Dynamic Scaling and Submodel Selection in Bundle Methods for Convex Optimization

Christoph Helmberg, Alois Pichler

Preprint 2017-4

Preprintreihe der Fakultät für Mathematik
ISSN 1614-8835

Impressum:
Herausgeber:
Der Dekan der
Fakultät für Mathematik
an der Technischen Universität Chemnitz
Sitz:
Reichenhainer Straße 39
09126 Chemnitz
Postanschrift:
09107 Chemnitz
Telefon: (0371) 531-22000
Telefax: (0371) 531-22009
E-Mail: dekanat@mathematik.tu-chemnitz.de
Internet:
http://www.tu-chemnitz.de/mathematik/
ISSN 1614-8835 (Print)

Dynamic Scaling and Submodel Selection in Bundle Methods for Convex Optimization

Christoph Helmberg* Alois Pichler*

August 25, 2017

Abstract

Bundle methods determine the next candidate point as the minimizer of a cutting model augmented with a proximal term. We propose a dynamic approach for choosing a quadratic proximal term based on subgradient information from past evaluations. For the special case of convex quadratic functions, conditions are studied under which this actually reproduces the Hessian. The approach forms the basis of an efficiently implementable variant that uses only the diagonal as dynamic scaling information. The second topic addresses the choice of the cutting model when minimizing the sum of several convex functions. We propose a simple rule for dynamically choosing a few functions that are each modeled by a separate cutting model while the others are subsumed in a common sum model and combine this with the scaling approach. Numerical experiments with a development version of the callable library ConicBundle illustrate the benefits of these techniques on a class of large scale instances of practical relevance.

Keywords: nonsmooth optimization, numerical methods
MSC 2010: 90C25; 90C06, 65K05

1 Introduction

Bundle methods are a basic algorithmic approach for minimizing nonsmooth convex functions. A thorough introduction to these methods and convex optimization is given in [1, 9], for a more recent survey see [2]. The following concise introduction focuses on basic concepts and notation needed in the context of this work.

Throughout, $f\colon \mathbb{R}^n \to \mathbb{R}$ will denote a closed proper convex function. For a point $x \in \mathbb{R}^n$ the function value $f(x)$ and any subgradient $g \in \partial f(x)$ in the subdifferential of f in x give rise to a global affine minorant of f via the subgradient inequality,

$$f(y) \geq f(x) + g^\top (y - x) \qquad \text{for all } y \in \mathbb{R}^n.$$

It will be convenient to collect the oracle information in a parameter $\omega = (\gamma, g) = (f(x) - g^\top x, g)$ and to denote the corresponding minorant by

$$f_\omega(y) = \gamma + g^\top y. \tag{1}$$

*Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany, {helmberg,alois.pichler}@mathematik.tu-chemnitz.de, www.mathematik.tu-chemnitz.de/{∼helmberg,∼alopi}

For brevity, f_ω is referred to as the minorant ω. The set of all affine minorants $W_f = \{(\gamma, g)\colon f(y) \geq \gamma + g^\top y \text{ for all } y \in \mathbb{R}^n\}$ is convex and, because f is closed, fully describes f via

$$f(y) = \sup_{\omega \in W_f} f_\omega(y) \qquad \text{for all } y \in \mathbb{R}^n.$$

Any subset $\widehat{W}_f \subseteq W_f$ yields a *cutting model* minorizing f,

$$f(y) \geq f_{\widehat{W}_f}(y) := \sup_{\omega \in \widehat{W}_f} f_\omega(y) \qquad \text{for all } y \in \mathbb{R}^n.$$

In iteration $k \in \mathbb{N}$ a bundle method determines for a given compact (often finite) model $\widehat{W}_f^k \subseteq W_f$ and a *center of stability* $\hat{y}^k$ a next *candidate* $\bar{y}^k$ by solving a *bundle subproblem*

$$\bar{y}^k = \operatorname*{argmin}_{y \in \mathbb{R}^n} f_{\widehat{W}_f^k}(y) + \frac{1}{2}\|y - \hat{y}^k\|_H^2,$$

where the quadratic (or proximal) term $\|y - \hat{y}^k\|_H^2 = (y - \hat{y}^k)^\top H (y - \hat{y}^k)$ with positive definite *scaling* matrix H serves as a kind of proximity or trust region control. Existence and uniqueness of the minimizer follow from strong convexity of the proximal term and compactness of $\widehat{W}_f^k$, because these ensure the existence of a unique saddle point $(\bar{y}^k, \bar{\omega}^k) \in \mathbb{R}^n \times \operatorname{conv} \widehat{W}_f^k$ to the Lagrangian

$$L(y, \omega{=}(\gamma, g)) \;=\; \gamma + g^\top y + \frac{1}{2}\|y - \hat{y}^k\|_H^2.$$

The saddle point minorant $\bar{\omega}^k = (\bar{\gamma}^k, \bar{g}^k) \in \operatorname{conv} \widehat{W}_f^k$ is called the *aggregate* and by optimality $\bar{y}^k = \hat{y}^k - H^{-1}\bar{g}^k$. The function is then evaluated at $\bar{y}^k$ by a first order oracle which returns $f(\bar{y}^k)$ and some subgradient $g^k \in \partial f(\bar{y}^k)$ giving the candidate minorant $\omega^k = (\gamma^k = f(\bar{y}^k) - (g^k)^\top \bar{y}^k, g^k) \in W_f$. If the *descent test*

$$f(\hat{y}^k) - f(\bar{y}^k) \geq \kappa[f(\hat{y}^k) - f_{\bar{\omega}^k}(\bar{y}^k)]$$

is successful for some given fixed $\kappa \in (0, 1)$, the method performs a *descent step* by moving the center of stability $\hat{y}^{k+1} = \bar{y}^k$ to the candidate. Otherwise, in a *null step*, the center is preserved, $\hat{y}^{k+1} = \hat{y}^k$, but the next model is required to satisfy $\{\bar{\omega}^k, \omega^k\} \subseteq \widehat{W}_f^{k+1}$ in order to improve the quality of the model near $\bar{y}^k$.

This framework suffices to ensure $f(\hat{y}^k) \to f^* := \inf_{y \in \mathbb{R}^n} f(y)$ for $k \to \infty$ and, if minimizers of f exist, the sequence $\hat{y}^k$ converges to some minimizer [1, Section 10.3.4]. Furthermore, if f is bounded from below, the norm of the gradients of the aggregates of a suitable subsequence converges to zero, $\liminf \|\bar{g}^k\| = 0$.

The same still holds if H of the quadratic term or the proximal term is dynamically updated with some care, see [1].

Practical efficiency strongly depends on the choice of the next model $\widehat{W}_f^{k+1}$ and the choice of H. In this work we propose a dynamic approach for choosing H and, in the case of f being a sum of convex functions, a strategy for forming the cutting model by separate submodels for a dynamically selected subset of the functions and a common model for the remaining ones.

Outline. The dynamic method for determining a suitable H will be introduced in Section 2. It attempts to exploit the nonsmooth subgradient information

independent of whether it is employed for the cutting model or not. In essence, the method constructs a quadratic term so that its addition to the aggregate ensures that none of the selected subgradient inequalities are violated by more than some appropriately chosen $\varepsilon^k > 0$. In the smooth case the result should be related to the Hessian and we will investigate this relation. In contrast to [10, 12], we do not aim for second order convergence but hope to contribute a practically efficient heuristic for choosing H in large scale problems. Numerical experiments will demonstrate the benefits of the dynamic scaling approach for a suitably simplified version on a class of large scale instances of practical relevance.

In these instances the function $f(y) = \sum_{i \in R} f_i(y)$ is a sum of proper closed convex functions f_i, $i \in R$, for a finite index set R. This situation arises frequently in Lagrangian relaxation when relaxing coupling constraints for example in scenario based stochastic optimization or scheduling with shared resources. On the implementational side we therefore also explore possibilities to exploit this structure in forming $\widehat{W}_f$. A subset $J^k \subseteq R$ of indices will be selected dynamically for which separate models $\widehat{W}_i^k$, $i \in J^k$, are used while a common *summodel* $\widehat{W}_{\bar{J}^k}^k$ is employed for the indices $i \in \bar{J}^k = R \setminus J^k$ in the complement. This approach is laid out in detail in Section 3. During the preparation of this work the idea of dynamically selecting submodels was independently mentioned as a possibility in [14].

Section 4 is devoted to further relevant aspects in implementing submodel selection and dynamic scaling. Numerical experiments exploring the use of the new techniques on instances of practical relevance will be given in Section 5. Section 6 concludes with some perspectives on further work.

Notation. The set of (real) symmetric matrices of order n will be denoted by S^n, the subset of positive semidefinite matrices by S^n_+. For $A, B \in S^n$ the Loewner partial order $A \succeq B$ ($A \succ B$) refers to $A - B \in S^n_+$ ($A - B$ positive definite). For $A, B \in \mathbb{R}^{m \times n}$ the standard (trace) inner product is $\langle A, B \rangle = \operatorname{tr} B^\top A = \sum_{i=1}^n (B^\top A)_{ii}$. For vectors $a, b \in \mathbb{R}^n$ this results in the canonical inner product $\langle a, b \rangle = a^\top b$. For the inner product associated with a positive definite H we write $\langle a, b \rangle_H := \langle a, Hb \rangle = b^\top H a$ and the associated norm reads $\|a\|_H = \langle a, a \rangle_H^{\frac{1}{2}}$, where $\|a\| = \|a\|_{I_n}$ is the usual Euclidean norm. For a matrix $A \in \mathbb{R}^{m \times n}$, $\|A\|_F = \langle A, A \rangle^{\frac{1}{2}}$ denotes the Frobenius norm, $\mathcal{R}(A)$ the range space and $\mathcal{N}(A)$ the null space. If vectors or matrices carry subscripts already like in $s_i \in \mathbb{R}^n$ or if they result from arithmetic expressions, a coordinate j is extracted by subscripting the bracketed expression $[s_i]_j$.

2 Dynamic Scaling

Bundle methods offer a lot of freedom in the choice of H for the proximal term. Having executed iteration k of the bundle method, it suffices to observe two simple rules in the selection of the next positive definite H^{k+1} so that validity of the convergence proofs is preserved, see [1]. For the purpose of stating them, fix $0 < \underline{\lambda} < \bar{\lambda} \in \mathbb{R}$. If iteration k gives rise to a descent step, one may choose any H^{k+1} satisfying $\underline{\lambda} I \preceq H^{k+1} \preceq \bar{\lambda} I$; if it results in a null step, it suffices to ensure $H^k \preceq H^{k+1} \preceq \bar{\lambda} I$. Algorithmically this will be easy to guarantee and we will refrain from stating these aspects explicitly.

Like in [7], for $k \in \mathbb{N}$ our choice will be of the form $H^k = u^k I + \bar{H}^k$ where $u^k I$

with $u^k \in [\underline{\lambda}, \bar{\lambda}]$ mimics a trust region and $0 \preceq \bar{H}^k \preceq \bar{\lambda} I$ should reflect the local geometry. Given the starting point $\hat{y}^0$ with evaluation $f(\hat{y}^0)$ and $\omega^0 = (\gamma^0, g^0)$, we start with $u^1 = \max\{1, \|g^0\|\}$ and $\bar{H}^1 = 0$ and then update u as in [6]. The main focus here is the choice of (or update to) the next $\bar{H}^{k+1}$ whenever a descent step occurred in iteration k.

As in any adaptive approach the main intention is to scale the space according to the local geometry of f at the next center of stability $\hat{y}^{k+1} = \bar{y}^k$. In the smooth case the appropriate choice would be the Hessian. In the nonsmooth setting considered here, the information revealed about f after iteration k is the piecewise linear max-function $\max_{i\in\{0,\ldots,k\}} f_{\omega^k}(\cdot) \leq f(\cdot)$. In order to keep the bundle subproblem manageable, $\widehat{W}_f^{k+1}$ will typically contain only a few of the subgradients returned by the oracle, but it may be possible to include more information in forming H.

In a descent step the aggregate $\bar{\omega}^k = (\bar{\gamma}^k, \bar{g}^k)$ gives rise to the new center $\hat{y}^{k+1} = \hat{y}^k - (H^k)^{-1}\bar{g}^k$, so we may consider $\bar{\omega}^k$ as a good model of f at $\hat{y}^{k+1}$. The aggregate is unlikely to touch the epigraph of f, because it is a convex combination of tangent hyperplanes to this epigraph. There is, however, an unknown $\epsilon \geq 0$ so that $(\bar{\gamma}^k + \varepsilon, \bar{g}^k)$ gives rise to a minorant of f with $\bar{g}^k \in \partial f(\tilde{y})$ in a point $\tilde{y} \in \operatorname{dom} f$ and $0 \leq \varepsilon = f(\tilde{y}) - \bar{\gamma}^k - (\bar{g}^k)^\top \tilde{y} \leq f(\hat{y}^{k+1}) - \bar{\gamma}^k - (\bar{g}^k)^\top \hat{y}^{k+1}$. It will not be important to determine this ε precisely. In practice we use

$$\varepsilon^k = f(\hat{y}^{k+1}) - \bar{\gamma}^k - (\bar{g}^k)^\top \hat{y}^{k+1}.$$

In our basic quadratic model $q(\cdot) = f_{\bar{\omega}^k}(\cdot) + \frac{1}{2}\|\cdot - \hat{y}^{k+1}\|^2_{H^{k+1}}$ of f at $\hat{y}^{k+1}$, we would like to design $\bar{H}^{k+1}$ so that all minorants (γ^j, g^j) that are inactive in the sense of $\gamma^j + (g^j)^\top \hat{y}^{k+1} < \bar{\gamma}^k + (\bar{g}^k)^\top \hat{y}^{k+1} + \varepsilon^k$ are "violated" by $q(\cdot)$ by at most ε^k. We call ε^k the *violation parameter*.

Lemma 1 *Given $\hat{y} \in \mathbb{R}^n$, violation parameter $\varepsilon > 0$ and $\bar{\omega} = (\bar{\gamma}, \bar{g}), \omega = (\gamma, g) \in \mathbb{R} \times \mathbb{R}^n$ with $\gamma + g^\top \hat{y} < \bar{\gamma} + \bar{g}^\top \hat{y} + \varepsilon$, a matrix $\bar{H} \in S^n_+$ ensures*

$$f_{\bar{\omega}}(y) + \frac{1}{2}(y - \hat{y})^\top \bar{H}(y - \hat{y}) \geq f_\omega(y) - \varepsilon \quad \text{for all } y \in \mathbb{R}^n \tag{2}$$

if and only if

$$\bar{H} \succeq \frac{1}{2}\frac{(\bar{g} - g)(\bar{g} - g)^\top}{\bar{\gamma} + \varepsilon - \gamma + (\bar{g} - g)^\top \hat{y}}. \tag{3}$$

Proof. Putting $\Delta y = y - \hat{y}$, $d = \bar{g} - g$, $\delta = \bar{\gamma} + \varepsilon - \gamma + (\bar{g} - g)^\top \hat{y} > 0$, and using the definition (1) of f_ω, eq. (2) computes to

$$\frac{1}{2}\Delta y^\top \bar{H} \Delta y + d^\top \Delta y + \delta \geq 0 \quad \text{for all } \Delta y \in \mathbb{R}^n.$$

Eq. (3) implies (2), because for all $\Delta y \in \mathbb{R}^n$

$$\tfrac{1}{2}\Delta y^\top \bar{H} \Delta y + d^\top \Delta y + \delta \overset{(3)}{\geq} \tfrac{1}{2}\Delta y^\top \tfrac{1}{2\delta} d d^\top \Delta y + d^\top \Delta y + \delta = \tfrac{1}{4\delta}(d^\top \Delta y + 2\delta)^2 \geq 0.$$

If (3) does not hold, then by $2\delta > 0$ and the Schur complement inequality, the matrix $Z = \begin{bmatrix} \bar{H} & d \\ d^\top & 2\delta \end{bmatrix}$ is not positive semidefinite. Hence, there is a vector $v \in \mathbb{R}^{n+1}$ with $v^\top Z v < 0$. Putting $\bar{v} = (v_1, \ldots, v_n)^\top$ and $\nu = v_{n+1}$ this reads

$$\bar{v}^\top \bar{H} \bar{v} + 2\nu \bar{v}^\top d + 2\delta\nu^2 < 0.$$

By $\bar{H} \succeq 0$ we get $\nu \neq 0$. Divide by $2\nu^2 > 0$ and put $\overline{\Delta y} = \frac{1}{\nu}\bar{v}$ to obtain

$$\frac{1}{2}\overline{\Delta y}^\top \bar{H}\overline{\Delta y} + d^\top \overline{\Delta y} + \delta < 0.$$

Thus, (2) does not hold. ■

In a "best" quadratic model, $\bar{H}$ would have minimal curvature in a sense yet to be specified with (2) holding for all relevant ω^i. If the curvature is measured by the trace inner product of $\bar{H}$ with some positive definite C (*e. g.*, for $C = I$ this is the sum of the eigenvalues of $\bar{H}$), $\bar{H}$ may in theory be determined by a semidefinite program. It is instructive to study this possibility, but for practical efficiency we will propose a simpler approach later.

Theorem 2 *Given $\hat{y} \in \mathbb{R}^n$, violation parameter $\varepsilon > 0$, $\bar{\omega} = (\bar{\gamma}, \bar{g}), \omega^i = (\gamma^i, g^i) \in \mathbb{R} \times \mathbb{R}^n$ with $\gamma^i + (g^i)^\top \hat{y} < \bar{\gamma} + \bar{g}^\top \hat{y} + \varepsilon$ for $i = 1, \ldots, k$ and a positive definite $C \in S^n$, let $\bar{H}$ be an optimal solution of the semidefinite program*

$$\begin{array}{ll} \textit{minimize} & \langle C, H\rangle \\ \textit{subject to} & H \succeq \frac{1}{2}\frac{(\bar{g}-g^i)(\bar{g}-g^i)^\top}{\bar{\gamma}+\varepsilon-\gamma^i+(\bar{g}-g^i)^\top\hat{y}}, \quad i = 1, \ldots, k, \\ & H \succeq 0. \end{array} \tag{4}$$

Then

$$f_{\bar{\omega}}(y) + \frac{1}{2}(y-\hat{y})^\top \bar{H}(y-\hat{y}) \geq \max_{i=1,\ldots,k} f_{\omega^i}(y) - \varepsilon \quad \textit{for all } y \in \mathbb{R}^n. \tag{5}$$

Furthermore, $\mathcal{D} := \operatorname{span}\{\bar{g} - g^i \colon i = 1, \ldots, k\} \subseteq \mathcal{R}(\bar{H})$. In case $C = I$, equality holds, $\mathcal{D} = \mathcal{R}(\bar{H})$.

Proof. Program (4) has at least one optimal solution, because for $\mu = \frac{1}{2}\max\{\|\bar{g}-g^i\|^2/(\bar{\gamma}+\varepsilon-\gamma^i+(\bar{g}-g^i)^\top\hat{y}) \colon i = 1, \ldots, k\}$ the matrix $H = \mu I$ is feasible, so the optimal solution lies within the compact set $\{H \succeq 0 \colon \langle C, H\rangle \leq \mu \operatorname{tr} C\}$. Feasibility and Lemma 1 ensure (5).

Suppose, there is an $i \in \{1, \ldots, k\}$ with $\bar{g} - g^i \notin \mathcal{R}(\bar{H})$, then with v the projection of $\bar{g}-g^i$ onto the null space $\mathcal{N}(\bar{H}) = \mathcal{R}(\bar{H})^\perp$, infeasibility of constraint i follows from $v^\top \bar{H} v = 0 < v^\top \frac{1}{2}\frac{(\bar{g}-g^i)(\bar{g}-g^i)^\top}{\bar{\gamma}+\varepsilon-\gamma^i+(\bar{g}-g^i)^\top\hat{y}} v$. In order to see that $\mathcal{R}(\bar{H}) \subseteq \mathcal{D}$ whenever $C = I$, assume $\dim \mathcal{D} = h$ and let $P \in \mathbb{R}^{n\times n}$ be an orthogonal matrix ($P^\top P = PP^\top = I$) with the first h columns of P forming an orthonormal basis of $\mathcal{D}$. Now consider the equivalent scaled semidefinite program with variable $H' = P^\top H P$

$$\begin{array}{ll} \text{minimize} & \langle P^\top I P, H'\rangle \\ \text{subject to} & H' \succeq \frac{1}{2}\frac{P^\top(\bar{g}-g^i)(\bar{g}-g^i)^\top P}{\bar{\gamma}+\varepsilon-\gamma^i+(\bar{g}-g^i)^\top\hat{y}}, \quad i = 1, \ldots, k \\ & H' \succeq 0. \end{array}$$

For each $i = 1, \ldots, k$ the semidefiniteness condition is restricted to the leading principal submatrix on indices $1, \ldots, h$. Thus, for an optimal $\bar{H}'$ there follows $\mathcal{R}(\bar{H} = P\bar{H}'P^\top) \subseteq \mathcal{D}$ because the objective $\langle I, H'\rangle$ ensures that the last $n - h$ rows and columns of $\bar{H}'$ are zero. ■

In fact, the proof shows that the range space of $\bar{H}$ equals $\mathcal{D}$ if $\mathcal{D}$ and $\mathcal{D}^\perp$ are invariant subspaces of C.

In the case of a smooth f the construction should relate in some meaningful way to the Hessian on the subspace explored. For gaining some intuition on this relation we study the special case of a convex quadratic f, assume $\hat{y}$ to correspond well to the aggregate in that $\bar{g} = \nabla f(\hat{y})$ and take ε large enough to include all minorants. In this setting, the next lemma ensures that the Hessian projected onto the subspace spanned by the generated gradients is feasible for (4).

Lemma 3 *Let $f(x) = \frac{1}{2}x^\top Ax + b^\top x + \rho$ with $A \in S^n_+$ and suppose $\hat{y} \in \mathbb{R}^n$, $\bar{\omega} = (\bar{\gamma}, \bar{g} = \nabla f(\hat{y}))$ and violation parameter $\varepsilon \geq 0$ satisfy $\bar{\gamma} + \bar{g}^\top \hat{y} \leq f(\hat{y}) \leq \bar{\gamma} + \bar{g}^\top \hat{y} + \varepsilon$. Given $y^i \in \mathbb{R}^n$ and $\omega^i = (\gamma^i, g^i) = (f(y^i) - \nabla f(y^i)^\top y^i, \nabla f(y^i))$, $i = 1, \ldots, k$, let $P \in \mathbb{R}^{n \times h}, P^\top P = I_h$ have range space $\mathcal{R}(P) = \operatorname{span}\{A(\hat{y} - y^i) \colon i = 1, \ldots, k\}$, then $PP^\top APP^\top$ is feasible for (4).*

Proof. Because $\bar{g} = A\hat{y} + b$ and $g^i = Ay^i + b$, constraint i of (4) takes the form

$$H \succeq \frac{A(\hat{y} - y^i)(\hat{y} - y^i)^\top A}{2(\bar{\gamma} + \epsilon - \gamma^i + (\bar{g} - g^i)^\top \hat{y})}.$$

Whenever $A(\hat{y} - y^i) = 0$ for some $i = 1, \ldots, k$, constraint i is redundant, so assume w.l.o.g. $A(\hat{y} - y^i) \neq 0$ for all i. By the assumptions on $\bar{\omega}$ and γ^i, $i = 1, \ldots, k$,

$$\begin{aligned} \bar{\gamma} + \epsilon - \gamma^i + (\bar{g} - g^i)^\top \hat{y} &\geq f(\hat{y}) - f(y^i) - (Ay^i + b)^\top (\hat{y} - y^i) \\ &= \tfrac{1}{2}\hat{y}^\top A\hat{y} + \tfrac{1}{2}(y^i)^\top Ay^i - \hat{y}^\top Ay^i \\ &= \tfrac{1}{2}(\hat{y} - y^i)^\top A(\hat{y} - y^i). \end{aligned} \tag{6}$$

Therefore

$$A^{\frac{1}{2}} \frac{A^{\frac{1}{2}}(\hat{y} - y^i)}{\|A^{\frac{1}{2}}(\hat{y} - y^i)\|} \frac{[A^{\frac{1}{2}}(\hat{y} - y^i)]^\top}{\|A^{\frac{1}{2}}(\hat{y} - y^i)\|} A^{\frac{1}{2}} \succeq \frac{A(\hat{y} - y^i)(\hat{y} - y^i)^\top A}{2(\bar{\gamma} + \epsilon - \gamma^i + (\bar{g} - g^i)^\top \hat{y})}. \tag{7}$$

Feasibility of $H = A = A^{\frac{1}{2}} I A^{\frac{1}{2}}$ now follows from

$$A - \frac{A(\hat{y} - y^i)(\hat{y} - y^i)^\top A}{(\hat{y} - y^i)^\top A(\hat{y} - y^i)} = A^{\frac{1}{2}} \left(I - \frac{A^{\frac{1}{2}}(\hat{y} - y^i)}{\|A^{\frac{1}{2}}(\hat{y} - y^i)\|} \frac{[A^{\frac{1}{2}}(\hat{y} - y^i)]^\top}{\|A^{\frac{1}{2}}(\hat{y} - y^i)\|} \right) A^{\frac{1}{2}} \succeq 0.$$

By $A(\hat{y} - y^i) = PP^\top A(\hat{y} - y^i)$ we conclude

$$0 \preceq PP^\top \left(A - \frac{A(\hat{y} - y^i)(\hat{y} - y^i)^\top A}{(\hat{y} - y^i)^\top A(\hat{y} - y^i)} \right) PP^\top = PP^\top APP^\top - \frac{A(\hat{y} - y^i)(\hat{y} - y^i)^\top A}{(\hat{y} - y^i)^\top A(\hat{y} - y^i)}.$$

Together with (7) this proves feasibility of the projected Hessian for (4). ■

For a convex quadratic f and an aggregate identical to the linear model in $\hat{y}$, *i. e.*, $(\bar{\gamma}, \bar{g}) = (f(\hat{y}) - \nabla f(\hat{y})^\top \hat{y}, \nabla f(\hat{y}))$, the proof shows that (3) simplifies to

$$H \succeq \frac{A(\hat{y} - y^i)}{\|A^{\frac{1}{2}}(\hat{y} - y^i)\|} \frac{[A(\hat{y} - y^i)]^\top}{\|A^{\frac{1}{2}}(\hat{y} - y^i)\|}.$$

In particular, for $A \succeq 0$ the A-normalized directions $d_i = A\frac{\hat{y} - y_i}{\|\hat{y} - y_i\|_A}$ are of importance while the lengths of the steps $(\hat{y} - y^i)$ are of little relevance.

The following theorem provides some structural properties of optimal solutions of (4). It will help to establish sufficient conditions on which directions have to be included in order to guarantee optimality of the projected Hessian of Lemma 3.

Theorem 4 *Given positive definite $C \in S^n$ and $d_i \in \mathbb{R}^n$, $i = 1, \ldots, k$, the program*

$$\begin{array}{ll} \textit{minimize} & \langle C, H\rangle \\ \textit{subject to} & H \succeq d_i d_i^\top, \quad i = 1, \ldots, k, \\ & H \succeq 0. \end{array} \tag{8}$$

and its dual

$$\begin{array}{ll} \textit{maximize} & \sum_{i=1}^k \left\langle d_i d_i^\top, Z_i \right\rangle \\ \textit{subject to} & \sum_{i=1}^k Z_i \preceq C, \\ & Z_i \succeq 0, i = 1, \ldots, k. \end{array} \tag{9}$$

satisfy strong duality and any primal and dual optimal solutions $\bar{H}$ and $\bar{Z}_i$, $i = 1, \ldots, k$ satisfy the complementarity conditions

$$(C - \sum_{i=1}^k \bar{Z}_i)\bar{H} = 0, \tag{10}$$

$$(\bar{H} - d_i d_i^\top)\bar{Z}_i = 0, \qquad i = 1, \ldots, k. \tag{11}$$

Any such $\bar{H}$ is also optimal for $C' = \sum_{i=1}^k \zeta_i \bar{Z}_i$ with $\zeta_i > 0$, $i = 1, \ldots, k$.

For $C = I$, let $P \in \mathbb{R}^{h \times n}$ have orthonormal columns ($P^\top P = I_h$) spanning the range space $\mathcal{R}(P) = \operatorname{span}\{d_i \colon i = 1, \ldots, k\}$, then

$$\sum_{i=1}^k P^\top \bar{Z}_i P = I_h. \tag{12}$$

Furthermore, $\operatorname{rank} P^\top \bar{Z}_i P \leq 1$, *$i = 1, \ldots, k$, and*

$$\bar{H} = \sum_{i=1}^k d_i d_i^\top \bar{Z}_i P P^\top.$$

In particular, $\bar{H}$ is unique.

Proof. Because $H = I_n \cdot 2\max\{\|d_i\|^2 \colon i = 1, \ldots, k\}$ is strictly primal feasible and $Z_i = C/(2k)$, $i = 1, \ldots, k$ is strictly dual feasible, standard semidefinite duality theory [15] ensures strong duality with attainment on both sides as well as the complementarity conditions (10), (11).

For cost matrix $C' = \sum_{i=1}^k \zeta_i \bar{Z}_i$, dual feasibility holds for $\bar{Z}_i' = \zeta_i \bar{Z}_i \succeq 0$, $i = 1, \ldots, k$, and the primal and dual objective values coincide by the explicit computation $\left\langle C', \bar{H}\right\rangle = \sum_{i=1}^n \zeta_i \left\langle \bar{Z}_i, \bar{H}\right\rangle \overset{(11)}{=} \sum_{i=1}^n \zeta_i \left\langle \bar{Z}_i, d_i d_i^\top\right\rangle$.

For $C = I$, Th. 2 implies that any primal optimal $\bar{H}$ satisfies $\mathcal{R}(\bar{H}) = \mathcal{R}(P)$. In particular $P^\top \bar{H} P \succ 0$, $\bar{H} = PP^\top \bar{H} PP^\top$ and $d_i d_i^\top = PP^\top d_i d_i^\top PP^\top$. Thus, complementarity (10) yields (12). The complementarity relation (11) implies

rank $P^\top \bar{Z}_i P \leq 1$, because any direction $d \in \mathcal{R}(P)$ orthogonal to d_i is not in the null space of $\bar{H} - d_i d_i^\top$. The formula for $\bar{H}$ follows from

$$\begin{aligned}
\bar{H} = PP^\top \bar{H} P I_h P^\top &\overset{(12)}{=} PP^\top \bar{H} P \sum_{i=1}^{k} P^\top \bar{Z}_i P P^\top \\
&= \sum_{i=1}^{k} \bar{H} \bar{Z}_i P P^\top \\
&\overset{(11)}{=} \sum_{i=1}^{k} d_i d_i^\top \bar{Z}_i P P^\top,
\end{aligned}$$

which completes the proof. ■

In the following we will no longer include projections onto the range space in our statements but simply perform all constructions in this space, i.e., we assume the range space is the entire space. The corollary below asserts, that in the case of a quadratic f the Hessian will be reconstructed correctly for conjugate directions and a matching choice of the cost matrix C in the objective.

Corollary 5 *Let $A \in S^n$ be positive definite and v_i, $i = 1, \ldots, n$, be a family of conjugate directions of A satisfying $v_i^\top A v_j = \delta_{ij}$ for $i, j = 1, \ldots, n$ and Kronecker delta δ_{ij}. Setting, in Th. 4, $k = n$, $d_i = Av_i$, and $C = \sum_{i=1}^n \zeta_i v_i v_i^\top$ for any choice of $\zeta_i > 0$, $i = 1, \ldots, n$, results in $\bar{H} = A$ as the optimal solution of (8).*

Proof. The matrix $V = [v_1, \ldots, v_n]$ is regular and $V^\top A V = I_n$. This implies $A^{-1} = VV^\top = \sum_{i=1}^n v_i v_i^\top$. In other words, the v_i, $i = 1, \ldots, n$ form an orthonormal basis with respect to the inner product $\langle u, v \rangle_A = v^\top A u$ and each vector u has a unique representation $u = VV^\top Au = \sum_{i=1}^n \langle u, v_i \rangle_A v_i$.

Now $\bar{H} = A \succ 0$ is feasible for (8), because for $j = 1, \ldots, n$ and any $u \in \mathbb{R}^n$ there holds $u^\top (\bar{H} - d_j d_j^\top) u = u^\top (A - A v_j v_j^\top A) u = \sum_{i \in \{1,\ldots,k\} \setminus \{j\}} (u^\top A v_i)^2 \geq 0$. In the dual (9) putting $\bar{Z}_i = v_i v_i^\top \succeq 0$, $i = 1, \ldots, n$ and $C = A^{-1}$ yields a feasible solution by $\sum_{i=1}^n \bar{Z}_i = A^{-1}$. Furthermore, primal and dual objective coincide, $\langle A^{-1}, \bar{H} \rangle = n = \sum_{i=1}^n \langle A v_i v_i^\top A, v_i v_i^\top \rangle = \sum_{i=1}^n \langle d_i d_i^\top, \bar{Z}_i \rangle$. The correctness for the other choices of C now follows from Th. 4. ■

Note that this still holds if further constraints based on other directions (within the same span) are added. The eigenvectors of an eigenvalue decomposition result in a particularly pleasant set of conjugate directions. In this sense the following may be seen as a corollary to the previous one, but a direct proof is equally short.

Corollary 6 *Let $A \in S^n$ be positive definite with eigenvalue decomposition $A = \sum_{i=1}^n \lambda_i v_i v_i^\top$ with orthonormal v_i, $i = 1, \ldots, n$. In Th. 4 the choice $C = I_n$, $k = n$, and $d_i = \sqrt{\lambda_i} v_i$, $i = 1, \ldots, n$, yields $\bar{H} = A$ as optimal solution of (8).*

Proof. $\bar{H} = A \succ 0$ is feasible for (8), because for $j = 1, \ldots, n$, $\bar{H} - d_j d_j^\top = \sum_{i \in \{1,\ldots,n\} \setminus \{j\}} \lambda_i v_i v_i^\top \succeq 0$. In the dual (9) putting $\bar{Z}_i = v_i v_i^\top \succeq 0$, $i = 1, \ldots, n$, yields a feasible solution by $\sum_{i=1}^n \bar{Z}_i = I_n$. Finally, primal and dual objective coincide, $\langle I_n, \bar{H} \rangle = \sum_{i=1}^n \lambda_i = \sum_{i=1}^n \langle \lambda_i v_i v_i^\top, v_i v_i^\top \rangle = \sum_{i=1}^n \langle d_i d_i^\top, \bar{Z}_i \rangle$. ■

In summary, Lemma 3 ensures that the Hessian of a quadratic function is a feasible solution of (4) for suitable ε, while corollaries 5 and 6 show that a

subset of the difference vectors $y^i - \hat{y}$ has to be close to conjugate directions or eigenvectors of the Hessian in order to exclude matrices with "smaller" curvature. As argued above, the actual length of the vectors $y^i - \hat{y}$ is of limited relevance for the quality of the solution in the case of quadratic functions. In tentative experiments with random data optimal solutions exhibited bigger curvature than the Hessian in well populated directions but smaller ones in lesser covered regions.

For practical purposes we consider solving (4) as computationally too expensive. Instead, we choose a $Q \in \mathbb{R}^{n\times h}$, whose columns form an orthonormal basis of the hopefully most relevant subspace of $\mathcal{D}$ of Th. 2. In the nonsmooth case a natural choice for this Q is to compute, for some parameter $h \in \mathbb{N}$, the eigenvectors to the h largest eigenvalues of the matrix $VV^\top$ for $V = [d_1, \ldots, d_k]$ with $d_i = (\bar{g} - g^i)/\sqrt{2(\bar{\gamma} + \varepsilon - \gamma^i + (\bar{g} - g^i)^\top \hat{y})}$ for $i = 1, \ldots, k$. Alternatively, compute the singular value decomposition of V and choose the left eigenvectors to the h largest singular values. Next, we restrict $\bar{H}$ to be of the form $\bar{H} = Q\Lambda_{\bar{H}}Q^\top$ with diagonal $\Lambda_{\bar{H}} = \operatorname{diag}(\lambda_1^{\bar{H}}, \ldots, \lambda_h^{\bar{H}})$. Finally, instead of (3), which would now read $\Lambda_{\bar{H}} \succeq Q^\top d_i d_i^\top Q$, we put

$$\lambda_j^{\bar{H}} = \max\{\operatorname{diag}(Q^\top d_i d_i^\top Q)_j = (Q^\top d_i)_j^2 \colon i = 1, \ldots, k\} \qquad j = 1, \ldots, h. \quad (13)$$

Again, for smooth f the selection of Q via the singular value decomposition of V may be justified to some extent. Indeed, consider the quadratic f of Lemma 3 and the $d_i = A(\hat{y} - y^i)/\|A^{\frac{1}{2}}(\hat{y} - y^i)\|$ appearing in its proof. If the generating directions $\hat{y} - y^i$ are spread rather randomly, the next result gives some hope that the columns of Q will be close to an eigenvector basis of A.

Theorem 7 *Given $\delta > 0$ and a positive definite $A \in S^n$, let directions $s_i \in \mathbb{R}^n \setminus \{0\}$, $i = 1, \ldots, k$, be chosen by a rotationally invariant distribution with second moments, let $V = [d_1, \ldots, d_k]$ with $d_i = As_i/\|A^{\frac{1}{2}}s_i\|$ and let $Q\Sigma_V P^\top = V$ denote a singular value decomposition with $Q^\top Q = I_n$. With probability going to one for $k \to \infty$ there is an orthogonal matrix $\bar{Q} = Q + O(\delta)$ diagonalizing $A = \bar{Q}\Lambda_A\bar{Q}^\top$.*

Proof. Let $A = Q_A\Lambda_A Q_A^\top$ with $Q_A^\top Q_A = I_n$ and $\Lambda_A = \operatorname{diag}(\mu_1, \ldots, \mu_n)$, $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n > 0$ be an eigenvalue decomposition of A, then

$$\frac{1}{k}VV^\top = Q_A\Lambda_A^{\frac{1}{2}}\left(\frac{1}{k}\sum_{i=1}^k \frac{\Lambda_A^{\frac{1}{2}}Q_A^\top s_i}{\|\Lambda_A^{\frac{1}{2}}Q_A^\top s_i\|}\frac{s_i^\top Q_A\Lambda_A^{\frac{1}{2}}}{\|\Lambda_A^{\frac{1}{2}}Q_A^\top s_i\|}\right)\Lambda_A^{\frac{1}{2}}Q_A^\top.$$

By the rotational symmetry of the distribution we may simplify $Q_A^\top s_i$ to s_i. Next we show that all offdiagonal elements of the middle random matrix go to zero for $k \to \infty$ with probability one. For this denote the elements of this random matrix by

$$x_{gh} = \left[\sum_{i=1}^k \frac{1}{k}\frac{\Lambda_A^{\frac{1}{2}}s_i(\Lambda_A^{\frac{1}{2}}s_i)^\top}{\|\Lambda_A^{\frac{1}{2}}s_i\|^2}\right]_{gh} = \sum_{i=1}^k \frac{1}{k}\frac{\mu_g^{\frac{1}{2}}\mu_h^{\frac{1}{2}}[s_i]_g[s_i]_h}{\|\Lambda_A^{\frac{1}{2}}s_i\|^2}, \quad 1 \leq g \leq h \leq n.$$

For $g < h$ and by the assumptions on the distribution, the density of $\frac{\mu_g^{\frac{1}{2}}\mu_h^{\frac{1}{2}}[s]_g[s]_h}{\|\Lambda_A^{\frac{1}{2}}s\|^2}$ is symmetric about the origin and its value is confined to $[-1, 1]$, thus the

expected value is 0 and the variance is bounded by 1. Therefore the expected value of x_{gh} is 0 and its variance bounded by $\frac{1}{k}$. Applying union bound and Chebyshev's inequality we obtain for any $\epsilon > 0$

$$\mathbb{P}(\max_{1\leq g<h\leq n} |x_{gh}| > \varepsilon) \leq n(n-1)\mathbb{P}(|x_{12}| > \varepsilon) \leq n(n-1)\frac{\mathrm{Var}(x_{12})}{\varepsilon^2} \leq \frac{n(n-1)}{\varepsilon^2 k}.$$

For ordering the diagonal elements, note that $\mu_g > \mu_h$ implies

$$\mathbb{E}\big(\tfrac{1}{\mu_g}\textstyle\sum_{j\in\{1,\ldots,n\}\setminus\{g\}} \mu_j [s]_j^2\big) < \mathbb{E}\big(\tfrac{1}{\mu_h}\textstyle\sum_{j\in\{1,\ldots,n\}\setminus\{h\}} \mu_j [s]_j^2\big)$$

and therefore $\mathbb{E}(x_{gg}) = \mathbb{E}(\frac{[s]_g^2}{[s]_g^2 + \frac{1}{\mu_g}\sum_{j\in\{1,\ldots,n\}\setminus\{g\}} \mu_j [s]_j^2}) \geq \mathbb{E}(x_{hh})$. Thus, with probability going to one for $k \to \infty$,

$$\Lambda_A^{\frac{1}{2}} \left(\frac{1}{k} \sum_{i=1}^{k} \frac{\Lambda_A^{\frac{1}{2}} Q_A^\top s_i}{\|\Lambda_A^{\frac{1}{2}} Q_A^\top s_i\|} \frac{s_i^\top Q_A \Lambda_A^{\frac{1}{2}}}{\|\Lambda_A^{\frac{1}{2}} Q_A^\top s_i\|} \right) \Lambda_A^{\frac{1}{2}} = D + Y$$

for some diagonal matrix D with $D_{11} \geq \cdots \geq D_{nn} > 0$, where $D_{gg} > D_{hh}$ if and only if $\mu_g > \mu_h$, and perturbation matrix $Y \in \mathbb{R}^{n\times n}$ of norm $\|Y\|_F < \delta$.

Now the result follows from [13, Lemma 4.3]. ∎

In practice the differences $\hat{y} - y^i$ are unlikely to be distributed randomly. Still, together with Cor. 6 the result indicates why in the smooth case the use of the singular value decomposition in the practical heuristic might indeed produce a reasonable estimate of the projected Hessian.

The construction in Th. 2 is designed for forming a "best" $\bar{H}$ after a descent step. It could be extended to updating $\bar{H}^k$ to $\bar{H}^{k+1}$ after null steps by memorizing $\bar{\omega}$, by replacing in (4) the constraint $H \succeq 0$ by $H \succeq \bar{H}^k$ and by requiring (3) only for those new $\omega = (\gamma, g)$ that satisfy the condition $\gamma + g^\top \hat{y} < \bar{\gamma} + \bar{g}^\top \hat{y} + \varepsilon$. In the heuristic approach, the new subgradients might be employed to form further inequalities for λ_j^H in (13), but we will not pursue this here.

3 Dynamic Submodel Selection for Sums of Convex Functions

Throughout this section, the function $f = \sum_{i\in R} f_i$ is a sum of proper closed convex functions $f_i \colon \mathbb{R}^n \to \mathbb{R}$, $i \in R$, for some finite index set R. When evaluating f at candidate $\bar{y}^k$, a separate oracle will be called for each f_i, $i \in R$. Each returns a function value $f_i(\bar{y}^k)$ and a subgradient $g_i^k \in \partial f_i(\bar{y}^k)$ giving rise to the minorant $f_{\omega_i^k}$ with $\omega_i^k := (\gamma_i^k := f_i(\bar{y}^k) - (g_i^k)^\top \bar{y}^k, g_i^k)$. For any index subset $J \subseteq R$, the function $f_J := \sum_{i\in J} f_i$ then has a subgradient $g_J^k := \sum_{i\in J} g_i^k \in \partial f_J(\bar{y}^k)$ and $\omega_J^k := (\gamma_J^k := \sum_{i\in J} \gamma_i^k, g_J^k)$ describes the corresponding minorant. In the same vein, the respective aggregates will be denoted by $\bar{\omega}_i^k = (\bar{\gamma}_i^k, \bar{g}_i^k)$ and $\bar{\omega}_J^k = (\bar{\gamma}_J, \bar{g}_J^k)$. The subgradient information of a subset of past iterations $K \subseteq \{0, \ldots, k\}$ is collected in (possibly multi-)sets $W_J^K := \{\omega_J^h \colon h \in K\}$ and we put $W_i^K := W_{\{i\}}^K$. Because

$$f_{W_R^K} = \max_{\omega_R^h \in W_R^K} f_{\omega_R^h} = \max_{h\in K} \sum_{i\in R} f_{\omega_i^h} \leq \sum_{i\in R} \max_{h\in K} f_{\omega_i^h} = \sum_{i\in R} f_{W_i^K} \leq \sum_{i\in R} f_i = f,$$

the use of separate cutting models for each function f_i gives a better overall cutting model. This is well known and in practice this difference is often significant, but even for moderate $|R|$ the cost of solving the corresponding bundle subproblem is quickly prohibitive. For sums of convex functions, however, there is some hope that close to $\hat{y}^k$ only a few of the f_i will exhibit relevant nonsmooth behavior in the sense that the subgradients differ strongly.

The idea is to iteratively identify a small but relevant index subset $J^k \subseteq R$ for which subgradients change rapidly. For the f_i, $i \in J^k$, separate models $\widehat{W}_i^k$ will be used, while a common *summodel* $\widehat{W}_{\bar{J}^k}^k$ represents $f_{\bar{J}^k}$, where $\bar{J}^k = R \setminus J^k$ comprises the complement of the indices. As usual, convergence is ensured by enforcing $\{\bar{\omega}_i^k, \omega_i^k\} \subseteq \widehat{W}_i^{k+1}$, $i \in J^{k+1}$, and $\{\bar{\omega}_{\bar{J}^{k+1}}^k, \omega_{\bar{J}^{k+1}}^k\} \subseteq \widehat{W}_{\bar{J}^{k+1}}^{k+1}$. In order to keep additional computations at a minium, the index selection approach chosen here starts with $J^1 = R$ and determines J^{k+1} on basis of the development of the relative sizes of the differences $f_i(\bar{y}^k) - f_{\bar{\omega}_i^k}(\bar{y}^k) = f_{\omega_i^k}(\bar{y}^k) - f_{\bar{\omega}_i^k}(\bar{y}^k)$ compared to $f(\bar{y}^k) - f_{\bar{\omega}_R^k}(\bar{y}^k)$. In detail, for some fixed size limit $\bar{r} \leq |R|$, up to $\bar{r}$ indices $i \in R$ having largest value (initially $\rho_i^1 = 0$).

$$\rho_i^{k+1} := 0.9 \cdot \rho_i^k + 0.1 \cdot \frac{f_{\omega_i^k}(\bar{y}^k) - f_{\bar{\omega}_i^k}(\bar{y}^k)}{f(\bar{y}^k) - f_{\bar{\omega}_R^k}(\bar{y}^k)}$$

form the set J^{k+1}. Except for evaluating the affine $f_{\bar{\omega}_i^k}(\bar{y}^k)$ for $i \in R$, no additional function or model evaluations are needed in this. Large values $\frac{f_{\omega_i^k}(\bar{y}^k) - f_{\bar{\omega}_i^k}(\bar{y}^k)}{f(\bar{y}^k) - f_{\bar{\omega}_R^k}(\bar{y}^k)}$ should indeed indicate quite well which $i \in R$ merit separate models due to a strong deviation between new subgradient and aggregate. In contrast, the combination with previous values is somewhat arbitrary. It is motivated by limited empiric evidence that without including some inertia based on past observations, oscillations in the index set inhibit the consolidation of the respective aggregates and seem to slow down convergence.

A further computationally relevant choice that is difficult to justify mathematically, is the size of the respective bundles in $\widehat{W}_{\bar{J}^{k+1}}^{k+1}$ and $\widehat{W}_i^{k+1}$ for $i \in J^{k+1}$. Attempts to assign the sizes in dependence of ρ_i^{k+1} produced inconclusive results in comparison to the trivial approach of allowing each model to use the same maximal number of bundle elements. The experiments of Section 5 therefore only explore the latter approach for various identical bundle sizes per model.

It remains to specify the selection of the actual bundle elements whenever the size exceeds two. For each $i \in R$ the last $\bar{s}$ minorants ω_i^h, $h \in K^k = \{\max\{0, k-\bar{s}+1\}, \ldots, k\}$, and the aggregate $\bar{\omega}_i^k$ are kept available. The summodel part $\widehat{W}_{\bar{J}^{k+1}}^{k+1}$ explicitly includes the aggregate $\bar{\omega}_{\bar{J}^{k+1}}^k$ and the new minorant $\omega_{\bar{J}^{k+1}}^k$, the remaining positions are filled up by minorants $\omega \in W_{J^{k+1}}^{K^k}$ that do not repeat and have largest value $f_\omega(\bar{y}^k)$. The minorants needed in this can be updated by subtracting and adding the minorants ω_i^h whose indices i enter ($i \in J^{k+1} \setminus J^k$) and leave ($i \in J^k \setminus J^{k+1}$) the separate models. For forming $\widehat{W}_i^{k+1}$ for $i \in J^{k+1}$ the standard routine of the callable library ConicBundle [5] is used. Besides the ideas above it also uses dual information, may generate several aggregates and has grown over some time; as it is rather cumbersome to state exactly, we refer directly to the code that is publically available.

4 Implementational Aspects

Additional implementational issues arise for scaling in combination with the dynamic selection of subproblems. Aspects related to submodel selection will be discussed first.

4.1 Submodel Selection

In order to explain efficiency aspects involved in dynamic submodel selection, it will be convenient to consider the example of relaxing coupling constraints for decomposing some problem. This is a typical technique employed for multicommodity flow problems or scheduling problems with coupling resource constraints. Our computational experiments use instances from practice, where trucks are scheduled for rearranging pallets between several warehouses in order to satisfy demand with high probability [8]. For our purposes there is no need to go into the details of the application; the following abstract setting will suffice. Consider resource constraints coupling several producers $i \in R$, each maximizing its profit over a compact feasible set $\mathcal{X}_i \subset \mathbb{R}^{n_i}$,

$$\max \left\{ \sum_{i \in R} c_i^\top x_i \colon \sum_{i \in R} A_i x_i \leq b,\ x_i \in \mathcal{X}_i,\ i \in R \right\}. \tag{14}$$

Penalizing violations of the coupling constraints by Lagrange multipliers in the respective cost functions decomposes the problem into independent subproblems,

$$f_i(y) := \max_{x_i \in \mathcal{X}_i} (c_i - A_i^\top y)^\top x_i, \quad i \in R. \tag{15}$$

For each function f_i the relevant compact set of minorants is formed by f_ω with $\omega \in W_i = \{(c^\top x_i, -A_i x_i) \colon x_i \in \operatorname{conv} \mathcal{X}_i\}$. The bundle method optimizes the multipliers $y \geq 0$ by solving

$$\min_{y \geq 0} f(y) = b^\top y + \sum_{i \in R} \max_{\omega \in W_i} f_\omega(y). \tag{16}$$

In our discussion we will ignore the sign constraints on y but refer to [6] for a detailed explanation of our approach to incorporate them. In practice, the dual bound $f(y)$ is often less important than obtaining approximate primal solutions $x_i \in \operatorname{conv} \mathcal{X}_i$ to (14), because these form the basis for rounding heuristics. Primal approximations are easily provided by extending ω to a triple $(c^\top x_i, -A_i x_i, x_i)$ and then aggregating the x_i along with ω. Thus, the aggregates $\bar{\omega}_i$ will now always be accompanied by a primal aggregate $\bar{x}_i$. In fact, in the current setting it seems worth to store the x_i and $\bar{x}_i$ alone and generate the respective minorants explicitly as explained next.

Compared to forming a joint common subgradient $g_R^k = \sum_{i \in R} g_i^k$ and aggregate $\bar{\omega}^k$ once per iteration, the separate handling of the subgradients g_i^k and aggregates $\bar{\omega}_i^k$ in every iteration may cause significant additional work if n and $|R|$ are large. The primal aggregates $\bar{x}_i^k$, however, have to be computed separately for each $i \in R$ in any case, so these costs do not differ whether submodels are used or not. Whenever $A_i x_i$ is easy to compute, *e. g.*, because the x_i and $\bar{x}_i$ are small or reasonably sparse, one can avoid computing and maintaining $\bar{g}_i^k \in \mathbb{R}^n$ for $i \in \bar{J}^k$ by aggregating the primal $\bar{x}_i$ and by updating the subgradients $g_{\bar{J}^k}^h$,

$h \in K^k$, and the aggregate $\bar{g}^k_{\bar{J}^k}$ of the summodel with Ax^h_i or $A\bar{x}^k_i$ for those $i \in R$ that move in and out of $\bar{J}^k$.

When implementing this within a general purpose code it seems preferable to offer the possibility of using affine transformations within the arguments. Then (16) turns into

$$\min_{y \geq 0} f(y) = b^\top y + \sum_{i \in R} f'_i(c_i - A_i^\top y)$$

with support functions

$$f'_i \colon \mathbb{R}^{n_i} \to \mathbb{R}, z \mapsto \max_{x_i \in \mathcal{X}_i} (x_i^\top z). \tag{17}$$

In this setting the primal aggregates $\bar{x}_i$ are the standard aggregates of the single support functions which give rise to the aggregate in the common design space by the chain rule, $\bar{g}_i = -A_i \bar{x}_i$. In our experiments we investigate the computational differences between using the f_i as before, where the affine transformation is inside the oracle, and the *affine transformation setting* calling f'_i with arguments affinely transformed within the bundle code.

Even if the affine transformation setting avoids maintaining the $\bar{g}_i \in \mathbb{R}^n$ for the summodel indices $i \in \bar{J}$, the additional work associated with handling each bundle $i \in R$ separately instead of operating with one large common object is significant if $|R|$ is large, see Section 5.

4.2 Scaling

We first describe the actual scaling heuristic based on the singular value decomposition in detail for a single convex function f, then discuss the extension to sums of convex functions with possibly affinely transformed arguments and dynamic submodel selection, and mention necessary adaptations for diagonal scaling.

For a single convex function f the input to the scaling routine comprises a (multi-)set of minorants $\{\omega^h = (\gamma^h, g^h) \colon h \in K\}$ for some index set $K \subseteq \{1, \ldots, k\}$, the current aggregate $\bar{\omega} = (\bar{\gamma}, \bar{g})$ and a point $\bar{y}$, which is typically the candidate obtained by $\bar{\omega}$ and the new center of stability $\hat{y} = \bar{y}$. In addition, it is given a violation parameter $\varepsilon > 0$, a maximum curvature parameter $\kappa_{\max} \leq 10^6$ and the low rank representation $\bar{G} \in \mathbb{R}^{n \times \bar{h}}$ of the previously computed scaling matrix $\bar{H} = \bar{G}\bar{G}^\top$ (initially $\bar{h} = 0$). From this it computes a new low rank representation G as follows.

In a first step all those minorants are selected that are reasonably close to the current function value in $\bar{y}$ and whose gradients differ sufficiently from the aggregate $\bar{g}$. For the precise statement, let

$$\begin{array}{rcll}
\hat{\delta} & = & \max\{\gamma^h + \bar{y}^\top g^h \colon h \in K\} & \text{largest value in } \bar{y}, \\
\bar{\delta} & = & \bar{\gamma} + \bar{y}^\top \bar{g} & \text{aggregate value in } \bar{y}, \\
\underline{\Delta} & = & \max\{1/\kappa_{\max}, 1.01 \cdot (\hat{\delta} - \bar{\delta})\} & \text{bound on minimum difference}, \\
\bar{\Delta} & = & \max\{10^{-3}, 100 \cdot (\hat{\delta} - \bar{\delta}), 100 \cdot \varepsilon\} & \text{bound on maximum difference}, \\
\eta & = & \min\left\{10^{-3}, \max\left\{10^{-12}, \frac{\hat{\delta} - \bar{\delta}}{|\hat{\delta}| + 1}\right\}\right\} & \text{relative precision}, \\
\delta^h & = & \max\{\underline{\Delta}, \varepsilon + \bar{\delta} - \gamma^h - \bar{y}^\top g^h\} & \text{difference used for } h \in K,
\end{array}$$

then the minorants selected are

$$\bar{K} = \left\{h \in K \colon (\delta^h < \bar{\Delta}) \wedge (\|\bar{g} - g^h\|^2 \geq 10^{-4} \cdot \max\{1, \|\bar{g}\|^2\})\right\}.$$

For this selection of minorants the "most important" subspace of the safe-guarded difference vectors $\tilde{d}_h = (\bar{g} - g^h)/\sqrt{2\delta^h}$, $h \in \bar{K}$, is identified by computing an orthogonal basis belonging to the largest singular values of the singular value decomposition of $D = [d_h]_{h\in K}$. In detail, if $\sigma_1 \geq \cdots \geq \sigma_{|\bar{K}|}$ are the singular values of the singular value decomposition $P\Sigma Q = D$, then the largest index $\bar{\jmath} \leq |\bar{K}|$ with $\sigma_{\bar{\jmath}} > \eta\sigma_1$ is used to select the first $j = 1, \ldots, \bar{\jmath}$ columns $P_{\bullet,j}$ of P as this orthogonal basis. The matching "eigenvalues" are then computed by

$$\tilde{\lambda}_j = \max\{\left\langle P_{\bullet,j}, \bar{g} - g^h\right\rangle^2 /(2\delta^h) \colon h \in \bar{K}\}, \quad j = 1, \ldots, \bar{\jmath}.$$

Among those, any index j with $\tilde{\lambda}_j < \eta \max\{\tilde{\lambda}_j \colon j = 1, \ldots, \bar{\jmath}\}$ is deleted. The remaining $\tilde{G} = \tilde{P}\tilde{\Lambda}^{\frac{1}{2}}$ forms the low rank representation of the scaling matrix $\tilde{H} = \tilde{G}\tilde{G}^\top$. When relative precision requirements are moderate (we use $\eta \geq 10^{-4}$) this $G = \tilde{G}$ is the result of the scaling heuristic. When relative precision requirements are high ($\eta < 10^{-4}$) we prefer to stabilize the scaling matrix via taking a kind of convex combination with the previous scaling matrix. For this we use the singular value decomposition on the matrix $[\tilde{G}/2, \bar{G}/2]$ and keep again only the singular values within a factor η of their largest singular value to form the final G.

In most applications f is optimized over a box constrained domain. As pointed out before, we employ the approach [6] for incorporating box constraints. With general scaling matrices the approach would involve solving one or more additional convex quadratic programs per step, while diagonal scaling results in a scalar comparison per bounded entry. Thus, for computational efficiency, the current implementation only uses the diagonal of $GG^\top$ in the presence of box constraints.

When optimizing sums of convex functions with dynamic submodel selection, the scaling heuristic is called after a descent step for the summodel $\bar{J}^k$ as well as for every selected model $i \in J^k$ that gave rise to this descent step, *i. e.*, it is called for every model, whose aggregate has been tuned to the subgradients presently in the model. The diagonals of all these scaling matrices are added up to form a joint scaling diagonal. Note that in the case of an affinely transformed $f_i'(c_i - A_i^\top y)$ the computation of G_i for f_i' allows to apply the chain rule before extracting the diagonal, *i. e.*, $\operatorname{diag}(A_i G_i G_i^\top A_i^\top)$ is the contribution of $i \in J$ in the presence of affine transformations.

Once all diagonal parts are summed up, the previous diagonal and the new diagonal are combined with weights 0.9 and 0.1, respectively, to form a diagonal matrix $\bar{D}^{k+1}$. As pointed out in the beginning of Section 2, the actual scaling matrix used will be of the form $H^{k+1} = u^{k+1}I + \bar{D}^{k+1}$ following the approach of [7] with u^k determined as in [6] for establishing a kind of trust region control. In general, the diagonal $\bar{D}$ may well turn out to represent the low rank information of $GG^\top$ rather badly, *e. g.*, if G is a vector of all ones all directions will be restricted significantly by its diagonal representation. Unfortunately, this effect appears rather frequently if the quadratic term approximates a nonsmooth valley tighter and tighter. The aim of diagonal scaling is mainly to correct the relative sizes of the variables while the role of u^k is to restrict the step size. For judging whether the current diagonal representation is overly restrictive in directions of importance, the current aggregate $\bar{g}^k$ seems to be the best candidate. Thus, if $u^k < 1$ indicates interest in larger steps but by $\tau := \frac{(\bar{g}^k)^\top (H^{k+1})^{-1}\bar{g}^k}{\|\bar{g}^k\|^2} < \frac{1}{2}$ the

diagonal enforces significantly smaller step sizes than the usual norm, then H^{k+1} is reset to use $\tau\bar{D}$ instead of $\bar{D}$.

5 Numerical Experiments

In order to explore the relative performance of scaling and dynamic submodel selection on instances of practical relevance, we use the truck scheduling instances of [8] that optimize the transportation of pallets of different products between three warehouses so as to maximize the probability that required products are available at the right place in time. All instances are linear programs, most can be solved quite efficiently by today's commercial packages and the optimal values are available. The reason for using these instances is that they offer rich possibilities for setting up realistic optimization problems that require minimizing the sum of a large number of convex functions. Without going into details, the model is based on time expanded networks for each product and truck representing the flow of the pallets and trucks between the warehouses over a day. With the need to transport typically between 500 to 1100 products by 4 to 6 trucks the number of arcs in all graphs ranges between 0.3 and 1.2 million. Lagrangian relaxation of 2500 to 5600 coupling constraints (this is the dimension n here) decomposes the instances into as many network flow problems as products and trucks and roughly three times as many convex, one dimensional, piecewise linear functions which may all appear as separate oracles. Handling the full resulting number of oracles separately turns out to be inefficient in any setting. In our experiments we therefore group the networks and the functions each into a parameter specified number of separate oracles and study the effects of changing this parameter. Let $r \in \mathbb{N}$ be this parameter, then the networks will be grouped into r oracles and the piecewise linear functions will also be grouped into r oracles by sequence of appearance, for a total of $2r$ oracles.

We will compare three implementation variants. They use the same modified version of the ConicBundle library [5] and the same network solver [11] but differ in the organization of the oracles. The traditional variant, denoted by `trd`, works with a single oracle subsuming all network problems and piecewise linear functions. The second variant, called `dyn`, uses dynamic submodel selection for the oracles as in (15). The third, called `aft`, employs the affine transformation setting in combination with dynamic submodel selection, *i. e.*, the oracles are now the pure support functions (17).

The variants `dyn` and `aft` are, unfortunately, heavily influenced by several parameters. At this point it seems difficult to develop suitable, mathematically sound heuristic procedures for choosing these automatically. Rather we will explore several variants in order to shed some light on the influence of the parameters in the hope to gain a better understanding of their role. One central aim is to study the dependence of the performance on the number of oracles. For this we use $r \in \{10, 50, 100\}$ resulting in instances with r network flow oracles and r piecewise linear function oracles grouped as described above. Further parameters of relevance are the maximum number of submodels $|J| \leq m \in \{5, 10, 20\}$ and the bundle size limit $b \in \{2, 5, 10\}$ for the sum- and submodels resulting in (dual) bundle subproblems of order at most $(m+1)b$. For comparison, the traditional variant `trd` will be run with maximum bundle size $b \in \{2, 50, 100\}$. For all three code variants we also experiment with the

number $p \in \{0, 10, 50\}$ of past subgradients employed in the scaling heuristic, where 0 indicates that no scaling is used. While the parameter ranges are quite limited, each original instance now gives rise to a total of $2 \cdot 3^4 + 3^2 = 191$ runs. Using a limit of 30 minutes CPU time per variant of an instance, computing all variants for one instance requires roughly 4 CPU days; this is the reason why the actual number of instances $i = 32$ tested per variant may seem moderate.

The codes were compiled and run on several identical Intel(R) Xeon(R) CPU E5-2620 v3 machines with 2.40GHz, 15360 KB cache, 64 GB memory and operating system Ubuntu 17.04 in purely sequential mode. For the current investigations actual running times are of little importance (the best settings require 8-9 minutes for a relative precision of 10^{-3} on average). The aim is to gain some experience on the relative improvement achievable by dynamic scaling and submodel selection and on the tradeoffs involved. For this we will use performance diagrams [3] to compare the various settings described above with respect to computation time and oracle calls. They are constructed as follows. Given an instance with its precomputed optimal value, we determine for each variant the first descent step with objective value within a relative precision of 10^{-3} of the precomputed optimal value and use the corresponding time and number of oracle calls for comparison. The performance diagram of a given subset of code variants now displays for each variant and factor $\rho \in [1, 2]$ the number of instances that were solved by this variant to the required precision within a factor ρ of the minimal time/calls recorded for the instance over all considered variants. Note that with an average computation time of 9 minutes and a time limit of 30 minutes, some instances will not be solved to the required precision and larger factors than 2 do not give useful information. By means of the performance profiles we will first determine a suitable parameter choice for `trd`, then jointly for `dyn` and `aft`, and finally compare these against each other.

In order to investigate, via the performance profiles, whether a single parameter influences the performance of the algorithm with a clear tendency independent of the other parameters, we choose the following nonstandard approach. For the different values of the parameter at stake we compare the performance with respect to all suitable "parameterized instances". These are formed by considering each original instance for every selection of the remaining parameters as a separate parameterized instance. If there is a clear best choice for the parameter at stake for all these parameterized instances, then we fix the corresponding parameter and continue with studying the influence of the other parameters on the reduced parameter space. The computational results indicate, that there is a clear winner among the scaling parameter values and so we consider this parameter first.

For each of the three scaling values $p \in \{0, 10, 50\}$, `trd` is tested on $3 \cdot i = 96$ parameterized instances (one for each choice of the bundle size $b \in \{2, 50, 100\}$) and the performance profiles for running time and number of oracle calls are displayed in Fig. 1. Note that not all of these instances are solved to the required precision within the time limit, therefore no code variant reaches the top; but there is a clear winner, $p = 50$. Fixing $p = 50$ we now compare the performance of the three bundle size variants $b \in \{2, 50, 100\}$ on the original i instances in Fig. 2. Again there is a clear winner, the minimal bundle $b = 2$. While this is not so surprising in terms of computation time, it is remarkable that this choice also requires the fewest oracle calls throughout. We are not aware of a solid mathematical explanation of this effect. Maybe the somewhat erratic exploration

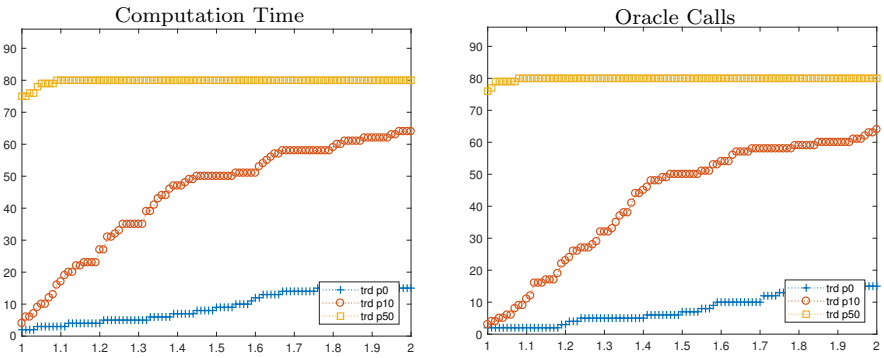

Figure 1: Performance profiles for `trd` using the scaling heuristic with $p \in \{0, 10, 50\}$ past subgradients on instances parameterized by $b \in \{2, 50, 100\}$.

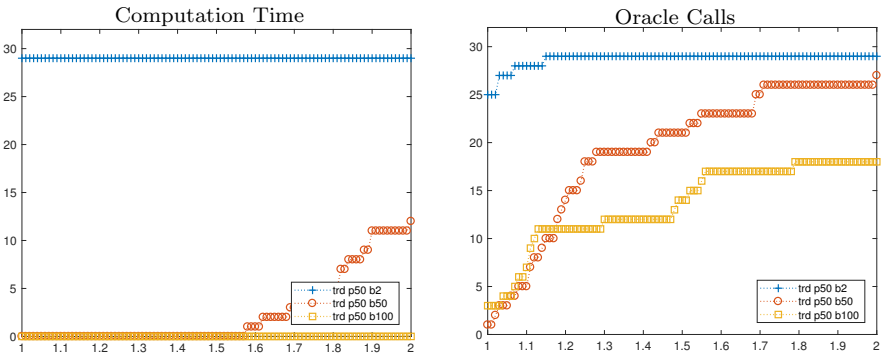

Figure 2: Performance profiles for `trd` using the scaling heuristic with $p = 50$ past subgradients and bundle size $b \in \{2, 50, 100\}$.

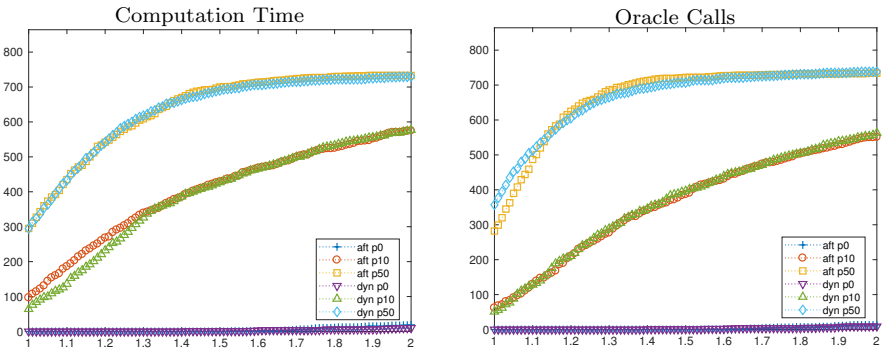

Figure 3: Performance profiles for `dyn` and `aft` using the scaling heuristic with $p \in \{0, 10, 50\}$ past subgradients on instances parameterized by r, m and b.

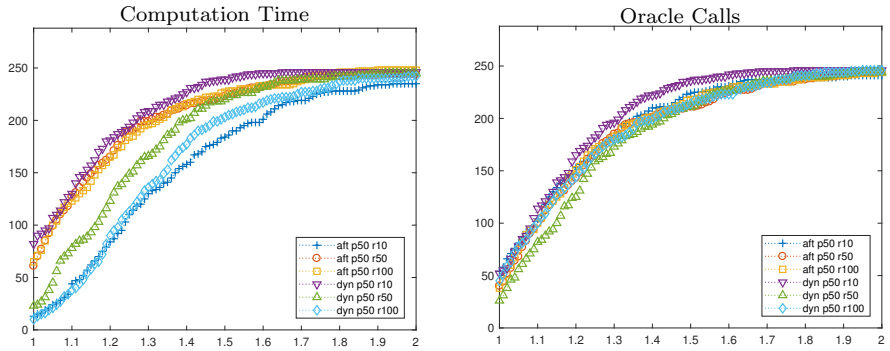

Figure 4: Performance profiles for `dyn` and `aft` with $p = 50$ and $r \in \{10, 50, 100\}$ on instances parameterized by m and b.

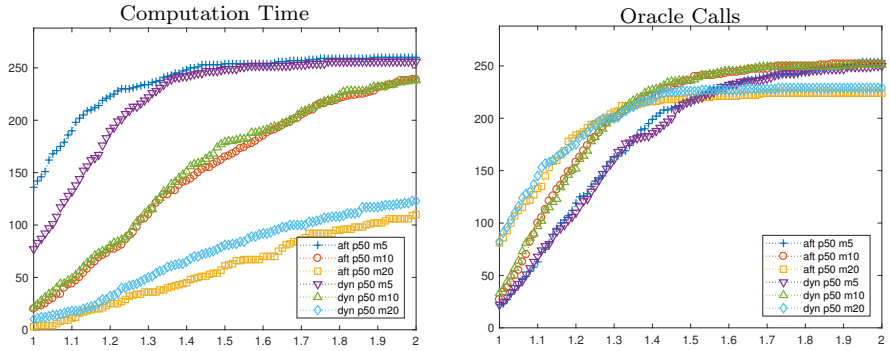

Figure 5: Performance profiles for `dyn` and `aft` with $p = 50$ and $m \in \{5, 10, 20\}$ on instances parameterized by r and b.

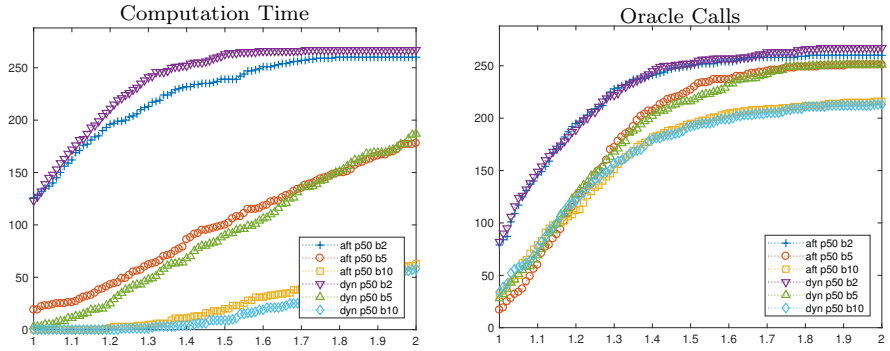

Figure 6: Performance profiles for `dyn` and `aft` with $p = 50$ and $b \in \{2, 5, 10\}$ on instances parameterized by r and m.

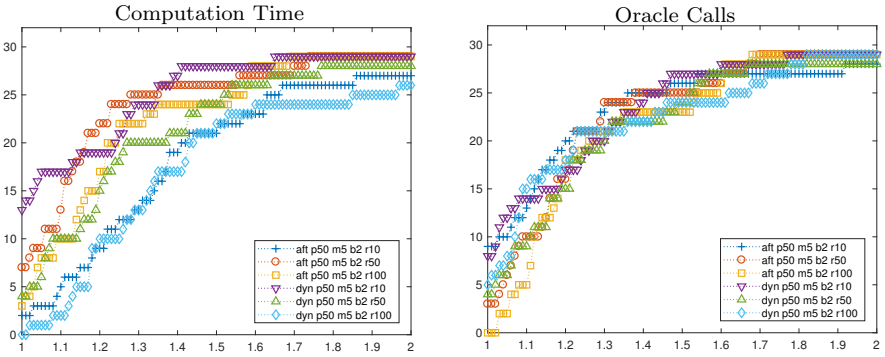

Figure 7: Performance profiles for `dyn` and `aft` with $p = 50$, $m = 5$, $b = 2$ and $r \in \{10, 50, 100\}$.

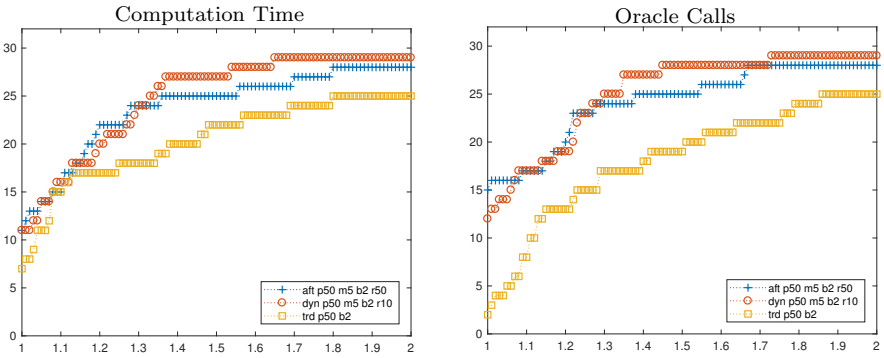

Figure 8: Performance profiles comparing `trd`, `dyn` and `aft` for the selected parameter settings.

by the minimal bundle is better if the bundle size is too small to fully control the relevant subspace. For `trd`, $p = 50$ and $b = 2$ are the winning choices on these instances.

Next we compare the use of scaling with $p \in \{0, 10, 50\}$ past subgradients for `dyn` and `aft` on $3^3 \cdot i = 864$ parameterized instances in Fig. 3. Again, $p = 50$ clearly dominates the variants without scaling and with only 10 past subgradients, so we keep $p = 50$ fixed in the following. In order to justify the next decision, the effect of each of the three parameters $r \in \{10, 50, 100\}$, $m \in \{5, 10, 20\}$ and $b \in \{2, 5, 10\}$ is displayed for its respective $3^2 \cdot i = 288$ parameterized instances in figures 4–6. For r the picture is least pronounced in Fig. 4, but note that `dyn` slows down for larger values – a significant part of this can be attributed to the work involved in forming the additional aggregates – while larger sizes in r are indeed better for `aft`. The influence of r on the number of calls seems moderate. For m (Fig. 5) and b (Fig. 6) the smallest values are clear winners in terms of computation time but for m the picture is less clear regarding the number of calls. For expensive oracles it might well be worth to consider a larger number of submodels. Fixing $m = 5$ and $b = 2$ we study $r \in \{10, 50, 100\}$ once more for the $i = 32$ original instances in Fig. 7. For `dyn` the choice $r = 10$ seems best, for

`aft` both larger values seem good with a slight advantage for $r = 50$.

Finally, Fig. 8 presents the performance profile of the winning parameter choices for `trd`, `dyn` and `aft` on the original $i =$ instances.

6 Conclusion

The proposed scaling heuristic works surprisingly well for the considered large scale real world instances. For minimizing sums of convex functions, dynamic submodel selection demonstrates some potential for improvements in computation time and number of oracle calls, but on the instances considered this improvement was comparatively moderate. Yet, in combination with affine argument transformations provided by the solver, dynamic submodel selection allows to deal efficiently with sums of a few hundred functions without the need to worry about grouping them into a few suboracles.

Several further aspects still need to be explored. On the implementational side this includes, *e. g.*, the use of general instead of diagonal scaling in combination with a suitable solver for the bundle subproblems, adapting the scaling information during null steps, and static or dynamic grouping of functions to form common oracles in order to limit the work in dynamic submodel selection. Is it possible to combine dynamic scaling and submodel selection with the asynchronous parallel bundle method of [4]? In terms of theory it would be interesting to know whether adaptations of the scaling approach allow to prove better convergence properties for bundle method, *e. g.*, for classes of smooth convex functions. A long standing open problem are solid mathematical guidelines for choosing the bundle size and its minorants. It is conceivable that the new scaling approach provides useful information for this task, *e. g.*, because relevant minorants should contribute significantly to the curvature of the scaling term, or because in conjunction with a quadratic term that is updated also in null steps the minimal model might indeed be preferable. These and similar aspects offer rich possibilities for future work.

References

[1] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical Optimization*. Springer, 2nd edition, 2006.

[2] W. de Oliveira and C. Sagastzábal. Bundle methods in the XXIst century: a bird's-eye view. *Pesquisa Operacional*, 34(3):647–670, 2014.

[3] E. Dolan and J. Moré. Benchmarking optimization software with performance profiles. *Math. Programming*, 91(2):201–213, 2002.

[4] F. Fischer and C. Helmberg. A parallel bundle framework for asynchronous subspace optimisation of nonsmooth convex functions. *SIAM J. Optim.*, 24(2):795–822, 2014.

[5] C. Helmberg. *ConicBundle 0.3*. Fakultät für Mathematik, Technische Universität Chemnitz, 2009. `http://www.tu-chemnitz.de/~helmberg/ConicBundle`.

[6] C. Helmberg and K. C. Kiwiel. A spectral bundle method with bounds. *Math. Programming*, 93(2):173–194, 2002.

[7] C. Helmberg, M. L. Overton, and F. Rendl. The spectral bundle method with second-order information. *Optimization Methods and Software*, 29(4):855–876, July 2014.

[8] C. Helmberg and S. Röhl. A case study of joint online truck scheduling and inventory management for multiple warehouses. *Operations Research*, 55(4):733–752, July 2007.

[9] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I+II*, volume 305 and 306 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1993.

[10] C. Lemaréchal, F. Oustry, and C. Sagastizábal. The $\mathcal{U}$-Lagrangian of a convex function. *Trans. Amer. Math. Soc.*, 352(2):711–729, 2000.

[11] A. Löbel. *MCF Version 1.2 – A network simplex Implementation*. Konrad-Zuse-Zentrum für Informationstechnik Berlin, Jan. 2000. Available at `http://www.zib.de/opt-long_projects/Software/Mcf` (free of charge for academic use).

[12] R. Mifflin and C. Sagastizábal. A $\mathcal{VU}$-algorithm for convex minimization. *Math. Programming*, 104(2-3):583–608, 2005.

[13] D. Sun and J. Sun. Strong semismoothness of eigenvalues of symmetric matrices and its application to inverse eigenvalue problems. *SIAM J. Numer. Anal.*, 40(6):2352–2367, 2003.

[14] W. van Ackooij and A. Frangioni. Incremental bundle methods using upper models. Technical report, Dipartimento di Informatica, Universitá di Pisa, 2016.

[15] H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors. *Handbook of Semidefinite Programming*, volume 27 of *International Series in Operations Research and Management Science*. Kluwer Academic Publishers, Boston/Dordrecht/London, 2000.