p}(\geq 1)$ iterations. This also gives us upper bounds to the optimal value in (2.1), used in the dual step length determination (of $\mathbf{v}$ ). The routine for accomplishing feasible solutions to (2.1a) is similar to the DynP forward recursion, but now all segments $s$ are treated simultaneously for one year at a time: from fixed states $\mathbf{x}_{s t}$ at the beginning of year $t$ one step of forward recursion determines optimal maintenance actions $u_{s t}^{*}$ for the relaxed problem, given $\mathbf{v}$, for this year. Depending on the maintenance cost in relation to the budget $b_{t}$ either of two strategies is used:

- If the overall maintenance cost for year $t$ exceeds budget, for each segment $s$ an action $u_{s t}^{c}$ with the least maintenance cost is considered. The segments are ranked, according to the (non-negative) difference $\Delta \varphi$ in overall cost between $u_{s t}^{c}$ and $u_{s t}^{*}$, and for the segments $s$ with the lowest $\Delta \varphi$-ranking $u_{s t}^{c}$ is chosen, until budget is reached. If this is impossible the heuristics fails.
- If the overall maintenance cost is below budget, the actions with the highest maintenance costs are considered for each segment, and selected according to a (non-negative) $1 / \Delta \varphi$-ranking. For the segments $s$ with the lowest ranking-values we stick to $u_{s t}^{*}$.

The choice procedure includes limited sorting: First the segments are distributed without search over a fixed array of $K+1$ bins, with entries corresponding to ordered intervals of rankingvalues. It is then sufficient to perform sorting in the bin $k, 0<k<K$, where the budget level is met. (If the ranking-value scale is inappropriate, the budget may be met in one of the two extreme bins, $k=0$ or $k=K$. In such a case another $(K+1)$-bin distribution is performed of the segments in that bin, before the final sorting.)

### 2.4 Application: PMS- and HIPS-based data

This chapter describes the application of our method, OPM, to the Värmland sub-network. In Sec 2.4.1 we describe the data sources and the implementation work. Since computing time will be critical in such a large-scale application we investigate, in Sec 2.4.2, some effects of differently sparse grids and of the two cost interpolation methods from Sec 2.3.2. Some optimisation results for the first run are analysed in Sec 2.4.3.

### 2.4.1 Data conversion to OPM

In the PMS-database the segments are identified by county, road No. and position along the road. For Värmland and the 9225 segments we consider, the average segment length is 499 m and the variation is large - see Tab 2.1. Our assumption in Sec 2.2.1, that the effects of coordinated maintenance can be ignored, might be questioned for the segments in the interval 0 - 400 m , comprising $15.2 \%$ of the total length. Laser-RST (Road Surface Tester) measurements have been performed since 1987. We use the following PMS data. Road constants: length, width, average AADT and AADTHeavy (the contribution to AADT from lorries and buses), and speed limit. Pavement data: pavement year, measurement year, average IRI and rut depth, both registered values and estimated evolutions. We handle missing segment data ( $1.0 \%$ of road constants, $8.9 \%$ of pavement data) by interpolating data for the same road, if possible. Otherwise the road segment is currently excluded ( $2.9 \%$ of total number, $8.0 \%$ of total road length).

| Interval $(\mathrm{m})$ | $0-$ | $400-$ | $800-$ | 1200 | -1600 | -2000 | $-2400-2800-3200-3600$ | $-4000-$ |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Frequency | 6752 | 976 | 482 | 278 | 172 | 113 | 97 | 70 | 62 | 47 | 176 |
| Length(\%) | 15.2 | 12.0 | 10.2 | 8.4 | 6.7 | 5.4 | 5.5 | 4.6 | 4.6 | 3.9 | 23.7 |

Table 2.1 Värmland: Length distribution of 9225 segments.
VV-HIPS (Highway Investment Programming System) is a Swedish implementation of the Finnish HIPS, see [Sikow et al (1994)]. In VV-HIPS the basic data of road user and maintenance costs, and transition probabilities, are given per rectilinear 4D cell - a combination of value intervals of 4 variables: Roughness IRI ( 3 intervals), Bearing Capacity (5), Cracks (3) and Rut Depth (3). In Sweden, Bearing Capacity and Cracks are not measured on a regular basis for each road segment, but calculated - as functions of IRI, Rut Depth, Pavement Age and AADT [Lang (1996), (1999), (2002)] - and in OPM they are replaced by one original (PMS registered) state variable, Pavement Age. We translate from cells to grid, by letting representative cell points determine the interior grid points in a basic discretization. In each state dimension two OPM grid-levels are added: a lower zero limit and an upper trigger level (acceptance limit) - suggested in Sweden for IRI and Rut Depth, for different traffic classes (and speed limits). In the basic OPM grid we use 5 grid levels per state dimension, 125 grid points in all. The grid point density can be automatically increased, by the introduction of intermediate grid levels (and using interpolated data from the basic grid).

Five maintenance actions (rather classes of actions, see [Lang (1996)]) are considered - routine maintenance and four rehabilitation actions of increasing extent and cost:

- routine maintenance (default action: crack sealing, patching, edge-repair, pothole fixing),
- simple action (inlays, e.g. treatment of rutting along wheel-paths),
- thin layer (open graded asphalt overlay for surface reconditioning),
- thick layer (dense graded asphalt overlay for, e.g., distributing load + open graded overlay), - reinforcement (reconstruction, reworking of the subgrade, new pavement layers).

In Värmland all roads belong to one and the same climate zone, and the 8 traffic classes define the road classes.

For the basic grid points in OPM the road user cost of the VV-HIPS database is replaced by formula-based computation [ibid.]. In the society cost model - partly estimated by regression and partly relying on HDM-4 submodels, see [HDM-4 (2000)] - VV distinguishes cost contributions from travel time, vehicle, comfort and accidents, through a valuation of the traffic effects in society-EUR.

According to VV, the HIPS maintenance costs cover the whole spectrum of real costs, from the lowest to the highest possible. Therefore we derive the costs of more extreme OPM-state variable values (outside the given data intervals) by projection onto the convex hull of the grid. Intermediate OPM-states are handled by the interpolation technique from Sec 2.3.2.

The stochastic HIPS cell transitions are replaced in OPM by deterministic expected statevariable values one year ahead. However, by our use of representative cell points, $44 \%$ of the expected values were detected unrealistic - and corrected. In our extensive checking we compared expected values for, e.g., a common action applied to two adjacent cells, the logic predicting a better initial state to be better off. These translation difficulties introduce some uncertainty about the results. Moreover, at the extraction from the large HIPS experimental database, we detected a lot of suspected errors, concerning missing or unknown or infeasible actions, illogically ordered road user costs and probabilities not summing up to 1 . Certain actions were prohibited in certain state cells, for unclear reasons. We identified a captive state, impossible to leave by any action, according to the transition probabilities. This db is experimental, so OK, but it makes us question the mere concept of immense data sets, in general, in putting heavy demands on data checking at input. A general remark is that a subdivision into three cell state levels per dimension does not allow conditions to get particularly bad: even in cells on the worst state level an eternal choice of routine maintenance does not produce particularly deterrent road user costs, etc.

The many question marks about the input data and the translation quality make us surrender our original thought of result comparisons with VV-HIPS. Such comparisons for the first year would be possible, although HIPS only presents probabilities for the various actions, since many probabilities are expected to be 1 , pointing out one particular initial action for all the segments of a road class, initially sharing a particular cell-state.

### 2.4.2 Discretization and interpolation experiments

In order to weigh the pros and cons of the two cost interpolation methods described in Sec2.3.2 and of different grid densities, we have performed experiments, using the data set described in Sec 2.4.1. We have, however, in these experiments applied ad hoc functions for the state transitions and costs in the basic grid points, see (2.8).

The ad hoc test functions that we used are ( $m=$ maintenance action, $r=$ road class, $s=$ segment, $d=$ state dimension):

$$
\begin{array}{rlr}
c_{s}(\mathbf{x}, m)= & (8-r) \cdot(m+1) \cdot\left(40+80 \cdot(m>0)+3 x_{1}+2 x_{2}+1 x_{3}\right) \cdot \frac{\text { Length }_{s} \cdot \text { Width }_{s}}{\text { averageWidth }} & \forall s, m: r=r(s) \\
& f_{s}(\mathbf{x})=(8-r) \cdot\left(300 x_{1}+200 x_{2}+100 x_{3}\right) \cdot \frac{\text { Length }_{s} \cdot \text { AADT }_{s}}{\text { averageAADT }} & \forall s: r=r(s) \\
& h_{r d}(\mathbf{x}, m)=\left\{\begin{array}{cl}
x_{d}+0.2 & m=0(\text { routine maintenance }) \\
x_{d} \cdot(1-m / 20) & m>0 \text { (rehabilitation action) }
\end{array}\right. & \forall s: r=r(s), d . \tag{2.8}
\end{array}
$$

The number of grid levels in each of the three state dimensions is taken as $2^{j}+1, j=2, \ldots, 5$, implying 6-7 times as many grid points for each $j$-step. The decision-cost interpolation, with $\left|\mathcal{U}_{r}\right|=4-5$, results in 8-10 times more $\bar{\varphi}$-interpolations than for linear interpolation, but the computational work in the DynP routine increases by less. Note that we have around $2 \cdot 10^{7}$ discrete decision-variables in the Lagrangean subproblems for the basic grid $j=2(T=15)$, and about $5 \cdot 10^{9}$ for $j=5$, quite sizeable numbers of discrete variables. For our experiments we use the initial dual prices $\mathbf{v}$ generated by the start routine, and compute the dual objective in two different ways, "Forward" and "Backward", for each grid level. In "Forward" the cost of the shortest paths are calculated by the DynP forward recursion, whereas in "Backward" their values are computed after the DynP backward recursion - by cost interpolation for the initial states $\mathbf{x}_{s 0}=\mathbf{a}_{0}$.


Figure 2.2 Calculated dual costs for two interpolation methods.
In Fig 2.2 we present the different dual objective values, scaled such that the reference case -decision-cost interpolation and $j=5$ - gets cost 1 . As can be seen, all "Forward" deviations are
very small - with a maximum of $0.22 \%$ for linear interpolation, $j=2$. In fact decision-cost interpolation for $j=2$ is superior to linear for $j=3$, e.g., the deviation ratio being 1:6. The "Backward" deviations are larger, indicating that discretization matters in some sense. This is further obvious for the choice of maintenance actions in Tab 2.2 below. The background of Tab 2.2 is that replanning is expected to be performed regularly, e.g. by a completely new set of RST-measurements every second year. We have counted the number of segments with action sequences different from the reference case during the $1^{\text {st }}$ year ( $1-\mathrm{yr}$ plan), during the first two years, etc. Out of 9225 segments, for a $1-\mathrm{yr}$ plan and $j=2$ ( 5 grid levels) more than $10 \%$ (987) deviate for linear, and more then $5 \%$ (492) for decision-cost interpolation. To accomplish similar deviation magnitudes for linear as for decision-cost interpolation we have to increase $j$ by 1 . Despite the many deviations between the plans, the dual objective values are essentially unaffected, as Fig 2.2 shows. Although the discretization has an influence on the computed costs, it looks like the deviations easily can be accepted economically - presumably mainly concerning segments with small deviations from the "true" optimal maintenance plans (e.g., rehabilitations moved one year or to an adjacent extent level, in comparison with the reference case). Computer runs like this, for different number of grid levels, can be useful in itself for pointing out the sensitive group of segments. For the rest, a majority of the segments, we expect the optimal plans to be more robust. As for the choice between the two cost interpolation methods and the different grid densities no definite conclusions can be drawn, but at least for this data set decision-cost interpolation and $j=2$ ( 5 grid levels) might do. The flat objective and the replanning philosophy are excuses for a moderate stopping criterion of the dual iteration process.

| Method | Linear cost interpolation | decision-cost interpol. |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
| \#grid levels | 5 | 9 | 17 | 5 | 9 | 17 |
| 1 yr plan | 987 | 570 | 357 | 492 | 319 | 121 |
| 2 yrs plan | 1911 | 1084 | 654 | 1102 | 589 | 253 |
| 3 yrs plan | 2751 | 1655 | 1015 | 1700 | 847 | 333 |

Table 2.2 Number of maintenance sequences different from the reference case: decision-cost interpolation, 33 grid levels.

### 2.4.3 Some results

The OPM method is implemented as a stand-alone C++ program. A first run of 94 iterations with the basic grid, the decision-cost interpolation, the trigger levels treated as constraints, the heuristics and an error checking in every iteration takes at present 55 hrs on a 1.6 GHz Pentium 4 PC. We use a Polyak relaxation coefficient, see [Polyak (1966)], geometrically decreasing with the iteration No. Iter as $2 \cdot 0.97^{\text {Iter }}$. Fig 2.3 gives the iteration histories for the best primal and dual values. Since problem (2.1) is non-convex a true duality gap surely exists - though small, due to the many segment contributions to the overall costs. For the dual prices from the start routine the initial deviation between the (initial) primal and dual objective values is $13 \%$. The remaining deviation satisfies the stopping criterion $\leq 0.1 \%$ (as $0.07 \%$ ) in iteration 94 , and is $0.04 \%$ after 150 iterations. By running 150 iterations we confirm that the subgradient procedure works, slightly improving the dual value, in 9 steps - although invisible in Fig 2.3, since the start routine works well: the difference between the dual values in iterations 0 and 150 is
only $0.05 \%$. However, due to our use of the basic grid - cf. Sec 2.4 .2 - such accuracy is somewhat illusionary. The difference between the dual objective function values calculated after the DynP forward and the backward recursions, respectively, are mostly of magnitude $2 \%$.


Figure 2.3 Evolution of best primal and dual values. Figure 2.4 Time evolution of costs and dual prices.

Fig 2.4 displays the overall costs and dual prices $v_{t}$ per year $t$ for the best achieved primal solution. For the road user costs, on average 75 MEUR/yr, to fit in, they are rescaled in Fig 2.4. The maintenance cost equals the budget level. The dynamics of the dual price and traffic cost evolution curves show that the initial road conditions are far from adjusted to the budget situation. The $v_{t}$-values are extremely high for the first years, much higher than the official tax factor $v_{B C R}$-level of magnitude one unit - cf. Sec 2.2.1 - a sign of neglected maintenance, or that the model does not catch all segment peculiarities. Yet we have chosen a budget (after yr 3) to get the long term $v_{t}$-values at a level with (or even somewhat below) $v_{B C R}$ and have performed the run for an initially increased budget level. A stationary $v$-level is not fully reached in 15 years.

The remaining dynamics is evident also from the time evolutions of two state-variables, IRI and Rut Depth, plotted in Fig 2.5a,b for two road classes, CI (AADT>12000) and M7 (AADT<250), as class averages and for one segment each. The final budget level admits improvements of the pavement conditions, although imperfections in the residual cost model might disturb the final evolution. Therefore ignore the values for the last years! The reduction of the Rut Depth averages is slower than for IRI, which is logical since the IRI-value (according to the input models) is more important for the total cost. The slowly expanding gap between the two average-curves in both a and b indicates an initial unbalance in road conditions between the two road classes, i.e. the suspected initial negligence from Fig 2.4 is tracked down to the heavy-traffic classes, such as CI, whereas the conditions of the sparsetraffic classes, like M7, are allowed to get worse. The trigger levels, IRI $6 \mathrm{~mm} / \mathrm{m}$ for segment 3893 and Rut Depth 25 mm for 9497, initiate one action each, in the low traffic class M7 (but
none in CI, as expected). By comparing Fig 2.5 a - b we realise that the rehabilitation frequencies vary within each traffic class.

The remaining dynamics at yr 15 in Fig 2.5 shows that the explicitly considered time horizon, 15 yrs, is too short. Since an infinite sum of equal annual costs $y$, discounted by factor $d$, amounts to $\frac{1}{1-d} \cdot y$, a rule of thumb is to use a $\frac{1}{1-d}$-yrs horizon, in our case the $4 \%$ discount rate $(d \approx 0.96)$ corresponding to 25 yrs. However, in the NOS- and HIPS- articles we have encountered, much shorter dynamic periods, of magnitude 10 yrs , are used.

As for the use of different maintenance actions, their cost portions are shown in Fig 2.6a,b for the best primal solution. Fig 2.6a depicts the maintenance costs ( $c$, grey shaded) and traffic costs ( $f$, white), as percentages of the regional totals, in common action order. The subdivided road user cost refers to the year when the respective action is applied. Surprisingly The Thick Layer and Reconstruction options are hardly visible. Apart from this result the picture is scattered: whereas Simple action is predominant in the low and high AADT classes, the more extensive Thin layer action stands for the major costs in the intermediate AADT classes. The contributions add up to an optimal split of 1:5 between routine maintenance and rehabilitation actions - cf. Sec 2.2.1. For class M7 of narrow, low-AADT roads, the total maintenance cost, portion $27 \%$, is slightly below its length proportion, $34 \%$ (shown next to the name), whereas the total traffic cost, $15 \%$, is far below (as expected): In the sparse-traffic classes the acceptance limits, rather than economy, determine the choice of actions (cf. Fig 2.5). Fig 2.6b displays the same information as Fig 2.6a, when the different road lengths (in metres) in each road class are accounted for. The plot data are received in two steps:
(1) Express the costs in EUR/m. The Värmland cost averages are 410 EUR $/ \mathrm{m}$ (road user, where the residual value represents more than $50 \%$ ) and $42 \mathrm{EUR} / \mathrm{m}$ (maintenance).
(2) Normalise such that both the Värmland cost averages correspond to 1.0 .

The road user cost per length unit in M7 is 8 times that of CI, whereas the maintenance costs (surprisingly) are of equal magnitude, somewhat more on the intermediate classes than on the high-AADT classes. We had expected the opposite relation. Thus a normalisation by road area instead of length, taking into account the wider roads in the more frequented traffic classes, would show that less money per $\mathrm{m}^{2}$ should be spent on these classes, according to this optimisation.


Figure 2.5a Time evolution of IRI.


Figure 2.6a Maintenance actions per road class.


Figure 2.5b Time evolution of Rut Depth.


Figure 2.6b Maintenance actions per road class.

### 2.5 Outlook

These first results indicate that our approach is feasible, right now on the regional planning level, although the runtimes need reductions, from both method and implementation improvements. The dual separability makes parallel computing a natural option, but in that case the intensely communicating primal heuristics - using simultaneous information about the current status of all segments - needs revision. The data translation difficulties documented in Sec 2.4.1 are mainly due to the large amount of cell based data, and might be overcome by, e.g., turning to function based descriptions of costs and state transitions, similar to those in (2.8). In general we hope to be part of a process of interaction between input and output data improvements. The many short road segments in Värmland either point towards some consideration of coordinated maintenance or an aggregation of segments.

## 3 Main study: Problem description

In Ch 1 we gave a conceptual description of the optimisation problem. The components to be specified here are the building blocks that were introduced in Sec. 1.1. As building blocks these components are replaceable: our optimisation method and computer program allows (in principle) a free choice of state variables, state evolution models and cost functions, as well as free combinations of discrete works types and continuous works extents (e.g. layer thickness). The practical limitations come from the increasing computing times of the more complex models, as well as the access to relevant, reliable data. The description here is a first attempt to catch the most important, easily available problem characteristics. The input data and functions in Sec 3.1-3.7 are mainly specified by VV, as documented in [Andersson et al (2004)]. Also see [Effektsamband 2000 (2001b)].

### 3.1 Road data

The dominant part of the input data concerns specific information about each road segment in the network considered. To be manageable, such a large data set must be almost automatically generated, in our case from the VV-PMS database. The laser-RST measuring and analysis (see, e.g., [Forsberg and Göransson (2000)]), regularly performed since 1987, are the keys to the data quality. A few question marks are noted below.

The road segments are aggregated into road classes. This subdivision of the road network is a consequence of the other input. In case, e.g., a general cost functional form applies to the whole database but there are function parameters - with values that differ between different subsets of segments - the principle is that one and the same set of parameter values should apply to all segments in a road class.

### 3.1.1 Road segments

Our analysis concerns PMS-database selection County Värmland, autumn 2002.
By definition the segments are homogeneous as to their pavement histories, and of varying length - see Tab 3.1. We exclude a segment from the optimisation, if

- the registered year of latest major maintenance operation is 1899 (excluded number of segments* 619 ( $6.5 \%$ ), excluded length $27.1 \%$; motivation: probable gravel road or other road administrator than VV)
- AADT (Annual Average Daily Traffic) traffic data are missing (excluded number* 135 (1.4\%), excluded length $0.5 \%$ )
- a road constant is missing or a pavement piece of information cannot be estimated by interpolation of neighbouring segments (excluded number 0).
Remaining (net) number of segments: 8749 (92.1\%), length: $3625.5 \mathrm{~km}(72.4 \%)$.
*) In the Initial study, Ch 2, these segments were not automatically excluded. Instead data were constructed by interpolation based on the neighbouring segments, whenever possible.


### 3.1 Road data

| Length class $(\mathrm{m})$ | -5 | -10 | -20 | -50 | -100 | -200 | -400 | -800 | -1500 | -3000 | $-\infty$ |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Frequency | 687 | 429 | 607 | 1144 | 1108 | 1380 | 1214 | 940 | 638 | 401 | 201 |
| Total length(\%) | 0.05 | 0.09 | 0.26 | 1.08 | 2.31 | 5.57 | 9.55 | 14.77 | 19.04 | 22.98 | 24.28 |

Table 3.1 Distribution of segment lengths, Värmland.

### 3.1.2 Road classes

In Värmland 8 AADT-based traffic classes are distinguished:

| 1 | $0-250$ | 5 | $2000-4000$ |
| :--- | ---: | :--- | :--- |
| 2 | $250-500$ | 6 | $4000-8000$ |
| 3 | $500-1000$ | 7 | $8000-12000$ |
| 4 | $1000-2000$ | 8 | $12000-$ |

The road classes are formed as every possible combination of traffic class and speed limit, where 30 (t for short), $50(\mathrm{f}), 70(\mathrm{~s}), 90(\mathrm{n}), 110(\mathrm{~h}) \mathrm{km} / \mathrm{h}$ are distinguished; 40 road classes in all. In Värmland there are segments for 29 of these combinations.

If a larger geographical region had been considered, also the climate zone would have had an influence on costs and effects, and on the road class concept.

### 3.1.3 Road constants and pavement data

In order to identify road segments and to correct for missing data (by interpolation) we use

- county
- road number
- lane (Belelem in the database)
- driving direction (Lrollord)
- accumulated length from a fixed calibration point (Lopa; missing number: 23 (0.2\%)).

As road constants in the optimisation we take

- segment length
- road width (missing number: 39 (0.4\%))
- traffic (see Sec 3.1.1)
- heavy vehicle traffic (lorries and buses)
- speed limit (missing number: 47 ( $0.5 \%$ )).

As initial segment conditions for the optimisation we assume

- year of latest major maintenance operation
- latest laser-RST-measuring year (missing number: 674 (7.7\%))
- measured IRI-average (see laser-RST year)
- regression parameter: IRI-value immediately after the latest major maintenance operation
- regression parameter: annual change of IRI-value
- measured rut depth average (see laser-RST year)
- regression parameter: rut depth immediately after the latest major maintenance operation
- regression parameter: annual change of rut depth.

Correction-completion (corrected number of segments: 677 (7.7\%)):

- measured IRI-average is missing: apply the IRI-regression line from the year of latest major maintenance operation
- measured rut depth average is missing: apply the rut depth regression line from the year of latest major maintenance operation
- (other) value is missing: interpolate data for the surrounding segments on the same road.


### 3.2 Pavement state

In the optimisation model the pavement condition is completely defined by the values of the state variables considered. In practice more aspects than those (ever?) being modelled is important for the pavement decisions. In this respect our variable Age is partly a proxy for neglected circumstances, e.g. texture, structure, and drainage.

### 3.2.1 State variables

The dimensions of our 5-dimensional state space stand for

- current IRI-value (IRI)
- IRI deterioration rate ( $\Delta I R I$ )
- current rut depth ( $R D$ )
- rut depth deterioration rate $(\Delta R D)$
- age (Age).

A description of IRI, an index of longitudinal unevenness (roughness), and rut depth, a measure of transversal unevenness, can be found in, e.g., [Ihs and Sjögren (2003)], also cf. Sec 1.1. Age means how old the road surface is from a maintenance point of view. It coincides with the pavement age for layer thickness 100 mm (see 3.5 .5 below). A choice of this value - the upper thickness limit in the model - implies the greatest possible "respite" (maximum latency period, here 5 years according to 3.6 .3 below) after a major maintenance operation, before degradation necessitates routine maintenance. For thinner pavement layers the degradation clock is put forward, so that the Age-value at 25 mm pavement will generate immediate routine maintenance costs. For the thinnest pavement layers the initial Age-value after major operations depends on the Age-value before. Our assumption here is that one and the same concept, the maintenance based age, is appropriate for a description of both the costs characterising routine maintenance and any trigger (acceptance) limit for the pavement ageing.

As initial Age-values for the segments we use the time from the latest major maintenance operation.

### 3.2 Pavement state

### 3.2.2 State limits

In the optimisation run the upper state variable bounds are interpreted in any of three ways: - absolute (must-) limits, with any violation forbidden,

- relative (ought to-) limits, any violation being penalised as an extra traffic cost,
- free (watch-) limits, used for statistical purposes only.

The choice of interpretation can be made differently for different state dimensions.
The motivations for state limits are fourfold:

- safe trafficability
- durability
- fairness
- cost effectiveness.

Our objective deviates from the listed VV-goals [Potucek and Rydén (2003)] insofar as the economic aspects are taken care of in the optimization routine.

In an optimisation run we structure our calculations by a grid of nodal states, defined by a userspecified or default subdivision of the admissible states per state dimension, free as to both the position and the number of node levels in each dimension.

### 3.2.3 IRI-value limits

The informal VV-standard for 20 m - and 400 m -averages has been completed to the following upper IRI-limits:

| AADT | >= $90 \mathrm{~km} / \mathrm{h}$ |  | <= $70 \mathrm{~km} / \mathrm{h}$ |  |
| :---: | :---: | :---: | :---: | :---: |
|  | 20 m | 400 m | 20 m | 400 m |
| > 4000 | 4 | 3 | 5 | 4 |
| 2000-4000 | 4.5 | 4 | 5.5 | 5 |
| <= 2000 | 5.5 | 5 | 6.5 | 6 |

Table 3.2 IRI-value limits ( $\mathrm{mm} / \mathrm{m}$ ) for different road classes and standard segment lengths.
As for segments of deviating length $l$ we use the IRI-limit $\bar{x}_{c}(l)=\mu_{0}+\frac{\mu_{1}}{\sqrt{l}}$, where the parameters $\mu_{0}, \mu_{1}$ are determined by curve fitting. See Sec 3.8.1 below.
Let $\Delta:=\bar{x}_{c}(20)-\bar{x}_{c}(400)$. Then it can be shown: $\mu_{1}=5.76 \Delta, \mu_{0}=\bar{x}_{c}(400)-0.288 \Delta$.

### 3.2.4 Rut depth limits

The informal VV-standard 20 mm is used for all road classes and segment lengths.

### 3.2.5 Age limits

Since the cost of routine maintenance increases by Age (see Sec 3.6 .2 below), major maintenance finally become economically beneficial. We use the state limit as an extra
safeguard against unrealistically long maintenance intervals and as a reasonable durability time for asphalt layers.

| Traffic class | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Maximum Age | 31 | 31 | 31 | 29 | 25 | 17 | 15 | 14 |

### 3.3 Maintenance

Among several term options, e.g. treatment and action, we have chosen the (somewhat more general) word works to describe the maintenance performance, also used in e.g. [HDM-4 (2000), p D2-34]. Consequently we speak about different works types (discrete variable) and different works extents (continuous variable, for amount).

Often, see e.g. [ibid, p D1-4], the works categorisation is made according to the financial sources. However, we assume that all maintenance is subject to a common budget. As for the major pavement works, variations of the layer thickness will stand for the most important cost variation. Our optimisation program will implicitly consider myriads of potential maintenance projects, candidating for implementation. Based on our limited input information, about IRI and rut depth only, we find it difficult to differentiate between more specified pavement works types. Such a differentiation is typically made after a maintenance project has been identified, e.g. by our computer program, and scrutinised.

The layer thickness is an example of a continuously varying pavement works extent. The user may control the accuracy of the floating-point calculations by specifying an acceptance limit for the works extent, here the maximum deviation (in mm ) from the optimal layer thickness.

### 3.3.1 Works types

Contrary to the initial study (Ch 2) we only use two types here, corresponding to two discrete variable values:

- routine maintenance (the no-action option, do the minimum of patching, crack sealing and edge repair),
- major maintenance (collective for all asphalt-paving).


### 3.3.2 Works extent: Layer thickness

For major maintenance we operate with a continuously varying layer thickness, of size 10-100 mm . As an interpretation of the different values we suggest:
$10-25 \mathrm{~mm}$ resurfacing (sealing/surface treatment, inlay),
$25-45 \mathrm{~mm}$ rehabilitation (mill \& replace, overlay of asphalt mix),
$45-100 \mathrm{~mm}$ reinforcement (of pavement base, plus overlay/surface treatment).

### 3.3 Maintenance

### 3.3.3 Restrictions on layer thickness

An initial validation of our model showed that the optimisation program frequently chooses (cheap) thin asphalt layers and avoids the thick layer options. A similar experience was referred to in Sec 1.4.2. The pavement state resulting from thin layers is controlled by an interpolation with the current state (see Sec 3.5 .5 below), which may be part of the trouble. Or the state transition models may favour thin layers economically. Another important aspect is that the accumulated effect of several thin layers is not technically equivalent to one thick layer, and is avoided in practice. We reduce this loophole by specifying a general lower thickness limit depending on the current Age-value $t$ and the upper Age-limit $T$, according to Sec 3.2.5:

Thickness $\geq 10 \cdot 2^{2 t / T}(\mathrm{~mm}) ;$
which means that the thickness must exceed 20 mm at half the Age-limit, and exceed 40 mm at the very limit - but the thin layer option 10 mm , also of repeated use, remains for "fresh" segments, of Age $t=0$.

### 3.4 Traffic effects and costs

The purpose is to value all the public traffic effects that are ascribable to a varying pavement quality, by measuring the additional traffic (road user) costs of departing from an ideal pavement state. Traditionally the maintenance is considered to have an influence upon travel time, vehicle operating cost, road safety, discomfort and environment. In this study only the first two contributions are considered.

The VV model is an adjustment of an international standard, the World Bank HDM-4 model [HDM-4 (2000)], to Swedish conditions. We base our model on the VV-specification [Odermatt (2001)]. Also see [Effektsamband 2000 (2001b)].

The traffic costs, as given in Appendix 1, include tax factor I, currently an additional 23\%. Tax factor I is an indirect cost, an average VAT-percent, included in the translation from producers' costs to consumers' costs, to reflect the consumers' willingness to pay for a product. (Although the resulting road user cost is a consumers' cost it involves goods and services that are subject to indirect costs.)

### 3.4.1 Travel time cost

See Appendix 1.1: Travel time costs.

### 3.4.2 Vehicle operating cost

See Appendix 1.2: Vehicle operating costs.

### 3.5 Maintenance effects (state transitions)

All the models describing the influence of maintenance upon the pavement state are estimated by applying multiple linear regressions to a historical data material on Värmland. The point estimates are for the averages along the segment. We comment on the results in Secs 3.8.2 and 3.9.3 below.
3.5.1 IRI-value immediately after a major maintenance operation, $I R I_{\text {after }}$

$$
I R I_{\text {after }}=c_{0}+c_{1} \cdot I R I_{\text {before }}+c_{2} \cdot \operatorname{Thickness}(\mathrm{~mm})
$$

$I R I_{\text {before }}$ refers to the IRI-average measured or estimated immediately before the operation.

| Traffic class | $n$ | $\mathrm{c}_{0}$ | $\mathrm{C}_{1}$ | $\mathrm{c}_{2}$ | Traffic class | $n$ | $\mathrm{c}_{0}$ | $\mathrm{c}_{1}$ | $\mathrm{C}_{2}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 7091 | 1.750507 | 0.297504 | -0.023511 | 5 | 7569 | 0.887289 | 0.263716 | -0.007945 |
| 2 | 7019 | 1.609943 | 0.29548 | -0.025013 | 6 | 8279 | 0.671925 | 0.349755 | -0.004974 |
| 3 | 8319 | 1.312917 | 0.277508 | -0.01808 | 7 | 3354 | 0.614376 | 0.377965 | -0.002479 |
| 4 | 7687 | 1.068788 | 0.266447 | -0.012363 | 8 | 2434 | 0.453251 | 0.479973 | -0.000951 |

Table 3.3 Sample size $n$ and parameters in $I R I_{\text {after }}$ regression model for different traffic classes. Data: Värmland. $\mathrm{R}^{2}$-values: 26-61 \%.
Limitation rule: $I R I_{\text {after }}$ must not exceed $I R I_{\text {before }}$.

### 3.5.2 IRI-deterioration rate after a major maintenance operation, $\Delta I R I_{\text {after }}$

The degradation rate (the annual IRI-increase) is assumed constant until the next major operation.

$$
\Delta I R I_{\text {after }}=c_{0}+c_{1} \cdot \Delta I R I_{\text {before }}+c_{2} \cdot I R I_{\text {after }}
$$

$\Delta I R I_{\text {before }}$ denotes the estimated degradation rate (immediately) before the operation.

| Traffic class | $n$ | $\mathrm{c}_{0}$ | $\mathrm{c}_{1}$ | $\mathrm{c}_{2}$ | Traffic class | $n$ | $c_{0}$ | $c_{1}$ | $c_{2}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 22 | -0.10859 | 0.445235 | 0.061882 | 5 | 293 | 0.067163 | 0.097267 | -0.001467 |
| 2 | 78 | 0.006928 | 0.126235 | 0.042098 | 6 | 148 | -0.071865 | 0.025507 | 0.089369 |
| 3 | 136 | 0.024607 | 0.083412 | 0.043083 | 7 | 71 | -0.121604 | 0.309196 | 0.114985 |
| 4 | 172 | 0.014645 | 0.114265 | 0.050106 | 8 | 25 | -0.116898 | 0.009766 | 0.107222 |

Table 3.4 Sample size $n$ and parameters in $\Delta I R I_{\text {after }}$ regression model for different traffic
classes Data: Värmland. Degrees-of-freedom corrected $R^{2}$-values: 3.9-43 \%.
Limitation rules: $\Delta I R I_{\text {after }}$ must not exceed $\Delta I R I_{\text {before }}$ and not be below 0.02 .
After the initial runs it became obvious that the restriction $\Delta I R I_{\text {after }} \leq \Delta I R I_{\text {before }}$ is unrealistic, since as soon as the ideal value 0.02 is reached, it costs nothing to maintain it. We abandoned the restriction.

### 3.5 Maintenance effects (state transitions)

3.5.3 Rut depth immediately after a major maintenance operation, $R D_{\text {after }}$

$$
\begin{equation*}
R D_{\text {after }}=c_{0}+c_{1} \cdot R D_{\text {before }}+c_{2} \cdot \operatorname{Thickness}(\mathrm{~mm}) \tag{3.1}
\end{equation*}
$$

$R D_{\text {before }}$ refers to the rut depth average as measured or estimated immediately before the oper.

| Traffic class | $n$ | $\mathbf{c}_{0}$ | $\mathbf{c}_{1}$ | $\mathbf{c}_{2}$ | Traffic class | $n$ | $\mathbf{c}_{0}$ | $\mathbf{c}_{1}$ | $\mathbf{c}_{2}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 7091 | 3.872777 | 0.241195 | -0.06833 | 5 | 7569 | 4.57046 | 0.033296 | -0.05119 |
| 2 | 7019 | 4.076478 | 0.212433 | -0.07664 | 6 | 8279 | 4.063815 | 0.061473 | -0.03608 |
| 3 | 8318 | 4.122393 | 0.143381 | -0.05826 | 7 | 3354 | 4.281419 | 0.048564 | -0.03854 |
| 4 | 7687 | 4.308123 | 0.070647 | -0.05265 | 8 | 2434 | 5.523075 | -0.01787 | -0.04959 |

Table 3.5 Sample size n and parameters in $R D_{\text {after }}$ regression model for different traffic
classes. Data: Värmland. R2-values: 4.9-28 \%.
Limitation rule: $R D_{\text {after }}$ must not exceed $R D_{\text {before }}$.

### 3.5.4 Deterioration rate of rut depth after a major maintenance operation, $\triangle R D_{\text {after }}$

The degradation rate (the annual rut depth increase) is assumed constant until the next major operation.

$$
\Delta R D_{\text {after }}=c_{0}+c_{1} \cdot \Delta R D_{\text {before }}+c_{2} \cdot R D_{\text {after }}
$$

$\Delta R D_{\text {before }}$ denotes the estimated degradation rate (immediately) before the operation.

| Traffic class | $n$ | $\mathrm{c}_{0}$ | $\mathrm{c}_{1}$ | $\mathrm{c}_{2}$ | Traffic class | $n$ | $\mathrm{c}_{0}$ | $\mathrm{c}_{1}$ | $\mathrm{c}_{2}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 11 | 0.387981 | -0.13582 | 0.040183 | 5 | 253 | 0.508778 | 0.122527 | 0.004533 |
| 2 | 44 | 0.454219 | 0.36949 | -0.20226 | 6 | 136 | 0.492265 | 0.233807 | 0.051564 |
| 3 | 98 | 1.355313 | 0.149224 | -0.02982 | 7 | 60 | 0.16946 | -0.164285 | 0.210995 |
| 4 | 134 | 0.987111 | -0.082187 | 0.000904 | 8 | 23 | 0.891889 | -0.049247 | -0.057367 |

Table 3.6 Sample size $n$ and parameters in $\Delta R D_{\text {after }}$ regression model for different traffic classes. Data: Värmland. Degrees-of-freedom corrected R ${ }^{2}$-values: 0.3-39 \% . Limitation rules: $\Delta R D_{\text {after }}$ must not exceed $\Delta R D_{\text {before }}$ and not be below 0.3 .

For the same reason as in $\operatorname{Sec} 3.5 .2$ we abandoned the restriction $\Delta R D_{\text {after }} \leq \Delta R D_{\text {before }}$.
Our initial runs showed quite many non-logical relations as to the expected future costs in different pavement states: We expect that out of two states for the same road class the currently better one should have lower future costs than the other. By simply putting the four negative $c_{1}$-coefficients to zero we have managed to reduce such anomalies considerably. Cf. Sec 3.9.3.

### 3.5.5 Age immediately after a major maintenance operation, $A g e_{\text {after }}$

The Age-concept was introduced in Sec 3.2.1. The thicker pavement, the longer latency time is generated before the degradation starts and routine maintenance is needed (from Age = 5 and onwards, according to Sec 3.6 .2 below). The parameters are specified in the figure below.

Below thickness 25 mm the Age-value is computed by linear interpolation with Age $_{\text {before }}$, the Age-value immediately before the operation - see Fig 3.1.


Figure 3.1 Age $_{\text {after }}$ vs. thickness model, illustrated for Age $_{\text {before }}=20$ years.

### 3.5.6 Updated IRI-value, rut depth and age after one year of routine maintenance

Since the degradation rates are assumed constant between major maintenance operations, the annual changes are confined to

$$
\begin{aligned}
& I R I:=I R I+\Delta I R I \\
& R D:=R D+\Delta R D \\
& \text { Age }:=\text { Age }+1
\end{aligned}
$$

### 3.6 Maintenance costs

Apart from the given cost values below, the maintenance costs are subject to add-ons by tax factor I - see Sec 3.4 - and tax factor II, currently $30 \%$, i.e. $53 \%$ in all. Tax factor II is a tax financial addition, applied to all investments that are financed by public Swedish budget means, reflecting the loss of efficiency in the Swedish economy that results from publicly financed activities, cf [Andersson et al (2001)].

Our maintenance cost models do not differentiate between posted speed limits, although these as well might have an influence on the choice of pavement quality - and costs.

### 3.6.1 Fixed major maintenance costs

Every maintenance project triggers a fixed initial/setup cost $K_{1}$, which depends on the traffic class according to Tab 3.7. This cost $K_{1}$ stands for general establishment and traffic-regulation expenses. Apart from $K_{1}$ an additional milling cost $K_{2}$ for asphalt removal is generated, per section of connected road segments simultaneously maintained. In practice a maintenance project may consist of several such sections, each one charged $K_{2}$, according to Tab 3.7. (In this model no extraordinary road user cost is taken into account.)

| Cost\|AADT | $0-250$ | $250-1000$ | $1000-4000$ | $4000-$ |
| :---: | ---: | ---: | ---: | ---: |
| $K_{1}($ SEK $)$ | 15000 | 15000 | 20000 | 25000 |
| $K_{2}($ SEK ) | 0 | 2000 | 5000 | 10000 |

Table 3.7 Fixed costs in different traffic classes.

### 3.6.2 Variable major maintenance costs

These costs are modelled proportional to the applied pavement volume. The basic cost 900 $\mathrm{SEK} / \mathrm{m}^{3}$ is modified according to both the road traffic and the pavement state immediately before the operation.

$$
\operatorname{Cost}\left(\mathrm{SEK} / \mathrm{m}^{2}\right)=0.9 \cdot \text { Thickness }(\mathrm{mm}) \cdot \rho_{\text {Traf }} \cdot\left(1+\frac{p_{I I I}}{100}\right) \cdot\left(1+\frac{p_{R D}}{100}\right)
$$

$\rho_{\text {Traf }}$ is a traffic dependent correction factor - see Fig 3.2.


Figure 3.2 Correction factor $\rho_{\text {Traf }}$ in variable cost model vs. traffic volume.
$p_{I R I}, p_{R D}$ denote percentage add-ons for preparatory works according to the IRI-value and rut depth immediately before the major maintenance operation. For state value $x$ (either $I R I$ or $R D$ ) immediately before the operation, the addition (in \%) has the piecewise linear form

$$
p=\left\{\begin{array}{cl}
0 & x \leq x_{0} \\
p_{1} \cdot\left(x-x_{0}\right) & x>x_{0}
\end{array}\right.
$$

where the parameter values $p_{1}$ (rate of increase, in $\%$ ) and $x_{0}$ ( $x$-activation point) according to Tab 3.8 give $p$ expressed in percent. As an example the $p$-value (in \%) for the IRI-value $x=10$ $(\mathrm{mm} / \mathrm{m})$, see Sec 1.1 for a comment on the unit, and for the rut depth $x=30(\mathrm{~mm})$, respectively, are presented.

| Traffic class | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| IRI $p 1$ | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $x 0$ | 4.8 | 4.4 | 4 | 3.6 | 3.2 | 2.8 | 2.4 | 2 |
| $p(x=10)$ | 26 | 28 | 30 | 32 | 34 | 36 | 38 | 40 |
| RD $p 1$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $x 0$ | 17 | 16 | 15 | 14 | 13 | 12 | 11 | 10 |
| $p(x=30)$ | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |

Table 3.8 Parameters in the preparatory works model for different traffic classes.

### 3.6.3 Cost of routine maintenance

The no-action cost increases exponentially, starting from 0 at Age $=5$, passing 1.79 at Age $=15,3.21$ at Age $=25$ and 5.74 at Age $=35$.

$$
\operatorname{Cost}\left(\mathrm{SEK} / \mathrm{m}^{2}\right)=(1+0.06)^{\max [\text { Age- }-0]}-1 .
$$

### 3.7 General data

A budget (in SEK) has to be specified for each year. Since we ignore important triggers in our state definition, this is to be considered as part of the whole maintenance budget for a year. An a priori separation between budgets for routine maintenance (operations budget) and major maintenance (maintenance budget) should not be made: the most cost-efficient subdivision is part of the optimisation.

### 3.7.1 Interest rate

A $4 \%$ annual interest rate is applied to the discounting of costs back to the base time January, 2003.

### 3.7.2 Capital scarcity factor

The budget means are to be spent on economically beneficial maintenance projects. In principle, see [Effektsamband 2000 (2001b), pp 37, 74], VV uses the cut-off value 1.0 for the benefit/cost-ratio, for the comparison between a candidate project and the no-action alternative. Since tax factor I is included in the traffic costs in Sec 3.4, and tax factors I + II are to be

### 3.7 General data

applied to the maintenance costs in Sec 3.6 in such a comparison, the cut-off value $1.0 \cdot 1.53 / 1.23=1.2439$ is consistent with our cost scales.

### 3.8 Stochastic submodels

### 3.8.1 Interpolation of state limits

For the acceptable pavement condition VV has specified target standards, in the form of state variable limits, cf. [Potucek and Rydén (2003)]. The IRI-value limits are given per road class as 20 m - and 400 m -averages in Sec 3.2.3. In order to determine relevant limits for road segments of other lengths the following interpolation-extrapolation procedure can be applied. We perform the analysis by using elementary statistical concepts - see, e.g., [Devore (2004)].

Consider the (laser-RST) measurement data for a state variable, e.g. IRI, registered at a certain time on a certain road segment. Look upon the data as a set of $n$ values $\left(x_{i}\right)_{i=1}^{n}$, determined every $e$ metres, with $e$ constant, so that the segment length $l$ fulfils $l \approx n \cdot e$. Thus for $e \approx 1$ the number of values coincides with the length in metres, $l \approx n$. Consider the measurement data set as a random sample of a continuous stochastic variable $X$, for measurements made at different times and/or in different (laser-RST) settings. We want to estimate the "true" mean state (IRIvalue) $\mu$ for the segment, from the data set; in presence of random noise, with a standard deviation $\sigma$ that either is known or can be estimated from the sample. For the state there exists a specified target standard $\mu_{0}$, such that if $\mu>\mu_{0}$ the road segment should be maintained, otherwise not. The question about the true segment state can be answered by a significance test, where the null- and alternative hypotheses are formulated as

$$
H_{0}: \mu \leq \mu_{0}, H_{1}: \mu>\mu_{0} .
$$

We reject $H_{0}$ if the sample mean $\bar{X}$ exceeds a certain limit $\bar{x}_{c}=\bar{x}_{c}(l)$ depending on the length $l$ (through the sample size $n$ ). This critical limit is fixed by the prescribed level of significance $\alpha$ from a general probability relation, where $P$ denotes probability:

$$
P\left[\bar{X} \leq \bar{x}_{c} \mid \mu=\mu_{0}\right]=1-\alpha .
$$

Especially if $X$ is Normal distributed, $X \in N(\mu, \sigma)$, we have $\bar{X} \in N(\mu, \sigma / \sqrt{n})$. If $X$ obeys another type of distribution with mean $\mu$ and standard deviation $\sigma$, we get $\bar{X} \in N(\mu, \sigma / \sqrt{n})$ as an approximation, according to the central limit theorem. (For symmetrical base distributions the approximation is good even for small samples, e.g. $n \geq 10$ for rectangular distributions.) Our additional assumption is $\bar{X} \in N(\mu, \sigma / \sqrt{n})$, at least as an approximation. If $\sigma$ is known, $\bar{x}_{c}$ is determined by the standard-Normal distribution value $\Phi_{1-\alpha}$ according to

$$
\frac{\bar{x}_{c}-\mu_{0}}{\sigma / \sqrt{n}}=\Phi_{1-\alpha}
$$

which for $\mu_{1}:=\Phi_{1-\alpha} \cdot \frac{\sigma}{\sqrt{e}}$ turns into

$$
\bar{x}_{c}=\mu_{0}+\Phi_{1-\alpha} \frac{\sigma}{\sqrt{l \cdot e}}=\mu_{0}+\frac{\mu_{1}}{\sqrt{l}} .
$$

If $\sigma$ is estimated by the sample standard deviation $s, \bar{x}_{c}$ is determined by the Student-tdistribution value $t_{1-\alpha}$ according to

$$
\frac{\bar{x}_{c}-\mu_{0}}{s / \sqrt{n}}=t_{1-\alpha}
$$

which for $\mu_{1}:=t_{1-\alpha} \cdot \frac{s}{\sqrt{e}}$ turns into

$$
\bar{x}_{c}=\mu_{0}+t_{1-\alpha} \frac{s}{\sqrt{l \cdot e}}=\mu_{0}+\frac{\mu_{1}}{\sqrt{l}}
$$

Hence the critical limit $\bar{x}_{c}=\bar{x}_{c}(l)$ takes the same mathematical form as a function of the length $l$, whether $\sigma$ or $s$ is viewed as given. From the two given state limits, $\bar{x}_{c}$ (20) and $\bar{x}_{c}$ (400) for $l=20$ and 400 m , respectively, a state limit can be computed for every possible segment length $l$. A system of two equations has to be solved:

$$
\left\{\begin{array}{c}
\bar{x}_{c}(20)=\mu_{0}+\frac{\mu_{1}}{\sqrt{20}} \\
\bar{x}_{c}(400)=\mu_{0}+\frac{\mu_{1}}{\sqrt{400}}
\end{array}\right.
$$

We let the difference be denoted $\Delta:=\bar{x}_{c}(20)-\bar{x}_{c}(400)$, whence the solution can be written

$$
\left\{\begin{array}{c}
\mu_{1}=5.76 \Delta \\
\mu_{0}=\bar{x}_{c}(400)-0.288 \Delta
\end{array}\right.
$$

As an example we apply the formulas for the IRI-value limit to the road classes characterized by AADT $>4000$ and posted speed limit $\geq 90$. With target values as in Sec 3.2.3, $\bar{x}_{c}(20)=4.0$ and $\bar{x}_{c}(400)=3.0$, we get

$$
\begin{aligned}
\Delta & =\bar{x}_{c}(20)-\bar{x}_{c}(400)=1.0, \\
\mu_{1} & =5.76 \Delta=5.76 \\
\mu_{0} & =\bar{x}_{c}(400)-0.288 \Delta=2.712 .
\end{aligned}
$$

The state limit $\bar{x}_{c}(l)=2.712+\frac{5.76}{\sqrt{l}}$ is illustrated in Fig 3.3 for different segment lengths $l$.

Figure 3.3 IRI acceptance limit vs. segment length.


### 3.8.2 Stochastic state transitions

We saw in Sec 3.5 that the deviations from the regression effect models are rather large, if measured by the $R^{2}$-values.

The regression models are implicitly based on the assumption that the residuals are Normal distributed. In consequence, an optional stochastic model is conceivable, where, e.g., $R D_{\text {after }}$ (cf. Sec 3.5.3) is considered as random, $R D_{\text {after }} \in N(\mu, \sigma)$. Stressing this, formula (3.1) in Sec 3.5.3 is a regression equation and its LHS is the expected value $\mu$, in this case a function of the regression variables $R D_{\text {before }}$ and Thickness. Whereas $\mu$ is estimated from (3.1) the estimation of the variance $\sigma^{2}$ is less obvious. We believe that the regression coefficients in Tab 3.5 have been found by least squares, weighted by the segment lengths.

Our intention is to use the remaining statistical uncertainty, in the form of the standard deviations $\sigma$ for the residuals, i.e. the differences between the regression model and the data, for an extension of the basic model - cf. Ch 7 below. However, we have no access to the regression input data. Thus for each model we have to rely on the standard regression output, e.g., the resulting residual sum of squares $q_{\text {RES }}$, number of observations $n$, total road length $L$ and number of regression variables $k$. We start by deriving an appropriate variance estimate.

Collect the (unknown) segment lengths $\left(\ell_{i}\right)_{i=1}^{n}$ in the $n \times n$ weight matrix $\mathbf{W}=\operatorname{diag}\left(\ell_{i}\right)$. Assume that the $n$ documented statistical trials of the dependent stochastic variable, e.g., $R D_{\text {after }}$ satisfy $\mathbf{Y} \in N_{n}\left(\boldsymbol{\mu}_{Y}, \mathbf{C}_{Y}\right)$, with mean vector $\boldsymbol{\mu}_{Y}=\mathbf{X} \boldsymbol{\beta}$ and covariance matrix $\mathbf{C}_{Y}$. Here $\boldsymbol{\beta}$ is the $k$-parameter vector and $\mathbf{X}$ is the observation matrix. Introducing the estimator vector $\mathbf{b}$ we write the residual sum of squares

$$
Q_{R E S}=\min _{\mathbf{b} \in \Re^{n}}(\mathbf{Y}-\mathbf{X b})^{T} \mathbf{W}(\mathbf{Y}-\mathbf{X b})
$$

By letting $\widetilde{\mathbf{Y}}:=\mathbf{W}^{1 / 2} \mathbf{Y}, \widetilde{\mathbf{X}}:=\mathbf{W}^{1 / 2} \mathbf{X}$ we turn the weighted problem into an unweighted one:

$$
Q_{R E S}=\min _{\mathbf{b} \in \Re^{n}}(\widetilde{\mathbf{Y}}-\widetilde{\mathbf{X}} \mathbf{b})^{T}(\tilde{\mathbf{Y}}-\widetilde{\mathbf{X}} \mathbf{b})
$$

The well-known least-squares solution, see e.g. [Draper and Smith (1981), p 87], is $\mathbf{b}=\boldsymbol{\beta}^{*}:=\left(\widetilde{\mathbf{X}}^{T} \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^{T} \widetilde{\mathbf{Y}}$, implying

$$
\mathbf{Z}:=\widetilde{\mathbf{Y}}-\widetilde{\mathbf{X}} \boldsymbol{\beta}^{*}:=\left(I-\widetilde{\mathbf{X}}\left(\widetilde{\mathbf{X}}^{T} \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^{T}\right) \widetilde{\mathbf{Y}}=: \mathbf{P} \widetilde{\mathbf{Y}}
$$

Here $\mathbf{P}$ is a projection matrix. The $\mathbf{Z}$-mean vector and covariance matrix satisfy

$$
\boldsymbol{\mu}_{Z}=\mathbf{0}, \quad \mathbf{C}_{Z}=\mathbf{P C}_{\widetilde{Y}} \mathbf{P}=\mathbf{P} \mathbf{W}^{1 / 2} \mathbf{C}_{Y} \mathbf{W}^{1 / 2} \mathbf{P}
$$

We consider two special cases for the covariance matrix $\mathbf{C}_{Y}$.

1. Weighing by length is natural if the $Y_{i}$-variances vary as $\mathbf{C}_{Y}=\sigma^{2} \cdot \mathbf{W}^{-1}$, in order to even out the variance differences. In this case we get the expected value, [ibid, p 120],

$$
E\left(Q_{R E S}\right)=E\left(Z^{T} Z\right)=\operatorname{trace}\left(\mathbf{C}_{Z}\right)=\sigma^{2} \cdot \operatorname{trace}\left(\mathbf{P}^{2}\right)=\sigma^{2} \cdot(n-k)
$$

whence $\hat{\sigma}^{2}=q_{\text {RES }} /(n-k)$ is an unbiased estimate - a well-known formula, [ibid, p 121].
However, the regression data in Sec 3.5 give too high $\sigma^{2}$-estimates for such a model to be realistic.
2. The common linear regression assumption is $\mathbf{C}_{Y}=\sigma^{2} \cdot \mathbf{I}$, i.e. the outcome is equally uncertain for all maintenance. This would lead to

$$
E\left(Q_{R E S}\right)=\sigma^{2} \cdot \operatorname{trace}(\mathbf{P W P})
$$

Since $\mathbf{P}, \mathbf{W}$ are unknown to us this is no useful basis for estimation, except in special cases. Therefore we also assume that all the segment lengths $\ell_{i}$ are approximately equal, $\ell_{i} \approx L / n \quad \forall i$. Then we get

$$
E\left(Q_{R E S}\right) \approx \sigma^{2} \cdot \frac{L}{n} \cdot \operatorname{trace}\left(\mathbf{P}^{2}\right)=\sigma^{2} \cdot \frac{L}{n} \cdot(n-k) ;
$$

whence

$$
\begin{equation*}
\hat{\sigma}^{2}=\frac{q_{R E S}}{L \cdot(1-k / n)} \tag{3.2}
\end{equation*}
$$

can be used.
Estimated standard deviations $\hat{\sigma}$, based on formula (3.2), are found in Tab 3.9. The values should be regarded with reservation, since the underlying regression results are somewhat incomplete (as to the used weighing).

| ModellTraffic class | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| IRI after | 0.840 | 0.723 | 0.600 | 0.555 | 0.408 | 0.353 | 0.277 | 0.229 |
| $\Delta I R I$ | 0.1569 | 0.1111 | 0.0953 | 0.0990 | 0.0929 | 0.0832 | 0.0639 | 0.0353 |
| $R D_{\text {after }}$ | 2.757 | 2.445 | 2.381 | 2.313 | 2.392 | 2.921 | 2.734 | 2.534 |
| $\Delta R D$ | 0.416 | 0.604 | 1.228 | 0.647 | 0.513 | 0.637 | 0.607 | 0.315 |

Table 3.9 Calculated standard deviations in the regression models for state transitions.
Contrary to the initial study and to NOS in its standard presentation, see [Alviti et al (1994), (1996)], where a huge amount of transition probabilities between each pair of state classes for each works type is needed in each road class, this stochastic model is far more transparent, relying on a few parameters only: the regression coefficients in Sec 3.5 and the standard deviations in Tab 3.9. This number might be further reduced - see Sec 3.9.3 below.

Observe that this model is a mixture between a deterministic model, as to routine maintenance (which dominates the calculations), and a probabilistic model, as to major maintenance. The standard deviations are not part of our basic model, but will be considered in Ch 7.

### 3.9 Discussion

### 3.9.1 Limitation

Our aim is not to build a new model system for VV but to specify reasonable demands for input models and parameters. By structuring the problem we also focus on the weaknesses, which might be reduced later on by an identification of existing useful data or by a collection of new pieces of information. Our ambition is to get meaningful results on the segment level, although we reckon that much of the normal project information about the pavement condition, e.g. microtexture, bearing capacity and drainage, is presently missing. However, we have a more detailed description than at the traditional planning and programming management levels - cf. [HDM-4 (2000), vol. 1, p 5] - where aggregations to road classes or works sections are performed and discrete sets of maintenance options are studied. And possible extensions, e.g. cracking initiation and propagation models have been developed, see [Wågberg (2001)]. We believe that the layer thickness is a useful concept for a basic description of essentially continuous maintenance costs, in view of the many asphalt-paving methods used in practice.

We emphasize that the quality of the output cannot be better than that of the input. Here several cost triggers are missing, e.g. certain functional performance (safety), all environmental performance (noise, etc.) and structural performance (bearing capacity, only caught by Age). Safety models might be based on, e.g., texture (monitored by Laser-RST) and transversal unevenness (monitored - at present Rut depth is included only as acceptance limits). Noise models might be based on, e.g., texture and bearing capacity on, e.g., transversal unevenness and cracking, etc. Structural deficiencies are important cost triggers that should be modelled, at least as acceptance limits.

### 3.9.2 Homogeneous segments

The length distribution, see Tab 3.1, is skewed; which means that much computer time is devoted to the many short segments - a non-optimal situation. By definition, each road segment is homogeneous as to its pavement history, not necessarily as to its state. In our model the pavement decisions are based on comparisons between IRI-averages for the whole segments, e.g. two segments of length 20 m and 400 m , respectively. Such comparisons are misleading, since the consecutive 20 m -averages for the long segment may fluctuate; which would lead to major maintenance for part of the long segment if it was subdivided into 20 m subsegments. This speaks for segments distinguished by a homogeneous state, apart from the pavement history. [Thomas (2001)] has devised a statistical method for the identification of such homogeneous road sections.

### 3.9.3 Robustness

Modelling is a trade-off between transparency and realism. From an analyst's point of view the preference is for robust, transparent models, controlled by a few parameters that are easy to vary and evaluate in a sensitivity analysis, i.e. contrary to the initial study (Ch 2). This is also for pedagogic reasons, in order to get a good understanding of the optimisation mechanisms.

In this respect some self-criticism is needed. The functional form of, e.g., the maintenance costs does not reflect any degree of uncertainty - a misleading precision. On the other hand, the regression models in Sec 3.5 can be scrutinised. This approach deserves praise in its direct relying on real historical data - contrary to most of the other models, which are expert opinions - but it also becomes open for criticism. It must be pointed out that these models are not final, since the generation of regression models for state transitions is an ongoing work at VV , see [Lang (2007)].

Whereas the regression model for the IRI-value immediately after a major maintenance operation is quite successful, showing high coefficients of multiple determination ( $R^{2}$-values) despite thousands of observations, the other models are weaker. Consider, e.g., the $\Delta R D_{\text {after }}$ model for the annual degradation rate. The number of observations is small in several traffic classes. Confidence intervals for the coefficients are not presented. The number of model parameters is large, 24 in all ( 3 per class), in view of the modest $R^{2}$-values. The $c_{1}$-coefficient is negative for 4 traffic classes. There is no logical reason why some coefficients should have deviating signs (here minus), for $c_{1}$ meaning that a segment of steep degradation slope is considered better off as to the future cost of the optimal choice than a segment of low degradation rate, all other factors equal. But this is what really happens in the optimisation runs - essentially in these 4 classes only. As was mentioned in Sec 3.5 we have handled this anomaly crudely by putting the 4 values to zero, without re-estimation.

Another objection is that there is no logical reason why the coefficients, e.g. $c_{0}$ and $c_{2}$, should jump up and down between the traffic classes, instead of showing a steady increase or decrease by increasing traffic $A A D T$. To cope with this anomaly a general recommendation addressed to

### 3.9 Discussion

VV is to deal with the interaction between traffic and, e.g., $R D_{\text {after }}$ in a controlled way, by assuming some specific functional form $f_{i}(A A D T)$ and turn to one general model (cf. Sec 3.5.4)

$$
\Delta R D_{\text {after }}=f_{0}(A A D T)+f_{1}(A A D T) \cdot \Delta R D_{\text {before }}+f_{2}(A A D T) \cdot R D_{\text {after }} .
$$

If, e.g., linear functions are tried,

$$
f_{i}(A A D T)=a_{i}+b_{i} \cdot A A D T \quad i=1,2,3
$$

we get a linear regression model in $\left(a_{i}, b_{i}\right)_{i=1}^{3}$, i.e. 6 parameters to estimate (instead of 24) - at most, on behalf of statistical significance.

The model structure suggested here applies to all the regression models in Sec 3.5. However, for a description of the deterioration rates $\Delta I R I, \Delta R D$ maybe all possible regression models are questionable, from a statistical point of view. Maybe (state dependent) road constants $\Delta I R I, \Delta R D$ could fully replace these two dimensions of the state space - and reduce the need for computer resources drastically.

### 3.9.4 Data corrections

As mentioned in Sec 3.1.3 data have to be supplied for 677 segments ( $7.7 \%$ ), primarily due to a missing IRI-average. The IRI-average data addition is made by an extrapolation from the year of the latest major maintenance operation, according to the registered regression line describing the IRI evolution. However, the two groups of corrected and uncorrected segments differ as to the IRI- and rut depth-parameters, but in no other respect. For the current IRI-averages the deviation is possibly due to a lower regression line slope for the corrected segments, in combination with a longer extrapolation time, on average 6.25 years longer, but this explanation does not apply to the current rut depth averages. The differences lead to deviations in start conditions: the current IRI-values are $0.52(17.2 \%)$ lower for the corrected segments, the $\Delta$ IRI-values are $0.026(21.9 \%)$ lower, and the pavement age is 0.51 years ( $8.1 \%$ ) higher. Since the corrections may have a certain impact on the optimisation results, (the cause of) the missing values should be further investigated.

## 4 Main study: Basic model and subnet results

The main study is based on the input data and functions described in Ch 3 . Compared to the initial study in Ch 2 this means a modification of our optimisation model and method, but the basic structure is kept. Here the focus is on model results, whereas major components of our optimisation method are examined in Ch 9 . A validation of our model, on part of the road network, road 63, is documented here, whereas the whole Värmland network is the database in Ch 9. Studies of subnets, e.g. one road at a time, mean a shortcut: At first the road class oriented start and residual-value routines are run on the whole network, for the determination of close-to optimal dual prices and residual values, respectively. By the residual values an infinite time horizon is considered. Contrary to the network use, where the dual prices are mere initial values - for improvement - in the subsequent main routine, the subnet application means that the determined prices are frozen, taken as a true optimum - without any tedious iterating. Instead the dual prices and residual values are applied to the road in question, for a direct generation of segment specific maintenance plans and road based statistics for various characteristics of the future pavement conditions. Or even better: If a full optimisation with the segment oriented main routine has been performed on the network level, these results are the obvious input to a special statistics run for the road. The start routine on its own can provide general information for the network, e.g. about the effects of budget changes.

The setup costs in Sec 3.6.1 for coordinated maintenance are fixed - apportioned by segment length - in the basic model, whereas these fixed costs will be correctly treated in Ch 6, devoted to coordinated maintenance. A motive for partly disregarding the common-cost reductions for simultaneous (coordinated) maintenance of a sequence of neighbouring segments is that the network (and total cost) is dominated by fairly long segments, at least (several times) 100 m . In case study Värmland 3975 segments ( $45.4 \%$ ), representing the total length $3.8 \%$, are shorter than 100 m (see Tab 3.1).

In $\operatorname{Sec} 4.1$ we present the segment and road class oriented models. For a major maintenance project to be implemented, it must obey given return rate (BCR, cf. Sec 1.1) restrictions. This type of dual constraints corresponds to extra terms in the primal objective. Sec 4.2 is devoted to the method. Because of the return rate restrictions, a Lagrangean relaxation methodology (cf. Sec 1.2.2) is of direct use. The dual subproblem is separable and the time structure makes DynP ( cf. Sec 1.2.5) the natural choice of solution method. Due to the partly discrete nature of the pavement works, we expect that approximation by smooth functions is less appropriate for describing the cost-by-state surfaces; instead a facet-like cost structure is expected. However, our approach - grid based DynP and multilinear interpolation - costs CPU time. We suggest a pair of alternatives to the subgradient algorithm for the dual price updating. One is based on the primal heuristics we use, solving a sequence of knapsack problems approximately. In Sec 4.3 the implementation is in focus. For our continuous state space the nodal grid approach introduces discretization errors, which we try to handle in two ways, in order to reduce the effects of the nonlinear cost functions. Secs $4.4-4.5$ contain results from our case study Värmland, on the network level and for road 63.

### 4.1 Model

In case the main optimisation routine is used on a subnet, i.e. run with fixed dual prices, computational aspects of the start routine are also important. Here both the road class oriented and the segment oriented models are presented. In order to distinguish these two models from other variants (in the following chapters) we will refer to them collectively as the basic model.

### 4.1.1 Return rate restrictions

The budget restriction on the total maintenance costs $G_{t}(\mathbf{z})$ in year $t$ can be written $G_{t}(\mathbf{z}) \leq b_{t}$, where $b_{t}$ as in Ch 2 denotes the maintenance budget, routine maintenance included, and $\mathbf{z}$ is the variable vector, unspecified for now. If we focus on these constraints, for a moment, and let $F(\mathbf{z})$ denote the total discounted traffic cost, the optimisation problem can be stated as to

$$
\underset{\mathbf{z}}{\operatorname{minimise}} F(\mathbf{z}) \text { s.t. }\left\{\begin{array}{c}
\mathbf{G}(\mathbf{z}) \leq \mathbf{b} \\
\mathbf{z} \in \mathcal{Z}
\end{array}\right.
$$

For our intended Lagrangean relaxation we introduce a non-negative multiplier (dual price) $v_{t}$ for every budget year $t$. The dual - the dual subproblem (cf. Sec 1.2.2) become $\underset{\mathbf{v} \geq \mathbf{0}}{\operatorname{maximise}} \Phi(\mathbf{v})$,

$$
\begin{equation*}
\Phi(\mathbf{v}):=\min _{\mathbf{z} \in \mathcal{Z}}\left(F(\mathbf{z})+\mathbf{v}^{T}[\mathbf{G}(\mathbf{z})-\mathbf{b}]\right) \tag{4.1}
\end{equation*}
$$

Our early runs confirmed that all the budget constraints are active at the optimum. However, the VV-restrictions on the capital scarcity factor $v_{\mathrm{BCR}}(>0$; see Sec 3.7.2), here written

$$
\begin{equation*}
v_{t} \geq v_{\mathrm{BCR}} \quad \forall t \tag{4.2}
\end{equation*}
$$

were not always fulfilled. Since this is a dual constraint it does not fit naturally in a primal model. What primal has (4.1) as dual subproblem and $\underset{\boldsymbol{v} \geq v_{B C R} \mathbf{1}}{\operatorname{maximise}} \Phi(\mathbf{v})$ as dual?
If too few candidate maintenance objects show acceptable return rates, we have to distinguish an utilizable (net) budget level $y_{t}$ below the nominal budget level $b_{t}$. This is the key to an improved primal formulation:

$$
\begin{align*}
& \underset{\mathbf{y}, \mathbf{z}}{\operatorname{minimise}} F(\mathbf{z})+v_{\mathrm{BCR}} \cdot \mathbf{1}^{T} \mathbf{y}  \tag{4.3a}\\
& \text { s.t. }\left\{\begin{array}{l}
\mathbf{G}(\mathbf{z}) \leq \mathbf{y} \\
\mathbf{y} \leq \mathbf{b} \\
\mathbf{z} \in \mathcal{Z}
\end{array}\right. \tag{4.3b}
\end{align*}
$$

In order to show that (4.3) has the wanted properties we formulate:
Lemma 4.1: For a Lagrangean relaxation of (4.3b) with dual prices $v_{t} \geq 0 \quad \forall t$, any dual optimum satisfies (4.2).

Proof: By the introduction of non-negative dual prices $\mathbf{v}$ the dual becomes $\underset{\mathbf{v} \geq \mathbf{0}}{\operatorname{maximise}} \widetilde{\Phi}(\mathbf{v})$, with dual subproblem

$$
\begin{equation*}
\widetilde{\Phi}(\mathbf{v}):=\min _{\mathbf{y} \leq \mathbf{b}, \mathbf{z} \in \mathcal{Z}}\left(F(\mathbf{z})+\mathbf{v}^{T} \mathbf{G}(\mathbf{z})+\left(v_{\mathrm{BCR}} \cdot \mathbf{1}-\mathbf{v}\right)^{T} \mathbf{y}\right) \tag{4.4}
\end{equation*}
$$

Here the $\mathbf{z}, \mathbf{y}$ separability means that we, while focusing on $\mathbf{y}$, may consider the subproblem solution $\mathbf{z}$ for a given $\mathbf{v}$ as fixed. As for the subproblem solution $\mathbf{y}$ one of the following cases applies to each component (time) $t$ :

$$
\left\{\begin{array}{ccc}
y_{t} \downarrow-\infty & \text { if } v_{\mathrm{BCR}}-v_{t}>0  \tag{4.5a}\\
\left.\left.y_{t} \in\right]-\infty, b_{t}\right] & \text { if } v_{\mathrm{BCR}}-v_{t}=0 \\
y_{t}=b_{t} & \text { if } v_{\mathrm{BCR}}-v_{t}<0
\end{array}\right.
$$

For $\mathbf{v}$ to be a dual $\widetilde{\Phi}$-optimum the first case (4.5a) can be ruled out.

## Theorem 4.1

(a): For any fixed solution $\mathbf{z}$ in the primal (4.3), such that $\mathbf{G}(\mathbf{z}) \leq \mathbf{b}, \mathbf{y}=\mathbf{G}(\mathbf{z})$ is optimal.
(b): For $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$ and any fixed solution $\mathbf{z}$ in the dual subproblem (4.4), $\mathbf{y}=\mathbf{b}$ is optimal.

Proof (a): Since $v_{\mathrm{BCR}}>0$ in (4.3a), each $y_{t}$ should be at its lower bound in (4.3b).
(b): According to (4.5c) and the proof of Lemma 4.1, $y_{t}=b_{t}$ is optimal if $v_{t}>v_{\mathrm{BCR}}$.

For $v_{t}=v_{\mathrm{BCR}}$ the objective in (4.4) is independent of $y_{t}$, e.g. $y_{t}=b_{t}$ is optimal.
Theorem 4.2: For a Lagrangean relaxation of (4.3b) with dual prices $\mathbf{v} \geq \mathbf{0}$, the dual subproblem is equivalent to (4.1) and the dual to $\underset{\mathbf{v} \geq v_{B C R} \cdot \mathbf{1}}{\operatorname{maximise}} \Phi(\mathbf{v})$.

Proof: According to Lemma 4.1, (4.2) is satisfied by any optimum of the Lagrangean dual of (4.3), i.e. can be added as constraints without changing the dual, to have $\underset{\mathbf{v} \geq v_{B C R} \cdot \mathbf{1}}{\operatorname{maximise}} \widetilde{\Phi}(\mathbf{v})$.

According to Thm 4.1(b), $\mathbf{y}=\mathbf{b}$ is optimum for any $\mathbf{z}$ in (4.4), i.e. can be inserted into the dual subproblem, to have

$$
\begin{aligned}
& \widetilde{\Phi}(\mathbf{v})=\min _{\mathbf{z} \in \mathcal{Z}, \mathbf{y}=\mathbf{b}}\left(F(\mathbf{z})+\mathbf{v}^{T} \mathbf{G}(\mathbf{z})+\left(v_{\mathrm{BCR}} \cdot \mathbf{1}-\mathbf{v}\right)^{T} \mathbf{b}\right)= \\
& =\min _{\mathbf{z} \in \mathcal{Z}, \mathbf{y}=\mathbf{b}}\left(F(\mathbf{z})+\mathbf{v}^{T}(\mathbf{G}(\mathbf{z})-\mathbf{b})\right)+v_{\mathrm{BCR}} \cdot \mathbf{1}^{T} \mathbf{b}=\Phi(\mathbf{v})+\mathrm{constant}
\end{aligned}
$$

This shows the equivalence with subproblem (4.1).
Since the consideration of $v_{t} \geq v_{\mathrm{BCR}}$ is part of the problem, present as an implicit constraint in (4.3), our Lagrangean relaxation approach is of direct use. In formulation (4.4) the nonutilizable part $b_{t}-y_{t}$ of $b_{t}$ (for $v_{t}=v_{\mathrm{BCR}}$ ) is lost for maintenance. In Ch 8 we will consider another model, where the non-used budget means of a year may be redistributed within the planning period, according to given discount and interest rates.

### 4.1.2 Segment oriented problem

The input model in Ch 3 characterises the maintenance works both as to type and extent. We let the $1^{\text {st }}$ component of the control variable vector $\mathbf{u}_{s t}$ catch the type, by letting $u_{s t 1}=1$ indicate that major maintenance is performed on segment $s$ in year $t$, otherwise (routine maintenance) $u_{s t 1}=0$. The $2^{\text {nd }}$ continuous component $u_{s t 2}$ describes the works extent, here the layer thickness.

### 4.1 Model

Fixed setup maintenance costs are admitted. On the road class level, e.g. in our start routine and in the NOS package [Alviti et al (1994), (1996)], such costs are necessarily standard (average) costs for all segments involved. Our cost subdivision means that the average setup cost per implemented maintenance project is estimated per road class, based on coordination statistics, and translated into a cost per metre by division with the average segment length. However, on the segment level these costs may be individually chosen, leading to a high standard cost for a short segment. Consider the maintenance cost contributions to the LHS of the budget constraints (4.6b) below. For each segment $s$, in state $\mathbf{x}_{s t}$ at the start of year $t$, we separate the setup cost $K_{s}\left(u_{s t 1}\right)=K_{s} \cdot u_{s t 1}$ and the variable cost $c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)$. (The indicator way of formulation, $K_{s} \cdot u_{s t 1}$, works if just two works types $u_{s t 1}=0,1$ are modelled.)

In the objective (4.6a) we recognise (cf. (2.1)) the traffic costs $f_{s}$ and the residual value $\varphi_{s T}$. at the horizon $t=T$, with $\varphi_{s T \bullet}\left(\mathbf{x}_{s T}\right)$ measuring the total discounted cost from state $\mathbf{x}_{s T}$ and onwards (the dot index means: irrespective of future works). All the cost contributions are expressed as net present values by the discount factor $d$. We assume (cf. Ch 1) that the $1-\mathrm{yr}$ traffic costs cover one year commencing at the preceding mid-year, whereas the maintenance cost is realised at the very mid-year - leading to the half-a-year discounting differences in (4.6a). We also reckon the return rate ( BCR ) term from (4.3a).

The state transitions (4.6d) are multidimensional and deterministic, characteristic for the segment $s$ road class $r=r(s)$. Finally, (4.6e) prescribes the initial state, and (4.6f) defines any possible, road class specific, maintenance restrictions due to, e.g., violated state acceptance limits. Full model:

$$
\begin{align*}
& \underset{\left(\mathbf{x}_{s t}\right)_{s, t},\left(\mathbf{u}_{s t} t_{s, t}\right.}{\operatorname{minimise}} \sum_{s}\left(\sum_{t=0}^{T-1} d^{t} \cdot f_{s}\left(\mathbf{x}_{s t}\right)+d^{T} \cdot \varphi_{s T \bullet}\left(\mathbf{x}_{s T}\right)\right)+v_{\mathrm{BCR}} \cdot \sum_{t=0}^{T-1} d^{t+\frac{1}{2}} y_{t}  \tag{4.6a}\\
& \text { subject to }\left\{\begin{array}{cc}
\sum_{s}\left[K_{s} \cdot u_{s t 1}+c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)\right] \leq y_{t} & \forall t \\
y_{t} \leq b_{t} & \forall t \\
\mathbf{x}_{s, t+1}=\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right) & \forall s, t \\
\mathbf{x}_{s 0}=\mathbf{a}_{s} & \forall s \\
\mathbf{u}_{s t} \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s t}\right) & \forall s, t .
\end{array}\right. \tag{4.6b}
\end{align*}
$$

By introducing set notations $\mathcal{S}, \mathcal{D}$ for the segments and the state dimensions, with cardinalities $S=|\mathcal{S}|$ and $D=|\mathcal{D}|$, respectively, and for 2D works variables (for type and extent), the number of variables can be written $(D+2) S T$. In our case study Värmland (see Ch 9 and Sec 4.4 below) we have $S=8749, D=5$ and $T=40$, i.e. 2.45 million variables.

### 4.1.3 Road class oriented problem

In our road class oriented formulation the segments are represented merely by their lengths as continuous variables, as described in Ch 2 . This segment anonymity and aggregation makes a node based formulation natural (except for the given initial segment states), using tildenotation $\widetilde{\lambda}_{r n t}$ for the total (segment) length in a road class $r$ specific nodal state $\widetilde{\mathbf{x}}_{r n}$ at time $t$.

The positions $\widetilde{\mathbf{x}}_{r n}$ of the nodes (grid points) $n \in \mathcal{N}$ in the $D$-dimensional state space are determined by the used discretization. This means that we have to reinterpret the models in Ch 3. As for the state transitions from $\widetilde{\mathbf{x}}_{r n}$, maintenance $\mathbf{u}=\widetilde{\mathbf{u}}_{r n t}=\left(\widetilde{u}_{r n t}, \widetilde{u}_{r n t 2}\right)$ will mostly lead to a non-nodal state $\overline{\mathbf{x}}$, according to the (deterministic) regression formulas in Sec 3.5. However, the corresponding length may be partitioned among the neighbouring nodal states $\widetilde{\mathbf{x}}_{r n^{\prime}}$ (transitions $n \rightarrow n^{\prime}$ ) as proportions $h_{r, n n^{\prime}}(\mathbf{u})$. The model describing such a split becomes analogous to a stochastic transition model, with the set of neighbouring states characterised by probabilities (summing up to 1, i.e. $\sum_{n^{\prime}} h_{r, n n^{\prime}}(\mathbf{u})=1$ ). If the works extent $u_{2}$ is assumed constant, e.g. for routine maintenance (type $u_{1}=0$ ), the probability distribution resulting from a given nodal state can be a priori computed, defining a constant transition matrix $\left(h_{r, n n^{\prime}}(\mathbf{u})\right)_{n, n^{\prime}}$.
In general, we apply the same weighing principle as in Ch 2 , in each state dimension $d$ guaranteeing that the weighted sum of the state variable values for the neighbouring nodes coincides with the dimension $d$ value $\bar{x}_{d}$ of the reached non-nodal state $\overline{\mathbf{x}}$. In essence this means a process oriented formulation, where the lengths of all the segments in a road class are distributed differently on the nodal states for different times $t$. At the end, at time horizon $2 T$, we get a mixture of most of the segment lengths in most of the nodal states.

The constraints in (4.6) will be doubled in (4.7) below, since we keep to the nodal states, except at $t=0$ where true segment data are used. In the first year we want to use the segment length Len $_{s}$ concentrated to the given (non-nodal) initial state $\mathbf{x}_{s 0}=\mathbf{a}_{s}$. For road class $r=r(s)$ the elements of the corresponding transition matrix $\left(h_{r n^{\prime}}\left(\mathbf{a}_{s}, \mathbf{u}_{s 0}\right)\right)_{s, n^{\prime}}$ define the proportions transferred from $\mathbf{a}_{s}$ to $\widetilde{\mathbf{x}}_{r n^{\prime}}$ during the start period, for $n^{\prime} \in \mathcal{N}, \mathrm{s} \in \mathcal{S}: r(s)=r$, a function of the initial maintenance - written $\mathbf{u}_{s 0}$ in analogy with (4.6).

The costs in road class $r$ are computed for an "average"-segment (bar notation), i.e. one of average width and traffic. Letting $\widetilde{f}_{r n}:=\bar{f}_{r}\left(\widetilde{\mathbf{x}}_{r n}\right), \widetilde{c}_{r n}(\mathbf{u}):=\bar{c}_{r}\left(\widetilde{\mathbf{x}}_{r n}, \mathbf{u}\right), \varphi_{r n, 2 T}:=\bar{\varphi}_{r, 2 T}\left(\widetilde{\mathbf{x}}_{r n}\right)$ we get the model

$$
\begin{align*}
& \left(\tilde{\lambda}_{r n t}\right)_{r, n, t} \underset{\left(, \mathbf{u}_{s 0}\right)_{s},\left(\tilde{\mathbf{u}}_{r n t}\right)_{r, n, t}}{\operatorname{minimise}} \\
& \sum_{r}\left(\sum_{s: r(s)=r} f_{s}\left(\mathbf{a}_{s}\right) \cdot \operatorname{Len}_{s}+\sum_{n}\left[\sum_{t=1}^{2 T-1} d^{t} \tilde{f}_{r n} \tilde{\lambda}_{r n t}+d^{2 T} \varphi_{r n, 2 T} \tilde{\lambda}_{r n, 2 T}\right]\right)+v_{\mathrm{BCR}} \cdot \sum_{t=0}^{2 T-1} d^{t+\frac{1}{2}} y_{t}  \tag{4.7a}\\
& \left(\begin{array}{lc}
\sum_{r} \sum_{s: r(s)=r}\left[K_{s} u_{s 01}+c_{s}\left(\mathbf{a}_{s}, \mathbf{u}_{s 0}\right)\right] \cdot \text { Len }_{s} \leq y_{0} & t=0 \\
\sum_{r} \sum_{n}\left[\bar{K}_{r} \widetilde{u}_{r n t 1}+\widetilde{c}_{r n}\left(\widetilde{\mathbf{u}}_{r n t}\right)\right] \cdot \widetilde{\lambda}_{r n t} \leq y_{t} & \forall t>0
\end{array}\right.  \tag{4.7b}\\
& \begin{array}{cc}
\sum_{r} \sum_{n}\left[\bar{K}_{r} \tilde{u}_{r n t 1}+\widetilde{c}_{r n}\left(\widetilde{\mathbf{u}}_{r n t}\right)\right] \cdot \tilde{\lambda}_{r n t} \leq y_{t} & \forall t>0 \\
y_{t} \leq b_{t} & \forall t
\end{array}  \tag{4.7c}\\
& \text { subject to }\left\{\begin{array}{l}
\tilde{\lambda}_{r n^{\prime} 1}=\sum_{s: r(s)=r} h_{t} \leq n^{\prime}\left(\mathbf{a}_{s}, \mathbf{u}_{s 0}\right) \cdot \text { Len }_{s} \quad t=0, \forall r, n^{\prime}, ~
\end{array}\right.  \tag{4.7~d}\\
& \tilde{\lambda}_{r n^{\prime}, t+1}=\sum_{n} \widetilde{h}_{r, n n^{\prime}}\left(\widetilde{\mathbf{u}}_{r n t}\right) \cdot \widetilde{\lambda}_{r n t} \quad \forall t>0, \forall r, n^{\prime}  \tag{4.7f}\\
& \begin{array}{lc}
\mathbf{u}_{s 0} \in \mathcal{U}_{r(s)}\left(\mathbf{a}_{s}\right) & \forall s \\
\widetilde{\mathbf{u}}_{r n t} \in \mathcal{U}_{r}\left(\widetilde{\mathbf{x}}_{r n}\right) & \forall t>0, \forall r, n .
\end{array} \tag{4.7~g}
\end{align*}
$$

By introducing set notations $\mathcal{R}, \mathcal{L}$ for the road classes and the node levels per state dimension, with cardinalities $R=|\mathcal{R}|$ and $L=|\mathcal{L}|$, respectively, and using the grid size per year $|\mathcal{N}|=L^{D}$ and two works variables plus one length variable per node and time, the total number of variables can be written $2 S+3(T-1) R \cdot L^{D}$. In our case study Värmland we have $R=29$, i.e. from 1.69 million variables for $L=3$ to 53.7 million variables for $L=6$.

### 4.2 Method

The methods we apply to (4.6) and (4.7) are almost identical. Below we describe the method for the segment oriented problem (4.6) and comment on the few differences. By a Lagrangean relaxation of the budget constraints (4.6b) (respectively (4.7b-c)) we receive a separable dual subproblem consisting of segment specific (road class specific) subproblems with time structure, which are solved by DynP. The optimal future costs, determined per nodal point in the dual DynP backwards routine, are used also for the generation of primally feasible solutions.

### 4.2.1 Dual optimisation

The budgetary constraints (4.6b) are relaxed by the introduction of (non-negative) Lagrangean multipliers $\left(v_{t}\right)_{t=0}^{T-1}$, implicitly restricted to $v_{t} \geq v_{\mathrm{BCR}}$ according to Thm 4.2. For their updating we use the same subgradient technique as in Ch 2 as standard method. By Thm 4.1(b) the residuals in (4.6b) for the subproblem optimum of the current dual iteration can be written $\mathbf{g}=\left(g_{t}\right)_{t}, g_{t}=G_{t}-b_{t}, G_{t}:=\sum_{s}\left[K_{s} \cdot u_{s t 1}+c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)\right]$. In terms of $\mathbf{g}$, the best found (highest) dual value $\underline{\Phi}$, current dual value $\hat{\Phi}$ and the best found (lowest) primal objective value $\bar{\Phi}$, the unconstrained multiplier subgradient step (cf. Sec 1.2.3) is taken as

$$
\Delta \mathbf{v}=\sigma \cdot \mathbf{g}, \quad \sigma:=\min \left[\rho \cdot(\underline{\Phi}+q \cdot(\bar{\Phi}-\underline{\Phi})-\hat{\Phi}) / \mathbf{g}^{T} \mathbf{g}, \sigma_{\max }\right]
$$

where $\sigma_{\max }$ is a given acceptance limit, $\rho$ is reduced geometrically from 2.0 and $q$ is increased from 0.5 towards 1.0 by the iteration number, such that $1-q$ decreases geometrically. In case of stagnation, i.e. no improvement of $\underline{\Phi}$ for a given number of dual iterations, $\rho$ is increased by a given factor. The updating involves only the time periods $t$ that have a non-optimal status, i.e. $g_{t}<0, v_{t}>v_{\mathrm{BCR}}$ or $g_{t}>0$. In case a bound $v_{t} \geq v_{\mathrm{BCR}}$ is violated, orthogonal projection onto $v_{t}=v_{\mathrm{BCR}}$ is applied, see [Larsson et al (1996)].

Some alternatives to the subgradient updating have been tried. The major test runs are documented in Ch 9 . Here we will comment upon two alternatives or supplements to the subgradient method and also show some test run results. One method approach emanates from our observations of the strong dependence between neighbour years, in so far as a change of the dual price $v_{t}$ in one year $t$ will have an effect not only on the maintenance cost $G_{t}$ in the very same year, but will often also lead to compensatory, opposite cost changes in the adjacent years, due to slightly postponed or pre-activated segment maintenance. A characteristic example is an initial huge budget violation moving one year ahead per dual iteration, if the
subgradient method is applied. Hence it may be advantageous, especially in the early iteration process, to damp the requests for changes in individual years, by applying a smooth price evolution curve approximating the preliminary points $\left(t, v_{t}\right)$ generated by the subgradient method. Moreover, we expect such a price curve to be transient cyclic, levelling out at the end of the time horizon, as the influence from the steady-state (residual-value) model becomes dominant. For the purpose we have applied trigonometric approximation in the following sense: First we generate new prices $\mathbf{v}$ according to the ordinary subgradient method but use these prices merely as trial values. For the final choice of prices in the next dual iteration we least squares approximate the points $\left(t, v_{t}\right)$ by the first terms in a Fourier series, i.e. by fitting a low-order trigonometric function. In Table 4.1 we present some resulting gaps, for two different orders of the approximating sin-cos-trigonometric function; letting Trig4 (Trig5) denote order $=4$ (5), corresponding to 9 (11) parameters. The results in Tab 4.1 refer to a level $L=5$ discretization in the start routine with a constant annual budget; Tab 4.1a was run with absolute acceptance limits for the worst acceptable state and Tab 4.1b with penalised limit violations. Tab 4.1 b shows an example of (theoretically impossible) negative gaps between the primal and dual objective values. We interpret this as a manifestation of discretization errors, more thoroughly discussed below in Sec 4.3.4. In the absolute limits run the ordinary subgradient method has difficulties in finding a feasible solution before iteration 100, whereas the Trig-method finds such ones in the early iterations. In the relative limits run (Tab 4.1b) the subgradient method outperforms the Trig-options from the start. We also notice that the Trigmethod early reaches a gap which is never improved thereafter. This is an inherent disadvantage of such approximation methods: Every least squares approximation means a projection onto a subspace of (or manifold in) the $\boldsymbol{v}$-space. As soon as a locally best point in the subspace is reached, no further improvements are possible. In order to guarantee convergence to a full optimum, by relying on corresponding results for the basic subgradient directions, the full $\boldsymbol{v}$-space must be covered - out of the question here. However, the initial power of the Trig-method in generating primally feasible solutions, i.e. suitable upper bounds for the optimal dual value, makes it conceivable as a general backup start method.

Our second approach is to extend the primal heuristics below in Sec 4.2 .3 by using the adjusted prices $v_{t}^{\prime}$, found to meet the budget demands per year $t$, also for the purpose of dual price updating. In this respect the Prim-method is a coordinate-wise search method, by adjusting one price at a time. We let the direction of net price changes $\mathbf{v}^{\prime}-\mathbf{v}$, damped by a factor $\left.p \in\right] 0,1[$, determine the next dual prices $\mathbf{v}+p \cdot\left(\mathbf{v}^{\prime}-\mathbf{v}\right)$. In Tab 4.1 the damp factor values $p=0.5$ (Prim0.5) and $p=0.25$ (Prim0.25) are used. Whereas the ordinary subgradient method shows smaller gaps in the final iterations in Tab 4.1a, the Prim0.5-method performs better (here more negative gaps) in almost all iterations in Tab 4.1b. However, since the Prim-method lacks the support from general theoretical convergence results and does not always outperform the subgradient method in practice we have not used the Prim-approach in the final runs.

| Methodllteration | 3 | 6 | 12 | 25 | 50 | 100 | 150 | 200 |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Subgradient |  |  |  |  |  | 1.446 | 0.195 | 0.162 |
| Trig4 | 9.025 | 8.984 | 8.984 | 8.984 | 8.984 | 8.984 | 8.984 | 8.984 |
| Trig5 |  | 16.845 | 16.845 | 16.845 | 16.845 | 16.845 | 16.845 | 16.845 |
| Prim0.5 |  |  |  | 0.738 | 0.686 | 0.351 | 0.294 |  |
| Prim0.25 |  |  |  |  |  | 0.703 | 0.703 | 0.703 |

Table 4.1a Gap (\%) after various numbers of iterations, for some dual price updating methods. Absolute state limits.

|  |  |  |  |  |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Methodllteration | 3 | 6 | 12 | 25 | 50 | 100 | 150 | 200 |
| Subgradient | 0.284 | 0.065 | 0.003 | $-9.4^{*} 10^{-7}$ | $-1.4^{*} 10^{-5}$ | $-1.5^{*} 10^{-5}$ | $-1.5^{*} 10^{-5}$ | $-1.5^{*} 10^{-5}$ |
| Trig4 | 0.736 | 0.736 | 0.736 | 0.736 | 0.736 | 0.736 | 0.736 | 0.736 |
| Trig5 | 0.672 | 0.601 | 0.491 | 0.382 | 0.381 | 0.381 | 0.381 | 0.381 |
| Prim0.5 | 0.208 | 0.024 | $-5.4^{*} 10^{-7}$ | $-3.4^{*} 10^{-6}$ | $-5.6^{*} 10^{-6}$ | $-1.6^{*} 10^{-5}$ | $-1.6^{*} 10^{-5}$ | $-1.6^{*} 10^{-5}$ |
| Prim0.25 | 0.424 | 0.171 | 0.023 | $-2.4^{* 1} 10^{-6}$ | $-4.2^{*} 10^{-6}$ | $-4.2^{*} 10^{-6}$ | $-1.7^{*} 10^{-5}$ | $-1.7^{*} 10^{-5}$ |

Table 4.1b Gap (\%) after various numbers of iterations, for some dual price updating methods. Relative state limits.

### 4.2.2 Dual subproblem solving

Recall the segment oriented problem (4.6). The objective of the separable dual is written

$$
\Phi(\mathbf{v})=: \sum_{\mathrm{s}} \Phi_{s}^{\mathrm{net}}(\mathbf{v})+\sum_{t=0}^{T-1} d^{t+\frac{1}{2}}\left(v_{\mathrm{BCR}}-v_{t}\right) y_{t}
$$

Here the dynamic, segment $s$ specific dual network subproblem is

$$
\left.\begin{array}{rl}
\Phi_{s}^{\mathrm{net}}(\mathbf{v}) & =\left(\min _{\left(\mathbf{x}_{s t} t_{t},\left(\mathbf{u}_{s t}\right)_{t}\right.} \sum_{t=0}^{T-1} d^{t} \cdot\left(f_{s}\left(\mathbf{x}_{s t}\right)+v_{t} \sqrt{d} \cdot\left[K_{s} u_{s t 1}+c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)\right]\right)+d^{T} \cdot \varphi_{s T} \cdot\left(\mathbf{x}_{s T}\right)\right.
\end{array}\right)
$$

In the DynP backwards routine we turn to a nodal formulation, analogous to (4.7), at the start of year $t$ considering the nodal states $\mathbf{x}_{s t}=\widetilde{\mathbf{x}}_{n}, n \in \mathcal{N}$ and maintenance $\mathbf{u}=\left(u_{1}, u_{2}\right)$. In practice we compute an optimal future cost $\varphi_{s t m}\left(\widetilde{\mathbf{x}}_{n}\right)$ for each $n$ and each works type (maintenance type) $m:=u_{1}=0,1$. The optimisation is performed with respect to the continuously varying works extent $u_{2}$ (cf. Sec 4.2 .4 below) and the works type $m^{\prime}$ in the next year $t+1$, in consideration of the optimal future cost registered for each $m^{\prime}$ at the end of year $t$. The dot index for the given residual values $\varphi_{s T}$ • in (4.8a) means that the succeeding maintenance type is irrelevant at the very horizon. The recursive formula, initiated at the horizon $t=T$, is

$$
\begin{aligned}
& \varphi_{s t m}\left(\widetilde{\mathbf{x}}_{n}\right)=\min _{u_{2}, m^{\prime}}\left(f_{s}\left(\widetilde{\mathbf{x}}_{n}\right)+v_{t} \sqrt{d} \cdot\left[K_{s} \cdot m+c_{s}\left(\widetilde{\mathbf{x}}_{n}, \mathbf{u}\right)\right]+d \cdot \bar{\varphi}_{s, t+1, m^{\prime}}\left(\mathbf{x}_{s, t+1}\right)\right) \\
& \text { subject to } \quad \mathbf{x}_{s, t+1}=\mathbf{h}_{r(s)}\left(\widetilde{\mathbf{x}}_{n}, \mathbf{u}\right), \mathbf{u}=\left(m, u_{2}\right) \in \mathcal{U}_{r(s)}\left(\widetilde{\mathbf{x}}_{n}\right), \mathbf{u}^{\prime}=\left(m^{\prime}, u_{2}^{\prime}\right) \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s, t+1}\right) .
\end{aligned}
$$

Here $\bar{\varphi}_{s, t+1, m^{\prime}}$ (bar notation) means that the value is computed by interpolation/extrapolation, based on the optimal future costs $\varphi_{s, t+1, m^{\prime}}\left(\widetilde{\mathbf{x}}_{n^{\prime}}\right)$ as determined at the end of year $t$ for maintenance type $m^{\prime}=u_{s, t+1,1}$ in the neighbouring nodal states $\widetilde{\mathbf{x}}_{n^{\prime}}-\mathrm{cf}$. Ch 2 . For $t=0$ the
node based computation is exchanged for the given (non-nodal) initial segment state $\mathbf{x}_{s 0}=\mathbf{a}_{s}$. A nodal formulation means the introduction of many help-variables, multiplying the problem complexity. The CPU-times are essentially controlled by the backwards routine and the total grid size $S T \cdot L^{D}$.

In the forwards routine, with the segment state $\mathbf{x}_{s t}$ reached at the start of year $t$ and considered as given, we perform a "two-year" optimisation for best works type and works extent. Since the traffic and maintenance costs $f_{s}\left(\mathbf{x}_{s t}\right), c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)$ are simple function evaluations, no interpolation is needed for their computation, in opposite to the initial study. Once again, the only need for interpolation/extrapolation concerns $\bar{\varphi}_{s, t+1, m^{\prime}}$ :

$$
\begin{aligned}
& \underset{u_{s t 2}, u_{s, t+1,1}}{\operatorname{minimise}}\left(f_{s}\left(\mathbf{x}_{s t}\right)+v_{t} \sqrt{d} \cdot\left[K_{s} u_{s t 1}+c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)\right]+d \cdot \bar{\varphi}_{s, t+1, u_{s, t+1,1}}\left(\mathbf{x}_{s, t+1}\right)\right) \\
& \text { subject to } \quad \mathbf{x}_{s, t+1}=\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right), \mathbf{u}_{s t} \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s t}\right), \mathbf{u}_{s, t+1} \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s, t+1}\right)
\end{aligned}
$$

The reason for our "two-year" optimisation, possible by determining and registering $\varphi_{s t m}\left(\widetilde{\mathbf{x}}_{n}\right)$ for each maintenance type $m$, is to get more accurate interpolation results in the forwards routine (cf. Ch 2, where the idea was introduced).

In the road class oriented start routine, where the segment lengths are partitioned among the nodal states, we consider the extra-work of "two-year" optimisation as less motivated (except in the initial time period, where the true segment states are to be used) and analogously to the residual values in (4.6a) we compute only the lowest future cost at each nodal state, letting both the type $m=\widetilde{u}_{r n t 1}$ and the extent $\widetilde{u}_{r n t 2}$ vary, i.e.

$$
\varphi_{r n t}:=\min _{\widetilde{\mathbf{u}}_{r n t}}\left(\bar{f}_{r}\left(\widetilde{\mathbf{x}}_{r n}\right)+v_{t} \sqrt{d} \cdot\left[\bar{K}_{r} \widetilde{u}_{r n t 1}+\bar{c}_{r}\left(\widetilde{\mathbf{x}}_{r n}, \widetilde{\mathbf{u}}_{r n t}\right)\right]+d \cdot \bar{\varphi}_{r, t+1}\left(\mathbf{h}_{r}\left(\widetilde{\mathbf{x}}_{r n}, \widetilde{\mathbf{u}}_{r n t}\right)\right)\right)
$$

Instead we distinguish the future traffic and future maintenance costs $\varphi_{r n t}^{\text {traf }}, \varphi_{r n t}^{\text {maint }}$ at each node and time, for a more accurate translation to the segment based dynamics in the first year. Since the traffic and width of a segment $s$ normally deviates from the averages $\overline{A a d t}_{r}$, $\overline{\text { AadtHeavy }}_{r}$ and $\overline{W i d}_{r}$ in its road class $r=r(s)$, the relevant future cost per metre at the end of the $1^{\text {st }}$ year is

$$
\varphi_{s 1} \bullet\left(\mathbf{x}_{s 1}\right):=\frac{\operatorname{Aadt}_{s}}{\overline{\operatorname{Aadt}}_{r(s)}} \varphi_{r(s), 1}^{\mathrm{traf}}\left(\mathbf{x}_{s 1}\right)+\frac{\text { Wid }_{s}}{\overline{W i d}_{r(s)}} \varphi_{r(s), 1}^{\operatorname{maint}}\left(\mathbf{x}_{s 1}\right)
$$

and at the start of the $1^{\text {st }}$ year - the basis for both backwards and forwards calculation -

$$
\varphi_{s 0}\left(\mathbf{a}_{s}\right)=\min _{\mathbf{u}_{s 0}}\left(f_{s}\left(\mathbf{a}_{s}\right)+v_{0} \sqrt{d} \cdot\left[K_{s} u_{s 01}+c_{s}\left(\mathbf{a}_{s}, \mathbf{u}_{s 0}\right)\right]+d \cdot \varphi_{s 1}\left(\mathbf{h}_{r(s)}\left(\mathbf{a}_{s}, \mathbf{u}_{s 0}\right)\right)\right)
$$

The road class oriented start routine covers $2 T$ years, double the horizon of the segment oriented main routine. The optimal dual prices $\left(v_{t}\right)_{t=T}^{2 T-1}$, as determined by the start routine, are used for the computation of residual values for the segments in the main routine - cf. Sec 4.2.5 below.

### 4.2 Method

### 4.2.3 Primal heuristics

In the case study described below in Sec 4.4 we generate primal solutions by a method similar to the heuristics used in the initial study (Ch 2), in its stepping forward one year $t$ at a time while adjusting the total maintenance costs $G_{t}$ to the budget $b_{t}$. Thus one nonlinear equation at a time is solved - or the lower bound $v_{t}=v_{\mathrm{BCR}}$ is established. This involves a costly sorting of all segments of interest.

We start from a given segment state distribution $\left(\mathbf{x}_{s t}\right)_{s}$ in the beginning of year $t$. For each segment $s$ the "two-year" optimisation is based on the traffic cost $f_{s}\left(\mathbf{x}_{s t}\right)$ and the optimal future costs $\varphi_{s, t+1, m^{\prime}}\left(\widetilde{\mathbf{x}}_{n}\right)$, as determined in the dual backwards routine for each maintenance type $m^{\prime}$ from each nodal state $\widetilde{\mathbf{x}}_{n}, n \in \mathcal{N}$, at the end of year $t$. The final choice is between one candidate $\mathbf{u}_{s t}^{m}=\left(m, u_{s t 2}^{m}\right)$ for each maintenance type $m$, where we determine the best works extent $u_{s t 2}^{m}$ according to Sec 4.2 .4 below. The corresponding maintenance cost is $G_{s t}^{m}:=K_{s} \cdot m+c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{m}\right)$ and the state transition satisfies $\mathbf{x}_{s, t+1}=\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{m}\right)$. Our strategy is to adjust the Lagrangean multiplier $v_{t}$ until the optimum maintenance type is changed for sufficiently many segments, minimizing the dual cost function for year $t$ and later. In a scarcity situation, by successively adjusting $v_{t}$ upwards, cheaper and cheaper maintenance types are favoured. We describe the details in the case of two maintenance types, indexed $m=1$ (major maintenance) and $m=0$ (routine maintenance), respectively. In this case the appropriate $v_{t}$ value can be found without iterating. Consider any segment $s$. Introducing abbreviations for the lowest future costs for maintenance $\mathbf{u}_{s t}^{m}$ registered at times $t+1, t$ as

$$
\begin{aligned}
& \bar{\Phi}_{s, t+1}^{m}:=\min _{m^{\prime}} \bar{\varphi}_{s, t+1, m^{\prime}}\left(\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{m}\right)\right), \\
& \Phi_{s t}^{m}:=f_{s t}\left(\mathbf{x}_{s t}\right)+v_{t} \sqrt{d} \cdot G_{s t}^{m}+d \cdot \bar{\Phi}_{s, t+1}^{m} \quad m=0,1
\end{aligned}
$$

the break-even change $\varepsilon_{s t}$ of $v_{t}$ is a change that makes the two competing future costs equal:

$$
f_{s t}\left(\mathbf{x}_{s t}\right)+\left(v_{t}+\varepsilon_{s t}\right) \sqrt{d} \cdot G_{s t}^{0}+d \cdot \bar{\Phi}_{s, t+1}^{0}=f_{s t}\left(\mathbf{x}_{s t}\right)+\left(v_{t}+\varepsilon_{s t}\right) \sqrt{d} \cdot G_{s t}^{1}+d \cdot \bar{\Phi}_{s, t+1}^{1}
$$

i.e.

$$
\begin{equation*}
\varepsilon_{s t}=\frac{\Phi_{s t}^{0}-\Phi_{s t}^{1}}{\sqrt{d}\left(G_{s t}^{1}-G_{s t}^{0}\right)} \tag{4.9}
\end{equation*}
$$

In case $\Phi_{s t}^{0} \leq \Phi_{s t}^{1}$, i.e. if routine maintenance initially rules, we expect $\varepsilon_{s t}<0$, since the maintenance costs should obey $G_{s t}^{1}>G_{s t}^{0}$, unless the segment state is extremely bad. Analogously, $\Phi_{s t}^{0}>\Phi_{s t}^{1}$ normally means $\varepsilon_{s t}>0$. Special handling is needed at every upper (acceptance) state limit, if any, and in the exceptional case $G_{s t}^{1} \leq G_{s t}^{0}$. Depending on the adjustment case, determined by the sign of the budget residual $g_{t}=G_{t}-b_{t}$, i.e. a violation or under-utilisation of the budget, the segments that are prone to adjustments are normally identified by the sign of $\varepsilon_{s t}$. By ranking all the segments (of the correct sign) according to the $\varepsilon_{s t}$-value and summing up the change of maintenance cost $G_{t}$, the necessary $v_{t}$-change $\hat{\varepsilon}_{t}$ can be calculated. The ranking procedure is described below.

Our heuristics makes it possible to explicitly consider the return rate constraint $v_{t} \geq v_{\mathrm{BCR}}$. In an under-utilisation situation, where $v_{t}\left(>v_{\mathrm{BCR}}\right)$ is to be decreased, this constraint means that all segments $s$ with $\varepsilon_{s t} \leq v_{\mathrm{BCR}}-v_{t}(<0)$ can be neglected (excluded from ranking).

Note that optimal works extents $\left(u_{s t 2}^{m}\right)_{s, m}$ are computed before the heuristics in year $t$ starts, for the states $\left(\mathbf{x}_{s t}\right)_{s}$ that are the result of the heuristics in the previous year - but no recalculation of works extents is performed during the process of finding $\hat{\varepsilon}_{t}$.

The ranking formula (4.9) can be motivated by well-known knapsack ranking: For a given discrete set of maintenance options $\left(\mathbf{u}_{s t}^{m}\right)_{s, m}$ the relevant local primal problem to solve (based on the total future costs if separate future traffic costs are non-utilizable) is

$$
\begin{array}{ll}
\operatorname{minimize} & \sum_{s, m} \Phi_{s t}^{m} \cdot z_{s m} \\
\text { s.t. } \quad \begin{cases}\sum_{s, m} G_{s t}^{m} \cdot z_{s m} \leq b_{t} \\
\sum_{m} z_{s m}=1 & \forall s \\
z_{s m} \in\{0,1\} & \forall s, m\end{cases}
\end{array}
$$

For simplicity, the return rate constraints are omitted here. In case there are two maintenance types only, we simplify the formulation by using $z_{s 0}=1-z_{s 1}$, introducing the minimumbudget $b_{t}^{0}:=b_{t}-\sum_{s} G_{s t}^{0}$ and omitting constants in the objective, to get

$$
\begin{array}{ll}
\operatorname{minimize} & \sum_{s}\left(\Phi_{s t}^{1}-\Phi_{s t}^{0}\right) \cdot z_{s 1} \\
\text { s.t. } & \left\{\begin{array}{cc}
\sum_{s}\left(G_{s t}^{1}-G_{s t}^{0}\right) \cdot z_{s 1} \leq b_{t}^{0} \\
z_{s 1} \in\{0,1\} & \forall s .
\end{array}\right.
\end{array}
$$

This is a knapsack problem, approximately solved by ranking the quotients between the objective and constraint coefficients for each variable $z_{s 1}$, i.e. (4.9).

Using $\hat{v}_{s t}:=v_{t}+\varepsilon_{s t}$ for the breakeven point, (4.9) can be written

$$
\begin{equation*}
\hat{v}_{s t}=\frac{\sqrt{d}\left(\bar{\Phi}_{s, t+1}^{0}-\bar{\Phi}_{s, t+1}^{1}\right)}{G_{s t}^{1}-G_{s t}^{0}} \tag{4.9'}
\end{equation*}
$$

This formula resembles the BCR-ratio (1.1). Remember that the denominator and numerator in (1.1) stand for the total discounted traffic and maintenance cost, respectively. For (4.9') to be useful as a general formula, the future costs $\bar{\Phi}$ and dual prices $v$ are needed. If no other information is available, we may assume $v_{t+\tau}=v_{B C R} \quad \forall \tau>0$ and calculate the total discounted future cost differences $\Delta f_{s, t+1}^{\mathrm{TOT}}, \Delta G_{s, t+1}^{\mathrm{TOT}}$ at the end of year $t$, for segment $s$, minor ( $m=0$ ) minus major ( $m=1$ ) maintenance. The breakeven point turns into

$$
\hat{v}_{s t}=\frac{\sqrt{d}\left(\Delta f_{s, t+1}^{\mathrm{TOT}}+v_{B C R} \cdot \Delta G_{s, t+1}^{\mathrm{TOT}}\right)}{-\Delta G_{s t}}
$$

whereas (1.1) corresponds to the assumption $v_{t+\tau}=\hat{v}_{s t} \quad \forall \tau>0$, providing

$$
\begin{equation*}
\hat{v}_{s t}=\frac{\sqrt{d} \cdot \Delta f_{s, t+1}^{\mathrm{TOT}}}{-\Delta G_{s t}-\sqrt{d} \cdot \Delta G_{s, t+1}^{\mathrm{TOT}}} \tag{4.10}
\end{equation*}
$$

In case full optimisation cannot be performed, we recommend (4.9') to be generally used for ranking instead of (1.1) and (4.10), since a potential project should carry the costs that reflect the future budget scarcities $\left(v_{t+\tau}\right)_{\tau>0}$ and a profitability (BCR) measure should answer the maintenance question, whether to choose major maintenance now or postpone the decision (at least) one year, cf. Fig 1.2.

The heuristics can be further improved: Since the resulting total cost $G_{t}$ normally does not exactly equal the budget $b_{t}$, and the segment $s$ maintenance cost varies continuously with the works extent $u_{s t 2}$, adjustments can be made for a group of segments with $\varepsilon_{s t} \approx \hat{\varepsilon}_{t}$. We have not implemented this option. Moreover, re-optimisation of the works extents $\left(u_{s t 2}^{m}\right)_{s, m}$ for the generated $v_{t}+\hat{\varepsilon}_{t}$ trial values might improve the solution quality (although there is no direct link between the dual and primal solution optimality in non-convex problems).

In case more than two maintenance types are present, multiple switching of maintenance types may occur for each segment, and a simple iterative procedure is needed for the determination of $\hat{\varepsilon}_{t}$.

A computational drawback to this heuristics is the need for handling data about all segments simultaneously. Implemented in this way the heuristics is dimensioning for the need of computer memory and, in case of parallelisation, it leads to intensive information transfer between the computers for the determination of $\left(\hat{\varepsilon}_{t}\right)_{t}$. An alternative is presented in Ch 9 .

We will now examine the ranking details. No complete ranking is needed - but partial ordering by repeated sorting. For each segment $s$ and the current dual price $v_{t}$ in year $t$ the optimal cost difference $\Delta G_{s t}:=G_{s t}^{1}-G_{s t}^{0}$ and the corresponding ranking quotient $\varepsilon_{s t}$ in (4.9) is computed. By summing up the costs for the optimum maintenance type for each segment we identify the case - under-utilisation or violation. In, e.g., a scarcity situation, where the budget level is exceeded by $g_{t}(>0)$ for the current dual price $v_{t}$, this Lagrangean multiplier has to be increased by an unknown amount $\hat{\varepsilon}_{t}$. Our search problem is to identify a segment $s$ of characteristics $\left(\Delta G_{s t}, \varepsilon_{s t}\right)$, with $\varepsilon_{s t}>0$, such that $\hat{\varepsilon}_{t}=\varepsilon_{s t}$ is the lowest change to accomplish $\sum_{s^{\prime}: \varepsilon_{s} t \leq \hat{\varepsilon}_{t}} \Delta G_{s^{\prime} t}+g_{t} \leq 0$. Formulated as an optimisation problem:
$\underset{\mathbf{y}, \hat{\varepsilon}_{t}}{\operatorname{minimise}} \hat{\varepsilon}_{t}$
subject to
$\left\{\begin{array}{cc}\sum_{s} \Delta G_{s t} \cdot y_{s}+g_{t} \leq 0 \\ \varepsilon_{s t} \cdot y_{s} \leq \hat{\varepsilon}_{t} \\ y_{s} \in\{0,1\} & \forall s\end{array}\right.$
This formulation focuses on the binary choice for each segment, trivial once the breakeven point $\hat{\varepsilon}_{t}$ has been determined. In the well-known 2-bin sorting procedure a sequence of trial
values $\hat{\varepsilon}_{t i}$ and remaining cost residuals $g_{t i}$ (initially $g_{t 0}=g_{t}$ ) is applied, for each trial (iteration) $i$ referring the remaining unsorted segments to either of two bins, according to $\varepsilon_{s t}$ greater or less than $\hat{\varepsilon}_{t i}$. In the latter bin, corresponding to a segment set $\mathcal{S}_{1 i}:=\left\{s: y_{s i}=1\right\}$, the total cost change $\sum_{s \in \mathcal{S}_{1 i}} \Delta G_{s t}$ is compared to $g_{t i}$. If $\sum_{s \in \mathcal{S}_{1 i}} \Delta G_{s t}+g_{t i}>0$, the trial $\hat{\varepsilon}_{t i}$-value was too small and we update $g_{t, i+1}:=g_{t i}+\sum_{s \in \mathcal{S}_{1 i}} \Delta G_{s t}$. In any case the bin containing the optimal $\hat{\varepsilon}_{t}-$ value can be identified, and the segments in that bin define the remaining segments to sort in the next iteration. In our implementation we multiply or divide the previous trial $\hat{\varepsilon}_{t i}$-value by a given factor, until the optimum $\hat{\varepsilon}_{t}$ is trapped. After that interval bisection is applied.

As an alternative to 2-bin sorting, $K$-bin sorting has been implemented, for any $K>2$. Here the two extreme bins, i.e. the ones containing the lowest and highest $\varepsilon_{s t}$-values, correspond to the two half-open value intervals of 2-bin sorting. These are supplied with $K-2$ intermediate, fixed value intervals. Instead of one dividing breakeven point value $\hat{\varepsilon}_{t i}$ per iteration $i$ we have to specify $K-1$ values $\left(\hat{\varepsilon}_{t i}^{k}\right)_{k=0}^{K-2}$. In practice the two extreme values $\hat{\varepsilon}_{t i}^{0}, \hat{\varepsilon}_{t i}^{K-2}$ are explicitly specified; the rest follows by dividing into equally large value intervals. For the corresponding scale $\kappa:=(K-2) /\left(\hat{\varepsilon}_{t i}^{K-2}-\hat{\varepsilon}_{t i}^{0}\right)$ each $\varepsilon_{s t}$ is sorted as follows:

- If $\varepsilon_{s t}<\hat{\varepsilon}_{t i}^{0}$, put $s$ in bin $k=0$.
- Otherwise, if $\varepsilon_{s t}>\hat{\varepsilon}_{t i}^{K-2}$, put $s$ in bin $k=K-1$.
- Otherwise compute $k:=\left\lceil\kappa \cdot\left(\varepsilon_{s t}-\hat{\varepsilon}_{t i}^{0}\right)\right\rceil$ and put $s$ in bin $k$.

An advantage of K-bin sorting is that once the wanted $\hat{\varepsilon}_{t}$-value becomes trapped inside an intermediate bin, a faster reduction of the number of unsorted segments is expected in the next iteration than from just two bins. A disadvantage is the additional numerical work for computing the correct intermediate value interval (bin). For a given computer system it is possible to judge the times needed for the various sorting activities and to mathematically formulate and find the optimal balance between number of bins and re-sorting iterations. However, with several computers collaborating, the information transfer time between the computers may be decisive for the wall-clock time. In our implementation each processor is responsible for a number of segments. After every performed re-sorting iteration, information about the total potential cost change per bin, $\sum_{s \in \mathcal{S}_{k i}} \Delta G_{s t}$ for $k=0,1, \ldots, K-1$, and the number of segments in each bin is transferred to the supervising root processor. At the root the decisive bin is identified and information about the bin number and a stop-indicator are transferred to all other processors. Stop is indicated when the number of segments in the decisive bin is below a given acceptance limit $u$ Hash. This triggers the transfer of $\left(\Delta G_{s t}, \varepsilon_{s t}\right)$ for all the remaining unsorted segments $s$ and a final ranking (of at most $u$ Hash elements) at the root. A high $u$ Hashvalue means few re-sorting iterations but much information to transmit.

Instead of relying on theoretical results we have tested various parameter sets of 2- and K-bin sorting. We have varied the number of bins $K$ as well as the stop-number $u H a s h$ of unsorted segments. On a 4 PC-cluster, for wide ranges of $u H a s h=5-80$ and $K=10-40$, a test run resulted in wall-clock times at or slightly above 1410 sec for the 2-bin and at or slightly below 1400 sec for the K-bin option, i.e. around $1 \%$ lower. Since our implementation is intended for several collaborating computers we stick to the K-bin option hereafter. In practice there are more parameters, e.g. the initial cost scale (initial $\hat{\varepsilon}_{t}^{0}, \hat{\varepsilon}_{t}^{K-2}$-values) and its change rate per resorting (a multiplicative change of scale, before $\hat{\varepsilon}_{t}$ is trapped in an interior bin), that may be inadequate, leading to repeated sorting of almost all elements while adjusting the cost scale. This trouble is shared with 2-bin sorting (if most of the $\varepsilon_{s t}$-values fall into one bin).

In the start routine, i.e. on the road class level, all computations are node based, whereas in the segment oriented routine the primally determined states may be non-nodal and may differ from the dually ( DynP -)determined states. This means that no re-optimisation of works extents is needed in the primal heuristics of the start routine, since the figures may be taken from the (node based) backwards routine of the dual iteration. Since we register the future traffic and maintenance costs separately in the start routine, a more natural knapsack ranking quotient (solely based on primal costs), with the denominator in (4.9) exchanged for the traffic cost difference, is conceivable.

### 4.2.4 Optimisation of works extent

The model accepts two or more works types, each with a continuously varying works extent (e.g. layer thickness) $z \in[a, b]$, where $a, b$ are given feasibility limits (cf. Sec 3.3). For segment $s$, in a given state $\mathbf{x}_{s t}$ at the beginning of year $t$, and for a given works type (maintenance type) $m=u_{s t 1}$ this means one-dimensional optimisation in order to determine the corresponding optimal works extent $z=u_{s t 2}$. We expect the future costs $\bar{\varphi}_{s, t+1, m^{\prime}}\left(\mathbf{x}_{s, t+1}\right)$ to be non-smooth functions of the resulting state $\mathbf{x}_{s, t+1}$, viz. non-differentiable wherever there is a change of optimum maintenance type for any of the following years in the time horizon. Therefore we let the 1D optimisation rely solely on function values. The stop criterion is that the remaining uncertainty interval, containing the optimum works extent, has an acceptable maximum width. Our general assumption, based on extensive testing, is that the cost function is unimodal on the given interval $[a, b]$. In general, we use and update four $z$-values in the 1Dinterval reduction algorithms, denoted $a$ (lowest possible value), $\alpha$ (e.g. lower golden section value), $\beta$ (e.g. upper golden section value) and $b$ (highest possible value).

One implemented alternative is to use golden section search, supplied by accelerating techniques near the feasibility limits. The initiation is normally controlled by the registered optimum works extent in similar situations. Thus in the backwards routine, if the optimum $u_{s t 2}$ for the same nodal state next year is close to, e.g., the upper bound $b$ we start by computing the summed up maintenance and future costs $C$ for $z=b$ and $z=\beta$, where $\beta:=a+\rho \cdot(b-a)$, $\rho:=(3-\sqrt{5}) / 2$. As long as $C(b)<C(\beta)$ and the remaining uncertainty interval is unacceptably wide, we update $\beta:=\beta+\rho \cdot(b-\beta)$ and recalculate $C(\beta)$. This accelerating
technique corresponds to an interval reduction by the factor $1-\rho \approx 0.382$ per computed value. Then, if necessary, we turn to ordinary golden section reduction (of reduction rate $\rho \approx 0.618$ ). In the forwards routine we start analogously, according to the registered optimum works extent of the preceding year.

The other implemented alternative is to use quadratic approximation. Due to the mentioned non-smoothness, we find higher order interpolation inappropriate. The initial trial values are chosen as for the first-mentioned method. If a minimum-estimate is close to any of the existing interpolation points $a, \alpha, \beta, b$ it is adjusted to a minimum percentage of the interpolation point distance. If we have a function which is judged (strictly) 4-point convex we perform quadratic approximation by computing the least squares solution based on all four points. It can be determined from an explicit formula. Otherwise, if we detect (strict) 3-point convexity we use quadratic interpolation through three points for a minimum-estimate. Otherwise we place the new point at a minimum distance from the best found point.

In the early computer runs we found that the overwhelming part of the 1D-optimisations resulted in an optimal layer thickness at the lower feasibility limit. Therefore the intended method comparison became rather pointless. Instead we supplied the two algorithms with special start routines, enclosing a lower limit optimum in a sufficiently narrow interval by two function evaluations only. In the few cases where the optimum could not be trapped we switch to either of the two main algorithms - see Tab 4.2, where the statistics from 10 dual iterations of a computer run with the golden-section approach is documented. Here the segment oriented main routine has been applied, with $L=3$ node levels per state dimension. In this run the stop criterion is fulfilled if the width of the uncertainty interval for the layer thickness is below 0.5 mm . Since the lower $z$-feasibility limit varies, according to Sec 3.3.3, we present the statistics both for three optimum position cases - lower limit, intermediate, upper limit (Tab 4.2a) - and for the number of 1D iterations (Tab 4.2b).

| Position case | lower limit | intermediate | upper limit |
| :--- | ---: | ---: | ---: |
| Frequency | 78269097 | 4438837 | 897510 |
| Relative freq. (\%) | 93.62 | 5.31 | 1.07 |

Table 4.2a 1D-optimum position frequencies in a test run.

| Iterations | 1 | 2 | $3-12$ | 13 | 14 | 15 | 16 | $m v$ | $s d$ |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Frequency | 2 | 78259420 | 0 | 8723414457688 | 8233 | 7760 | 2.76 | 2.90 |  |
| Relative freq. (\%) | 0.000 | 93.606 | 0.000 | 1.043 | 5.332 | 0.010 | 0.009 |  |  |

Table 4.2b 1D-iteration statistics. Frequencies, average ( $m v$ ) and standard deviation ( $s d$ ).
The classical golden section method, applied to a case where the uncertainty interval is to be reduced from length 90 to 0.5 , would (always) require 12 function evaluations. Thus the special start routine means slightly more 1D-iterations in the intermediate and upper limit cases, according to Tab 4.2b, but the average number $m v$ is much reduced.

### 4.2.5 Residual values

The maintenance options are explicitly decided upon during a given time horizon. At the end of the horizon the reached segment state is important for the future costs. We determine this residual (cost) value by applying a sequence of steps.

1) A special residual-value routine is applied to the network, on the road class level. For a given long-term annual budget level (or a given dual price) we compute residual values for each road class $r$ on a grid of nodal states $\widetilde{\mathbf{x}}_{n}, n \in N$, separating each value into future traffic costs $\bar{f}_{r \infty}\left(\widetilde{\mathbf{x}}_{n}\right)$ and future maintenance costs $\bar{G}_{r \infty}\left(\widetilde{\mathbf{x}}_{n}\right)$ per road metre - see Ch 5 for details. (For future use, where e.g. traffic volume and some costs increase exponentially, also a residual-value model of average costs on a finite horizon is implemented.)
2) Approximate the discounted future costs at time $2 T$ by the computed stationary costs: $\bar{f}_{r, 2 T}\left(\widetilde{\mathbf{x}}_{n}\right) \approx \bar{f}_{r \infty}\left(\widetilde{\mathbf{x}}_{n}\right), \bar{G}_{r, 2 T}\left(\widetilde{\mathbf{x}}_{n}\right) \approx \bar{G}_{r \infty}\left(\widetilde{\mathbf{x}}_{n}\right)$, and apply the usual road class oriented start routine between times 0 and $2 T$ for optimisation. Here $T$ denotes the time horizon to be used in the segment oriented model.
3) Also translate the residual values into future segment costs (cf. Sec 4.2.2): the maintenance costs $G_{s, 2 T}\left(\widetilde{\mathbf{x}}_{n}\right)$ are corrected for discrepancies between the segment $s$ width Wid $_{s}$ and the average road width $\overline{W i d}_{r(s)}$ in its road class $r(s)$, as $G_{s, 2 T}\left(\widetilde{\mathbf{x}}_{n}\right) \approx \frac{\operatorname{Wid}_{s}}{\overline{W i d}_{r(s)}} \cdot \bar{G}_{r(s), 2 T}\left(\widetilde{\mathbf{x}}_{n}\right)$, and the traffic costs $f_{s, 2 T}\left(\widetilde{\mathbf{x}}_{n}\right)$ for discrepancies between the segment $s$ annual average daily traffic Aadt $_{s}$ and the road class average $\overline{\operatorname{Aadt}}_{r(s)}$, as
$f_{s, 2 T}\left(\widetilde{\mathbf{x}}_{n}\right) \approx \frac{\operatorname{Aadt}_{s}}{\overline{\operatorname{Aadt}_{r(s)}}} \cdot \bar{f}_{r(s), 2 T}\left(\widetilde{\mathbf{x}}_{n}\right)$. These corrections will catch all first order model discrepancies. The remaining discrepancy comes from ignoring the segment specific proportions of heavy vehicles in the traffic cost model - see App 1. Another imperfection comes from our approximating the setup costs for coordinated maintenance (cf. Sec 3.6.1) by constants.
4) The DynP backwards routine (cf. Sec 4.2.2) is applied once for each segment, between times $T$ and $2 T$, by use of the translated residual values $f_{s, 2 T}\left(\widetilde{\mathbf{x}}_{n}\right), G_{s, 2 T}\left(\widetilde{\mathbf{x}}_{n}\right)$ at time $2 T$ and the calculated close-to-optimal dual prices found by the start routine, in order to determine residual values at time $T$. The overall purpose is to translate the essentially road class based residual values into more accurate segment specific residual values $\varphi_{s T}\left(\widetilde{\mathbf{x}}_{n}\right)$ at the end of the horizon $T$.
5) The values $\varphi_{s T}\left(\widetilde{\mathbf{x}}_{n}\right)$ are used as given in the subsequent iteration process of the Lagrangean dual, which is confined to explicit computations for times $0, \ldots, T$.

### 4.3 Implementation

### 4.3.1 Running on a PC-cluster

We have made some comments on code parallelisation in the previous sections. The dual subproblem, separable per road class and/or per segment, is ideal for parallelisation but the primal heuristics generates intense inter-computer communication and uses data on all road classes and/or segments simultaneously. Before an optimisation run can be made, the input data and model structures must be split among the processors. Ideally all the collaborating processors handle equally many road classes and equally many segments. In reality we apply either of two principles.

- If the number of processors exceeds the number of road classes, we equalise the number of segments per processor, by splitting the most frequent road classes on two or more processors. For database Värmland, with $R=29$ road classes and 8749 segments, for $R$ or more processors this means at most two road classes per processor.
- Otherwise we do not split the road classes but combine them for equalisation.

This data allocation problem can itself be formulated and solved as an optimisation problem. However, we have used different heuristics for the two equalisation principles. Since the residual value and start routines are less time consuming and memory demanding than the main routine, they can be run on a smaller PC-cluster, for one and the same level of discretization $L$. The results are then recombined for the main run.

### 4.3.2 Computer memory

We have implemented hierarchical model structures. Cf. Sec 1.3. In the start routine we generate one processor specific "chain" of road classes and one "chain" of time points (years), together identifying the nodal grid components for each road class in time and space. Depending on the system properties and the wanted level $L$, we have prepared two extreme variants.

- The whole structure is kept in main (primary) memory, for fastest access and computations. Moreover, all the $2^{D}$ interpolation weights and values are registered, per road class state (static data for routine maintenance) and road class time state (dynamic data for major maintenance), respectively. For $D=L=5$ and $2 T=80$ this corresponds to 3.4 GB memory, i.e. an acceptable amount for a cluster ("Penta") of four parallel computers of 1 GB CPU and 10 GB secondary storage each, which we have access to.
- Temporary binary data files are used whenever possible. The minimum need for structure is a complete grid (of $2 T$ years) for one road class. In the primal heuristics this is reshaped as a grid for the current time for $R(<2 T)$ parallel road classes (plus one extra for updating). Instead of saving complete interpolation weights and values such are recalculated whenever needed from the registered underlying factors, one per state dimension, and one reference node (as in Ch 2). Such data compression allows


### 4.3 Implementation

much higher $L$-values and the practical limits come from CPU-times, e.g. for the increased reading-writing from files, and/or data addressability in the program code.

In the main routine we have implemented the first variant only, since we have access to a PCcluster of several hundreds of processors. The total structure for $D=L=5$ and $T=40$ amounts to 48.8 GB . For this we have utilised at least 33 nodes of 2 processors and 2 GB each.

### 4.3.3 Absolute and relative state limits

An open question (cf. Sec 3.2.2) is how to interpret the acceptance limits for the worst pavement states: to prohibit or to penalise (or to just notice) any violation. For direct result comparisons between these options, the grids should coincide within any common acceptance intervals. A basic grid may consist of $L$ equidistant levels per state dimension (and time period), from the ideal value, e.g. 0 , to the very acceptance limit. Moreover, one external level is needed for the penalty option, in order to reduce the number of extrapolation cases. Our program admits a free number of "density-rounds", for the generation of additional grid node levels to the given ones, in each round adding a mid level between the existing ones. For, e.g., the penalty option and $L=5$ given levels (with one level above the acceptance limit) and one "density-round" we get $L=5+4=9$ levels per dimension, with two levels above the acceptance limit, and $9^{5} \approx 590005 \mathrm{D}$ nodal states.

Even in the interpretation as absolute limits, the penalisation option can be useful: In some runs we have noticed difficulties in generating any feasible solution at all in the absolute limits case. Then a relevant set of dual prices can be identified by pre-optimising for a high penalty instead.

Our implementation of an absolute limit may need some clarification. In the start routine, where the registered future costs per node do not differentiate between the succeeding maintenance types, routine maintenance is forbidden for any node on the highest level in the corresponding state dimension. In the main routine, where such differentiation is made, routine maintenance is forbidden if the limit is violated, i.e. if routine maintenance is applied on or close to the highest node level it must be followed by major maintenance. (A future cost for each node and works type is needed for interpolation purposes.) If the discrepancies between the start and main routines are considered too big some kind of differentiation for the future costs is conceivable also in the start routine, e.g. between routine maintenance and all other works types (here major maintenance).

### 4.3.4 Discretization errors

In each segment subproblem the DynP backwards and forwards routines are applied - cf. Ch 2. In the backwards routine the calculations are performed on a grid of $L^{D}$ nodes, positioned at the intersection of $L$ node levels in $D$ state dimensions. In the DynP forwards routine the actual, continuously varying, segment state at each time $t$ is used, for determining the correct traffic and maintenance costs. But the optimum decision relies on the calculated future costs from time $t+1$ and onwards, established in the grid based backwards routine.

The residual value and backwards routines, both grid based, will introduce discretization errors. We recognize such errors in different ways. One manifestation is the differences as to the calculated future costs, between the backwards and forwards routines, measured from the given initial segment states. The size of these differences is a major drawback to our implementation. On an aggregated level, as measured by the total cost in the dual subproblem, it is realized to be a few percent - cf. Tab 4.3, which presents overall dual cost differences as relative errors. The negative signs mean that the linearized, backwards computed cost is higher than the correct, forwards computed cost - a sign of mainly convex cost functions. The data are for the best found solutions (on the segment level) in a given budget run, made for various numbers $L$ of node levels. In the start routine (small) discrepancies arise, only in the start year - where the state space is continuous (to permit the use of the initial segment states).

| Number of node levels $L$ | 3 | 4 | 5 | 6 |
| :--- | ---: | ---: | ---: | ---: |
| Number of states per yr | 243 | 1024 | 3125 | 7776 |
| Relative error (\%) | -5.47 | -3.80 | -2.40 | -2.09 |

Table 4.3 Relative dual cost error (\%) backwards vs. forwards for best found solution.
Although the (segment-length-weighted) error gets smaller as $L$ increases the computational price is big, as the decisive number $L^{D}$ of states per year in Tab 4.3 show. However, the discretization error is expected to decrease linearly with the average node level distance in each state dimension - cf. Ch 9 - i.e. to be inversely proportional to $L-1$. The results in Tab 4.3 do not contradict such behaviour, although the error for $L=6$ is somewhat higher than expected.

| LI Quotient | $0.5-0.6$ | $0.6-0.7$ | $0.7-0.8$ | $0.8-0.9$ | $0.9-1.0$ | $1.0-1.1$ | $1.1-1.2$ | $1.2-1.3$ | $n$ | $m v$ | $s d$ | sdRoadd |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 3 | 0.23 | 1.57 | 4.31 | 19.66 | 50.44 | 21.44 | 2.31 | 0.04 | 87490 | 0.942 | 0.092 |  |
| 4 | 0 | 0.55 | 2.35 | 10.41 | 66.52 | 19.95 | 0.22 | 0 | 87490 | 0.954 | 0.067 |  |
| 5 | 0 | 0.21 | 1.58 | 7.18 | 63.86 | 26.47 | 0.71 | 0 | 87490 | 0.967 | 0.065 | 0.043 |
| 6 | 0 | 0.03 | 1.13 | 6.02 | 67.90 | 24.86 | 0.06 | 0 | 87490 | 0.967 | 0.058 | 0.037 |

Table 4.4 Percentage distribution of cost quotients backwards over forwards for various levels of discretization $L$. Ideal quotient 1.0. $n=$ number of observations, $m v=$ mean value, $s d=$ standard deviation, sdRoadcl $=$ average within-road-class standard deviation.

The errors per segment show a greater variation. As $L$ varies we have noticed that the (crucial) forwards routine costs are stable, whereas the (supporting) backwards routine costs show more variation. This robustness speaks for the use of a low $L$-value, despite the errors. In Tab 4.4 we present some statistics for the cost quotients backwards over forwards. As can be seen the quotients vary between below 0.7 and above 1.1 for all discretization choices $L=3, \ldots, 6$. The mean values $(m v)$ and standard deviations ( $s d$ ) approach 1.0 and 0.0 , respectively, slower than expected. For comparison we have computed the average within-road-class standard deviations (sdRoadcl), which show to be much smaller. However, the statistics in Tab 4.4 is for 10 iterations, i.e. the $s d$ Roadcl-values may be biased by possible repetitions of unchanged maintenance plans for the segments. To handle this, we formulate an extreme case:

Theorem 4.3 If the outcomes $\left(x_{i}\right)_{i=1}^{n}$ of $n$ statistical trials are 10 -fold duplicated, the sample standard deviation is reduced by the factor $\sqrt{(n-1) /(n-0.1)}$.

Proof The original standard deviation is $\sqrt{\frac{n}{n-1}\left(\overline{x^{2}}-\bar{x}^{2}\right)}$, where bar $\left(^{-}\right)$denotes sample mean. If all outcome frequencies are increased 10 -fold, the very same means $\bar{x}, \overline{x^{2}}$ will result but the standard deviation becomes $\sqrt{\frac{10 n}{10 n-1}\left(\overline{x^{2}}-\bar{x}^{2}\right)}$.

In Tab 4.4 the underlying within-class standard deviations have been corrected according to Thm 4.3. Despite this extreme correction the average within-class variance is clearly smaller than the overall variance, and we anticipate systematic between-class differences (an idea supported by clear class $m v$-differences). The most obvious cause is the data and model differrences listed in Ch 3 and App 1 - rather than implementation faults. Also cf. the following:

In Ch 9 we will present some alternatives replacing the ordinary multilinear interpolation scheme, intended to reduce the errors. Here we investigate some other possibilities. On behalf of the interpolation errors as being controlled by the grid density, for a given number $L$ of node state levels the equidistant grid is natural. On the other hand the size of the value function is equally important for the errors, speaking for a denser grid towards the bad states where cost is high. Since we also suspect that the nonlinearity of the discounted future costs becomes greater as the segment state gets worse, we have tried letting the grid density vary with the state level. In the first place such a differentiation should be made on the road class level, with road classes of high AADT-values expected to have decent/good average states, and low-traffic road classes expected to spend more time close to the upper state limits. In our implementation we control this density variation by an exponent $p$, as the relative inter-level distance $(L-\ell)^{p}$ for node level $\ell=1, \ldots, L-1$ (where $\ell=1$ refers to the ideal state), e.g. $p=0$ corresponding to equidistant node levels. In Tab 4.5 we have investigated the effects of some $p$-values on the segments in a small road class $\mathbf{t} 2$, consisting of just eight segments.

|  | $\mathrm{p}=0$ (equidistant) |  |  |  |  | $p=0.5$ |  |  |  |  |  | $p=1$ |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| segmlL | 3 | 4 | 5 | 6 | 7 | 3 | 4 | 5 | 6 | 7 | 8 | 3 | 4 | 5 | 6 | 7 |
| 3345 | -6.6 | -5.3 | -5.4 | -4.7 | -4.4 | -5.8 | -5.7 | -5.0 | -3.8 | -3.2 | -2.9 | -5.8 | -3.3 | -3.8 | -3.7 | -3.4 |
| 3346 | 2.1 | -2.5 | 2.7 | 1.9 | 2.1 | 2.7 | 5.9 | 7.4 | 9.3 | 10.2 | 9.5 | 2.9 | 9.4 | 10.8 | 11.5 | 12.6 |
| 8851 | -3.8 | -4.0 | -2.5 | -2.5 | -2.4 | -2.9 | -2.0 | 0.0 | -0.4 | 0.1 | 0.0 | -2.1 | -0.4 | 1.8 | 1.9 | 2.3 |
| 8852 | -4.3 | -4.1 | -1.6 | -2.5 | -1.9 | -4.1 | -3.0 | -0.8 | -1.5 | -0.7 | -1.4 | -4.2 | -2.9 | -0.2 | -1.2 | -0.2 |
| 8853 | -7.7 | -7.2 | -4.8 | -4.8 | -4.5 | -6.3 | -3.2 | 0.6 | 1.1 | 1.8 | 2.3 | -5.3 | 2.2 | 5.1 | 6.2 | 6.7 |
| 8854 | -30.5 | -22.5 | -19.0 | -18.2 | -16.9 | -26.6 | -15.2 | -12.9 | -9.4 | -8.9 | -7.3 | -20.8 | -13.2 | -6.4 | -5.9 | -3.0 |
| 8855 | 6.5 | 4.8 | 3.5 | 3.6 | 2.3 | 11.1 | 8.6 | 11.7 | 10.7 | 10.5 | 10.7 | 15.7 | 11.7 | 16.7 | 16.7 | 17.5 |
| 9480 | -3.9 | -4.9 | -3.3 | -2.0 | -3.3 | -4.5 | -1.3 | -0.4 | -0.2 | 0.5 | 0.2 | -5.3 | 0.4 | 0.0 | 1.6 | 1.4 |
| mv | -6.0 | -5.7 | -3.8 | -3.6 | -3.6 | -4.6 | -2.0 | 0.1 | 0.7 | 1.3 | 1.4 | -3.1 | 0.5 | 3.0 | 3.4 | 4.2 |
| sd | 10.9 | 7.7 | 6.9 | 6.6 | 6.0 | 10.6 | 7.2 | 7.4 | 6.6 | 6.5 | 6.1 | 10.1 | 7.8 | 7.6 | 7.7 | 7.5 |

Table 4.5 Relative dual cost errors (\%) backwards vs. forwards for best found solution in a small road class, $\mathbf{t 2}$, for three different dicretization strategies characterized by $p$.

Here another property becomes visible: especially for $p=1$, i.e. linearly decreasing node distances, the dominant trend, as the number $L$ of node levels increases, is that the error goes from negative, via null, to positive, i.e. does not stop at null, as expected. The same behaviour applies to the square-root decrease $p=0.5$ for segments 3346 and 8855 . It shows that the possible success of a non-constant grid density depends on the $L$-value being used. For the two
segments 3346 and 8855 the choice $p=0$ is superior. For the rest a higher exponent $p$-value is preferred, here mostly $p=0.5$. For that reason we have run all of our tests with this $p$-value (mostly using $L=5$ or $L=6$; with small average errors $m v$ in Tab 4.5), although we reckon that error reductions are possible by fine tuning the $p$-value for each segment. The standard deviations ( $s d$ in Tab 4.5) are rather insensitive to the $p$-value being used.


Figure 4.1 Relative dual cost errors backwards vs. forwards for best found solution for two segments. Different numbers $L$ of node levels; logarithmic cost scale vs. straight costs.

Another possibility of error reduction is illustrated in Fig 4.1. In the two most frequent road classes, $\mathbf{s 2}$ and $\mathbf{n 5}$, the most extreme segments, in this respect, both have length 1 m . Whereas an (asymptotic) error shrinkage as the inverted average node level distance is expected, i.e. as $(L-1)^{-1}$, the reduction is much slower, e.g., more than $40 \%$ of the errors remaining from $L=3$ to $L=9$ instead of the expected $25 \%$. We suspect that the reason is partly the nonlinearity of the cost functions involved. Therefore we have performed multilinear interpolation with the cost logarithms instead of the straight cost values, whenever the costs to interpolate are above 1 (since then $\Delta \log C<\Delta C$ ). (Or we might have used $\log (1+C)$ always.) Also this result is displayed in Fig 4.1. The errors for the $\log$ scale, $L=4$ are competitive with the straight cost scale, $L=9$. For a run on the whole network, with a common set of dual prices - close-to-optimal for the straight cost interpolation - and $L=3$ we get the total backwards vs. forwards cost error (cf. Tab 4.4) $-5.54 \%$ (straight) and $+5.56 \%$ (log). The segment based results (cf. Tab 4.5) are $m v=0.942, s d=0.093$ and sdRoadcl $=0.062$ (straight) and $m v=1.059, s d=0.110$ and $s d R o a d c l=0.064(\log )$. Thus the within-road-class standard deviations are clearly less than the overall values, i.e. the road classes seem to be homogeneous in this respect. A cyclic pattern is clear in Fig 4.2, where the average values per road class are plotted, with local maxima for traffic class $\mathbf{1}$ (see Sec 3.1.2) and different speed limits. The deviations between the two curves point at a systematic effect of the changed cost

### 4.3 Implementation

scale. The ideal value, quotient 1.0 in Fig 4.2, is in between the curves for a majority of road classes. This suggests road class specific scales, e.g. powers with varying exponents. The price to pay for these improvements is an increased computational burden of cost transformations at almost every interpolation.


Figure 4.2 Cost quotients backwards vs. forwards per road class. Ideal quotient 1.0.
Logarithmic cost scale vs. straight costs.

### 4.3.5 Reservations as to traffic evolution

Although the VV database includes information about traffic changes no specification is made in Ch 3. In our implementation we do not consider any such data. This can be interpreted as if all segments implicitly obey one and the same annual change rate $\bar{\alpha}$ and that all maintenance costs, as well as the budgets, implicitly show the same annual change rate $\bar{\beta}=\bar{\alpha}$. "Implicitly" here means that it is a hidden factor of the discount factor $d$. Although reasonable on the network level such a view may seem too restrictive to hold for each single segment. This objection speaks for a consideration of the segment deviations from the average behaviour. Moreover, $\bar{\beta} \neq \bar{\alpha}$ might be of interest for a study. However, the VV segment data cannot be applied without additional assumptions. For each segment $s$ the change rate $\alpha_{s}$ must satisfy $\frac{\alpha_{s}}{\bar{\alpha}} d<1$ for the future traffic costs to be finite, if the change rate is for ever. (Here $\bar{\alpha}$ is part of d.) Otherwise some levelling trend model is needed. As an alternative we have implemented a residual value model based on average costs for a finite horizon, admitting free evolutions of $\alpha_{s}, \beta_{s}$. Anyhow the Ch 3 effect models, subdividing the segments into traffic based road classes, become irrelevant. Here some "road class free" models, as we suggested in Sec 3.9.3, would be more natural.

### 4.4 Case study: Värmland

### 4.4.1 General results

Our run time saving strategy is to avoid generating and storing statistical results during the very solution process. Instead we recommend a succeeding special statistics run for the one or two iterations of interest, corresponding to the best found dual and primal solutions. We generate overall and road class based statistics for maintenance, state and cost. Cf. the initial study, Ch 2 . Here we illustrate the output diagrams with two examples - see Fig 4.3.


In Fig 4.3a the distributions of Age immediately before a major maintenance operation are given as histograms per traffic class, for two velocity classes: $70 \mathrm{~km} / \mathrm{h}$ to the left of the vertical lines, and $90 \mathrm{~km} / \mathrm{h}$ to the right. Some Age-values are over-represented, e.g. the unexpected Age $=0.5$. But remember (from Sec 3.5.5) that Age is not the same as the time since the previous major operation and that $I R I$ is also a triggering factor. This is a run with absolute state limits, shown by the dominant boxes at the highest feasible Age-level per road class. (In fact this is from Run 3 in the road 63 study - see Sec 4.5 below.) The short horizontal bars mark the averages. Thus in this run (early in the project) the average is only slightly dependent upon the road class, despite the clearly different Age-limits. In Fig 4.3b the time evolution of the averages of two state variables, $I R I$ and Age, is shown for a (recent) run with the start routine. (In fact this is from the budget run in Sec 4.4.2 corresponding to the long term dual price $v_{\infty}=v_{\mathrm{BCR}}$.) The initial state in Värmland is found inappropriate in two different ways: whereas the IRI-average can be decreased by time, the Age-average is admitted to increase by time, as a more cost-effective solution according to the models in Ch 3.

### 4.4.2 Budget runs

We illustrate the possibilities by some comparison runs performed by the start routine with $L=4$ levels per state dimension, using absolute state limits and various budget levels. In Fig4.4 several comparisons are shown. Figs 4.4 a - b are almost identical, meaning that there is an almost constant affine transformation from long term dual price (a) to total traffic cost (b). In a very narrow budget interval of just $0.6 \mathrm{MSEK} / \mathrm{yr}$, i.e. less than $1 \%$, the long term dual price $v_{\infty}$ varies between $v_{\infty}=v_{\mathrm{BCR}}=1.2439$ and $v_{\infty}=15$, and the traffic cost shows a $40 \%$ increase. Thus the modelled maintenance system is most sensitive to budget reductions and extraordinary costs in this region, rapidly increasing the return rate demands (here $v_{\infty}$ ) for a maintenance project to be accepted.


Figs $4.4 \mathrm{c}-\mathrm{d}$ display the corresponding effects upon two state variables, IRI and Age, as measured after 10 yrs. Whereas the IRI-average decreases by around $4 \%$ as the budget level increases, the Age-average increases in most of the budget interval. We interpret this surprising result by the non-optimal initial Age-distribution, as in Fig 4.3b. The curve jumps in Fig 4.4d show that the best found solutions are of varying quality.

### 4.5 Case study: road 63

A possible future use of the program is the following: Since the run times become high if all regional or national segments are considered, and since (cf. Ch 2 ) the dual prices determined in the faster start routine are rather good, a short cut is to run the main routine one iteration only, on a subset of the road segments of immediate interest, while activating all output data routines prepared. This means full optimisation in the start routine, and a kind of consequence computation in the main routine. We apply this run strategy for validating the basic model. A full-scale case study is reported in Ch 9.

### 4.5.1 Run strategy

The budget shadow prices and the future costs at the time horizon, as determined by the start and residual value routines, are assumed fixed in the main routine. In the start routine we apply standard input cost values for all setup costs. There is no incentive for iterating in the main routine, and the dually determined maintenance plans are taken as the solution, since no overall checking for feasibility is possible on this subset of road segments. The segments will carry the correct setup costs, except those for coordinated maintenance. By setting up different program assumptions we accomplish a kind of sensitivity analysis.

### 4.5.2 Results

The target for our validation was road 63, in Värmland running 81 km Karlstad - Molkom Filipstad - province border (towards Hällefors). It was chosen as a well-known national road, consisting of segments of various road classes. As for the registered VV-data, all segments except one (of length 1 m ) had acceptable information, leaving 212 segments for investigation. A special validation meeting was held, with participants from VV and VTI. The purpose was not to sanction the results, rather to analyze, question and suggest further improvements, since the meeting was held soon after a first version of the model in Ch 3 had been formulated and implemented.

Five runs were scrutinized at the validation meeting. We will comment upon the comparison runs between the three possible interpretations of the upper state limit concept (see Sec 3.2.2). These runs utilized the same budget assumptions, on the network level corresponding to a decreasing annual funding from initially 66 MSEK to 64 MSEK after 5 years. This was also used in Fig 4.3a above.

The magnitude of these annual budgets is itself a remarkable result. Although optimisation means cost efficiency we do not claim that these figures are comparable with the annual VV budget level for Värmland - almost 2.5 times our values. Some discrepancy is expected, since the VV-budget also finances pavement maintenance generated by other state parameters than our variables IRI, rutting and Age. Thus, according to VV judgements [Lang (2004)], 40-60\% of the maintenance is referable to surface deficiencies as to, e.g., texture and cracking, and structural conditions such as bearing capacity. In practice also non-modelled obstacles may prevent maintenance strictly according to the optimal plans. After the validation meeting several model changes have been implemented but our standard budget level is still around 75 MSEK for a (stationary) dual price equal to $v_{\mathrm{BCR}}$ (cf. Fig 4.4a for $L=4$ ). The long term ambition is of course to include all maintenance activities and also to improve the cost-effect model, database and method for better coincidence.

The comparison runs are named Run3Abs (absolute IRI, rutting and Age limits), Run4Rel (relative IRI, rutting and Age limits, any violation penalized) and Run5Free (free limits, only for observation purposes). Before each run the start routine was applied, with the corresponding meaning of the limit concept.

The return rate constraint (cf. Sec 3.7.2) was not considered and not satisfied in all runs (Run5Free). This facility was implemented later on.

For presentation the segments were sorted according to the physical order along the road. For each segment the main output is a maintenance plan and other results by the year: state variable values, traffic and maintenance costs, future costs for each works type, layer thickness for an optional major maintenance operation and chosen maintenance. These results are also plotted see Fig 4.5 for Run3Abs and Fig 4.6 for Run5Free. The horizontal axes show time, the horizon being 21 years here. The vertical axes are for two different road sections, imitating the road direction. Each horizontal line bounds a segment, with number and road class to the left, e.g. in Figs a, c beginning with № 2233 (in Karlstad), road class s6. The distances between the horizontal lines reflect the real segment lengths, except for the shortest segments. The greyscales to the right are for the state variable in question, here IRI in Figs a, b and Age in Figs c, d, spanning from the ideal white to the worse blackish conditions. Notice that the road classes have different upper state acceptance limits. The thick vertical lines mark a major maintenance operation, positioned at the end of the operation year, e.g. at time 1 for a major operation during the $1^{\text {st }}$ year. The short, thick horizontal marks illustrate the relative layer thickness, from the lowest admissible value at the bottom to the maximum thickness (here 100 mm ) at the top for each segment.

For the validation we had access to a defect-inspection of the pavement conditions in Värmland [Daradian (2003)], made at approximately the same time as our VV-data were extracted, in the autumn 2002. In this defect-inspection the roads and road sections were classified according to the judged requirements, as maintenance categories Full, Major, Minor and None. The time points for the $1^{\text {st }}$ major maintenance operation in our output data were compared with the classifications. The comparison result is shown in Tab 4.6. The Run3Abs-results in parenthesis
in Tab 4.6 denote sparsely occurring values. Thus for all occasions of defect-inspection category None, Run3Abs generates a high time value, at least the $9^{\text {th }}$ yr for $1^{\text {st }}$ major operation: good agreement. For category Minor, our computer run states at most the $9^{\text {th }}$ yr: good agreement. For category Major, the Abs run mostly says $5^{\text {th }}$ yr: rather good agreement. Finally, for category Full, Run3Abs mostly suggests $2^{\text {nd }}$ yr: rather good agreement.

| sourcelLopa | $0-7000$ | -11000 | -12900 | -15280 | -27670 | -34770 | -58120 | -61740 | -67530 | -81000 |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Defect-insp. | None | Major | Full | Major | Full | None | Major | Minor | Major | None |
| Run3Abs | 9 | $5(9,16)$ | $2(5,9)$ | $5(9)$ | $2,5(9)$ | 9 | $9(2,4,5)$ | $9(2,5)$ | $5,9(1,2)$ | 9,18 |

Table 4.6 Road 63. Comparison between requirements (maintenance category), according to Defect-inspection [Darabian (2003)], and year for $1^{\text {st }}$ major maintenance operation, according to Run3Abs. Lopa $=$ distance (in metres) from Karlstad.

As for the general results, objections were made for unrealistically short periods between the major operations, especially in the Free and Rel runs (the latter not shown here). This phenomenon is accompanied with thin pavement layers (a necessity the other way round: thin layers must soon be replenished), with the horizontal thickness-marks at the bottom in many cases in Figs 4.5-4.6, especially in the Free run. The modelling trouble is that the quality of several thin layers is not the same as for one thick layer. Later on Age-differentiated lower limits (cf. Sec 3.3.3) were introduced for the layer thickness. However, the objection remains (cf. Sec 4.2.4) - but after the differentiation was made, the "bottom marks" normally represent thicker layers than before. Moreover, we might argue that the initial Age-averages in Figs 4.3b and 4.4 b indicate even more frequent major maintenance operations performed in practice.

In the three runs the possible state variable values (Age excluded) immediately after a major maintenance operation were bounded between 0 and the value immediately before the operation. Since the restrictions on the degradation rates $d I, d R$ were realized to lead to ideal states 0 (without extra costs), these were later on replaced according to Sec 3.5.

As for the simultaneous maintenance of a sequence of segments, although our approximating the coordination setup costs with constants means that the segments lack an economic incentive, Figs $4.5-4.6$ show synchronization, with some exceptions. Since the pavement histories of adjacent segments are expected to be similar, this is a natural cause of simultaneous maintenance in the runs - especially if Age is the triggering factor. In fact this is the case in quite many cases in Run3Abs, as can be understood from the dark grey Age-states in Figs 4.5 c , d immediately before major operations. The acceptance limit is, e.g., 17 yrs for traffic class 6 and 15 years for class 7 . However, there are some short segments not maintained when the neighbours are. Such question marks are part of the validation. For the five deviating segments in Figs 4.5 a, c - № 2233, 2234, 2236, 2239 and 2263 - we turn to the detailed output lists (not shown).

- № 2233 belongs to road class s6, like № 2238. The initial IRI-state is better for the former and its length is one third of the latter, tripling the setup cost per metre. This speaks for postponed maintenance for № 2233. Moreover, the future cost shows an unexpected jump near the end of the horizon. This also happens to № 2236, 2239 and 2263. As a consequence we later on improved the residual-value computations, according to Sec 4.2.5.
- № 2234 belongs to road class $\mathbf{s 5}$, like № 2237. The initial IRI-state is better for the former; yet major maintenance is one year before the latter. However, the degradation rate $d I$ is much higher for № 2234 - which may explain the economic advantage of an earlier operation. Its double length in comparison with № 2237, halving the setup cost, also speaks for an earlier realization.
- № 2239 has an initial rutting above the acceptance limit, necessitating an immediate major operation. № 2263 has a $d I$-value much higher than the neighbour segments, motivating early major maintenance.
- In the Free run № 2295 and № 2296 might be compared, as well as № 2297 and № 2298 (all n6-segments). In both pairs one segment is much shorter than the other, at least ten folding the setup cost. These are cases where a more careful model for coordinated maintenance should make a difference - see Ch 6 below.

For two n6-segments, № 2297 and 2299, the $2^{\text {nd }}$ major maintenance operation was postponed one year (until yr 17 - see Figs 4.5 b, d), the Age-values exceeding 18 yrs, instead of just the limit 17, before the major operation. The code was revised later on, in order to prevent such program errors. (Similar objections can be made on Fig 4.3a above.)

If the respective Figs $\mathrm{c}, \mathrm{d}$ are compared we notice that the Age-values, often triggers in the Abs-run, become higher before the first major operation in the Free run, as expected. Notice that the differences between the runs are bigger than it seems, since the grey-scales are different. In general, the first major maintenance is performed later in Fig 4.6 than in Fig 4.5; reasonable when no state limits are active. Then why are Figs d lighter than Figs c, although all include, e.g., many n6-segments? The motivation comes from Figs a, b where the greyness differences are slightly reversed for the $\mathbf{n 6}$-segments: The initial IRI-values are worse and act as triggers.

A general question mark was put on the computed traffic costs. As an extra check-up we computed such costs in SEK per metre and year for some segments on road 63, for some IRIvalues. These costs were then recalculated by VV. No differences were found.

As a validation tool the plots (Figs 4.5-4.6) received some criticism. It was discovered that some segments were simply missing from the output. Moreover, since the VV-database is based on Lopa (see Sec 3.1.1), here the distance from Karlstad, it was recommended to print these values at the segment numbers. All this has been taken care of.

In summary, the validation meeting was very constructive and resulted in several model and program improvements. A still open question concerns the predominance of thin layers. Our hypothesis is that it will be answered by a minor revision of the models in Ch 3 . An insight into the optimisation mechanism is that simultaneous maintenance of consecutive segments often but not always - results quite naturally. However, whenever a subnet, like road 63, is studied and it is possible to add information about a suitable road sectioning for optional coordination, the coordinated maintenance model in Ch 6 is conceivable, in order to achieve an increased confidence in our results.


Figure 4.5a,b Run 3Abs.Two sections of road 63. Time evolution of IRI-values, according to grey-scale


Figure 4.5c,d Run 3Abs.Two sections of road 63.Time evolution of Age-values, according to grey-scale
4.5 Case study: road 63


Figure 4.6a,b Run5Free.Two sections of road 63.Time evolution of IRI-values, according to grey-scale


Figure 4.6c,d Run5Free.Two sections of road 63.Time evolution of Age-values, according to grey-scale

## 5 Residual values

In road maintenance planning the residual values should measure the future costs to be expected after the explicitly considered planning period. One possibility is to use zero residual values and a long planning period. For positive discount rates, an alternative - admitting a reduction of the considered time period - is to replace the zeroes by steady-state values, for an infinite subsequent time. The determination of such steady-state values is the topic of this chapter.

### 5.1 Background

Let us first describe the basic model structure (cf. Ch 4) in a slightly different setting. In our road class oriented maintenance applications we utilise a finite set of nodal states to characterise the pavement condition, and use discrete time for its evolution. In a first approximation the current state of a road segment and the immediate maintenance operation together determine the subsequent state transition (whereas the earlier states are irrelevant) the Markov property. (This is partly a question of state definition: in Ch 3 we included IRI- and rutting deterioration rates to achieve this property.) We consider the distribution of road lengths between the nodal states at $1-\mathrm{yr}$ times, in different road classes. We perform discretization of the continuous layer thickness variable in Sec 3.3, to have a finite set of works options and we apply deterministic models for the deterioration (by routine maintenance) and improvement (by major maintenance) of the state variable values. If a node based maintenance operation leads to a non-nodal state one year later, the maintained road length in our model is split among the neighbouring nodal states. Except for extrapolation situations the corresponding relative length distribution is equivalent to a probability distribution for the state transitions in a finite Markov chain (MC), with our distribution weights ("probabilities") chosen so that the average ("expected") state variable values coincide with those of the true succeeding state. Problems where the evolution is (partly) controlled by decisions, MDPs (cf. Sec 1.4.1), were formulated and solved by [Bellman (1957)] and [Howard (1960)], devising the two standard iteration methods. For, e.g., the sequential decision process with/without discounting [ibid] the optimal operation ("policy") and the optimal value per state are simultaneously determined, both for general dynamic evolutions and for their limit, steady-state problems with nonzero and zero discount rates. Howard [ibid, p 83] characterises the resulting policy as statewise optimal. We can use the optimal values as residual values, provided that our annual budget restrictions are met. For a budget evaluation, the optimal steady-state distribution of the road lengths among the nodal states is needed. Such a distribution is also the result of the long-term model in Alviti et al (cf. Secs 1.4.2-1.4.3) - a road class based model maximising the traffic benefit, subject to an overall annual budget condition and road class specific constraints. Whereas they use the optimal long-term distribution of the road lengths themselves as terminal conditions, i.e. constraints, in a short-term model, we will use the residual values (i.e. the expected future costs per metre) for price-controlling the evolution during the explicitly considered planning period. Alviti et al apply Dantzig-Wolfe decomposition to the whole long-term problem, whereas we will build on the Bellman - Howard approaches. We will formulate two similar models, each providing both residual values and length distributions. We will, e.g., combine the two standard

### 5.1 Background

solution methods, policy and value iteration, as well as apply general LP and Newton's method - which we prove will converge. At solving the Lagrangean dual by the Dantzig-Wolfe method we introduce a modification, for improved convergence in non-convex problems.

### 5.2 Models

What steady-state assumptions are reasonable? How can we motivate an assumption of a constant works policy (= decision rule), i.e. a stationary choice of works option for each pavement nodal state? On behalf of an assumed stationary budget level and our models in Ch4, this assumption is linked to a stationary total maintenance cost and a stationary utilizable budget. The two last-mentioned properties restrict the overall length distribution, permitting the lengths per nodal state to fluctuate. However, a stationary total maintenance cost is an unwanted condition, introducing coupling constraints between the road classes / segments that would be a computational challenge. Instead we may control both properties, by assuming that all the relaxed budget constraints correspond to a stationary Lagrangean multiplier value reasonable, if the total maintenance cost and the utilizable budget are stationary. This is a dual type of constraint and we call this model, (5.1) below, semi-stationary (because of the remaining length variation). Another possibility would be to assume that the time evolution has led to a constant length distribution as well; we call it the steady-state model, (5.5) below.

In this chapter we will use time index $k$ (and occasionally $\kappa$ ), denoting the terminal time by $K$, to distinguish from transposing ${ }^{T}$.

We will determine the residual values, at time $K$, and the length distributions in two separate models. This separation is possible, due to our main assumptions of the stationary works policy and multiplier value.

### 5.2.1 Residual values model

In the road class oriented, dynamic model (4.7) we applied a residual value $\varphi_{r n K}$ to each nodal- $n$ state $\mathbf{x}_{r n}$ of road class $r$ at the time horizon $K(=2 T)$. These values are substitutes for the optimal future costs in an implicit infinite succeeding time period. The first, semi-stationary model (5.1) below is formulated in order to establish the form of these future costs. We think of the time evolution as one dynamic period, described by (4.7), switching to the infinite-time semi-stationary model at some unspecified transition time point $\kappa$ (instead of an asymptotic approach). Hypothetically we presume that the influence of (4.7) is transferred as an initial length distributions $\left(\lambda_{r n}^{0}\right)_{r, n}$ given at $\kappa$. We repeat the semi-stationary (heuristic) model assumptions:

- $\kappa$ begins a stabilised budget situation, meaning a stationary annual budget $b_{\infty}$ and a stationary (optimal) multiplier value $v_{\infty}$ for the relaxed annual budget constraints,
- $\kappa$ begins a stabilised pavement works situation, meaning a stationary (optimal) decision rule $\left(\mathbf{U}_{r}\right)_{r}=\left(\mathbf{u}_{r n}\right)_{r, n}$ each year.

In comparison with (4.7), in (5.1) we keep the 2D works description $\mathbf{u}_{r n}$ of type and extent from Sec 3.3 , let $k$ for simplicity measure the future time from the switch point $\kappa(k=0)$ and onwards, and will write the nodal based quantities without tilde notations (since we risk no confusion with any segment quantities here). By writing the traffic and maintenance costs $f_{r n}$ and $g_{r n}\left(\mathbf{u}_{r n}\right)$, respectively, and denoting the transition functions $p_{r n n^{\prime}}\left(\mathbf{u}_{r n}\right)$, to emphasise their character of transition probabilities to the different succeeding states $n^{\prime}$, the semistationary model becomes

$$
\begin{align*}
& \underset{\left(\lambda_{r n k}\right)_{r, n, k}\left(\mathbf{u}_{r n}\right)_{r, n}}{\operatorname{minimis}} \sum_{r} \sum_{n} \sum_{k=0}^{\infty} d^{k} f_{r n} \lambda_{r k n}+v_{\mathrm{BCR}} \cdot \sum_{k=0}^{\infty} d^{k+\frac{1}{2}} y_{k}  \tag{5.1a}\\
& \text { subject to }\left\{\begin{array}{cc}
\sum_{r} \sum_{n} g_{r n}\left(\mathbf{u}_{r n}\right) \cdot \lambda_{r k n} \leq y_{k} & \forall k \\
y_{t} \leq b_{\infty} & \forall k \\
\lambda_{r, k+1, n^{\prime}}=\sum_{n} p_{r n n^{\prime}}\left(\mathbf{u}_{r n}\right) \cdot \lambda_{r k n} & \forall r, k, n^{\prime} \\
\lambda_{r 0 n}=\lambda_{r n}^{0} & \forall r, n \\
\mathbf{u}_{r n} \in \mathcal{U}_{r}\left(\mathbf{x}_{r n}\right) & \forall r, n \\
\lambda_{r k n} \geq 0 & \forall r, k, n .
\end{array}\right. \tag{5.1b}
\end{align*}
$$

Since in (5.1) the budget constraints (5.1b) are the only explicit links between the road classes, Lagrangean relaxation of (5.1b) is a natural means for the determination of the residual values. The first steady-state assumption means that one and the same Lagrangean multiplier value $v_{k}=v_{\infty}$ applies each year $k$. As in Ch 4, the Lagrangean dual separates into a number of road class specific network subproblems and a trivial budget subproblem. Letting $c$ denote the 1year sum of traffic and relaxed maintenance costs in (5.1), i.e.

$$
c_{r n}\left(\mathbf{u}_{r n}\right):=f_{r n}+v_{\infty} \cdot \sqrt{d} \cdot g_{r n}\left(\mathbf{u}_{r n}\right)
$$

the road class $r$-subproblem runs

$$
\underset{\left(\lambda_{r k n}\right)_{k, n}\left(\mathbf{u}_{r n}\right)_{n}}{\operatorname{minimise}} \sum_{n} \sum_{k=0}^{\infty} d^{k} \cdot c_{r n}\left(\mathbf{u}_{r n}\right) \cdot \lambda_{r k n},\left\{\begin{array}{cc}
\lambda_{r k n^{\prime}}=\sum_{n} p_{r n n^{\prime}}\left(\mathbf{u}_{r n}\right) \cdot \lambda_{r, k-1, n} & \forall k \geq 1, n^{\prime} \\
\lambda_{r 0 n^{\prime}}=\lambda_{r n^{\prime}}^{0} & \forall n^{\prime}  \tag{5.2b}\\
\mathbf{u}_{r n} \in \mathcal{U}_{r}\left(\mathbf{x}_{r n}\right) & \forall n \\
\lambda_{r k n} \geq 0 & \forall k, n .
\end{array}\right.
$$

By the optimal residual values $\varphi_{r}=\left(\varphi_{r n}\right)_{n}$ for road class $r$ we mean the optimal future costs from the nodal states at time $\kappa$ (in general from any fixed (finite) time $\geq \kappa$ ) in (5.2).

Since $\mathbf{P}_{r}\left(\mathbf{U}_{r}\right) \equiv\left(p_{r n n^{\prime}}\left(\mathbf{u}_{r n}\right)\right)_{n, n^{\prime}}$ is a probability matrix, i.e. nonnegative with all row sums equal to 1, the Frobenius-Perron theorem, see e.g. [Luenberger (1979), p 193], guarantees that the inverse $\left[\mathbf{I}-d \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right]^{-1}$ exists for every discount factor value $d \in[0,1[$.

### 5.2 Models

Lemma 5.1: Given $\mathbf{U}_{r}$ and $d<1$, the residual values in (5.2) are

$$
\widetilde{\boldsymbol{\varphi}}_{r}\left(\mathbf{U}_{r}\right):=\left[\mathbf{I}-d \cdot \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right]^{-1} \mathbf{c}_{r}\left(\mathbf{U}_{r}\right) .
$$

Proof: Let $\mathbf{c} \equiv \mathbf{c}_{r}\left(\mathbf{U}_{r}\right)$ and $\mathbf{P} \equiv \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)$. The residual values at time $\tau$ reflect the future costs emanating from $\mathbf{c}$. The worst possible case for any state is to get $c_{\max }:=\max _{n} c_{r n}\left(\mathbf{u}_{r n}\right)$ with probability 1 in every succeeding time period, i.e.

$$
\widetilde{\varphi}_{r m n}\left(\mathbf{U}_{r}\right) \leq c_{\max }+d \cdot c_{\max }+d^{2} \cdot c_{\max }+\cdots=\frac{c_{\max }}{1-d}
$$

The same upper bound applies to every node $n$ and every succeeding time period $\kappa+k, k>0$, i.e. $\left(\widetilde{\boldsymbol{\varphi}}_{r, \kappa+k}\left(\mathbf{U}_{r}\right)\right)_{k}$ are uniformly bounded, as $k \uparrow \infty$. The residual values $\widetilde{\boldsymbol{\varphi}}_{r \kappa}\left(\mathbf{U}_{r}\right)$ at time $\kappa$ are transferred to the succeeding years through

$$
\begin{aligned}
& \widetilde{\boldsymbol{\varphi}}_{r k}\left(\mathbf{U}_{r}\right)=\mathbf{c}+d \cdot \mathbf{P} \widetilde{\boldsymbol{\varphi}}_{r, \kappa+1}\left(\mathbf{U}_{r}\right)=\mathbf{c}+d \cdot \mathbf{P}\left(\mathbf{c}+d \cdot \mathbf{P} \widetilde{\boldsymbol{\varphi}}_{r, K+2}\left(\mathbf{U}_{r}\right)\right)=\cdots= \\
& =\left(\mathbf{I}+d \cdot \mathbf{P}+d^{2} \cdot \mathbf{P}^{2}+\cdots+d^{k-1} \cdot \mathbf{P}^{k-1}\right) \mathbf{c}+d^{k} \cdot \mathbf{P}^{k} \widetilde{\boldsymbol{\varphi}}_{r, \kappa+k}\left(\mathbf{U}_{r}\right)= \\
& =\left((\mathbf{I}-d \cdot \mathbf{P})^{-1}-d^{k} \cdot \mathbf{P}^{k}(\mathbf{I}-d \cdot \mathbf{P})^{-1}\right) \mathbf{c}+d^{k} \cdot \mathbf{P}^{t} \widetilde{\boldsymbol{\varphi}}_{r, \kappa+k}\left(\mathbf{U}_{r}\right)= \\
& =\left(\mathbf{I}-d^{k} \cdot \mathbf{P}^{k}\right)(\mathbf{I}-d \cdot \mathbf{P})^{-1} \mathbf{c}+d^{k} \cdot \mathbf{P}^{k} \widetilde{\boldsymbol{\varphi}}_{r, \kappa+k}\left(\mathbf{U}_{r}\right) .
\end{aligned}
$$

Hence

$$
\left\|\widetilde{\boldsymbol{\varphi}}_{r \kappa}\left(\mathbf{U}_{r}\right)-(\mathbf{I}-d \cdot \mathbf{P})^{-1} \mathbf{c}\right\| \leq d^{k} \cdot\left(\left\|\mathbf{P}^{k}\right\| \cdot\left\|\widetilde{\boldsymbol{\varphi}}_{r, \kappa+k}\left(\mathbf{U}_{r}\right)\right\|+\left\|\mathbf{P}^{k}\right\| \cdot\left\|(\mathbf{I}-d \cdot \mathbf{P})^{-1} \mathbf{c}\right\|\right) .
$$

Here $\mathbf{P}$ is a stochastic matrix with $\left\|\mathbf{P}^{k}\right\|_{1} \leq\|\mathbf{P}\|_{1}^{k}=1,\left\|\widetilde{\boldsymbol{\varphi}}_{r, \kappa+k}\left(\mathbf{U}_{r}\right)\right\|$ is bounded and $d^{k} \downarrow 0$ as $k \uparrow \infty$. The result is unchanged for every finite initial time $\geq \kappa$. Hence the lemma follows. For a given $\mathbf{U}_{r}$ the length distributions in (5.2) are determined by the state transition constraints (5.2b), in matrix $k$-power notation $\lambda_{r k}=\mathbf{P}_{r}\left(\mathbf{U}_{r}\right)^{T} \boldsymbol{\lambda}_{r, k-1}=\cdots=\mathbf{P}_{r}{ }^{k}\left(\mathbf{U}_{r}\right)^{T} \lambda_{r}^{0}$.

Lemma 5.2: For any given $\mathbf{U}_{r}$, with $\lambda_{r k}=\mathbf{P}_{r}{ }^{k}\left(\mathbf{U}_{r}\right)^{T} \lambda_{r}^{0}$, the objective (5.2a) can be written

$$
\sum_{n} \sum_{k=0}^{\infty} d^{k} c_{r n}\left(\mathbf{U}_{r}\right) \lambda_{r k n}=\widetilde{\varphi}_{r}\left(\mathbf{U}_{r}\right)^{T} \lambda_{r}^{0}
$$

Proof: By use of Lemma 5.1 and letting $\mathbf{c} \equiv \mathbf{c}_{r}\left(\mathbf{U}_{r}\right), \mathbf{P} \equiv \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)$ we get

$$
\begin{aligned}
& \sum_{n} \sum_{k=0}^{\infty} d^{k} c_{r n}\left(\mathbf{U}_{r}\right) \lambda_{r k n}=\mathbf{c}^{T} \boldsymbol{\lambda}_{r 0}+d \cdot \mathbf{c}^{T} \boldsymbol{\lambda}_{r 1}+d^{2} \cdot \mathbf{c}^{T} \boldsymbol{\lambda}_{r 2}+\cdots= \\
& =\mathbf{c}^{T}\left(\lambda_{r 0}+d \cdot \mathbf{P}^{T} \boldsymbol{\lambda}_{r 0}+d^{2} \cdot\left(\mathbf{P}^{T}\right)^{2} \boldsymbol{\lambda}_{r 0}+\cdots\right)=\mathbf{c}^{T}\left(\mathbf{I}-d \mathbf{P}^{T}\right)^{-1} \boldsymbol{\lambda}_{r 0} \equiv \\
& \equiv\left[\left(\mathbf{I}-d \cdot \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right)^{-1} \mathbf{c}_{r}\left(\mathbf{U}_{r}\right)\right]^{T} \lambda_{r}^{0}=\widetilde{\boldsymbol{\varphi}}_{r}\left(\mathbf{U}_{r}\right)^{T} \boldsymbol{\lambda}_{r}^{0} .
\end{aligned}
$$

Lemma 5.2 shows that the $\lambda_{r n}^{0}$-coefficient in the objective sums up the future costs per metre, in nodal state $n$ of decision rule $\mathbf{U}_{r}$ (at time $\kappa$ ), i.e. for an optimum decision the coefficient $\widetilde{\varphi}_{r n}\left(\mathbf{U}_{r}\right)$ coincides with the optimal residual value.

From Lemmas 5.1-5.2 we realise that (5.2) can be expressed in terms of residual values instead of the length distributions, as

$$
\begin{align*}
& \underset{\mathbf{U}_{r}}{\operatorname{minimise}}  \tag{5.3a}\\
& \text { subject to }  \tag{5.3b}\\
& \widetilde{\varphi}_{r}\left(\mathbf{U}_{r}\right)^{T} \lambda_{r}^{0} \\
& \left\{\begin{array}{r}
\widetilde{\varphi}_{r}\left(\mathbf{U}_{r}\right)=\mathbf{c}_{r}\left(\mathbf{U}_{r}\right)+d \cdot \mathbf{P}_{r}\left(\mathbf{U}_{r}\right) \widetilde{\varphi}_{r}\left(\mathbf{U}_{r}\right) \\
\mathbf{U}_{r} \in \prod_{n} \mathcal{U}_{r}\left(\mathbf{x}_{r n}\right)
\end{array}\right.
\end{align*}
$$

With the result of Lemma 5.1 written like (5.3b) it is logical that the future costs at two consecutive times coincide.

In the objective (5.3a) only the initial lengths $\lambda_{r}^{0}$ (at time $\tau$ ) matter. The question is: do they? Above we claimed that, due to our assumption of a stationary decision rule, the residual values do not depend on the initial lengths. This is verified by Cor 5.1 below.

Definition 5.1: A decision rule $\mathbf{U}_{r}^{*}$ is statewise optimal if $\widetilde{\varphi}_{r n}\left(\mathbf{U}_{r}^{*}\right)=\min _{\mathbf{U}_{r}} \widetilde{\varphi}_{r n}\left(\mathbf{U}_{r}\right)$ for each $n$. For general compact sets $\mathcal{U}_{r}$ and for (lower) semicontinuous functions $\mathbf{c}_{r}\left(\mathbf{U}_{r}\right), \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)$, the use of the Banach fixed-point theorem, e.g. [Puterman (2004), p 154], shows the existence of a statewise optimal deterministic decision rule (in the general class of randomised $=$ stochastic rules) satisfying (5.4a) - (5.4b) below, which are the optimality conditions in discounted Markov decision problems (MDP's). We formulate it, without proof, in a weaker sense, sufficient for our purposes.

Proposition 5.1: In any road class $r$ assume that $\mathbf{c}_{r}\left(\mathbf{U}_{r}\right), \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)$ are continuous functions of $\mathbf{U}_{r}$ and that $\prod_{n} \mathcal{U}_{r}\left(\mathbf{x}_{r n}\right)$ is a compact set. A decision rule $\mathbf{U}_{r}^{*} \equiv\left(\mathbf{u}_{r n}^{*}\right)_{n}$ exists, solving

$$
\begin{equation*}
\mathbf{u}_{r n}^{*} \in \underset{\mathbf{U}_{r}}{\arg \min } \widetilde{\varphi}_{r n}\left(\mathbf{U}_{r}\right) \quad \forall n \tag{5.4a}
\end{equation*}
$$

where

$$
\begin{equation*}
\widetilde{\boldsymbol{\varphi}}_{r}\left(\mathbf{U}_{r}\right):=\left[\mathbf{I}-d \cdot \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right]^{-1} \mathbf{c}_{r}\left(\mathbf{U}_{r}\right) . \square \tag{5.4b}
\end{equation*}
$$

Corollary 5.1: $\mathbf{U}_{r}^{*}$ in Prop 5.1 solves (5.3) (and (5.2)).
Proof: Since $\lambda_{r}^{0} \geq 0$ we have

$$
\min _{\mathbf{U}_{r}} \widetilde{\boldsymbol{\varphi}}_{r}\left(\mathbf{U}_{r}\right)^{T} \lambda_{r}^{0} \geq \sum_{n} \min _{u_{r n}} \widetilde{\boldsymbol{\varphi}}_{r n}\left(\mathbf{u}_{r n}\right) \cdot \lambda_{r n}^{0}=\sum_{n} \widetilde{\boldsymbol{\varphi}}_{r n}\left(\mathbf{u}_{r n}^{*}\right) \cdot \lambda_{r n}^{0}=\widetilde{\boldsymbol{\varphi}}_{r}\left(\mathbf{U}_{r}^{*}\right)^{T} \lambda_{r}^{0}
$$

i.e. $\mathbf{U}_{r}^{*}$ is an optimum solution of (5.3) (and $\mathbf{U}_{r}^{*}, \boldsymbol{\lambda}_{r k}=\mathbf{P}_{r}{ }^{k}\left(\mathbf{U}_{r}^{*}\right)^{T} \boldsymbol{\lambda}_{r}^{0}$ of (5.2)).

### 5.2.2 Length distributions model

Model (5.1) is inconsequent insofar as we have assumed an implicit constant dual price $v_{\infty}$ and a constant decision rule $\left(\mathbf{U}_{r}\right)_{r}$, whereas the utilizable budget $y_{k}$ and the length distributions $\left(\boldsymbol{\lambda}_{r k}\right)_{r}$ may vary by $k$. An alternative is to assume $K \gg 0$, i.e. that $\left(\mathbf{U}_{r}\right)_{r}$ is applied many times prior to the terminal time, and that $\left(\lambda_{r K}\right)_{r}$ approximate limit distributions.

### 5.2 Models

By the use of a Cesaro limit, e.g. [Puterman (2004), p 591], the limit transition matrix

$$
\mathbf{P}_{r}^{\infty}\left(\mathbf{U}_{r}\right):=\operatorname{Clim}_{K \rightarrow \infty} \mathbf{P}_{r}^{K}\left(\mathbf{U}_{r}\right) \equiv \lim _{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{P}_{r}^{k}\left(\mathbf{U}_{r}\right)
$$

becomes well-defined, coincident with $\lim _{K \rightarrow \infty} \mathbf{P}_{r}{ }^{K}\left(\mathbf{U}_{r}\right)$ whenever the ordinary limit exists, i.e. for aperiodic (regular) MCs - in which case $\mathbf{P}_{r}{ }^{k}\left(\mathbf{U}_{r}\right)$ is positive for some $k>0$, see e.g. [Berman and Plemmons (1979), p 219]. The corresponding limit length distribution is

$$
\lambda_{r \infty}:=\mathbf{P}_{r}^{\infty}\left(\mathbf{U}_{r}\right)^{T} \lambda_{r}^{0}
$$

In practice we expect that the ordinary limit exists, providing

$$
\lambda_{r \infty}=\mathbf{P}_{r}\left(\mathbf{U}_{r}\right)^{T} \lambda_{r \infty}
$$

i.e. $\boldsymbol{\lambda}_{r \infty}$ is an eigenvector corresponding to the maximum eigenvalue 1 of $\mathbf{P}_{r}{ }^{T}$.

As independent of the initial length distribution, this $\lambda_{r \infty}$ is the key to our second, steady-state model, used for approximating the length distribution at time $K$. It corresponds to an additional (heuristic) assumption about the distance from transition time $\kappa$, written with the index-value $\infty$ implicit:

- time $K \gg 0$ occurs in a stabilised pavement state situation, meaning a stationary length distribution $\boldsymbol{\Lambda}=\left(\boldsymbol{\lambda}_{r}\right)_{r}=\left(\lambda_{r n}\right)_{r, n}$ in every succeeding year.

With this additional assumption and with the given initial length distributions (5.1e) replaced by total length conditions plus non-negativity, the semi-stationary model (5.1) can be simplified to a full steady-state model (5.5) below. In the semi-stationary model the total length conditions are redundant, since they follow from the initial length distributions (5.1e) and the flow balance conditions (5.1d). In the steady-state model we need just one utilizable-budget variable $y_{\infty}$ and can evaluate the sums over $k$ (arbitrarily from time $K$ or from $k=0$ ) in the objective (5.1a), to have

$$
\begin{align*}
& \underset{\left(\lambda_{r n}\right)_{r, n}\left(\mathbf{u}_{r n}\right)_{r, n}}{\operatorname{minimise}} \frac{1}{1-d} \cdot\left(\sum_{r} \sum_{n} f_{r n} \lambda_{r n}+v_{\mathrm{BCR}} \sqrt{d} \cdot y_{\infty}\right)  \tag{5.5a}\\
& \text { subject to }\left\{\begin{array}{cc}
\sum_{r} \sum_{n} g_{r n}\left(\mathbf{u}_{r n}\right) \cdot \lambda_{r n} \leq y_{\infty} \\
y_{\infty} \leq b_{\infty} & \\
\lambda_{r n^{\prime}}=\sum_{n} p_{r n n^{\prime}}\left(\mathbf{u}_{r n}\right) \cdot \lambda_{r n} & \forall r, n^{\prime} \\
\sum_{n}^{n} \lambda_{r n}=L e n_{r} & \forall r \\
\mathbf{u}_{r n} \in \mathcal{U}_{r}\left(\mathbf{x}_{r n}\right) & \forall r, n \\
\lambda_{r n} \geq 0 & \forall r, n .
\end{array}\right. \tag{5.5b}
\end{align*}
$$

In (5.5), the equations (5.5d) are linearly dependent:

$$
\sum_{n^{\prime}}\left(\lambda_{r n^{\prime}}-\sum_{n} p_{r n n^{\prime}} \cdot \lambda_{r n}\right)=\sum_{n^{\prime}} \lambda_{r n^{\prime}}-\sum_{n} \lambda_{r n} \cdot \sum_{n^{\prime}} p_{r n n^{\prime}}=0
$$

Model (5.5) is nonlinear. By discretizing the works extent, layer thickness, for the works type major maintenance, the 2D pavement works variables $\mathbf{u}_{r n} \in \mathcal{U}_{r}\left(\mathbf{x}_{r n}\right)$ turn into a finite set of 1D maintenance options $m=u_{r n} \in \mathcal{M}_{r n}$ (where routine maintenance is included). By adding index $m$ to the maintenance costs $g_{r n m} \equiv g_{r n}(m)$, the transition probabilities $p_{r n n^{\prime} m} \equiv p_{r n n^{\prime}}(m)$ and the length variables $\lambda_{r n m}\left(\equiv \lambda_{r \infty n m}\right)$, letting the time index $\infty$ be implicit, we get the linearized steady-state model:

$$
\begin{align*}
& \underset{\left.\left(\lambda_{r n}\right)_{r, n}, \mathbf{u}_{r n}\right)_{r, n}}{\operatorname{minimise}} \frac{1}{1-d} \cdot\left(\sum_{r} \sum_{n} \sum_{m \in \mathcal{M}_{r n}} f_{r n} \lambda_{r n m}+v_{\mathrm{BCR}} \sqrt{d} \cdot y_{\infty}\right)  \tag{5.5a'}\\
& \text { subject to }\left\{\begin{array}{cc}
\sum_{r} \sum_{n} \sum_{m \in \mathcal{M}_{r n}} g_{r n m} \cdot \lambda_{r n m} \leq y_{\infty} \\
y_{m} \leq b_{\infty} \\
\sum_{m \in \mathcal{M}_{r n^{\prime}}} \lambda_{r n^{\prime} m}=\sum_{n} \sum_{m \in \mathcal{M}_{r n}} p_{r n n^{\prime} m} \cdot \lambda_{r n m} & \forall r, n^{\prime} \\
\sum_{n} \sum_{m \in \mathcal{M}_{r n}} \lambda_{r n m}=\text { Len }_{r} & \forall r \\
\lambda_{r n m} \geq 0 & \forall r, n, m .
\end{array}\right.
\end{align*}
$$

Using the presumed dual price $v_{\infty}$ for a relaxation of ( $5.5 b^{\prime}$ ), the dual subproblem separates into road class $r$ specific subproblems and a trivial budget subproblem. Letting $c_{r n m}:=f_{r n}+v_{\infty} \sqrt{d} \cdot g_{r n m}, \widetilde{c}_{r n m}:=\frac{1}{1-d} c_{r n m}$ the dual $r$-subproblem becomes

$$
\begin{array}{ll}
\underset{\left(\lambda_{r n m}\right)_{n, m}}{\operatorname{minimise}} & \sum_{n} \sum_{m \in \mathcal{M}_{r n}} \widetilde{c}_{r n m} \lambda_{r n m} \\
\text { subject to } & \left\{\begin{array}{cc}
\sum_{m \in \mathcal{M}_{r n}} \lambda_{r n^{\prime} m}=\sum_{n} \sum_{m \in \mathcal{M}_{r n}} p_{r n n^{\prime} m} \lambda_{r n m} & \forall n^{\prime} \\
\sum_{n} \sum_{m \in \mathcal{M}_{r n}} \lambda_{r n m}=\text { Len }_{r} \\
\lambda_{r n m} \geq 0 & \forall n, m \in \mathcal{M}_{r n}
\end{array}\right. \tag{5.6b}
\end{array}
$$

Mathematically model (5.6) is an example of an average cost MDP, see [Puterman (2004), $\mathrm{p} 391]$. If $\lambda_{r n m}$ is in metres we identify its coefficient $\widetilde{c}_{r n m}$ in (5.6a) as the residual value per metre, according to this model. Due to the model logics (and the simplicity) it is conceivable to use (5.6) instead of (5.2) for the residual values computation.
Viewing (5.6) as an average cost LP-primal, introducing LP-dual variables $\mathbf{v}=\left(\mathrm{v}_{n}\right)_{n}$ and $w$ for the flow balance constraints (5.6b) and the length condition (5.6c), respectively, the corresponding LP-dual, cf. [ibid., p 391], becomes a $w$-maximin problem

$$
\begin{array}{ll}
\underset{\mathbf{v}, w}{\operatorname{maximise}} & \text { Len }_{r} \cdot w \\
\text { subject to } & w+\mathrm{v}_{n}-\sum_{n^{\prime}} p_{r n n^{\prime} m} \mathrm{v}_{n^{\prime}} \leq \widetilde{c}_{r n m} \quad \forall n, m \in \mathcal{M}_{r n} . \tag{5.7b}
\end{array}
$$

Notice the resemblance between (5.7b) and (5.4) for the residual values $\left(\varphi_{r n}\right)_{n}$. Introducing $\mathbf{R}_{r}$ the optimality conditions (5.4) can be written

$$
\mathbf{R}_{r}\left(\mathbf{U}_{r}\right):=\left[\mathbf{I}-d \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right] \boldsymbol{\varphi}_{r}-\mathbf{c}_{r}\left(\mathbf{U}_{r}\right) \leq \mathbf{0} \quad \forall \mathbf{U}_{r} \in \prod_{n} \mathcal{U}_{r}\left(\mathbf{x}_{r n}\right),
$$

### 5.2 Models

where equality holds for some $\mathbf{U}_{r}=\mathbf{U}_{r}{ }^{*}$ according to (5.4a). For a linearization of $\left(\mathbf{u}_{r n}\right)_{n}$ in (5.4) - (5.4'), corresponding to (5.5'), we have

Proposition 5.2: The optimal residual values $\left(\varphi_{r n}\right)_{n}$ of (5.4) provide a lower bound $\operatorname{Len}_{r} \cdot \min _{n} \varphi_{r n}$ for the optimal value of the LP-dual (5.7) (and primal (5.6)).

Proof: Take $\mathrm{v}_{n}:=\frac{d}{1-d} \varphi_{r n}+C \quad \forall n, C$ arbitrary, and $w:=\min _{n} \varphi_{r n}$ in the LP-dual (5.7).
Since the optimal $\boldsymbol{\varphi}_{r}=\left(\varphi_{r n}\right)_{n}$ for any (linearized) $\mathbf{U}_{r}$ satisfies (5.4b') and since $\mathbf{P}_{r}\left(\mathbf{U}_{r}\right) \mathbf{1}=\mathbf{1}$,

$$
\begin{aligned}
& {\left[\mathbf{I}-\mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right] \mathbf{v}+w \mathbf{1}-\widetilde{\mathbf{c}}_{r}\left(\mathbf{U}_{r}\right)=} \\
& =\frac{d}{1-d}\left(\mathbf{I}-\mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right) \boldsymbol{\varphi}_{r}+C \cdot\left(\mathbf{I}-\mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right) \mathbf{1}+\min _{n} \varphi_{r n} \cdot \mathbf{1}-\frac{1}{1-d} \mathbf{c}_{r}\left(\mathbf{U}_{r}\right)= \\
& =\frac{1}{1-d}\left(\left[\mathbf{I}-d \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right] \boldsymbol{\varphi}_{r}-(1-d) \cdot \boldsymbol{\varphi}_{r}\right)+\min _{n} \varphi_{r n} \cdot \mathbf{1}-\frac{1}{1-d} \mathbf{c}_{r}\left(\mathbf{U}_{r}\right)= \\
& =\frac{1}{1-d}\left(\left\{\left[\mathbf{I}-d \mathbf{P}_{r}\left(\mathbf{U}_{r}\right)\right] \boldsymbol{\varphi}_{r}-\mathbf{c}_{r}\left(\mathbf{U}_{r}\right)\right\}-(1-d) \cdot\left(\boldsymbol{\varphi}_{r}-\min _{n} \varphi_{r n} \cdot \mathbf{1}\right)\right)= \\
& =\frac{1}{1-d}\left(\mathbf{R}_{r}\left(\mathbf{U}_{r}\right)-(1-d) \cdot\left(\boldsymbol{\varphi}_{r}-\min _{n} \varphi_{r n} \cdot \mathbf{1}\right)\right) \leq \frac{1}{1-d}\left(\mathbf{R}_{r}\left(\mathbf{U}_{r}\right)-\mathbf{0}\right) \leq \mathbf{0},
\end{aligned}
$$

i.e. a dually feasible solution with a lower objective bound $\operatorname{Len}_{r} \cdot w=\operatorname{Len}_{r} \cdot \min _{n} \varphi_{r n}$.

Definition 5.2: At node $n$ in road class $r$ the set of works $m$-options that by certainty lead to the very same (absorbing) state $n$ is $\mathcal{M}_{r n}^{\text {absorb }}:=\left\{m \in \mathcal{M}_{r n}: p_{r n n m}=1\right\}$.

As can be seen from the dual constraints (5.7b) an obvious upper objective bound is received from the possible absorbing states $n$, as $\bar{w}:=\min _{n, m \in \mathcal{M}_{r n}^{\text {absorb }}} \widetilde{c}_{r n m}$.

In the opposite model direction we have:
Proposition 5.3: If $(\mathbf{v}, w)$ is feasible in (5.7), then

$$
\varphi_{r n}:=\frac{1-d}{d}\left(\mathrm{v}_{n}-\max _{n} \mathrm{v}_{n}\right)+w \quad \forall n
$$

satisfies (5.4b').
Proof: (5.7b) corresponds to

$$
(1-d) w+d \varphi_{r n} \leq c_{r n m}+d \cdot \sum_{n^{\prime}} p_{r n n^{\prime} m} \varphi_{r n^{\prime}} \quad \forall n, m \in \mathcal{M}_{r n}
$$

Since

$$
\varphi_{r n}=\frac{1-d}{d}\left(\mathrm{v}_{n}-\max _{n} \mathrm{v}_{n}\right)+w \leq 0+w \quad \forall n
$$

the previous LHS satisfies

$$
(1-d) w+d \varphi_{r n} \geq(1-d) \cdot \max _{n} \varphi_{r n}+d \cdot \varphi_{r n} \geq \varphi_{r n}
$$

Hence $\left(\varphi_{r n}\right)_{n}$ satisfies (5.4b').

Whereas Prop 5.2 assumes $w=\min _{n} \varphi_{r n}$, Prop 5.3 is about $w=\max _{n} \varphi_{r n}$. For such choices the normal case is that one node $n$ (i.e. one $\varphi_{r n}$ ) provides the extreme $w$-value, implying that one constraint in (5.7b) and (5.4b'), respectively, is satisfied with equality. However, for the optimum we expect at least one binding ( $n, m$ )-constraint per node $n$ - which tells us that an optimum solution of (5.7) or (5.4) does not normally creates an optimum for the other problem. Anyhow, it may serve as a feasible, initial solution.

### 5.2.3 Markov chains

Consider (5.1) and (5.5) - (5.5') for a given decision rule $\hat{\mathbf{U}}_{r}$, and corresponding probabilities $\hat{\mathbf{P}}_{r} \equiv \mathbf{P}_{r}\left(\hat{\mathbf{U}}_{r}\right)$. The matrix $\hat{\mathbf{P}}_{r}$ is stochastic, i.e. each row sum equals 1, and $\mathbf{I}-\hat{\mathbf{P}}_{r}$ is singular. By a re-indexing of the states, the MC can be put on the canonical form $\hat{\mathbf{P}}_{r}=\left(\begin{array}{ll}\mathbf{R} & \mathbf{O} \\ \mathbf{S} & \mathbf{Q}\end{array}\right)$, where $\mathbf{O}$ is a zero matrix, $\mathbf{R}$ is block-diagonal with a block $\mathbf{R}^{c}=\hat{\mathbf{P}}_{r}^{c}$ for each closed (absorbing), communicating (=comm.) class $c$ of states, the quadratic matrix $\mathbf{Q}$ corresponds to the transient comm.classes and $\mathbf{S}$ contains nonzero elements. Since $\mathbf{Q}$ is a substochastic matrix, i.e. a stochastic matrix with at least one row sum less than 1.0, the fundamental matrix $(\mathbf{I}-\mathbf{Q})^{-1}$ is well-defined - and positive, according to [Luenberger (1979), p 240]. This positivity claim (as well as [ibid, Thm 2, p 198] is wrong, as the following (counter-)example shows.

Example 5.1: Let $\hat{\mathbf{P}}:=\left(\begin{array}{c|cc}1 & 0 & 0 \\ \hline 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5\end{array}\right)$. We identify $\mathbf{Q} \equiv\left(\begin{array}{cc}0.5 & 0 \\ 0 & 0.5\end{array}\right)$ and get $\mathbf{I}-\mathbf{Q} \equiv\left(\begin{array}{cc}0.5 & 0 \\ 0 & 0.5\end{array}\right),(\mathbf{I}-\mathbf{Q})^{-1} \equiv\left(\begin{array}{ll}2 & 0 \\ 0 & 2\end{array}\right)$, i.e. the inverse is nonnegative but not positive.

Lemma 5.3: The fundamental matrix $(\mathbf{I}-\mathbf{Q})^{-1}$ exists, nonnegative.
Proof: For any fixed $t>0$ let $\mathbf{B}_{k}:=\mathbf{I}+\mathbf{Q}+\cdots+\mathbf{Q}^{k-1}$, whence $(\mathbf{I}-\mathbf{Q}) \mathbf{B}_{k}:=\mathbf{I}-\mathbf{Q}^{k}$.
According to [ibid, p 238 ] the probability of remaining in the transient set of states goes to zero with the number of time steps, i.e. $\mathbf{Q}^{k} \rightarrow \mathbf{O}$ as $t \uparrow \infty$. Hence $(\mathbf{I}-\mathbf{Q}) \mathbf{B}_{k} \rightarrow \mathbf{I}$, i.e. $\mathbf{B}_{k} \rightarrow(\mathbf{I}-\mathbf{Q})^{-1}$.
In the series expansion, $\mathbf{Q} \geq \mathbf{O}$ implies $(\mathbf{I}-\mathbf{Q})^{-1} \geq \mathbf{O}$.
Since the nondiagonal elements of $\mathbf{I}-\mathbf{Q}$ are nonpositive and $(\mathbf{I}-\mathbf{Q})^{-1} \geq \mathbf{O}, \mathbf{I}-\mathbf{Q}$ is an $M$ matrix, see [Ortega and Rheinboldt (1970), p 54].

Lemma 5.4: Form $\hat{\mathbf{P}}_{r+}^{c}$ from $\hat{\mathbf{P}}_{r}^{c}$ by excluding the row and column of some node in the closed comm.class $c$. $\hat{\mathbf{P}}_{r+}^{c}$ is substochastic, $\left(\hat{\mathbf{P}}_{r+}^{c}\right)^{t} \rightarrow \mathbf{O}$ as $t \uparrow \infty$ and $\left(\mathbf{I}-\hat{\mathbf{P}}_{r+}^{c}\right)^{-1}$ exists, nonnegative.

### 5.2 Models

Proof: Since $c$ is a closed comm.class the remaining nodes in $\hat{\mathbf{P}}_{r+}^{c}$ communicate with the excluded node, i.e. at least one row sum in $\hat{\mathbf{P}}_{r+}^{c}$ is less than 1. Thus $\hat{\mathbf{P}}_{r+}^{c}$ is substochastic - and behaves like the transient matrix $\mathbf{Q}$ in Lemma 5.3.

Irreducible (=ergodic) MCs is the special case when all states communicate in a single class, and corresponds to an irreducible matrix $\hat{\mathbf{P}}_{r}$. Although we expect irreducible MCs, or MCs with exactly one closed comm.class and a transient part, to be the normal case in our applications, we will comment on the degeneracy cases below.

The question of reducible MCs is linked to the road length distributions. In (4.7) these are given per road class at time 0. Apart from our assumption in Sec 5.2.2 that ordinary limit matrices $\left(\mathbf{P}_{r}^{\infty}\left(\mathbf{U}_{r}\right)\right)_{r}$ and length distributions $\left(\boldsymbol{\lambda}_{r \infty}\right)_{r}$ exist, we assume that it is possible, during the initial dynamic planning period of $K$ years, to reach every node from every other node, i.e. possible to control the evolution (e.g., by the residual prices) to the terminal states we want, those leading to minimum total cost in (5.5). The general distribution freedom is between the closed classes, not within them. See Sec 5.3.6 below.

### 5.3 Method

The methods we apply to the dual subproblems (5.2) and (5.6) are almost identical, both involving two steps:

- Determine an optimal works solution $\mathbf{U}^{*}$ and optimal residual values $\varphi$, either as a discounted MDP, by solving (5.4), or as an average cost MDP, by solving (5.7) and identifying $\varphi=\widetilde{\mathbf{c}}$ in (5.6a) for $\mathbf{U}^{*}$.
- Determine optimal length distributions $\lambda$ for $\mathbf{U}^{*}$, either as an average cost MDP, by solving (5.6), or as a discounted MDP, by computing a Cesaro or ordinary limit based $\lambda_{r \infty}=\mathbf{P}_{r}^{\infty}\left(\mathbf{U}_{r}^{*}\right)^{T} \lambda_{r}^{0}$ from the initial length distribution $\lambda_{r n}^{0}=\operatorname{Len}_{r} /|\mathcal{N}| \quad \forall n$.

Below we describe the method proposals for the average cost MDP (5.6) and comment on the few differences vs. the discounted MDP. The classical methods are the value and policy iteration methods, see [Howard (1960)]. A lot of variants have been proposed. We concentrate on the Gauss-Seidel variants of the value iteration method - cf. [Bertsekas (1995)] for the average cost and [Puterman (2004), p 166], for the discounted cost MDPs. For the linearized versions of (5.2) and (5.6) we also apply LP and for all the models a Newton type method is developed. Both these methods can be viewed as policy iteration in a more general setting. In the dual network $r$-subproblem below the road class index $r$ is implicit.

### 5.3.1 Dual optimisation

In the Lagrangean relaxations of both (5.1b) and (5.2b) the first steady-state assumption means that one and the same Lagrangean multiplier value $v \equiv v_{\infty}$ applies each year. As in Ch 4 the
dual subproblem separates into a network subproblem and a budget subproblem. The latter has the trivial solution $y_{\infty}=b_{\infty}$ for every $v_{\infty} \geq v_{\mathrm{BCR}}$ (and $y_{\infty} \downarrow-\infty$ for $v_{\infty}<v_{\mathrm{BCR}}$; which therefore is neglected in the dual) and the contribution to the dual objective is
$\frac{\sqrt{d}}{1-d} b_{\infty}\left(v_{\mathrm{BCR}}-v_{\infty}\right)$ in both relaxations. We use a common tensor notation $\boldsymbol{\Lambda}$ for the respective collection of length variables, in the residual values model $\boldsymbol{\Lambda}:=\left(\lambda_{r k n}\right)_{r, k, n}$ and in the length distributions model $\boldsymbol{\Lambda}:=\left(\lambda_{r n}\right)_{r, n}$ (or, linearized, $\boldsymbol{\Lambda}:=\left(\lambda_{r n m}\right)_{r, n, m}$ ), and let $\boldsymbol{\Lambda}^{*}(v)$ hold the optimal road lengths for a given $v \equiv v_{\infty}$ in the network subproblem. Moreover, we use common notations for the respective total discounted traffic cost and maintenance cost-budget difference, e.g. in the residual values model

$$
F(\mathbf{\Lambda}):=\sum_{k} d^{k} \cdot \sum_{r} \sum_{n} \sum_{m \in \mathcal{M}_{r n}} f_{r n} \cdot \lambda_{r k n} \text { and } G(\mathbf{\Lambda}):=\sum_{k} d^{k+\frac{1}{2}} \cdot\left(\sum_{r} \sum_{n} g_{r n}\left(\mathbf{u}_{r n}\right) \cdot \lambda_{r k n}-b_{\infty}\right)
$$

where $\mathbf{U}=\left(\mathbf{u}_{r n}\right)_{r, n}$ solves the dual subproblem.
The common form of the Lagrangean-dual problems becomes

$$
\underset{v \geq v_{\mathrm{BCR}}}{\operatorname{maximise}} \Phi(v):=F\left(\mathbf{\Lambda}^{*}(v)\right)+v \cdot G\left(\boldsymbol{\Lambda}^{*}(v)\right)
$$

The updating of the only multiplier $v$ can be made in several ways - see Fig 5.1. One possibility is to use the Dantzig-Wolfe approach (cf. Sec 1.2.4): Given any two points $v_{1}, v_{2}, v_{1}<v_{2}$, surrounding the optimum and with dual values $\Phi_{1}, \Phi_{2}$ and subgradients $G_{1}>0, G_{2}<0$, linear approximation at each of the two points gives an optimistic (upper) bound for the dually optimal value as

$$
\Phi=\frac{\Phi_{2} G_{1}-\Phi_{1} G_{2}-G_{1} G_{2}\left(v_{2}-v_{1}\right)}{G_{1}-G_{2}}
$$

and the next iterate

$$
v=\frac{G_{1} v_{1}+\left|G_{2}\right| \cdot v_{2}}{G_{1}+\left|G_{2}\right|}-\frac{\Phi_{1}-\Phi_{2}}{G_{1}+\left|G_{2}\right|} .
$$

By a successive reducing of the uncertainty interval ] $v_{1}, v_{2}$ [, the process goes on until the difference between the solutions $\boldsymbol{\Lambda}^{*}\left(\nu_{i}\right)$ for $i=1,2$ is due to, e.g, a single maintenance choice.

Whenever $G_{1}, G_{2}$ have magnitudes of different orders it may be advantageous to replace the Dantzig-Wolfe updating by interpolation, in order to solve $G=0$. Linear $G$-interpolation:

$$
v=\frac{G_{1} v_{2}+\left|G_{2}\right| \cdot v_{1}}{G_{1}+\left|G_{2}\right|}
$$

or cubic $(\Phi, G)$-interpolation:

$$
v=b_{1} \pm \sqrt{b_{1}^{2}-b_{0}}
$$

where

$$
b_{0}=v_{1} v_{2}+\frac{1}{3} \cdot \frac{G_{1} v_{2}^{2}+G_{2} v_{1}^{2}-v_{1} v_{2}\left(G_{1}+G_{2}\right)}{G_{1}+G_{2}-2 \cdot \frac{\Phi_{2}-\Phi_{1}}{v_{2}-v_{1}}}, \quad b_{1}=\frac{1}{2}\left(v_{1}+v_{2}\right)-\frac{1}{6} \cdot \frac{\left(G_{2}-G_{1}\right)\left(v_{2}-v_{1}\right)}{G_{1}+G_{2}-2 \cdot \frac{\Phi_{2}-\Phi_{1}}{v_{2}-v_{1}}} .
$$

Here the $v$-sign should be chosen such that the new iterate falls between $v_{1}$ and $v_{2}$.


Figure 5.1 Dual price updating methods. a (left): Dantzig-Wolfe, $\mathbf{b}$ (middle): linear $G$ interpolation, $\mathbf{c}$ (right): cubic ( $\Phi, G$ )-interpolation.

A cubic interpolation might be successful, if the dual function surface consists of a large number of (small) facets of constant subproblem solutions. The DW method might be superior, especially in a linearized model and if few facets are important for the dual optimum.

For the DW method, in general, we have prepared a modification to accomplish convergence in practice, also in non-convex problem instances. In a 1D case like here it simply means that we do not accept the new iterate to fall too close to $v_{1}$ or $v_{2}$. The Lagrangean multipliers can be occasionally bounded from above (for one dual iteration; or permanently, since extremely high values would be interpreted as unrealistic demands for the benefit/cost ratio (1.1) - in practice unsolvable). Such bounds, providing a compact $\mathbf{v}$-domain, eliminate the risk of unbounded solutions. The remaining convergence trouble is that the method might get stuck, taking infinitesimal or giant steps and never approaching the dual optimum.

Our aim is to separate the dual iterates. The optimal solution of the latest master problem is the top $\mathbf{v}^{*}$ - see Fig 5.2 for a 2D illustration - close to the current iterate $\mathbf{v}^{(I)}$. (Cf. Fig 1.3; in Fig 5.2 we visualise a $3^{\text {rd }} \Phi$-dimension and a tetrahedron bounded by three majorant planes and the level plane $\Phi=\Phi^{*}$.) Denoting the dual iteration number by $i$ and the generated affine majorants by

$$
(\Phi(\mathbf{v}) \leq) \hat{\Phi}\left(\mathbf{v} ; \mathbf{v}_{(i)}\right) \equiv \Phi\left(\mathbf{v}_{(i)}\right)+\left(\mathbf{v}-\mathbf{v}_{(i)}\right)^{T} \mathbf{G}\left(\mathbf{v}_{(i)}\right)=F\left(\mathbf{v}_{(i)}\right)+\mathbf{v}^{T} \mathbf{G}\left(\mathbf{v}_{(i)}\right) \quad i=1, \ldots, I
$$

we consider the linearization level set $\mathcal{V}:=\left\{\mathbf{v}: \hat{\Phi}\left(\mathbf{v} ; \mathbf{v}_{\left(i^{*}\right)}\right) \geq \Phi^{*}\right\}$ in Fig 5.2, where the level value $\Phi^{*} \equiv \Phi\left(\mathbf{v}_{\left(i^{*}\right)}\right)$ is attained in iteration $i^{*}$, as determined in either of two ways.
(1): $i^{*} \in \underset{i}{\arg \max } \Phi\left(\mathbf{v}_{(i)}\right)$ or (2): $i^{*} \in \underset{i}{\arg \min }\left\|\mathbf{v}_{(i)}-\mathbf{v}^{*}\right\|$ (cf. formula below).

Letting $F^{*} \equiv F\left(\mathbf{v}_{\left(i^{*}\right)}\right)$ and $\mathbf{G}^{*} \equiv \mathbf{G}\left(\mathbf{v}_{\left(i^{*}\right)}\right)$, we compute the step length $\sigma_{(i)}$ along the subgradient $\mathbf{G}^{*}$ from $\mathbf{v}^{*}$ needed to reach each majorant $i$ on level $\Phi^{*}$ :

$$
\left\{\begin{aligned}
& \Phi^{*}= F\left(\mathbf{v}_{(i)}\right)+\mathbf{v}^{T} \mathbf{G}\left(\mathbf{v}_{(i)}\right) \\
& \mathbf{v}=\mathbf{v}^{*}+\sigma_{(i)} \cdot \mathbf{G}^{*}
\end{aligned} \text { for } i=1, \ldots, I\right.
$$

For $\mathbf{G}\left(\mathbf{v}_{(i)}\right)^{T} \mathbf{G}^{*} \neq 0$ we get $\sigma_{(i)}=\frac{\Phi^{*}-\left(F\left(\mathbf{v}_{(i)}\right)+\mathbf{v}^{* T} \mathbf{G}\left(\mathbf{v}_{(i)}\right)\right)}{\mathbf{G}\left(\mathbf{v}_{(i)}\right)^{T} \mathbf{G}^{*}}$.
Introducing the dimensioning step lengths $\sigma_{C}=\min _{i}\left[\sigma_{(i)}: \sigma_{(i)}>0\right], \sigma_{A}=\max _{i}\left[\sigma_{(i)}: \sigma_{(i)}<0\right]$ we identify the two intersections with level set $\mathcal{V}$ (see Fig 5.2) $\mathbf{v}_{j}=\mathbf{v}^{*}+\sigma_{j} \cdot \mathbf{G}^{*} \quad j=A, C$. If, say, $\left|\sigma_{A}\right|<\sigma_{C}$ (as in Fig 5.2) we focus on $\mathbf{v}_{A}$. A modification is performed, if the relative distance $\left\|\mathbf{v}^{*}-\mathbf{v}_{A}\right\| /\left\|\mathbf{v}_{C}-\mathbf{v}_{A}\right\|=\sigma_{A} /\left(\sigma_{C}-\sigma_{A}\right)$ is less than a prescribed lower bound $p$, $0<p<\frac{1}{2}$, i.e. if $\mathbf{v}^{*}$ is close to the boundary of $\mathcal{V}$. (In Fig $5.2 p=0.2$ is used.) In such a case we add a constraint to the master problem, which excludes the greyish region and is derived as follows.


Figure 5.2 Illustrative example of convergence improving modification of DW-iterate.

Let $\sigma:=(1-p) \sigma_{A}+p \sigma_{C}$ and $\mathbf{v}_{B}:=\mathbf{v}_{A}+p \cdot\left(\mathbf{v}_{C}-\mathbf{v}_{A}\right)=\mathbf{v}^{*}+\sigma \cdot \mathbf{G}^{*}$, on an acceptable (relative) $p$-distance from $\mathbf{v}_{A}$. We let the majorant of iteration $i^{*}$ and its level value at $\mathbf{v}_{B}$ determine an occasional "cut":

$$
\hat{\Phi}\left(\mathbf{v} ; \mathbf{v}_{\left(i^{*}\right)}\right) \geq \hat{\Phi}\left(\mathbf{v}_{B} ; \mathbf{v}_{\left(i^{*}\right)}\right) \Leftrightarrow \mathbf{G}^{* T} \mathbf{v} \geq \mathbf{G}^{* T} \mathbf{v}^{*}+\sigma \cdot\left\|\mathbf{G}^{*}\right\|^{2}
$$

By adding one or more such constraints to the optimal simplex tableau of the master problem and applying the dual simplex method, see [Dantzig (1963), p 243], the modified solution, $\mathbf{v}^{(I+1)}$ in Fig 5.2, is usually derived in a few LP-iterations. By solving the dual subproblem for $\mathbf{v}=\mathbf{v}^{(I+1)}$ in Fig 5.2, a new real "cut" (affine majorant) is found. If the first proposal $\mathbf{v}^{*}$ is correct, all but the greyish region might be cut off from the level set $\mathcal{V}$. It should be pointed out that the added constraints are occasional, utilised in one dual iteration only. Near the $\mathbf{v}$ bounds, a special modification is needed.

### 5.3.2 Value iteration

Consider the network $r$-subproblem (with $r$ implicit). As for the average cost MDP (5.6), the Gauss-Seidel variant means that from a given initial solution $\left(\mathbf{v}^{i}, w^{i}\right)$ in iteration $i=0$, e.g. chosen according to Prop 5.2 or, letting $c_{\min }:=\min _{n} \min _{\mathbf{u}_{n} \in \mathcal{U}\left(\mathbf{x}_{n}\right)} \widetilde{c}_{n}\left(\mathbf{u}_{n}\right), \mathbf{v}^{0}=\mathbf{0}, w^{0}=c_{\min }$ or $\mathrm{v}_{n}^{0}=\min _{\mathbf{u}_{n} \in \mathcal{U}\left(\mathbf{x}_{n}\right)} \widetilde{c}_{n}\left(\mathbf{u}_{n}\right)$ together with $w^{0}=c_{\min }$ if $c_{\text {min }}<0$, otherwise $w^{0}=0$. All three choices are feasible in (5.7)), a fact that we prove for one of them.

Lemma 5.5: $\mathrm{v}_{n}^{0}=\min _{\mathbf{u}_{n} \in \mathcal{U}\left(\mathbf{x}_{n}\right)} \widetilde{c}_{n}\left(\mathbf{u}_{n}\right)$ together with (a): $w^{0}=c_{\text {min }}$ if $c_{\text {min }}<0$, otherwise (b): $w^{0}=0$, provide feasibility in (5.7b).

Proof: Consider the difference between LHS and RHS for an arbitrary pair ( $n, m$ ) in (5.7b).

$$
f_{n m}\left(\mathbf{v}^{0}, w^{0}\right):=w^{0}+\min _{m^{\prime} \in \mathcal{M}_{n}} \widetilde{c}_{r n m^{\prime}}-\sum_{n^{\prime}} p_{r n n^{\prime} m} \cdot \min _{m^{\prime} \in \mathcal{M}_{n^{\prime}}} \widetilde{c}_{r n^{\prime} m^{\prime}}-\widetilde{c}_{r n m} \leq w^{0}+0-\sum_{n^{\prime}} p_{r n n^{\prime} m} \cdot \min _{m^{\prime} \in \mathcal{M}_{n^{\prime}}} \widetilde{c}_{r n^{\prime} m^{\prime}}
$$

Case (a) $p_{r n n^{\prime} m} \cdot \min _{m^{\prime} \in \mathcal{M}_{n^{\prime}}} \widetilde{c}_{r n^{\prime} m^{\prime}} \geq p_{r n n^{\prime} m} \cdot c_{\text {min }}$ implies

$$
f_{n m}\left(\mathbf{v}^{0}, w^{0}\right) \leq c_{\min }-c_{\min } \cdot \sum_{n^{\prime}} p_{r n n^{\prime} m} \leq c_{\min }-c_{\min } \cdot 1=0, \text { feasible }
$$

Case (b) $\min _{m^{\prime} \in \mathcal{M}_{n^{\prime}}} \widetilde{c}_{r n^{\prime} m^{\prime}} \geq c_{\text {min }}$ implies

$$
f_{n m}\left(\mathbf{v}^{0}, w^{0}\right) \leq 0+\sum_{n^{\prime}} p_{r n n^{\prime} m}\left(-\widetilde{c}_{r n^{\prime} m^{\prime}}\right) \leq c_{\min }+\sum_{n^{\prime}} p_{r n n^{\prime} m}\left(-c_{\min }\right) \leq c_{\min }-c_{\min } \cdot 1=0
$$

feasible.
From the given start solution $\left(\mathbf{v}^{0}, w^{0}\right)$ the succeeding values are computed sequentially:

$$
\text { For } n=1,2, \ldots,|\mathcal{N}|: \mathrm{v}_{n}^{i+1}=\min _{\mathbf{u}_{n} \in \mathcal{U}\left(\mathbf{x}_{n}\right)}\left[\widetilde{c}_{n}\left(\mathbf{u}_{n}\right)+\sum_{n^{\prime}<n} p_{n n^{\prime}}\left(\mathbf{u}_{n}\right) \cdot \mathrm{v}_{n^{\prime}}^{i+1}+\sum_{n^{\prime} \geq n} p_{n n^{\prime}}\left(\mathbf{u}_{n}\right) \cdot \mathrm{v}_{n^{\prime}}^{i}\right]-w^{i}
$$

For the updating of the "gain" $w^{i}$, [Bertsekas (1995)] gives the details. He suggests a Jacobian updating, i.e. postponing the use of the latest computed values $\mathrm{v}_{n^{\prime}}^{i+1}, n^{\prime}<n$, until all nodes $n \in \mathcal{N}$ have been considered, is recommended e.g. every tenth iteration. The iterating continues until the values converge in norm, as $\left\|\mathbf{v}^{i+1}-\mathbf{v}^{i}\right\|+\left\|w^{i+1}-w^{i}\right\|<\varepsilon$, and - if the works $\mathbf{U}^{i}=\left(\mathbf{u}_{n}^{i}\right)_{n}$ that provides $\mathbf{v}^{i+1}$ is registered - also until stagnation $\mathbf{U}^{i+1}=\mathbf{U}^{i}$. In the discounted MDP case (5.2) we exchange $\mathbf{v}^{i}$ for $\varphi^{i}$ and exclude $w^{i}$. Although the two sums that represent the succeeding costs become suppressed by the discount factor $d$, the direct annual cost becomes less influential, since $\widetilde{\mathbf{c}}_{r}\left(\mathbf{U}_{r}\right)=\frac{1}{1-d} \mathbf{c}_{r}\left(\mathbf{U}_{r}\right)$ is replaced by $\mathbf{c}_{r}\left(\mathbf{U}_{r}\right)$ and $d>1-d$ is expected.

### 5.3.3 Policy iteration

Consider the network $r$-subproblem (with $r$ implicit). For the average cost MDP an initial pavement works solution $\mathbf{U}^{i}=\left(\mathbf{u}_{n}^{i}\right)_{n}$ in iteration $i=0$ is given, e.g. $\mathbf{U}^{0} \in \underset{\mathbf{U}}{\arg \min } \widetilde{\mathbf{c}}(\mathbf{U})$ for the current $v_{\infty}$-value. For a given $\mathbf{U}^{i}$ the values $\mathbf{v}=\mathbf{v}^{i}, w=w^{i}$ solve the linear system of equations

$$
\left\{\begin{array}{l}
{\left[\mathbf{I}-\mathbf{P}\left(\mathbf{U}^{i}\right)\right] \mathbf{v}+w \cdot \mathbf{1}=\widetilde{c}\left(\mathbf{U}^{i}\right)} \\
\mathbf{v}_{n_{0}}=0
\end{array} \quad\right. \text { (policy evaluation = value determination) }
$$

where the choice $\mathrm{v}_{n_{0}}=0$ is motivated by the occurrence of the arbitrary constant $C$ in the proof of Prop 5.2. For a given value iterate $\left(\mathbf{v}^{i}, w^{i}\right)$ the next works iterate is

$$
\mathbf{U}^{i+1} \in \underset{\mathbf{U} \in \prod_{n} \mathcal{U}_{n}\left(\mathbf{x}_{n}\right)}{\arg \min }\left[\widetilde{\mathbf{c}}(\mathbf{U})+\mathbf{P}(\mathbf{U}) \mathbf{v}^{i}\right] \quad \quad \text { (policy improvement) }
$$

The stop criterion is stagnation $\mathbf{U}^{i+1}=\mathbf{U}^{i}$. As for degeneracy - see the following subsections.
For the discounted MDP, $\mathbf{c}$ replaces $\widetilde{\mathbf{c}}, d \cdot \mathbf{P}$ replaces $\mathbf{P}, \varphi$ replaces $\mathbf{v}$, and $w$ exits.
[Howard (1960)] proves that every policy improvement step leads to an improved value $\mathrm{v}_{n}$ (or $\varphi_{n}$ ) for at least one node $n$, and that - when the policy improvement routine converges (in a finite number of iterations) - no other $\mathbf{U}$ can provide better values: statewise optimal.

### 5.3.4 LP iteration

Our use of LP comes from the observation that a simplified simplex pivoting can be performed, without any explicit primal length variables $\lambda$, without any computation of the current basis inverse and of the leaving basic variable. Consider the network $r$-subproblem (with $r$ implicit) and the linearized version (5.6) of the average cost MDP. In each LP-iteration we assume that one works option $\hat{m}_{n} \in \mathcal{M}_{n}$ per node $n$ is basic and that exactly one closed comm.class (of cardinality $\geq 2)$ exists. In matrix-vector notation, letting $\hat{\mathbf{P}} \equiv\left(p_{n n^{\prime} \hat{m}_{n}}\right)_{n, n^{\prime}}, \hat{\lambda} \equiv\left(\lambda_{n \hat{m}_{n}}\right)_{n}$ and using 1 as the one-vector, the iterate $\hat{\lambda}$ satisfies the steady-state $r$-subproblem constraints

$$
\left\{\begin{array}{c}
\left(\mathbf{I}-\hat{\mathbf{P}}^{T}\right) \hat{\lambda}=\mathbf{0} \\
\mathbf{1}^{T} \hat{\lambda}=\text { Len } \\
\hat{\lambda} \geq \mathbf{0} .
\end{array}\right.
$$

Since the equations in (5.6b') are linearly dependent we omit one, for some node $n_{0}$ in the only closed comm.class, and distinguish the corresponding $\hat{\lambda}$-component $\hat{\lambda}_{0}$. By partitioning $\hat{\lambda}=\binom{\hat{\lambda}_{+}}{\hat{\lambda}_{0}}$ we get a reduced $\hat{\mathbf{P}}=\binom{\hat{\mathbf{P}}_{+}}{\hat{\mathbf{p}}_{0}{ }^{T}}$, without column $\left(p_{n n_{0}}\right)_{n}$, and a reduced transposed

### 5.3 Method

coefficient matrix $\mathbf{G}:=\left(\begin{array}{cc}\mathbf{I}-\hat{\mathbf{P}}_{+} & \mathbf{1} \\ -\hat{\mathbf{p}}_{0}{ }^{T} & 1\end{array}\right)$.
In standard LP $\mathbf{G}^{T} \hat{\lambda}=\binom{\mathbf{0}}{$ Len } is solved for $\hat{\lambda}$, and the pivoting involves checking $\hat{\lambda} \geq 0$. Here this is unnecessary, since:

Proposition 5.4: Assume that the MC consists of exactly one closed comm.class $c$ and possibly also a (collective) transient class. $\hat{\lambda}$ solving $\mathbf{G}^{T} \hat{\lambda}=\binom{\mathbf{0}}{$ Len } is nonnegative, positive in $c$ and zero otherwise.

Proof: Letting $\hat{\mathbf{P}}_{+} \equiv\left(\begin{array}{cc}\hat{\mathbf{P}}_{+}^{c} & \mathbf{O} \\ \mathbf{S} & \mathbf{Q}\end{array}\right)$, Lemmas 5.3, 5.4 imply that $(\mathbf{I}-\mathbf{Q})^{-1},\left(\mathbf{I}-\hat{\mathbf{P}}_{+}^{c}\right)^{-1}$ exist, nonnegative. We get

$$
\mathbf{H}:=\left(\mathbf{I}-\hat{\mathbf{P}}_{+}\right)^{-1}=\left(\begin{array}{cc}
\left(\mathbf{I}-\hat{\mathbf{P}}_{+}^{c}\right)^{-1} & \mathbf{O} \\
(\mathbf{I}-\mathbf{Q})^{-1} \mathbf{S}\left(\mathbf{I}-\hat{\mathbf{P}}_{+}^{c}\right)^{-1} & (\mathbf{I}-\mathbf{Q})^{-1}
\end{array}\right), \text { nonnegative. }
$$

Making the ansatz $\left(\begin{array}{cc}\mathbf{A} & \mathbf{b} \\ \mathbf{c}^{T} & d\end{array}\right):=\mathbf{G}^{-T}$, the solution $\hat{\lambda}$ can be written $\binom{\hat{\lambda}_{+}}{\hat{\lambda}_{0}}=\binom{\mathbf{b}}{d} \cdot$ Len .
We get $d=1 /\left(1+\mathbf{1}^{T} \mathbf{H}^{T} \hat{\mathbf{p}}_{0}\right), \mathbf{b}=\mathbf{H}^{T} \hat{\mathbf{p}}_{0} \cdot d$.
Since $\mathbf{H}, \hat{\mathbf{p}}_{0}$ are nonnegative we get $d>0$ and $\mathbf{b}, \hat{\lambda}$ nonnegative.
Partitioning into the comm.class $c$ and the transient states, as $\hat{\mathbf{p}}_{0} \equiv\binom{\hat{\mathbf{p}}_{0}^{c}}{\mathbf{0}}$, we get

$$
\mathbf{b} \equiv\binom{\mathbf{b}^{c}}{\mathbf{b}^{\text {trans }}}=d \cdot \mathbf{H}^{T} \hat{\mathbf{p}}_{0}=\binom{d \cdot\left(\mathbf{I}-\hat{\mathbf{p}}_{+}^{c^{T}}\right)^{-1} \hat{\mathbf{p}}_{0}^{c}}{\mathbf{0}} .
$$

The zero length in the transient states is obvious. Using Lemma 5.4, $\mathbf{b}^{c}$ can be expanded as

$$
\mathbf{b}^{c}=d \cdot\left(\mathbf{I}-\hat{\mathbf{P}}_{+}^{c^{T}}\right)^{-1} \hat{\mathbf{p}}_{0}^{c}=d \cdot\left(\hat{\mathbf{p}}_{0}^{c}+\hat{\mathbf{P}}_{+}^{c^{T}} \hat{\mathbf{p}}_{0}^{c}+\left(\hat{\mathbf{P}}_{+}^{c}\right)^{2^{T}} \hat{\mathbf{p}}_{0}^{c}+\cdots\right) .
$$

Since the distinguished state $n_{0}$ communicates with at least one other state in $c, \hat{\mathbf{p}}_{0}^{c}$ has a positive component for some $n$. The meaning of a closed $c$ is that every state $n^{\prime}$ in $c$ can be reached in a finite number $k$ of time steps, i.e. $\left(\hat{\mathbf{P}}_{+}^{c}\right)^{k^{T}} \hat{\mathbf{p}}_{0}^{c}$ has a (first) positive $n^{\prime}$-component. Hence every $\mathbf{b}^{c}$-component $n^{\prime}$ gets positive contributions in the series expansion.

As for the choice of a dually feasible, initial works solution $\mathbf{U}^{0}$ - see Policy iteration. The dual constraints (5.7b) are used for LP-pivoting. For each node $n$ we use one basic $\mathbf{u}_{n}$-variable, here indexed $m \in \mathcal{M}_{n}$, and let $\mathrm{v}_{n_{0}}=0$ for some node $n_{0}$. The reduced costs

$$
\bar{c}_{n m}:=\widetilde{c}_{n m}+\sum_{n^{\prime}} p_{n n^{\prime} m} \mathrm{v}_{n^{\prime}}-\mathrm{v}_{n}-w
$$

determine the $(n, m)$-variables entering the basis. If one per iteration enters, we have the classical LP-case. If one entering variable per node $n$ is admitted, the one with most negative
reduced cost, the LP-iteration method coincides with policy iteration (compare the $\bar{c}_{n m}$ expression, for constant $\mathrm{v}_{n}+w$, with the policy improvement formula in Sec 5.3.3). By adjusting the entering rule to, e.g., the most negative $\bar{c}_{n m}$-values (for different states $n$ ), a more general method is obtained. Given a basis of $m$-values $\hat{m}_{n}$ for all the nodes $n \in \mathcal{N}$, the solving of the linear system $\bar{c}_{n \hat{m}_{n}}=0 \quad \forall n$ for the dual prices $(\mathbf{v}, w)$ coincides with the value determination step of policy iteration. The updating and registration of the LP-primal variables $\boldsymbol{\lambda}$ in (5.6), needed for any budget evaluation, is postponed until the length distributions step see below.
During the LP-iterating (and policy iterating) we try to circumvent any possible trouble with reducible MCs. Encountered absorbing states and multiple comm.classes are handled by occasional manipulation, in changing the current works choices by force. The purpose is to pass by the non-optimal, irrelevant cases. If the final works choice is manipulated, this is signalised as a degeneracy case to the length distributions step - see Sec 5.3.6 below. We distinguish two types of (possibly co-existing) non-optimal solutions.

- $1^{\text {st }}$ kind of degeneracy: a reducible MC with at least two closed comm.classes is identified in the final solution, by a singular system of eqns in the value determination step.
- $2^{\text {nd }}$ kind of degeneracy: a reducible MC with at least one absorbing state is identified in the final solution, by the $w$-value satisfying $w=\bar{w}$ for some upper bound $\bar{w}$ coming from the possible (pre-identified) absorbing states $n\left(\right.$ with $\mathcal{M}_{r n}^{\text {absorb }} \neq \varnothing$ ) in Def 5.2.

As for the discounted MDP, an adequate LP formulation (cf. (5.3)) is

$$
\begin{array}{ll}
\underset{\left(\varphi_{n}\right)_{n}}{\operatorname{maximise}} & \sum_{n} \lambda_{n}^{0} \cdot \varphi_{n} \\
\text { subject to } & \varphi_{n}-d \cdot \sum_{n^{\prime}} p_{n n^{\prime} m} \varphi_{n^{\prime}} \leq c_{n m} \quad \forall n, m \in \mathcal{M}_{r n} . \tag{5.8b}
\end{array}
$$

The similarities with (5.7) makes the LP-handling evident - but the argument for a simplified pivoting is slightly different. Therefore we view (5.8) as an LP-dual, with primal

$$
\begin{array}{ll}
\underset{\left(\lambda_{n m}\right)_{n, m}}{\operatorname{minimise}} & \sum_{n} \sum_{m \in \mathcal{M}_{r n}} c_{n m} \lambda_{n m} \\
\text { subject to } & \left\{\begin{array}{lc}
\sum_{m \in \mathcal{M}_{r n}} \lambda_{n m}-d \cdot \sum_{\substack{n^{\prime} \\
\lambda_{n m} \geq 0}} \sum_{m \in \mathcal{M}_{r n}} p_{n^{\prime} n m} \lambda_{n^{\prime} m}=\lambda_{n}^{0} & \forall n \\
\lambda_{n m} & \forall n, m \in \mathcal{M}_{r n} .
\end{array}\right. \tag{5.9b}
\end{array}
$$

Here (5.9b) implies

$$
\begin{aligned}
& \sum_{n}\left(\sum_{m \in \mathcal{M}_{r n}} \lambda_{n m}-d \cdot \sum_{n^{\prime}} \sum_{m \in \mathcal{M}_{r n}} p_{n^{\prime} n m} \lambda_{n^{\prime} m}\right)=\sum_{n} \sum_{m \in \mathcal{M}_{r n}} \lambda_{n m}-d \cdot \sum_{n^{\prime}} \sum_{m \in \mathcal{M}_{r n}} 1 \cdot \lambda_{n^{\prime} m}=(1-d) \cdot \sum_{n} \sum_{m \in \mathcal{M}_{r n}} \lambda_{n m} \\
& \sum_{n} \sum_{m \in \mathcal{M}_{r n}} \lambda_{n m}=\frac{1}{1-d} \cdot \sum_{n} \lambda_{n}^{0}
\end{aligned}
$$

i.e. a constant total length (in each road class).

With the current basis consisting of one works choice $\hat{m}_{n} \in \mathcal{M}_{r n}$ per node $n$, corresponding to length vector $\hat{\lambda}$ and transition matrix $\hat{\mathbf{P}}$, the iterate $\hat{\lambda}$ solves $\left(\mathbf{I}-d \cdot \hat{\mathbf{P}}^{T}\right) \hat{\lambda}=\hat{\lambda}_{0}$.

### 5.3 Method

Proposition 5.5: Assume that the the MC consists of one closed comm.class $c$ and possibly also a transient class. $\hat{\lambda}$ solving $\left(\mathbf{I}-d \cdot \hat{\mathbf{P}}^{T}\right) \hat{\lambda}=\hat{\lambda}_{0}$ is nonnegative, positive in $c$.

Proof: $\mathbf{I}-d \hat{\mathbf{P}}$ is strictly diagonally dominant with positive diagonal, otherwise nonpositive. Hence, see [Ortega and Rheinboldt (1970), p 55], $\mathbf{I}-d \hat{\mathbf{P}}$ is an M-matrix, i.e.
$\mathbf{H}:=(\mathbf{I}-d \hat{\mathbf{P}})^{-1} \geq \mathbf{O}$. Letting $\hat{\mathbf{P}} \equiv\left(\begin{array}{cc}\hat{\mathbf{P}}^{c} & \mathbf{O} \\ \mathbf{S} & \mathbf{Q}\end{array}\right), \hat{\mathbf{P}}^{c}$ is irreducible, and thus also $\mathbf{I}-d \hat{\mathbf{P}}^{c}$. Then,
[Ortega (1987), p 221], $\left(\mathbf{I}-d \hat{\mathbf{P}}^{c}\right)^{-1}>\mathbf{O}$. The whole inverse $\mathbf{H}$ takes the form (cf. Prop 5.4)

$$
\mathbf{H}=\left(\begin{array}{cc}
\left(\mathbf{I}-d \hat{\mathbf{P}}^{c}\right)^{-1} & \mathbf{O} \\
d \cdot(\mathbf{I}-d \mathbf{Q})^{-1} \mathbf{S}\left(\mathbf{I}-d \hat{\mathbf{P}}^{c}\right)^{-1} & (\mathbf{I}-d \mathbf{Q})^{-1}
\end{array}\right) .
$$

Partitioning $\hat{\lambda} \equiv\binom{\hat{\lambda}^{c}}{\hat{\lambda}^{\text {trans }}}, \hat{\lambda}_{0} \equiv\binom{\hat{\lambda}_{0}^{c}}{\hat{\lambda}_{0}^{\text {trans }}}$ we get

$$
\begin{aligned}
& \hat{\lambda}^{c}=\left(\mathbf{I}-d \hat{\mathbf{P}}^{T}\right)^{-1}\left(\hat{\lambda}_{0}^{c}+d \cdot \mathbf{S}^{T}\left(\mathbf{I}-d \mathbf{Q}^{T}\right)^{-1} \hat{\lambda}_{0}^{\text {trans }}\right), \\
& \hat{\lambda}^{\text {trans }}=\left(\mathbf{I}-d \mathbf{Q}^{T}\right)^{-1} \hat{\lambda}_{0}^{\text {trans }} .
\end{aligned}
$$

Since $\hat{\lambda}_{0} \geq \mathbf{0}$ and $\mathbf{H} \geq \mathbf{O}$ we get $\hat{\lambda} \geq \mathbf{0}$. For $\hat{\lambda}_{0}^{c} \not \equiv \mathbf{0},\left(\mathbf{I}-d \hat{\mathbf{P}}^{c^{T}}\right)^{-1}>\mathbf{O}$ implies $\hat{\lambda}^{c}>\mathbf{0}$.
For $\hat{\lambda}_{0}^{c} \equiv \mathbf{0}$ we have $\hat{\lambda}_{0 n}^{\text {trans }} \neq \mathbf{0}$ for some node $n$. We may expand

$$
\left(\mathbf{I}-d \mathbf{Q}^{T}\right)^{-1} \hat{\lambda}_{0}^{\text {trans }}=\hat{\lambda}_{0}^{\text {trans }}+d \mathbf{Q}^{T} \hat{\lambda}_{0}^{\text {trans }}+d^{2} \mathbf{Q}^{2^{T}} \hat{\lambda}_{0}^{\text {trans }}+\cdots
$$

Since the transient part is not closed, at least one transient state $n^{\prime}$ communicates with $c$, i.e. the $n^{\prime}$-row of $\mathbf{S}$ is non-zero and $n^{\prime}$ can be reached from $n$ in a finite number $k$ of time steps corresponding to a (first) positive $n^{\prime}$-component of $d^{k} \mathbf{Q}^{k^{T}} \hat{\lambda}_{0}^{\text {trans }}$ in the series expansion. $\left(\mathbf{I}-d \hat{\mathbf{P}}^{c^{T}}\right)^{-1}>\mathbf{O}$ implies $\hat{\lambda}^{c}>\mathbf{0}$.

### 5.3.5 Newton iteration

We avoid the $2^{\text {nd }}$ kind of degeneracy during the very solution process, as in Sec 5.3.4, by confining the considered works options to $m \in \mathcal{M}_{r n}^{\text {transf }}:=\mathcal{M}_{r n} \backslash \mathcal{M}_{r n}^{\text {absorb }}$ (cf. Def 5.2). The LP-dual constraints (5.7b) can be lumped into piecewise linear, convex constraints

$$
\begin{equation*}
f_{n}(\mathbf{v}, w):=\mathrm{v}_{n}+w-\min _{m \in \mathcal{M}_{r n}^{\text {trans }}}\left[\tilde{c}_{r n m}+\sum_{n^{\prime}} p_{r n n^{\prime} m} \mathbf{v}_{n^{\prime}}\right] \leq 0 \quad \forall n \tag{5.10}
\end{equation*}
$$

By assuming that each encountered MC consists of a single closed comm.class (of cardinality $\geq 2$ ) plus possible transient states we also avoid the $1^{\text {st }}$ kind of degeneracy. According to Prop 5.4, a characteristic for such MCs is that the limit length distribution in the LP-primal (5.6) has a positive $\lambda_{r n m}$-value for each node $n$ in road class $r$, corresponding to one active LP-dual constraint for each $n$. In (5.10) we might therefore look for a solution of the nonlinear system $\mathbf{f}=\mathbf{0}$. The Jacobian matrix $\mathbf{G}:=\nabla \mathbf{f}$ is well defined almost everywhere in the $(\mathbf{v}, w)$-space. The G-discontinuities are easily identifiable during calculation by (5.10): at least two $m$ options for some $n$ provide the minimum in $f_{n}$.

In order to guarantee convergence with Newton's method we replace each discontinuities by a smooth approximation in an arbitrarily small region, as follows. The convex function $f_{n}$ is piecewise affine. The smooth corrections are made by convex combinations of the affine facets involved in the discontinuity, so that convexity remains. Since the number of works options is finite, a switch of locally best option can occur in a finite number of ways $a$, each way represented by one arbitrary switch point $\mathbf{x}_{a}, \mathbf{x} \equiv(\mathbf{v}, w)$. (In our application we expect at most 2 (-3) minimising $m$-options per node, routine maintenance and 1(-2) layer thicknesses.) Consider the switch zone of $\mathbf{x}_{a}$, the set of points having $\mathbf{x}_{a}$ closest among all the switch points of $a$, and at a maximum affine distance $\rho$ from $\mathbf{x}_{a}$. See Figure 1a for an illustration of a switch $a$ between two locally best works options, corresponding to two affine pieces $z=f_{n}(\mathbf{x})=\max \left[\hat{\mathbf{q}}_{1}{ }^{T} \mathbf{x}-\hat{r}_{1}, \hat{\mathbf{q}}_{2}{ }^{T} \mathbf{x}-\hat{r}_{2}\right]$ and a switch zone characterised by $t \in[0,1]$. In the switch zone $f_{n}$ is replaced by a smooth convex combination

$$
z=\widetilde{f}_{n}(\mathbf{x}(t))=\alpha(t) \cdot\left(\hat{\mathbf{q}}_{1}{ }^{T} \mathbf{x}(t)-\hat{r}_{1}\right)+(1-\alpha(t)) \cdot\left(\hat{\mathbf{q}}_{2}{ }^{T} \mathbf{x}(t)-\hat{r}_{2}\right), \quad \alpha(t) \in[0,1] \forall t .
$$

Smoothness put demands on the parameter function $\alpha$ as illustrated in Figure 1b.
One and the same form of the approximation $\widetilde{f}_{n}$ is used for all the switch points of $a$.


Figure 5.3a: Approximation of $f_{n}$ on switch zone


Figure 5.3b Convex combination-parameter on switch zone

At a G-discontinuity data takes the same form as otherwise, analogous to a stochastic choice with probabilities $\alpha, 1-\alpha$ and formally written $\hat{m}_{n}$. The data are denoted $\hat{\mathbf{P}}:=\left(p_{r n n^{\prime} \hat{m}_{n}}\right)_{n, n^{\prime}}$, $\hat{\mathbf{c}}:=\left(\widetilde{c}_{r n \hat{m}_{n}}\right)_{n}$. Motivated by the arbitrary $C$-constant in the proof of Prop 5.2 we take $\mathrm{v}_{n_{0}}=0$ for one (excluded $\hat{\mathbf{P}}$-column) node $n_{0}$ in the assumed closed comm.class and use the $n_{0}$-data notations $\hat{\mathbf{p}}_{0}, \hat{c}_{0}$. All the remaining node-prices are collected in $\mathbf{v}_{+}$and the corresponding data in $\hat{\mathbf{P}}_{+}, \hat{\mathbf{c}}_{+}$. We get the linearization

$$
\hat{\mathbf{f}}(\mathbf{v}, w):=\binom{\left(\mathbf{I}-\hat{\mathbf{P}}_{+}\right) \mathbf{v}_{+}+w \mathbf{1}-\hat{\mathbf{c}}_{+}}{-\hat{\mathbf{p}}_{0}{ }^{T} \mathbf{v}_{+}+w-\hat{c}_{0}}=: \hat{\mathbf{G}}(\mathbf{v}, w)\binom{\mathbf{v}_{+}}{w}-\binom{\hat{\mathbf{c}}_{+}}{\hat{c}_{0}}, \quad \hat{\mathbf{G}}(\mathbf{v}, w)=\nabla \hat{\mathbf{f}}=\left(\begin{array}{cc}
\mathbf{I}-\hat{\mathbf{P}}_{+} & \mathbf{1}  \tag{5.11}\\
-\hat{\mathbf{p}}_{0}{ }^{T} & 1
\end{array}\right) .
$$

Lemma 5.6: $\hat{\mathbf{G}}$ in (5.11) is non-singular.

Proof: By assumption, $\hat{\mathbf{P}}_{+}$is of type $\hat{\mathbf{P}}_{+}=\left(\begin{array}{cc}\hat{\mathbf{P}}_{+}^{c} & \mathbf{O} \\ \mathbf{S} & \mathbf{Q}\end{array}\right)$, with $\hat{\mathbf{P}}_{+}^{c}$ substochastic and $\mathbf{Q}$ transient as in Prop 5.4. The constructive proof of Prop 5.4 shows the existence.

The average cost MDP (5.7) is to $\underset{\mathbf{v}, w}{\operatorname{maximise}} \operatorname{Len}_{r} \cdot w$ subject to $\mathbf{f}(\mathbf{v}, w)=\mathbf{0}$. A correspondence to Cor 5.1 is:

Proposition 5.6: Any solution of $\mathbf{f}=\mathbf{0}$ solves the average cost MDP (5.7).
Proof: Any solution ( $\hat{\mathbf{v}}, \hat{w}$ ) is a feasible solution in the LP-dual (5.7). The corresponding MC and data $\hat{\mathbf{G}}, \hat{\mathbf{c}}$ define a LP-primally feasible solution $\lambda$ as in Prop 5.4. Writing $\mathbf{f}(\hat{\mathbf{v}}, \hat{w}) \equiv \hat{\mathbf{G}}\binom{\hat{\mathbf{v}}}{\hat{w}}-\hat{\mathbf{c}}=\mathbf{0}$ the primal objective value becomes

$$
\hat{\mathbf{c}}^{T} \boldsymbol{\lambda}=\left(\hat{\mathbf{v}}^{T}, \hat{w}\right) \hat{\mathbf{G}}^{T} \boldsymbol{\lambda}=\left(\hat{\mathbf{v}}^{T}, \hat{w}\right)\binom{\mathbf{0}}{\text { Len }_{r}}=\hat{w} \cdot \text { Len }_{r},
$$

i.e. it coincides with the LP-dual objective value for $(\hat{\mathbf{v}}, \hat{w})$. Strong LP duality applies.

Supported by Prop 5.6 we concentrate on solving the system of equations $\mathbf{f}=\mathbf{0}$. The Newton updating, formally well defined everywhere by our extended $\hat{m}_{n}$-definition and Lemma 5.6, is $(\Delta \mathbf{v}, \Delta w)=-\hat{\mathbf{G}}^{-1} \hat{\mathbf{f}}$. Since we prefer having a flexible step length, we turn to unconstrained optimisation

$$
\begin{equation*}
\underset{\mathbf{v}, w}{\operatorname{minimise}} F(\mathbf{v}, w):=\frac{1}{2} \cdot \sum_{n} f_{n}^{2}(\mathbf{v}, w) \tag{5.12}
\end{equation*}
$$

Introducing $\hat{F}:=\frac{1}{2} \cdot \hat{\mathbf{f}}^{T} \hat{\mathbf{f}}$, writing $\nabla \hat{F}$ as a column vector, and using (5.11) we have

$$
\begin{equation*}
\nabla \hat{F}=\hat{\mathbf{G}}^{T} \hat{\mathbf{f}}, \quad \nabla^{2} \hat{F}=\hat{\mathbf{G}}^{T} \hat{\mathbf{G}} \tag{5.13}
\end{equation*}
$$

Using Lemma 5.6 the Newton updating for (5.12), formally well defined everywhere, becomes

$$
(\Delta \mathbf{v}, \Delta w)=-\left(\hat{\mathbf{G}}^{T} \hat{\mathbf{G}}\right)^{-1} \hat{\mathbf{G}}^{T} \hat{\mathbf{f}}=-\hat{\mathbf{G}}^{-1}\left(\hat{\mathbf{G}}^{T}\right)^{-1} \hat{\mathbf{G}}^{T} \hat{\mathbf{f}}=-\hat{\mathbf{G}}^{-1} \hat{\mathbf{f}}
$$

i.e. the same as for the system of equations.

The convergence of a Newton updating for (5.12) is guaranteed by Thm 5.1 below.
Lemma 5.7: In every (convex) $(\mathbf{v}, w)$-set where $\mathbf{f}(\mathbf{v}, w) \geq \mathbf{0}, F$ is convex.
Proof: For each $n, f_{n}(\mathbf{v}, w)$ is generated by affine minorants ( $\min []$ defines a concave function - cf. a dual function generated by affine majorants - and the minus-sign turns it upside down), thus convex.
For $f_{n}(\mathbf{v}, w) \geq 0$ also $f_{n}{ }^{2}$ becomes convex, and the sum over $n$ of convex functions $f_{n}{ }^{2}$ is convex, see e.g. [Minoux (1986), p 10]).

Lemma 5.8: The region $\mathcal{F}_{+}:=\{(\mathbf{v}, w): \mathbf{f}(\mathbf{v}, w) \geq \mathbf{0}\}$ is simply connected and has a continuous boundary function $w=\eta(\mathbf{v})$.

Proof: For each $n$ and $\mathbf{v}$, (5.10) uniquely defines (a finite) $w=\eta_{n}(\mathbf{v})$ such that $f_{n}(\mathbf{v}, w)=0$. For any $w<\eta_{n}(\mathbf{v})$ we have $f_{n}(\mathbf{v}, w)<0$, otherwise $f_{n}(\mathbf{v}, w) \geq 0$. In any compact $\mathbf{v}$-set the terms in $f_{n}$ vary continuously with $\mathbf{v}$, uniformly bounded from above, and hence also the boundary function.
For any fixed $\mathbf{v}, \eta(\mathbf{v}):=\max _{n} \eta_{n}(\mathbf{v})$ is attained, such that $(\mathbf{v}, w) \notin \mathcal{F}_{+}$for any $w<\eta(\mathbf{v})$, otherwise $(\mathbf{v}, w) \in \mathcal{F}_{+}$. As $\mathbf{v}$ varies in any compact $\mathbf{v}$-set, the continuity of $\eta$ is implied by the continuity of each $\eta_{n}$. Hence $\eta$ is uniformly bounded from above there.
The simple connectedness of $F_{+}$follows from the uniform boundedness of $\eta$ from above.
Theorem 5.1: For any given $\mathbf{v}=\mathbf{v}^{0}$ a $w=w^{0}$ exists, such that $\left(\mathbf{v}^{0}, w^{0}\right) \in \mathcal{F}_{+}$.
Starting from ( $\mathbf{v}^{0}, w^{0}$ ) Newton's method, with Armijo step lengths $\leq 1$, converges to a solution of $\mathbf{f}(\mathbf{v}, w)=\mathbf{0}$.

Proof: The Hadamard theorem, e.g. [Ortega \& Rheinboldt (1970), p 137], guarantees that a unique solution of $\mathbf{f}(\mathbf{v}, w)=\mathbf{0}$ exists. (The existence of a solution follows from Props 5.1 and 5.6.) As the constraint index $n$ varies we realise that the overall number of representative switch points $\mathbf{x}_{a}$ (see above) for $\mathbf{f}$ is finite. Let $\mathcal{D} \subset \mathfrak{R}^{N}$ be a closed sphere including all these switch points, together with their switch zones - where the Jacobian is denoted $\widetilde{\mathbf{G}}$. Outside the switch zones $\widetilde{\mathbf{f}}=\mathbf{f}, \widetilde{\mathbf{G}}=\mathbf{G}$ are used. By construction $\widetilde{\mathbf{G}}$ is continuous on $\mathfrak{R}^{N}$ and $\widetilde{\mathbf{G}}^{-1}$ exists on $\Re^{N}$, according to Lemma 5.6. According to [ibid, p 46], $\widetilde{\mathbf{G}}^{-1}$ is continuous at each point in $\mathcal{D}$, compact, and some $\gamma>0$ exists such that $\left\|\widetilde{\mathbf{G}}^{-1}\right\|<\gamma$ on $\mathcal{D}$. By construction $\mathcal{D}$ includes all representative switch points $\mathbf{x}_{a}$ and their switch zones. Thus no other $\widetilde{\mathbf{G}}$-value is received outside $\mathcal{D}$ than those on $\mathcal{D}$, i.e. $\left\|\widetilde{\mathbf{G}}^{-1}\right\|<\gamma$ on $\mathfrak{R}^{N}$. Now the Hadamard theorem applies: $\widetilde{\mathbf{f}}$ is a homeomorphism of $\mathfrak{R}^{N}$ onto $\mathfrak{R}^{N}$, and $\widetilde{\mathbf{f}}=\mathbf{0}$ has a unique root. By construction, the local approximation error fulfils $\left|\widetilde{f}_{n}(\mathbf{x}(t))-f_{n}(\mathbf{x}(t))\right|<\rho \forall t \in[0,1]$, i.e. for the solution we have $\mathbf{f}^{T} \mathbf{f}<N \cdot \rho^{2}$. Since $\rho>0$ is arbitrary, any stop criterion $\mathbf{f}^{T} \mathbf{f}<\varepsilon$ can be met.

The existence of a $w^{0}$, such that $\left(\mathbf{v}^{0}, w^{0}\right) \in \mathcal{F}_{+}$, follows as in Lemma 5.8. For a given iterate $\left(\mathbf{v}^{i}, w^{i}\right)$ with minimising policy $\mathbf{U}_{r}^{i}$ and $\mathbf{f}\left(\mathbf{v}^{i}, w^{i}\right) \geq \mathbf{0}$, the search direction solves

$$
\hat{\mathbf{G}}\binom{\Delta \mathbf{v}_{+}}{\Delta w}=-\hat{\mathbf{f}}=-\hat{\mathbf{G}}\binom{\mathbf{v}_{+}}{w}+\binom{\hat{\mathbf{c}}_{+}}{\hat{c}_{0}} \Leftrightarrow \hat{\mathbf{G}}\binom{\mathbf{v}_{+}^{\text {new }}}{w^{\text {new }}}=\binom{\hat{\mathbf{c}}_{+}}{\hat{c}_{0}} .
$$

The last system of equations coincides with the policy evaluation system in Sec 5.3.3. Let the affine expressions in (5.10), as defined by $\mathbf{U}_{r}^{i}$, be denoted $\mathbf{f}\left(\mathbf{v}, w ; \mathbf{U}_{r}^{i}\right)$. Thus $\mathbf{f}\left(\mathbf{v}^{i}, w^{i}\right)=\mathbf{f}\left(\mathbf{v}^{i}, w^{i} ; \mathbf{U}_{r}^{i}\right) \geq \mathbf{0}$. Solving the Newton system for $\left(\mathbf{v}^{\text {new }}, w^{\text {new }}\right)$ means $\mathbf{f}\left(\mathbf{v}^{\text {new }}, w^{\text {new }} ; \mathbf{U}_{r}^{i}\right)=\mathbf{0}$ and step length 1 . By taking a shorter step to the next iterate $\left(\mathbf{v}^{i+1}, w^{i+1}\right)$, a convex combination of $\left(\mathbf{v}^{i}, w^{i}\right)$ and ( $\left.\mathbf{v}^{\text {new }}, w^{\text {new }}\right)$, the affine expressions defined by $\mathbf{U}_{r}^{i}$ satisfies $\mathbf{f}\left(\mathbf{v}^{i+1}, w^{i+1} ; \mathbf{U}_{r}^{i}\right) \geq \mathbf{0}$. However, $\mathbf{U}_{r}^{i+1} \neq \mathbf{U}_{r}^{i}$ may occur. In such a case the $\min []$-expression in (5.10) for some $n$ is lower for $\mathbf{U}_{r}^{i+1}$ than for $\mathbf{U}_{r}^{i}$, i.e.
$f_{n}\left(\mathbf{v}^{i+1}, w^{i+1}\right)=f_{n}\left(\mathbf{v}^{i+1}, w^{i+1} ; \mathbf{U}_{r}^{i+1}\right) \geq f_{n}\left(\mathbf{v}^{i+1}, w^{i+1} ; \mathbf{U}_{r}^{i}\right) \geq 0$. Hence the nonnegativity, $\left(\mathbf{v}^{i+1}, w^{i+1}\right) \in \mathcal{F}_{+}$, is kept for arbitrary step lengths in [0,1]. According to Lemma 5.7, $F$ is convex there, and the convergence follows, e.g. by the damped Gauss-Newton theorem, [ibid, p 504].

Thm 5.1 is also an argument for the convergence of policy iteration:
Corollary 5.2: With possible exception for the initial point $\left(\mathbf{v}^{0}, w^{0}\right)$, policy iteration is entirely performed in $\mathcal{F}_{+}$.

Proof: A full Newton step is always taken, in the notations from Thm 5.1 implying

$$
f_{n}\left(\mathbf{v}^{i+1}, w^{i+1}\right)=f_{n}\left(\mathbf{v}^{i+1}, w^{i+1} ; \mathbf{U}_{r}^{i+1}\right) \geq f_{n}\left(\mathbf{v}^{i+1}, w^{i+1} ; \mathbf{U}_{r}^{i}\right)=0 .
$$

An Armijo based step length $\sigma$ reduction is based on the linear approximation

$$
F^{\text {new }} \approx \hat{F}+\sigma \cdot \nabla \hat{F}^{T}(\Delta \mathbf{v}, \Delta w)=\hat{\mathbf{f}}^{T} \hat{\mathbf{f}}+\sigma \cdot\left(\hat{\mathbf{G}}^{T} \hat{\mathbf{f}}\right)^{T}\left(-\hat{\mathbf{G}}^{-1} \hat{\mathbf{f}}\right)=(1-\sigma) \cdot \hat{F} .
$$

The step length reduction is controlled by three parameters: the number of reduction steps $J_{A}$, the target reduction $p_{A}$ and the reduction rate per step $q_{A}$. Starting from $\sigma=1$ and $F^{0} \equiv F^{\text {new }}, \mathbf{v}^{0} \equiv \mathbf{v}^{\text {new }}$ for reduction steps $j=0, \ldots J_{A}-1$ we break as soon as $F^{j} \leq \hat{F}+p_{A} \cdot q_{A}{ }^{j} \cdot(0-\hat{F})$ and otherwise update $\mathbf{v}^{j+1}=\mathbf{v}^{j}+\left(1-q_{A}\right) \cdot\left(\mathbf{v}^{j}-\hat{\mathbf{v}}\right)$.

In practice the convergence of Newton's method is not expected to cause any trouble, even if the start point is outside $\mathcal{F}_{+}:$As soon as a full step is taken, Cor 5.1 applies - the succeeding iterates will be in $\mathcal{F}_{+}$. Otherwise, the full step is reduced, because at least one $\mathbf{f}$-component has a high positive value. For the accepted reduced step length, the Armijo criterion guarantees a reduction of the objective $F$-value, and every $f_{n}\left(\mathbf{v}^{i}, w^{i} ; \mathbf{U}_{r}^{i}\right)<0$ will get a less negative $f_{n}\left(\mathbf{v}^{i+1}, w^{i+1} ; \mathbf{U}_{r}^{i}\right)$. Convergence follows as for general descent methods.

### 5.3.6 Length distributions

We start from the works solution $\mathbf{U}_{r}^{*}$ found by any of methods for the dual $r$-subproblem, in Secs 5.3.2-5.3.5. In any non-degenerate case of the average cost MDP, $\mathbf{U}_{r}^{*}$ is the optimum solution and the corresponding length distributions are found in the start iteration of the LPprimal (5.6). Also the degenerate cases are taken care of -cf . Sec 5.2.3. In principle it is possible to handle the trouble in two steps:

- determine the road length distribution $\hat{\lambda}_{r \infty}^{c}$ within each comm.class $c$, provided that (5.5e) is replaced with a condition that $c$ shall occupy the whole length $\operatorname{Len}_{r}$ (and without the redundant non-negativity $(5.5 \mathrm{~g})$ ), and compute the corresponding class- $c$ $\operatorname{cost} h_{r \infty}^{c}:=\frac{1}{1-d} \cdot \hat{c}_{r}^{c^{T}} \hat{\lambda}_{r \infty}^{c}$.
- The resulting knapsack problem is to choose the optimum shares $\left(\alpha^{c}\right)_{c}$ among the closed comm-classes: $\underset{\left(\alpha^{c}\right)_{c}}{\operatorname{minimise}} \sum_{c} h_{r \infty}^{c} \cdot \alpha^{c}$ subject to $\left\{\begin{array}{cc}\sum_{c} \alpha^{c}=1 & \\ \alpha^{c} \geq 0 & \forall c .\end{array}\right.$ The optimum solution is to give the cheapest class full share 1 .

In practice the two steps can be taken care of in one and the same LP problem. Contrary to Sec 5.3.4 we now perform full (non-simplified) LP-iterations, if necessary.

In the discounted MDP case an implemented option is to approximate the Cesaro limit length distribution. We work with two length series $\left(\lambda_{r}^{(K)}\right)_{K=0}^{\infty}$ and $\left(\bar{\lambda}_{r}^{(K)}\right)_{K=0}^{\infty}$, starting from $\bar{\lambda}_{r}^{(0)}=\mathbf{S}_{r}^{(0)}=\lambda_{r}^{(0)}=\lambda_{r}^{0}$.
For $K=1,2, \ldots:\left\{\begin{array}{c}\lambda_{r}^{(K)}:=\mathbf{P}_{r}\left(\mathbf{U}_{r}^{*}\right)^{T} \lambda_{r}^{(K-1)} \\ \mathbf{S}_{r}^{(K)}:=\mathbf{S}_{r}^{(K-1)}+\lambda_{r}^{(K)} \\ \bar{\lambda}_{r}^{(K)}:=\frac{1}{K+1} \mathbf{S}_{r}^{(K)}\end{array}\right.$
We stop as soon as two consecutive iterates $\bar{\lambda}_{r}^{(K-1)}, \bar{\lambda}_{r}^{(K)}$ are sufficiently close.

### 5.4 Results

The models and methods above were implemented and run on the Värmland data set, where 29 road classes are active. We used a discretization of $L=6$ levels per state dimension, totally 7776 states. In the policy evaluation steps this figure is also the number of linear equations to solve. In these sparse systems, the non-zero matrix elements are approximately $0.4 \%$ out of theoretically $60 \cdot 10^{6}$. After system triangulation typically $1.4 \%$ non-zero elements result. We used a fixed annual budget of 74.7 MSEK in all the runs. Two cases were considered: the discounted MDP together with penalised state violation and the average cost MDP with absolute acceptance limits. In the discounted MDP case we then determined the (Cesaro- or ordinary) limit length distribution in Sec 5.2.2; in the average cost MDP case the primal LPproblem (5.6) was solved a posteriori. The works types and works extents were discretised into totally 20 maintenance options: one for routine and nineteen for major maintenance. The road class computations were split among four parallel processors. Five dual iterations were performed, i.e. five different $v$-values were tried, in the process of finding an appropriate dual price for the budget constraint(s).

The main results are found in Tab 5.1. The wall clock time in sec is registered. The number of dual $r$-subproblems are one per road class $r$ and $v$-value. The dual $r$-iterations denote the iterations in each $r$-subproblem. The primal $r$-problems refer to the length distribution routines, applied to the resulting dual $r$-subproblem solution after five $v$-iterations. In the average cost MDP case no primal (a posteriori) LP-iteration was needed; just a calculation of the stationary lengths. The discounted MDP is much easier to solve than average cost MDP. As for the algorithms described in Secs 5.3.2-5.3.5, the value iteration method is superior in the
discounted MDP case. We tried two start solutions, based on zero cost and least cost, respectively, as described in Sec 5.3.2. This choice is arbitrary. We also tried a switch to the policy iteration method, as soon as the road class $r$ policy $\mathbf{U}_{r}$ was constant for three $r$ subproblem iterations. In the discounted MDP case this option lead to $35 \%$ longer run times. From the striking differences in Tab 5.1 between switch and no switch, as to the number of dual $r$-subproblem iterations, we realise that most time is spent on getting the correct values when almost all optimal policy components are found. In the average cost MDP case we received no convergence for the value iteration method (after four hours), unless a switch to policy iteration was performed. With this mixed strategy the value iteration method outperforms full policy iteration also in the average cost MDP case. In all the runs only one closed comm.class was detected.

As for our LP-iteration method in Sec 5.3.4, we tried three strategies for the number of simultaneously entering basic variables. If full policy iteration corresponds to $100 \%$ of the candidates entering (according to negative reduced costs), we also tried 80,90 and $95 \%$. In the discounted MDP case the run times were worse than for policy iteration for all these strategies, as Tab 5.1 shows (for 90 and $95 \%$ ). Apart from that, the penalisation option (used in the discounted MDP case) means a risk of extrapolation, i.e. negative interpolation weights. This is encountered if routine maintenance is applied to a state on the highest discretization level in at least one of the state dimensions current IRI, current rutting and current Age. Such negative weights may lead to negative lengths, infeasible in the LP-primal. Since no checking of nonnegativity is performed in our simplified LP-iterations these may lead to infeasible final solutions. In fact this occurred for one particular state in the 90 and $80 \%$ runs for one and two road classes, respectively, corresponding to the highest level in all the three critical state dimensions (and the second level in the two remaining dimensions). In such cases, pure chance controls whether the iteration process should lead to a feasible subproblem optimum or not. Although the few failures may be handled manually, as we have done for the $90 \%$ run in Tab 5.1, this is not acceptable for a general routine - if the penalisation option is used. In the average cost MDP case (been run with absolute limits, i.e. without extrapolation), the LPiteration method is up to $25 \%$ faster than full policy iteration - for the best tried strategy $95 \%$ but $11 \%$ slower than the combined value-policy iteration strategy.

Also for our second extension of policy iteration, the Newton method, the results are decent. In the discounted MDP case we tried different vales of the three Armijo parameters defined in Sec 5.3.5. Whereas the results were rather insensitive to different reduction target values $p_{A}$, as

Tab 5.1 shows, the best values of the maximum number of steps $J_{A}$ and the reduction rate $q_{A}$ were found to be related through a common size of the maximum reduction $q_{A}{ }^{J_{A}}$. The wall time reduction was $24 \%$ for the best parameter set, in comparison to ordinary policy iteration, and the reduction of the number of dual subproblem iterations $20 \%$. In the average cost MDP case the corresponding figures were $11 \%$ and $19 \%$, respectively. Here the given maximum number of Armijo steps $J_{A}=5$ was never utilised: e.g., in the run with $p_{A}=0.5, q_{A}=0.45$ we registered 1 out of 555 cases of 2 reduction steps ( $=0.2 \%$ ) and 66 cases of 1 reduction steps ( $=12 \%$ ), usually in the initial Newton iteration. Thus the $12.2 \%$ cases of deviations from policy iteration had an amplified general (positive) influence on the total number of $r$ -
subproblem iterations. As for the convergence of Newton's method, we checked both the average cost and discounted MDP models, starting from points in the nonnegativity region $\mathcal{F}_{+}$ and outside. The run time differences were confined to a few seconds.

| $\begin{aligned} & \text { MDP } \\ & \text { type } \end{aligned}$ | Method type | Parameter No. 1 | Parameter No. 2 | Param No. 3 | $\begin{array}{r} \text { wall clock } \\ \text { time }(\mathrm{s}) \end{array}$ | \#Dual subproblems | \#Dual $r$-iter.s mean value | \#Dual $r$-iter.s std dev | \#Primal problems | \#Primal $r$-iter.s mean value | \#Primal $r$-iter.s std dev |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| discounted | Value | zero cost | no switch |  | 2800 | 145 | 611,26 | 11,13 | 145 | 340,86 | 144,72 |
| discounted | Value | zero cost | switch to Policy |  | 3816 | 145 | 36,57 | 7,91 | 145 | 340,86 | 144,72 |
| discounted | Value | least cost | no switch |  | 2798 | 145 | 610,26 | 11,13 | 145 | 340,86 | 144,72 |
| discounted | Value | least cost | switch to Policy |  | 3820 | 145 | 35,57 | 7,91 | 145 | 340,86 | 144,72 |
| discounted | Policy |  |  |  | 7049 | 145 | 7,26 | 0,70 | 145 | 340,86 | 144,72 |
| discounted | LP | enter 90\% |  |  | (10471) | 145 | 8,29 | 0,73 | 145 | 340,86 | 144,72 |
| discounted | LP | enter 95\% |  |  | 9306 | 145 | 8,06 | 0,63 | 145 | 340,86 | 144,72 |
| discounted | Newton | $\mathrm{p}_{\mathrm{A}}=0.3$ | $\mathrm{q}_{\mathrm{A}}=0.45$ | $\mathrm{J}_{\mathrm{A}}=5$ | 5382 | 145 | 5,82 | 0,77 | 145 | 340,86 | 144,72 |
| discounted | Newton | $\mathrm{p}_{\mathrm{A}}=0.6$ | $\mathrm{q}_{\mathrm{A}}=0.45$ | $J_{A}=5$ | 5414 | 145 | 5,88 | 0,82 | 145 | 340,86 | 144,72 |
| discounted | Newton | $\mathrm{p}_{\mathrm{A}}=0.6$ | $\mathrm{q}_{\mathrm{A}}=0.35$ | $J_{A}=5$ | 6217 | 145 | 5,65 | 1,34 | 145 | 340,86 | 144,72 |
| discounted | Newton | $\mathrm{p}_{\mathrm{A}}=0.6$ | $\mathrm{q}_{\mathrm{A}}=0.55$ | $J_{A}=5$ | 5530 | 145 | 6,10 | 0,78 | 145 | 340,86 | 144,72 |
| discounted | Newton | $\mathrm{p}_{\mathrm{A}}=0.5$ | $\mathrm{q}_{\mathrm{A}}=0.05$ | $J_{A}=2$ | 9442 | 145 | 6,99 | 6,48 | 145 | 340,86 | 144,72 |
| discounted | Newton | $p_{\text {A }}=0.5$ | $\mathrm{q}_{A}=0.1$ | $J_{A}=2$ | 5732 | 145 | 5,93 | 0,75 | 145 | 340,86 | 144,72 |
| discounted | Newton | $\mathrm{p}_{\mathrm{A}}=0.5$ | $\mathrm{q}_{\mathrm{A}}=0.25$ | $J_{A}=2$ | 6133 | 145 | 6,20 | 1,00 | 145 | 340,86 | 144,72 |
| average | Value | zero cost | no switch |  | >14400 |  |  |  |  |  |  |
| average | Value | zero cost | switch to Policy |  | 9339 | 145 | 99,86 | 39,33 | 145 | 0,00 | 0,00 |
| average | Value | least cost | no switch |  | >14400 |  |  |  |  |  |  |
| average | Value | least cost | switch to Policy |  | 9430 | 145 | 108,59 | 82,76 | 145 | 0,00 | 0,00 |
| average | Policy |  |  |  | 14427 | 145 | 4,65 | 0,53 | 145 | 0,00 | 0,00 |
| average | LP | enter 80\% |  |  | 14304 | 145 | 5,91 | 0,29 | 145 | 0,00 | 0,00 |
| average | LP | enter 90\% |  |  | 12028 | 145 | 4,59 | 0,49 | 145 | 0,00 | 0,00 |
| average | LP | enter 95\% |  |  | 10859 | 145 | 4,05 | 0,32 | 145 | 0,00 | 0,00 |
| average | Newton | $\mathrm{p}_{\mathrm{A}}=0.5$ | $\mathrm{q}_{\mathrm{A}}=0.45$ | $J_{A}=5$ | 13105 | 145 | 3,83 | 0,53 | 145 | 0,00 | 0,00 |
| average | Newton | $\mathrm{p}_{\mathrm{A}}=0.5$ | $\mathrm{q}_{\mathrm{A}}=0.6$ | $J_{A}=5$ | 13303 | 145 | 3,76 | 0,52 | 145 | 0,00 | 0,00 |
| average | Newton | $\mathrm{p}_{\mathrm{A}}=0.5$ | $\mathrm{q}_{\mathrm{A}}=0.3$ | $J_{A}=5$ | 13360 | 145 | 3,94 | 0,64 | 145 | 0,00 | 0,00 |
| average | Newton | $p_{\text {A }}=0.7$ | $\mathrm{q}_{\mathrm{A}}=0.45$ | $J_{A}=5$ | 13449 | 145 | 3,94 | 0,58 | 145 | 0,00 | 0,00 |
| average | Newton | $p_{A}=0.3$ | $q_{A}=0.45$ | $J_{A}=5$ | 12904 | 145 | 3,79 | 0,50 | 145 | 0,00 | 0,00 |
| average | Newton | $\mathrm{p}_{\mathrm{A}}=0.5$ | $\mathrm{q}_{\mathrm{A}}=0.1$ | $J_{A}=2$ | 16692 | 145 | 4,86 | 1,88 | 145 | 0,00 | 0,00 |
| average | Newton | $\mathrm{p}_{\mathrm{A}}=0.5$ | $\mathrm{q}_{\mathrm{A}}=0.25$ | $J_{A}=2$ | 13694 | 145 | 4,02 | 0,74 | 145 | 0,00 | 0,00 |

Table 5.1: Results from residual values calculation with application of different solution methods.
As for the updating of the dual price $v$, both the Dantzig-Wolfe approach and the cubic interpolation are sensitive to numerical errors (non-concavity and non-symmetry, respectively). Hence only a combination of the three updating methods in Sec 5.3.1 is recommended for general use, but only if the new $v$-value is restricted to the central part of the remaining uncertainty interval (to avoid slow convergence), i.e. with the modification described in Sec 5.3.1. In the documented runs cubic interpolation is initially used, with linear interpolation as a backup, followed by DW updating. How many dual iterations are needed to determine the correct $v$-value for the applied budget? In the discounted MDP case, starting from $v=v_{\mathrm{BCR}}=1.2439$, after the five trial $v$-values in Tab 5.1 we get $\left.v \in\right] 1.6589,1.6689[$, the interval bounds corresponding to budget errors 776 and -5956 SEK ( $+0.001 \%$ and $-0.008 \%$ ), respectively. In the average cost MDP case (with absolute state acceptance limits), the correct $v$-value is not trapped after five trial values: $v$ is increased by a factor $\leq \sqrt{2}$ in each dual iteration; the final fifth value $v=4.75967$ violating the budget by 144638 SEK.

### 5.4 Results

At the end of the test series we discovered that one of the four processors worked around $40 \%$ slower than the others. This means that the registered wall clock times measure the solution time for the allocated 8 road classes, instead of all 29.

## 6 Coordinated maintenance

We have realised (cf. Tab 3.1) that the segment lengths in the VV-PMS database vary between 1 m and several km . In practice pavement works are planned and performed on several kilometres at a time. VV considers the road sectioning for such coordinated maintenance as relatively steady over time. Coordinated maintenance is optional, not mandatory, depending on the current pavement state. The number of possible maintenance projects, i.e. sets of simultaneously maintained section segments, obeys the combinatorial laws. Full consideration is impossible on the road class level, e.g. in our start routine and in NOS, see [Alviti et al (1994), (1996)], since the coordination decisions should be based on segment specific conditions and since a road section may include segments from different road classes. In Ch 4 we admitted a constant setup maintenance cost per segment, which may stand for an average of anticipated project costs. Here these averages are exchanged for the actual, condition dependent costs.

In Sec 6.1 we introduce some new concepts, e.g. the synonym component for (connected) subsection, suitable for describing the special coordination costs. The basic model for the segment oriented problem in Sec 4.1.2 is extended by a new variable type, coordination indicator variables, linking the segment specific maintenance variables to form a project for each section and time. Here the empty-set project corresponds to routine maintenance for all the section segments. In Sec 6.2 the method is described. We stick to the Lagrangean relaxation philosophy, supplying the network subproblem of Sec 4.2 .2 with a coordination subproblem for each section and time, by relaxing the project-segment linking constraints. In the primal heuristics the dual budget price for one year at a time is adjusted in an iterative procedure, for each trial budget price performing a local optimisation on behalf of coordination. Both the dual coordination subproblem and the primal local optimisation involve cost evaluation problems for the maintenance projects, where the combinatorial effects become visible. Two different cost evaluation methods, of different scope and complexity are described. In Sec 6.3 we describe a case study performed for validation purposes. Since the study is confined to one road, without a separate budget, any budget price updating is irrelevant. This means that the computer run times become modest and that the input data can easily be checked; therefore a recommendable approach for future use. Finally in Sec 6.4 we give some comments on the full network optimisation and in Sec 6.5 we present an alternative dual price updating method based on the primal heuristics (cf. Sec 4.2.1) - for the relaxed project-segment linking constraints.

### 6.1 Mathematical formulation

In the basic model (Ch 4) the major maintenance cost for a segment was considered independent of the cost for any other major operations performed simultaneously on the same road section. Here the setup cost for such coordinated pavement works is to be distributed among all the segments involved.

The pavement state and the cost mechanism control which parts of a road section to maintain simultaneously. We will not distinguish between different types of major maintenance (cf. Sec 3.3), i.e. the main choice for each segment and year is between major maintenance or not. We will admit a maintenance project to consist of a varying layer thickness along the section.

### 6.1.1 Indices and data

The cost advantages of large-scale production are quantified in Sec 3.6.1. Coordination means that several close segments are simultaneously maintained. In theory there are myriads of possible segment combinations. We assume that the given network or road is partitioned into a fixed set $\mathcal{N}$ of road sections, letting segments of one and the same section be entitled to coordination cost reductions. This is a restriction since a free sectioning would permit cost reductions also across section boundaries. On the other hand the historical coordination, as registered in the initial data, will have a unifying effect on the following state evolution (cf. Sec 4.2.2) and makes a corresponding sectioning natural.

Definition 6.1: A road section $n \in \mathcal{N}$ is an ordered, connected set of segments.
Since all computations in this chapter will be section oriented, it is often natural to enumerate the segments of the overall set $\mathcal{S}$ as a set of segments $s \in \mathcal{S}_{n}$ per section $n \in \mathcal{N}$ and apply the local segment order $i_{s} \equiv i(s)=1, \ldots,\left|\mathcal{S}_{n}\right|$. The fixed sectioning also restricts the set of candidate maintenance projects to each section:

Definition 6.2: A (potential) maintenance project $p \in \mathcal{P}_{n}$ is a set of segments $s \in \mathcal{S P}{ }_{p} \subseteq \mathcal{S}_{n}$.
For implementation purposes we characterise the maintenance projects $p \in \mathcal{P}_{n}$ of each road section $n$ by the local index $j_{p} \equiv j(p)=1, \ldots,\left|\mathcal{P}_{n}\right|$. Below we will specify the precise meaning of $j_{p}=1$, etc.

The maintenance cost for a project $p$ will depend on the closeness between its segments, especially on whether the segments are connected or not.

Definition 6.3: A section $n$ component $c \in C_{n}$ is a subsection of $n \in \mathcal{N}$.
Definition 6.4: A project $p$ component $c \in C \mathcal{P}_{p}$ is a maximum subsection of maintained segments $s \in \mathcal{S P}_{p}$, i.e. $c$ cannot be extended at any end by a segment in $p$.

For example, a section $n$ of at least $\left|\mathcal{S}_{n}\right|=5$ segments has a project $p$ with $\mathcal{S P}_{p}=\{1,2,3,5\}$.
There are two project components, $C \mathcal{P}_{p}=\{(1,2,3),(5)\}$.
For a given component $c$ we denote the set of (connected) segments by $s \in \mathcal{S C} C_{c}$.
As for the problem complexity we state two simple facts:

Theorem 6.1: For a road section $n \in \mathcal{N}$ of segment cardinality $\left|\mathcal{S}_{n}\right|$ there are $2^{\left|\mathcal{S}_{n}\right|}$ candidate projects and $\frac{1}{2}\left|\mathcal{S}_{n}\right| \cdot\left(\left|\mathcal{S}_{n}\right|+1\right)$ section components.

Proof: The segments $i_{s}=1, \ldots,\left|\mathcal{S}_{n}\right|$ can be combined freely into maintenance projects and for each segment the choice is between major maintenance or not. A (section) component is fully specified by its first and last segment. If its first segment is № 1, its last segment can be chosen in $\left|\mathcal{S}_{n}\right|$ different ways. If its first segment is № 2 , its last segment can be chosen in $\left|\mathcal{S}_{n}\right|-1$ ways. Etc.

According to Tab 3.7 and assuming given principles for the identification of the cost carrying road classes, a project $p$ is charged an AADT-dependent setup cost $K_{1, r p(p)}$ for a characteristic road class $r=r p(p)$, and each component $c \in C \mathcal{P}_{p}$ of $p$ receives an additional AADTdependent cutting cost $K_{2, r c(c)}$ for a characteristic road class $r=r c(c)$. These two contributions are summed up to the coordination cost $K_{p}$ for project $p$,

$$
K_{p}:=K_{1, r p(p)}+\sum_{c \in C \mathcal{P}_{p}} K_{2, r c(c)} .
$$

The precise meaning of characteristic road class is not specified in Ch 3 . We will discuss this matter below.

The local project enumeration order $j_{p}=1, \ldots,\left|\mathcal{P}_{n}\right|$ for section $n$ is defined by the constant 0/1-matrix $\mathbf{A}_{n}=\left(a_{i_{s} j_{p}}\right)_{s, p}:$ Let $a_{i_{s} j_{p}}=1$ if segment $s \in \mathcal{S P} \mathcal{P}_{p}, a_{i_{s} j_{p}}=0$ otherwise. To have a common enumeration method for the maintenance projects $p$, as $n$ and its number of segments $\left|\mathcal{S}_{n}\right|$ varies, we let the columns $j=j_{p}$ of $\mathbf{A}_{n}$ correspond to the binary counting order 0000, $0001,0010,0011,0100, \ldots$ with the rows in reverse order, to have the last binary digit correspond to the first segment $i_{s}=1$ along section $n$, etc, i.e.

$$
\mathbf{A}_{n}=\left(\begin{array}{cccccc}
0 & 1 & 0 & 1 & 0 & \cdots \\
0 & 0 & 1 & 1 & 0 & \cdots \\
0 & 0 & 0 & 0 & 1 & \cdots \\
0 & 0 & 0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}\right)
$$

Here the no-action option along section $n$ is included, as column $j=1$. According to Thm 6.1 a section $n$ consisting of $\left|\mathcal{S}_{n}\right|$ segments (the rows in $\mathbf{A}_{n}$ ) will generate the $2^{\left|\mathcal{S}_{n}\right|} \mathbf{A}_{n}$-columns.

### 6.1.2 Model

As mentioned we will describe the model in case of two works types, as in Sec 3.3.1. The formulation is a natural modification/extension of (4.6). We let the first component of the control variable vector $\mathbf{u}_{s t}$ by $u_{s t 1}=1$ indicate major maintenance of segment $s$ in year $t$, otherwise (routine maintenance) $u_{s t 1}=0$. Moreover, we introduce the coordination indicator variables $\left(z_{p t}\right)_{p, t}$, using $z_{p t}=1$ if maintenance project $p \in \mathcal{P}_{n}$ is chosen for road section $n$ in year $t, z_{p t}=0$ otherwise. Since the no-action option is included (as $j_{p}=1$ ) exactly one
indicator variable should be positive for each $n, t\left((6.1 \mathrm{~h})\right.$ below). The $z_{p t}$-variables are used for replacing the average setup maintenance costs $\bar{K}_{s} u_{s t 1}$ in Ch 4 ((6.1b) below) and for activating $u_{s t 1}=1$, for all segments $s \in \mathcal{S P}{ }_{p}$ that correspond to the chosen project $p((6.1 \mathrm{~g})$ below). The full model is:

$$
\begin{align*}
& \operatorname{mix}_{\left(\mathbf{x}_{s t}\right)_{s, t},\left(\mathbf{u}_{s t}\right)_{s, t},\left(z_{p t}\right)_{p, t}}^{\operatorname{minisin}} \sum_{s}\left(\sum_{t=0}^{T-1} d^{t} \cdot f_{s}\left(\mathbf{x}_{s t}\right)+d^{T} \cdot \varphi_{s T \bullet}\left(\mathbf{x}_{s T}\right)\right)+v_{\mathrm{BCR}} \cdot \sum_{t=0}^{T-1} d^{t+\frac{1}{2}} y_{t}  \tag{6.1a}\\
& \left\{\begin{array}{cc}
\sum_{r} \sum_{p \in \mathcal{P}_{n}} K_{p} \cdot z_{p t}+\sum_{s} c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right) \leq y_{t} & \forall t \\
y_{t} \leq b_{t} & \forall t \\
\mathbf{x}_{s, t+1}=\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right) & \forall s, t \\
\mathbf{x}_{s 0}=\mathbf{a}_{s} & \forall s \\
\mathbf{u}_{s t} \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s t}\right) & \forall s, t \\
\sum_{p \in \mathcal{P}_{n}} a_{i_{s} j_{p}} \cdot z_{p t}=u_{s t 1} & \forall s \in \mathcal{S}_{n}, \forall n \in \mathcal{N}, \forall t \\
\sum_{p \in \mathcal{P}_{n}} z_{p t}=1 & \forall n \in \mathcal{N}, \forall t \\
z_{p t} \in\{0,1\} & \forall p \in \mathcal{P}_{n}, \forall n \in \mathcal{N}, \forall t
\end{array}\right. \tag{6.1b}
\end{align*}
$$

### 6.2 Method

Our strategy is to extend the method for the basic model, by regarding coordinated maintenance as an optional facility in the program system. This means that we stick to the overall relaxation methodology, identifying and solving separable network subproblems by DynP.

### 6.2.1 Dual optimisation

As in the main study Ch 4 the budgetary constraints (6.1b) are relaxed, by the introduction of nonnegative Lagrangean multipliers $\left(v_{t}\right)_{t=0}^{T-1}$. In Sec 4.2.2 the dual subproblem was separable into segment specific network subproblems. For a correct treatment of the co-operative effects in a DynP approach here, every road section $n$ should be treated as an entity, defining a product state space of the state spaces of all the $\left|\mathcal{S}_{n}\right|$ segments involved. Due to the huge amount of lattice points we find such an approach out of the question. Instead, since ( 6.1 g ) is the only link between the segment $s$ specific control variables $\mathbf{u}_{s t}$ and the coordination project $p$ variables $z_{p t}$, it is natural to relax these constraints as well, introducing additional Lagrangean multipliers $\left(\mu_{s t}\right)_{s, t}$, free in sign. The multipliers are scaled (discounted), in order to reflect the time in question. By a relaxation of both (6.1b) and (6.1g) the dual objective becomes

$$
\begin{aligned}
& \Phi(\boldsymbol{\mu}, \mathbf{v}):= \\
& \min _{\left(\mathbf{x}_{s t}\right),\left(\mathbf{u}_{s t}\right),\left(z_{p t}\right)} \sum_{t=0}^{T-1} d^{t} \cdot\left(\sum_{s} f_{s}\left(\mathbf{x}_{s t}\right)+\sqrt{d} \cdot v_{t} \cdot\left[\sum_{n} \sum_{p \in \mathcal{P}_{n}} K_{p} z_{p t}+\sum_{s} c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)-y_{t}\right]\right) \\
&+\sum_{t=0}^{T-1} d^{t+\frac{1}{2}} \cdot \sum_{n} \sum_{s \in \mathcal{S}_{n}} \mu_{s t} \cdot\left[u_{s t 1}-\sum_{p \in \mathcal{P}_{n}} a_{i_{s} j_{p}} z_{p t}\right]+d^{T} \cdot \sum_{s} \varphi_{s T \bullet}\left(\mathbf{x}_{s T}\right)+v_{\mathrm{BCR}} \cdot \sum_{t=0}^{T-1} d^{t+\frac{1}{2}} y_{t}= \\
&=\min _{\left(\mathbf{x}_{s t}\right),\left(\mathbf{u}_{s t}\right),\left(z_{p t}\right)} \sum_{t=0}^{T-1} d^{t} \cdot\left(\sum_{s} f_{s}\left(\mathbf{x}_{s t}\right)+\sqrt{d} \cdot\left(\sum_{s}\left[v_{t} c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)+\mu_{s t} u_{s t 1}\right]-\sqrt{d} \cdot v_{t} y_{t}\right)\right) \\
&+\sum_{t=0}^{T-1} d^{t+\frac{1}{2}} \cdot\left(\sum_{n} \sum_{p \in \mathcal{P}_{n}}\left[v_{t} K_{p}-\sum_{s \in \mathcal{S}_{n}} \mu_{s t} a_{i_{s} j_{p}}\right] \cdot z_{p t}\right)+d^{T} \cdot \sum_{s} \varphi_{s T \bullet}\left(\mathbf{x}_{s T}\right)+v_{\mathrm{BCR}} \cdot \sum_{t=0}^{T-1} d^{t+\frac{1}{2}} y_{t}= \\
&= \min _{\left(\mathbf{x}_{s t}\right),\left(\mathbf{u}_{s t}\right),\left(z_{p t}\right)} \sum_{\mathrm{s}}\left(\sum_{t=0}^{T-1} d^{t} \cdot\left(f_{s}\left(\mathbf{x}_{s t}\right)+\sqrt{d} \cdot\left[v_{t} c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)+\mu_{s t} u_{s t 1}\right]\right)+d^{T} \cdot \varphi_{s T \bullet}\left(\mathbf{x}_{s T}\right)\right) \\
&+\sum_{n}^{T-1} \sum_{t=0}^{T+\frac{1}{2}} d^{t \in \mathcal{P}_{n}}\left[v_{t} K_{p}-\sum_{s \in \mathcal{S}_{n}} \mu_{s t} a_{s_{s} j_{p}}\right] \cdot z_{p t}+\sum_{t=0}^{T-1} d^{t+\frac{1}{2}} \cdot\left(v_{\mathrm{BCR}}-v_{t}\right) y_{t}
\end{aligned}
$$

The dual is separable according to the three main terms in the final expression:

$$
\Phi(\boldsymbol{\mu}, \mathbf{v})=: \sum_{\mathrm{s}} \Phi_{s}^{\mathrm{net}}(\boldsymbol{\mu}, \mathbf{v})+\sum_{n} \sum_{t=0}^{T-1} d^{t+\frac{1}{2}} \cdot \Phi_{n t}^{\mathrm{coo}}(\boldsymbol{\mu}, \mathbf{v})+\sum_{t=0}^{T-1} d^{t+\frac{1}{2}} \cdot\left(v_{\mathrm{BCR}}-v_{t}\right) y_{t}
$$

The dynamic, segment-specific network subproblem $s$ is:

$$
\begin{align*}
& \Phi_{s}^{\mathrm{net}}(\boldsymbol{\mu}, \mathbf{v})=\min _{\left(\mathbf{x}_{s t}\right)_{t},\left(\mathbf{u}_{s t}\right)_{t}} \sum_{t=0}^{T-1} d^{t} \cdot\left(f_{s}\left(\mathbf{x}_{s t}\right)+\sqrt{d} \cdot\left[v_{t} c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)+\mu_{s t} u_{s t 1}\right]\right)+d^{T} \varphi_{s T \bullet}\left(\mathbf{x}_{s T}\right)  \tag{6.2a}\\
& \text { subject to }\left\{\begin{array}{cl}
\mathbf{x}_{s, t+1}=\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right) & \forall t \\
\mathbf{x}_{s 0}=\mathbf{a}_{s} \\
\mathbf{u}_{s t} \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s t}\right) & \forall t .
\end{array}\right. \tag{6.2b}
\end{align*}
$$

Formulation (6.2) is identical to (4.8), except for the relaxation term $\mu_{s t} u_{s t 1}$, which replaces $v_{t} \bar{K}_{s} u_{s t 1}$ and can be viewed as the fair proportion of the realised project setup cost $K_{p}$ attributable to segment $s \in \mathcal{S}_{n}$. However, note that there is no explicit link to the road sectioning. The two formulations are equally handled - cf. the DynP backward and forward routines in Ch 4.

The coordination subproblem $n, t$ (assuming $v_{t}>0$ ) is:

$$
\begin{align*}
& \Phi_{n t}^{\mathrm{coo}}(\boldsymbol{\mu}, \mathbf{v})=\min _{\left(z_{p t}\right)_{p \in \mathcal{P}_{n}}} v_{t} \cdot \sum_{p \in \mathcal{P}_{n}}\left[K_{p}-\sum_{s \in \mathcal{S}_{n}} \frac{\mu_{s t}}{v_{t}} a_{i_{s} j_{p}}\right] \cdot z_{p t}=: v_{t} \cdot \sum_{p \in \mathcal{P}_{n}} k_{p t} \cdot z_{p t}  \tag{6.3a}\\
& \text { subject to }\left\{\begin{array}{l}
\sum_{p \in \mathcal{P}_{n}} z_{p t}=1 \\
z_{p t} \in\{0,1\} \quad p \in \mathcal{P}_{n} .
\end{array}\right. \tag{6.3b}
\end{align*}
$$

This subproblem has a trivial solution. Introducing vector notations, letting $\mathbf{e}_{p}$ denote the unit vector in coordinate direction $p$ of the $\left|\mathcal{P}_{n}\right|$-dimensional real space and letting $\mathbf{z}_{n t}$ have the
vector components $j_{p}=1, \ldots,\left|\mathcal{P}_{n}\right|$ corresponding to $\left(z_{p t}\right)_{p \in \mathcal{P}_{n}}$, the optimum is written

$$
\mathbf{z}_{n t}=\left\{\begin{array}{cc}
\mathbf{0} & \text { if } k_{p t} \geq 0 \text { for every } p \in \mathcal{P}_{n}  \tag{6.4}\\
\mathbf{e}_{p^{*}} & \text { otherwise; where } p^{*} \in \underset{p \in \mathcal{P}_{n}}{\arg \min _{p t}} k_{p t} .
\end{array}\right.
$$

Let us compare the segment oriented model with the road class oriented start routine. In the start routine, here and in Ch 4, the dual prices $\left(\mu_{s t}\right)_{s, t}$ are replaced by fixed costs $\left(m_{r}\right)_{r}$. Since the start routine handles length variables, all unit costs are expressed in SEK/m. The proxy $m_{r}$ is an a priori estimate of the optimal setup and coordination costs per metre for the road class $r$ maintenance projects. We use

$$
m_{r}:=\left(K_{1 r}+K_{2 r} \cdot n C_{r}\right) / \lambda \mathcal{P}_{r}
$$

where $n C_{r}, \lambda \mathcal{P}_{r}$ denote estimates of the average number of components and of the average length, respectively, for all chosen projects involving road class $r$. The estimates are taken from historical data or from test runs of the main routine. The setup costs $K_{1 r}, K_{2 r}$ are defined in Sec 3.6.1.

As for the updating of the dual variables $\boldsymbol{\mu}, \mathbf{v}$, their characters and scopes are quite different. Whereas the budget prices $\mathbf{v}$ affect all maintenance and segments, $\boldsymbol{\mu}$ has an influence merely on the coordination cost savings. Therefore we have implemented a two-level updating, on the outer iteration level updating $\mathbf{v}$ and on the inner iteration level updating $\boldsymbol{\mu}$, for given $\mathbf{v}$. On both levels we use subgradient technique as the standard updating method. The stop criterion on the inner level should depend on the advancement of the outer process, avoiding full optimisation in the early $\mathbf{v}$-iterations. But observe that both the network and the coordination subproblems must be resolved for each dual trial point $(\boldsymbol{\mu}, \mathbf{v})$.

If only a subnet is studied, as we will in Sec 6.3 below, we assume that the budgetary dual prices $\mathbf{v}=\left(v_{t}\right)_{t}$ are fixed. This means a simplification of the method, in the sense that the coordination dual prices $\boldsymbol{\mu}=\left(\mu_{s t}\right)_{s, t}$ alone are updated in the dual routine. One dual iteration runs:

- The dual iteration number is given. For given $\boldsymbol{\mu}, \boldsymbol{v}$, solve the dual network and coordination subproblems.
- Determine a primally feasible solution (as to (6.1g)), according to the heuristics in Sec 6.2 .2 below but without any $v_{t}$-adjustments in year $t$.
- Increment the dual iteration number and find a new trial $\boldsymbol{\mu}$-iterate by subgradient technique.


### 6.2.2 Primal heuristics

At first we present the strategy. In order to satisfy the two sets of dually relaxed constraints in (6.1) we have chosen a two-level approach: As for the budget restrictions (6.1b) we keep to the idea from Sec 4.2.3 of adjusting the dual price for one year at a time, until the budget level is met. In our implementation the adjustment is made in an outer routine, based on information
from all the segments. Unlike the situation in Ch 4, where the correct price could be directly determined, the presence of the relaxed coordination constraints $(6.1 \mathrm{~g})$ will necessitate an iterative procedure - however in only one variable $v_{t}$ for time $t$. The reason is that the maintenance projects (and costs) to be determined in the inner routine may vary with $v_{t}$. Since the dually relaxed link $(6.1 \mathrm{~g})$ is limited to each road section, the inner procedure is separable into (primal) heuristics subproblems for each $n, t$ - to be solved for each trial $v_{t}$-value: For fixed $t$ we determine compatible $\left(z_{p t}\right)_{p \in \mathcal{P}_{n}}$ and $\left(\mathbf{u}_{s t}\right)_{s \in \mathcal{S}_{n}}$ in one and the same heuristics subproblem.

For the generation of feasible maintenance plans in the inner routine we have implemented two alternatives, in fact two different cost evaluation methods. The first alternative handles a general coordination cost structure, at the price of a computational work that increases exponentially by $\left|\mathcal{S}_{n}\right|$. The second - DynP - approach, suggested by [Lindberg et al (1997)], presumes that the costs can be recursively computed but has the advantage of a quadratically increasing computational work. The conflict can be exemplified by the cost structure as described in Ch 3: If the overall AADT-dependent coordination cost of a maintenance project $p$ would be interpreted, e.g., as the average or maximum for the maintained segments $s \in \mathcal{S P}_{p}$ instead of an average or maximum for the section segments $s \in \mathcal{S}_{n}$ (our implementation), we have a non-recursive case that excludes our second approach. But DynP also applies if, e.g., the first segment of a project is made representative to carry the AADT-related project setup cost $K_{1 r}$.

Now we regard the heuristics details. The approach is common to both cost evaluation methods. From a given state reached in the beginning of year $t$, step one year forward by performing a local optimisation in year $t$, relying on the future costs after year $t$ as computed for the lattice points in the segment $s$ specific dual backward routines.

- The primal iteration is $i$. For a given price $v_{t}=v_{t}^{i}$ determine the total optimal maintenance costs $G_{t}=G_{t}^{i}$ for the network, in view of coordination.
- Stop iterating if $b_{t}-\varepsilon<G_{t}^{i} \leq b_{t}, \quad v_{t}^{i}>v_{\mathrm{ret}}$ or $G_{t}^{i} \leq b_{t}, \quad v_{t}^{i}=v_{\mathrm{ret}}$.
- Otherwise increment the iteration number $i:=i+1$ and choose a new trial value $v_{t}^{i}$ by linear or nonlinear interpolation of $\left(v_{t}^{i^{i}}, G_{t}^{i^{\prime}}\right)_{i^{\prime}<i}$, in order to solve the equation $G_{t}^{i}=b_{t}$.

Here we focus on the first point. We will show that the primal problem to solve, given $\mathbf{v}$, is trivial - but that we have to distinguish two cases. At first we determine, as in Ch 4, the optimal works extent (layer thickness) $u_{s t 2}^{1}$ for major maintenance $\mathbf{u}_{s t}^{1}$ (of type $u_{s t 1}=u_{s t 1}^{1}=1$ ) from the reached segment state $\mathbf{x}_{s t}$, by relying on the current $\mathbf{v}$. We use the same objective as in the dual DynP forward routine but, since major maintenance is presumed, the optimal layer thickness $u_{s t 2}^{1}$ is independent of the dual $\boldsymbol{\mu}$-term (distributed setup cost) and indeed

### 6.2 Method

independent of whether coordination is applied or not. Let $\hat{c}_{s t}$ denote the cost difference (without any coordination dependent setup cost in year $t$ ) in state $\mathbf{x}_{s t}$ between performing major maintenance $\mathbf{u}_{s t}^{1}$ and not, i.e. routine maintenance $\mathbf{u}_{s t}^{0}$ (of type $u_{s t 1}=u_{s t 1}^{0}=0$ ). Since the traffic cost $f_{s}\left(\mathbf{x}_{s t}\right)$ is common to both choices, $\hat{c}_{s t}$ concerns the relaxed direct maintenance $\operatorname{costs} c_{s}$ in year $t$ and the succeeding optimal costs $\Phi_{s, t+1}^{\mathrm{net}}$ after year $t$. The succeeding costs $\Phi^{\text {net }}$ are determined in (6.2) by the DynP backward routine, cf. Ch 4, i.e. by the consideration of all coordination options (through $\boldsymbol{\mu}$ ). After rescaling, for an easier final expression, $\hat{c}_{s t}$ is defined (for $v_{t}>0$ ) from

$$
\begin{aligned}
& v_{t} \cdot \hat{c}_{s t}= \\
& =v_{t} \cdot\left[c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{1}\right)-c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{0}\right)\right]+d \cdot\left(\Phi_{s, t+1}^{\mathrm{net}}\left[\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{1}\right)\right]-\Phi_{s, t+1}^{\mathrm{net}}\left[\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{0}\right)\right]\right)
\end{aligned}
$$

where $\Phi_{s t}^{\text {net }}\left(\mathbf{x}_{s t}\right)$ denotes the optimal future cost from state $\mathbf{x}_{s t}$. But the local optimisation problem here is to consider the earlier relaxed link between the segment specific dual subproblems $s \in \mathcal{S}_{n}$ and the coordination subproblem for $n, t$. Thus the primal $n, t$ subproblem in the given state $\mathbf{x}_{s t}$ is:

$$
\begin{align*}
& \underset{\left.\left(z_{p t}\right)_{p \in \mathcal{P}_{n}} \operatorname{minimize}_{s t}\right)_{s \in \mathcal{S}_{n}}}{\min } v_{t} \cdot\left(\sum_{p \in \mathcal{P}_{n}} K_{p} \cdot z_{p t}+\sum_{s \in \mathcal{S}_{n}} \hat{c}_{s t} u_{s t 1}\right)  \tag{6.5a}\\
& \text { subject to }\left\{\begin{array}{cc}
\sum_{p \in \mathcal{P}_{n}} a_{i_{s} j_{p}} z_{p t}=u_{s t 1} & s \in \mathcal{S}_{n} \\
\sum_{p \in \mathcal{P}_{n}} z_{p t}=1 & \\
\mathbf{u}_{s t} \in \mathcal{U}_{r(s)}\left(\mathbf{x}_{s t}\right) & s \in \mathcal{S}_{n} \\
z_{p t} \in\{0,1\} & p \in \mathcal{P}_{n} .
\end{array}\right. \tag{6.5b}
\end{align*}
$$

The corresponding matrix-vector formulation, without the given scale factor $v_{t}$, is written

$$
\underset{\mathbf{z}_{n t}, \mathbf{u}_{n t 1}}{\operatorname{minimize}} \mathbf{K}_{n}{ }^{T} \mathbf{z}_{n t}+\hat{\mathbf{c}}_{n t}{ }^{T} \mathbf{u}_{n t 1}
$$

In case the $\mathbf{u}_{n t 1}$-part of ( $6.5 \mathrm{~d}^{\prime}$ ) stands for the standard binary $u_{s t 1}$-constraints $\mathbf{u}_{n t 1} \in\{0,1\}^{\left|\mathcal{S}_{n}\right|}$, without further restrictions due to violated acceptance limits, the simple structure of $\mathbf{A}_{n}$ and ( $6.5 \mathrm{c}^{\prime}$ ) makes ( $6.5 \mathrm{~d}^{\prime}$ ) redundant. Eliminating $\mathbf{u}_{n t 1}$ from ( $6.5 \mathrm{a}^{\prime}$ ) by the use of ( $6.5 \mathrm{~b}^{\prime}$ ) gives

$$
\begin{equation*}
\underset{\mathbf{z}_{n t}}{\operatorname{minimize}}\left(\mathbf{K}_{n}+\mathbf{A}_{n}^{T} \hat{\mathbf{c}}_{n t}\right)^{T} \mathbf{z}_{n t}=: \hat{\mathbf{k}}_{n t}^{T} \mathbf{z}_{n t} \tag{6.5a"}
\end{equation*}
$$

subject to $\left\{\begin{array}{c}\mathbf{1}^{T} \mathbf{z}_{n t}=1 \\ \mathbf{z}_{n t} \in\{0,1\}^{\left|P_{n}\right|} .\end{array}\right.$

This is the same type of problem as the dual coordination subproblem $n, t$ (6.3), with $\hat{\mathbf{k}}$ replacing $\mathbf{k}$, and has the trivial solution according to (6.4).

The result (6.4) presumes that the binary control variables $\left(u_{s t 1}\right)_{s \in \mathcal{S}_{n}}$ may vary freely. According to ( 6.5 d ) there are situations where the $u_{s t 1}$-value is restricted, namely at an absolute upper bound for some state dimension $d$ variable $x_{s t, d}$. To have stronger dual results we should make the constraints ( 6.5 b )-(6.5f) state dependent, by restricting the set of candidate projects $p$ in case any upper bound is reached, making $\mathbf{u}_{s t}^{0}=\mathbf{0}$ infeasible. By introducing the section state $\mathbf{X}_{n t}:=\left(\mathbf{x}_{s t}\right)_{s \in \mathcal{S}_{n}}$ and the set of feasible maintenance projects

$$
\Omega_{n}\left(\mathbf{X}_{n t}\right):=\left\{p \in \mathcal{P}_{n} \mid a_{i_{s} j_{p}}=0 \quad \forall s \in \mathcal{S}_{n}: \mathbf{0} \notin \mathcal{U}_{r(s)}\left(\mathbf{x}_{s t}\right)\right\}
$$

the reformulation of (6.5) runs

$$
\begin{align*}
& {\underset{\left(z_{p t} t\right.}{ } \operatorname{minimize}_{n}}^{p \in \mathcal{P}_{n}} \sum_{p \in \mathcal{C}}\left(K_{1, r p(p)}+\sum_{c \in \mathcal{P}_{p}} K_{2, r c(c)}+\sum_{s \in \mathcal{S} \mathcal{P}_{p}} \hat{c}_{s t} a_{i_{s} j_{p}}\right) \cdot z_{p t}=: \sum_{p \in \mathcal{P}_{n}} \hat{k}_{p t} \cdot z_{p t}  \tag{6.5a"'}\\
& \text { subject to }\left\{\begin{array}{cc}
\sum_{p \in \Omega_{n}\left(\mathbf{X}_{n t}\right)} a_{i_{s} j_{p}} \cdot z_{p t}=u_{s t 1} & \forall s \in \mathcal{S}_{n}, \forall n, t \\
\sum_{p \in \Omega_{n}\left(\mathbf{X}_{n t}\right)} z_{p t}=1 & \forall n, t \\
z_{p t} \in\{0,1\} & \forall p \in \Omega_{n}\left(\mathbf{X}_{n t}\right), \forall n, t .
\end{array}\right. \tag{6.5b"'}
\end{align*}
$$

In practice this means that all $u_{s t 1}$-restrictions are taken care of in advance, according to the reached state. The general optimum is

$$
\mathbf{z}_{n t}=\left\{\begin{array}{cc}
\mathbf{0} & \text { if } \hat{\mathbf{k}}_{n t} \geq \mathbf{0} \text { and if } \mathbf{z}_{n t}=\mathbf{0} \text { is feasible }  \tag{6.6}\\
\mathbf{e}_{p^{*}} & \text { otherwise, where } p^{*} \in \arg \min _{p \in \Omega_{n}\left(\mathbf{x}_{n t}\right)} \hat{k}_{p t} .
\end{array}\right.
$$

The primal heuristics is the strongest possible, in the following sense:
Theorem 6.2: If, hypothetically, the dual solution $(\boldsymbol{\mu}, \boldsymbol{v})$ is primally feasible, i.e. also satisfies $(6.1 \mathrm{~b}, \mathrm{~g})$, the primal heuristics will (in any non-degenerate case) produce the same (optimal) solution.

Proof: As for the $\mathbf{v}$-part, the stop criterion ( $2^{\text {nd }}$ point above) in the heuristics means that no iteration is performed, if the dual solution is feasible. As for $\boldsymbol{\mu}$ consider the dual subproblems associated with an arbitrary index pair $n, t$. Let $\varphi_{s t m}\left(\mathbf{x}_{s t}\right), \Phi_{s t}^{\text {net }}\left(\mathbf{x}_{s t}\right)=\min _{m} \varphi_{s t m}\left(\mathbf{x}_{s t}\right)$ denote the optimal costs from state $\mathbf{x}_{s t}$ and onwards, for applying maintenance type $u_{s t 1}=m$ and overall, respectively. This is the situation in the DynP forward routine, for a computed optimal layer thickness, notation $\mathbf{u}_{s t}^{m}$ for type $m$. Thus we have

$$
\begin{align*}
& \varphi_{s t m}\left(\mathbf{x}_{s t}\right)=f_{s}\left(x_{s t}\right)+\sqrt{d} \cdot\left[v_{t} \cdot c_{s}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{m}\right)+\mu_{s t} \cdot u_{s t 1}^{m}\right]+d \cdot \Phi_{s, t+1}^{\mathrm{net}}\left[\mathbf{h}_{r(s)}\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}^{1}\right)\right] \quad m=0,1, \\
& \Phi_{s t}^{\mathrm{net}}\left(\mathbf{x}_{s t}\right)=\min _{m} \varphi_{s t m}\left(\mathbf{x}_{s t}\right)=\varphi_{s t 0}\left(\mathbf{x}_{s t}\right)+\sqrt{d} \cdot \min _{m}\left[\left(v_{t} \hat{c}_{s t}+\mu_{s t}\right) \cdot u_{s t 1}^{m}\right] . \tag{6.7}
\end{align*}
$$

Assume that the dual and the heuristics partial solutions coincide up to time $t$ and state $\left(\mathbf{x}_{s t}\right)_{s}$. Then, by construction, the heuristics will find the very same optimum thickness $u_{s t 2}$ for $m=1$
as the dual does, and the cost differences $\hat{c}_{s t}$ will be uniquely determined (i.e. will be the same for the dual and the heuristics).
Let $\mathbf{u}_{n t 1} \times \mathbf{z}_{n t} \in \mathcal{M}_{1}, \mathbf{u}_{n t 1} \in \mathcal{M}_{2}$ and $\mathbf{z}_{n t} \in \mathcal{M}_{3}$ denote the feasible domains defined by constraints ( $6.1 \mathrm{~b}-\mathrm{i}),(6.2 \mathrm{~b}-\mathrm{d})$ and $(6.3 \mathrm{~b}-\mathrm{c})$, respectively. Due to $(A)$ : the separability of the dual subproblems and $(B)$ : especially for feasible dual solutions (as to $(6.1 \mathrm{~b}, \mathrm{~g})$ ) we have

$$
\begin{aligned}
& \sum_{s \in \mathcal{S}_{n}} \Phi_{s t}^{\mathrm{net}}\left(\mathbf{x}_{s t}\right)+\Phi_{n t}^{\mathrm{coo}}=\sum_{s \in \mathcal{S}_{n}} \min _{\mathbf{u}_{n t} \in \mathcal{M}_{2}} \varphi_{s t, u_{s t 1}}\left(\mathbf{x}_{s t}\right)+\min _{\mathbf{z}_{n t} \in \mathcal{M}_{3}} \sum_{p \in \mathcal{P}_{n}}\left(v_{t} K_{p}-\sum_{s \in \mathcal{S}_{n}} \mu_{s t} a_{i_{s} j_{p}}\right) z_{p t}= \\
& \stackrel{(A)}{=} \sum_{s \in \mathcal{S}_{n}} \varphi_{s t 0}\left(\mathbf{x}_{s t}\right)+\min _{\left(\mathbf{u}_{n t 1}, \mathbf{z}_{n t}\right) \in \mathcal{M}_{2} \times \mathcal{M}_{3}}\left(\sum_{s \in \mathcal{S}_{n}}\left(v_{t} \hat{c}_{s t}+\mu_{s t}\right) \cdot u_{s t 1}+\sum_{p \in \mathcal{P}_{n}}\left(v_{t} K_{p}-\sum_{s \in \mathcal{S}_{n}} \mu_{s t} a_{i_{s} j_{p}}\right) \cdot z_{p t}\right) \\
& \stackrel{(B)}{\leq} \sum_{s \in \mathcal{S}_{n}} \varphi_{s t 0}\left(\mathbf{x}_{s t}\right)+v_{t} \cdot \min _{\left(\mathbf{u}_{n t 1}, \mathbf{z}_{n t}\right) \in \mathcal{M}_{1}}\left(\sum_{s \in \mathcal{S}_{n}} \hat{c}_{s t} u_{s t 1}+\sum_{p \in \mathcal{P}_{n}} K_{p} z_{p t}\right) .
\end{aligned}
$$

Apart from a constant, corresponding to the "mandatory" cost of routine maintenance, the last expression coincides with the primal problem (6.5). Since a feasible dual solution satisfies (B) with equality, the dual solution solves the corresponding primal problem as well.
Thus the dual and heuristics partial solutions will coincide also for the next time period. Due to a given, common starting point for $t=0$ the theorem is proved by induction.

Finally we will comment on the two different cost evaluation methods. The issue is the heavy computational work caused by $\left|\mathcal{P}_{n}\right|$ increasing exponentially by $\left|\mathcal{S}_{n}\right|$, according to Thm 6.1. As for our first, general-cost heuristics we realise from ( $6.5 \mathrm{a} \mathrm{a}^{\prime \prime}$ ) that the component based setup costs, i.e. the first sum over $c \in C \mathcal{P}_{p}$, are needed. However, the second sum, over $s \in \mathcal{S P}{ }_{p}$, can be omitted, by first forming $\hat{K}_{c}:=K_{2, r c(c)}+\sum_{s \in \mathcal{S C} C_{c}} \hat{c}_{s t}$ for each $c \in C_{n}$, and then summing up $\hat{K}_{c}$ over $c$ for each $p \in \mathcal{P}_{n}$; which takes $\left|C \mathcal{P}_{p}\right|$ additions for $p$, and $\left(\left|\mathcal{S}_{n}\right|+1\right) \cdot 2^{\left|\mathcal{S}_{n}\right|-2}$ additions overall, in case all maintenance projects are feasible.

The same computation strategy applies to the dual coordination subproblems (6.3) too.
Our second, recursive-cost heuristics is to use DynP for solving (6.5) as well. The assumption is that the project specific setup cost $K_{1, r p(p)}$ is the same for all $p \in \mathcal{P}_{n}$, i.e. there is a characteristic road class $r$ such that $r=r p(p) \forall p \in \mathcal{P}_{n}$. The network in Fig 6.1 illustrates the shortest route optimisation, here from node 0 to node 4 b , in case $\left|\mathcal{S}_{n}\right|=4$ and all potential projects are feasible. There are two node levels, the upper one (a-level) characterising the first segment $s_{a}$ of each component and the lower one (b-level) containing the last segment $s_{b}$. The downward $\operatorname{arc} c=\left(s_{a}, s_{b}\right)$ carries the component specific cost $\hat{K}_{c}$. The (dashed) upward arcs, admitting a new component to start after one segment of routine maintenance, are free of charge. Also the horizontal arcs are free, except for the $K_{1 r}$-charged leading arc. The extra zero-cost arc is for the no-action option. As for the arithmetic operations, both the formation of the arc costs $\hat{K}_{c}$ and the DynP solution takes magnitude $\left|C_{n}\right|$ additions, i.e. according to Thm 6.1, $2\left|\mathcal{C}_{n}\right|=\left|\mathcal{S}_{n}\right| \cdot\left(\left|\mathcal{S}_{n}\right|+1\right)$ additions in all. For, e.g., $\left|\mathcal{S}_{n}\right|=20$ this modest increase by
$\left|S_{n}\right|$ means magnitude $4 \cdot 10^{2}$ operations for the DynP approach vs. $10^{6}$ operations for the general-cost method. In practice most administrative road sections of a few km may be managed in this way. The relatively small setup costs make longer segments of mere academic interest. (To avoid that several paths describe one and the same maintenance plan, all the horizontal arcs on the lower level, except the last one, can be replaced by a single 0 -cost downward (4a, 4b)-arc.)


Figure 6.1 Primal heuristics: Cost evaluation by DynP.
If, instead, we interpret the concept characteristic road class as if the first segment of each project $p$ is to carry the project setup cost $K_{1, r(p)}$, then the costs of the upper horizontal arcs have to be revised: $\operatorname{Cost}(0,1 \mathrm{a})=K_{1, r\left(s_{1}\right)}$ where $i_{s_{1}}=1, \operatorname{Cost}(1 \mathrm{a}, 2 \mathrm{a})=K_{1, r\left(s_{2}\right)}-K_{1, r\left(s_{1}\right)}$ where $i_{s_{2}}=2$, etc. (In this case all horizontal arcs on the lower level are needed.)

### 6.3 Case study: road 63

### 6.3.1 Run strategy

A validation study was performed on road 63 , similar to the corresponding study for the basic model in $\operatorname{Sec} 4.5$. As a preparation step, relevant future costs at the time horizon and budgetary Lagrangean multipliers for each year were computed, by running the residual value and start routines for the whole Värmland network to near-optimum. Then as in $\operatorname{Sec} 4.5$, the budgetary dual prices $\mathbf{v}=\left(v_{t}\right)_{t}$ were frozen, while running the main routine one dual iteration for road 63, now with the coordination program-facility activated.

We use the same road data as in Sec 4.5. Without any given administrative coordination road sectioning in the VV-PMS database we have to assume such a sectioning here, based on the registered pavement works and initial states. Such an ad hoc sectioning must be a compromise, since ideally each section should be short, excluding any combination of distant sub-sections from cost reductions, and the section boundaries should be natural, not cutting off any sub-
section of coordination potential. For the purpose we have used an $\left|\mathcal{S}_{n}\right|$-limitation to between 6 and 11 consecutive segments per road section $n$, by manually judging the pavement conditions. Contrary to the corresponding runs in Ch 4 , now the segments should carry the fully correct coordination costs, according to the data model in Ch 3.

### 6.3.2 Results

As with the basic model in Ch 4 a special validation meeting was held, with participation of project group members from VV and VTI, with the purpose of scrutinizing various output data. The meeting was held soon after a first version of the coordination routine was implemented. We used the same two means of output as in Ch 4, tables and diagrams, for presenting the state, maintenance and cost for each segment and year. But now the result was after a $\boldsymbol{\mu}$ optimisation for each coordination section, for a horizon of 21 yrs. The subgradient method made use of order 100 dual iterations per section. The stop criterion was based on the remaining gap between the best primal and dual solutions. We noticed some cases of resulting negative gaps, which we ascribe to discretization errors - cf. Sec 4.3.4. With the run times dominated by the backward and forward routines of the dual network subproblems, any comparison between the two cost evaluation methods described in Sec 6.2 .2 was considered less meaningful, as having a marginal effect on the CPU-times here.

Two runs, Run3Abs and Run4Rel, were presented, corresponding to absolute and relative state variable limits, respectively. The discussion focused on Run4Rel, since the penalisation option of state violations was considered more realistic than the direct prohibition option. The diagrammatic results for approximately the same two parts of road 63 as in Figs 4.5-4.6 are shown in Fig 6.2. Now the accumulated road length - up to and including the segment in question - from the start point in Karlstad is presented (in m) at the segment number, together with the road class. The road parts in Figs 6.2a,c (starting from accumulated road length 0) cover three coordination sections, consisting of segments 2233-2243, 2251-2260 and 22612271, and Figs 6.2b,d (starting from length 16144) show two sections, of segments 2292-2300 and 2301-2308 (2310), out of the totally 24 sections defined for road 63. In the diagrams we see examples of all kinds of coordinated major maintenance (wide vertical lines), involving: all segments in the section, one sub-section of connected segments (component), several components and a single segment.

The average number of major maintenance operations is somewhat less than for the corresponding run in Ch 4 , although the relative layer thickness (short horizontal line) in most cases is at the lowest, as before. This was still the main objection to our results. For, e.g., road class n6 VV expects a pavement works period of around 10 years, whereas Run4Rel includes 2 - 4 major operations in 21 years. However, as we have pointed out (in Sec 4.5.2), our network optimisation often (and also in Run4Rel) means that the average Age-value actually increases from the initial value found in the database (cf. Fig 4.3b). The same trend also applies to the average Age-value immediately before a major operation. Anyhow, the main concern is of course that several thin pavement layers do not accomplish the same quality as one thick layer does. We have tried to meet the main objection in three different ways: As a result of the
validation meeting the differentiated lower limits for the layer thickness (see Sec 3.3.3) were introduced, depending on the relative Age-value immediately before the major operation. Another consequence was the introduction of the return rate restrictions (see Sec 3.7.2). In Run4Rel the budgetary dual prices were around 0.3 after yr 10, i.e. much lower than $v_{\mathrm{BCR}}=1.2439$. This model change will make a significant difference. Let us look at the four occasions of single-segment maintenance in Fig 6.2 and consult the corresponding table results (not shown)! Neither of the pavement works for segment 2263 in yr 6, for 2260 in yr 19, for 2294 in yr 18 and for 2303 in yr 20 would have been feasible in a corresponding situation, in presence of the constraints $v_{t} \geq v_{\mathrm{BCR}}$ for each year $t$. The IRI-value for 2263 in yr 6 is much worse than for the rest of the segments in the coordination section (see Fig 6.2a) and triggers a major operation here. Yet another incentive for repeated pavement works is the upper bounds for the two state variables IRI degradation rate and rutting degradation rate, restricted to rates at least as good as before the operation (cf. Secs 3.5.2 and 3.5.4). These constraints modify/replace the regression effect models also for the cheapest major maintenance, permitting the rates that have reached the lower bounds, here 0.02 and 0.3 , respectively, to remain there at low costs. In all the runs made after the meeting the upper bound restrictions were neglected.

The maintenance plans for a few segments in Run4Rel drew special attention. Why does segment 2295 get such a high future cost for routine maintenance in yr 3 ? The output table shows that the rut depth value is 19.95 at the start of yr 3 and will reach 20.56 at the end of the year, violating the penalisation limit 20.0. This means that the future cost for routine maintenance will be penalised and explains the high value. However, it also raises a new question of how to choose a reasonable penalisation. Here the sum of the relative violations, i.e. state variable value over acceptance limit minus 1, is penalised linearly, by a user defined factor (apart from segment traffic and length). Maybe the chosen factor 0.1 is too high and/or maybe VV would prefer a nonlinear penalty model.

Why do the plans and the initial future costs for segments 2300 and 2301 differ so much? Both are n6-segments. The main reason for the different plans is that they belong to two different coordination sections, since the plans are influenced by the states of all segments in the corresponding section. The initial IRI-degradation rate is worse for 2301 (cf. Fig 6.2b), whereas the initial Age-value is higher for 2300 (cf. Fig 6.2d). Thus 2300 reaches the penalisation Age-limit 17.0 in the major operation year 5 (and also in the two major operation years 15 and 21), generating high future costs there. In this respect segment 2300 is a representative for the unwanted thin layer solution, repeatedly triggered by its Age-value or (in major operation yr 8 ) by the section behaviour. (The unrealistically dense major operations for the section as a whole may be due to a non-optimal solution and/or to the low budgetary dual prices.) For segment 2301 routine maintenance in the major operation year 5 would lead fairly close (value 19.1) to the rutting penalisation limit 20. Our backward interpolation routine means that the future costs also for such close-to penalised variable values are affected by the neighbouring true penalised states. Moreover, the rather high IRI-values for many segments in the coordination section speak for a major operation in yr 5 . Now for the different initial future costs, $1461 \mathrm{SEK} / \mathrm{m}$ in yr 0 for 2301 but just $362 \mathrm{SEK} / \mathrm{m}$ for 2300 . The difference is partly
linked to the different segment lengths 2 m and 435 m , respectively, since these costs include the future distributed project setup costs $\left(\mu_{s t} / \text { Len }_{s}\right)_{t \geq 0}$, on average higher for a short segment, like the $L e n_{s}=2 \mathrm{~m}$ segment $s=2301$. However, the total predicted future cost is just 2922 SEK for 2301 but 157470 SEK for 2300 . For 2301 the future costs computed in yrs $0,1, \ldots$ vary more than expected. Although 2301 is the far most extreme segment among all in this respect, we see the same variation tendency for some other short segments. In Sec 4.3 .4 we received the most extreme discretization errors for the shortest segments. Our only explanation is that the short segments are more prone to such errors, due to the distributed setup costs, here and in the basic model, since the cost variation (differences between routine and major maintenance) becomes greater. A comfort is that their influence on the total cost is modest.

We discovered that one segment, 2279 (not shown in Fig 6.2), had a small negative maintenance cost, $-4 \mathrm{SEK} / \mathrm{m}$, computed in yr 3 . Also this has to do with the distribution of setup costs and the shortness of $2279,1 \mathrm{~m}$. If, e.g., a major operation is cost-efficient in yr 3 for each of the two surrounding segments, then a component setup cost can be saved by including the intermediate segment 2279 in one connected sub-section (component). This is recognised by negative coordination dual prices $\left(\mu_{s t}\right)_{s}$ for some of the segments $s$ in the section. (The purpose of these dual prices is to guarantee that the correct cost is paid by the project.) In this case 7 out of 10 section segments have negative dual prices but a negative net cost is received only for the extremely short segment 2279 , where the distributed coordination cost dominates the direct maintenance cost per metre. (But the dual price is very high for one of the two surrounding segments, with a rut depth violating the penalisation limit in yr 3.)

In summary the validation meeting led to several model improvements, although the main objection of repeated thin pavement layers needs further treatment.

Is the coordination routine really needed? Let us compare Run4Rel in Fig 6.2 and Run3Abs in Fig 4.5, although the run assumptions are somewhat different. The occurrences of coordinated pavement works are similar and do not motivate the extra $\boldsymbol{\mu}$-iteration routine. However, this is just one example. And a coordination run has the obvious advantage of using the correct coordination costs, not just some standard values.


Figure 6.2a,b Run4Rel. Two sections of road 63. Time evolution of $I R I$-values, according to grey-scale.


Figure 6.2c,d Run4Rel. Two sections of road 63. Time evolution of Age-values, according to grey-scale.

### 6.4 Discussion

In principle, a complete regional network, e.g. Värmland, can be handled like road 63 in Sec 6.3. An appropriate sectioning can be made semi-automatically, in consideration of the initial state and maintenance history of every segment in the VV-PMS database. However, in such a case a full optimisation (i.e. including $\mathbf{v}$-updating) is tempting, meeting the corresponding regional budget. What run times are to be expected? Compared to a run of the basic model, as documented in Ch 9 below, each budget price iteration $i$, for fixed $\mathbf{v}=\mathbf{v}^{i}$, now corresponds to the computations of Sec 6.3:

- Solve the dual $\max _{\boldsymbol{\mu}} \Phi\left(\boldsymbol{\mu}, \mathbf{v}^{i}\right)$ approximately, e.g. by performing a constant number $I$ of $\boldsymbol{\mu}$-iterations according to Sec 6.2.1.
- Generate primally feasible solutions, as to $(6.1 \mathrm{~b}, \mathrm{~g})$ and according to $\operatorname{Sec} 6.2 .2$, at least for the last $\boldsymbol{\mu}$-iteration. (In Sec 6.3 only ( 6.1 g ) had to be considered.)

Therefore we anticipate the run times to become at most $I$ times the values found in Ch 9 . This is the only way to decide, e.g., how good the $\mathbf{v}$-results from the start routine are - and sanction their use on separate roads like road 63.

### 6.5 An alternative dual price updating method

As pointed out in Sec 6.3 the initial use of the standard subgradient method needed quite many dual ( $\boldsymbol{\mu}$-)iterations before an acceptable primal-dual gap was reached, each iteration meaning that the run time consuming network subproblem had to be resolved. Therefore we also formulated and implemented an alternative for the $\boldsymbol{\mu}$-updating, using the special problem structure as we did for the primal heuristics in updating $\mathbf{v}$ (cf. Sec 6.2.2). Whereas we have to guess the solution of an updated network subproblem the solution of the coordination subproblem is guaranteed. The additional computational work is limited to cost evaluation. The idea is based on the observation that for any section $n$, any time $t$ and any subproblem solution $\left(\mathbf{x}_{s t}, \mathbf{u}_{s t}\right)_{s \in \mathcal{S}_{n}}, p^{*} \equiv p_{n t}^{*}$ (such that $z_{p^{*} t}=1$ ) of (6.2) and (6.3), respectively, the RHS - LHS $(s, t)$-difference in $(6.1 \mathrm{~g})$ takes three values only:

$$
\begin{equation*}
u_{s t 1}-\sum_{p \in \mathcal{P}_{n}} a_{i_{s} j_{p}} z_{p t}=u_{s t 1}-a_{i_{s} j_{p^{*}}} \in\{-1,0,+1\} \quad s \in \mathcal{S}_{n} \tag{6.8}
\end{equation*}
$$

Consistency between the network and coordination subproblems means coincident major maintenance or coincident routine maintenance, i.e. 0-differences for each segment $s \in \mathcal{S}_{n}$. For time $t$ we therefore update the ( 6.1 g )-dual price $\mu_{s t}$ only for the non-zero difference segments $s$, i.e. part of the $\boldsymbol{\mu}_{n t}$ vector. We split $\mathcal{S}_{n}$ according to (6.8) into three sign sets

$$
\mathcal{S}_{n t}^{i}:=\left\{s \in \mathcal{S}_{n} \mid u_{s t 1}-a_{i_{s} j_{p^{*}}}=i\right\} \text { for } i=-, 0,+(\text { short for } i=-1,0,+1)
$$

By how much does $\mu_{s t}$ have to be changed in order to accomplish a transition for segment $s$ from $\mathcal{S}_{n t}^{-}$or $\mathcal{S}_{n t}^{+}$to $\mathcal{S}_{n t}^{0}$ ? We want to avoid a switch from $\mathcal{S}_{n t}^{-}$to $\mathcal{S}_{n t}^{+}$or vice versa, i.e. exactly one of the two members (sides) in $(6.1 \mathrm{~g})$ for $s$ should be affected.

- As for the effect on the network subproblem (6.2), an increased $\mu_{s t}$-value will favour a transition from major to routine maintenance, i.e. from $u_{s t 1}=1$ to $u_{s t 1}=0$, and a decreased $\mu_{s t}$-value the opposite transition. No other segment is directly influenced, but for $s$ itself also the solutions for times $t^{\prime} \neq t$ are. However, (6.8) implies that there is a threshold effect meaning that $\mathbf{u}_{s t}$ (and the maintenance for $t^{\prime} \neq t$ ) is unchanged in a $\mu_{s t}$-interval. The switch value can be predicted by comparing the future costs for the two maintenance types $m:=u_{s t 1}=0,1$ in the given state $\mathbf{x}_{s t}$. By using the expressions for the future costs $\varphi_{s t m}\left(\mathbf{x}_{s t}\right)$ in (6.7), determined for $m=0,1$ in the forward routine, and by letting $\Delta_{s t}$, $\sigma_{s t}$ denote the prospective absolute $\mu_{s t}$-change and its sign, respectively, the anticipated switch value corresponds to equal costs

$$
\varphi_{s t 1}\left(\mathbf{x}_{s t}\right)+\sqrt{d} \cdot u_{s t 1}^{1} \cdot \sigma_{s t} \Delta_{s t}=\varphi_{s t 0}\left(\mathbf{x}_{s t}\right)+\sqrt{d} \cdot u_{s t 1}^{0} \cdot \sigma_{s t} \Delta_{s t}
$$

i.e. identifying the maintenance type notations $u_{s t 1}^{m}=m$ the decisive absolute change is

$$
\begin{equation*}
\Delta_{s t}:=\sigma_{s t} \cdot\left(\varphi_{s t 0}\left(\mathbf{x}_{s t}\right)-\varphi_{s t 1}\left(\mathbf{x}_{s t}\right)\right) / \sqrt{d} \tag{6.9}
\end{equation*}
$$

Here the fixed sign fulfils

$$
\sigma_{s t}:= \begin{cases}-1 & s \in \mathcal{S}_{n t}^{-}  \tag{6.10}\\ +1 & s \in \mathcal{S}_{n t}^{+}\end{cases}
$$

- The $(s, t)$-difference in (6.8) can also be changed in the coordination $(n, t)$-subproblem by a sufficiently strong change of $\mu_{s t}$. If $\mu_{s t}$ is increased, it favours any project $p$ such that $s \in \mathcal{S P}{ }_{p}$. On behalf of the project component $c$ based cost $K_{2, r c(c)}$ such an increase may also lead to adjacent segments becoming part of the optimal project. In the same way a decreased $\mu_{s t}$-value may lead to both $s$ and its adjacent segments leaving the optimal project $p_{n t}^{*}$. A change of $\left(\mu_{s t}\right)_{s \in \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}}$affects the costs as follows. Consider the set of projects $\mathcal{P}_{n t}^{ \pm}$corresponding to the different segment subsets of $\mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}$, i.e.

$$
p \in \mathcal{P}_{n t}^{ \pm} \Leftrightarrow \mathcal{S} \mathcal{P}_{p} \subseteq \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+} .
$$

Theorem 6.3: For arbitrary project pairs $p^{\prime}, p^{\prime \prime} \in \mathcal{P}_{n}$, the different segment subsets of $\mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}$define an equivalence relation $R$ :

$$
p^{\prime} R p^{\prime \prime} \Leftrightarrow \mathcal{S} \mathcal{P}_{p^{\prime}} \cap\left(\mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}\right)=\mathcal{S P}_{p^{\prime \prime}} \cap\left(\mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}\right)
$$

Proof: This follows directly from the definition of equivalence relation, see e.g. [Luenberger (1979), p 249].

In an equivalence class the project $p \in \mathcal{P}_{n t}^{ \pm}$, as satisfying $\mathcal{S P}_{p} \cap\left(\mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}\right)=\mathcal{S} \mathcal{P}_{p}$, is the natural class representative:

$$
p R p^{\prime} \Rightarrow \mathcal{S} \mathcal{P}_{p} \subseteq \mathcal{S} \mathcal{P}_{p^{\prime}} \quad \forall p^{\prime} \in \mathcal{P}_{n}
$$

For $p^{\prime} \in \mathcal{P}_{n}$ the modified cost in (6.3a) for an absolute deviation $\boldsymbol{\delta}_{n t}$ from the current $\mu_{n t}$ is

$$
K_{p^{\prime}}-\sum_{s \in \mathcal{S}_{n}} \frac{\mu_{s t}+\sigma_{s t} \delta_{s t}}{v_{t}} \cdot a_{i_{s} j_{p^{\prime}}}=k_{p^{\prime} t}-\sum_{s \in \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}} \delta_{s t} \cdot \frac{\sigma_{s t}}{v_{t}} \cdot a_{i_{s} j_{p^{\prime}}}
$$

In the final expression the $\boldsymbol{\delta}_{n t}$-dependent $2^{\text {nd }}$ term is a constant for each equivalence class. Hence an a priori minimisation of the $1^{\text {st }}$ cost term can be made within each class:

$$
\begin{equation*}
\bar{k}_{p t}:=\min _{p^{\prime} \in \mathcal{P}_{n}: p^{\prime} R p} k_{p^{\prime} t} \quad \forall p \in \mathcal{P}_{n t}^{ \pm} \tag{6.11}
\end{equation*}
$$

After that only the projects $p \in \mathcal{P}_{n t}^{ \pm}$have to be distinguished. The restriction of $p_{n t}^{*}$ in $\mathcal{P}_{n t}^{ \pm}$is denoted $\bar{p}^{*} \equiv \bar{p}_{n t}^{*}$ (i.e. $\mathcal{S P}_{\bar{p}^{*}} \subseteq \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}$).

Before formulating the simple $\boldsymbol{\mu}_{n t}$-updating algorithm we will motivate the introduction (6.10) of the signs $\left(\sigma_{s t}\right)_{s \in \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}}$of the changes. They provide a unified form when we consider the cost deviations from the currently optimal project $p_{n t}^{*}$. Just observe:

$$
a_{i_{s} j_{\bar{p}^{*}}}=\left\{\begin{array}{ll}
1 & s \in \mathcal{S}_{n t}^{-} \\
0 & s \in \mathcal{S}_{n t}^{+}
\end{array} \quad \text { and } \quad a_{i_{s} j_{p}}=\left\{\begin{array}{cc}
1 & s \in \mathcal{S P}_{p} \\
0 & s \in\left(\mathcal{S P} \mathcal{P}_{p}\right)^{c}
\end{array} \quad p \in \mathcal{P}_{n t}^{ \pm},\right.\right.
$$

where ${ }^{c}$ denotes the complementary set.
In consideration of (6.10) the cost difference between some project $p \in \mathcal{P}_{n t}^{ \pm}$and $\bar{p}^{*}$ is

$$
\begin{align*}
& \left(\bar{K}_{\bar{p}^{*}}-\sum_{s \in \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}} \delta_{s t} \cdot \frac{\sigma_{s t}}{v_{t}} \cdot a_{i_{s} j_{\bar{p}^{*}}}\right)-\left(\bar{K}_{p}-\sum_{s \in \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}} \delta_{s t} \cdot \frac{\sigma_{s t}}{v_{t}} \cdot a_{i_{s} j_{p}}\right)= \\
& =\bar{K}_{\bar{p}^{*}}-\bar{K}_{p}+\sum_{s \in \mathcal{S}_{n t}^{-}} \delta_{s t} \cdot \frac{-1}{v_{t}} \cdot\left(a_{i_{s} j_{p}}-1\right)+\sum_{s \in \mathcal{S}_{n t}^{+}} \delta_{s t} \cdot \frac{+1}{v_{t}} \cdot\left(a_{i_{s} j_{p}}-0\right)= \\
& =\bar{K}_{\bar{p}^{*}}-\bar{K}_{p}+\left(\sum_{s \in \mathcal{S}_{n t}^{-} \cap\left(\mathcal{S} \mathcal{P}_{p}\right)^{c}} \delta_{s \in \mathcal{S}_{n t}^{+} \cap \mathcal{S} \mathcal{P}_{p}}+\delta_{s t}\right) \cdot \frac{1}{v_{t}} . \tag{6.12}
\end{align*}
$$

In (6.12) only the project differences matter and all the non-zero $\delta_{s t}$-coefficients are equal.

- Determine the network switch values $\Delta_{s t}$ in (6.9) for $s \in \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}$. Compute the lowest coordination cost among $p$ in (6.11) for $\delta_{s t}=\Delta_{s t}$ and register the corresponding project $p_{\Delta} \in \mathcal{P}_{n t}^{ \pm}$. (This cost evaluation corresponds to resolving the coordination subproblem as in (6.4), now restricted to projects in $\mathcal{P}_{n t}^{ \pm}$.)
- If $p_{\Delta}=\bar{p}_{n t}^{*}$ we are done: Use $\left(\Delta_{s t}\right)_{s}$. Otherwise, assuming that $p_{\Delta}$ is a unique minimiser we take a step of length $\rho$ along the direction of steepest ascent for the cost difference (6.12), i.e. orthogonally towards the hyperplane in the $\left(\delta_{s t}\right)_{s \in \mathcal{S}_{n t}^{-}} \mathcal{S}_{n t}^{+}$-space corresponding to the cost difference 0 , using

$$
\delta_{s t}=\left\{\begin{array}{cc}
\Delta_{s t}-\rho & s \in\left(\mathcal{S}_{n t}^{-} \cap\left(\mathcal{S} \mathcal{P}_{p_{\Delta}}\right)^{c}\right) \cup\left(\mathcal{S}_{n t}^{+} \cap \mathcal{S} \mathcal{P}_{p_{\Delta}}\right)  \tag{6.13}\\
\Delta_{s t} & \text { otherwise }
\end{array}\right.
$$

The step length is chosen such that $p_{\Delta}$ still provides the lowest cost in (6.11), but close to the step length boundary where it no longer does.

By registering not only $p_{\Delta}$ but also some close-to lowest-cost projects $p$ for $\boldsymbol{\delta}=\boldsymbol{\Delta}$, the step length $\rho=\rho_{\left(p_{\Delta}, p\right)}$ where the cost order between $p_{\Delta}$ and $p$ is reversed (derived as (6.12)) is
in practice making any step unnecessary (but for verification). (The numerator is 0 if the cost order between $p_{\Delta}$ and $p$ is never changed.)

In Fig 6.3 we illustrate the alternative updating method for two segments in $\mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}$, assuming $\mathcal{S}_{n t}^{+}=\{1,2\}$. The two segments generate three projects $p$ in $\mathcal{P}_{n t}^{ \pm}$, corresponding to the $\mathcal{S P}{ }_{p}$-sets, $\{1\},\{2\}$ and $\{1,2\}$, apart from $\mathcal{S P}_{\bar{p}_{n t}^{*}}=\varnothing$. Fig 6.3 shows the switch lines of the cost difference 0 in (6.12) between $\bar{p}_{n t}^{*}$ and each of the three other projects, for two different positions of the $\{1,2\}$-switch line. In the rectangle determined by the origin and $\left(\Delta_{1 t}, \Delta_{2 t}\right)$ we paint the region grey where $\bar{p}_{n t}^{*}$ provides the lowest cost, whereas the $\{1\}$-determining region is white and the $\{1,2\}$-determining region is striped.
In Fig 6.3a the $\{1,2\}$-set defines $p_{\Delta}$ and the lowest cost (6.11) at $\left(\Delta_{1 t}, \Delta_{2 t}\right)$. By stepping orthogonally to its switch line the next iterate is determined as $\left(\Delta_{1 t}-\rho, \Delta_{2 t}-\rho\right)$ for some $\rho>0$. Because it satisfies $\delta_{s t}<\Delta_{s t} \quad s=1,2$ in (6.13) we predict and hope for an unchanged solution of the next network subproblem, since $p_{\Delta}$ will result with certainty in the next coordination subproblem.
In Fig 6.3b the $\{1\}$-set defines $p_{\Delta}$. In this case $\delta_{2 t}=\Delta_{2 t}$ is held constant in (6.13) while changing $\delta_{1 t}$ downwards close to the $\{1\}$-switch line. Here we predict and hope for a switch in the next network subproblem by (6.9) to $u_{2 t 1}=0$.


Figure 6.3 An illustration of the $\boldsymbol{\mu}_{n t}$-updating by the alternative method: a Left, $\mathbf{b}$ Right.

In case of degeneracy, i.e. if the lowest cost project $p_{\Delta}$ at $\left(\Delta_{1 t}, \Delta_{2 t}\right)$ is non-unique we may choose one (with most segments) and even step orthogonally away from the cost switch line, in order to guarantee uniqueness - at the price of risking a switch from $\mathcal{S}_{n t}^{-}$to $\mathcal{S}_{n t}^{+}$or vice versa for some segments.

A bonus of the alternative method in the above form is that it also determines the optimum solution of the next coordination subproblem. In practice we may want to parameterise the method, e.g. as we did for the Prim-method in the $\mathbf{v}$-updating in Sec 4.2.1, by a common factor reducing all the chosen $\left(\delta_{s t}\right)_{s \in \mathcal{S}_{n t}^{-} \cup \mathcal{S}_{n t}^{+}}$-values, or at least the components intended to change the optimum of the network subproblem. As we did for the Prim-method we recommend the alternative method primarily as a start method, backed up by the subgradient method for the final convergence, if necessary.

## 7 Stochastic model

As for the input data we saw in Ch 3 that the deviations from part of the regression effect models are rather large, as measured by the coefficient-of-determination. This speaks for a stochastic approach. On the other hand, a result description solely in terms of state-transition probabilities may seem too abstract, e.g. to many traffic engineers, due to the lack of real maintenance plans to check for inconsistencies, etc. However, in such models, e.g. NOS [Alviti et al (1994), (1996)], an expected time evolution can be calculated, and with our approach using individual segment data for the initial state also in the road class oriented model - initial optimum decisions can be computed for each segment. A weak point of probabilistic models, that we got aware of in the initial study (Ch 2), is the difficulty to avoid inconsistencies with a huge amount of input state-transition probability data. To avoid this we have proposed (Sec 3.8.2) a transparent stochastic model with a few, controllable parameters - the standard deviations - implemented here.

In Sec 1.4 several probabilistic models were referred to, the majority restricted to discrete state spaces, whereas the state space in our segment oriented model version is continuous.
Moreover, our works options are mixed discrete/continuous. In Sec 7.1 we suggest a role in PMS for the stochastic approach. In Sec 7.2 the model is presented as a natural extension of the basic model in Ch 4. In this study we neglect the systematic influence that between-years weather variations might have on the road network. By assuming that the state evolutions for different segments are weakly linked - only through the overall budget constraints - we handle the segments by independent stochastic variables. Since the number of segments in a road network is large we assume that the law of large numbers is applicable. In Sec 7.3 the method is described. By relying on the same approach as in Ch 4 and on collective standard distributions instead of segment specific random numbers the stochastics is limited to an extended multilinear interpolation scheme in the DynP routines. In Sec 7.4 we document a comparison run between the road class oriented deterministic and stochastic models, and in Sec 7.5 we discuss possible extensions of the stochastic approach for handling systematic effects due to winter weather fluctuations, also for coordinated maintenance.

### 7.1 Introduction

Although we all agree that random phenomena affect and even control many things - the pavement conditions included - a stochastic optimisation model is not per se to be preferred: the descriptive qualities of a stochastic model might be unrealistic to achieve on present day computers, or the result differences might be negligible. We raise the last question here: does a stochastic model lead to clear systematic differences as to the optimal pavement works, states and costs?

Remember from the preceding chapters that an optimisation run of a deterministic model has a pedagogic and checkable output: for each road segment a full maintenance plan is generated, whereas a stochastic model will specify the optimal decisions on segment level for the first year only. A new plan is needed for each year, after replacing an expected evolution with fresh
measurements. In practice, the differences are less dramatic: we would not trust a deterministic model to hold for ever, but replace it after some time, thus generating revised maintenance plans based on fresh data. A combination of both model types may be advantageous, with the initial pavement works based on the results of a stochastic model, and the deterministic results presented as plausible plans, to be implemented if things develop as expected.

### 7.2 Model

We will formulate a model formally identical to the corresponding deterministic model, the difference lying in the interpretation of the quantities involved. This is partly for pedagogic reasons: we can easily switch the optimisation program between two options relying on the same model foundation, generating maintenance plans for each segment as well as somewhat more realistic stochastic results. Another, more pragmatic reason is that our grid based linear interpolation technique remains applicable.

### 7.2.1 Stochastics

As a bi-product of the regression models in Sec 3.5, e.g. for the rut depth immediately after a major maintenance operation, in Sec 3.8.2 we formulated $R D_{\text {after }} \in N\left(\mu_{R D}(\mathbf{u}, \mathbf{x}), \sigma_{R D}\right)$, motivated by the regression assumptions. By analogy, also the three other state variables $I R_{\text {after }}, \Delta I R_{\text {after }}$ and $\Delta R D_{\text {after }}$ obey Normal distributions. In total this means $D^{\prime}=4$ new parameters, i.e. four standard deviations $\sigma_{R D}$ etc, per traffic class, where $D^{\prime}$ denotes the number of stochastic state dimensions (or fewer parameters, if we turn into regression models that are common to all traffic classes, as suggested in Sec 3.9.3).

In this model we focus on the randomness of the state transitions that result from major maintenance operations. Notice that the future degradation rates $\Delta I R_{\text {after }}$ and $\Delta R D_{\text {after }}$ are part of the stochastics. The randomness of degradation by age is further treated in Discussion below. The difference from the deterministic approach is that the state immediately after a major maintenance operation now is viewed as the result of a random event, whereas the noaction (routine maintenance) evolution is deterministic, as before (according to the current degradation rates). We will assume that the $D^{\prime}$ state variable values immediately after a major maintenance operation are independently and random. This means that the multivariate Normal distribution assumed for the combined resulting state is characterised by a $D^{\prime}$-dimensional mean vector $\boldsymbol{\mu}(\mathbf{x}, \mathbf{u})$, with components $\mu_{R D}(\mathbf{x}, \mathbf{u})$, etc, and a diagonal co-variance matrix, with diagonal elements $\sigma_{R D}{ }^{2}$, etc, i.e. by $2 D^{\prime}$ parameters in all. The independence property can be used to simplify the calculations, as described in Sec 7.3 below. The actual property that we will use is the factorability of the multivariate density function, making it possible to consider one marginal distribution at a time.

### 7.2.2 Truncation of distributions

In the basic model (Chs 3-4) the regression means $\boldsymbol{\mu}_{\text {regr }}$ were supplied by ideal state constraints $\mathbf{x} \geq \mathbf{x}^{-}$, specific to each traffic class. In a stochastic model we want the resulting $\boldsymbol{\mu}(\mathbf{x}, \mathbf{u})$ to denote a mean vector, and since the restrictions $\mathbf{x} \geq \mathbf{x}^{-}$imply truncations from below of symmetric distributions, we will put symmetric upper bounds $\mathbf{x} \leq \mathbf{x}^{+}$, $x_{d}^{+}-\mu_{d}=\mu_{d}-x_{d}^{-} \quad \forall d \in D^{\prime}$. Here $\boldsymbol{\mu}(\mathbf{x}, \mathbf{u})$ is chosen such that $\mu_{d}=\max \left[\mu_{d, \text { regr }}, x_{d}^{-}\right]$. This means that in (the rare) cases where the regression model does not apply, i.e. $\mu_{d, \text { regr }}<x_{d}^{-}$, we will continue using a deterministic model $\mu_{d}=x_{d}^{-}=x_{d}^{+}$; otherwise a symmetrically truncated Normal distribution is applied on the interval $x_{d} \in\left[x_{d}^{-}, x_{d}^{+}\right]$, together with lump probabilities for the truncated distribution tails.

We will use the a priori estimated, regression standard deviations $\sigma_{R D}$, etc, as parameters of the (untruncated) Normal distributions. This means that our randomly generated probabilistic $x_{d}$-values will show a smaller variance. In order to eliminate this inconsistency the given regression effect models could be exchanged for ours at the very parameter estimation.

Ceilings $\mathbf{x}^{+}$for the state immediately after a major maintenance operation are not only our ad hoc constructions but may be useful also in practice in order to guarantee the quality of a job, as measured by the IRI- and RD-values in particular. If the road administrator and the maintenance provider agree on the stochastic model, then $\boldsymbol{\mu}$ should contain the target values and $\mathbf{x}^{+}$the acceptance limits for the resulting state. Our interpretation is that all major maintenance leading to $x_{d}>x_{d}^{+}$for some $d$ are corrected to the very limit. Such corrections will lead to some additional cost. Our interpretation is that the expected additional cost is included in the fixed (agreed) maintenance cost.

In summary we will use multidimensional intervals $\left[\mathbf{x}^{-}, \mathbf{x}^{+}\right]$for the state immediately after a major maintenance operation. In each of the $D^{\prime}$ stochastic dimensions $d$ a marginal, truncated Normal density function will apply for the stochastic variable $X_{d}$, plus surrounding lump probabilities $P\left(X_{d} \leq x_{d}^{-}\right), P\left(X_{d} \geq x_{d}^{+}\right)$.

### 7.2.3 Problem formulation

We will limit the description to the road class oriented model; the segment oriented model differs exactly as in Ch 4 . This formulation formally coincides with the basic deterministic model (4.7), for convenience repeated as (7.1) below. Remember that any optional additional cost (for violating state ceilings) is hidden in the ordinary maintenance costs $c_{s}\left(\mathbf{a}_{s}, \mathbf{u}_{s 0}\right)$ and $\bar{c}_{r}\left(\widetilde{\mathbf{x}}_{n}, \widetilde{\mathbf{u}}_{r n t}\right)$ for road segment $s$ and road class $r$, respectively. Here $\mathbf{a}_{s}$ is the given initial state of segment $s$, and $\widetilde{\mathbf{x}}_{n}$ denotes the nodal state of grid node $n$. The pavement works performed at time $t$ is written $\mathbf{u}_{s 0}(t=0)$ and $\widetilde{\mathbf{u}}_{r n t}(t>0)$, respectively. The time horizon is $2 T$. As was

### 7.2 Model

mentioned in Sec 4.1.3 the state transition matrices $\left(h_{r, s n}\left(\mathbf{a}_{s}, \mathbf{u}_{s 0}\right)\right)_{s, n}$ and $\left(\bar{h}_{r, n n^{\prime}}\left(\widetilde{\mathbf{x}}_{n}, \widetilde{\mathbf{u}}_{r n t}\right)\right)_{n, n^{\prime}}$ can be interpreted as probability distributions. Due to the return rate (benefit/cost ratio) limit $v_{\text {BCR }}$ the model is

$$
\begin{align*}
& \underset{\left(\tilde{\lambda}_{r r t}\right)_{r, n, t}\left(\mathbf{u}_{s 0}\right)_{s},\left(\widetilde{\mathbf{u}}_{r t}\right)_{r, n, t}}{\operatorname{minimis}} \sum_{r}\left(\sum_{s \cdot r(s)=r} f_{s}\left(\mathbf{a}_{s}\right) \cdot \text { Len }_{s}+\sum_{n}\left[\sum_{t=1}^{2 T-1} d^{t} f_{r n} \widetilde{\lambda}_{r n t}+d^{2 T} \boldsymbol{\varphi}_{r n, 2 T} \widetilde{\lambda}_{r n, 2 T}\right]\right)+v_{\mathrm{BCR}} \cdot \sum_{t=0}^{2 T-1} d^{t+\frac{1}{2}} y_{t} \\
& \left(\begin{array}{ll}
\sum_{r} \sum_{s: r(s)=r}\left[K_{s} u_{s 01}+c_{s}\left(\mathbf{a}_{s}, \mathbf{u}_{s 0}\right)\right] \cdot \text { Len }_{s} \leq y_{0} & t=0 \\
\sum_{r} \sum_{K_{r}}\left[\bar{u}_{r n t 1}+c_{r n}\left(\widetilde{\mathbf{u}}_{r n t}\right)\right] \cdot \tilde{\lambda}_{r n t} \leq y_{t} & \forall t>0
\end{array}\right.  \tag{7.1b}\\
& \left.\tilde{\lambda}_{r n^{\prime} 1}^{n}=\sum_{s: r(s)=r, s n^{\prime}} h_{\mathbf{a}_{s}}, \mathbf{u}_{s 0}\right) \cdot \text { Len }_{s} \quad t=0, \forall r, n^{\prime}  \tag{7.1d}\\
& \tilde{\lambda}_{r n^{\prime}, t+1}=\sum_{n} \widetilde{h}_{r, n n^{\prime}}\left(\widetilde{\mathbf{u}}_{r n t}\right) \cdot \tilde{\lambda}_{r n t} \quad \forall t>0, \forall r, n^{\prime}  \tag{7.1e}\\
& \mathbf{u}_{s 0} \in \mathcal{U}_{r(s)}\left(\mathbf{a}_{s}\right) \quad \forall s  \tag{7.1f}\\
& \widetilde{\mathbf{u}}_{r n t} \in \mathcal{U}_{r}\left(\widetilde{\mathbf{x}}_{n}\right)  \tag{7.1g}\\
& y_{t} \leq b_{t} \tag{7.1h}
\end{align*}
$$

### 7.3 Method

We will use the DynP approach for the dual subproblem here as well. Remember that in the DynP backward iterations of the deterministic models above, we base the future cost at the start of year $t$ on a node based interpolation of the optimal future cost $\varphi_{t+1}\left(\mathbf{x}_{t+1}\right)$ computed one year later, where $\mathbf{x}_{t+1}=\mathbf{h}\left(\mathbf{x}_{t}, \mathbf{u}_{t}\right)$ is the (partly continuously varying) transition end point. In the stochastic model this relationship is replaced by the expected value transition $E\left(\mathbf{x}_{t+1}\right)=\boldsymbol{\mu}\left(\mathbf{x}_{t}, \mathbf{u}_{t}\right)$ plus a multivariate Normal-based disturbance, a mixture of a truncated Normal distribution and two lump probabilities for the truncated tails. In both types of models the $\mathbf{x}_{t+1}$-contributions are referred to the surrounding nodal states and the results can be interpreted as expected future costs, but the interpolation scheme is far more ambitious here. In the deterministic model our interpolation is based on $2^{D}$ neighbouring nodes in $D$ state dimensions, whereas in the stochastic model the whole node lattice may be involved.

### 7.3.1 Computation of stochastic elements

In order to speed up the computations, we use the linear property of the interpolation method, as described in Ch 2. Consider the marginal distribution of some state variable $x$, with the corresponding $\boldsymbol{\mu}$-component $\mu$ and the standard deviation $\sigma$. Such a distribution is illustrated in Fig 7.1 below. The nodal state levels in such a stochastic dimension $d$ (implicit here) are denoted $x^{0}, x^{1}, \ldots$ and the distribution truncation bounds are $x^{-}, x^{+}$. For simpler enumerations we occasionally let $L$ denote the highest node level. For each non-nodal state $x$ the value is split between the two surrounding node levels (interpolation) or between the two nearest node levels on one side (extrapolation). Specifically, for any two consecutive node levels $x^{\ell}, x^{\ell+1}, \ell=0, \ldots, L-1$, the contributions from the intermediate states $x$ are controlled
by the linear interpolation property. Thus the weights $w\left(x ; x^{\ell}\right)$ and $w\left(x ; x^{\ell+1}\right)$ referred to, respectively, $x^{\ell}$ and $x^{\ell+1}$ satisfy

$$
\begin{equation*}
w\left(x ; x^{\ell}\right)=\left(x^{\ell+1}-x\right) /\left(x^{\ell+1}-x^{\ell}\right), \quad w\left(x ; x^{\ell+1}\right)=\left(x-x^{\ell}\right) /\left(x^{\ell+1}-x^{\ell}\right) \tag{7.2}
\end{equation*}
$$

i.e. the total contribution to level $\ell$ from the interval $x \in] x^{\ell}, x^{\ell+1}$ [ becomes

$$
W\left(\ell ; x^{\ell}, x^{\ell+1}\right)=\int_{x^{\ell}}^{x^{\ell+1}} w\left(x ; x^{\ell}\right) \cdot f(x ; \mu, \sigma) d x
$$

where $f(x ; \mu, \sigma)$ denotes the Normal density function. By introducing the standard-Normal density function $\varphi(z), x=\mu+\sigma \cdot z$, letting $x^{\ell}=\mu+\sigma \cdot z^{\ell}$, etc, be implied and defining

$$
\begin{aligned}
& \Phi\left(z^{\ell}\right):=\int_{-\infty}^{z^{\ell}} \varphi(z) d z \\
& \Psi\left(z^{\ell}\right):=\int_{-\infty}^{z^{\ell}} z \cdot \varphi(z) d z=\int_{-\infty}^{z^{\ell}} z \cdot \frac{1}{\sqrt{2 \pi}} e^{-z^{2} / 2} d z=\left[-\frac{1}{\sqrt{2 \pi}} e^{-z^{2} / 2}\right]_{-\infty}^{z^{\ell}}=-\varphi\left(z^{\ell}\right),
\end{aligned}
$$

the contribution to level $\ell$ turns into

$$
\begin{align*}
& W\left(\ell ; x^{\ell}, x^{\ell+1}\right)=\int_{x^{\ell}}^{x^{\ell+1}} \frac{\left(x^{\ell+1}-\mu\right)-(x-\mu)}{x^{\ell+1}-x^{\ell}} \cdot f(x ; \mu, \sigma) d x=\int_{z^{\ell}}^{z^{\ell+1}}\left(\frac{x^{\ell+1}-\mu}{x^{\ell+1}-x^{\ell}}-\frac{\sigma \cdot z}{x^{\ell+1}-x^{\ell}}\right) \cdot \varphi(z) d z= \\
& \quad=\frac{x^{\ell+1}-\mu}{x^{\ell+1}-x^{\ell}} \cdot\left(\Phi\left(z^{\ell+1}\right)-\Phi\left(z^{\ell}\right)\right)-\frac{\sigma}{x^{\ell+1}-x^{\ell}} \cdot\left(\Psi\left(z^{\ell+1}\right)-\Psi\left(z^{\ell}\right)\right) \tag{7.3a}
\end{align*}
$$

The corresponding contribution to level $\ell+1$ from $x \in] x^{\ell}, x^{\ell+1}$ [ is

$$
\begin{equation*}
W\left(\ell+1 ; x^{\ell}, x^{\ell+1}\right)=\Phi\left(z^{\ell+1}\right)-\Phi\left(z^{\ell}\right)-W\left(\ell ; x^{\ell}, x^{\ell+1}\right) \tag{7.3b}
\end{equation*}
$$

Introducing interval bounds $\pm \infty$ to incorporate the discrete probability contributions, in case $x^{-} \in\left[x^{0}, x^{1}\right]$ the total weights at levels 0 and 1 reflect the lump probability at $x^{-}$:

$$
\begin{align*}
& W\left(0 ;-\infty, x^{1}\right)=\frac{x^{1}-\mu}{x^{1}-x^{0}} \cdot\left(\Phi\left(z^{1}\right)-\Phi\left(z^{-}\right)\right)-\frac{\sigma}{x^{1}-x^{0}} \cdot\left(\Psi\left(z^{1}\right)-\Psi\left(z^{-}\right)\right)+\frac{x^{1}-x^{-}}{x^{1}-x^{0}} \cdot \Phi\left(z^{-}\right) \\
& W\left(1 ;-\infty, x^{1}\right)=\Phi\left(z^{1}\right)-W\left(0 ;-\infty, x^{1}\right) \tag{7.4}
\end{align*}
$$

(If the discretization accompanies the ideal state, as $x^{0} \equiv x^{-}$, (7.4) can be slightly simplified.) In case $x^{+} \in\left[x^{L-1}, x^{L}\right]$ the total weights at levels $L-1 \& L$ reflect the lump probability at $x^{+}$

$$
\begin{align*}
& W\left(L-1 ; x^{L-1}, \infty\right)= \\
& =\frac{x^{L}-\mu}{x^{L}-x^{L-1}} \cdot\left(\Phi\left(z^{+}\right)-\Phi\left(z^{L-1}\right)\right)-\frac{\sigma}{x^{L}-x^{L-1}} \cdot\left(\Psi\left(z^{+}\right)-\Psi\left(z^{L-1}\right)\right)+\frac{x^{L}-x^{+}}{x^{L}-x^{L-1}} \cdot\left(1-\Phi\left(z^{+}\right)\right) \\
& \quad W\left(L ; x^{L-1}, \infty\right)=1-\Phi\left(z^{L-1}\right)-W\left(L-1 ; x^{L-1}, \infty\right) \tag{7.5}
\end{align*}
$$

In a stochastic dimension $d$ the outcome $x \equiv x_{d \text {,after }}$ fulfils $x \in\left[x_{d}^{-}, x_{d}^{+}\right]$. In order to get correct averages we will apply extrapolation whenever $x \notin\left[x_{d}^{0}, x_{d}^{L}\right]$. Extrapolation due to $x<x_{d}^{0}$ will not occur, since $x_{d}^{-}$is a given ideal value, and the lowest discretization level $x_{d}^{0}$ for the road
class is a priori chosen such that $x_{d}^{0} \leq x_{d}^{-}$. However, the extrapolation case $x>x_{d}^{L}$ cannot be ruled out, since $\mu_{d, \text { regr }}$ will vary according to the state immediately before the major maintenance operation and the layer thickness, and thus will $x_{d}^{+}$as well. Any state $x$ above $x_{d}^{L}$ corresponds to weight factors (cf. (7.2), index $d$ implicit)

$$
w\left(x ; x^{L-1}\right)=\left(x^{L}-x\right) /\left(x^{L}-x^{L-1}\right)<0, \quad w\left(x ; x^{L}\right)=\left(x-x^{L-1}\right) /\left(x^{L}-x^{L-1}\right)>1 .
$$

By analogy with (7.3) the total contributions from the interval $x \in] x^{L}, x^{+}$[ become

$$
\begin{align*}
& W\left(L-1 ; x^{L}, x^{+}\right)=\frac{x^{L}-\mu}{x^{L}-x^{L-1}} \cdot\left(\Phi\left(z^{+}\right)-\Phi\left(z^{L}\right)\right)-\frac{\sigma}{x^{L}-x^{L-1}} \cdot\left(\Psi\left(z^{+}\right)-\Psi\left(z^{L}\right)\right), \\
& W\left(L ; x^{L}, x^{+}\right)=\Phi\left(z^{+}\right)-\Phi\left(z^{L}\right)-W\left(L-1 ; x^{L}, x^{+}\right) \tag{7.6}
\end{align*}
$$

If we form $W\left(L-1 ; x^{L-1}, \infty\right), W\left(L ; x^{L}, \infty\right)$ by summing up the lump probability at the truncation bound and the corresponding contribution in (7.6), we get expression (7.5) also in the extrapolation case. This means that the whole weight referred to, e.g., the highest level $L$ (with $x_{d}^{+} \geq x_{d}^{L-1}$ ) in any case formally is written

$$
W\left(L ; x^{L-1}, \infty\right)=W\left(L ; x^{L-1}, x^{+}\right)+\frac{x^{+}-x^{L-1}}{x^{L}-x^{L-1}} \cdot\left(1-\Phi\left(z^{+}\right)\right) .
$$

(On behalf of the extrapolation case we prefer the concept weight instead of probability.)
Theorem 7.1: In the interpolation case the nodal weights, as defined in (7.3) - (7.5), can be used as probabilities, and provide the correct expected state variable values, i.e. the vector $\boldsymbol{\mu}$.

Proof: For the interpolation case we may, without loss of generality, assume that $x^{-} \in\left[x^{0}, x^{1}\right]$ and $x^{+} \in\left[x^{L-1}, x^{L}\right]$. As probabilities the weights should be non-negative and sum up to 1 .

Since the lump probabilities are nonnegative, like the point-based weights $w\left(x ; x^{\ell}\right)$ and $w\left(x ; x^{\ell+1}\right.$ ) (defined by (7.2)) for every $\left.x \in\right] x^{\ell}, x^{\ell+1}[$, also the integrated weights $W$ in (7.3) (7.5) become non-negative. The weight sum in an arbitrarily chosen stochastic dimension $d$ (implicit) is:

$$
\begin{aligned}
& W\left(0 ;-\infty, x^{1}\right)+W\left(1 ;-\infty, x^{1}\right)+\sum_{\ell=1}^{L-2}\left[W\left(\ell ; x^{\ell}, x^{\ell+1}\right)+W\left(\ell+1 ; x^{\ell}, x^{\ell+1}\right)\right]+W\left(L-1 ; x^{L-1}, \infty\right)+ \\
& +W\left(L ; x^{L-1}, \infty\right)=\Phi\left(z^{1}\right)+\sum_{\ell=1}^{L-2}\left[\Phi\left(z^{\ell+1}\right)-\Phi\left(z^{\ell}\right)\right]+1-\Phi\left(z^{L-1}\right)=1 .
\end{aligned}
$$

The nodal weight is a product of 1D-weights. The linear interpolation (multiplicity) property means that the free combinations of weights from all the different stochastic dimensions, i.e. the nodal weights, sum up to 1 . (In order to show this, fix the weights of all dimensions except one. The only remaining degree-of-freedom corresponds to all the nodes in the remaining dimension with weights summing up to 1 , etc.)

As for the expected value we consider the state variable values in stochastic dimension $d$ (implicit). Each node level value $x^{\ell}$ is to be multiplied by the weights of all nodes
corresponding to this level value. The linear interpolation property means that the 1D-weights of all other dimensions but $d$ sum up to 1 . By use of (7.3) - (7.5) the correspondence to the expected value becomes

$$
\begin{aligned}
& x^{0} \cdot W\left(0 ;-\infty, x^{1}\right)+x^{1} \cdot\left[W\left(1 ;-\infty, x^{1}\right)+W\left(1 ; x^{1}, x^{2}\right)\right]+\sum_{\ell=2}^{L-2} x^{\ell} \cdot\left[W\left(\ell ; x^{\ell-1}, x^{\ell}\right)+W\left(\ell ; x^{\ell}, x^{\ell+1}\right)\right] \\
& +x^{L-1} \cdot\left[W\left(L-1 ; x^{L-2}, x^{L-1}+W\left(L-1 ; x^{L-1}, \infty\right)\right]+x^{L} \cdot W\left(L ; x^{L-1}, \infty\right)=\cdots=\right. \\
& =x^{-} \cdot \Phi\left(z^{-}\right)+x^{+} \cdot\left[1-\Phi\left(z^{+}\right)\right]+\mu \cdot\left[\Phi\left(z^{+}\right)-\Phi\left(z^{-}\right)\right]+\sigma \cdot\left[\Psi\left(z^{+}\right)-\Psi\left(z^{-}\right)\right] .
\end{aligned}
$$

Our choice of symmetric truncation intervals now means $\Phi\left(z^{+}\right)+\Phi\left(z^{-}\right)=1, \Psi\left(z^{+}\right)=\Psi\left(z^{-}\right)$ and $x^{-}+x^{+}=2 \mu$. Hence the final expression turns into $\mu$.


Figure 7.1 Marginal density function $f$ and lump probabilities $P$. Node levels $x^{0}, x^{1}, \ldots$ and truncation bounds $x^{-}, x^{+}$for state variable value $x$ immediately after a major maintenance op.


Figure 7.2 Marginal node level weights in a 2D case. Six node levels per stochastic dimension.

The weights $W\left(\ell ; x^{\ell}, x^{\ell+1}\right)$, etc., can be determined from tabulated $\Phi-, \Psi$ - values. Thus the additional computational burden is modest, so far. In this manner the weights are referred to the node levels of each separate state dimension - see Fig 7.2 for a 2D-illustration. In a succeeding step the weights are referred to the full $D^{\prime}$-dimensional nodal points, by simple multiplication of the marginal weight contributions, on behalf of the assumed statistical independence (and the multiplicative property of our interpolation method - cf. Ch 2).

The additional computational work comes from handling far more nodes in the interpolation process than just the neighbouring $2^{D}$ ones, for $D$ state dimensions. With a lattice of, e.g., $L=6$ node levels (the standard interpretation of $L$ ) per $D^{\prime}=4$ stochastic dimension, this computational part of the DynP backward iterations means up to $L^{D^{\prime}} / 2^{D^{\prime}}=3^{4}=81$ times as many nodes to weigh as in the corresponding deterministic model. The average extra work will be lower, due to truncation. The DynP backward iterations will dominate the CPU-time, because of the one-year mixed integer nonlinear optimisation performed at each node and time.

As for implementation, due to the increased storage requirement it is less appropriate to store all the individual weights that are involved in a weighing, at maximum $2^{D-D^{\prime}} \cdot L^{D^{\prime}}$ weights. Instead the underlying node level weights per state dimension should be stored, at maximum $\left(D-D^{\prime}\right) \cdot 2+D^{\prime} \cdot L$ weights. For, e.g., $L=D=5, D^{\prime}=4$ we get the figures 1250 and 22, respectively. A disadvantage of the latter strategy is that the full weights have to be generated each time they are needed, although this is easy for multilinear interpolation.

### 7.4 Results

We have run the road class oriented versions of the deterministic (Ch 4) and stochastic models, with five node levels per state dimension and absolute limits for the worst acceptable state. A fixed annual maintenance budget level is applied, corresponding to the long term dual price $v_{\infty}=v_{\mathrm{BCR}}$, according to the deterministic model. (The residual values have not been recalculated for the stochastic model but are taken from the deterministic model in Ch 5 .) The standard deviation input values in Tab 3.9 are somewhat uncertain, since the regression models were based on weighing by segment lengths, where the weight sums, i.e. the total segment length per traffic class sample, are unknown to us. We have to assume that the segments lengths in the samples coincide with the population mean lengths. Some results from 500 iterations can be found in Table 7.1.

|  | Traffic cost | Primal cost | Primal-Dual Gap |
| :--- | ---: | ---: | ---: |
| Deterministic | 448157068 | 2620810336 | 0,000195 |
| Stochastic | 455137984 | 2634871685 | 0,000068 |

Table 7.1 Model comparison run. Costs for the best found primal solutions and resulting gaps after 500 iterations.

In Tab. 7.1 the model differences are modest, e.g., the traffic cost is $1.5 \%$ higher in the stochastic model. Remember that the primal solutions in the road class oriented model are determined by splitting segment-metres on the neighbouring nodal states in the DynP forward
routine. In the stochastic model this splitting obeys the extended interpolation scheme, involving most of the nodal states. Due to the nonlinear costs (cf. Sec 4.3.4) even the decent states will risk costly time evolutions, raising the total expected cost levels as in Tab 7.1.

By comparing the chosen works types for each segment in the first year we get the statistics in Tab 7.2.

| DeterministiclStochastic | Routine maintenance | Major maintenance |
| :--- | ---: | ---: |
| Routine maintenance | 7538 | 7 |
| Major maintenance | 8 | 1196 |

Table 7.2 Model comparison run. Works types per segment in $1^{\text {st }}$ year for best found primal solutions.

According to Tab. 7.2 almost all works types coincide; deviating for less than $0.2 \%$ of the road segments. For the segments where major maintenance is performed in both models, also the layer thicknesses (in mm) coincide - almost always at the lowest feasible value (cf. Sec 4.2.4).

We have also compared the state distributions in each road class at the end of year 10 - see Figure 7.3 below, where state dimension IRI and four road classes are treated. There are no dramatic differences between the results of the two models, the systematic differences confined to somewhat wider distributions for the stochastic model - as expected from the introduced random states immediately after a major maintenance operation.

In summary the result differences between the deterministic and the stochastic models are very small in this run. Since this may be due to our underestimating the standard deviations involved, further regression modelling and optimisation should be made for clarification.

### 7.5 Discussion

The above modeled stochastic variation is separate for each segment and road class, thus making independent contributions to the total cost and overall state distribution in the network. However, an important cause of real co-variation exists. The Swedish winters have a firm impact on the degradation of road surfaces. This is a kind of between-year weather variation that is essentially common to all the road segments in a road class. Apart from having an influence on the current deterioration rates it will cause the state distribution in the whole network to oscillate between above and below average. With a constant annual budget we anticipate this to lead to a flatter state distribution due to more frequent major maintenance and thin pavement layers, i.e. a clear systematic influence on the results. On the other hand, with a redistributable budget (cf. Ch 8 below) strategies for funding and pre-use might be derived from the use of a stochastic model similar to the one above, in order to cope with the winter fluctuations in a cost-efficient way that will reduce the result differences. The robustness of different works types might be another area of study. However, before any such study is conducted a realistic model should be formulated, for the winter weather impact on the road surfaces.

The segment oriented version of our approach in Sec 7.3 is an example of discrete event (micro) simulation, see e.g. [Fishman (1973)], in individually modelling the decision process of

### 7.5 Discussion

each segment (pavement object). Although the model (7.1) is stochastic, our method represents deterministic simulation, according to [Kleijnen (1974), p 13], since no sampling by random number generation is involved. This is acceptable, if many segments make small independent contributions to the overall state and cost. If a winter weather model is to be added, we have two options: turn to stochastic (Monte Carlo) simulation, by generating scenarios as series of random numbers, or stick to our present approach. We evolve the latter option:

- Satisfy the Markov property, by expanding the pavement state space by one or more weather state dimensions.
- Quantify the rules that control the evolution in the resulting weather-pavement state space, e.g. the influence upon deterioration rates of different weather states, and the correlation between the weather states in any two consecutive years.
- If the stochastic influence is Normal, then apply the formulas in Sec 7.3; otherwise derive similar expressions. Apply the method in Sec 7.3, with initial values including the current winter weather.

If coordinated maintenance is admitted (on the segment oriented level), the models and methods in Ch 6 still hold, modified according to the "stochastic interpolation scheme" in Sec 7.3, with or without winter weather state dimensions. This will work, since coordination has a direct influence upon the cost model (cf. Sec 3.6.1), whereas the state effect model is unaffected.

In summary, our approach can be extended in several ways.
Figure 7.3 State distribution IRI after 10 years


## 8 Redistributable budget

According to the initial study (Ch 2) as well as the main study (Ch 4), the initial state distribution shows imbalances between the road classes and budget shortages in the whole network. In the optimisation runs this is recognised as a need for increased budget means during the very first years, and an initial major spending on some road classes, whereas the average states of others are deliberately worsened. In contrast to real life we may, on a computer, easily allow redistributions of budgetary means between the years of the planning period, for better overall utilisation, if possible. Another reason for redistribution would be to optimise the use of idle capital, unavailble in the budget year due to return rate (BCR) restrictions on the maintenance investments.

### 8.1 Model

The discount rate (interest rate of discounting) was introduced in Sec 3.7.1. In the previous chapters the traffic and maintenance costs for different years are made comparable as present worth, through the discount rate, determining the cost discount factor $d$, for each year of backward discounting. Now assume that unspent budget money can be used later on, capitalised according to a given interest rate corresponding to the capital growth factor $\vec{d}$ per year, e.g. a $3 \%$ interest rate corresponding to $\vec{d}=1.03$. Since this is just one alternative use of money, the interest rate should be at most the same as the discount rate, i.e. we restrict $d \leq 1 / \vec{d}$. Moreover, assume that budgetary means can be utilised prior to the budget year, according to a given capital discount factor $\bar{d}$ per year, e.g. $6 \%$ interest rate a year corresponding to $\bar{d}=1 / 1.06$, and in general, since $\bar{d}$ typically stands for an external loan rate otherwise not utilised by VV, we restrict $\bar{d} \leq d$. As in earlier chapters the nominal budget level in year $t$ is $b_{t}$ and the total maintenance cost in year $t$ is $G_{t}\left(\mathbf{x}_{\bullet t}, \mathbf{u}_{\bullet}\right)$, where . (dot) index marks an implicit vector component, here for segments or road classes.

The budget flow for year $t$ is illustrated in Fig 8.1, where the variable $\bar{w}_{t+1}$ stands for the amount flowing (carried) backward from the succeeding year $t+1$ and $\vec{w}_{t-1}$ is the amount carried forward from the preceding year $t-1$. Besides redistribution it should be possible to leave money unspent - marked as slack in Fig 8.1 - due to the return rate (BCR) restrictions.


Figure 8.1 Budgetary flow balance

### 8.1 Model

When used in year $t$ the carried budget values are $\vec{d} \cdot \vec{w}_{t-1}$ and $\bar{d} \cdot \bar{w}_{t+1}$. The vertical arrows at the period boundaries illustrate the capital losses from $\vec{d} \cdot \vec{w}_{t-1}$ (vs. normal discounting $\left.\vec{w}_{t-1} / d\right), \bar{d} \cdot \bar{w}_{t}\left(\right.$ vs. $\left.d \cdot \bar{w}_{t}\right)$, etc. Motivated by practical restrictions, e.g. on partly fixed nonmonetary maintenance resources, we assume that the net budget means must not deviate from the nominal budget $b_{t}$ by more than two given factors $\alpha \leq 1, \beta \geq 1$, i.e. the net budget $z_{t}$ in year $t=0,1, \ldots, T-1$ should satisfy $z_{t}:=b_{t}+\vec{d} \vec{w}_{t-1}+\bar{d} \bar{w}_{t+1}-\vec{w}_{t}-\bar{w}_{t} \in\left[\alpha \cdot b_{t}, \beta \cdot b_{t}\right]$. Here we introduce $\vec{w}_{-1}=\bar{w}_{0}=0, \vec{w}_{T-1}=\bar{w}_{T}=0$ to get a unified form for these constraints.

If the utilizable budgets $\left(y_{t}\right)_{t}$ are introduced, to cope with return rate restrictions (cf. Sec 4.1.1), the model becomes (cf. (4.3))

$$
\underset{\mathbf{x}, \mathbf{u}, \mathbf{y}, \mathbf{z}, \overline{\mathbf{w}}, \overline{\mathbf{w}}}{\operatorname{minimise}}\left[F(\mathbf{x}, \mathbf{u})+v_{\mathrm{BCR}} \cdot \mathbf{1}^{T} \mathbf{y}\right] \text { s.t. }\left\{\begin{array}{c}
\mathbf{G}(\mathbf{x}, \mathbf{u}) \leq \mathbf{y}  \tag{8.1a}\\
\mathbf{y} \leq \mathbf{z} \\
z_{t}=b_{t}+\vec{d} \vec{w}_{t-1}+\bar{d} \bar{w}_{t+1}-\vec{w}_{t}-\bar{w}_{t} \quad \forall t \\
\mathbf{z} \in[\alpha \mathbf{b}, \beta \mathbf{b}] \\
\overrightarrow{\mathbf{w}} \geq \mathbf{0}, \overline{\mathbf{w}} \geq \mathbf{0} \\
(\mathbf{x}, \mathbf{u}) \in \mathcal{X U}
\end{array}\right.
$$

As in Ch 4 this model can be further specialized into a road class oriented and a segment oriented model, respectively.

### 8.2 Method

As in earlier chapters our strategy is to start from and modify the method for the basic model. This means that we retain the overall relaxation methodology, identifying and solving separable, road network based dual subproblems by DynP. However, the supplementary dual subproblem for the utilizable budgets $\mathbf{y}$ now becomes non-trivial. Moreover, for the generation of feasible solutions we need to extend the primal heuristics, in order to accomplish consistency between the network and budget subproblems.

### 8.2.1 Dual optimisation

As in the main study in Ch 4 the budgetary constraints (8.1a) are relaxed, by the introduction of nonnegative Lagrangean multipliers $\left(v_{t}\right)_{t=0}^{T-1}$. The dual subproblem becomes

$$
\Phi(\mathbf{v}):=\min _{\mathbf{x}, \mathbf{u}, \mathbf{y}, \mathbf{z}, \overline{\mathbf{w}}, \overline{\mathbf{w}}}\left[F(\mathbf{x}, \mathbf{u})+\mathbf{v}^{T} \mathbf{G}(\mathbf{x}, \mathbf{u})+\left(v_{\mathrm{BCR}} \cdot \mathbf{1}-\mathbf{v}\right)^{T} \mathbf{y}\right] \text { s.t. (8.1b) }-(8.1 \mathrm{f})
$$

As before (cf. Lemma 4.1), every dual $\mathbf{v}$-optimum must satisfy $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$. For any such given $\mathbf{v}$, the objective and the only $\mathbf{y}$-constraints (8.1b) imply that $\mathbf{y}=\mathbf{z}$ is optimum for the dual subproblem. We state this property as a general result (cf. Thm 4.1(b)):

Theorem 8.1 (a): For any fixed solution ( $\mathbf{x}, \mathbf{u}$ ) in the primal (8.1), such that $\mathbf{G}(\mathbf{x}, \mathbf{u}) \leq \beta \mathbf{b}$, $\mathbf{y}=\mathbf{G}(\mathbf{x}, \mathbf{u})$ is optimum.
(b): For $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$ and any fixed solution ( $\mathbf{x}, \mathbf{u}$ ) in the dual subproblem (8.2), $\mathbf{y}=\mathbf{z}$ is optimum.
(c): Assume that, as $\mathbf{v}$ varies, unique subproblem optima exist such that the maintenance costs $\overline{\mathbf{G}}(\mathbf{v})$ and the net budgets $\overline{\mathbf{z}}(\mathbf{v})$ are continuous functions of $\mathbf{v}$. Then, for any dual optimum $\mathbf{v}$, each component $t$ satisfies

$$
v_{t}>v_{\mathrm{BCR}}, \quad \bar{G}_{t}(\mathbf{v})=\bar{z}_{t}(\mathbf{v})
$$

or

$$
v_{t}=v_{\mathrm{BCR}}, \quad \bar{G}_{t}(\mathbf{v}) \leq \bar{z}_{t}(\mathbf{v}) .
$$

Proof (a): $\mathbf{G}(\mathbf{x}, \mathbf{u}) \leq \beta \mathbf{b}$ admits feasible solutions of the budget subproblem. For $\mathbf{y}=\mathbf{G}(\mathbf{x}, \mathbf{u})$ the constraints (8.1b) coincide with (8.1a), i.e. this choice of $\mathbf{y}$ does not restrict the feasible region in the budget subproblem. The rest of (a) and (b) follows as in the proof of Thm 4.1. (c): $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$ for any optimum $\boldsymbol{v}$ is proved as in Lemma 4.1.

Case cl: $v_{t}>v_{\mathrm{BCR}}$. Assume the opposite, e.g. $\bar{g}_{t}(\mathbf{v}):=\bar{G}_{t}(\mathbf{v})-\bar{z}_{t}(\mathbf{v})=\varepsilon>0$ for optimum $\mathbf{v}$. Consider $\widetilde{\mathbf{v}}$ such that $\bar{g}_{t}(\widetilde{\mathbf{v}})>0, \widetilde{v}_{t}>v_{t}$ and $\widetilde{v}_{t^{\prime}}=v_{t^{\prime}} \quad \forall t^{\prime} \neq t$. The existence of such a $\widetilde{\mathbf{v}}$ follows from the continuity assumption. The subgradient property (cf. Sec 1.2) implies

$$
\Phi(\mathbf{v}) \leq \Phi(\widetilde{\mathbf{v}})+\bar{g}_{t}(\widetilde{\mathbf{v}}) \cdot\left(v_{t}-\widetilde{v}_{t}\right)<\Phi(\widetilde{\mathbf{v}}),
$$

a contradiction.
Case c2: $v_{t}=v_{\mathrm{BCR}}$. Assume the opposite, $\bar{g}_{t}(\mathbf{v})=\varepsilon>0$. Consider $\widetilde{\mathbf{v}}$ chosen as in case c 1 . A contradiction follows analogously.

The uniqueness and continuity assumptions in Thm 8.1c are not expected to be satisfied here, but are a basis for our primal heuristics below.

As in Ch 4 the dual subproblem separates into a number of dynamic road class specific or segment specific subproblems, plus a budget subproblem. By letting $\Phi^{\text {net }}(\mathbf{v}):=\min _{\mathbf{x}, \mathbf{u}}\left[F(\mathbf{x}, \mathbf{u})+\mathbf{v}^{T} \mathbf{G}(\mathbf{x}, \mathbf{u})\right]$ catch the network part and focusing on the budget subproblem, without loss of generality using the optimum (Thm 8.1(b)) relation $\mathbf{y}=\mathbf{z}$ and (8.1c), the separable dual objective can be written

$$
\begin{aligned}
& \Phi(\mathbf{v})=\min _{\mathbf{x}, \mathbf{u}, \mathbf{z}, \overrightarrow{\mathbf{w}}, \overline{\mathbf{w}}}\left[F(\mathbf{x}, \mathbf{u})+\mathbf{v}^{T} \mathbf{G}(\mathbf{x}, \mathbf{u})+\sum_{t=0}^{T-1} d^{t+\frac{1}{2}} \cdot\left(v_{\mathrm{BCR}}-v_{t}\right)\left(b_{t}+\vec{d} \cdot \vec{w}_{t-1}+\bar{d} \cdot \bar{w}_{t+1}-\vec{w}_{t}-\bar{w}_{t}\right)\right]= \\
& =\Phi^{\mathrm{net}}(\mathbf{v})+\min _{\mathbf{z}, \overrightarrow{\mathbf{w}}, \overline{\mathbf{w}}} \sum_{t=0}^{T-1} d^{t+\frac{1}{2}} \\
& \quad \cdot\left(\left[v_{t}-d \vec{d} v_{t+1}-(1-d \vec{d}) v_{\mathrm{BCR}}\right] \cdot \vec{w}_{t}+\left[v_{t}-\frac{\bar{d}}{d} \cdot v_{t-1}-\left(1-\frac{\bar{d}}{d}\right) v_{\mathrm{BCR}}\right] \cdot \bar{w}_{t}-\left(v_{t}-v_{\mathrm{BCR}}\right) b_{t}\right)= \\
& =: \Phi^{\mathrm{net}}(\mathbf{v})+\Phi^{\mathrm{bud}}(\mathbf{v})-\sum_{t=0}^{T-1} d^{t+\frac{1}{2}}\left(v_{t}-v_{\mathrm{BCR}}\right) b_{t}
\end{aligned}
$$

Here the dynamic, road class or segment oriented, network subproblem with objective function $\Phi^{\text {net }}(\boldsymbol{v})$ is formally identical to (4.7) and (4.6), respectively; the only implicit difference is the

### 8.2 Method

interpretation of the dual prices $\left(v_{t}\right)_{t=0}^{T-1}$. The two respective formulations are handled equally cf. the DynP backward and forward iteration routines in Ch 4.

The budget subproblem, with constraints (8.1c) - (8.1e), is slightly reformulated (to get nonnegative variables only), by the introduction of $\mathbf{z}^{+}:=\mathbf{z}-\alpha \mathbf{b}$ :

$$
\begin{align*}
& \Phi^{\mathrm{bud}}(\mathbf{v})=\min _{\mathbf{z}^{+}, \overrightarrow{\mathbf{w}}, \overline{\mathbf{w}}} \sum_{t=0}^{T-1} d^{t+\frac{1}{2}}\left(\left[v_{t}-d \vec{d} v_{t+1}-(1-d \vec{d}) v_{\mathrm{BCR}}\right] \cdot \vec{w}_{t}+\left[v_{t}-\frac{\bar{d}}{d} \cdot v_{t-1}-\left(1-\frac{\bar{d}}{d}\right) v_{\mathrm{BCR}}\right] \cdot \bar{w}_{t}\right) \\
& \quad \text { subject to }\left\{\begin{array}{cl}
z_{t}^{+}-\vec{d} \cdot \vec{w}_{t-1}-\bar{d} \cdot \bar{w}_{t+1}+\vec{w}_{t}+\bar{w}_{t}=(1-\alpha) \cdot b_{t} & \forall t \\
z_{t}^{+} \leq(\beta-\alpha) \cdot b_{t} & \forall t \\
z_{t}^{+}, \vec{w}_{t}, \bar{w}_{t} \geq 0 & \forall t
\end{array}\right. \tag{8.2a}
\end{align*}
$$

Problem (8.2) has a certain structure that can be utilised for solving it. [Jewell (1962)] formulated and solved it as a minimum-cost flow with gains problem. To show the network flow structure, see Fig 8.2. We specify gains, see e.g. [Ahuja et al (1993)] and [Evans and Minieka (1992), p 151], for all arcs, as amplification factors $\vec{d}(>1)$ for all forward redistribution arcs and the damp factors $\bar{d}(<1)$ for all backward redistribution arcs. The arcs starting from the source node so generate the nominal budgets, and the arcs ending at the sink node si control the net budgets variation.

[Dantzig (1963), p 413] formulated and solved the dual subproblem as an LP-problem with special structure, the weighted distribution problem. We will use our general LP-code plus
special start solutions, since this is a problem with just $2 T$ constraints, if the upper $z_{t}^{+}$bounds are included, i.e. 80 in the main and 160 in the start routine for our standard study. Moreover, our primal heuristics below will solve the same kind of problem and utilise information from the optimal LP-basis.

Let the objective $\vec{w}_{t}$, $\bar{w}_{t}$-coefficients in (8.2a) be denoted $\vec{c}_{t}, \bar{c}_{t}$, respectively. We devote the rest of this section to characterising the optimum of the budget subproblem (8.2). The $1^{\text {st }}$ optimum property concerns the three potentially optimal $z_{t}$-levels that can last over several consecutive time periods.

Proposition 8.1: For any $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}, \vec{c}_{t}, \bar{c}_{t+1}$ in (8.2a) cannot both be negative.
Proof: Assume the opposite, i.e. that both coefficients are negative. Neglect the common coefficient factor $d^{t+\frac{1}{2}}(>0)$, multiply the $\bar{w}_{t+1}$-coefficient by $d \vec{d}(>0)$ and add to the $\vec{w}_{t}$ coefficient, to have

$$
\begin{aligned}
& {\left[v_{t}-d \vec{d} v_{t+1}-(1-d \vec{d}) v_{\mathrm{BCR}}\right]+d \vec{d} \cdot\left[v_{t}-\frac{\bar{d}}{d} \cdot v_{t-1}-\left(1-\frac{\bar{d}}{d}\right) v_{\mathrm{BCR}}\right]<0} \\
& \Leftrightarrow(1-\vec{d} \vec{d})\left(v_{t}-v_{\mathrm{BCR}}\right)<0
\end{aligned}
$$

Here, $\bar{d} \leq d \leq \frac{1}{\bar{d}}$ and $v_{t} \geq v_{\mathrm{BCR}}$, i.e. the negativity assumption is contradicted.
A deterrent against redistribution is:
Lemma 8.1: For any given $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$, the interval

$$
\begin{equation*}
v_{t+1} \in\left[v_{\mathrm{BCR}}+\frac{\bar{d}}{d}\left(v_{t}-v_{\mathrm{BCR}}\right), v_{\mathrm{BCR}}+\frac{1}{d \vec{d}}\left(v_{t}-v_{\mathrm{BCR}}\right)\right] \tag{8.3}
\end{equation*}
$$

is non-empty and both the $\vec{w}_{t}, \bar{w}_{t+1}$-coefficients $\vec{c}_{t}, \bar{c}_{t+1}$ in (8.2a) are nonnegative there.
Proof: The form of the interval end points follows directly from (8.2a). The non-emptiness corresponds to $(1-\vec{d} \bar{d})\left(v_{t}-v_{\mathrm{BCR}}\right) \geq 0-$ cf. the opposite inequality in the proof of Prop 8.1.

Theorem 8.2: For any given $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$, if $\vec{c}_{t}, \bar{c}_{t+1}$ in (8.2a) satisfy (8.3) for some $t$, then any feasible solution of (8.2) with $\vec{w}_{t}>0$ and/or $\bar{w}_{t+1}>0$ and $\left.z_{\tau} \in\right] \alpha b_{\tau}, \beta b_{\tau}[$ for $\tau=t, t+1$ is nonoptimal.

Proof: If $\vec{w}_{t}>0$, then use that $z_{t}$ can be increased and $z_{t+1}$ decreased in the solution. Hence $\vec{w}_{t}$ can be reduced until either of $\vec{w}_{t}=0, z_{t}=\beta b_{t}$ and $z_{t+1}=\alpha b_{t+1}$ occurs. By Lemma 8.1 the new solution gets a lower cost. Otherwise, i.e. if $\vec{w}_{t}=0, \bar{w}_{t+1}>0$, the argumentation is analogous.

For certain problem instances the net budget optimum of (8.1) turns into the nominal budget optimum of Thm 4.1(b):

Proposition 8.2: If the $\vec{w}_{t}, \bar{w}_{t+1}$-coefficients in (8.2a) satisfy (8.3) for each $t$, then $\mathbf{z}=\mathbf{b}$ is a budget subproblem optimum.

Proof: $\operatorname{In}(8.2) \mathbf{z}^{+}=(1-\alpha) \mathbf{b}, \overrightarrow{\mathbf{w}}=\overline{\mathbf{w}}=\mathbf{0}$ is a feasible solution, with objective value 0 . Since (by Lemma 8.1) all coefficients are nonnegative, like all the variables, this value is optimal.

Corollary 8.1: If all dual prices $v_{t}$ are equal, then $\mathbf{z}=\mathbf{b}$ is a budget subproblem optimum.
Proof: Since $v=v_{\mathrm{BCR}}+1 \cdot\left(v-v_{\mathrm{BCR}}\right)$ and $\frac{\bar{d}}{d} \leq 1 \leq \frac{1}{d \vec{d}},(8.3)$ is fulfilled for each $t$.
Cor 8.1 is an example of interior solutions in the $z_{t}$-intervals (8.1d). If (8.3) is not fulfilled, then the optimum budget is at either of the bounds, in the following sense:

Theorem 8.3 For any given $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$,
(a): if $\vec{c}_{t}<0$, then $z_{t}^{+}>0, z_{t+1}^{+}<(\beta-\alpha) b_{t+1}$ is non-optimal for (8.2),
(b): if $\bar{c}_{t+1}<0$, then $z_{t}^{+}<(\beta-\alpha) b_{t}, z_{t+1}^{+}>0$ is non-optimal for (8.2).

Proof: (a) According to Lemma $8.1 \vec{c}_{t}<0$ implies $\bar{c}_{t+1} \geq 0$.
Consider a candidate $\mathbf{z}^{+}$-optimum $\hat{\mathbf{z}}^{+}$with $\hat{z}_{t}^{+}>0, \hat{z}_{t+1}^{+}<(\beta-\alpha) b_{t+1}$.
Case a1 $\hat{z}_{t+1}^{+}+\vec{d} \cdot \hat{z}_{t}^{+} \leq(\beta-\alpha) b_{t+1}$. Compare the solution $\hat{\mathbf{z}}^{+}$with a new solution, only differing from $\hat{\mathbf{z}}^{+}$as to $z_{t}^{+}=0, z_{t+1}^{+}=\hat{z}_{t+1}^{+}+\vec{d} \cdot \hat{z}_{t}^{+}$and the additional $\vec{w}_{t}$-flow $\hat{z}_{t}^{+}$.
The new solution means an additional cost $\vec{c}_{t} \hat{z}_{t}^{+}<0$, i.e. $\hat{\mathbf{z}}^{+}$is non-optimal.
Case $\mathbf{a} \mathbf{2} \hat{z}_{t+1}^{+}+\vec{d} \cdot \hat{z}_{t}^{+}>(\beta-\alpha) b_{t+1}$. Compare $\hat{\mathbf{z}}^{+}$with a new solution with $z_{t+1}^{+}=(\beta-\alpha) b_{t+1}$, $z_{t}^{+}=\hat{z}_{t}^{+}-\left((\beta-\alpha) b_{t+1}-\hat{z}_{t+1}^{+}\right) / \vec{d}$ and $\vec{w}_{t}=\left((\beta-\alpha) b_{t+1}-\hat{z}_{t+1}^{+}\right) / \vec{d}$.
The new solution has the additional cost $\vec{c}_{t} \vec{w}_{t}<0$, i.e. $\hat{\mathbf{z}}^{+}$is non-optimal.
(b) is proved in the same way.

If a number of consecutive periods satisfy the condition in Thm 8.3 (a), we expect just one of the periods to have an intermediate (interior) net budget $\left.z_{t}^{+} \in\right] 0,(\beta-\alpha) b_{t}[$, likewise for the condition in Thm 8.3 (b), i.e. a bang-bang behaviour. If every pair of consecutive components of the price vector $\mathbf{v}$ shows a non-negligible increase or decrease, the optimum budgets are redistributed to one end of the planning period:

Corollary 8.2 For any given $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$,
(a): if $\vec{c}_{t}<0 \forall t$, then a $t_{0}$ exists, such that $\left\{\begin{array}{cc}z_{t}^{+}=0 & \forall t<t_{0} \\ z_{t}^{+}=(\beta-\alpha) b_{t} & \forall t>t_{0}\end{array}\right.$ is optimal for (8.2),
(b): if $\bar{c}_{t+1}<0 \forall t$, then a $t_{0}$ exists, such that $\left\{\begin{array}{cc}z_{t}^{+}=(\beta-\alpha) b_{t} & \forall t<t_{0} \\ z_{t}^{+}=0 & \forall t>t_{0}\end{array}\right.$ is optimal for (8.2).

Proof: (a) According to Thm 8.3 it is advantageous to send as much flow as possible forwards,
filling up to the upper budget limit from the last time period and backwards, one period at a time. (In such a case the net budget level $z_{t_{0}}$ in the switch period $t_{0}$ can take any optimum value, not only the standard levels $\alpha b_{t_{0}}, b_{t_{0}}$ or $\beta b_{t_{0}}$.)
(b) follows in the same way.

The $2^{\text {nd }}$ (and last) optimum property for subproblem (8.2) concerns the conditions - Thm 8.4 below - for absence of flow cycles. (As a general property this is well-known.)

We expect that at most one of the flows $\vec{w}_{t}, \bar{w}_{t+1}$ between times $t$ and $t+1$ is positive in the optimum, i.e. the flow is one-way across every period boundary (cf. Fig 8.1). To show this, from now on let $\mathcal{T}$ denote an arbitrary set of consecutive time periods, starting at some $t_{1}$ and ending at some $t_{2}$, i.e. $\mathcal{T}=\left\{t_{1}, t_{1}+1, \ldots, t_{2}-1, t_{2}\right\}$.

Definition 8.1: A flow cycle is a set $\mathcal{T}$ of consecutive time periods $t$ with positive two-way inter-period flow except at the end periods $t=t_{1}$ (as to the boundary $\left(t_{1}-1, t_{1}\right)$ ) and $t=t_{2}$ (as to the boundary $\left(t_{2}, t_{2}+1\right)$ ).

Except for financial abundance situations flow cycles are out of the question:
Lemma 8.2: Assume $\mathbf{v}>v_{\mathrm{BCR}} \cdot \mathbf{1}$ and $\vec{d} \bar{d}<1$. Any flow cycle $\mathcal{T}$ is non-optimal for (8.2) unless $\mathbf{z}$ in (8.1) satisfies $z_{t}=\beta b_{t} \quad \forall t \in \mathcal{T}$.

Proof: Consider an arbitrary boundary in $\mathcal{T}$ between periods $t\left(<t_{2}\right)$ and $t+1\left(>t_{1}\right)$. Since $\mathcal{T}$ is a flow cycle, we have $\vec{w}_{t}>0, \bar{w}_{t+1}>0$. For a sufficiently small $\varepsilon(>0)$ the modified flow

$$
\begin{equation*}
\vec{w}_{t}^{\prime}=\vec{w}_{t}-\varepsilon \cdot \vec{q}, \bar{w}_{t+1}^{\prime}=\bar{w}_{t+1}-\varepsilon \cdot \bar{q} \tag{8.4}
\end{equation*}
$$

is nonnegative. The $\varepsilon$-coefficients in (8.4), $\vec{q}$ and $\bar{q}$, are chosen to satisfy

$$
\begin{equation*}
\bar{d} \leq \frac{\vec{q}}{\bar{q}} \leq \frac{1}{\vec{d}} \tag{8.5}
\end{equation*}
$$

For the $\varepsilon$-modifications not to have an influence on other boundary crossing flow ( $\vec{w}_{t+1}, \bar{w}_{t}$ etc), the balancing quantities $z_{t}^{+}, z_{t+1}^{+}$in (8.2b), corresponding to the period net budgets $z_{t}$, $z_{t+1}$, must equal

$$
z_{t}^{\prime+}=z_{t}^{+}+\varepsilon \cdot(\vec{q}-\bar{d} \bar{q}), z_{t+1}^{\prime+}=z_{t+1}^{+}+\varepsilon \cdot(\bar{q}-\vec{d} \vec{q})
$$

By (8.5) these modifications are nonnegative, and feasible unless both $z_{t}^{+}$and $z_{t+1}^{+}$are at the upper bound: if $z_{t}=\beta b_{t}$ we must choose $\vec{q} / \bar{q}=\bar{d}$, and if $z_{t+1}=\beta b_{t+1}$ we must choose $\vec{q} / \bar{q}=1 / \vec{d}$. As for the objective function value the $\varepsilon$-modifications mean the change $-\varepsilon \cdot d^{t+\frac{1}{2}} \cdot\left[\left(v_{t}-v_{\mathrm{BCR}}\right)(\vec{q}-\bar{d} \bar{q})+d \cdot\left(v_{t+1}-v_{\mathrm{BCR}}\right)(\bar{q}-\vec{d} \vec{q})\right]$.
Because of the assumptions $\mathbf{v}>v_{\mathrm{BCR}} \cdot \mathbf{1}$ and $\vec{d} \bar{d}<1$ this change is strictly negative. This means that the $\varepsilon$-modifications provide strictly improved solutions, until $\vec{w}_{t}=0$ or $\bar{w}_{t+1}=0$ is reached, whence the flow cycle is broken, or else both $z_{t}$ and $z_{t+1}$ reach their upper budget bounds. Here the period boundary was arbitrarily chosen in $\mathcal{T}$..

### 8.2 Method

As for the remaining flow cycle possibilities we observe:
Lemma 8.3: Assume at least one strict inequality (out of three) in $\bar{d} \leq d \leq \frac{1}{\vec{d}}, \beta \geq 1$. If $z_{t}=\beta b_{t} \quad \forall t \in \mathcal{T}$, then the set of consecutive time periods $\mathcal{T}$ must have a net inflow from outside $T$.

Proof: Assume the opposite, i.e. $\mathcal{T}=\left\{t_{1}, \ldots, t_{2}\right\}$ has zero-flow or outflow at the end periods $t=t_{1}$ and $t=t_{2}$. For a net budget at the upper limits, i.e. $z_{t}=\beta b_{t} \forall t \in \mathcal{T}$, it follows from (8.2b), the strict inequality assumption and $\bar{w}_{t_{1}} \geq 0, \vec{w}_{t_{1}-1}=0, \vec{w}_{t_{2}} \geq 0, \bar{w}_{t_{2}+1}=0$ that $\beta \cdot \sum_{t \in T} d^{t+\frac{1}{2}} b_{t}=\sum_{t \in T} d^{t+\frac{1}{2}} \cdot z_{t}=\sum_{t \in T} d^{t+\frac{1}{2}} \cdot\left[b_{t}+\vec{d} \vec{w}_{t-1}+\bar{d} \bar{w}_{t+1}-\vec{w}_{t}-\bar{w}_{t}\right]=$ $=\sum_{t \in T} d^{t+\frac{1}{2}} b_{t}-\sum_{t=t_{1}}^{t_{2}-1} d^{t+\frac{1}{2}} \cdot(1-d \vec{d}) \vec{w}_{t}-\sum_{t=t_{1}+1}^{t_{2}} d^{t+\frac{1}{2}} \cdot(1-\bar{d} / d) \bar{w}_{t}-d^{t_{1}+\frac{1}{2}} \bar{w}_{t_{1}}-d^{t_{2}+\frac{1}{2}} \vec{w}_{t_{2}}<\beta \cdot \sum_{t \in T} d^{t+\frac{1}{2}} b_{t}-0$ a contradiction.

Theorem 8.4: For any given $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$, an optimum solution for (8.2) without flow cycles always exists. Especially, if $\mathbf{v}>v_{\mathrm{BCR}} \cdot \mathbf{1}$ and $\vec{d} \bar{d}<1$, then any flow cycle is non-optimal.

Proof: On behalf of Lemmas 8.2 and 8.3 the only remaining possibility is a positive net inflow at either or both end periods $t=t_{1}, t=t_{2}$. Assume $\vec{w}_{t_{1}-1}>0$; the remaining possibility $\vec{w}_{t_{1}-1}=0, \bar{w}_{t_{2}+1}>0$, is treated analogously. For an $\varepsilon$-modification with $\vec{q} / \bar{q}=1 / \vec{d}$ at the boundary $\left(t_{1}, t_{1}+1\right)$ the net budget $z_{t_{1}+1}\left(=\beta b_{t_{1}+1}\right)$ becomes unchanged (cf. the proof of Lemma 8.2), whereas the corresponding increase of $z_{t_{1}}$ is neutralized by a decreased $\vec{w}_{t_{1}-1}^{\prime}=\vec{w}_{t_{1}-1}-\varepsilon \cdot \frac{1}{\vec{d}}(\vec{q}-\bar{d} \bar{q})$. Such a change is feasible, unless $z_{t_{1}-1}=\beta b_{t_{1}-1}$. If the net budget $z_{t_{1}-1}$ already is at its upper bound, we continue backwards neutralizing by means of further reduced inflows $\vec{w}_{t_{1}-2}^{\prime}=\vec{w}_{t_{1}-2}-\varepsilon \cdot \frac{1}{\vec{d}^{2}}(\vec{q}-\bar{d} \bar{q})$, where $\vec{w}_{t_{1}-2}>0$ must hold according to Lemma 8.3, etc., until we reach a time period with its net budget below its upper limit. The existence of such a time period $t_{0} \in\left\{0, \ldots, t_{1}-1\right\}$ follows from Lemma 8.3, since otherwise $\mathcal{T}^{\prime}=\left\{0, \ldots, t_{1}-1\right\}$ would form a set with upper limit net budget and net outflow (at $t=t_{1}-1$ ). The $\varepsilon$-modifications mean the following net change for the objective (at $t \in\left\{t_{0}, \ldots, t_{1}+1\right\}$ ):

$$
\begin{aligned}
& -\varepsilon \cdot \sum_{t=t_{0}}^{t_{1}-1} d^{t+\frac{1}{2}}\left[v_{t}-d \vec{d} v_{t+1}-(1-d \vec{d}) v_{\mathrm{BCR}}\right] \cdot \frac{\vec{q}}{\vec{d}^{t_{1}-t}}(1-\vec{d} \bar{d})-\varepsilon \cdot d^{t_{1}+\frac{1}{2}}\left(v_{t_{1}}-v_{\mathrm{BCR}}\right) \vec{q}(1-\vec{d} \bar{d})= \\
& =-\varepsilon \cdot d^{t_{0}+\frac{1}{2}} \frac{\vec{q}}{\vec{d}^{t_{1}-t_{0}}}(1-\vec{d} \bar{d}) v_{t_{0}}+\varepsilon \cdot \frac{\vec{q}}{\vec{d}^{t_{1}}}(1-d \vec{d})(1-\vec{d} \bar{d}) v_{\mathrm{BCR}} d^{\frac{1}{2}} \sum_{t=t_{0}}^{t_{1}-1}(d \vec{d})^{t}+\varepsilon \cdot \vec{q}(1-\vec{d} \bar{d}) d^{t_{1}+\frac{1}{2}} v_{\mathrm{BCR}}= \\
& =-\varepsilon \cdot d^{t_{0}+\frac{1}{2}} \frac{\vec{q}}{\vec{d}^{t_{1}-t_{0}}}(1-\vec{d} \bar{d})\left(v_{t_{0}}-v_{\mathrm{BCR}}\right) .
\end{aligned}
$$

Since $\mathbf{v} \geq v_{\mathrm{BCR}} \cdot \mathbf{1}$ and our general assumption is $\vec{d} \bar{d} \leq 1$, the last expression is non-positive, i.e. the cycle free solution is at least as good. In the special case we get a cost reduction by eliminating the given flow cycle.

For any fixed $\mathbf{v}$, according to the $1^{\text {st }}$ optimum property for (8.2), most net budgets are expected to be on any of the three $z_{t}$-levels $\alpha b_{t}, b_{t}$ and $\beta b_{t}$. In contrast, an immense number of optimum levels for the network subproblem (by $G$ in (8.1a)) are expected. Therefore iterative adjustments of $\mathbf{v}$ are needed, in order to solve the Lagrangean dual. Such are performed, e.g., by the standard subgradient technique, see Ch 4.

### 8.2.2 Primal heuristics

We keep to our strategy from earlier chapters in finding a primal solution by modifying the solution of the current dual subproblem. As for the solutions of two dual subproblem parts, i.e. the network subproblem for $\Phi^{\text {net }}(v)$ s.t. (8.1f) and the budget subproblem (8.2), primal feasibility means that the resulting maintenance cost $\bar{G}_{t}$ and net budget $\bar{z}_{t}$ in year $t$ also must satisfy (8.1a) - (8.1b), i.e. $\bar{G}_{t} \leq \bar{z}_{t}$. As in earlier chapters our strategy is to accomplish primally feasible solutions by stepping one year at a time. From a given state $\mathbf{x}_{\bullet t}$ at the start of yr $t$ we match cost and budget in $\mathrm{yr} t$ by appropriate adjustments of the dual price $v_{t}$, while relying on the dually computed costs and dual prices for year $t+1$ and onwards.

Definition 8.2 (a): The heuristics network problem (in year $t$ ) is the network subproblem (cf. (4.7), (4.6)) restricted to given $\left(\mathbf{x}_{\boldsymbol{t}^{\prime}}\right)_{t^{\prime} \leq t},\left(u_{\bullet t^{\prime}}\right)_{t^{\prime}<t}$.
(b): The heuristics budget problem (in year $t$ ) is the budget subproblem (8.2) restricted to given $\left(z_{t^{\prime}}\right)_{t^{\prime}<t},\left(\vec{w}_{t^{\prime}}\right)_{t^{\prime}<t},\left(\bar{w}_{t^{\prime}}\right)_{t^{\prime} \leq t}$.
(c): The heuristics consistency problem (in year $t$ ) is to find

$$
v_{\text {Prim }, t}:=\arg \min \left(v_{t} \in\left[v_{B C R,}, v_{\infty}\right]: \bar{G}_{t}\left(v_{t}\right) \leq \bar{z}_{t}\left(v_{t}\right)\right)
$$

for at least one optimum solution cost $\bar{G}_{t}=\bar{G}_{t}\left(v_{t}\right)$ of the heuristics network problem and at least one optimum solution net budget $\bar{z}_{t}=\bar{z}_{t}\left(v_{t}\right)$ of the heuristics budget problem for $v_{t}$.

In order to solve the heuristics consistency problem the following two properties are crucial:
Proposition 8.3: The optimal maintenance cost $\bar{G}_{t}$ of the heuristics network problem is a nonincreasing set-valued function of (the one and only variable) $v_{t}, \bar{G}_{t}=\bar{G}_{t}\left(v_{t}\right)$.

Proof: Since the heuristics network problem is the Lagrangean dual subproblem for some primal, restricted according to Def 8.2(a), Lagrangean duality (cf. Sec 1.2.2) applies. Thus the dual objective $\Phi$ is a concave function of the dual price $v_{t}$ of (8.1a), i.e. wherever $\Phi$ is twice differentiable with respect to $v_{t}$ we have

$$
\frac{d \Phi}{d v_{t}}=\bar{G}_{t}, \frac{d \bar{G}_{t}}{d v_{t}}=\frac{d^{2} \Phi}{d v_{t}^{2}} \leq 0
$$

Otherwise we make use of the subgradient property (1.7) of $\bar{G}_{t}$ at $v_{t}$ and at any $v_{t}^{\prime}>v_{t}$ :

### 8.2 Method

$$
\Phi\left(v_{t}^{\prime}\right) \leq \Phi\left(v_{t}\right)+\bar{G}_{t}\left(v_{t}\right) \cdot\left(v_{t}^{\prime}-v_{t}\right), \quad \Phi\left(v_{t}\right) \leq \Phi\left(v_{t}^{\prime}\right)+\bar{G}_{t}\left(v_{t}^{\prime}\right) \cdot\left(v_{t}-v_{t}^{\prime}\right)
$$

Hence

$$
\bar{G}_{t}\left(v_{t}^{\prime}\right) \cdot\left(v_{t}^{\prime}-v_{t}\right) \leq \Phi\left(v_{t}^{\prime}\right)-\Phi\left(v_{t}\right) \leq \bar{G}_{t}\left(v_{t}\right) \cdot\left(v_{t}^{\prime}-v_{t}\right)
$$

and $v_{t}^{\prime}>v_{t}$ implies

$$
\bar{G}_{t}\left(v_{t}^{\prime}\right) \leq \bar{G}_{t}\left(v_{t}\right) .
$$

Proposition 8.4: The optimum net budget $\bar{z}_{t}\left(=\alpha b_{t}+z_{t}^{+}\right)$of the heuristics budget problem is a non-decreasing set-valued function of (the one and only variable) $v_{t}, \bar{z}_{t}=\bar{z}_{t}\left(v_{t}\right)$, attaining all values between the lowest and highest possible.

Proof: Assume the opposite of non-decreasing, i.e. for a pair of LP-problem instances $A, B$ such that $v_{t}^{B}>v_{t}^{A}$ we get an optimum solution $a$, with net budget $\bar{z}_{t}^{a}$, of problem $A$ and an optimum solution $b$, with net budget $\bar{z}_{t}^{b}$, of problem $B$, such that $\bar{z}_{t}^{b}<\bar{z}_{t}^{a}$. Both solutions $a, b$ are feasible in both problems $A, B$. By using the notation $\bar{b}_{t}$ in (8.6) below for the utilizable budget and the net redistribution $w_{t}:=\vec{w}_{t}-\bar{d} \cdot \bar{w}_{t+1}$, the flow balance (8.2b) in time period $t$ becomes

$$
\bar{z}_{t}+w_{t}=\bar{b}_{t}
$$

According to this equation and the assumption, the solutions $a, b$ satisfy $w_{t}^{b}>w_{t}^{a}$. Let $\Phi^{A a}$ denote the objective function value for solution $a$ in problem $A$, etc. The optimality for solution $a$ in problem $A$ means

$$
\Phi^{A b} \geq \Phi^{A a}
$$

The $v_{t}$-change $\Delta v_{t}:=v_{t}^{B}-v_{t}^{A}(>0)$ will have an influence on the objective coefficient of $w_{t}$ in (8.2a) only; whence

$$
\begin{aligned}
& \Phi^{B a}=\Phi^{A a}+d^{t+\frac{1}{2}} \cdot \Delta v_{t} \cdot w_{t}^{a} \\
& \Phi^{B b}=\Phi^{A b}+d^{t+\frac{1}{2}} \cdot \Delta v_{t} \cdot w_{t}^{b}>\Phi^{A b}+d^{t+\frac{1}{2}} \cdot \Delta v_{t} \cdot w_{t}^{a} \geq \\
& \geq \Phi^{A a}+d^{t+\frac{1}{2}} \cdot \Delta v_{t} \cdot w_{t}^{a}=\Phi^{B a}
\end{aligned}
$$

Thus $\Phi^{B b}>\Phi^{B a}$, i.e. solution $b$ is non-optimal in problem $B$ - a contradiction.
As for the attained values of $\bar{z}_{t}=\bar{z}_{t}\left(v_{t}\right)$, a change of $v_{t}$ only affects the objective coefficients in (8.2a), linearly and continuously, i.e. the set of LP-basic solutions is constant for every $v_{t}$. If two or more LP-basic solutions provide the optimal value (8.2a) in the heuristics budget problem for some $v_{t}$, then its LP-problem character implies that also any convex combination of the minima does. Specifically this applies to the $\bar{z}_{t}$-component.

In the heuristics network problem, which is a mixed-integer, nonlinear optimisation problem, one and the same maintenance solution may be optimum on some narrow $v_{t}$-interval (narrow, due to the many segments in (4.6) and the many nodal states in (4.7), respectively). For the heuristics budget problem follows (as in the proof of Prop 8.4), from LP sensitivity analysis,
see e.g. [Dantzig (1963), p 275], that any basic solution (with a fixed $\bar{z}_{t}$-value) is optimum on some $v_{t}$-interval. In between these intervals, discontinuous $\bar{G}_{t}$ and basic solution $\bar{z}_{t}$-steps are expected. This motivates the formulation of the heuristics consistency problem in Def 8.2c: we should look for the lowest $v_{t}$-value such that $\bar{G}_{t} \leq \bar{z}_{t}$ (cf. Thm 8.1(c)), rather than hope for a unique $v_{t}$-value such that $\bar{G}_{t}=\bar{z}_{t}$.

In time step $t$ all quantities $\left(\vec{w}_{t^{\prime}}\right)_{t^{\prime}<t},\left(\bar{w}_{t^{\prime}}\right)_{t^{\prime} \leq t}$ are fixed by the earlier steps. The given budget level $b_{t}$ in (8.2b) is replaced by the utilizable budget

$$
\begin{equation*}
\bar{b}_{t}:=b_{t}+\vec{d} \cdot \vec{w}_{t-1}-\bar{w}_{t} \tag{8.6}
\end{equation*}
$$

and $\bar{z}_{t}=\bar{b}_{t}$ replaces $z_{t}=b_{t}$ as one of the three potential optimum levels (infeasible as to (8.1d) if $\bar{b}_{t}>\beta b_{t}$ or $\bar{b}_{t}<\alpha b_{t}$ ). As was mentioned at Cor 8.2 other $\bar{z}_{t}$-values may occasionally be optimal. Thus, as $v_{t}$ varies the two limit levels, normally $z_{t}^{0}=\alpha b_{t}$ and $z_{t}^{1}=\beta b_{t}$, may become optimum $\bar{z}_{t}$-levels, together with zero or more other levels, e.g. $\bar{z}_{t}=\bar{b}_{t}$. In Fig. 8.3 we illustrate the issue, assuming one additional optimum level $\widetilde{b}_{t}$ on the characteristic step curve. The drawn $\bar{G}_{t}$-curves are simplifications of functions with many small steps downwards (cf. Fig 8.4 below). As for the choice of $v_{t}$ in $\left.v_{t} \in\right] v_{\mathrm{BCR}}, v_{\infty}[$ two main cases are distinguished:

- In the $1^{\text {st }}$ case, from an implementation point of view and illustrated by curves a, c, e, the optimal maintenance cost $\bar{G}_{t}$ for consistency with $\bar{z}_{t}$ is adjusted to (slightly below) the appropriate budget level $\alpha b_{t}, \widetilde{b}_{t}$ or $\beta b_{t}$. In this case a maintenance cost ranking of the potential projects - one per road class and nodal state or one per road segment - is necessary. This gives the adjusted $v_{t}$-value, possibly a $\bar{G}_{t}$-discontinuity.
- In the $2^{\text {nd }}$ case, see curves $\mathbf{b}, \mathbf{d}$, the variation of $v_{t}$ also leads to a switch of budget level in (8.2); this admits an intermediate (non-LP-basic) $\bar{z}_{t}$-value $\left(=\bar{G}_{t}\right)$ in the primal solution (cf. Fig 8.4 below). In this case the $\bar{G}_{t}$-value is simply evaluated for the appropriate switch $v_{t}$-value, without any CPU-time-consuming cost ranking of the maintenance projects.

As a result of this primal step, also the variables $\left(\mathbf{u}_{\bullet t}\right), \vec{w}_{t}$ and $\bar{w}_{t+1}$ become fixed. During the step we do not perform any recalculation of optimal works extents $\left(u_{s t 2}\right)_{s}$. In the $1^{\text {st }}$ main case (cf. Fig 8.3, curves a, $\mathbf{c}, \mathbf{e}$ ) this means that the $v_{t}$-value that solves the heuristics consistency problem can be calculated directly, as in Sec 4.2.3. In the $2^{\text {nd }}$ main case (curves $\mathbf{b}, \mathbf{d}$ ) the expected $v_{t}$-switch points can be estimated, simplest if all the constraints in (8.2b) are ignored, except the two for year $t$. In any case the $v_{t}$-value giving consistency can be identified after a few trials, as described below.

The constraints $v_{t} \geq v_{\mathrm{BCR}}$ are handled as in Ch 4 .


Figure 8.3 Maintenance cost curves and net budget levels vs. dual price in year $t$.
Let us look into the details of our method for finding a $v_{t}$-value that solves the heuristics consistency problem, Def 8.2c. The initial $v_{t}$-value comes from the current Lagrangean dual iteration (cf. Sec 8.2.1). As for the switch points $v_{t}^{-}, v_{t}^{+}$in Fig 8.3, the deviations $\Delta v_{t}$ from the current price $v_{t}$ can be estimated by sensitivity analysis:

Theorem 8.5: Let $\overline{\mathbf{c}}_{N}$ and $\overline{\mathbf{N}}_{i}$, respectively, denote the vector of current reduced costs and row $i$ of the current coefficient matrix for the non-basic variables in the optimum LP-solution of the heuristics budget problem for time $t$ (for fixed $v_{t}$ ). The optimum LP-basis remains unchanged for the following $v_{t}$-deviations $\Delta v_{t}$.
If both $\vec{w}_{t}$ and $\bar{w}_{t+1}$ are non-basic: $-\frac{1}{d^{t+\frac{1}{2}}} \cdot \bar{c}_{\bar{w}_{t}} \leq \Delta v_{t} \leq \frac{1}{\bar{d} d^{t+\frac{1}{2}}} \cdot \bar{c}_{\bar{w}_{t+1}}$.
If $\vec{w}_{t}$ is non-basic and $\bar{w}_{t+1}$ is basic: $\overline{\mathbf{c}}_{N}^{T}+\bar{d} d^{t+\frac{1}{2}} \cdot \Delta v_{t} \overline{\mathbf{N}}_{\bar{w}_{t+1}} \geq \mathbf{0}^{T}$.
If $\bar{w}_{t+1}$ is non-basic and $\vec{w}_{t}$ is basic: $\overline{\mathbf{c}}_{N}^{T}-d^{t+\frac{1}{2}} \cdot \Delta v_{t} \overline{\mathbf{N}}_{\bar{w}_{t}} \geq \mathbf{0}^{T}$.
Proof: In the objective function (8.2a) $v_{t}$ affects the coefficients of the four variables $\vec{w}_{t-1}$, $\bar{w}_{t}, \vec{w}_{t}$ and $\bar{w}_{t+1}$. Here the first two variables $\vec{w}_{t-1}, \bar{w}_{t}$ have been fixed in the preceding primal substep $t-1$. The results, based on the two remaining coefficients, then follow from standard sensitivity analysis.

The results in Thm 8.6 might be used for a direct estimation of the switch values $v_{t}^{-}, v_{t}^{+}$. We will, however, take the corresponding $\Delta v_{t}$-limits merely as initial values and re-solve the heuristics budget problem until the true switch points are found (within prescribed accuracy). We illustrate this procedure by an example. Consider Fig 8.4 where the price $\hat{v}_{t}$ generates the maintenance cost $\hat{G}_{t}$ in the heuristics network problem (unfilled point), and the budget level,
here $\widetilde{b}_{t}$, of the heuristics budget problem (filled point). This is a situation where we have to identify at least one switch point, here $v_{t}^{+}$. For any switch point we start from the most recent LP-optimum and use the sensitivity results in Thm 8.5 to estimate a trial switch $v_{t}$-value. If necessary, a sequence of trial $v_{t}$-values is generated, based on sensitivity results for the most recent LP-solution, each value requiring at most a few primal LP-iterations. In Fig 8.4 we search for the curve intersection point $\left(v_{t}^{+}, G_{t}^{+}\right)$- unique whenever the $\bar{G}_{t}$-curve has no step downwards exactly at $v_{t}^{+}$. The case $\widetilde{b}_{t}=\bar{b}_{t}$ (in (8.6)) corresponds to both $\vec{w}_{t}$ and $\bar{w}_{t+1}$ being zero and non-basic - whence $v_{t}^{+}=\hat{v}_{t}+\frac{1}{\bar{d}} \cdot \bar{c}_{\bar{w}_{t+1}}$ is expected, according to Thm 8.5. In general we enclose the correct switch value $v_{t}^{+}$by re-solving the budget problem twice, for $v_{t}^{+} \pm \varepsilon$, and find the corresponding $G_{t}^{+}$-value in the network problem, by comparing $v_{t}^{+}$with the calculated $v_{t}$-switch points for each maintenance project (i.e. the $v_{t}$-value where the preference order between routine and major maintenance for a segment is reversed; cf. Sec 4.2.3). The final solution of the heuristics consistency problem is $v_{t}^{+}$and $\bar{z}_{t}=G_{t}^{+}$.


Figure 8.4 Example maintenance cost curve and net budget curve vs. dual price in year $t$.

A simpler solution alternative in our example would be to adjust the free budget level (here by means of $\bar{w}_{t+1}$ ) up to $\hat{G}_{t}$ at the given price $\hat{v}_{t}$, by leaving the budget step curve in Fig 8.4. But such a feasible solution is clearly non-optimal (for $\hat{v}_{t}>V_{\mathrm{BCR}}$ ), according to Def 8.2c, and we therefore stick to our more ambitious choice, here $\left(v_{t}^{+}, G_{t}^{+}\right)$, which might be optimal.

A shortcut occurs if the current $v_{t}$-value coincides with either of the two limit values, $v_{t}=v_{\infty}$ (a given proxy for infinity) or $v_{t}=v_{\mathrm{BCR}}$, and the corresponding computed $\bar{G}_{t}, \bar{z}_{t}$-values demand a still more extreme $v_{t}$-value: the former $\left(v_{t}=\infty, \bar{G}_{t}>\bar{z}_{t}\right)$ indicates that the

### 8.2 Method

heuristics consistency problem has no feasible solution (cf. Fig 8.5d below) and the latter $\left(v_{t}=v_{\mathrm{BCR}}, \bar{G}_{t} \leq \bar{z}_{t}\right)$ means a binding return rate constraint (cf. Fig 8.5a).


Figure 8.5 Examples of deviations from the standard heuristics solution, due to reduced budget limits $z_{t}=z_{t}^{0}, z_{t}^{1}(\mathbf{b}, \mathbf{c})$, and due to prescribed multiplier limits $v_{t}=v_{\mathrm{BCR}}, v_{\infty}(\mathbf{a}, \mathbf{d})$.

There are situations where the full, step curve $\bar{z}_{t}$ vs. $v_{t}$ cannot be generated:

- one or more switch points are below the lower feasibility limit $v_{t}=v_{\mathrm{BCR}}$,
- due to deficit financing in the preceding time periods (primal steps) more than the maximum possible $\bar{d} \bar{w}_{t+1}$-value is reserved by $\bar{w}_{t}$ :
$(\beta-1) b_{t}+\bar{w}_{t}-\vec{d} \vec{w}_{t-1}>\sum_{k=1}^{T-1-t}(1-\alpha) b_{t+k} \bar{d}^{k}$ implying that $\bar{z}_{t}=\beta b_{t}$ is infeasible, and in general $\bar{z}_{t} \leq z_{t}^{1}:=\min \left[\beta b_{t}, \bar{b}_{t}+\sum_{k=1}^{T-1-t}(1-\alpha) b_{t+k} \bar{d}^{k}\right] \quad(\mathrm{cf}$. Fig 8.5c),
- due to surplus financing in the preceding time periods (primal steps) more than the maximum possible $\vec{w}_{t}$-value is reserved by $\vec{w}_{t-1}$ :
$(1-\alpha) b_{t}+\vec{d} \vec{w}_{t-1}-\bar{w}_{t}>\sum_{k=1}^{T-1-t}(\beta-1) b_{t+k} / \vec{d}^{k}$ implying that $\bar{z}_{t}=\alpha b_{t}$ is infeasible, and in general $\bar{z}_{t} \geq z_{t}^{0}:=\max \left[\alpha b_{t}, \bar{b}_{t}-\sum_{k=1}^{T-1-t}(\beta-1) b_{t+k} / \vec{d}^{k}\right]$ (cf. Fig 8.5b).

In the heuristics budget problem an initial feasible LP basic solution can easily be found: in time step $t$, with the variables $\left(\vec{w}_{\tau}\right)_{\tau=0}^{t-1},\left(\bar{w}_{\tau}\right)_{\tau=1}^{t}$ of the preceding years fixed, we construct a feasible solution by stepping forwards one year $\tau=t, t+1, \ldots, T-1$ at a time and focusing on the net budget (8.6) in year $\tau$. At time $\tau$ normally two basic variables are chosen for the time $\tau$-constraints in (8.2b)-(8.2c), out of the four candidate variables $\vec{w}_{\tau}, \bar{w}_{\tau}, z_{\tau}^{+}\left(=z_{\tau}-\alpha b_{\tau}\right)$ and the slack variable $s_{\tau}$ of constraint ( 8.2 c ). In general we choose

$$
\bar{b}_{\tau}<z_{\tau}^{0}:\left\{\begin{array}{c}
\vec{w}_{\tau}=0 \\
\bar{w}_{\tau+1}=\left(z_{\tau}^{0}-\bar{b}_{\tau}\right) / \bar{d} \\
\bar{z}_{\tau}=z_{\tau}^{0} \\
s_{\tau}=\beta b_{\tau}-z_{\tau}^{0}
\end{array} \quad z_{\tau}^{0} \leq \bar{b}_{\tau} \leq z_{\tau}^{1}:\left\{\begin{array}{c}
\vec{w}_{\tau}=\bar{w}_{\tau+1}=0 \\
\bar{z}_{\tau}=\bar{b}_{\tau} \\
s_{\tau}=\beta b_{\tau}-\bar{b}_{\tau}
\end{array} \quad \bar{b}_{\tau}>z_{\tau}^{1}:\left\{\begin{array}{c}
\vec{w}_{\tau}=\bar{b}_{\tau}-z_{\tau}^{1} \\
\bar{w}_{\tau+1}=0 \\
\bar{z}_{\tau}=z_{\tau}^{1} \\
s_{\tau}=\beta b_{\tau}-z_{\tau}^{1}
\end{array}\right.\right.\right.
$$

The need for three instead of two basic variables in a period $\tau$ arises only if $z_{\tau}^{0}>\alpha b_{\tau}$ or $z_{\tau}^{1}<\beta b_{\tau}$. Our choices $\bar{z}_{\tau}=z_{\tau}^{0}$ and $\bar{z}_{\tau}=z_{\tau}^{1}$, respectively, then mean that the flows in all the succeeding periods $\tau^{\prime}>\tau$ satisfy $\bar{z}_{\tau^{\prime}}=\beta b_{\tau^{\prime}}$ and $\bar{z}_{\tau^{\prime}}=\alpha b_{\tau^{\prime}}$, respectively, and the extra basic variable in time $\tau$ may replace $s_{\tau^{\prime}}=0$ and $z_{\tau^{\prime}}^{+}=0$, respectively.

### 8.3 Results

We have implemented the method and performed a small study with a constant annual budget, corresponding to the stationary budget level for the return rate limit $v=v_{\mathrm{BCR}}$ according to the residual value routine, using 5 levels of discretization, and assuming absolute feasibility limits for worst acceptable state. For each case we have run the road class oriented routine for 601 dual iterations, using a subgradient technique for the updating of $\mathbf{v}$ (cf. Ch 4), with time horizon $T=80$. From the fixed budget case we have made changes, allowing $5 \%$ or $10 \%$ budget deviations and using the backward financial discount factor 0.95 , whereas the forward interest rate is put to the interest rate of calculation, $4 \%$. A summary is found in Tab 8.1. Here the constant budget case (run 0 ) is the basis for the cost comparisons: the relative objective $\Phi(\mathbf{v})$-differences ( $\Phi / \Phi_{0}-1$ ) for the best dual iterations and the relative traffic cost $F$ differences $\left(F / F_{0}-1\right)$ for the best primal iterations are shown. Moreover, the present worth values of the unused ("slack") capital are presented for the best primal solutions. For each run the remaining gaps between the best found values of the primal and dual objective values are also listed.

In these runs the difference between the costs computed in the DynP backward and forward iteration routines is around $0.2 \%$.

### 8.3 Results

The primal heuristics is equipped with a supplementary step length routine: if the enclosing of a switch point for the net budget curve by the repeated use of sensitivity analysis (see Thm 8.5) fails in two steps, then we turn to extrapolation with increased $v_{t}$-steps. When the switch point becomes enclosed, the remaining $v_{t}$-uncertainty interval is reduced by use of Thm 8.5, modified to avoid too close trial $v_{t}$-values.

| Run | $\vec{d}=d^{-1}$ | $\bar{d}$ | $\beta=2-\alpha$ | $\Delta \Phi / \Phi_{0}$ | $\Delta F / F_{0}$ | $\Delta($ slack $) /(\text { slack })_{0}$ | gap |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | - | - | $(1)$ | 0 | 0 | 0 | $19 \cdot 10^{-5}$ |
| 1 | 1.04 | 0.95 | 1.05 | $-4.6 \cdot 10^{-4}$ | $-0.8 \cdot 10^{-3}$ | $1.1 \cdot 10^{-2}$ | $3.1 \cdot 10^{-5}$ |
| 2 | 1.04 | 0.95 | 1.10 | $-4.8 \cdot 10^{-4}$ | $-1.7 \cdot 10^{-3}$ | $0.6 \cdot 10^{-2}$ | $0.4 \cdot 10^{-5}$ |

Table 8.1 Run characteristics and cost comparisons with basic run (0).
The reduction of the best dual cost $\Phi$ is modest but the traffic cost $F$ gains somewhat from redistribution. A bonus for government is the side effect that despite the (small) improvements of $\Phi$ and $F$ the capital is used more efficiently in case of redistribution, increasing the unused ("slack") money by around $1 \%$. Thus everyone has something to gain from budget redistribution, although it might mean a more varying volume of maintenance.

The redistribution option has a clear effect on the resulting dual prices $\mathbf{v}$ for the best found dual solutions - see Fig 8.6. Whereas the rigid budget case (run 0 ) shows capital scarcity over most of the horizon, the variations are completely evened out in run 2 , except for the initial year.

Fig 8.6 Lagrangean multiplier vs. year. Redistributable budget


If redistribution is admitted the budget levels will vary - see Fig 8.7 below, where the results for the best found primal solutions are drawn. The run characteristics are the same as in Tab 8.1.



### 8.3 Results

Note the fine-tuned agreement in Fig 8.7 between runs 1 and 2 in the years $10-40$. Before and after that, the extreme net budget values $\bar{z}_{t}=\alpha b_{t}$ and $\bar{z}_{t}=\beta b_{t}$ in (8.1d) are used. The bangbang behaviour is somewhat illusory, as Fig 8.8 for the corresponding maintenance costs shows. All three maintenance cost curves approach the stationary cost level corresponding to $v_{t}=v_{\mathrm{BCR}}$, i.e. the nominal budget level in these runs. For all three runs the cost curves show a deep dip after the initial years, deepest after an initial redistribution. This means that the volume of maintenance will vary strongly between years, irrespective of redistribution possibilities, due to the return rate restrictions. Thus flexible resources are needed, irrespective of budget strategy, and should not be an argument against budget redistribution. The runs also show that the general maintenance state of the road network is rather good, after an initial extra effort to handle the roads close to the given state feasibility limits. The cost curve for run 1 has a curious peak after 40 years and decreases somewhat in the end, showing that this is an imperfect solution: the remaining gap in Tab 8.1 is higher in this run than in run 2.

There are discrepancies between the budget levels and costs in the second half of the time horizon for these primal solutions. This is due to a redistribution of slack money towards the end of the horizon. How is this possible? We have not registered the modified dual prices $\mathbf{v}_{\text {Prim }}$ for the heuristics, just the dual optimum $\mathbf{v}$ in Fig 8.6. Let us scrutinise, e.g., run 1. Time $t$ intervals, coefficients $\vec{c}_{\mathrm{t}}, \bar{c}_{\mathrm{t}+1}$-signs and redistributions $\vec{w}_{\mathrm{t}}, \bar{w}_{\mathrm{t}+1}$-values are collected in Tab 8.2. Here $t_{1}, t_{2}$ are some unknown redistribution switch points (see below).

| $t$ | $v_{t}$ | $\vec{c}_{t}$ | $\bar{c}_{t+1}$ | $\vec{w}_{t}$ | $\overleftarrow{w}_{t+1}$ | Exp. $\vec{w}_{t}$ | Exp. $\overleftarrow{w}_{t+1}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $0-1$ | steep descent | + | - | 0 | + | 0 | + |
| $2-\mathrm{t} 1$ | $\nu_{\mathrm{BCR}}$ | 0 | 0 | $?$ | $?$ | 0 | + |
| $\mathrm{t} 1-8$ | $v_{\mathrm{BCR}}$ | 0 | 0 | $?$ | $?$ | + | 0 |
| $9-10$ | ascent | - | + | + | 0 | + | 0 |
| $11-\mathrm{t} 2$ | constant | 0 | + | $?$ | 0 | + | 0 |
| $\mathrm{t} 2-79$ | constant | 0 | + | $?$ | 0 | 0 | 0 |

Table 8.2 Analysis of best primal solution for the heuristics budget problem in run 1.
The $\vec{c}_{\mathrm{t}}, \bar{c}_{\mathrm{t}+1}$-signs are determined by the definitions and the used $d, \vec{d}, \bar{d}$-values. The possible $\vec{w}_{\mathrm{t}}, \bar{w}_{\mathrm{t}+1}$-signs are implied by the corresponding coefficient signs. The expected ("Exp.") $\vec{w}_{\mathrm{t}}$, $\bar{w}_{\mathrm{t}+1}$-signs follow from regarding the coefficient signs for the whole 80 yrs period. Since $\bar{c}_{\mathrm{t}+1}$ is negative for $t=0,1$ and zero for $t \in[2,8]$, maximum feasible backward redistribution is profitable on $t \in[2, t 1]$, for some $t 1$. Since $\vec{c}_{\mathrm{t}}$ is negative for $t=9,10$ and zero for $t>10$, maximum feasible forward redistribution is profitable on $t \in[t 1,8]$. Since any forward redistribution means capitalisation, the budget means from, e.g., $t=5$ to $t=75$ will increase by the factor $1.04^{70} \approx 15.6$, which easily creates budget surpluses later on. Therefore it is not surprising that no switch value $t 2(<80)$ exists, i.e. that the maximum net budget values $\bar{z}_{t}=\beta b_{t}$ are attained for every $t>42$ in the primal solution. The discrepancy between $\bar{z}_{t}$ in Fig 8.7 and the maintenance cost $\bar{G}_{t}$ in Fig 8.8 is then a consequence of the return rate constraints, and we conclude that $v_{t, \text { Prim }}=v_{\mathrm{BCR}}$ must hold for $t>42$, according to Fig 8.5a.

Note that, e.g., run 1 in Fig 8.7 could not be an optimum LP-basic solution (corresponding to $\mathbf{v}$ in Fig 8.6). There are simply too many LP-basic variables: three variables $\left(z_{t}{ }^{+}, s_{t}, \vec{w}_{t}\right)$ for $t \in[9,42]$ and two variables otherwise - but only on average two constraints per $t$-value in (8.2). From Thm 8.3(a) we expect at most one of yrs 9, 10 to have an interior net budget $\left.\bar{z}_{t} \in\right] \alpha b_{t}, \beta b_{t}[$.

As for the special iterative methods involved we present some statistics for runs $1+2$ in Tab8.3, characterising the varying number of $v_{t}$-trial values per time step $t$ of the primal heuristics, denoted $v_{\text {Prim }}$, as well as the number of simplex iterations per LP-problem in the dual subproblem and the primal heuristics. In Tab 8.3 \#problem denotes the number of cases where the corresponding iterative method is applied, whereas mean and std dev are the observed iteration mean values and standard deviations, respectively. Moreover, we present the observed percentage distribution of the number of iterations. Remember that the dual subproblem (8.2) contains $2 T=160$ constraints ( $8.2 \mathrm{~b}, \mathrm{c}$ ), whereas the number of constraints in the heuristics budget (LP-)problem varies between this number and essentially 0 , with on average 80 constraints. The low iteration mean values for the heuristics LP- and $v_{\text {Prim }}{ }^{-}$ iterations reflect the predominance of $v=v_{\mathrm{BCR}}$. On behalf of future applications it may be worthwhile to investigate the causes of the occurring $v_{\text {Prim }}$-iterations around 40 , signalling that the sensitivity based updating has been replaced/modified by the slower but more robust extrapolation and interval reduction updating.

|  | \#problem | mean | std dev | $\%: 0-19$ | $20-39$ | $40-59$ | $60-79$ | $80-99$ | $100-119$ | $120-1$ |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Dual subprobl:LP | 1202 | 84,4 | 15,2 | 0,1 | 2,0 | 4,7 | 16,8 | 63,9 | 12,1 | 0,4 |
| Primal heur:LP | 411211 | 6,3 | 17,7 | 90,2 | 2,7 | 3,0 | 2,5 | 1,4 | 0,1 | 0,0 |

Table 8.3 Iteration statistics for the redistribution runs 1-2.
8.3 Results

## 9 Methods for road network application

Is it possible to derive trustworthy optimisation results for a whole geographical region in reasonable time? With our approach, solving the Lagrangean dual, every dual iteration costs several CPU hours. Therefore our focus for the network applications is on the method for updating the dual "prices" (variable values).

In Sec 9.1 we present our prerequisites in case study Värmland, as to computer resources, and discuss the best choice of ambition level, here synonymous to discretization level. In Sec 9.2 our main updating method, by subgradients, is applied both by solving the dual subproblem for all segments before a full updating and by solving for one segment group at a time, in a cyclic pattern of partial updatings. Apart from the updating method our main concern is about result precision. The standard multi-linear interpolation can be improved in two different ways discussed in Sec 9.3. As an alternative to subgradient updating we investigate the DantzigWolfe method in Sec 9.4, implementing a compact solver for our problem and using a linear, bundle-like modification plus flexible boxes. A main reason for parallel computing, subdividing the large data/model set into smaller sets per processor, which can be kept in primary memory, is to have fast access to all information in the primal heuristics. In Sec 9.5 we examine the proc and cons of a posterior heuristics, for the generation of feasible maintenance plans. Finally, in Sec 9.6 we give some concluding comments.

### 9.1 Parallelisation

For our Värmland runs we had access to the 400 processor Monolith cluster at The National Supercomputer Center in Sweden (NSC). Each Monolith node is a PC with two (dual) Intel 2.2 GHz Xeon processors sharing 2 GBytes of memory and 80 GBytes of disk space. For parallelisation we have chosen the LAM implementation of Message Passing Interface (MPI), using FastEthernet. Our code is a stand alone package of $\mathrm{C}++$ routines, including a setup routine for the assignment of the input segments data to the available number of processors (cf. Sec 4.3.1). If the number of processors exceeds the number of road classes ( $R=29$ here), our primary aim is to split the data for equalising the number of segments per processor (totally $S=8749$ here), in such a way that the number of road classes per processor is the lowest possible; here two road classes per processor will do. The hierarchical, object oriented (memory demanding) model structure, which we generate dynamically during the run, is dominated by the dynamic programming grid, one for each road segment. There are more memory saving alternatives, which we discuss in Secs 1.3 and 4.3.2. The advantage of having each segment structure active during the whole solution process is that most file data transfer is omitted and that information about the simultaneous state of all segments is directly available during the primal heuristics, where it is repeatedly needed. For the same reason we compute in advance and store the interpolation weights for the fixed routine-maintenance option at each grid node, although there are more memory saving alternatives - see Secs 1.3 and 4.3.2. Thus we use around 16 GB memory for $L=4$ node levels per state dimension ( $D=5$ here) and 48 GB for $L=5$.

Since each pair of Monolith processors share 2 GB the ideal number of processors ought to equal the number of used GB. However, this is for an ideal situation where all computer work is locally handled at each processor. In practice we use a controlling root processor, collecting information from all others, performing computations based on overall information, e.g. finding a new dual search direction, and passing this information to all others. The data transfer itself introduces an unproductive communication time and leads to waiting, mainly due to an imbalance of the cluster work load between the computers. For each problem we have to distinguish one root processor time $T_{\text {root }}$ (non-dividable, i.e. wasted, inactive time for the other processors) and the rest, the dividable time $T_{\text {loc }}$ (where all processors are active). In the first approximation, presuming a perfectly balanced cluster, $N$ processors will result in the wall clock time $T_{\text {root }}+T_{\text {loc }} / N$, i.e. although this time decreases by $N$ the total processor run time $N \cdot T_{\text {root }}+T_{\text {loc }}$ increases. Hence a good utilisation of the computers speaks for a small $N$-value. (In our case, allocated a limited number of CPU hours per month by NSC, we had such an incentive.) However, if the $N$-value is too small the limited memory per cluster node will accomplish time-consuming swapping of data - and the simple formula must be revised. In reality this means that in most cases there is an optimal number $N$ of parallel processors ( $N>1$ ), minimising the total run time. In Fig 9.1 we have varied $N$ for two typical jobs, both using $L=5$ node levels, as in the majority of the network runs to be presented below. One method is the subgradient algorithm applied with partial updating - see Sec 9.2 - and the other is the Dantzig-Wolfe approach, also applied with partial updating - see Sec 9.4. Due to the presence of both swapping and competing network communication the CPU times will vary from run to run, i.e. the numbers in Fig 9.1 are mere examples.

The results for the subgradient method refer to 26 dual iterations, of 2 full and 24 partial updating iterations (with 10 updatings per iteration). The value for 88 processors is the average of two run times of 137 and 154 CPU hrs and the value for 44 processors has been translated from $20 \frac{2}{10}$ to 26 iterations (run stop due to a reached wall clock limit, corresponding to 220 CPU hours). In the subgradient runs the communication with the root takes place at the end of every partial (and full) dual iteration and in the primal heuristics after every completed dual iteration. We notice a steep increase at 44 processors, which we refer to swapping, and a slow increase by $T_{\text {root }}$ at and above 66 processors, where all swapping is expected to have disappeared (making the results more stabile). From these numbers and the least squares method we estimate (dotted line in Fig 9.1) $T_{\text {loc }}=98.1, T_{\text {root }}=0.638$, i.e. the dividable time $T_{\text {loc }}$ corresponds to $99.35 \%$ in cases of negligible swapping - a measure of utilisation efficency. The fewest CPU hours are registered for 88 processors and in the runs below for $L=5$ we have used this $N$-value.

In the Dantzig-Wolfe runs in Fig 9.1 the curve minimum is less pronounced. Here the results are for 8 dual iterations, of 2 full and 6 partial iterations ( 10 updatings per iteration). The value for 66 processors is translated from the 149 CPU hours received after $2+7$ iterations. The data transfer with the root processor is more intense in this method, and $T_{\text {root }}$ is expected to constitute a bigger proportion, since at the end of every partial (and full) iteration we also solve an LP-problem of several thousands of LP-iterations mainly at the root, communicating about
entering columns, etc., with the other processors - see Sec 9.4 for details. The time variation is partly due to the total number of LP-iterations. These, in turn, depend on how the segments are grouped per partial iteration - differently for different $N$-values. For, e.g., $N=88$, the total number of LP-iterations is $I=433750$ but for $N=98$ we get $I=391261$. These disturbances make us refrain from any estimation of the $T_{\text {root }}, T_{\text {loc }}$-proportions for DW from Fig 9.1.

Figure 9.1 Run time vs. number of processors for two dual updating methods


### 9.2 Subgradient method with full and partial updating

Our implementation of the subgradient method is described in Sec 4.2.1, also cf. Sec 1.2.3. To facilitate a comparison between different methods and method variants we confined the study to one main case of a constant annual budget level $b=75.5 \mathrm{MSEK}$, penalising any violation of the upper state bounds and using $L=5$ node levels per state dimension. Results for this reference case can be found in tables throughout this chapter.

### 9.2.1 Input data precision, discretization errors and run strategy

In Sec 4.3.4 we discussed the imperfect node lattices that were used in the dynamic programming routine for solving the dual subproblem. In the reference case the difference between the total costs computed backwards and forwards is typically around $2 \%$. As for the input data and functions described in Ch 3 , in many cases formulated in a first version, we cannot expect those to be perfect. This double lack of precision speaks for a less ambitious stop criterion. On the other hand we do not want that interruptions of our solution method should

### 9.2 Subgradient method with full and partial updating

contribute to the overall uncertainty. This conflict is probably reduced to a question about processor resources and CPU times. What counts here is the quality and cost of the best found feasible maintenance solution. A drawback for method comparisons is that the primal solutions are derived by a heuristics, more exposed to chance. Therefore also the best dual values are of interest as a measure of the dual method convergence. We recommend a strategy where one starts from a coarse grid, e.g. the start value $L=3$ as in Tab 9.1 below, and then use the best found primal solution as a start point for the next $L$-value. In our implementation, where the residual values are computed with the same grid density as in the main routine, we have to translate the residual value part of the total cost for the registered maintenance solutions (in a prior routine), when we change $L$-value. In Tab 9.1 the non-translated primal values are shown, i.e. the results are not directly comparable but illustrate the cost variation caused by the different grid densities. In all shown cases the primal start solution was improved during the run. The relative discretization errors are found in Tab 4.3.

| \#Levels L | \#States | \#Processors | CPU-hrs | \#Dual iterations | Best dual(kSEK) | Best primal(kSEK) | Relative gap |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 3 | 243 | 4 |  | 51 | 2545478 | 2543993 | -0.000583 |
| 4 | 1024 | 30 | 90 | 100 | 2545752 | 2537831 | -0.003111 |
| 5 | 3125 | 66 | 264 | 76 | 2541856 | 2541485 | -0.000146 |
| 6 | 7776 | 166 | 664 | 71 | 2559947 | 2559993 | 0.000018 |

Table 9.1 Run results received with different numbers of node levels and processors.
In Tab 9.1 we receive negative gaps, i.e. differences between the best found primal and dual solutions, although this trouble decreases as the grid density increases. In order to prevent stagnation of the dual method in such cases we correct the target value in the subgradient routine, whenever necessary, in such a way that a given (positive) minimum gap size is always guaranteed. In all runs we applied a (first or increased) correction, whenever the difference between the best found dual objective value $\underline{\Phi}$ and primal objective value $\bar{\Phi}$ satisfies $\underline{\Phi}>0.9999 \cdot(\bar{\Phi}+\Delta \Phi)$, where $\Delta \Phi$ is the current correction.


In Fig 9.2 the resulting primal and dual costs are given per iteration for the reference case. The vertical bars in dual iterations $31,35,57$ and 74 mark all the correction changes performed
after iteration 30. In the enlargement Fig 9.2 b it is realised that the pattern after around 30 iterations becomes rather scattered, with no significant improvements in this region of overlapping cost curves. We suspect that this happens when the discretization errors become of the same magnitude as the true dual gap. Therefore we judged that entering such a region would be a natural stop criterion and in most runs below we will use around 30 dual iterations.

The question is: does it pay to use higher $L$-values? The generated maintenance plans, rather than total costs, provide the most useful information. In Tab 9.2 we present statistics for the pairwise $L$-value comparisons that can be formed out of the runs in Tab 9.1, as to the sum of absolute deviations in number of yrs and cm:s for the pavements works on each segment. In case a segment receives different numbers of pavings according to the two models it is referred to column $n$ DiffSegm. Otherwise the difference (total absolute deviation) is referred to the corresponding distribution to the left. The difference averages are found in column $m v$. The pavement thickness shows small differences, as expected from Tab 4.2a. Despite the repeated maintenance works made during the horizon on each segment the total absolute deviation in years is only around 1 yr , on average, for all comparisons, even for $L=3$; apart from nDiffSegm. However, we notice a significant improvement for the pair 5 vs. 6 ; which speaks for the use of $L=5$.

|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | $>9$ | $m v$ | $n$ DiffSegm |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 3 vs. 4 yr | 2724 | 2070 | 1261 | 667 | 267 | 105 | 62 | 37 | 18 | 9 | 5 | 1.22 | 1524 |
| cm | 6947 | 176 | 39 | 50 | 11 | 2 | 0 | 0 | 0 | 0 | 0 | 0.13 | 1524 |
| 3 vs. 5 yr | 3756 | 2464 | 968 | 406 | 118 | 80 | 33 | 19 | 8 | 3 | 0 | 0.84 | 894 |
| cm | 7638 | 109 | 40 | 51 | 13 | 3 | 1 | 0 | 0 | 0 | 0 | 0.09 | 894 |
| 3 vs. 6 yr | 3135 | 2249 | 1276 | 548 | 268 | 90 | 53 | 20 | 3 | 6 | 0 | 1.08 | 1101 |
| cm | 7377 | 130 | 40 | 71 | 21 | 4 | 4 | 1 | 0 | 0 | 0 | 0.12 | 1101 |
| 4 vs. 5 yr | 2950 | 2036 | 1284 | 649 | 284 | 118 | 51 | 19 | 9 | 1 | 2 | 1.13 | 1346 |
| cm | 7182 | 166 | 29 | 21 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0.11 | 1346 |
| 4 vs. 6 yr | 3050 | 1969 | 1453 | 630 | 275 | 67 | 33 | 12 | 5 | 1 | 0 | 1.09 | 1254 |
| cm | 7287 | 135 | 26 | 37 | 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0.10 | 1254 |
| 5 vs. 6 yr | 4062 | 2808 | 917 | 261 | 69 | 18 | 5 | 0 | 3 | 1 | 0 | 0.65 | 605 |
| cm | 7989 | 96 | 30 | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.07 | 605 |

Table 9.2 Differences as to paving time (yr) and thickness (cm) for different pairs of $L$-values.

### 9.2.2 Full and partial updating

The full updating, documented for $L=5$ in Tab 9.1 above, is the basis for our comparisons. Partial updating means that for each processor the allocated segments are split in a number $M$ of (ideally) equally sized updating groups. In each partial updating, made for new dual prices $\mathbf{v}$, the dual subproblem is solved only for one updating group. For the others we extrapolate the dual cost for the previous solution to the current $\mathbf{v}$-value. Out of a subproblem solution in iteration $i=i(s)$ for segment $s$ we use the dual price $\mathbf{v}_{s i}$, the dual cost $\varphi\left(\mathbf{v}_{s i}\right)$ and the maintenance costs per year $\mathbf{g}_{s i}$ for constructing an affine majorant (cf. Sec 1.2.3)

$$
\begin{equation*}
\left(\varphi_{s i}(\mathbf{v}) \leq\right) \hat{\varphi}_{s i}\left(\mathbf{v} ; \mathbf{v}_{s i}\right) \equiv \varphi_{s i}\left(\mathbf{v}_{s i}\right)+\mathbf{g}_{s i}^{T}\left(\mathbf{v}-\mathbf{v}_{s i}\right) . \tag{9.1}
\end{equation*}
$$

The majorant sum, defined as $\sum_{s} \hat{\varphi}_{s, i(s)}\left(\mathbf{v} ; \mathbf{v}_{s, i(s)}\right)$, is used as an estimate of the total dual cost $\sum_{s} \varphi_{s, i(s)}(\mathbf{v})$. Here the fresh information for the current updating group is included. This
information is used for the next $\mathbf{v}$-updating, and we get $M$ such updatings per full round of segment subproblems (= one dual iteration). For the subgradient method this means two things:

- Since we get no complete solution for the current $\mathbf{v}$-value in a partial updating, the resulting total cost cannot be used as a guaranteed lower bound for the best primal solution. Therefore we normally start and finish every run with a full iteration.
- The usual step length formula in Sec 4.2.1 does not necessarily make sense. If it was used in partial updating it would mean that the influence on the budget constraint residuals in (4.6b), of the maintenance costs for each group, would be repeated $M$ times; which speaks for a reduction by the factor $1 / M$. On the other hand our estimate of the current dual cost in the updating formula is based on affine majorants, i.e. we overestimate the dual cost and underestimate the true gap. This speaks for an increased step length, in comparison with the formula in $\operatorname{Sec} 4.2 .1$. To find out the most useful choice we have tried different reduction factors.

If, as an approximation, the majorant sum for all updating groups but the current group $\hat{m}$ coincides with the corresponding best primal value and if all $M$ updating groups contribute equally to the estimated dual gap, then in the step length formula in $\operatorname{Sec} 4.2$.1 for $q \approx 1$ we get

$$
\underline{\Phi}+q \cdot(\bar{\Phi}-\underline{\Phi})-\hat{\Phi} \approx \bar{\Phi}-\hat{\Phi}=\sum_{m}\left(\bar{\Phi}_{m}-\hat{\Phi}_{m}\right) \approx \bar{\Phi}_{\hat{m}}-\hat{\Phi}_{\hat{m}} \approx(\bar{\Phi}-\hat{\Phi}) / M
$$

i.e. we get a natural step length reduction by factor $1 / M$ and should not reduce at all by force. The primal heuristics (cf. Sec 4.2.3) is applied once per round (dual iteration). In Tab 9.3a below we have varied the number of partial updatings $M$ per round, as well as the reduction factor - here $1 / M^{q}$ for $q=0, \frac{1}{2}, 1, \frac{3}{2}, \ldots$. All runs concern the reference case, for 26 dual iterations and 88 processors. The full updating corresponds to $M=1$. In the table also the $\|\mathbf{g}\|_{2}:=\sqrt{\sum_{t} d^{t+\frac{1}{2}} g_{t}^{2}}$-values are presented, where $g_{t}:=G_{t}-b_{t}$ is the budget residual for time $t$, i.e. maintenance cost minus budget, and the sum is over $t$ with $v_{t}>v_{\mathrm{BCR}}$, on behalf of the optimality condition (4.2). Tab 9.3a consists of three divisions; the first one containing results for runs where the affine majorant in (9.1) is based on the latest dual iteration $i$. In the second and third divisions all previous iterations are admissible. By choosing a majorant with the lowest extrapolated value at $\mathbf{v}$ our aim is to reduce the amount of overestimation. However, due to the discretization errors there is a risk of over- or underestimation on large distances $\left\|\mathbf{v}-\mathbf{v}_{s i}\right\|$. The (erroneous) overestimation may even lead to negative gaps and is a second cause of corrections. The correction purpose is the same as before: to guarantee a (positive) minimum gap in the step length formula. In the second division the iteration is chosen per processor $p$, common for every segment handled there (except for the current updating group); in the third division the iteration is chosen per segment. The use of information for all the previous iterations resembles a Dantzig-Wolfe approach - see Sec 9.4 below. But here the extrapolation is to the current point $\mathbf{v}$, not to the next intended point, as in DW.

| \#Partial up- <br> datings $M$ | Reduction <br> factor | Considered <br> iterations | CPU-hrs | Best dual <br> $(\mathrm{kSEK})$ | Best primal <br> $(\mathrm{kSEK})$ | Relative gap |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | | $\\|\mathbf{g}\\|_{2}$ |
| ---: |
| $(\mathrm{kSEK})$ |

Table 9.3a Results for the subgradient method with full and partial updating, reference case.
The use of bigger $M$-values means more communication with the root processor and an increased central run time $T_{\text {root }}$ for the $\mathbf{v}$-updating. This is evident from the run times for $M=50$ in Tab 9.3a (although its cost values - within parenthesis - are for a slightly different model). As for $M=10$, there are no dramatic differences between the different choices of reduction factor $1 / M^{q}$ but the results for $q>1$ are slightly inferior, as expected. It is somewhat surprising that the $p$ - and $s$-specific variants do better for a $1 / 1$-reduction, since the gap ought to be more accurate here. On the other hand the $M$-repetition argument, practised for variant latest, is irrelevant here: In the runs for variant latest, the angle between any two consecutive $\mathbf{v}$-directions is small, on average $24^{\circ}$ for $q=\frac{1}{2}$, i.e. the partial updating accomplishes a slow steering. For the $s$-specific variant the angles are bigger, on average $47^{\circ}$ for $q=\frac{1}{2}$, i.e. not only the updating group contributes to the steering, when all groups are open for changes at every updating. The best choice according to the costs is to use $s$-specific iterations and $q=0$, i.e. no reduction in comparison with full updating. This choice also outperforms the usual full updating method - but for the run time. The additional computational work and communication with the root processor, that results from $s$-specific updating, amounts to $8-10 \%$ extra CPU-time - time which the usual full updating method would have used to find better solutions. But if we choose from the closeness to optimality, according to $\|\mathbf{g}\|_{2}$, the winner is found among those using the latest iteration for extrapolation.
This is reasonable, since the average extrapolation distance ought to be shorter; whereas the advantage of using all iterations is mainly for the gap size. In Fig 9.1 we used the $s$-specific variant with $q=0$.

### 9.2 Subgradient method with full and partial updating

### 9.2.3 Camerini - Fratta - Maffioli (CFM) modification

The CFM-method is described in [Camerini, Fratta and Maffioli (1975)]. Its purpose is to avoid zigzag behaviour for the subgradient method, when consecutive iterates, e.g., fall on opposite sides of a sharp ridge, leading to slow convergence, back and forth across the ridge.

Consider the latest two dual (partial) iterates $\left(\boldsymbol{v}_{i}\right)_{i=0,1}$ and the corresponding objective values $\Phi_{i} \equiv \Phi\left(\mathbf{v}_{i}\right)$ and subgradients $G_{i} \equiv G\left(\mathbf{v}_{i}\right)$ for $i=0,1$. The corresponding affine majorants are

$$
\left(\Phi(\mathbf{v}) \approx \hat{\Phi}_{i}\left(\mathbf{v} ; \mathbf{v}_{i}\right) \equiv \Phi_{i}+\mathbf{G}_{i}^{T}\left(\mathbf{v}-\mathbf{v}_{i}\right) \quad i=0,1\right.
$$

The current iterate $\mathbf{v}_{1}$ is possibly (but not necessarily) the result of a preceding CFMmodification, $\mathbf{v}_{1}=\hat{\mathbf{v}}_{0}$. We determine if the angle between the two subgradients $\left(\mathbf{G}_{i}\right)_{i=0,1}$ is obtuse - or with $\Delta \mathbf{v}:=\mathbf{v}_{1}-\mathbf{v}_{0}$, rather if $\mathbf{G}_{1}{ }^{T} \Delta \mathbf{v}<0, \mathbf{G}_{0}{ }^{T} \Delta \mathbf{v}>0$ and $\mathbf{G}_{1}{ }^{T} \mathbf{G}_{0}<\left\|\mathbf{G}_{1}\right\| \cdot\left\|\mathbf{G}_{0}\right\|$ (Cauchy's inequality for non-aligned vectors $\mathbf{G}_{1}, \mathbf{G}_{0}$, i.e. the step from $\mathbf{v}_{0}$ to $\mathbf{v}_{1}$ is (somewhat) along $\mathbf{G}_{0}$ and (somewhat but not quite) opposite to $\mathbf{G}_{1}$ ). If so, we look for a direction in the ( $\mathbf{v}, \Phi)$-space, in the "ridge" subspace defined by the two majorants, and starting at a point $\hat{\mathbf{v}}_{1}$ on the line between $\mathbf{v}_{0}$ and $\mathbf{v}_{1}$ :

$$
\left\{\begin{array}{c}
\Phi_{1}+\mathbf{G}_{1}^{T}\left(\hat{\mathbf{v}}_{1}-\mathbf{v}_{1}\right)=\Phi_{0}+\mathbf{G}_{0}^{T}\left(\hat{\mathbf{v}}_{1}-\mathbf{v}_{0}\right) \\
\hat{\mathbf{v}}_{1}=\mathbf{v}_{1}+t \cdot\left(\mathbf{v}_{0}-\mathbf{v}_{1}\right)
\end{array}\right.
$$

Letting $\Delta \Phi: \equiv \Phi_{1}-\Phi_{0}$ and $\Delta \mathbf{G}:=\mathbf{G}_{1}-\mathbf{G}_{0}$, we get the step length

$$
t=\frac{\Delta \Phi-\mathbf{G}_{0}^{T} \Delta \mathbf{v}}{\Delta \mathbf{G}^{T} \Delta \mathbf{v}}
$$

where the assumptions $\Delta \mathbf{G}^{T} \Delta \mathbf{v}=\mathbf{G}_{1}{ }^{T} \Delta \mathbf{v}-\mathbf{G}_{0}{ }^{T} \Delta \mathbf{v}<0$ and the majorant property

$$
\Phi\left(\mathbf{v}_{1-i}\right) \leq \Phi_{i}+\mathbf{G}_{i}^{T}\left(\mathbf{v}_{1-i}-\mathbf{v}_{i}\right) \quad i=0,1
$$

imply $t \in[0,1]$.
Especially, we look for a direction $\hat{\mathbf{G}}_{1}$ in the convex cone formed by $\left(\mathbf{G}_{i}\right)_{i=0,1}$, as described by

$$
\hat{\mathbf{G}}_{1}:=\alpha \mathbf{G}_{0}+(1-\alpha) \mathbf{G}_{1} \quad \alpha \in[0,1] .
$$

Since also $\hat{\mathbf{v}}_{1}+\hat{\mathbf{G}}_{1}$ shall belong to the ridge subspace,

$$
\Phi_{1}+\mathbf{G}_{1}^{T}\left(\hat{\mathbf{v}}_{1}+\hat{\mathbf{G}}_{1}-\mathbf{v}_{1}\right)=\Phi_{0}+\mathbf{G}_{0}^{T}\left(\hat{\mathbf{v}}_{1}+\hat{\mathbf{G}}_{1}-\mathbf{v}_{0}\right)
$$

we get $\Delta \mathbf{G}^{T} \hat{\mathbf{G}}_{1}=0$, in that sense conjugate directions, and

$$
\alpha=\frac{\Delta \mathbf{G}^{T} \mathbf{G}_{1}}{\Delta \mathbf{G}^{T} \Delta \mathbf{G}}
$$

The next iterate is taken as $\mathbf{v}_{2}:=\hat{\mathbf{v}}_{1}+\sigma \cdot \hat{\mathbf{G}}_{1}$ where the step length $\sigma$ is chosen as in Sec 4.2.1, with $\|g\|^{2}$ exchanged for $\left\|\hat{\mathbf{G}}_{1}\right\|^{2}$ or $\left\|\hat{\mathbf{G}}_{1}\right\| \cdot\|g\|$. The assumption corresponding to $\mathbf{G}_{0}{ }^{T} \Delta \mathbf{v}>0$ in the next dual iteration is $\mathbf{G}_{1}{ }^{T}\left(\mathbf{v}_{2}-\mathbf{v}_{1}\right)>0$. The automatic satisfaction of this inequality follows from the assumed non-alignment of $\mathbf{G}_{1}, \mathbf{G}_{0}$.

We have tested both norm variants. Since partial updating generates acute angles, according to Sec 9.2.2, we have only applied the CFM-modification to full updating. In Tab 9.3b the main results are documented, and the standard run repeated from Tab 9.3a, for convenience.

| CFM- | Norm | Reduction <br> factor | Considered <br> iterations | CPU-hrs | Best dual <br> (kSEK) | Best primal <br> (kSEK) | Relative gap | $\\|\mathbf{g}\\|_{2}$ <br> (kSEK) |
| ---: | :---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| No | $\\|\mathbf{g}\\|^{2}$ | $1 / 1$ | latest | 127 | 2541764 | 2541544 | -0.000087 | 11287 |
| Yes $\\|\hat{\mathbf{G}}\\| \cdot\\|\mathbf{g}\\|$ | $1 / 1$ | latest | 125 | 2539582 | 2542974 | 0.001336 | 70859 |  |
| Yes | $\\|\hat{\mathbf{G}}\\|^{2}$ | $1 / 1$ | latest | 125 | 2541328 | 2542319 | 0.000390 | 24827 |

Table 9.3b Subgradient method with full updating and two variants of CFM-modification.
As we can see the CFM-modifications do not mean any improvement, quite the opposite.

### 9.3 Interpolation improvements

The two interpolation schemes presented here are intended to supplement all the price updating methods discussed in this chapter. The aim of both schemes is to improve the accuracy, by offering an alternative to the expensive increase of the node density - cf. Tab 4.3. Our first way is to apply higher order polynomials than in standard multi-linear interpolation. The results for pairs of quadratics are mixed into multi-linear-quadratic interpolation for better utilisation of the existing lattice information, catching more of the global function behaviour, rather than to increase the node density. Our second way is the opposite, i.e. to introduce more nodes in an economic way, $(L-1)^{D}$ nodes in between the existing $L^{D}$ ones, and stick to multi-linear interpolation. The resulting double grid may be viewed, e.g., as $L=5 \frac{1}{2}$ levels per state dimension.

### 9.3.1 Multi-linear-quadratic interpolation

Consider the interpolation problem of finding a function value $f(\hat{\mathbf{x}})$ for a point $\hat{\mathbf{x}}$ in a $D$ dimensional state space, based on given function values in a surrounding regular lattice. We demand that the estimated $\hat{f}(\hat{\mathbf{x}})$-values should define a continuous function throughout the state space. This means that, if we think of full multi-quadratic interpolation analogous to multi-linear interpolation, it cannot be based on just three levels per state dimension. (For example, assuming integer node levels, the three levels $x_{d}=0,1,2$ are natural to use for $\hat{x}_{d}<1.5$ and $x_{d}=1,2,3$ for $\hat{x}_{d}>1.5$. But the exchange does not guarantee continuity, since the interpolation weights for $x_{d}=0$ and $x_{d}=3$, respectively, at $\hat{x}_{d}=1.5$ are non-zero.) On the other hand, using four node levels per dimension (which guarantees continuity, since she shift takes place on the very node levels where just one weight is non-zero) would increase the number of interpolation points from $2^{D}$ to $4^{D}$, i.e. 32 times for $D=5$. Instead we have tried to find a compromise - but the demands for continuity still makes it necessary to use four node levels, whenever possible. Our scheme consists of two steps:
a) Consider the hyperbox, say hypercube, enclosing $\hat{\mathbf{x}}$ and having lattice nodes as corner points $\mathbf{p}$ and no node in its interior. For each state dimension $d$ and each edge line $e_{d}(\mathbf{p})$ in the $d$-coordinate direction from $\mathbf{p}$, determine by (mixed) quadratic interpolation the function value for the target point on $e_{d}(\mathbf{p})$, i.e. the one having $x_{d}=\hat{x}_{d}$.
b) Given the determined function values for all the hypercube edges, compute the function value at $\hat{\mathbf{x}}$ by edge based multi-linear interpolation.

In Fig 9.3a we illustrate step (a) for one of the edge lines $e=e_{d}(\mathbf{p})$ in $D=3$ dimensions, with four interpolation points (white balls). The target point (black ball on $e$ ) is on the (hyper-) plane containing point $\hat{\mathbf{x}}$. In Fig 9.3b all the edge points used in step (b) are marked. In standard multi-linear interpolation the weight points are at the surrounding hypercube corners. Here the weight points on the edges $e_{d}(\cdot)$ in state dimension $d$ form the hypercube corners in a $D-1$ dimensional subspace or manifold, i.e. step (b) corresponds to $D$ ordinary multi-linear interpolations in $D-1$ dimensions.

Figure 9.3a Mixed quadratic interpolation Figure 9.3b Edge based multi-linear interpolation


In summary, our strategy for generalising the multi-linear interpolation method is to use quadratic polynomials for the edges in each state dimension, i.e. multi-linear-quadratic interpolation. For the variation along each edge line in a considered hypercube (hyperbox) we normally use 4 points in the state space - one grid point on each side of two hypercube corner points; the latter ones indexed 0 and 1 . At or outside the grid surface we only have access to grid points on one side of the hypercube, and use 3 points (without jeopardising continuity).

In order to judge the computational work, an optimistic view is to consider the computational time for step (a), i.e. the quadratic interpolation for getting the target point function values, as negligible. The remaining step (b), i.e. the edge based multi-linear interpolation, is based on the $D \cdot 2^{D-1}$ target points (edge lines). For $D=5$ this is 2.5 times the usual (multi-linear interpolation) number.

A pessimistic view is to focus on the 4 grid points used per edge line, i.e. $D \cdot 2^{D+1}$ quadratic interpolation points in all. Thus, for $D=5$ we will handle 10 times the usual number of points per hypercube. Somewhere in between these two factors, 2.5 and 10 , is a realistic figure for the increase of the computational burden.

Continuity is guaranteed when shifting hypercube, since when moving orthogonally to a common edge line, an unchanged set of interpolation points on the edge are used and when
moving parallel to an edge line the shift of hypercube coincides with the exchange of one set of interpolation points - with weight zero at the shift, since the nodes on the boundary hyperplane will have all weight when passing it.

In order to derive the interpolation point weights, we assume that the points on an edge line $e=e_{d}(\cdot)$ in dimension $d$ are described by a normalized state variable $x_{d}$, where $x_{d}=0,1$ correspond to the two hypercube corners. Since we work with a regular grid the node positions $x_{d}$ are common to all parallel edge lines (in dimension $d$ ). We make one or two quadratic interpolations, each based on 3 points. For each quadratic interpolation the ansatz is the function form

$$
\hat{f}_{e}\left(\hat{x}_{d}\right)=c_{e 0} \cdot\left(1-\hat{x}_{d}\right)^{2}+c_{e 1} \cdot \hat{x}_{d}^{2}+c_{e 2} \cdot 2 \hat{x}_{d}\left(1-\hat{x}_{d}\right)
$$

symmetric as to $\hat{x}_{d}, 1-\hat{x}_{d}$.
Case 1: 3 interpolation points $x_{d}=0,1, x_{d 2}$ with computed function values $f_{e 0}, f_{e 1}, f_{e 2}$. The ansatz implies

$$
\begin{align*}
& c_{e 0}=f_{e 0}, c_{e 1}=f_{e 1}, c_{e 2}=\left(f_{e 2}-f_{e 0} \cdot\left(1-x_{d 2}\right)^{2}-f_{e 1} x_{d 2}^{2}\right) /\left(2 x_{d 2}\left(1-x_{d 2}\right)\right), \\
& \hat{f}_{e}\left(\hat{x}_{d}\right)=f_{e 0} \cdot\left(1-\hat{x}_{d}\right)\left(1-\frac{\hat{x}_{d}}{x_{d 2}}\right)+f_{e 1} \cdot \hat{x}_{d} \frac{\hat{x}_{d}-x_{d 2}}{1-x_{d 2}}+f_{e 2} \cdot \hat{x}_{d} \frac{1-\hat{x}_{d}}{x_{d 2}\left(1-x_{d 2}\right)} \tag{9.2a}
\end{align*}
$$

where we identify the $f_{e i}$-coefficients as the weights $w_{d i}\left(\hat{x}_{d}\right) \quad i=0,1,2$ of the computed function values for every edge $e=e_{d}(\cdot)$ in dimension $d$. The weights on each edge sum up to 1 .

Case 2: Two sets of 3 interpolation points $x_{d}=0,1, x_{d 2}$ and $x_{d}=0,1, x_{d 3}$, with computed function values $f_{e 0}, f_{e 1}, f_{e 2}$ and $f_{e 0}, f_{e 1}, f_{e 3}$, respectively.
For each set we receive a $\hat{f}_{e}\left(\hat{x}_{d}\right)$-expression as in case 1 . By mixing the two expressions into a final arithmetic mean we get

$$
\begin{align*}
& \hat{f}_{e}\left(\hat{x}_{d}\right)=f_{e 0} \cdot\left(1-\hat{x}_{d}\right)\left(1-\frac{\hat{x}_{d}}{2} \cdot\left[\frac{1}{x_{d 2}}+\frac{1}{x_{d 3}}\right]\right)+f_{e 1} \hat{x}_{d}\left(1-\frac{1-\hat{x}_{d}}{2} \cdot\left[\frac{1}{1-x_{d 2}}+\frac{1}{1-x_{d 3}}\right]\right)  \tag{9.2b}\\
& \quad+f_{e 2} \frac{\hat{x}_{d}\left(1-\hat{x}_{d}\right)}{2 x_{d 2}\left(1-x_{d 2}\right)}+f_{e 3} \frac{\hat{x}_{d}\left(1-\hat{x}_{d}\right)}{2 x_{d 3}\left(1-x_{d 3}\right)}
\end{align*}
$$

The $f_{e i}$-coefficients act as weights $w_{d i}\left(\hat{x}_{d}\right) \quad i=0,1,2,3$ on each edge $e=e_{d}(\cdot)$ in state dimension $d$. If, especially, $f$ denotes state variable values, $f_{e i}=x_{d i} \forall i$ implies $\hat{f}_{e}\left(\hat{x}_{d}\right)=\hat{x}_{d}$, without error; likewise if $f$ is quadratic (but step (b) will introduce errors in the quadratic case).

Theorem 9.1 If $\hat{\mathbf{x}}$ is strictly in between the two hyperplane boundaries of the circumscribed hyperbox that are orthogonal to dimension $d$, then the interpolation point weights in (9.2b) satisfy

$$
w_{d 0}\left(\hat{x}_{d}\right), w_{d 1}\left(\hat{x}_{d}\right)>0, w_{d 2}\left(\hat{x}_{d}\right), w_{d 3}\left(\hat{x}_{d}\right)<0 .
$$

Proof By assumption $0<\hat{x}_{d}<1$. The outer interpolation points satisfy $x_{d 2}<0, x_{d 3}>1$. Hence the negativity of the $f_{e 2}$ - and $f_{e 3}$-coefficients is evident by inspection. The $f_{e 0}$-weight is

$$
\begin{aligned}
& \left(1-\hat{x}_{d}\right)\left(1-\frac{\hat{x}_{d}}{2} \cdot\left[\frac{1}{x_{d 2}}+\frac{1}{x_{d 3}}\right]\right)=\left(1-\hat{x}_{d}\right)\left(1+\frac{\hat{x}_{d}}{2 \cdot\left|x_{d 2}\right|}-\frac{\hat{x}_{d}}{2 x_{d 3}}\right)> \\
& >\left(1-\hat{x}_{d}\right)\left(1+0-\frac{1}{2}\right)>0
\end{aligned}
$$

analogously for the $f_{e 1}$-weight.
In consideration of all $D$ state dimensions the multi-linear interpolation weight for a normalized corner point $\mathbf{p}$, i.e. taking $0 / 1$-coordinates, can be written

$$
\begin{equation*}
W_{\mathbf{p}}(\hat{\mathbf{x}}):=\prod_{d: p_{d}=0}\left(1-\hat{x}_{d}\right) \cdot \prod_{d: p_{d}=1} \hat{x}_{d} . \tag{9.3}
\end{equation*}
$$

All the interpolation points on an edge in dimension $d$ share all coordinates but $x_{d}$. For interpolation step (b) we have:

Lemma 9.1: Let $e_{d}(\mathbf{p})$ denote the edge in dimension $d$ passing through corner point $\mathbf{p}$.
The $e_{d}(\mathbf{p})$-weight in step (b) is

$$
W_{\mathbf{p}}(\hat{\mathbf{x}}) /\left(1-\hat{x}_{d}\right) \text { if } p_{d}=0, \hat{x}_{d} \neq 1 \text { and } W_{\mathbf{p}}(\hat{\mathbf{x}}) / \hat{x}_{d} \text { if } p_{d}=1, \hat{x}_{d} \neq 0
$$

Proof: In the subspace $d^{\prime} \neq d$ of $D-1$ dimensions all the points on $e_{d}(\mathbf{p})$ correspond to one and the same corner point $\left(p_{d^{\prime}}\right)_{d^{\prime} \neq d}$. For such a point the multi-linear interpolation weight in step (b) - cf. (9.3) - is $\prod_{d^{\prime} \neq d: p_{d^{\prime}}=0}\left(1-\hat{x}_{d^{\prime}}\right) \cdot \prod_{d^{\prime} \neq d: p_{d^{\prime}}=1} \hat{x}_{d^{\prime}}$.

Since each edge contains two nodal points of the hypercube we choose the one closest to $\hat{\mathbf{x}}$ for handling the edge in step (b):

Theorem 9.2: The $\hat{f}(\hat{\mathbf{x}})$-value estimated in step (b) is

$$
\frac{1}{D} \cdot \sum_{\mathbf{p}} W_{\mathbf{p}}(\hat{\mathbf{x}}) \cdot\left(\sum_{d: \hat{x}_{d} \leq \frac{1}{2}, p_{d}=0} \hat{f}_{e_{d}(\mathbf{p})}\left(\hat{x}_{d}\right) /\left(1-\hat{x}_{d}\right)+\sum_{d: \hat{x}_{d}>\frac{1}{2}, p_{d}=1} \hat{f}_{e_{d}(\mathbf{p})}\left(\hat{x}_{d}\right) / \hat{x}_{d}\right)
$$

Proof: The computed value is the arithmetic mean of $D$ multi-linear interpolations in $D-1$ dimensions. For each interpolation (9.2) and Lemma 9.1 apply.

What advantages does this interpolation scheme have? It does not increase the storage demands, in comparison with multi-linear interpolation. (In practice $10 \%$ less storage is used, since we for the resulting states of routine maintenance prefer storing weight factors per state dimension $d$ instead of pre-computed full weights $W_{\mathbf{p}}$ per neighbour node.) Moreover, if the function $f$ is linear or quadratic, the edge target points will get correct $f$-values and the distance between a point at the centre of the hypercube will have the average interpolation distance (to the black balls in Fig 9.3b) reduced by the factor $1-1 / D\left(\ell_{1}\right.$-norm) or $\sqrt{1-1 / D}$ ( $\ell_{2}$-norm); for $D=5$ meaning factor 0.8 and 0.89 , respectively. If the alternative option is to increment the grid density from $L$ to $L+1$ node levels per state dimension and to use multi-linear interpolation, the corresponding reduction factor is $1 /(1+L)$; for $L=5$ factor 0.83 and for
$L=60.86$, at the price of multiple storage demands. Both alternatives mean multiple CPU times.

In Tab 9.4 two runs Quad with multi-linear-quadratic interpolation are reported. We also repeat three runs of interest from Tab 9.1, Std, where standard multi-linear interpolation were used. ( $D b l$ will be treated below.) Remember that the best dual and primal values for different number of node levels $L$ cannot be directly compared (due to differing residual values; see comment at Tab 9.1). In column "CPU-hrs" we notice that the Quad runs took $2.2-3$ times as long as the corresponding multi-linear interpolation runs for the same number of dual iterations. Whereas the best primal value is better for $Q u a d$, the best dual and residual $\|\mathbf{g}\|_{2}$ is inferior. However, the discretization error, as measured by the backward-forward cost differences for $L=4$ is $-2.8 \%$ for Quad and $-3.8 \%$ for $S t d$ (cf. Tab 4.3); for $L=5-1.9 \%$ and $-2.4 \%$, respectively. These Quad-errors match the $S t d$-errors on the next $L$-level. Also the CPU-times are of equal size. Thus in a future situation where processor capacity rather than CPU time is decisive, a higher-order interpolation for, e.g., $L=5$ may be preferred to standard multi-linear interpolation for $L=6$.

| \#Node levels L | Interpol. method | \#States | $\begin{aligned} & \text { \#Pro- } \\ & \text { cessors } \end{aligned}$ | CPU-hrs | \#Dual iterations | Best dual (kSEK) | Best primal (kSEK) | Relative gap | $\begin{array}{r} \\|\boldsymbol{g}\\|_{2} \\ \text { (KSEK) } \end{array}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4 | Std | 1024 | 30 | 90 | 100 | 2545752 | 2537831 | -0.003111 | 17743 |
| 5 | Std | 3125 | 66 | 264 | 76 | 2541856 | 2541485 | -0.000146 | 11933 |
| 6 | Std | 7776 | 166 | 664 | 71 | 2559947 | 2559993 | 0.000018 | 7039 |
| 4 | Quad | 1024 | 30 | 270 | 50 | 2534045 | 2528263 | -0.002282 | 21487 |
| 5 | Quad | 3125 | 66 | 592 | 26 | 2537217 | 2537136 | -0.000032 | 23236 |
| $41 / 2$ | Dbl | 1267 | 30 | 127 | 26 | 2573109 | 2569333 | -0.001467 | 12272 |
| $51 / 2$ | Dbl | 4149 | 66 | 457 | 26 | 2546861 | 2546382 | -0.000188 | 15297 |

Table 9.4 Results for different interpolation methods, number of node levels and processors.

### 9.3.2 Double grid interpolation

In this subsection we will examine multi-linear interpolation using an extra grid inside an existing grid of nodal states; e.g., in a 3D cubic lattice, where each cube has nodal corners and lacks interior nodes, we put an extra node at every cube centre. By using information about the value function from both grids the average interpolation distance, i.e. the distance between a given point $\mathbf{x}$ and the neighbour grid nodes, is reduced. We start by formulating some general properties for interpolation distances, using two different metrics. Then we focus on the double multi-linear interpolation, stating some properties and discussing some implementation results for Värmland. For a simple description we partly assume an odd number of state dimensions $D$. (In case study Värmland $D=5$.)

## Maximin interpolation distance

Consider, e.g., a hypercube box of side $a$ with nodal corners. Here the mean $\ell_{1}$-distance between the set of enclosed, uniformly distributed points and the closest corner node is $\frac{a D}{4}$. To realize this we introduce a coordinate system, placing the box in the first hyper-octant, with the origin $\mathbf{0}$ and $a \cdot \mathbf{1}$ as opposite box corners. Subdividing the box into $2^{D}$ equally sized subboxes, each extended between the hypercube mid point $\frac{a}{2} \cdot \mathbf{1}$ and a box corner, we focus on an

### 9.3 Interpolation improvements

arbitrary sub-box, e.g. one having the origin and the hypercube mid point as opposite corners. All the enclosed states $\mathbf{x}$ share $\mathbf{0}$ as the closest node. The $\ell_{1}$-distance between $\mathbf{x}$ and $\mathbf{0}$ is $\sum_{d=1}^{D} x_{d}$, and each $x_{d}$ varies independently in $0 \leq x_{d} \leq \frac{a}{2}$. Thus the average distance is $D \cdot \frac{a}{4}$.

The interpolation error is expected to go to zero linearly with $a$. To accomplish an error halving we need twice as many node levels per state dimension, i.e. $2^{D}$ as many nodes (and computational work), e.g. for $D=5$ meaning 32 times as many nodes.

What if the distance to the closest node has an important influence on the rate of convergence? Then what is the worst case, i.e. the maximum possible such distance, for a double grid?

Theorem 9.3: Assume that the basic inifinite grid consists of hypercubes of side length $a$, supplied with an extra nodal point in the middle. In $\mathcal{R}^{D}$ with Euclidean metric the distance to the nearest grid point is at most $\frac{a}{4} \cdot \sqrt{D+2 \cdot\left[\frac{1}{2} D\right]}$ (where $\left.L\right\rfloor$ denotes integer part).

Proof: Consider a hypercube between $\mathbf{0}$ and $\mathbf{a}$, with nodal mid point $\frac{1}{2} \cdot \mathbf{a}=\frac{a}{2} \cdot \mathbf{1}$. The hypercube consists of $2^{D}$ equally sized sub-hypercubes (subboxes), each with the midpoint node and one of the corner nodes as opposite corner points, and no other node points. Each point $\mathbf{x}$ in the hypercube belongs to at least one subbox, and by symmetry a point belonging to more than one subbox has coincident distances to the two corresponding subbox nodes ( $\frac{1}{2} \mathbf{a}$ is a common node and the distances to the respective corner nodes coincide). Therefore, without loss of generality, we may consider $\mathbf{x}$ belonging to one specific subbox (or more), e.g. the one with nodes $\frac{1}{2} \mathbf{a}$ and $\mathbf{a}$.

At first we show that these two nodes are closest to $\mathbf{x}$, among all the grid nodes. For simplicity (and without loss of generality) we consider $a=2$, to have $\frac{1}{2} \mathbf{a}=(1,1, \ldots, 1), \mathbf{a}=(2,2, \ldots, 2)$ and $x_{d} \in[1,2] \forall d$. For $a=2$ all the corner nodal points have even integer co-ordinates and all the mid points have odd integer co-ordinates, and every subbox has one node of each type. $\ell_{2}$ norm means that (the square-root of) a sum of co-ordinate differences, squared, is considered.

For the closest grid point, a first contradictory assumption would be that some co-ordinate is 3 , instead of 1 . But since $x_{d} \in[1,2]$ we have $\left(x_{d}-3\right)^{2} \geq\left(x_{d}-1\right)^{2}$ and get a shorter distance by replacing 3 for 1 .
A second contradictory assumption would be that some co-ordinate is 0 , instead of 2 . But since $x_{d} \in[1,2]$ we have $\left(x_{d}-0\right)^{2} \geq\left(x_{d}-2\right)^{2}$ and get a shorter distance by replacing 0 for 2 . Other integer values fit worse. Hence only $(1,1, \ldots, 1)$ and $(2,2, \ldots, 2)$ remain as candidates.

It is obvious for a maximum distance case that (at least) two node points should be at one and the same distance. Without loss of generality we consider all states $\mathbf{x}$ having two specific closest node points: the mid node point $\frac{1}{2} \cdot \mathbf{a}$ and the outer end point $\mathbf{a}$ on the main diagonal, i.e. $\left\|\mathbf{x}-\frac{1}{2} \mathbf{a}\right\|=\|\mathbf{a}-\mathbf{x}\|$. For the Euclidian norm (metric), squared, this corresponds to (the
constraint) $\sum_{d}\left(x_{d}-\frac{1}{2} a\right)^{2}=\sum_{d}\left(a-x_{d}\right)^{2}$, i.e. $\sum_{d} x_{d}=\frac{3 a D}{4}$. Since we have assumed $\frac{1}{2} \cdot \mathbf{a}$ and a to be closest, let us consider the neighbour nodes of $\mathbf{a}$, i.e. those connected to a by edge lines on the hypercube surface. Since they differ from a in one dimension only, taking 0 instead of $a$ as one coordinate, we realize that the unknown point $\mathbf{x}$ coordinates are confined to $\frac{1}{2} a \leq x_{d} \leq a$. Equivalently working with the norm squared, and preferably with a minimization formulation, we write the maximum distance (between $\mathbf{x}$ and $\mathbf{a}$ ) problem:

$$
\text { minimize }-\sum_{d}\left(x_{d}-a\right)^{2} \text { s.t. }\left\{\begin{array}{l}
\sum_{d} x_{d}=\frac{3 a D}{4} \\
\frac{1}{2} a \leq x_{d} \leq a \quad d=1, \ldots, D
\end{array}\right.
$$

By introducing the Lagrangean

$$
L(\mathbf{x} ; \lambda, \boldsymbol{\mu}, \mathbf{v}):=-\sum_{d}\left(x_{d}-a\right)^{2}+\lambda \cdot\left(\frac{3 a D}{4}-\sum_{d} x_{d}\right)+\sum_{d} \mu_{d} \cdot\left(x_{d}-a\right)+\sum_{d} v_{d} \cdot\left(\frac{1}{2} a-x_{d}\right),
$$

three Karush-Kuhn-Tucker cases can be identified:

1. $x_{d}=a, v_{d}=0, \mu_{d}=\lambda(\geq 0)$
2. $x_{d}=\frac{1}{2} a, \mu_{d}=0, \quad v_{d}=a-\lambda(\geq 0)$
3. $\quad \frac{1}{2} a<x_{d}<a \quad \mu_{d}=v_{d}=0 \quad x_{d}=a-\frac{1}{2} \lambda$.

Thus there are only three possible values for each $x_{d}$ to take. Varying $d$ we let $m_{1}, m_{2}, m_{3}$ denote the number of (state) dimensions for which the respective cases 1,2 and 3 apply. Let us express the problem in $\left(m_{k}\right)_{k=1}^{3}$ instead. The primary constraints are

$$
\left\{\begin{array}{c}
m_{1}+m_{2}+m_{3}=D \\
m_{1} \cdot a+m_{2} \cdot \frac{1}{2} a+m_{3} \cdot\left(a-\frac{1}{2} \lambda\right)=\frac{3 a D}{4}
\end{array}\right.
$$

By eliminating $m_{1}$ we get $m_{2}+m_{3} \cdot \frac{\lambda}{a}=\frac{1}{2} D$ and insert into the distance squared:

$$
\begin{aligned}
& \text { maximize } w:=\sum_{d}\left(x_{d}-a\right)^{2}=\frac{1}{4} a^{2} \cdot\left(m_{2}+m_{3} \cdot\left(\frac{\lambda}{a}\right)^{2}\right)= \\
& =\left\{\begin{array}{cl}
\frac{1}{4} a^{2} \cdot \frac{1}{2} D & m_{3}=0, \\
\frac{1}{4} a^{2} \cdot\left(m_{2}+\frac{\left(\frac{1}{2} D-m_{2}\right)^{2}}{m_{3}}\right) & m_{3}>0 .
\end{array}\right.
\end{aligned}
$$

Thus for $m_{3}=0$ the only optimum candidate is $m_{1}=m_{2}=\frac{1}{2} D$, integer valued only for $D$ even. Otherwise, we temporarily consider the expression for $m_{3}>0$ as if $m_{2}$ is a continuous variate. By differentiating the distance function $w$ for $m_{2}$ and putting the partial derivative to zero we establish (partial) stationarity for $m_{2}=\frac{1}{2} \cdot\left(D-m_{3}\right)$, i.e. for $m_{3}=D-2 m_{2}$. Since the latter equality can be fully treated as an integer relation, it specifies the remaining optimum candidates without rounding off. Also the case $m_{3}=0$ can be incorporated into this general relation. By inserting into the objective we get

$$
w=\frac{1}{4} a^{2} \cdot\left(\frac{1}{4} D+\frac{1}{2} m_{2}\right)
$$

### 9.3 Interpolation improvements

i.e. $w$ increases with $m_{2}$. Since $m_{3} \geq 0$ the optimum solution is $m_{2}=\left\lfloor\frac{1}{2} D\right\rfloor$, $m_{3}=D-2 m_{2}=D-2 \cdot\left\lfloor\frac{1}{2} D\right\rfloor, m_{1}=D-m_{2}-m_{3}=\left\lfloor\frac{1}{2} D\right\rfloor$.

As for the maximum distances, what is to be preferred: a refined hyper-cubic lattice or the addition of mid point nodes? With a basic grid of $L$ node levels and $D$ state dimensions an extra node level per dimension means $(L+1)^{D}-L^{D}$ extra nodes and (equidistant) node level distances $a \cdot \frac{L}{L+1}$ instead of $a$. The maximum possible $\ell_{2}$-distance to the nearest node, attained at the (empty) mid point, is $a \cdot \frac{L}{L+1} \cdot \frac{1}{2} \sqrt{D}$, whereas the addition of mid point nodes means $(L-1)^{D}$ extra nodes and a maximum distance according to Thm 9.3. For, e.g., $D=5$ and $L=4$ the former means maximum distance $0.8944 a$ and 2101 more nodes, and the latter $0.75 a$ and 243 , respectively. For $D=5$ and $L=5$ the corresponding figures are $0.9317 a$ and 4651 for the former, and $0.75 a$ and 1024 for the latter. These results speak clearly for the latter choice. We expect nonlinear behaviour, especially close to the boundary of the state region, with discretization errors (linked to maximum distance) that might have a heavy impact on the overall precision. However, also the mean distance should be relevant for the general accuracy. Moreover, interpolation does not rely on the closest node only.

## Expected interpolation distance

Consider the states $\mathbf{x}$ in a rectilinear box occupying the $D$-dimensional interval set $\left[\mathbf{p}_{0}, \mathbf{p}_{0}+\mathbf{a}\right]$, where $\mathbf{a}$ is the main diagonal vector and $\mathbf{p}_{0}$ is the basic corner nodal state. Nodes are at all the box corners. We turn to normalized state coordinates $\mathbf{y} \in[0,1]^{D}$ by the affine transformation $y_{d}=\left(x_{d}-p_{0 d}\right) / a_{d} \quad \forall d$. The matrix of corner nodal states turns into $\mathbf{Q}=\left(\mathbf{q}_{n}\right)_{n=1}^{2^{D}}$, where the columns $\left(\mathbf{q}_{n}\right)$ consist of all the different sequences of $0 / 1$-coordinates. For a given corner node $n$ let $\mathcal{D}_{n 0}, \mathcal{D}_{n 1}$ denote the two subsets of the $D$ state dimensions whose coordinates are 0 and 1 , respectively. In multi-linear interpolation the weight assigned to $\mathbf{q}_{n}$ can be written

$$
w\left(\mathbf{y}, \mathbf{q}_{n}\right):=\prod_{d \in \mathcal{D}_{n 1}} y_{d} \cdot \prod_{d \in \mathcal{D}_{n 0}}\left(1-y_{d}\right)
$$

The weight sum for the $2^{D}$ corner points equals 1 .
For Lipschitz continuous value functions we expect the multi-linear interpolation errors to go to zero linearly with a characteristic "interpolation distance" between the normalized state $\mathbf{y}$ and the interpolation basis $\mathbf{Q}$. We will use a distance function both for comparing different interpolation methods and for direct method specification. For simplicity we apply $\ell_{1}$-norm, evaluating the distance function $\rho$ between $\mathbf{y}$ and $\mathbf{q}_{n}$ as

$$
\rho\left(\mathbf{y}, \mathbf{q}_{n}\right):=\sum_{d}\left|y_{d}-q_{n d}\right|=\sum_{d \in \mathcal{D}_{n 1}}\left(1-y_{d}\right)+\sum_{d \in \mathcal{D}_{n 0}} y_{d} .
$$

For every $\mathbf{y}$ enclosed in the convex hull of the $\mathbf{Q}$-states, a hypercube, the unweighted (arithmetic) mean $\ell_{1}$-distance to the surrounding $2^{D} \mathbf{Q}$-states equals $\frac{D}{2}$. However, interpolation means weighing. Introducing weights $w$, by how much can the non-weighted mean be reduced?

Definition 9.1: By the interpolation distance between a normalized state $\mathbf{y}$ and the basis $\mathbf{Q}$ we mean

$$
\bar{\rho}(\mathbf{y}, \mathbf{Q}):=\sum_{n=1}^{2^{D}} w\left(\mathbf{y}, \mathbf{q}_{n}\right) \cdot \rho\left(\mathbf{y}, \mathbf{q}_{n}\right)
$$

Theorem 9.4: For multi-linear interpolation the interpolation distance between $\mathbf{y} \in[0,1]^{D}$ and the interpolation basis $\mathbf{Q}$ is

$$
\bar{\rho}(\mathbf{y}, \mathbf{Q})=2 \cdot \sum_{d=1}^{D} y_{d}\left(1-y_{d}\right)
$$

Proof: The distance can be written

$$
\begin{aligned}
& \bar{\rho}(\mathbf{y}, \mathbf{Q}):=\sum_{n=1}^{2^{D}} w\left(\mathbf{y}, \mathbf{q}_{n}\right) \cdot \rho\left(\mathbf{y}, \mathbf{q}_{n}\right)=\sum_{n=1}^{2^{D}} \prod_{d \in \mathcal{D}_{n 1}} y_{d} \cdot \prod_{d \in \mathcal{D}_{n 0}}\left(1-y_{d}\right) \cdot\left(\sum_{d \in \mathcal{D}_{n 1}}\left(1-y_{d}\right)+\sum_{d \in \mathcal{D}_{n 0}} y_{d}\right)= \\
& =\sum_{n}\left(\sum_{d \in \mathcal{D}_{n 1}}\left(1-y_{d}\right) \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 1}} y_{d^{\prime}} \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 0}}\left(1-y_{d^{\prime}}\right)+\sum_{d \in \mathcal{D}_{n 0}} y_{d} \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 0}}\left(1-y_{d^{\prime}}\right) \prod_{d^{\prime} \in \mathcal{D}_{n 1}} y_{d^{\prime}} \cdot\right)= \\
& =\sum_{n}\left(\sum_{d \in \mathcal{D}_{n 1}}\left(1-y_{d}\right) y_{d} \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 1} \backslash\{d\}} y_{d^{\prime}} \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 0}}\left(1-y_{d^{\prime}}\right)+\sum_{d \in \mathcal{D}_{n 0}} y_{d}\left(1-y_{d}\right) \prod_{\left.d^{\prime} \in \mathcal{D}_{n 0} \backslash d\right\}}\left(1-y_{d^{\prime}}\right) \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 1}} y_{d^{\prime}} \cdot\right)= \\
& =\sum_{d} \sum_{n: d \in \mathcal{D}_{n 1}}\left(1-y_{d}\right) y_{d} \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 1} \backslash\{d\}} y_{d^{\prime}} \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 0}}\left(1-y_{d^{\prime}}\right)+\sum_{d} \sum_{n: d \in \mathcal{D}_{n 0}} y_{d}\left(1-y_{d}\right) \prod_{d^{\prime} \in \mathcal{D}_{n 0} \backslash\{d\}}\left(1-y_{d^{\prime}}\right) \prod_{d^{\prime} \in \mathcal{D}_{n 1}} y_{d^{\prime}}= \\
& =\sum_{d: d \in \mathcal{D}_{n 1}} y_{d}\left(1-y_{d}\right) \cdot\left(\prod_{d^{\prime} \in \mathcal{D}_{n 1} \backslash\{d\}} y_{d^{\prime}} \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 0}}\left(1-y_{d^{\prime}}\right)+\sum_{n: d \in \mathcal{D}_{n 0}} \prod_{d^{\prime} \in \mathcal{D}_{n 1}} y_{d^{\prime}} \cdot \prod_{d^{\prime} \in \mathcal{D}_{n 0} \backslash\{d\}}\left(1-y_{d^{\prime}}\right)\right) .
\end{aligned}
$$

Here we identify the terms of the two interior sums in the last expression as the multi-linear interpolation weights in a ( $D-1$ )-dimensional state subspace (each interior sum has the $q_{n d}$ value fixed, respectively 1 and 0 ). Since each weight sum equals 1 the theorem follows.

According to Thm 9.4 the interpolation distance varies between 0 and $\bar{\rho}\left(\frac{1}{2} \cdot \mathbf{1}, \mathbf{Q}\right)=2 \cdot D \cdot \frac{1}{4}=\frac{D}{2}$.

For randomly distributed y's we use $\overline{\bar{\rho}}$ to denote the overall mean. We have:
Theorem 9.5: For uniformly distributed states within the $D$-dimensional hypercube spanned by the basis $\mathbf{Q}$, the expected interpolation distance for multi-linear interpolation is $\overline{\bar{\rho}}(\mathbf{Q})=\frac{D}{3}$.

Proof: By assumption the stochastic variables $\left(Y_{d}\right)_{d=1}^{D}$ vary equally, uniformly and independently in the rectilinear box, each on $Y \in[0,1]$. Due to the symmetry and the

### 9.3 Interpolation improvements

irrelevance of the other state dimensions but the one in focus, we get ( $E=$ expected value)

$$
\overline{\bar{\rho}}(\mathbf{Q}):=E(\bar{\rho}(\mathbf{Y}, \mathbf{Q}))=2 \cdot D \cdot \int_{0}^{1} y(1-y) d y \cdot\left(\int_{0}^{1} 1 d y\right)^{D-1}=2 D \cdot \frac{1}{6} \cdot 1 .
$$

In comparison with an arithmetic-mean interpolation (mentioned at Def 9.1), the multi-linear formula is expected to reduce the errors by the factor $\frac{2}{3}$.

## Double multilinear interpolation

We will consider a double rectilinear grid, with the nodal states of an inner lattice placed at the mid points of an outer grid - see Fig 9.4 for a 2D illustration. In Fig 9.4 the outer grid has white nodes and the inner grid black nodes. The purpose is to perform a multi-linear interpolation in two steps, thus keeping the calculations simple, in a way that guarantees continuity of any interpolated value function. If no explicit lower and upper state variable bounds are given, we must cover ourselves against the risk of extrapolation. (Extrapolation is unwanted also if normal multi-linear interpolation is used.) In such cases we extend the responsibility domain of the normal rectilinear boxes to outside the most extreme node levels for each of the two grids. In this way every state belongs to one (possibly extended) rectilinear box of each grid, and the final interpolate will be based on more global information than from one grid only.


By a transformation to normalized states we accomplish a double hypercube lattice - cf. Fig. 9.4, right, for 2D. Consider, e.g., the central quadratic box having the 4 interior (black grid) nodes as corners. For interpolation purposes we subdivide the box into $2^{D}=4$ equally sized sub-boxes, each extended between the hypercube mid point (white grid) node and a box corner. We study an arbitrary sub-box, e.g. the grey one in Fig. 9.4. Since this is to become the final multi-linear interpolation hypercube we transform the original state coordinates, such that the "lower left" state turns into $\mathbf{0}$ and the opposite corner into $\mathbf{1}$, in general into sub-box states $\mathbf{y} \in[0,1]^{D}$. In order to have a unified presentation below, if necessary we make a change of variables, $y_{d}^{\prime}:=1-y_{d}$, to get the outer grid corner point as $\mathbf{0}$ and the inner grid corner point as

1, as in the grey sub-box in Fig. 9.4, right, illustrates. Then the two basic grids will have nodes at $y_{d}=0, \pm 2, \pm 4, \ldots \quad \forall d$ (outer grid) and at $y_{d}= \pm 1, \pm 3, \ldots \quad \forall d$ (inner grid), respectively.

Interpolation steps:

1. Estimate by linear interpolation the function values at the $2^{D}-2$ sub-box corners that do not correspond to a node in any of the two original grids (small balls in Fig. 9.4, right, and in Fig. 9.5 below).
2. Estimate by multilinear interpolation the function value at $\mathbf{y}$, based on the sub-box corners.


In step 1 we use one of the two computed nodes, i.e. $\mathbf{0}$ or $\mathbf{1}$, plus another node of the same original grid, such that the wanted sub-box node is the midpoint between the two interpolation nodes, e.g. $(1,1,1,0,0)=\frac{1}{2} \cdot((1,1,1,1,1)+(1,1,1,-1,-1))$ in 5D. In general, every level $\ell$ of the opposite grid (here $\ell=0$ ) is incremented in one interpolation point and decremented in the other, formally $\ell=\frac{1}{2} \cdot((\ell+1)+(\ell-1))$. (A more time consuming alternative would be to weigh all the symmetrically placed neighbour nodes, instead of just two.) For the final choice of interpolation nodes in step 1 we define

- Interpolation nodal set $\mathbf{R}$ : For a sub-box node $n$ choose the grid that provides the least expected interpolation distance in step 2, assuming a uniform distribution for $\mathbf{y} \in[0,1]^{D}$


### 9.3 Interpolation improvements

For every sub-box node, set $\mathbf{R}$ is always determined by the smallest number of deviating 1's and 0 's from the two computed nodes $\mathbf{0}, \mathbf{1}$, in this example does ( $1,1,1,0,0$ ) deviate in 3 digits from $\mathbf{0}$ and in 2 digits from $\mathbf{1}$ (preferred). We prove this:

Theorem 9.6: The grid whose computed node, $\mathbf{0}$ or $\mathbf{1}$, has the least $\ell_{1}$-distance from a considered sub-box node also gives the least expected interpolation distance between uniformly, independently distributed states $\mathbf{y} \in[0,1]^{D}$ and that sub-box node.

Proof: Consider a sub-box node $n$ with $\left(\left|\mathcal{D}_{n 1}\right|=\right) D_{1}$ 1-coordinates, for simplicity the $D_{1}$ first ones, and $\left(\left|\mathcal{D}_{n 0}\right|=\right) D-D_{1} 0$-coordinates. From $\mathbf{0}$ the $\ell_{1}$-distance is $D_{1}$ and from $\mathbf{1}$ it is $D-D_{1}$. The corresponding step 2 distances from a state $\mathbf{y}$ are

$$
\begin{aligned}
& \frac{1}{2}(|\mathbf{y}-\mathbf{0}|+|\mathbf{y}-(2, \ldots, 2,0, \ldots, 0)|)=D_{1}+\sum_{d=D_{1}+1}^{D} y_{d}, \\
& \frac{1}{2}(|\mathbf{y}-\mathbf{1}|+|\mathbf{y}-(1, \ldots, 1,-1, \ldots,-1)|)=D-D_{1}+\sum_{d=1}^{D_{1}}\left(1-y_{d}\right) .
\end{aligned}
$$

The expected interpolation distances, as the $\mathbf{y}$-coordinates vary uniformly and independently in $[0,1]^{D}$, are

$$
\begin{aligned}
& E\left[D_{1}+\sum_{d=D_{1}+1}^{D} Y_{d}\right]=D_{1}+\frac{1}{2}\left(D-D_{1}\right)=D-\frac{1}{2}\left(D-D_{1}\right), \\
& E\left[D-D_{1}+\sum_{d=1}^{D_{1}}\left(1-Y_{d}\right)\right]=D-D_{1}+\frac{1}{2} D_{1}=D-\frac{1}{2} D_{1} .
\end{aligned}
$$

Thus if $D-D_{1}<D_{1}$, i.e. if $D_{1}>\frac{D}{2}$, then node $\mathbf{1}$ both has a shorter $\ell_{1}$-distance and provides a lower expected value to the sub-box node $n$ than node $\mathbf{0}$ does.

A better choice would be to choose the grid individually for each state, the one on the least interpolation distance. However, we realize from the proof of Thm 9.6 that for a given state $\mathbf{y} \in[0,1]^{D}$ and a sub-grid node $n$ with $\left|\mathcal{D}_{n 1}\right|=D_{n 1}$ 1-coordinates, the grid providing the shortest interpolation distance to $\mathbf{y}$ is the $\mathbf{0}$-grid if $\sum_{d=1}^{D} y_{d} \leq D-D_{n 1}$ and the $\mathbf{1}$-grid if $\sum_{d=1}^{D} y_{d} \geq D-D_{n 1}$. The implication is devastating: for any $\varepsilon>0$ there exist pairs of interior states $\left(\mathbf{y}_{1}, \mathbf{y}_{2}\right)$ such that $\left\|\mathbf{y}_{1}-\mathbf{y}_{2}\right\|<\varepsilon$ and the grids chosen for node $n$ are different. This means that our demand for continuous interpolation results fails - and we abandon the idea. (However, a weighing of the two grids, according to the interpolation distances, would do.)

We encounter extrapolation of two kinds. One kind occurs for all multi-linear interpolation, whenever a state falls outside all the (non-extended) rectilinear boxes in some dimensions. This is not treated here. The second kind is specific to our double grids, with an interior grid leaving an outmost region of sub-boxes, where the usual midpoint interpolation in step 1 cannot be performed for the inner grid. This is illustrated in Fig. 9.4, right, by the striped 2D sub-box with a lower-left corner that has all the inner grid nodes (black) to the right. For simplicity we
avoid this kind of extrapolation, by relying on outer grid interpolation in every such case. The Thms below assume unrestricted bases $\mathbf{R}$, i.e. with an empty set $\mathcal{E}$ of nodes where this kind of extrapolation trouble occurs.

How efficient is double grid interpolation? First we simplify:
Theorem 9.7: If an unweighted (arithmetic mean) interpolation is used in step 2, then the interpolation distance for every enclosed state $\mathbf{y}$ for $D$ odd, $\mathcal{E}=\varnothing$ is

$$
\frac{3 D}{4}-D \cdot\binom{D-1}{(D-1) / 2} \cdot 2^{-D-1}, \text { which for } D \text { big is approximately } \frac{D}{4} \cdot\left(3-\frac{1}{\sqrt{\pi(D-1) / 2}}\right)
$$

Proof: By sorting the surrounding sub-box nodes according to the number of 1-coordinates, using the formula in the proof of Thm 9.6 and symmetry, and letting $M:=(D-1) / 2$, the common (arithmetic mean) nodal weight $2^{-D}$ determines the interpolation distance

$$
\begin{aligned}
& \frac{1}{2^{D}} \cdot\left(\sum_{d} y_{d}+\sum_{d}\left(1+\sum_{d^{\prime} \neq d} y_{d^{\prime}}\right)+\sum_{d} \sum_{d^{\prime}>d}\left(2+\sum_{d^{\prime \prime} \neq d, d^{\prime}} y_{d^{\prime \prime}}\right)+\cdots\right. \\
& +\sum_{d} \sum_{d^{\prime}>d} \cdots \sum_{d^{(M-1)}>d, d^{\prime}, \ldots, d^{(M-2)}}\left[M+\sum_{d^{(M)} \neq d, d^{\prime}, \ldots, d^{(M-1)}} y_{d^{(M)}}\right] \\
& +\sum_{d}\left(1-y_{d}\right)+\sum_{d}\left[1+\sum_{d^{\prime} \neq d}\left(1-y_{d^{\prime}}\right)\right]+\sum_{d} \sum_{d^{\prime}>d}\left[2+\sum_{d^{\prime \prime} \neq d, d^{\prime}}\left(1-y_{d^{\prime \prime}}\right)\right]+\cdots \\
& \left.+\sum_{d} \sum_{d^{\prime}>d} \cdots \sum_{d^{(M-1)}>d, d^{\prime}, \ldots, d^{(M-2)}}\left[M+\sum_{d^{(M)} \neq d, d^{\prime}, \ldots, d^{(M-1)}}\left(1-y_{d^{(M)}}\right)\right]\right)= \\
& =\frac{1}{2^{D}} \cdot\left(D+\left[\binom{D}{1} \cdot 1+\binom{D}{2} \cdot 2+\cdots+\binom{D}{M} \cdot M+\binom{D}{1} \cdot 1+\binom{D}{2} \cdot 2+\cdots+\binom{D}{M} \cdot M\right]+\right. \\
& \left.+\left[\binom{D}{1} \cdot(D-1)+\binom{D}{2} \cdot(D-2)+\cdots+\binom{D}{M} \cdot(D-M)\right]\right)= \\
& =\frac{1}{2^{D}} \cdot\left(2 D \cdot\left[\binom{D-1}{0}+\binom{D-1}{1}+\cdots+\binom{D-1}{M-1}\right]+D \cdot\left[1+\binom{D-1}{1}+\binom{D-1}{2}+\cdots+\binom{D-1}{M}\right]\right)= \\
& =\frac{1}{2^{D}} \cdot\left(2 D \cdot \frac{1}{2} \cdot\left[2^{D-1}-\binom{D-1}{M}\right]+D \cdot \frac{1}{2} \cdot\left[2^{D-1}+\binom{D-1}{M}\right]\right)=\frac{3 D}{4}-D \cdot\binom{D-1}{M} \cdot 2^{-D-1} \text {. }
\end{aligned}
$$

For $D$ big we approximate the factorials according to Stirling's formula, using $D-1=2 M$.
If these results are to be compared with the corresponding single grid result, we must remember that the state scale is doubled, i.e. the upper bound $\frac{3 D}{4}$ here means a reduction factor $\frac{3 D}{2 \cdot 4} / \frac{D}{2}=\frac{3}{4}$.

For the two steps, double grid multi-linear interpolation we compute the expected interpolation distance:

Theorem 9.8: If multi-linear interpolation on basis $\mathbf{R}$ is used in step 2, then the expected interpolation distance for uniformly, independently distributed states $\mathbf{y} \in[0,1]^{D}$ for $D$ odd,

### 9.3 Interpolation improvements

$\mathcal{E}=\varnothing$ is
$\overline{\bar{\rho}}(\mathbf{R})=\frac{2 D}{3} \cdot\left[1-\binom{D-1}{(D-1) / 2} \cdot 2^{-D}\right]$, for $D$ big turning into $\overline{\bar{\rho}}(\mathbf{R}) \approx \frac{2 D}{3} \cdot\left(1-\frac{1}{\sqrt{2 \pi(D-1)}}\right)$.
Proof: Following the proof strategy of Thm 9.7 and using the multi-linear interpolation weights we get the interpolation distance

$$
\begin{aligned}
& \bar{\rho}(\mathbf{y}, \mathbf{R})=\sum_{n=1}^{2^{D}} \bar{\rho}\left(\mathbf{y}, \mathbf{r}_{n}\right) \cdot w\left(\mathbf{y}, \mathbf{r}_{n}\right)= \\
& =\sum_{d} y_{d} \cdot \prod_{d^{\prime}}\left(1-y_{d^{\prime}}\right)+\sum_{d}\left(1+\sum_{d^{\prime} \neq d} y_{d^{\prime}}\right) \cdot y_{d} \cdot \prod_{d^{\prime} \neq d}\left(1-y_{d^{\prime}}\right)+ \\
& +\sum_{d} \sum_{d^{\prime}>d}\left(2+\sum_{d^{\prime \prime} \neq d, d^{\prime}} y_{d^{\prime \prime}}\right) \cdot y_{d} y_{d^{\prime}} \cdot \prod_{d^{\prime \prime} \neq d, d^{\prime}}\left(1-y_{d^{\prime \prime}}\right)+\cdots+\sum_{d}\left(\cdots \sum_{d^{(M-1)}>d, d^{\prime}, \ldots, d^{(M-2)}}[M+\right. \\
& \left.\left.+\sum_{d^{(M)} \neq d, d^{\prime}, \ldots, d^{(M-1)}} y_{d^{(M)}}\right] \cdot y_{d} \cdots y_{d^{(M-1)}} \cdot \prod_{d^{(M)} \neq d, d^{\prime}, \ldots, d^{(M-1)}}\left(1-y_{d^{(M)}}\right)\right)+ \\
& +\sum_{d}\left(1-y_{d}\right) \cdot \prod_{d^{\prime}} y_{d^{\prime}}+\sum_{d}\left[1+\sum_{d^{\prime} \neq d}\left(1-y_{d^{\prime}}\right)\right] \cdot\left(1-y_{d}\right) \cdot \prod_{d^{\prime} \neq d} y_{d^{\prime}}+ \\
& +\sum_{d} \sum_{d^{\prime}>d}\left[2+\sum_{d^{\prime} \neq d, d^{\prime}}\left(1-y_{d^{\prime \prime}}\right)\right] \cdot\left(1-y_{d}\right)\left(1-y_{\left.d^{\prime}\right)}\right) \cdot \prod_{d^{\prime \prime} \neq d, d^{\prime}} y_{d^{\prime \prime}}+\cdots+\sum_{d}\left(\cdots \sum_{d^{(M-1)}>d, d^{\prime}, \ldots, d^{(M-2)}}^{\sum_{d^{\prime}}[M+}\right. \\
& \left.\left.+\sum_{d^{(M)} \neq d, d^{\prime}, \ldots, d^{(M-1)}}\left(1-y_{d^{(M)}}\right)\right]\left(1-y_{d}\right) \cdots\left(1-y_{d^{(M-1)}}\right) \cdot \prod_{d^{(M)} \neq d, d^{\prime}, \ldots, d^{(M-1)}} y_{d^{(M)}}\right)
\end{aligned}
$$

As for the expected interpolation distance, for independently and uniformly distributed s.v.'s $\left(Y_{d}\right)_{d=1}^{D}$, we simply evaluate the two types of integrals $\int_{0}^{1} y(1-y) d y=\frac{1}{6}$ and $\int_{0}^{1} y d y=\int_{0}^{1}(1-y) d y=\frac{1}{2}$ to have

$$
\begin{aligned}
& \overline{\bar{\rho}}(\mathbf{R})=2 \cdot\left(D \cdot \frac{1}{6} \cdot\left(\frac{1}{2}\right)^{D-1}+\binom{D}{1} \cdot\left[1 \cdot\left(\frac{1}{2}\right)^{D}+(D-1) \cdot \frac{1}{6} \cdot\left(\frac{1}{2}\right)^{D-1}\right]+\right. \\
& \left.+\binom{D}{2} \cdot\left[2 \cdot\left(\frac{1}{2}\right)^{D}+(D-2) \cdot \frac{1}{6} \cdot\left(\frac{1}{2}\right)^{D-1}\right]+\cdots+\binom{D}{M} \cdot\left[M \cdot\left(\frac{1}{2}\right)^{D}+(D-M) \cdot \frac{1}{6} \cdot\left(\frac{1}{2}\right)^{D-1}\right]\right)= \\
& =2 \cdot\left(\left(\frac{1}{2}\right)^{D} \cdot\left[1 \cdot\binom{D}{1}+2 \cdot\binom{D}{2}+\cdots+M \cdot\binom{D}{M}\right]+\right. \\
& \left.+\frac{1}{6} \cdot\left(\frac{1}{2}\right)^{D-1} \cdot\left[D+(D-1) \cdot\binom{D}{1}+(D-2) \cdot\binom{D}{2}+\cdots+(D-M) \cdot\binom{D}{M}\right]\right)= \\
& =2 \cdot\left(\left(\frac{1}{2}\right)^{D} \cdot D \cdot\left[\binom{D-1}{0}+\binom{D-1}{1}+\cdots+\binom{D-1}{M-1}\right]+\frac{1}{6} \cdot\left(\frac{1}{2}\right)^{D-1} \cdot D \cdot\left[\binom{D-1}{0}+\binom{D-1}{1}+\cdots+\binom{D-1}{M}\right]\right)=
\end{aligned}
$$

$=2 \cdot\left(\frac{1}{2}\right)^{D} \cdot D \cdot\left(\frac{1}{2} \cdot\left[2^{D-1}-\binom{D-1}{M}\right]+\frac{1}{6} \cdot\left[2^{D-1}+\binom{D-1}{M}\right]\right)=\frac{D}{3} \cdot\left[2-\binom{D-1}{M} \cdot 2^{1-D}\right]$.
For $D$ big we approximate the factorials according to Stirling's formula, using $D-1=2 M$.
Theorem 9.9: Double multi-linear interpolation, basis $\mathbf{R}$, provides continuity for the interpolated value function.

Proof: Continuity in the interior of each sub-box is evident. When passing a sub-box boundary the only nodes to have positive weights are some boundary nodes, the same active node set for all adjacent sub-boxes. Since step 2 in itself means multi-linear interpolation it guarantees continuity across the boundary, whenever the function values are unique (i.e. coincide for all adjacent sub-boxes). These values are determined for each grid and sub-box in step 1, either directly or by linear interpolation between two grid points that are predefined for each given midpoint node, i.e. unique for each grid. As for the final choice of grid each node has a unique characterisation by its odd and even state levels, according to Thm 9.6, i.e. the choice is unique. (For $D$ even the two grids may perform equally, and we supply a rule that guarantees a unique choice.)

Some results are found as $D b l$ in Tab 9.4 above. In comparison with the standard multi-linear interpolation $S t d$ the run times are $40-70 \%$ longer per iteration but they are significantly shorter than for the higher-order method Quad. The unique Dbl-discretization makes it difficult to compare the derived costs with the other interpolation methods (as to the residual values). However, the backward-forward cost discretization error becomes $-2.0 \%$ for $L=4 \frac{1}{2}$ and $-1.2 \%$ for $L=5 \frac{1}{2}(D b l)$; even less than for $L=6$ in Tab 4.3 (for Std) and for $Q u a d, L=5$. For the future we therefore recommend that the $D b l$-option should be used in a final run, always or at least if the available number of processors are fully utilised, preventing an incremented $L$ value, and if the prospective CPU-hrs are acceptable.

### 9.4 Dantzig-Wolfe decomposition

The presence of discrete maintenance options for each segment and year makes our maintenance optimisation problem non-convex. The estimates of the unknown dual gap that can be formed from iteration results are interfered with discretization errors. Here the DW method is useful by generating upper bounds for the optimal dual value. However, also these upper bounds will suffer from discretization errors. Since we want to apply parallelisation, running the program on a cluster of co-operating processors, an important aspect is that the special structure of our DW master problem makes it possible to handle just a small amount of data at the supervising root processor, by careful implementation.

### 9.4.1 Master problem

The DW method was introduced in Sec 1.2.4. In this subsection we formulate the LP-problem. Since every dual iteration is costly we want to avoid extreme $\mathbf{v}$-changes. One way is to enclose

### 9.4 Dantzig-Wolfe decomposition

the acceptable $\mathbf{v}$-values of an updating in a box; further to restrict the objective target of an updating by an upper bound. To counteract the expected side effects of such a bound - multiple optimal solutions - a weak, bundle-like preference for small $\boldsymbol{v}$-changes can be used. For the formulated problem we try to get acquainted with its solutions, by stating some theorems.

After a relaxation of the budgetary constraints for each year $t$ of the planning period of length $T$, we consider the updating of the dual prices $\mathbf{v}=\left(v_{t}\right)_{t=0}^{T-1}$. In Ch 4 we formulated the Lagrangean dual problem (cf. (4.3) - (4.4))

$$
\underset{\mathbf{v} \geq v_{\text {BCR }} \cdot \mathbf{1}}{\operatorname{maximise}} \Phi(\mathbf{v})
$$

where the dual subproblem is on the form

$$
\Phi(\mathbf{v})=\underset{\mathbf{z}, \mathbf{y}}{\operatorname{minimum}} F(\mathbf{z})+\mathbf{v}^{T} \mathbf{G}(\mathbf{z})+\left(v_{\mathrm{BCR}} \cdot \mathbf{1}-\mathbf{v}\right)^{T} \mathbf{y} \quad \text { s.t. }\left\{\begin{array}{c}
\mathbf{G}(\mathbf{z}) \leq \mathbf{y} \\
\mathbf{y} \leq \mathbf{b} \\
\mathbf{z} \in \mathcal{Z}
\end{array}\right.
$$

Here $\mathbf{b}=\left(b_{t}\right)_{t}$ is the budget vector. Due to separability the subproblem is partitioned into a separate (budget) sub-problem for $\mathbf{y}$ and separate (network) subproblems for each $\mathbf{z}$-component $z_{s}$, where $s$ stands for road segment. The subproblem for $\mathbf{y}$ has optimum solution $\mathbf{y}=\mathbf{b}$ for every $\mathbf{v} \geq v_{B C R} \cdot \mathbf{1}-\mathrm{cf}$. Thm 4.1. The concavity property of the dual problem also applies to each subproblem, for any segment $s$ making every subgradient computed at the (segment) subproblem optimum an affine majorant for the optimal value at every other $\mathbf{v}$-value.

In parallelisation runs the aim is to split the computational work evenly between the processors. Before each run we make an appropriate subdivision of the road segments $s \in \mathcal{S}$, $|\mathcal{S}| \gg 1$, for the available processors. By iteratively solving the dual subproblem for different $\boldsymbol{v}$-values we get subgradient information. For each segment $s$ we distinguish those dual iterations $i \in \mathcal{I}(s)$ leading to different affine majorants, as defined by the discounted total traffic costs $f_{s i}$ and annual maintenance costs $\left(g_{s i t}\right)_{t}$. For prices $\mathbf{v}=\mathbf{v}_{i}$ and the calculated optimal value $\varphi_{s i}\left(\mathbf{v}_{i}\right)$ the corresponding affine majorant, extrapolated to a candidate $\mathbf{v}$, is

$$
\varphi_{s i}(\mathbf{v})=\varphi_{s i}\left(\mathbf{v}_{i}\right)+\sum_{t} g_{s i t} d^{t+\frac{1}{2}}\left(v_{t}-v_{i t}\right)=f_{s i}+\sum_{t} g_{s i t} d^{t+\frac{1}{2}} v_{t} .
$$

Thus for extrapolation purposes, information about the underlying $\mathbf{v}_{i}$ is irrelevant. Any combination $\varphi(\mathbf{v})=\sum_{s} \varphi_{s, i_{s}}(\mathbf{v})+\left(\nu_{\mathrm{BCR}} \cdot \mathbf{1}-\mathbf{v}\right)^{T} \mathbf{b}$ of majorants constructed from individual segment $s$ iterations $i_{s}$ becomes an affine majorant for the full dual $\Phi(\mathbf{v})$. Thus from $I$ registered iterations per segment we get $I^{|\mathcal{S}|}$ different majorants, e.g., $I=10$ and $|\mathcal{S}|=9 \cdot 10^{3}$ meaning $10^{9000}$ combinations - the potential columns in our master problem. Our updating maximin master problem is to find a $\mathbf{v}$-value such that the lowest affine majorant of all combinations is as high as possible. By turning to the linearization

$$
\Phi(\mathbf{v}) \approx \sum_{s} \min _{i \in \bar{I}(s)} \varphi_{s i}(\mathbf{v})-\sum_{t} b_{t} d^{t+\frac{1}{2}} v_{t}+v_{\mathrm{BCR}} \cdot \sum_{t} b_{t} d^{t+\frac{1}{2}}
$$

and under box constraints $\mathbf{v} \in[\mathbf{r}, \mathbf{R}]$ we get the master problem

$$
\begin{align*}
& q=\underset{\mathbf{u}, \mathbf{v}}{\operatorname{maximum}} \sum_{s} u_{s}-\sum_{t} b_{t} d^{t+\frac{1}{2}} v_{t}\left(+v_{\mathrm{BCR}} \cdot \sum_{t} b_{t} d^{t+\frac{1}{2}}\right)  \tag{9.4}\\
& \text { s.t. } \begin{cases}\varphi_{s i}(\mathbf{v}) \geq u_{s} & \forall(s, i) \\
r_{t} \leq v_{t} \leq R_{t} & \forall t .\end{cases}
\end{align*}
$$

In (9.4) the optimal majorant and the dual optimum $\mathbf{v}^{*}$ satisfy $q \geq \Phi\left(\mathbf{v}^{*}\right)$. The lower box bounds reflect the real return rate conditions $\mathbf{v} \geq \nu_{\mathrm{BCR}} \cdot \mathbf{1}$ or artificial, movable bounds ( $r_{t} \geq v_{\mathrm{BCR}}$ ), whereas the (fixed or movable) upper box bounds $R_{t}$ always are artificial, useful primarily in the early iterations for avoiding infinite iterates. In the fixed box case, we use $\mathbf{r} \equiv v_{\mathrm{BCR}} \cdot \mathbf{1}$, and the components of $\mathbf{R}=\left(R_{t}\right)$ are chosen a priori - ideally as small as possible without influencing the optimum solution. We will return to that.

Formulation (9.4) can be further extended. Especially in the early iterations an occasionally found good primal solution may act as a ceiling $c$ for the optimal value $q$ ( $q \leq c$ as constraint (9.5c)). (However, such an additional constraint will introduce multiple optimal master solutions, and an unchanged risk of choosing the redundant upper bounds $\mathbf{R}$ as the next iterate. By penalizing any deviations $\left(\delta_{t}\right)_{t}$ from the current dual iterate $\mathbf{v}^{0}$ linearly with a coefficient $\varepsilon>0$, introducing $\left|v_{t}-v_{t}^{0}\right| \leq \delta_{t}$ as constraints (9.5e) - (9.5f), as well as keeping the box bounds $v_{t} \in\left[r_{t}, R_{t}\right]$ as (9.5f) and the majorant conditions $u_{s} \leq \varphi_{s i}(\mathbf{v})$ as (9.5b), we get our final, bundle-like formulation, (9.5) below. Here the discounted maintenance costs and budgets are written as $\widetilde{g}_{s i t}:=g_{s i t} d^{t+\frac{1}{2}}$ and $\tilde{b}_{t}:=b_{t} d^{t+\frac{1}{2}}$, respectively. Model:

$$
\begin{align*}
& q=\underset{\mathbf{u}, \mathbf{v}, \mathbf{\delta}}{\operatorname{maximum}} \sum_{s} u_{s}-\sum_{t} \widetilde{b}_{t} v_{t}-\varepsilon \cdot \sum_{t} \delta_{t}\left(+v_{\mathrm{BCR}} \cdot \sum_{t} \widetilde{b}_{t}\right)  \tag{9.5a}\\
& \text { s.t. }\left\{\begin{array}{cl}
u_{s}-\sum_{t} \widetilde{g}_{s i t} v_{t} \leq f_{s i} & \forall(s, i) \\
\sum_{s} u_{s}-\sum_{t} \widetilde{b}_{t} v_{t} \leq c \\
-v_{t}-\delta_{t} \leq-v_{t}^{0} & \forall t \\
v_{t}-\delta_{t} \leq v_{t}^{0} & \forall t \\
r_{t} \leq v_{t} \leq R_{t} & \forall t
\end{array}\right. \tag{9.5b}
\end{align*}
$$

By introducing LP-dual variables $\mathbf{X}=\left(x_{s i}\right)_{s, i}, w, \mathbf{z}^{-}=\left(z_{t}^{-}\right)_{t}, \mathbf{z}^{+}, \mathbf{y}^{-}$and $\mathbf{y}^{+}$, the LP-dual problem becomes

$$
\begin{equation*}
\underset{\mathbf{X}, w, \mathbf{z}^{-}, \mathbf{z}^{+}, \mathbf{y}^{-}, \mathbf{y}^{+}}{\operatorname{minimise}} \sum_{s, i} f_{s i} x_{s i}+c w+\sum_{t} v_{t}^{0} \cdot\left(z_{t}^{+}-z_{t}^{-}\right)+\sum_{t} R_{t} y_{t}^{+}-\sum_{t} r_{t} y_{t}^{-} \tag{9.6a}
\end{equation*}
$$

$$
\text { s.t. }\left\{\begin{array}{cc}
\sum_{i} x_{s i}+w=1 & \forall s  \tag{9.6b}\\
-\sum_{s, i} \widetilde{g}_{s i t} x_{s i}-\widetilde{b}_{t} w-\left(z_{t}^{-}+y_{t}^{-}\right)+\left(z_{t}^{+}+y_{t}^{+}\right)=-\widetilde{b}_{t} & \forall t \\
-z_{t}^{-}-z_{t}^{+}=-\varepsilon & \forall t \\
x_{s i} \geq 0 & \forall(s, i) \\
w \geq 0 & \forall t .
\end{array}\right.
$$

Let us discuss some problem properties. The variable types $\mathbf{y}^{-}$and $\mathbf{y}^{+}$are linked to the fixed/movable price bounds (9.5f), whereas $\mathbf{z}^{-}$and $\mathbf{z}^{+}$are coupled to the penalized bounds (9.5d) - (9.5e). We have

Theorem 9.10: Assume $0<r_{t}<v_{t}^{0}<R_{t}$. In any LP-optimum

$$
\begin{aligned}
& y_{t}^{+}>0 \Rightarrow z_{t}^{+}=\varepsilon, y_{t}^{-}=z_{t}^{-}=0, \\
& y_{t}^{-}>0 \Rightarrow z_{t}^{-}=\varepsilon, y_{t}^{+}=z_{t}^{+}=0 .
\end{aligned}
$$

Proof: In (9.6c) the two variables of each pair $y_{t}^{+}, z_{t}^{+}$and $y_{t}^{-}, z_{t}^{-}$are indistinguishable.
However, in the objective the assumption $0<r_{t}<v_{t}^{0}<R_{t}$ means that $z_{t}^{+}>0$ is preferred to $y_{t}^{+}>0$, and $z_{t}^{-}>0$ is preferred to $y_{t}^{-}>0$. Thus either of $z_{t}^{+}$or $z_{t}^{-}$reach its upper bound $\varepsilon$ in (9.6d) before the corresponding $y_{t}^{+}$or $y_{t}^{-}$becomes positive. According to $r_{t}<R_{t}$ and (9.5f) both $y_{t}^{+}$and $y_{t}^{-}$cannot be positive in an LP-optimum.

An implication of Thm 9.10 for the LP-iterating is that small budget violations are penalized by $v_{t}^{0}$ per unit and greater ones with $R_{t}$. Whereas $y_{t}^{+}>0, y_{t}^{-}>0$ is non-optimal according to Thm 9.10 the case $z_{t}^{+}>0, z_{t}^{-}=\varepsilon-z_{t}^{+}>0$ cannot be ruled out: all the split solutions correspond to one and the same price $v_{t}=v_{t}^{0}$.

In (9.6), w $=0$ indicates that the objective bound (9.5b) is redundant. On the other extreme is:
Theorem 9.11: In any LP-optimum

$$
w=1 \Rightarrow \mathbf{v}=\mathbf{v}^{0} .
$$

Proof: From (9.6b), (9.6g) and $w=1$ we get $\mathbf{x}=\mathbf{0}$. For every $t$-value (9.6c) turns into $z_{t}^{+}+y_{t}^{+}=z_{t}^{-}+y_{t}^{-}$.
According to Thm 9.10, at most one of $y_{t}^{+}, y_{t}^{-}$is positive. Assume, e.g., $y_{t}^{+}>0, y_{t}^{-}=0$. Thm 9.10 implies $z_{t}^{+}=\varepsilon$ and $z_{t}^{+}+y_{t}^{+}>\varepsilon$, as well as $z_{t}^{-}=0$ and $z_{t}^{-}+y_{t}^{-}=0-\mathrm{a}$ contradiction. Thus $y_{t}^{+}=y_{t}^{-}=0$ and $z_{t}^{+}=z_{t}^{-}=\frac{1}{2} \varepsilon>0$. By complementarity both ( 9.5 d ) and (9.5e) are active for the considered $t$-value; from which $\delta_{t}=0, v_{t}=v_{t}^{0}$ follows.

We realise that if the penalty $\varepsilon$ in (9.5a) is too big, it will restrict the $\boldsymbol{v}$-changes undesirably. How big might $\varepsilon$ be? For answering the question we need some concepts: Since the objective $\Phi(\boldsymbol{v})$ of the Lagrangean dual is a concave function, also its linearization in (9.4) is. Denote the affine majorant by

$$
\hat{\Phi}(\mathbf{v}):=\sum_{s} \min _{i \in \bar{I}(s)} \varphi_{s i}(v)-\sum_{t} \widetilde{b}_{t} v_{t}
$$

Hence for any fixed target value $\gamma$ the objective level set

$$
C(\gamma):=\{\mathbf{v} \in[\mathbf{r}, \mathbf{R}]: \hat{\Phi}(\mathbf{v}) \geq \gamma\}
$$

is convex. Due to constraint $(9.5 \mathrm{c})$ the set $C(\gamma)$ is empty for $\gamma>c$. Since the generating majorants of $\hat{\Phi}(\mathbf{v})$ are affine functions of $\mathbf{v}, C(\gamma)$ is a polyhedron with hyperplanes as boundary facets. Since the $\mathbf{v}$-domain (9.5f) is bounded and the function $\hat{\Phi}(\mathbf{v})$ is continuous, the maximum $\gamma$-value is attained for some $\gamma=\gamma^{*}: C\left(\gamma^{*}\right) \neq \varnothing, C(\gamma)=\varnothing \quad \forall \gamma>\gamma^{*}$. Fig 9.6 is a 2D illustration for $\gamma^{*}=c$. All solutions in the grey region reach $\hat{\Phi}(\mathbf{v}) \geq c$. The additional points in the horizontally striped region satisfy $\hat{\Phi}(\mathbf{v}) \geq \gamma$ for some $\gamma<c$.


Figure 9.6 Example of box constraints, objective level sets $C$ and distance sets $\mathcal{D}$ for $\gamma^{*}=c$.
As for the correction term in (9.5a), proportional to the $\ell_{1}$-distance $\sum_{t} \delta_{t}$ from the current dual iterate $v_{t}^{0}$, the distance set for (9.5d)-(9.5e),

$$
\mathcal{D}(\Delta):=\left\{\mathbf{v}:\left|v_{t}-v_{t}^{0}\right| \leq \delta_{t}, \sum_{t} \delta_{t} \leq \Delta\right\}
$$

is a regular polyhedron, centred at $\mathbf{v}^{0}$. There exists a minimum distance $\Delta=\Delta^{*}$ such that the intersection $\mathcal{C}(\gamma) \cap \mathcal{D}(\Delta)$ is non-empty. The $\mathcal{D}\left(\Delta^{*}\right)$-boundary hyperplanes have normal vectors with components $\pm 1$ in all possible combinations. Only in rare cases the intersection set $C\left(\gamma^{*}\right) \cap \mathcal{D}\left(\Delta^{*}\right)$ will contain more than one point, i.e. by the penalization we avoid multiple

LP-optimum solutions, in practice. The unique (basic) LP-optimum will occur at a corner point of $C\left(\gamma^{*}\right)$ and/or $\mathcal{D}\left(\Delta^{*}\right)$. In Fig $9.6 \mathcal{D}\left(\Delta^{*}\right)$ is the grey-dotted region. If $\varepsilon$ is too big the vertically striped region $\mathcal{D}(\Delta)$ determines the $\mathbf{v}$-optimum, for some $\gamma<c, \Delta<\Delta^{*}$.

Theorem 9.12: Assume that the minimum distance $\Delta^{*}=\Delta^{*}(\gamma)$, as a function of the level value $\gamma$, satisfies $\left.\frac{d}{d \gamma}\right|_{ \pm} \Delta^{*}(\gamma)<\frac{1}{\varepsilon}$. Then the left and right derivatives are well-defined everywhere. and the LP-optimum corresponds to the maximum possible level value $\gamma=\gamma^{*}$.

Proof: The objective (9.5a) can be written maximise $\hat{\Phi}(\mathbf{v})-\varepsilon \cdot \sum_{t} \delta_{t}$.
At first consider the set of feasible LP-solutions restricted to an arbitrary fixed value $\hat{\Phi}(\mathbf{v})=\gamma$. The LP-optimisation will lead to the minimum distance $\Delta^{*}(\gamma)$. Then letting $\gamma$ vary, the unrestricted objective becomes to
maximise $\gamma-\varepsilon \cdot \Delta^{*}(\gamma)$.
Since a finite number of affine facets will generate the full majorant function, the distance function $\Delta^{*}(\gamma)$ has a well-defined two-sided derivative almost everywhere; otherwise each of the left and the right derivatives exist. By the derivative assumption the unrestricted objective increases by $\gamma$ everywhere. Hence the maximum possible value $\gamma=\gamma^{*}$ is attained in the LPoptimum.

Notice that if the current iterate $\mathbf{v}^{0}$ belongs to $C\left(\gamma^{*}\right)$, then the minimum distance $\Delta^{*}=\Delta^{*}\left(\gamma^{*}\right)$ is attained for $\mathbf{v}=\mathbf{v}^{0}$ (cf. Thm 9.11).

### 9.4.2 Implicit simplex pivoting

A dual (outer) iteration consists of solving the dual subproblem for given $\mathbf{v}$ and from this information solving the master problem by a number of (inner) simplex iterations. In this subsection we document the implementation of a solver for the DW master problem (9.5) and its role in the Lagrangean dual problem. Since the task of constructing a program that can challenge the subgradient method is non-trivial, we describe our efforts in detail. The idea is to exploit the fact that any basic solution of (9.6) permits at most $|\mathcal{S}|+2 T$ positive variable values. Thus for at most $2 T$ segments in the GUB-like constraints $(9.6 \mathrm{~b})$ the $x_{s i}$-values are split. These constraints plus those of (9.6c) with (split) basic $x_{s i}$ 's or with $w$ basic are the only ones that we handle explicitly in the simplex tableau. For an intended cluster of co-working processors we decentralise the detailed segment information, reducing the centralised handling of an LP-iteration of (9.5) - (9.6) into solving two small linear systems of equations. For the choice of start solution we state some further problem properties.

Formulation (9.6) is suitable for parallel computing, since iteration information about each segment can be kept and handled locally by the processor; only data for the simplex entering
candidate $x_{s i}$ of each processor are transferred to the supervising processor (the root). Before going into details we introduce some set notations. For a given iteration we partition the segments $s \in \mathcal{S}$ into disjoint sets $\mathcal{S}_{1}$ and $\mathcal{S}_{2}$, for those segments $s$ having exactly one respectively more than one basic $x_{s i}$-variable in the convexity condition (9.6b) for $s$, and partition the time periods $t \in \mathcal{T}$ into three disjoint sets according to the current type of basic variable in (9.6c): $\mathcal{T}_{1}$ (some $x_{s i}$ or $w$ basic), $\mathcal{T}_{2}\left(z_{t}^{+}\right.$or $z_{t}^{-}$basic; the other one basic in ( 9.6 d )) and $\mathcal{T}_{3}$ ( $y_{t}^{+}$or $y_{t}^{-}$basic). As pointed out, the $2 T$ cost constraints ( 9.6 c$)-(9.6 \mathrm{~d})$ permit at most $2 T$ extra $x_{s i}$ being basic, apart from the mandatory one per segment $s$ in (9.6b). Since $2 T$ is much less than the number of segments we will have the majority of basic variables in set $\mathcal{S}_{1}$, and in between 0 and $2 T$ variables in $\mathcal{S}_{2}$. Because of the bipartitioning of $\mathcal{S}\left(\mathcal{S}_{1}=\mathcal{S}_{2}^{c}\right.$, where ${ }^{c}$ denotes complementary set) we need to keep record of $\mathcal{S}_{2}$ only. We consider a current inner LP-iteration. For every $s \in \mathcal{S}$ one mandatory variable $x_{s, i_{s}^{*}}$ is LP-basic and information about the corresponding dual iteration $i_{s}^{*} \equiv i^{*}(s)$, when the affine majorant was generated (as a subproblem solution), is registered locally. This defines the iteration set $I_{B}(s)=\left\{i_{s}^{*}\right\}$ for every $s \in \mathcal{S}_{1}$ and additional information about the iteration set $I_{B}(s)$ for $s \in \mathcal{S}_{2}$ is registered at the root.

With emphasis on parallelisation, an iteration of the primal simplex method, applied to the LPdual (9.6), involves the following steps:

1. Determine the shadow prices $\mathbf{v}$ in (9.5) at the root. For $t \in \mathcal{T}_{2}$ (both $z_{t}^{+}$and $z_{t}^{-}$are basic in (9.6)) we immediately get $\delta_{t}=0, v_{t}=v_{t}^{0}$ from the equalities in (9.5d)-(9.5e). For $t \in \mathcal{T}_{3}\left(y_{t}^{+}\right.$or $y_{t}^{-}$is basic in (9.6)), we immediately get $v_{t}$ from the corresponding equality in (9.5f). As for the rest of $\mathbf{v}$, formulation (9.5) is not appropriate for parallel computing, since to fully solve ( 9.5 c ) with equality, in case $w$ is basic in (9.6), we need the complete vector $\mathbf{u}$, i.e. it involves all segments simultaneously. Instead we use the equalities in ( 9.5 b ) for the one registered (basic) iteration $i_{s}^{*}$ for each segment $s$. By subtracting all these equations from (9.6c) we get

$$
\begin{equation*}
\sum_{t}\left(\sum_{s} \widetilde{g}_{s, i_{s}^{*}, t}-\widetilde{b}_{t}\right) \cdot v_{t}=c-\sum_{s} f_{s, i_{s}^{*}} \tag{9.7}
\end{equation*}
$$

Thus we just have to update the cost sums $F:=\sum_{s} f_{s, i_{s}^{* *}}$ and $\left(\sum_{s} \widetilde{g}_{s,,_{s}^{*}, t}\right)_{t}$ at the root, whenever a registered $x_{s, i_{s}^{* *}}$-variable enters or leaves the basis. In practice we distinguish the two segment sets, and update

$$
\begin{equation*}
\beta_{t}^{T O T}:=\widetilde{b}_{t}-\sum_{s} \widetilde{g}_{s, i_{s}^{*}, t} \text { and } \beta_{t}:=\widetilde{b}_{t}-\sum_{s \in \mathcal{S}_{1}} \widetilde{g}_{s, i_{s}^{*}, t} \quad \forall t . \tag{9.8}
\end{equation*}
$$

In summary, we solve the system

$$
\left\{\begin{array}{l}
u_{s}-\sum_{t \in \mathcal{T}_{1}} \widetilde{g}_{s i t} v_{t}=f_{s i}+\sum_{t \in \mathcal{T}_{2} \cup \mathcal{T}_{3}} \widetilde{g}_{s i t} v_{t} \quad \forall i \in \mathcal{I}_{B}(s), s \in \mathcal{S}_{2} \\
-\sum_{t \in \mathcal{T}_{1}} \beta_{t}^{T O T} v_{t}=c-F+\sum_{t \in \in \mathcal{T}_{2} \cup \mathcal{T}_{3}} \beta_{t}^{T O T} v_{t}
\end{array}\right.
$$

for $\left(u_{s}\right)_{s \in \mathcal{S}_{2}}$ and $\left(v_{t}\right)_{t \in \mathcal{T}_{1}}$. The LHS coefficient matrix and the corresponding variable vector take the general form

$$
\left(\begin{array}{cc}
\mathbf{E}^{T} & -\widetilde{\mathbf{G}}_{12}^{T} \\
\mathbf{0}^{T} & \boldsymbol{\gamma}^{T}
\end{array}\right)\binom{\mathbf{u}_{2}}{\mathbf{v}_{1}}
$$

Here the $\mathbf{u}_{2}$-part of the coefficient matrix (after column renumbering) is almost triangular, and we solve as a sparse linear system of equations.
2. Determine the entering basic variable. First information about $\mathbf{v}$ is transmitted to each processor from the root and, for each processor segment $s, u_{s}$ is determined from (9.5b) for the locally registered basic iteration $i_{s}^{*}$. For the rest of all the registered iterations $i \in I(s)$ the reduced costs $\bar{c}\left(x_{s i}\right)$ are computed as the difference between the RHS and LHS of ( 9.5 b ). General data for the best reduced cost candidate at each processor and the total sum of $u_{s}$ are transmitted to the root. At the root the remaining candidates are evaluated, by the computation of the RHS - LHS differences in (9.5c)(9.5f). In the final choice of entering variable we consider the different variable scales: whereas the reduced costs for $x_{s i}$ and $w$ are typically of order $10^{6}$ (discounted total future costs in SEK), the reduced costs for the other variable types are of order 1 or less. Therefore we compare the reduced costs multiplied by a relevant variable upper bound: 1 for $x_{s i}$ and $w$, according to (9.6b), but $\beta_{t}^{T O T}$ in (9.8) for the other types, according to ( 9.6 c ). A message for the resulting choice of entering variable is passed to every processor. If a segment variable $x_{s i}$ is entering, detailed variable data are transmitted from the host processor of $s$ to the root. The sums $F,\left(\beta_{t}^{\text {TOT }}\right)_{t}$ in (9.7)-(9.8) are updated whenever any $i_{s}^{*}$ is exchanged, and $\left(\beta_{t}\right)_{t}$ whenever this occurs in $\mathcal{S}_{1}$ or a pivoting means $s \in \mathcal{S}_{1} \rightarrow s \in \mathcal{S}_{2}$ or vice versa. However, basic data about $\mathcal{S}_{1}$ are not registered at the root. Therefore if $x_{s i}$ is a entering the basis for an iteration $i \in I(s)$, then $f_{s i},\left(\widetilde{g}_{s i t}\right)_{t}$ data as well as the corresponding data for $i_{s}^{*}$ are transferred to the root.
3. Determine the basic solution (without the new entering variable) and the basic variable to leave. We make this by first computing the current RHS and the current column of the entering variable and then forming the standard quotients between the two. It is advantageous to consider the structure of the linear system to solve: we distinguish the coefficient basic matrix and the two RHS-vectors - the given RHS's of the basic equations in (9.6) and the given column $\mathbf{n}$ of the entering variable - and the different types of basic variables.

| $x_{s i}, s \in \mathcal{S}_{1}$ | $\mathbf{I}$ | $\mathbf{O}$ | $\mathbf{1}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{1}$ | $\mathbf{n}_{1}$ | $(9.6 \mathrm{~b})$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $x_{s i}, s \in \mathcal{S}_{2}$ | $\mathbf{O}$ | $\mathbf{E}$ | $\mathbf{1}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{1}$ | $\mathbf{n}_{2}$ |  |
| $\mathcal{T}_{1}: x_{s i}, s \in \mathcal{S}_{2}, w$ | $-\widetilde{\mathbf{G}}_{11}$ | $-\widetilde{\mathbf{G}}_{12}$ | $-\widetilde{\mathbf{b}}_{1}$ | $\mathbf{H}_{1}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{O}$ | $-\widetilde{\mathbf{b}}_{1}$ | $\mathbf{n}_{3}$ |  |
| $\mathcal{T}_{2}: z_{t}^{-}, z_{t}^{+}$ | $-\widetilde{\mathbf{G}}_{21}$ | $-\widetilde{\mathbf{G}}_{22}$ | $-\widetilde{\mathbf{b}}_{2}$ | $\mathbf{O}$ | $\mathbf{H}_{2}$ | $\mathbf{O}$ | $\mathbf{O}$ | $-\widetilde{\mathbf{b}}_{2}$ | $\mathbf{n}_{4}$ | (9.6c) |
| $\mathcal{T}_{3}: y_{t}^{-}, y_{t}^{+}$ | $-\widetilde{\mathbf{G}}_{31}$ | $-\widetilde{\mathbf{G}}_{32}$ | $-\widetilde{\mathbf{b}}_{3}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{H}_{3}$ | $\mathbf{K}_{3}$ | $-\widetilde{\mathbf{b}}_{3}$ | $\mathbf{n}_{5}$ |  |
| $\mathcal{T}_{1}: z_{t}^{-}, z_{t}^{+}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{0}$ | $-\mathbf{I}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{O}$ | $-\boldsymbol{\varepsilon}_{1}$ | $\mathbf{n}_{6}$ |  |
| $\mathcal{T}_{2}: z_{t}^{-}, z_{t}^{+}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{0}$ | $\mathbf{O}$ | $-\mathbf{L}_{2}$ | $\mathbf{O}$ | $\mathbf{O}$ | $-\boldsymbol{\varepsilon}_{2}$ | $\mathbf{n}_{7}$ | (9.6d) |
| $\mathcal{T}_{3}: z_{t}^{-}, z_{t}^{+}$ | $\mathbf{O}$ | $\mathbf{O}$ | $\mathbf{0}$ | $\mathbf{O}$ | $\mathbf{O}$ | $-\mathbf{I}$ | $\mathbf{O}$ | $-\boldsymbol{\varepsilon}_{3}$ | $\mathbf{n}_{8}$ |  |

Here the submatrices $\mathbf{H}_{1}, \mathbf{H}_{3}$ and $\mathbf{K}_{3}$ are diagonal, with elements $\pm 1$ according to the basic variable in the corresponding $t$-constraint, $z_{t}^{+}$or $z_{t}^{-}$in (9.6d), $y_{t}^{+}$or $y_{t}^{-}$in (9.6c). The submatrices $\mathbf{H}_{3}$ and $\mathbf{L}_{2}$ contain two non-zero elements per row, whereas $\mathbf{E}$ contains at least two, all the non-zero elements occurring in different submatrix columns.

In the $1^{\text {st }}$ equation part, for $s \in \mathcal{S}_{1}$, we can express $x_{s, i_{s}^{*}}=1-w$ for the $\mathcal{S}_{1}$-part $\mathbf{x}_{1}$. In the $6^{\text {th }}$ and $8^{\text {th }}$ (last) parts, where $z_{t}^{+}$or $z_{t}^{-}$is basic in (9.6d) for $t$, we immediately get its value $\varepsilon$. Solving the $2^{\text {nd }}$ and $3^{\text {rd }}$ parts of equations for the $\mathcal{S}_{2}$-part $\mathbf{x}_{2}$ and $w$, we make use of these facts plus notation (9.7), i.e. $\boldsymbol{\beta}_{i}:=\widetilde{\mathbf{b}}_{i}-\widetilde{\mathbf{G}}_{i 1} \mathbf{1}$ for $i=1,2,3$, to have a reduced system of equations

$$
\left\{\begin{array}{c}
\mathbf{E x} \mathbf{x}_{2}+\mathbf{1} w=\mathbf{1}  \tag{9.9}\\
-\widetilde{\mathbf{G}}_{12} \mathbf{x}_{2}-\boldsymbol{\beta}_{1} w=-\boldsymbol{\beta}_{1}-\mathbf{H}_{1} \boldsymbol{\varepsilon}_{1} .
\end{array}\right.
$$

The upper part of this system is almost triangular (after renumbering), and we solve it as sparse.

The remaining parts of the whole system of equations are easily computed: since $\mathbf{K}_{3}$ is diagonal, $y_{t}^{+}$or $y_{t}^{-}$for $t \in \mathcal{T}_{3}$ ( $5^{\text {th }}$ part) immediately follows, whereas the $4^{\text {th }}$ and $7^{\text {th }}$ part determine the split between $z_{t}^{+}$and $z_{t}^{-}$for $t \in \mathcal{T}_{2}$ :

$$
\left\{\begin{array}{cl}
z_{t}^{+}=\frac{1}{2} \cdot\left(\varepsilon+\sum_{s \in \mathcal{S}_{2}} \sum_{i \in I_{B}(s)} \tilde{g}_{s i t} x_{s i}-\beta_{t} \cdot(1-w)\right) & \forall t \in \mathcal{T}_{2} \\
z_{t}^{-}=\varepsilon-z_{t}^{+} & \forall t \in \mathcal{T}_{2}
\end{array}\right.
$$

As for the current column of the entering variable, the procedure is analogous: for the $2^{\text {nd }}$ and $3^{\text {rd }}$ parts, with RHS's $\mathbf{n}_{2}$ and $\mathbf{n}_{3}$, we make use of the same reduced coefficient matrix and solve

$$
\left(\begin{array}{cc}
\mathbf{E} & \mathbf{1}  \tag{9.10}\\
-\widetilde{\mathbf{G}}_{12} & -\boldsymbol{\beta}_{1}
\end{array}\right)\binom{\overline{\mathbf{n}}_{2}}{\bar{n}_{3}}=\binom{\mathbf{n}_{2}}{\mathbf{n}_{3}+\widetilde{\mathbf{G}}_{11} \mathbf{n}_{1}+\mathbf{H}_{1} \mathbf{n}_{6}} .
$$

Since the coefficient matrices in (9.9) and (9.10) coincide, we solve the systems simultaneously. From the solutions $\left(\mathbf{x}_{2}, w\right),\left(\overline{\mathbf{n}}_{2}, \bar{n}_{3}\right)$ and using $a$ for either of,+- we get

$$
\left\{\begin{array} { c } 
{ \mathbf { x } _ { 1 } = ( 1 - w ) \cdot \mathbf { 1 } } \\
{ \mathbf { z } _ { 1 } ^ { a } = \boldsymbol { \varepsilon } _ { 1 } } \\
{ \mathbf { z } _ { 2 } ^ { - } = \frac { 1 } { 2 } ( \boldsymbol { \varepsilon } _ { 2 } - \widetilde { \mathbf { G } } _ { 2 2 } \mathbf { x } _ { 2 } + ( 1 - w ) \cdot \boldsymbol { \beta } _ { 2 } ) } \\
{ \mathbf { z } _ { 2 } ^ { + } = \frac { 1 } { 2 } ( \boldsymbol { \varepsilon } _ { 2 } + \widetilde { \mathbf { G } } _ { 2 2 } \mathbf { x } _ { 2 } - ( 1 - w ) \cdot \boldsymbol { \beta } _ { 2 } ) } \\
{ \mathbf { z } _ { 3 } ^ { a } = \boldsymbol { \varepsilon } _ { 3 } } \\
{ \mathbf { K } _ { 3 } \mathbf { y } _ { 3 } ^ { a } = - ( 1 - w ) \cdot \boldsymbol { \beta } _ { 3 } + \widetilde { \mathbf { G } } _ { 3 2 } \mathbf { x } _ { 2 } - \mathbf { H } _ { 3 } \boldsymbol { \varepsilon } _ { 3 } }
\end{array} \left\{\begin{array}{c}
\overline{\mathbf{n}}_{1}=\mathbf{n}_{1}-\bar{n}_{3} \cdot \mathbf{1} \\
\overline{\mathbf{n}}_{4}=-\mathbf{n}_{6} \\
\overline{\mathbf{n}}_{5}^{-}=-\frac{1}{2}\left(\mathbf{n}_{4}+\mathbf{n}_{7}+\widetilde{\mathbf{G}}_{21} \mathbf{n}_{1}+\widetilde{\mathbf{G}}_{22} \overline{\mathbf{n}}_{2}+\bar{n}_{3} \cdot \boldsymbol{\beta}_{2}\right) \\
\overline{\mathbf{n}}_{5}^{+}=\frac{1}{2}\left(\mathbf{n}_{4}-\mathbf{n}_{7}+\widetilde{\mathbf{G}}_{21} \mathbf{n}_{1}+\widetilde{\mathbf{G}}_{22} \overline{\mathbf{n}}_{2}+\bar{n}_{3} \cdot \boldsymbol{\beta}_{2}\right) \\
\overline{\mathbf{n}}_{4}=-\mathbf{n}_{8} \\
\mathbf{K}_{3} \overline{\mathbf{n}}_{7}=\mathbf{n}_{5}+\widetilde{\mathbf{G}}_{31} \mathbf{n}_{1}+\widetilde{\mathbf{G}}_{32} \overline{\mathbf{n}}_{2}+\bar{n}_{3} \cdot \boldsymbol{\beta}_{3}+\mathbf{H}_{3} \mathbf{n}_{8}
\end{array}\right.\right.
$$

In the matrix-vector multiplications above, we take care of the characteristic sparseness for each type of entering variable, e.g. for $x_{s i}, s \in \mathcal{S}_{1}$, entering we get $\bar{n}_{1 s^{\prime}}=-\bar{n}_{3} \leq 0$ $\forall s^{\prime} \in \mathcal{S}_{1}, s^{\prime} \neq s$, i.e. not the candidates for leaving the basis; etc.
4. Make pivoting. This solely means that information is transmitted from the root to all the processors, about the new prices $\mathbf{v}$ and the leaving basic variable - optionally none: the stop signal for the LP-problem.

In summary the computational work of each LP-iteration is essentially limited to our solving two linear systems of size $0-2 T$ (with 1 and 2 RHS-vectors, respectively). What to pay for this shortcut is merely book-keeping: At every LP-dual iteration the set membership of $\mathcal{S}_{1}$, $\mathcal{S}_{2}$ and $\mathcal{T}_{1}, \mathcal{T}_{2}, \mathcal{T}_{3}$ must be updated at the root.

The choice of start solution is an initial step:
0 . In the very first Lagrangean dual iteration we solve the full dual subproblem for a given $\boldsymbol{v}$. Then we let the initial $\boldsymbol{\beta}$ in (9.8) hold the discounted budget surpluses for the dual subproblem optimum, and take $\mathbf{v}^{0}$ as the subproblem prices $\mathbf{v}$. For each $\beta_{t}$-value we classify $t \in \mathcal{T}_{2}$ or $t \in \mathcal{T}_{3}$ (since, with just one iteration performed per segment, $\mathcal{S}_{2}=\mathcal{T}_{1}=\varnothing$ : With $\mathbf{x}=0, w=0$, (9.6c) runs

$$
z_{t}^{+}-z_{t}^{-}+y_{t}^{+}-y_{t}^{-}=-\beta_{t} .
$$

Most components of the initial $\mathbf{v}$, determined from the zero-reduced costs in (9.5), are expected to deviate from $\mathbf{v}^{0}$. In this very first Lagrangean dual iteration there are no alternative iterations $i \in I(s)$ to choose, i.e. the start solution $\mathbf{x}$ is also an LP-optimum. However, the $w$-value is free to vary on $0 \leq w \leq 1$, according to (9.6b). Before continuing some clarification is needed.

Our implementation will focus on the following two problem properties; stated as theorems:
Theorem 9.13: In any LP-dual basic solution with $w<1$ basic, (9.7) is equivalent to (9.5c).
Proof: In (9.6b) the assumption $w<1$ implies $\sum_{i} x_{s i}>0$ for each segment $s$. Hence for each $s$ there exists an iteration $i=i_{s}^{*}$ such that $x_{s, i_{s}^{*}}$ is basic, and since the corresponding reduced cost is $0,(9.5 \mathrm{~b})$ holds with equality for $\left(s, i_{s}^{*}\right)$. These equations for every $s$ can be subtracted from $(9.5 \mathrm{c})$ - an equality since $w$ is basic by assumption - to get an equivalent constraint (9.7).

As for the uncovered cases in Thm 9.13, $w$ non-basic means no difficulty, since (9.5c) (and (9.7)) will then be absent from the zero-reduced cost equations. However, $w=1$ means massive degeneration in (9.6b). This is the extreme case handled by Thm 9.11, corresponding to $\mathbf{x}=\mathbf{0}$ and $\mathbf{v}=\mathbf{v}^{0}$. We have:

Theorem 9.14: If

$$
\begin{equation*}
u_{s}^{0}=f_{s, i_{s}^{*}}+\sum_{t} \tilde{g}_{s, i_{s}^{*}, t} v_{t}^{0}:=\min _{i \in \mathcal{I}(s)}\left[f_{s i}+\sum_{t} \tilde{g}_{s i t} v_{t}^{0}\right] \quad \forall s \tag{9.11}
\end{equation*}
$$

satisfies

$$
\begin{equation*}
\sum_{s} u_{s}^{0}-\sum_{t} \tilde{b}_{t} v_{t}^{0} \geq c \tag{9.12}
\end{equation*}
$$

then $\mathbf{v}=\mathbf{v}^{0}$ is LP-optimum.

Proof: Consider
$\gamma:=\sum_{s} u_{s}^{0}-\sum_{t} \tilde{b}_{t} v_{t}^{0}-c$,
with $\gamma \geq 0$, by assumption.
If $\gamma=0$, then the LP-primal solution $(\mathbf{u}, \mathbf{v}, \boldsymbol{\delta})=\left(\mathbf{u}^{0}, \mathbf{v}^{0}, \mathbf{0}\right)$ is feasible and has objective value $c$, where the maximum value in $(9.5 \mathrm{a})$ is $q \leq c$, i.e. we are done.

If $\gamma>0$, then apply an arbitrary reduction vector $\boldsymbol{\rho} \geq \mathbf{0}, \mathbf{u}=\mathbf{u}^{0}-\boldsymbol{\rho}$, such that $\sum_{S} \rho_{s}=\gamma$. Then the LP-primal solution $(\mathbf{u}, \mathbf{v}, \boldsymbol{\delta})=\left(\mathbf{u}^{0}-\boldsymbol{\rho}, \mathbf{v}^{0}, \mathbf{0}\right)$ satisfies $(9.5 \mathrm{~b})$, since $u_{s} \leq u_{s}^{0}$ for each $s$, and $(9.5 \mathrm{c})$ with equality, and has the maximum objective value $q=c$.
0. Initial step, continuation: initiation and LP-dual start iteration.

Compute $\mathbf{u}^{0}$ as in (9.11) and examine (9.12). If (9.12) is satisfied we have found an LP-optimum directly, with $\mathbf{v}=\mathbf{v}^{0}$. Otherwise we distinguish three alternatives $-\mathrm{A}, \mathrm{B}$ or C - for the initial LP-dual basic solution:

Alt. A. Consider an LP-dual basis consisting of $w, \mathbf{z}^{-}, \mathbf{z}^{+}$and $x_{s, i_{s}^{*}}$ for each $s \neq \hat{s}$, where $\hat{s}$ is an arbitrarily chosen segment, and $i=i_{s}^{*}$ determines $u_{s}^{0}$ for $\mathbf{v}=\mathbf{v}^{0}$ as in Thm 9.14. The zeroreduced costs in (9.5) for this basis means $u_{s}=u_{s}^{0}$ for each $s \neq \hat{s}$, and $u_{\hat{s}}-u_{\hat{s}}^{0}+\gamma=0$ with $\gamma$ as in (9.13). The non-optimum assumption means $\gamma<0$, i.e. $u_{\hat{s}}>u_{\hat{s}}^{0}$. Thus the reduced cost in $(9.5 \mathrm{~b})$ for $x_{\hat{s}, i_{\hat{s}}^{*}}$ is

$$
\bar{c}\left(x_{\hat{s}, i_{\hat{s}}^{*}}\right)=f_{\hat{s}, i_{\hat{s}}^{*}}+\sum_{t} \widetilde{g}_{\hat{s}, \hat{i}_{\hat{s}}^{*}, t^{*}} v_{t}^{0}-u_{\hat{s}}=u_{\hat{s}}^{0}-u_{\hat{s}}<0
$$

Therefore $x_{\hat{s}, i_{\hat{s}}^{*}}$ is taken as entering basic variable. As for the leaving basic variable the initial basic solution ( $\mathbf{x}_{1}, w, \mathbf{z}^{-}, \mathbf{z}^{+}$) and the current matrix column $\left(\overline{\mathbf{n}}_{1}, \bar{n}_{3}, \overline{\mathbf{n}}_{5}\right)$ for the entering variable are the solutions of the linear system of equations, with notations as above,

$$
\left(\begin{array}{ccc|cc}
\mathbf{I} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\
\mathbf{0}^{T} & 1 & 0 & 1 & 1 \\
-\widetilde{\mathbf{G}}_{21} & -\widetilde{\mathbf{b}}_{2} & \mathbf{H}_{2} & -\widetilde{\mathbf{b}}_{2} & -\widetilde{\mathbf{g}}_{2, s, i_{s}^{*}} \\
\mathbf{O} & \mathbf{0} & -\mathbf{L}_{2} & -\boldsymbol{\varepsilon}_{2} & \mathbf{0}
\end{array}\right)
$$

We get $\left(\mathbf{x}_{1}, w, \mathbf{z}^{-}, \mathbf{z}^{+}\right)=\left(\mathbf{0}, 1, \frac{1}{2} \boldsymbol{\varepsilon}, \frac{1}{2} \boldsymbol{\varepsilon}\right)$, as expected, and $\left(\overline{\mathbf{n}}_{1}, \bar{n}_{3}, \overline{\mathbf{n}}_{5}^{-}, \overline{\mathbf{n}}_{5}^{+}\right)=\left(-\mathbf{1}, 1,-\frac{1}{2} \boldsymbol{\beta}, \frac{1}{2} \boldsymbol{\beta}\right)$, where $\beta_{t}$ is defined in (9.8) and $\mathcal{S}_{1}=\mathcal{S}$ is used. Thus $w$ is a leaving basic variable, if $\left|\beta_{\hat{t}}\right|:=\max _{t}\left|\beta_{t}\right|$ satisfies $\left|\beta_{\hat{t}}\right| \leq \varepsilon$. Otherwise either $z_{\hat{t}}^{-}$or $z_{\hat{t}}^{+}$leaves the basis. In this case we formally switch basic variables before continuing; letting $x_{\hat{s}, \hat{i}_{\hat{s}}}$, be basic on the corresponding $\mathcal{S}_{1}$-row in (9.6b) and $w$ basic on the $\mathcal{T}_{1}$-row $\hat{t}$ in (9.6c). In all cases the next iteration will bring $x_{\hat{s}, i_{s}^{*}}>0$ and $w<1$ (and in fact $x_{\hat{s}, i_{s}^{*}}>0$ for every $s$ ). Hence pivoting according to (9.7) and Thm 9.13 is well-defined.

Alt. B. Use the lowest affine majorant iterations $i=i_{s}^{*}$ as defined in (9.11) for each segment s in order to determine the total traffic $\operatorname{cost} F$ (cf. (9.7)) and discounted maintenance costs $\widetilde{G}_{t}$ per year $t$. Let $\mathcal{S}_{1}=\mathcal{S}$; whence $\beta_{t}$ in (9.8) satisfies $\beta_{t}=\beta_{t}^{T O T}=\widetilde{b}_{t}-\widetilde{G}_{t}$. Apply the following rules for the choice of an initial basis:
If $\left|\beta_{t}\right|<\varepsilon$, then take $z_{t}^{-}, z_{t}^{+}$basic, i.e. $t \in \mathcal{T}_{2}$;
else if $\beta_{t}>0$, then take $z_{t}^{-}, y_{t}^{-}$basic, i.e. $t \in \mathcal{T}_{3}$;
else $\left(\beta_{t}<0\right)$ take $z_{t}^{+}, y_{t}^{+}$basic, i.e. $t \in \mathcal{T}_{3}$.
This means that we always start from empty sets $\mathcal{S}_{2}, \mathcal{T}_{1}$.
An advantage for alternative B vs. A is in LP-problem instances where $w=0$ is optimal, since in A it should take some LP-dual iterations, by small $w$-decrements $\varepsilon /\left|\beta_{\hat{t}}\right|$ per iteration, until $w=0$ is reached. In B the basic set $I_{B}=\left\{i_{s}^{*}\right\}$ coincides with the dual subproblem optimum, i.e. $F$ and $\boldsymbol{\beta}=\boldsymbol{\beta}^{T O T}=\widetilde{\mathbf{b}}-\widetilde{\mathbf{G}}$ are known at the root, without any information transfer.

Alt. C. Use the LP-dual optimum basis of the previous Lagrangean dual iteration as a start basis.

Thus after the Lagrangean dual initial iteration we start from the set $\mathcal{S}, \mathcal{T}$ partitions of the optimum LP-basis for the previous iteration (also determining $\mathbf{v}^{0}$ ). Therefore $F, \boldsymbol{\beta}$ and $\boldsymbol{\beta}^{T O T}$ are unchanged. If movable upper bounds $\mathbf{R}$ in (9.5f) are used, any basic $y_{t}^{+}$corresponds to $v_{t}=R_{t}$ in (9.5f), i.e. start values $v_{t} \neq v_{t}^{0}$ are possible.

A disadvantage of alternative C vs. B is in the early Lagrangean dual iterations, where many new subproblem $s$ solutions, i.e. new $i=i_{s}^{*}$, are generated, leading to tedious pivoting in C for the explicit updating of these, mostly in $\mathcal{S}_{1}$. On the other hand, an advantage of C vs. B is in the final iterations, where few new $i_{s}^{*}$ are generated. Our implementation means that we start
with alternative B and turn to C when fewer than a prescribed percentage of the resolved segment $s$ subproblems lead to a new $i_{s}^{*}$.

For alternatives A and B we notice that, since $\mathbf{v}=\mathbf{v}^{0}$ is used in the preceding Lagrangean dual subproblem, $i=i_{s}^{*}$ should be updated for every re-solved segment $s$ subproblem. For the non-re-solved $s$ subproblems (in a partial updating) since $\mathbf{v}=\mathbf{v}^{0}$ was also the previous LPoptimum, the old registrations $i=i_{s}^{*}$ still apply. For alternative C all of $\mathcal{I}_{B}=\left\{i_{s}^{*}\right\}$ is unchanged.

### 9.4.3 Results

The DW-results for the reference case (constant budget, penalised state bound violations, $L=5,88$ processors) are found in Tab 9.5. Initially, one full updating (= one dual iteration) was performed. In all the partial updatings a full dual step length was applied, i.e. $q=0$ according to Sec 9.2.2. In the Tab 9.5 column "Upper target bound $c$ " the RHS-value in (9.5c) is given, as an additional percentage of the best found primal value (for avoiding giant dual steps as well as negative gaps due to discretization errors). In column " $\#$ Partial updatings $M$ " (10), (stagn) within parenthesis means that a full updating is performed every $10^{\text {th }}$ iteration and only at stagnation, respectively, and $M=1$ means full updating in every iteration, as in Tab 9.3 above. In column "Upper $\mathbf{v}$-bound $R$ " the percentages outside and within parenthesis are the flexible upper (and lower) box bounds used in the full and the partial updatings, respectively. In column "CPU-hrs" the values within parenthesis refer to (initial) runs with extensive error checking - otherwise not used in Tab 9.5, for fair comparison with the subgradient method, e.g. in Fig 9.1. The partial updating run with $R=100$ (i.e. $R_{t}=100 \quad \forall t$ ) imitates a formulation (9.5) without upper box bounds, and $R=2$ denotes a fixed upper bound. For $R=100$ the maximum admitted run time (wall time) was reached after many LP-iterations, without any dual \& primal improvements. For $R=2$ the best primal value among all these runs was found and the run was close to stagnation at wall time interruption. During the run series we realised that such stagnations occurred after around 10 iterations for most runs (except for the tightest flexible box). It means that the linearization (9.5) - (9.6) leads to a price point $\mathbf{v}$ where we get stuck - an approximate dual optimum solution. We changed the full updating and stop criterions accordingly in the final runs for $M=1$ and $M=10$, respectively. If the full and partial updating runs are compared, we notice clear CPU time differences. Contrary to the subgradient method, the DW-method does not gain at all from partial updating. This is due to the LP-iteration differences, for the best runs meaning around 54122 and 433750 LP- iterations for $M=1$ and $M=10$, respectively, and the consequences for processor communication. The qualities of the full and partial updating result, as to best dual and primal and residual $\|\mathbf{g}\|_{2}$, are equivalent. The results (for full updating) are not sensitive to the upper target bound $c$ value (maybe not needed at all), whereas the flexible box bounds matter, outside the interval $1-4 \%$. The general impression is a robust method, the best values all approximately equal - and received due to stagnation. Comparing DW in Tab 9.5 with the subgradient method in Tab 9.3 the run times are around 140 hrs (subgradient) and 80 hrs (DW, full updating), and the solution qualities are comparable. Thus all implementation effort has provided us with a superior solver.

| \#Partial updatings M | Upper v bound $R$ | Upper target bound $c$ | CPU-hrs | \#Dual iterations | Stop criterion | Best dual (kSEK) | Best primal (kSEK) | Relative gap | $\begin{array}{r} \\|\mathbf{g}\\|_{2} \\ (\mathrm{kSEK}) \end{array}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 20\% | +5\% | (256) | 21 | (stagn 12 iter) | 2541754 | 2541865 | 0.000044 | 42532 |
| 1 | 10\% | +1\% | 94 | 9 | stagnation | 2541951 | 2541739 | -0.000083 | 40100 |
| 1 | 5\% | +1\% | 93 (\&148) | 11 | stagnation | 2542193 | 2541837 | -0.000140 | 34785 |
| 1 | 2.5\% | +5\% | 78 | 10 | stagnation | 2542037 | 2541492 | -0.000215 | 32404 |
| 1 | 2.5\% | +1\% | 81 | 11 | stagnation | 2542208 | 2541426 | -0.000308 | 27497 |
| 1 | 2.5\% | +0.0001\% | 81 | 11 | stagnation | 2542208 | 2541426 | -0.000308 | 27497 |
| 1 | 1.25\% | +1\% | 83 | 13 | stagnation | 2542043 | 2541584 | -0.000181 | 25070 |
| 1 | 0.625\% | +1\% | 110 | 20 | stagnation | 2542065 | 2541516 | -0.000216 | 38562 |
| 10(10) | 100 | +100\% | (264) | 3.7 | wall time | 2539582 | 2546765 | 0.002828 | 71561 |
| 10(10) | 2 | +1\% | (264) | 6.1 | wall time | 2539582 | 2541486 | 0.000750 | 71561 |
| 10(10) | 20\%(6\%) | +5\% | (352) | 17.3 | wall (stagn 7 it) | 2542120 | 2541654 | -0.000183 | 23776 |
| 10(10) | 10\%(3\%) | +5\% | (352) | 18.1 | wall (stagn 7 it) | 2542165 | 2541534 | -0.000248 | 28886 |
| 10(stagn) | 10\%(3\%) | +1\% | 219 | 10 | stagnation | 2542127 | 2541479 | -0.000255 | 26000 |
| 10(stagn) | 5\%(1.5\%) | +1\% | 182 (\&232) | 8 | stagnation | 2542258 | 2541581 | -0.000266 | 26321 |
| 10(stagn) | 5\%(5\%) | +1\% | 211 | 10 | stagnation | 2542194 | 2541563 | -0.000248 | 24785 |
| 10(stagn) | 2.5\%(2.5\%) | +1\% | 208 | 8 | stagnation | 2542149 | 2541487 | -0.000261 | 25054 |

Table 9.5 Results for the Dantzig-Wolfe method with full and partial updating, reference run.
At the registration of a new subproblem solution for a segment we check if DW-extrapolation from any previous solution (of a partial or full updating) shows a lower majorant value. This is theoretically impossible but occurs in practice due to discretization errors, primarily in the early dual iterations (far from the optimum). For, e.g., the last run in Tab 9.5 around $10 \%$ of the segments in dual iteration № 3 shows such errors, although each relative error is small - of order $0.1 \%$, as a rule - and after iteration № 3 a vanishing amount occurs.

In some LP-iterations with $w \approx 1$ (cf. Thm 9.11) we got numerical trouble. A more robust approach would be to introduce a variable $\rho \geq 0$, replace ( 9.5 c ) with a penalised constraint

$$
\sum_{s} u_{s}-\sum_{t} \widetilde{b}_{t} v_{t}-\rho \leq c
$$

and add the penalisation term $-\gamma \cdot \rho$ to the objective (9.5a). In the LP-dual (9.6) this corresponds to an additional upper bound $w \leq \gamma$; where a $\gamma$-value clearly below 1, e.g. $\gamma=0.5$, should be chosen to avoid the numerical trouble.

### 9.5 Primal heuristics

Running our basic model, the standard heuristics in Sec 4.2 .3 presumes simultaneous and repeated access to data for all segments in the current dual iteration. This means either slow, intense swapping, if the information is on file, or the use of a large cluster of co-working processors with large primary memories. A short cut might be to concentrate the dual iterating entirely to the dual problem, making the necessity minimal for message passing between the computers, and to postpone the generation of primally feasible maintenance plans to afterwards, based on the dually generated maintenance plans. Although we must have one plan per road segment, we are free to mix the registered dual iterations (when the plans were generated) for different segments, in order to accomplish primal feasibility and lowest cost on the network level. Since every dual iteration means several CPU-hours it may be worthwhile to save more than one (close to subproblem-optimal) plan per iteration and segment.

### 9.5.1 Problem

For any segment $s$ each dual iteration $i \in I(s)$, the set of registered solutions, is characterised by the total discounted traffic + penalty costs $\widetilde{f}_{s i}$ and discounted maintenance costs $\left(\widetilde{g}_{s i t}\right)_{t}$. The discounted budgets are $\left(\widetilde{b}_{t}\right)_{t}$. Since all the registered solutions have passed the dual return rate constraints in (4.6), such considerations are implicit here. The simplified primal problem is to

$$
\begin{align*}
& \underset{\left(w_{s i}\right)_{s, i}}{\operatorname{minimise}} \sum_{s, i} \widetilde{f}_{s i} w_{s i}  \tag{9.14a}\\
& \text { subject to } \begin{cases}\sum_{s, i} \widetilde{g}_{s i t} \cdot w_{s i} \leq \widetilde{b}_{t} & \forall t \\
\sum_{i} w_{s i}=1 & \forall s \\
w_{s i} \in\{0,1\} & \forall s, i .\end{cases} \tag{9.14b}
\end{align*}
$$

The binary variable constraints make the problem non-convex. In practice an LP-relaxation of (9.14d) would find a solution with at most as many fractional segments as there are budget constraints. However, this is far from an integer solution and we will try another way. The idea is to modify (9.14), directing the search for feasible solutions to regions with greater chances of success. By a Lagrangean relaxation of the budget constraints (9.14b), introducing Lagrangean multipliers (dual variables) $v_{t} \geq 0$, the Lagrangean dual becomes to

$$
\begin{equation*}
\underset{v \geq 0}{\operatorname{maximise}} \varphi(v), \tag{9.15}
\end{equation*}
$$

where the dual objective value is the optimal value of the dual subproblem:

$$
\varphi(\mathbf{v})=\min _{\left(w_{s i}\right)_{s, i}}\left[\sum_{s, i}\left(\tilde{f}_{s i}+\sum_{t} v_{t} \tilde{g}_{s i t}\right) w_{s i}-\sum_{t} v_{t} \widetilde{b}_{t}\right] \text { s.t. } \sum_{i} w_{s i}=1 \quad \forall s, \quad w_{s i} \in\{0,1\} \quad \forall s, i
$$

For fixed $\mathbf{v}$-values the dual subproblem becomes separable into segment $s$-specific subproblems

$$
\begin{equation*}
\underset{\left(w_{s i}\right)_{i}}{\operatorname{minimise}} \sum_{i}\left(\tilde{f}_{s i}+\sum_{t} v_{t} \tilde{g}_{s i t}\right) w_{s i} \text { s.t. } \sum_{i} w_{s i}=1, \quad w_{s i} \in\{0,1\} \quad \forall i . \tag{9.16}
\end{equation*}
$$

The $s$-subproblem has the obvious optimum $w_{s i_{s}^{*}}=1: i_{s}^{*} \equiv i_{s}^{*}(\mathbf{v}) \in \underset{i \in \mathcal{I}(s)}{\arg \min }\left(\tilde{f}_{s i}+\sum_{t} v_{t} \widetilde{g}_{s i t}\right)$.
Denoting the total traffic and maintenance costs for the subproblem optimum by, respectively, $F(\mathbf{v}):=\sum_{s} \widetilde{f}_{s i_{s}^{*}}$ and $g_{t}(\mathbf{v}):=\sum_{s} \widetilde{g}_{s i_{s}^{*} t} \quad \forall t$, the dual problem (9.15) can be written

$$
\begin{equation*}
\underset{\mathbf{v} \geq \mathbf{0}}{\operatorname{maximise}} \quad \varphi(\mathbf{v})=F(\mathbf{v})+\sum_{t} v_{t} \cdot\left(g_{t}(\mathbf{v})-\widetilde{b}_{t}\right) \tag{9.17}
\end{equation*}
$$

Since we are to find primally feasible solutions by solving a dual problem we will tamper with the budget levels, using $b_{t}^{\kappa} \leq \widetilde{b}_{t}$, instead of $\widetilde{b}_{t}$ in (9.17), in order to increase the chances of feasibility. Here $\kappa$ denotes the heuristics iteration number. If some $\mathbf{v}=\mathbf{v}^{\kappa}$ would provide a primally feasible subproblem solution, we may, e.g., choose $b_{t}^{K}=g_{t}\left(\mathbf{v}^{K}\right) \quad \forall t$, without jeopardising the feasibility. In fact this choice maximises the chances for $\mathbf{v}^{\kappa}$, in the sense of Thm 9.15 below. For fixed resources $\mathbf{b}$ the linear (subgradient based) $\varphi$-extrapolation $\hat{\varphi}^{\kappa}$
from $\mathbf{v}^{K}$ satisfies

$$
\begin{aligned}
& \hat{\varphi}^{\operatorname{lin}, \kappa}(\mathbf{v} ; \mathbf{b}):=\varphi\left(\mathbf{v}^{\kappa}\right)+\left(\mathbf{v}-\mathbf{v}^{\kappa}\right)^{T}\left(\mathbf{g}\left(\mathbf{v}^{\kappa}\right)-\mathbf{b}\right)= \\
& =F\left(\mathbf{v}^{\kappa}\right)+\mathbf{v}^{\kappa^{T}}\left(\mathbf{g}\left(\mathbf{v}^{\kappa}\right)-\mathbf{b}\right)+\left(\mathbf{v}-\mathbf{v}^{\kappa}\right)^{T}\left(\mathbf{g}\left(\mathbf{v}^{\kappa}\right)-\mathbf{b}\right)= \\
& =F\left(\mathbf{v}^{\kappa}\right)+\mathbf{v}^{T}\left(\mathbf{g}\left(\mathbf{v}^{\kappa}\right)-\mathbf{b}\right) .
\end{aligned}
$$

Theorem 9.15: For the subproblem optimum at $\mathbf{v}^{\kappa}$, characterised by costs $F\left(\mathbf{v}^{\kappa}\right)$, $\mathbf{g}\left(\mathbf{v}^{\kappa}\right)=\mathbf{b}^{\kappa}$, and for any $\mathbf{b} \geq \mathbf{b}^{\kappa}, \mathbf{v} \geq \mathbf{0}$ we have

$$
\hat{\varphi}^{\operatorname{lin}, \kappa}(\mathbf{v} ; \mathbf{b}) \leq \hat{\varphi}^{\operatorname{lin}, \kappa}\left(\mathbf{v}^{\kappa} ; \mathbf{b}^{\kappa}\right) .
$$

Proof: $\hat{\varphi}^{\operatorname{lin}, \kappa}(\mathbf{v} ; \mathbf{b})=F\left(\mathbf{v}^{\kappa}\right)+\mathbf{v}^{T}\left(\mathbf{g}\left(\mathbf{v}^{\kappa}\right)-\mathbf{b}\right)$. Here $\mathbf{v} \geq \mathbf{0}$ and $\mathbf{b} \geq \mathbf{b}^{\kappa}=\mathbf{g}\left(\mathbf{v}^{\kappa}\right)$, i.e.

$$
\hat{\varphi}^{\operatorname{lin}, K}(\mathbf{v} ; \mathbf{b}) \leq F\left(\mathbf{v}^{\kappa}\right) .
$$

For $\mathbf{v}^{\kappa}, \mathbf{b}^{\kappa}$ we get

$$
\hat{\varphi}^{\operatorname{lin}, \kappa}\left(\mathbf{v}^{\kappa} ; \mathbf{b}^{\kappa}\right)=\varphi\left(\mathbf{v}^{\kappa}\right)=F\left(\mathbf{v}^{\kappa}\right)+\mathbf{v}^{\kappa^{T}}\left(\mathbf{g}\left(\mathbf{v}^{\kappa}\right)-\mathbf{b}^{\kappa}\right)=F\left(\mathbf{v}^{\kappa}\right)+0 .
$$

The $s$-subproblem will have an unchanged optimum registration $i_{s}^{*}$ in a whole $\mathbf{v}$-region, where the optimal value $\tilde{f}_{s_{s}^{*}}+\sum_{t} v_{t} \widetilde{g}_{s_{s}^{*} t}$ varies linearly. Since $F(\mathbf{v}), \mathbf{g}(\mathbf{v})$ are sums over $s$ also the Lagrangean dual problem has $\boldsymbol{v}$-regions where the objective $\varphi=\varphi(\boldsymbol{v})$ in (9.20) varies linearly, i.e. forms affine hyperplane facets in the $(\mathbf{v}, \varphi)$-space. A switch of $(\mathbf{v}, \varphi)$-facet corresponds to a switch of $i_{s}^{*}$ for some $s$. By keeping track of all the relevant $i_{s}^{*}$-switches we can control the full dual.

### 9.5.2 Method

We use a doubly iterative search method.

- Outer iteration № $\kappa$ : For fixed $\mathbf{b}^{\kappa}$ we have determined the dual (9.17) optimum $\mathbf{v}=\mathbf{v}^{\kappa}$, specified by a set $\mathcal{A}$ of intersecting affine majorants characterised by $\left(F\left(\mathbf{v}_{a}\right), \mathbf{g}\left(\mathbf{v}_{a}\right)\right)_{a \in \mathcal{A}}$, the optimal costs for the dual subproblems at $\left(\mathbf{v}_{a}\right)_{a \in A}$. Choosing one $\mathbf{v}_{a}$ with the smallest number of budget violations $g_{t}\left(\mathbf{v}_{a}\right)>b_{t}^{\kappa}$, we update

$$
b_{t}^{\kappa+1}:=\min \left[b_{t}^{\kappa}-\rho^{\kappa} \cdot\left(e^{g_{t}\left(\mathbf{v}_{a}\right)-b_{t}^{\kappa}}-1\right), \widetilde{b}_{t}\right]
$$

- The inner iteration consists of the following steps:

1. Based on a given number $A=|\mathcal{A}|$ of affine majorants at a given iterate $\mathbf{v}=\hat{\mathbf{v}}$, determine the search direction $\mathbf{r}$ that maximises $\varphi$ in the unit-sphere surrounding $\hat{\mathbf{v}}$. Decrement $A$ if a majorant is redundant for the maximum $\mathbf{r}$.
2. By going through all segment registrations $(I(s))_{s}$, determine the $N$ step lengths $\left(\hat{\sigma}_{n}\right)_{n=1}^{N}$ closest to $\hat{\mathbf{v}}$ along $\mathbf{r}$, where a segment switches to a new subproblem optimum.
3. Choose step length $\sigma=\frac{1}{2}\left(\hat{\sigma}_{N-1}+\hat{\sigma}_{N}\right)$ or, if the dual objective $\varphi$ shows descent at $\hat{\sigma}_{N}$, choose $\sigma=\hat{\sigma}_{n}$ where $\varphi$ reaches the top along $\mathbf{r}$, and increment $A$ by the (two) surrounding affine majorants.

The use of the exponential in the outer iteration for updating $\mathbf{b}^{\kappa}$ means that a budget violation leads to a bigger absolute resource change than a cost below budget. The inner iteration process stops if a point $\mathbf{v}^{K}$ with $\mathbf{g}\left(\mathbf{v}^{K}\right)=\mathbf{b}^{K}$ is reached.

The inner iterating means that we initially use iterates $\mathbf{v}$ such that the subproblem (9.17) optimum has a single hyperplane active in $\mathcal{A}$, and later on turn to a ridge based climbing.

The meaning of $A$ non-redundant majorants is that $\mathbf{r}$ leads to a point where all majorant conditions $\varphi(\mathbf{v}) \leq F\left(\mathbf{v}_{a}\right)+\mathbf{v}^{T}\left(\mathbf{g}\left(\mathbf{v}_{a}\right)-\mathbf{b}\right)$ are satisfied with equality. The choice of search direction $\mathbf{r}$ guarantees $\varphi$-ascent, as we will prove in Cor 9.1 below. The choice of step length guarantees that the next iterate gets an improved $\varphi$-value. With $T$ budget constraints the inner iterative procedure goes on (at least) until $A=T+1$; which means that we get stuck in a point - the majorant based top, by construction. If the tampered budgets are successfully chosen the reached top (or some of the preceding inner iterates) is a primally feasible point, close to the primal optimum. Otherwise, further tampering is needed.

We look at the method details, first the choice of direction. For each segment $s$ denote by $i=i_{s}^{*}$ the $s$-iteration that solves the dual $s$-subproblem (9.16) for a given inner iterate $\mathbf{v}=\hat{\mathbf{v}}$. Letting $F_{a}:=F\left(\mathbf{v}_{a}\right), \mathbf{g}_{a}:=\mathbf{g}\left(\mathbf{v}_{a}\right)-\mathbf{b}$ denote the traffic cost and maintenance cost/budget residuals, respectively, at some registered point $\mathbf{v}=\mathbf{v}_{a}$, the extrapolated $\varphi$-value to the current iterate $\hat{\mathbf{v}}$ from $\mathbf{v}_{a}$ is written

$$
\hat{\varphi}_{a} \equiv \varphi_{a}^{\operatorname{lin}}(\hat{\mathbf{v}}):=\varphi\left(\mathbf{v}_{a}\right)+\mathbf{g}_{a}^{T}\left(\hat{\mathbf{v}}-\mathbf{v}_{a}\right)=F_{a}+\mathbf{g}_{a}{ }^{T} \hat{\mathbf{v}} \quad \forall a \in \mathcal{A}
$$

Using this and the search direction $\mathbf{r}:=\mathbf{v}-\hat{\mathbf{v}}$ from $\hat{\mathbf{v}}$ to a candidate $\mathbf{v}$, the extrapolated value at $\mathbf{v}$ can be written

$$
\begin{array}{ll}
\varphi_{a}^{\operatorname{lin}}(\mathbf{v})=\varphi\left(\mathbf{v}_{a}\right)+\mathbf{g}_{a}{ }^{T}\left(\mathbf{v}-\mathbf{v}_{a}\right)=\varphi\left(\mathbf{v}_{a}\right)+\mathbf{g}_{a}{ }^{T}(\mathbf{v}-\hat{\mathbf{v}})+\mathbf{g}_{a}{ }^{T}\left(\hat{\mathbf{v}}-\mathbf{v}_{a}\right)= \\
=\varphi_{a}^{\operatorname{lin}}(\hat{\mathbf{v}})+\mathbf{g}_{a}{ }^{T} \mathbf{r} \equiv \hat{\varphi}_{a}+\mathbf{g}_{a}{ }^{T} \mathbf{r} . & \forall a \in \mathcal{A} . \tag{9.18}
\end{array}
$$

The majorant property for the dual (9.17) means that the optimal value $\varphi(\mathbf{v})$ at the end point $\mathbf{v}$ of $\mathbf{r}$ must satisfy

$$
\varphi(\mathbf{v}) \leq \hat{\varphi}_{a}+\mathbf{g}_{a}{ }^{T} \mathbf{r} \quad \forall a \in \mathcal{A}
$$

Assuming an active set $\mathcal{A}$ of $A=|\mathcal{A}| \geq 1$ affine majorants in $(\mathbf{v}, \varphi)$-space, we choose $\mathbf{r}$ as to

$$
\begin{align*}
& \text { maximise } \varphi  \tag{9.19a}\\
& \text { subject to }\left\{\begin{array}{r}
\varphi \leq \hat{\varphi}_{a}+\mathbf{g}_{a}{ }^{T} \mathbf{r} \quad a=1, \ldots A \\
\mathbf{r}^{T} \mathbf{r} \leq 1 .
\end{array}\right. \tag{9.19b}
\end{align*}
$$

In practice we expect $\|\mathbf{r}\|_{2}=1$.
Letting $\hat{\varphi}:=\left(\hat{\varphi}_{a}\right)_{a}, \mathbf{G}:=\left(\mathbf{g}_{1}, \ldots, \mathbf{g}_{A}\right)$, we have
Theorem 9.16: Assume that $\mathbf{G}$ has full rank $A$. The optimum direction $\mathbf{r}$ for (9.19) satisfies

$$
\begin{align*}
& \mathbf{r}=\mathbf{G} \mathbf{y}, \text { where } \mathbf{y}=\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1}(\varphi \cdot \mathbf{1}-\hat{\varphi}) \text { and the target value } \varphi \text { is } \\
& \varphi=\frac{\mathbf{1}^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \hat{\boldsymbol{\varphi}}}{\mathbf{1}^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \mathbf{1}}+\sqrt{\left(\frac{\mathbf{1}^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \hat{\varphi}}{\mathbf{1}^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \mathbf{1}}\right)^{2}+\frac{1-\hat{\varphi}^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \hat{\varphi}}{\mathbf{1}^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \mathbf{1}}} . \tag{9.20}
\end{align*}
$$

(For implementation purposes introducing $\mathbf{z}$, solving $\left(\mathbf{G}^{T} \mathbf{G}\right) \mathbf{z}=\mathbf{1}$, and $\mathbf{x}$, solving $\left(\mathbf{G}^{T} \mathbf{G}\right) \mathbf{x}=\hat{\boldsymbol{\varphi}}$, we compute $\varphi=\frac{\mathbf{1}^{T} \mathbf{x}}{\mathbf{1}^{T} \mathbf{z}}+\sqrt{\left(\frac{\mathbf{1}^{T} \mathbf{x}}{\mathbf{1}^{T} \mathbf{z}}\right)^{2}+\frac{1-\hat{\boldsymbol{\varphi}}^{T} \mathbf{x}}{\mathbf{1}^{T} \mathbf{z}}}$ and let $\mathbf{y}$ solve $\left.\mathbf{y}=\varphi \cdot \mathbf{z}-\mathbf{x}.\right)$

Especially if all conditions ( 9.19 b ) are active (equalities) at the current point $\hat{\mathbf{v}}$, then $\mathbf{y}=\frac{1}{\sqrt{\mathbf{1}^{T} \mathbf{z}}} \cdot \mathbf{z}$ and $\varphi=\hat{\varphi}+\frac{1}{\sqrt{\mathbf{1}^{T} \mathbf{z}}}$, where $\mathbf{z}$ solves $\left(\mathbf{G}^{T} \mathbf{G}\right) \mathbf{z}=\mathbf{1}$.

Proof: Introduce non-negative Lagrangean multipliers $\lambda$ and $\frac{1}{2} \mu$ for (9.19b) - (9.19c). The Lagrangean is

$$
L(\mathbf{r}, \varphi ; \lambda, \mu)=-\varphi+\lambda^{T}\left(\varphi \cdot \mathbf{1}-\hat{\boldsymbol{\varphi}}-\mathbf{G}^{T} \mathbf{r}\right)+\frac{1}{2} \mu \cdot\left(\mathbf{r}^{T} \mathbf{r}-1\right) .
$$

By differentiating for $\varphi$ and each $r_{t}$, and putting the partial derivatives to zero we get

$$
\lambda^{T} \mathbf{1}=1, \mu \cdot \mathbf{r}=\mathbf{G} \lambda .
$$

First assume $\mu=0$, implying $\mathbf{0}=\mathbf{G} \lambda$. By assumption $\mathbf{G}$ has full rank; hence $\lambda=\mathbf{0}$, contradicting $\lambda^{T} \mathbf{1}=1$. Thus $\mu \neq 0$ and we have $\mathbf{r}=\frac{1}{\mu} \cdot \mathbf{G} \lambda,\|\mathbf{r}\|_{2}=1$. The meaning of (9.19b) active are the equalities (9.18) for every $a \in \mathcal{A}$. Using all this we get

$$
\varphi \cdot \mathbf{1}=\hat{\varphi}+\frac{1}{\mu} \cdot \mathbf{G}^{T} \mathbf{G} \lambda, \text { i.e. } \lambda=\mu \cdot\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1}(\varphi \cdot \mathbf{1}-\hat{\varphi})
$$

Here $\lambda^{T} \mathbf{1}=1$ implies $\mu=1 /(\varphi \cdot \mathbf{1}-\hat{\boldsymbol{\varphi}})^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \mathbf{1}$,

$$
\lambda=\frac{1}{(\varphi \cdot \mathbf{1}-\hat{\varphi})^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \mathbf{1}} \cdot\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1}(\varphi \cdot \mathbf{1}-\hat{\varphi}) \quad \text { and } \quad \mathbf{r}=\mathbf{G}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1}(\varphi \cdot \mathbf{1}-\hat{\varphi})
$$

Finally (9.19c) corresponds to

$$
1=\mathbf{r}^{T} \mathbf{r}=(\varphi \cdot \mathbf{1}-\hat{\varphi})^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1}(\varphi \cdot \mathbf{1}-\hat{\varphi})
$$

This is a $2^{\text {nd }}$ order equation for $\varphi$, in general having two solutions $\pm$ in (9.20); the maximal $\varphi$-value corresponding to + .
Especially, if the entire $\mathcal{A}$ is active at $\hat{\boldsymbol{\varphi}}$, i.e. for $\mathbf{r}=\mathbf{0}$, then all the component values there must coincide, i.e. $\hat{\varphi}=\hat{\varphi} \cdot \mathbf{1}$. Using this in all the general results above, the special result follows. Since in the general $\mu$-expression $\mathbf{1}^{T}\left(\mathbf{G}^{T} \mathbf{G}\right)^{-1} \mathbf{1}$ is a quadratic form, by assumption positive, $\varphi>\hat{\varphi}$ is equivalent to $\mu>0$.
9.5 Primal heuristics

## 10 References

AASHTO (1993): Guide for design of pavement structures, American Association of State Highway and Transportation Officials, Washington DC.

Ahuja RK, Magnanti TL and Orlin JB (1993): Network flows, Prentice-Hall, Upper Saddle River, NJ.
Altman E (1998): Constrained Markov decision processes with total cost criteria: Lagrangian approach and dual linear program, Mathematical methods of Operations Research 48:387-417.
Alviti E, Kulkarni RB, Johnson EG, Clark N, Walrafen V, Nazareth L and Stone JC (1994):
Enhancements to the Network Optimization System, Proc. $3^{\text {rd }}$ international conference on managing pavements, 2:190-194, National Academy Press, Washington DC.

Alviti E, Johnson EG, Kulkarni RB, Nazareth L and Stone JC (1996): A pavement network optimisation system using Dantzig-Wolfe decomposition, Proc. Network Optimization, 1-16, Florida.
Also in: Pardalos PM et al, eds. (1997), Network optimization, Lecture notes in economics and mathematical systems (LNEMS) 450, Springer-Verlag, Berlin.

Andersson M, Hultkrantz L and Nilsson J-E (2001): Drift och underhåll av det statliga vägnätet på samhällsekonomiska grunder, Institutet för transportforskning (TFK) and CTEK, Borlänge.
Andersson P-Å, Enocksson C-G and Lang J (2004): OPM-applikation R2 - basdata och funktionssamband, unpublished PM, LiU and VV.
Artman DH, Liebman JS and Darter MI (1983): Optimization of long-range major rehabilitation of airfield pavements, Transportation Research Record 938:1-11.

Bellman RE (1957): Dynamic programming, Princeton University Press, Princeton NJ.
Berman A and Plemmons RJ (1979): Nonnegative matrices in the mathematical sciences, Academic Press, New York.

Bertsekas DP (1995): Dynamic programming and optimal control I, Athena Scientific, Belmont Mass.
Beutler FJ \& Ross KW (1985): Optimal policies for Controlled Markov Chains with a Constraint, Journal of Mathematical Analysis and Applications 112:236-252.
Beutler FJ \& Ross KW (1986): Time-average optimal constrained Semi-Markov Decision Processes, Advances in Applied Probability 18:341-359.

Butt AA, Shahin MY, Carpenter SH and Carnahan JV(1994): Application of Markov process to pavement management systems at network level, Proc. $3^{\text {rd }}$ international conference on managing pavements, 2:159-172, National Academy Press, Washington DC.

Camerini PM, Fratta L and Maffioli F (1975): On improving relaxation methods by modified gradient techniques, Mathematical Programming study 3:26-34.

Carnahan JV, Davis WJ, Shihin MY, Keane PL and Wu MI (1987): Optimal maintenance decisions for pavement management, Journal of Transportation Engineering 113(5):554-572.

Carnahan JV (1988) Analytical framework for optimizing pavement maintenance, Journal of Transportation Engineering 114(3):307-322.

Chan WT, Fwa TF and Hoque KZ (2001): Constraint handling methods in pavement maintenance programming, Transportation Research Part C 9:175-190.
Chasey AD (1995): A framework for determining the impact of deferred maintenance and/or obsolescence of a highway system, Thesis UMI 9528583,Virginia Polytechnic Institute and State Univ.

## 10 References

Chua DKH (1996): Framework for PMS using mechanistic distress submodels, Journal of Transportation Engineering 122(1):29-40.
Dantzig GB (1963): Linear programming and extensions, Princeton university press, Princeton.
Darabian F (2003): Distrik-karlstad-pmsRUH, VV Region Väst, Vägteknik, Göteborg.
Davis WJ and Carnahan JV (1987): Decision support for road surface maintenance, OMEGA Int. J. of Mgmt Sci. 15(4):313-322.
de Ghellinck GT (1960): Les problèmes de decisions séquentielles, Cahiers du Centre de Recherche Opérationelle 2:161-179.

Dekker R (1995): The use of operations research models for maintenance decision making, Microelectron. Reliab. 35(9-10):1321-1331.

Dekker R (1996): Applications of maintenance optimization models: a review and analysis, Reliability Engineering and System Safety 51:229-240.

Dekker R and Scharf PA (1998): On the impact of optimization models in maintenance decision making: the state of the art, Reliability Engineering and Systems Safety 60:111-119.
d'Epenoux F (1960): Sur un problème de production et de stockage dans l'aléatoire, Revue Française de Recherche Opérationelle 14:3-16.
d'Epenoux F (1963): A probabilistic production and inventory problem, Management Science 10:98108.

Denardo EV \& Fox BL (1968): Multichain Markov renewal programs, SIAM J Appl Math 16(3):468487.

Denardo EV (1970): On Linear programming in a Markov decision problem, Management Science 16(5):281-288.

Devore JL (2004): Probability and statistics for engineering and the sciences, $6^{\text {th }}$ ed, Duxbury, Toronto.
Draper N R and Smith H (1981): Applied regression analysis, $2^{\text {nd }}$ ed, Wiley, New York.
Effektsamband 2000 (2001a): Gemensamma förutsättningar, publ. 2001:75, VV, Borlänge.
Effektsamband 2000 (2001b): Drift och Underhåll - handledning, publ. 2001:79, VV, Borlänge.
Evans JR and Minieka E (1992): Optimization algorithms for networks and graphs, Dekker, NY.
Feighan KJ, Shahin MY, Sinha KC and White TD (1988): Application of dynamic programming and other mathematical techniques to pavement management systems, Transportation Research Record 1200:90-98.

Feinberg EA \& Schwartz A (1995): Constrained Markov decision models with weighted discounted rewards, Mathematics of Operations Research 20(2):302-320.

Fishman GS (1973): Concepts and methods in discrete event digital simulation, Wiley, New York.
Flintsch GW, Zaniewski JP and Medina A (1998): Development of a knowledge-based formula to prioritize pavement rehabilitation projects, Transportation Research Record 1643.
Forsberg I and Göransson N-G (2000): Tillståndsmätning av observationssträckor med Laser RST, VTI notat 12-2000, VTI.
Frid EB (1972): On optimal strategies in control problems with constraints, Theory of Probability and its Applications 17(1):188-192.

Fwa TF, Chan WT and Tan CY (1994): Optimal programming by genetic algorithms for pavement management, Transportation Research Record 1455:31-41.

Golabi K, Kulkarni RB and Way GB (1982): A statewide pavement management system, Interfaces 12(6):5-21.

Grivas DA and Schultz BC (1994): Integration of pavement and bridge management systems: a case study, Proc. $3^{\text {rd }}$ internat. conf. on managing pavements, 1:22-28, Nat. Acad. Press, Washington DC.

Guo X \& Shi P \& Zhu W (2000): A new strong optimality criterion for nonstationary Markov decision processes, Mathematical methods of Operations Research 52:287-306.

Haas R, Hudson WR and Zaniewski J (1994): Modern pavement management, Krieger, Malabar, Fl.
Haas R, coord. (1997): Pavement design and management guide, Transportation Association of Canada, Ottawa.

Hagquist RF (1994): An 'Apollo project' for management science, Special report, OR/MS Today, Oct 1994:14-16.

Hanna PB, Hanna AS and Papagiannakis TA (1993): Knowledge-based advisory system for flexible pavement routine maintenance, Can. J. Civ. Eng. 20:154-163.

Haugødegård T, Johansen JM, Bertelsen D and Gabestad K (1994): Norwegian public roads administration: a complete pavement management system in operation, Proc. $3^{\text {rd }}$ international conference on managing pavements 2:25-33, National Academic Press, Washington DC.

Haviv M (1996): On constrained Markov decision processes, Operations Research Letters 19:25-28.
HDM-4 Highway Development \& Management (2000): Series Coll. ISBN 2-84060-058-7, The World Road Ass. (PIARC), Paris, and The World Bank, Washington DC.

Holland JH (1975): Adaptation in natural and artificial systems, Univ. of Michigan Press, Ann Arbor.
Hordijk A \& Kallenberg LCM (1979): Linear Programming and Markov Decision Chains, Management Science 25(4):352-362.

Howard RA (1960): Dynamic programming and Markov processes, Technology Press of MIT, Cambridge Mass. and Wiley, New York.

Hudson WR and Haas RCG (1994): What are the true costs and benefits of pavement management, Proc. $3^{\text {rd }}$ international conference on managing pavements 3:29-36, Nat. Acad. Press, Washington DC.

Ihs A and Magnusson G (2000): Betydelsen av olika karakteristika hos beläggningsytan för trafik och omgivning, VTI notat 71-2000, VTI, Linköping.

Ihs A and Sjögren L (2003): An overview of HDM-4 and the Swedish Pavement Management System (PMS), workshop pres., VTI, Linköping.

Jansen JM and Schmidt B (1994): Performance models and prediction of increase in overlay need in the Danish state highway pavement management system BELMAN, Proc. $3^{\text {rd }}$ international conference on managing pavements 1:74-84, National Academic Press, Washington DC.
Jewell WS (1962): Optimal flow through networks with gains, Operations Research 10 (4):476-499.
Johnson EG, Connor BG and Kulkarni RB (1994): Organizational implementation and application of Alaska's pavement management system, Proc. $3^{\text {rd }}$ international conference on managing pavements 1:198-202, National Academy Press, Washington DC.

Kleijnen (1974): Statistical techniques in simulation, Dekker, New York.

## 10 References

Knudsen F and Simonsen P (1994): How decision makers at various levels use output from the Danish pavement management system BELMAN, Proc. $3^{\text {rd }}$ international conference on managing pavements 2:74-82, National Academy Press, Washington DC.

Kulkarni RB (1984): Dynamic decision model for a pavement management system, Transportation Research Record 997:11-18.

Lang J (1996): Översiktlig analys av det statliga belagda vägnätet, Technical report, VV, Borlänge Lang J(1997): Underhållskostnader för investeringskalkyler - belagd väg, PM VV, Borlänge.

Lang J (1999): Översiktlig analys av det statliga belagda vägnätet 1999, Technical report, VV. Lang J and Dahlgren JM (2001): Prediction models in the Swedish PMS, Proc. $5^{\text {th }}$ international conference on managing pavements, 2001.
Lang J and Potucek M (2001): Pavement management systems in Sweden, Proc. $5^{\text {th }}$ international conference on managing pavements.
Lang J (2002): Förändringar i Hips 1995-2002, Technical report, VV, Borlänge.
Lang, J (2003): Private discussion, VV Hk, Borlänge.
Lang J (2004): Personal communication at CDU:R2 meetings 2003-2004, VV Hk, Borlänge.
Lang J (2007): Karaktärisering av vägytans jämnhet, progressive doctoral project CDU:T24, KTH and VV.

Larsson T, Patriksson M and Strömberg A-B (1996): Conditional subgradient optimization - theory and applications, European Journal of Operations Research 88:382-403.

Li N, Huot M and Haas R (1997a): Cost-effectiveness-based priority programming of standardized pavement maintenance, Transportation Research Record 1592:8-16.

Li N, Haas R and Xie W-C (1997b): Investigation of relationship between deterministic and probabilistic prediction models in pavement management, Transportation Research Record 1592:70-79.
Li N, Haas R and Huot M (1998): Integer programming of maintenance and rehabilitation treatments for pavement networks, Transportation Research Record 1629:242-248.
Li Y and Madanat S (2002): A steady-state solution for the optimal pavement resurfacing problem, Transportation Research Part A, 36:525-535.

Lindberg PO, Lundgren J, Mattsson L-G and Strömberg A-B (1997): Optimeringsmetoder för resursfördelning inom drift och underhåll av vägar, LiTH-MAT-R-97-14, LiU.

Lindberg PO, Larsson T and Hector S (2001): Optimization methods for the allocation of resources in road operation and maintenance activities, State of the art report, LiU and University of Gävle.

Luenberger DG (1979): Introduction to dynamic systems, Wiley \& Sons, New York
Lytton RL (1994): Optimization techniques - tutorial session, $3^{\text {rd }}$ international conference on managing pavements, National Academy Press, Washington DC
Madanat S (1993a): Incorporating inspection decisions in pavement management, Transpn Res. B, 27B(6):425-438.

Madanat S (1993b): Optimal infrastructure management decisions under uncertainty, Transpn Res. C, 1(1):77-88.

Madanat S and Ben-Akiva M (1994): Optimal inspection and repair policies for infrastructure facilities, Transportation Science, 28(1):55-62.

Mamlouk MS, Zaniewski JP and He W (2000): Analysis and design optimization of flexible pavement, Journal of Transportation Engineering 126(2):161-167.
Manne AS (1960): Linear programming and sequential decisions, Management Science 6(3):259-267.
Meyer CD (2000): Matrix analysis and applied linear algebra, Siam, Philadelphia, PA.
Minoux M (1986): Mathematical programming: Theory and algorithms, Wiley, Chichester.
Nazareth JL and Kulkarni RB (1986): Linear programming formulations of Markov decision processes, Operations Research Letters 5(1):13-16.

Nazareth JL (2000): The $D_{L} P$ decision support system and its extension to stochastic programming, Optimization Meth. \& Soft., 13:117-154.

Nesbitt DM, Sparks GA and Neudorf RD (1993): A semi-Markov formulation of the pavement maintenance optimization problem, Can. J. Civ. Eng. 20:436-447.
Odermatt N (2001): Restids- och fordonskostnader, unpublished PM, VV, Borlänge.
Olsonen R (1988): Finland's pavement management system: data and models for condition of roads, transl. of technical rep. TVH/Stk Sarja B № 1/1988, Roads and Waterways Administration, Finland.
Ortega JM \& Rheinboldt WC (1970): Iterative solution of nonlinear equations in several variables, Academic Press, New York and London.

Ortega JM (1987): Matrix theory, Plenum Press, New York.
Paterson WDO (1987): Road deterioration and maintenance effects - models for planning and management, John Hopkins University Press, Baltimore.

Philips DT and Lytton RL (1977): A resource constrained capital budgeting model for state highway rehabilitation and maintenance, Texas Engineering Experiment station, Tech. Bull. 77-3:29-40.

Polyak BT (1966): A general method for solving extremum problems, Soviet Mathematics, 8:593-597.
Potucek J and Rydén C-G (2003): Den goda resan - förslag till nationell plan för vägtransportsystemet 2004-2015, publ. 2003:99, VV, Borlänge.
Puterman ML (1994), (2004): Markov decision processes - discrete stochastic dynamic programming, Wiley, New York.
Ravirala V and Grivas DA (1996): Project-level evaluation of highway maintenance using state increment dynamic programming, Civil Eng. Syst. 13:223-238.

Ritchie SG, Asce AM, Yeh C-I, Mahoney JP and Jackson NC (1987): Surface condition expert system for pavement rehabilitation planning, Journal of Transportation Engineering 113(2):155-167.

Ritchie SG (1987): Expert systems in pavement management, Transpn Res. A 21A(2):145-152.
Robinson R, Danielson U and Snaith M (1998): Road maintenance management - concepts and systems, Macmillan Press, Basingstoke and London.

Ross SM (1970): Applied probability models with optimization applications, Holden-Day, San Fransisco.

Ross KW (1989): Randomized and past-dependent policies for Markov Decision Processes with multiple constraints, Operations Research 37(3):474-477.
Ross KW \& Varadarajan R (1989): Markov Decision Processes with sample path constraints: the communicating case, Operations Research 37(5):780-790.

## 10 References

Schweitzer PJ (1971): Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming, Journal of Mathematical Analysis and Applications 34:495-501.
Sikow C, Tikka KA and Äijö JS (1994): Strategic tools in Finland, Proc. $3^{\text {rd }}$ international conference on managing pavements, 1:203-210, National Academy Press, Washington DC.

Smadi OG and Maze TH (1994): Network pavement management system using dynamic programming: application to Iowa State interstate network, Proc. $3^{\text {rd }}$ international conference on managing pavements, 1:195-203, National Academy Press, Washington DC.

Smilowitz KS and Madanat S (2000): Optimal inspection and maintenance policies for infrastructure networks, Computer-Aided Civil and Infrastructure Engineering 15:5-13.

Sundin S and Braban-Ledoux C (2001): Artificial intelligence-based decision support technologies in pavement management, Computer-aided civil and infrastructure engineering, 16(2):143-157.

Tapio R and Piirainen A (1994): Performance indicators in product-based management in Finnish national road administration, Proc. $3^{\text {rd }}$ international conference on managing pavements, 1:134-141, National Academy Press, Washington DC.

Tarricone P (1993): So much pavement, so little time, Civil engineering 63(11):40-43.
Thomas F (2001): A Bayesian approach to retrospective detection of change-points in road surface measurements, Diss. Stockholm University.

Thompson PD, Olsonen R, Talvitie A and Tapio R (1989): A micro-computer Markov model for optimal pavement rehabilitation policy, Proc. World conference of transportation research (WCTR), Yokohama, Japan.

Thompson PD (1994): Making optimization practical in pavement management systems: lessons from leading-edge projects, $3^{\text {rd }}$ international conference on managing pavements, 2:184-189, National Academy Press, Washington DC.

Toyoda Y (1978): A simplified algorithm for obtaining approximate solutions of 0-1 programming problems, Management Science 21:1417-1426.

Tsunokawa K and Schofer JL (1994): Trend curve optimal control model for highway pavement maintenance: case study and evaluation, Transpn Res. A, 28A(2):151-166.

Wallman C-G, Björckebaum F and Yngvesson T (1995): System för beläggningsunderhåll (PMS) - en litteratursammanställning, Technical report 22-1995, VTI, Linköping.

Wang KCP, Zaniewski J and Delton J (1994): Design of project selection procedure based on expert systems and network optimization, $3^{\text {rd }}$ international conference on managing pavements, 2:173-183, National Academy Press, Washington DC.

Wang KCP and Zaniewski JP (1996): 20/30 hindsight: the new pavement optimization in the Arizona state highway network, Interfaces 26(3):77-89.

Wang H (2002): A survey of maintenance policies of deteriorating systems, European Journal of Operational Research 139:469-489.

Watanatada T, Harral CG, Paterson WDO, Dhareswar AM, Dhandari A, Tsunokawa K (1987): HDMIII - the Highway Design and Maintenance Standards Model, John Hopkins Univ. Press for the World Bank, Baltimore.

White DJ (1972): Dynamic Programming and Probabilistic Constraints, Operations Research 22(5): 654-664.

White DJ (1985): Real applications of Markov decision processes, Interfaces 15(6):73-83.

White DJ (1993a): A survey of applications of Markov decision processes, J Opl Res. Soc. 44(11):10731096.

White DJ (1993b): Markov decision processes, Wiley, Chichester.
Virtala P (1996): Analysing alternative maintenance strategies of Swedish paved road network using the network level pavement management system (HIPS), Finnish National Road Administration, Road and Traffic Research.

Worm JM and van Harten A (1996): Model based decision support for planning of road maintenance, Reliability Engineering and System Safety 51:305-316.

Wågberg L-G (2001): Utveckling av nedbrytningsmodeller - sprickinitiering och sprickpropagering, VTI meddelande 916, VTI, Linköping.
Vägverket (1997): Vägverkets samhällsekonomiska kalkylmodell - ekonomisk teori och värderingar, publ 1997:130.

Yang JD, Lu JJ, Gunaratne M and Xiang QJ (2003): Forecasting overall pavement condition with neural networks - application on Florida highway network, Transportation Research Record 1853:3-12.

Äijö J, Männistö V, Thompson P and Tapio R (1990): Minimum of social cost as a guideline of Finnish pavement management, pres. PTCR Summer congress, Brighton, and technical report, Viasys Oy and Road and Waterways Administration, Finland.

Äijö J (1995): Översiktlig vägnätsanalys I Sverige - test av system, Technical report, Svenska vägverket, Finska vägverket and 100 Gen Ltd.

Öberg G (2001): Statligt belagda vägar - tillståndet på vägytan och i vägkroppen, effekter och kostnader, VTI notat 44-2001, VTI, Linköping.

## A1 Road user costs

The VV principle is to clearly distinguish between the technical effect models and the economic valuation of these effects, in terms of unit time costs and tax factors. IRI is the only state variable occurring in these traffic effect models. The models are simplifications of the HDM-4 model system, see [HDM-4 (2000), vol 4, part E]. The predecessor model system HDM-III is described in [Watanatada T et al (1987)].

## A1.1 Travel time costs

In the travel time model, see [Effektsamband 2000 (2001b), p 84] and [Odermatt (2001), p 1], VV differs between passenger cars (index value $p$ ) and lorries/trucks ( $l$ ). A model summary is also found in [Ihs and Sjögren (2003)]. The average speeds $\bar{v}_{p}, \bar{v}_{l}$ (in km/h $=\mathrm{kph}$ ) are determined by the posted speed limit $H G(\mathrm{~km} / \mathrm{h}$ ), the speed law enforcement factor $L F$ (default $=1.0$ ) and the average $I R I$-value, as

$$
\begin{aligned}
& \bar{v}_{p}=\frac{3.6}{\left[\left(\frac{3.6}{H G \cdot L F}\right)^{\frac{1}{0.151}}+\left(\frac{1.5 \cdot I R I}{203}\right)^{\frac{1}{0.151}}\right]^{0.151}} \\
& \bar{v}_{l}=\frac{3.6}{\left[\left(\frac{3.6}{H G \cdot L F}\right)^{\frac{1}{0.11}}+\left(\frac{1.5 \cdot I R I}{180}\right)^{\frac{1}{0.11}}\right]^{0.11}}
\end{aligned}
$$

For $I R I=0$ the average speeds turn into the ideal value $\bar{v}_{p}=\bar{v}_{l}=H G \cdot L F$. Hence the time losses (in $\mathrm{h} / \mathrm{km}$ ) ascribable to the non-ideal pavement state are

$$
\Delta t_{i}=\frac{1}{\bar{v}_{i}}-\frac{1}{H G \cdot L F} \text { for } i=p, l
$$

For given traffic volumes $A A D T_{i}$ for $i=p, l$ (in veh/day) and unit time costs $K_{p}=120$, $K_{l}=150(\mathrm{SEK} / \mathrm{veh} \cdot \mathrm{h})$, the total travel time cost (in SEK/day $\cdot \mathrm{km}$ ) becomes

$$
K_{p} \cdot \Delta t_{p} \cdot A A D T_{p}+K_{l} \cdot \Delta t_{l} \cdot A A D T_{l} .
$$

( $A A D T_{l}$ has been denoted AADTHeavy in the chapters above.)

## A1.2 Vehicle operating costs

In this model, see [Effektsamband 2000 (2001b), p 91 ] and [Odermatt (2001), p 3], VV distinguishes two - three vehicle types: passenger cars (index value $p$ ) and lorries/trucks ( $l$ ), articulated ( $l l$ ) and not ( $l 0$ ). For given traffic volumes $A A D T_{i}$ for $i=p, l$ (in veh/day) and computed operating costs $V O C_{i}$ for $i=p, l 0, l 1$ (in SEK/1000veh $\cdot \mathrm{km}$ ), the total vehicle operating cost (in SEK/day $\cdot \mathrm{km}$ ) is
$V O C_{p} \cdot A A D T_{p}+\frac{1}{2} \cdot\left(V O C_{l 0}+V O C_{l 1}\right) \cdot A A D T_{l}$.

Here the vehicle type specific costs $V O C_{i}$ have contributions from fuel consumption, tyre consumption, parts consumption, labour hours and depreciation (capital cost). The corresponding cost relationships are

$$
V O C_{i}=F C_{i}+T C_{i}+P C_{i}+L C_{i}+C C_{i} \quad i=p, l 0, l 1
$$

Each of the RHS terms are described in the sub-sections below.
The common principle is to calculate the incremental costs, in comparison with an ideal state.

## A1.2.1 Fuel consumption

The fuel consumption model is based on (hidden) uphill and downhill submodels for power requirement, fuel-to-power efficiency and instantaneous fuel consumption. The average speeds $\bar{v}_{p}$ and $\bar{v}_{l 0}=\bar{v}_{l 1}=\bar{v}_{l}(\mathrm{in} \mathrm{m} / \mathrm{s})$ are computed as in App 1.1. For each vehicle type the coefficients in
Tab A1.1 below and the average $I R I$-value will determine the rolling resistance to motion $F R$, the specific fuel consumption (litre/1000veh $\cdot \mathrm{km}$ ) $S F C$ and the fuel cost $F C$ (SEK/1000veh•km) as

$$
\begin{aligned}
& F R_{i}=a_{i}+b_{i} \cdot I R I+c_{i} \cdot \bar{v}_{i}^{2}+d_{i} \cdot I R I \cdot \bar{v}_{i}^{2} \\
& S F C_{i}=e_{i} \cdot F R_{i}+f_{i} \cdot F R_{i}^{2} \cdot \bar{v}_{i} \quad \text { for } i=p, l 0, l 1 . \\
& F C_{i}=\left(S F C_{i}-S F C_{i}^{0}\right) \cdot F P_{i}
\end{aligned}
$$

Here $S F C^{0}$ denotes the specific fuel consumption for the ideal state $I R I=0$, and $F P$ is the fuel price (SEK/litre).

| Veh.typelCoeff | $a$ | $b$ | $c$ | $d$ | $e$ | $f$ | $F P$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $p$ | 552.253528 | 5.198864 | 0.13719467 | 0.00319733 | 0.07444444 | $2.95414 \mathrm{e}-07$ | 2.8 |
| 10 | 2082.91539 | 27.0555129 | 0.0832849 | 0.00277616 | 0.06627906 | $7.70686 \mathrm{e}-08$ | 1.88 |
| 11 | 6156.5253 | 95.41751 | 0.24985469 | 0.00832849 | 0.06395348 | $2.47882 \mathrm{e}-08$ | 1.88 |

Table A1.1 Parameters in the fuel consumption model. The $e, f$-values are doubled in comparison with [Odermatt (2001), p 7], after discussion with [Lang (2003)].

## A1.2.2 Tyre consumption

The tyre consumption model is based on energy requirements (cf. Sec 1.2.1) and forces acting on the wheels. For each vehicle type the coefficients in Tab A1.2 below and the average IRIvalue will determine the number of (possible) retreads per tyre carcass $N R$, the number of tyres consumed per 1000 veh $\cdot \mathrm{km} S T C$, and the tyre cost $T C$ as

$$
\begin{gathered}
N R_{i}=\max \left[0, m_{i} \cdot e^{-0.03224 \cdot I R I}-1\right] \\
S T C_{i}=n_{i} \cdot\left(\frac{\left(h_{i}+j_{i} \cdot F R_{i}^{2}\right)\left(1+k_{i} \cdot N R_{i}\right)}{l_{i} \cdot\left(1+N R_{i}\right)}+0.0027\right) \text { for } i=p, l 0, l 1 . \\
T C_{i}=\left(S T C_{i}-S T C_{i}^{0}\right) \cdot T P_{i}
\end{gathered}
$$

Here $F R$ is the rolling resistance to motion, computed in App 1.2.1, $S T C^{0}$ denotes the specific
tyre consumption for $I R I=0$, and $T P$ is the tyre price (SEK/tyre).
Observe that $p$ in Tab A1.2 both denotes a vehicle type index and a coefficient type.

| Veh.typelCoeff | $h$ | $j$ | $k$ | 1 | $m$ | $n$ | $T P$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $p$ | 0.02616 | $4.33231 \mathrm{e}-08$ | 0.15 | 1.4 | 1.3 | 0.8 | 500 |
| 10 | 0.02585 | $4.55318 \mathrm{e}-09$ | 0.15 | 6 | 1.3 | 3 | 3400 |
| 11 | 0.03988 | $6.29015 \mathrm{e}-10$ | 0.15 | 8 | 1.3 | 9 | 3400 |

Table A1.2 Parameters in the tyre consumption model.

## A1.2.3 Parts consumption

The parts consumption model considers the effects of vehicle age (in km), road roughness and speed-change cycles. Here an adjusted roughness measure $R I_{a d j}(\mathrm{~mm} / \mathrm{m})$ is used:

$$
R I_{a d j}=\max \left[I R I, \min \left(3.25,3+5.54 \cdot 10^{-8} \cdot I R I^{13}\right)\right]
$$

For each vehicle type the parts consumption per $1000 \mathrm{veh} \cdot \mathrm{km}, S P C$, is first expressed as a fraction, to be taken of the average new (or replacement) vehicle price $N P$ (in SEK), and then the full parts cost $P C$ is computed (in SEK/1000veh $\cdot \mathrm{km}$ ), by the use of the coefficients in Tab A1.3. $S P C^{0}$ denotes the specific parts consumption for $I R I=0$.

$$
\begin{aligned}
& S P C_{i}=0.4 \cdot\left(o_{i}^{p_{i}} \cdot\left(q_{i} \cdot 10^{-6}+r_{i} \cdot 10^{-6} \cdot R I_{a d j}\right)\right) \text { for } i=p, l 0, l 1 . \\
& P C_{i}=\left(S P C_{i}-S P C_{i}^{0}\right) \cdot N P_{i}
\end{aligned}
$$

| Veh.typelCoeff | 0 | $p$ | $q$ | $r$ | $N P$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
| $p$ | 115000 | 0.308 | 36.94 | 6.2 | 162000 |
| 10 | 240000 | 0.371 | 11.58 | 2.96 | 922000 |
| 11 | 602000 | 0.371 | 13.58 | 2.96 | 1957000 |

Table A1.3 Parameters in the parts consumption model.

## A1.2.4 Labour hours

The labour hours are for fitting spare parts and repairing vehicles. The model relies on the parts consumption SPC in App 1.2.3, the labour wage rates $L P$ (in SEK/h) and the coefficients in Tab A1.4 below. The maintenance labour hours $L H$ per $1000 \mathrm{veh} \cdot \mathrm{km}$ and the labour costs $L C$ (in SEK/1000veh $\cdot \mathrm{km}$ ) are determined per vehicle type. $L H^{0}$ denotes the labour hours for the ideal state $I R I=0$.

$$
\begin{gathered}
L H_{i}=s_{i} \cdot S P C_{i}^{t_{i}} \\
L C_{i}=\left(L H_{i}-L H_{i}^{0}\right) \cdot L P_{i}
\end{gathered} \text { for } i=p, l 0, l 1 .
$$

| Veh.typelCoeff | $s$ | $t$ | $L P$ |
| ---: | ---: | ---: | ---: |
| $p$ | 77.14 | 0.547 | 190 |
| 10 | 242.03 | 0.519 | 190 |
| 11 | 652.51 | 0.519 | 190 |

Table A1.4 Parameters in the labour hours model. The LP-values are taken from [Effektsamband 2000 (2001a), p 115], for passengers on work-purpose journey, after discussion with [Lang (2003)].

## A1.2.5 Capital costs

In general, capital costs comprises depreciation and interest costs. Here only the depreciation costs are considered. At first the depreciation cost factor $D E P$ is determined for each vehicle type per $1000 \mathrm{veh} \cdot \mathrm{km}$, as a fraction of the average new (or replacement) vehicle price $N P$ (in Tab A1.3 above), by the use of the coefficients in Tab A1.5 below and the adjusted roughness $R I_{a d j}$ in App 1.2.3. Finally the capital costs $C C$ are determined. $D E P^{0}$ denotes the depreciation for $I R I=0$.

$$
\begin{gathered}
D E P_{i}=\frac{1000 \cdot(1-0.01 \cdot \max [2,15-\max (0, I R I-5)]) \cdot\left(1+\exp \left(-65.8553 \cdot R I_{a d j}^{-1.9194}\right)\right)}{u_{i} \cdot w_{i}}, \\
C C_{i}=\left(D E P_{i}-D E P_{i}^{0}\right) \cdot N P_{i} \quad \text { for } i=p, l 0, l 1 .
\end{gathered}
$$

| Veh.typelCoeff | $u$ | $w$ |
| ---: | ---: | ---: |
| $p$ | 23000 | 10 |
| 10 | 40000 | 12 |
| 11 | 86000 | 14 |

Table A1.5 Parameters in the depreciation model.

