Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

C. HELMBERG[1] AND K.C. KIWIEL[2]

A Spectral Bundle Method with Bounds

Revised September 26, 2001

To appear in Mathematical Programming

[1]Konrad-Zuse-Zentrum für Informationstechnik Berlin, Takustraße 7, D-14195 Berlin, Germany, helmberg@zib.de, http://www.zib.de/helmberg

[2]Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland, kiwiel@ibspan.waw.pl

Preprint SC 99-37 (December 1999)

A Spectral Bundle Method with Bounds*

C. Helmberg† K.C. Kiwiel‡

December 22, 1999
revised September 26, 2001

Abstract

Semidefinite relaxations of quadratic 0-1 programming or graph partitioning problems are well known to be of high quality. However, solving them by primal-dual interior point methods can take much time even for problems of moderate size. The recent spectral bundle method of Helmberg and Rendl can solve quite efficiently large structured equality-constrained semidefinite programs if the trace of the primal matrix variable is fixed, as happens in many applications. We extend the method so that it can handle inequality constraints without seriously increasing computation time. In addition, we introduce inexact null steps. This abolishes the need of computing exact eigenvectors for subgradients, which brings along significant advantages in theory and in practice. Encouraging preliminary computational results are reported.

Key Words. Eigenvalue optimization, convex optimization, semidefinite programming, proximal bundle method, large-scale problems.
MSC 1991. 65F15, 90C25; Secondary 52A41, 90C06.

1 Introduction

It is well known [22, 10, 23, 8] that semidefinite programming allows to design powerful relaxations for quadratic 0-1 programming and graph partitioning problems. To describe such relaxations, let $\langle A, B\rangle := \operatorname{tr} A^T B = \sum a_{ij} b_{ij}$ denote the inner product of matrices $A, B \in \mathbb{R}^{m\times n}$, let $\mathcal{S}^n$ denote the space of symmetric matrices of order n, and $\mathcal{S}^n_+ := \{A \in \mathcal{S}^n : A \succeq 0\}$ its cone of positive semidefinite matrices. Let $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ be a linear operator and $\mathcal{A}^T : \mathbb{R}^m \to \mathcal{S}^n$ its adjoint, defined by $\langle \mathcal{A}X, y\rangle = \left\langle X, \mathcal{A}^T y\right\rangle \; \forall X \in \mathcal{S}^n, y \in \mathbb{R}^m$,

*A preliminary version of this paper appeared in the proceedings of IPCO '98 [13].

†Konrad-Zuse-Zentrum für Informationstechnik Berlin, Takustraße 7, D-14195 Berlin, Germany, helmberg@zib.de, http://www.zib.de/helmberg

‡Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland, kiwiel@ibspan.waw.pl

having the form

$$\mathcal{A}X = \begin{bmatrix} \langle A_1, X\rangle \\ \vdots \\ \langle A_m, X\rangle \end{bmatrix} \quad \text{and} \quad \mathcal{A}^T y = \sum_{i=1}^{m} y_i A_i,$$

with given matrices $A_i \in \mathcal{S}^n$. For a fixed $\mathcal{I} \subset \{1,\dots,m\}$, $\mathcal{A}_\mathcal{I}$ and $y_\mathcal{I}$ refer to rows $\mathcal{I}$ of $\mathcal{A}$ and $y \in \mathbb{R}^m$, respectively. Finally, let $\bar{\mathcal{I}} := \{1,\dots,m\} \setminus \mathcal{I}$. Then the primal-dual pair of semidefinite programs is written in standard form as

$$\text{(P)} \qquad \begin{array}{ll} \max & \langle C, X\rangle\,, \\ \text{s.t.} & \mathcal{A}_{\bar{\mathcal{I}}} X = b_{\bar{\mathcal{I}}}, \\ & \mathcal{A}_\mathcal{I} X \le b_\mathcal{I}, \\ & X \succeq 0, \end{array} \qquad \text{(D)} \qquad \begin{array}{ll} \min & \langle b, y\rangle\,, \\ \text{s.t.} & Z = \mathcal{A}^T y - C, \\ & Z \succeq 0, y_\mathcal{I} \ge 0, \end{array}$$

where the cost matrix $C \in \mathcal{S}^n$ and the right-hand side vector $b \in \mathbb{R}^m$ are given. In most applications, A_i are (sparse) matrices of rank one or two and C is sparse.

Primal-dual interior point methods [1, 17, 20, 25], which are most commonly used for solving problem (P), offer few opportunities to exploit its structure. Their work per iteration is typically dominated by the factorization of a dense symmetric positive definite matrix of order m, and by one or more factorizations of the variable X during the line search. The solution times for problems with $m \ge 5000$ or $n \ge 500$, say, are prohibitive. The recent purely dual approach of [3](see also [2]) is able to exploit the sparsity of C and A_i for the dual variable Z, but in general still needs to factorize a dense positive definite matrix of order m.

The alternative approach of [16] transforms (D) into the eigenvalue optimization problem

$$\min_{y\in\mathbb{R}^m, y_\mathcal{I}\ge 0} a\lambda_{\max}(C - \mathcal{A}^T y) + \langle b, y\rangle\,, \tag{1.1}$$

where $\lambda_{\max}(\cdot)$ is the maximum eigenvalue of $\cdot$ and $a > 0$ is such that $\mathcal{A}_{\bar{\mathcal{I}}} X = b_{\bar{\mathcal{I}}}$ implies $\operatorname{tr} X = a$ (as happens in most combinatorial applications). The spectral bundle method [16] is tailored to the structure of (1.1) with $\mathcal{I} = \emptyset$. In contrast to standard bundle methods [19, 30], it constructs a non-polyhedral semidefinite cutting surface model of the objective by accumulating eigenvectors instead of subgradients. Minimization of the model augmented with a regularizing quadratic term yields the next iterate. The model is enriched by one or more eigenvectors corresponding to the largest eigenvalues at this iterate, computed efficiently by Lanczos methods (see, e.g., [9]) via matrix-vector multiplications that do not require the matrix in explicit form. Thus structural properties of C and A_i can be exploited. Aggregation of the past eigenvector information ensures bounded storage and efficient solvability of consecutive subproblems even for large m and n.

So far the spectral bundle method could handle only equality constraints in (P), i.e., $\mathcal{I} = \emptyset$, because sign constraints on the dual variables y might increase significantly the solution times for its semidefinite subproblems. In this paper we employ Lagrangian relaxation to approximate the solution of sign constrained subproblems. Surprisingly, just one update of Lagrange multipliers per subproblem suffices to ensure convergence. The subproblems can be solved as efficiently as in the unconstrained case, thus rendering our method an attractive choice for large-scale semidefinite cutting-plane algorithms.

The Lanczos process approximates the maximal eigenvalue iteratively from below. If the Lanczos vector providing the current lower bound suffices to improve the bundle model sufficiently, the process may be stopped prematurely. First numerical experiments based on this idea were presented in [14] and exhibited significant improvements on the performance reported in [16]. Here, we provide a proper theoretical foundation for this approach. By abolishing the need of computing exact eigenvectors in theory we ensure practical implementability. Our revised implementation following the theory is equally efficient.

Like most first-order methods, the spectral bundle method exhibits fast progress in the beginning, but shows a strong tailing-off effect as the optimal solution is approached. Fortunately, many semidefinite relaxations do not have to be solved exactly. Rather, an approximate solution is used to improve the current relaxation (e.g., by cutting planes), which is then resolved.

Since the first writing of the paper, promising results have been published for nonlinear approaches. These are based on factorizing the positive semidefinite variables and working with the factors as new variables [5, 6, 4, 32, 7]. Regarding the spectral bundle method various alternatives for formulating and updating the bundle have been compared in [24], some exhibited remarkable improvements in computation time. A C++-implementation of the spectral bundle method along the lines of this paper has recently been made available [12], possible approaches for using it in a cutting plane framework are described in [11].

The paper is organized as follows. Section 2 gives our extension of the spectral bundle method to inequality constraints. Convergence of our method is established in §3, and its modifications are described in §4. Section 5 discusses efficiency of the subproblem solution. Finally, §6 gives computational results. The paper is rather heavy in notation. For the convenience of the reader we provide an appendix listing the most important objects and definitions.

2 Extension of the spectral bundle method

In order to simplify notation we assume that $a = 1$ and $\mathcal{I} = \{1, \ldots, m\}$ in problem (1.1). In addition, we employ the indicator function $\imath_Y$ of $Y := \mathbb{R}^m_+$ ($\imath_Y(y) = 0$ if $y \in Y$, ∞ otherwise) to extend the definition of the objective function to $\mathbb{R}^m$,

$$\min_{y \in \mathbb{R}^m} \Big\{ f(y) := \lambda_{\max}(C - \mathcal{A}^T y) + \langle b, y \rangle + \imath_Y(y) \Big\}. \tag{2.1}$$

We start by giving a dual formulation of the "nontrivial" components $\lambda_{\max}(\cdot)$ and $\imath_Y$ of f and use this to derive a family of convex minorants as well as the subdifferential of f.

Concerning the maximum eigenvalue function $\lambda_{\max}(\cdot)$, we note that, with

$$\mathcal{W} := \Big\{ W \in \mathcal{S}^n_+ : \operatorname{tr} W = 1 \Big\} = \operatorname{conv} \Big\{ vv^T : \|v\| = 1 \Big\} \tag{2.2}$$

for any $X \in \mathcal{S}^n$, the maximum eigenvalue (v. [26, Thm 1])

$$\lambda_{\max}(X) = \max \{ \langle X, W \rangle : W \in \mathcal{W} \} = \max \Big\{ v^T X v : \|v\| = 1 \Big\}, \tag{2.3}$$

can be computed by finding its maximizing (normalized) eigenvector v. Thus, the function

$$\lambda_{\max}(C - \mathcal{A}^T y) = \max\left\{ \left\langle C - \mathcal{A}^T y, W \right\rangle : W \in \mathcal{W} \right\} \tag{2.4}$$

is a convex finite-valued (and hence continuous) max-type function. Using (2.2) and (2.3) it can be worked out that the set of maximizers in (2.4) is the convex hull of the dyadic products vv^T of all normalized eigenvectors v to the maximum eigenvalue of $C - \mathcal{A}^T y$.

Likewise, the indicator function $\imath_Y$ can be described in dual form by

$$\imath_Y(y) = \sup\{ -\eta^T y : \eta \in \mathbb{R}^m_+ \}. \tag{2.5}$$

Combining (2.3) and (2.5) we may rewrite f of (2.1) as

$$f(y) = \sup\{ \left\langle C - \mathcal{A}^T y, W \right\rangle + \langle b - \eta, y \rangle : (W, \eta) \in \mathcal{W} \times \mathbb{R}^m_+ \}. \tag{2.6}$$

Therefore, any particular choice of $W \in \mathcal{W}$ and $\eta \in \mathbb{R}^m_+$ gives rise to an affine function that minorizes f on $\mathbb{R}^m$,

$$f_{W,\eta}(y) := \langle C, W \rangle + \langle b - \eta - \mathcal{A}W, y \rangle \leq f(y) \qquad \forall y \in \mathbb{R}^m. \tag{2.7}$$

Further convex minorants of f are conveniently defined by taking the pointwise supremum over a family of affine functions $f_{W,\eta}$, where the family is described by subsets $\widehat{\mathcal{W}} \subseteq \mathcal{W}$ and $\widehat{Y} \subseteq \mathbb{R}^m_+$,

$$f_{\widehat{\mathcal{W}},\widehat{Y}}(y) := \sup\left\{ f_{W,\eta}(y) : (W, \eta) \in \widehat{\mathcal{W}} \times \widehat{Y} \right\} \leq f(y) \qquad \forall y \in \mathbb{R}^m. \tag{2.8}$$

Instead of $f_{\{W\},\hat{Y}}$ and $f_{\widehat{\mathcal{W}},\{\eta\}}$ we often write $f_{W,\hat{Y}}$ and $f_{\widehat{\mathcal{W}},\eta}$, respectively. We will make heavy use of this notation in the following. It is worth to get acquainted with it by verifying that $f_{\mathcal{W},Y} = f$, $f_{\mathcal{W},0}(y) = f(y)$ for $y \in \mathbb{R}^m_+$, and that on the domain of f the subdifferential of f is described by

$$\partial f(y) = \left\{ \nabla f_{W,\eta} = b - \eta - \mathcal{A}W : f_{W,\eta}(y) = f(y), (W, \eta) \in \mathcal{W} \times \mathbb{R}^m_+ \right\} \quad \forall y \in \mathbb{R}^m_+. \tag{2.9}$$

We turn to the description of our algorithmic approach. The nonsmooth convex minimization problem (2.1) could be solved over $y \in \mathbb{R}^m_+$ by the proximal bundle method [19]. We first sketch this method to introduce some useful concepts.

The method generates a *trial point* $y^+ := \operatorname{argmin}_Y \hat{f}(\cdot) + \frac{u}{2}\| \cdot - \hat{y}\|^2$ from the *current iterate* $\hat{y}$, where $\hat{f}$ is an accumulated cutting plane *model* of f and the *weight* $u > 0$ keeps y^+ near $\hat{y}$. A *descent step* $\hat{y}^+ = y^+$ occurs if $f(\hat{y}) - f(y_+) \geq \kappa[f(\hat{y}) - \hat{f}(y^+)]$, where $\kappa \in (0, 1)$. Otherwise a *null step* $\hat{y}^+ = \hat{y}$ is made but the next model is improved with a new cutting plane computed at y^+.

Our spectral bundle method is tailored for the eigenvalue problem (2.1). Instead of the usual polyhedral $\hat{f}$, it uses a *semidefinite cutting surface model* $f_{\widehat{\mathcal{W}},0}$ of f (v. (2.8)) where $\widehat{\mathcal{W}} \neq \emptyset$ is an appropriate closed convex subset of $\mathcal{W}$ (v. (2.2)). Since $\mathcal{W}$ is compact, $\widehat{\mathcal{W}}$ is compact as well, and for $\hat{Y} = \{0\}$ the sup in (2.8) turns into a finite-valued max, so $f_{\widehat{\mathcal{W}},0}$ makes sense.

For large m, the usual bundle subproblem of finding the standard trial point

$$\hat{y}^* := \operatorname{argmin}\left\{ f_{\widehat{\mathcal{W}},Y}(\cdot) + \tfrac{u}{2}\|\cdot - \hat{y}\|^2 \right\} = \operatorname{argmin}\left\{ f_{\widehat{\mathcal{W}},0}(y) + \tfrac{u}{2}\|y - \hat{y}\|^2 : y \geq 0 \right\} \tag{2.10}$$

is expensive to solve by interior point methods because of the constraints on y (v. §5). It is much easier to find an approximation y^+ to $\hat{y}^*$ via Lagrangian relaxation, i.e., we introduce the Lagrange multipliers η of (2.5) for the sign constraints on y and define the *Lagrangian* $L : \mathbb{R}^m \times \widehat{\mathcal{W}} \times \mathbb{R}^m_+ \to \mathbb{R}$ of (2.10) by

$$L(y; W, \eta) := f_{W,\eta}(y) + \tfrac{u}{2}\|y - \hat{y}\|^2 = \langle C, W\rangle + \langle b - \eta - \mathcal{A}W, y\rangle + \tfrac{u}{2}\|y - \hat{y}\|^2. \tag{2.11}$$

The objective function of (2.10) may be expressed as

$$f_{\widehat{\mathcal{W}},Y}(\cdot) + \tfrac{u}{2}\|\cdot - \hat{y}\|^2 = \sup\left\{ L(\cdot; W, \eta) : (W, \eta) \in \widehat{\mathcal{W}} \times \mathbb{R}^m_+ \right\}. \tag{2.12}$$

For each fixed y the primal function is obtained by taking the supremum of $L(y; \cdot, \cdot)$ over W and η. The (Lagrangian) *dual function* $\psi : \widehat{\mathcal{W}} \times \mathbb{R}^m_+ \to \mathbb{R}$ is determined, for each fixed pair (W, η), by minimizing $L(\cdot; W, \eta)$ over $y \in \mathbb{R}^m$. Looking at the right hand side of (2.11) we see that the latter optimization problem is strictly convex and quadratic. Thus, for each fixed pair (W, η) the (unique) optimal y can be computed explicitly,

$$y_{W,\eta} := \operatorname{argmin} L(\cdot; W, \eta) = \operatorname{argmin}\left\{ f_{W,\eta}(\cdot) + \tfrac{u}{2}\|\cdot - \hat{y}\|^2 \right\} = \hat{y} - \tfrac{1}{u}(b - \eta - \mathcal{A}W). \tag{2.13}$$

Substituting $y_{W,\eta}$ for y into the right hand side of (2.11) we obtain the dual function ψ,

$$\psi(W, \eta) := \min L(\cdot; W, \eta) = \langle C, W\rangle + \langle b - \mathcal{A}W - \eta, \hat{y}\rangle - \tfrac{1}{2u}\|b - \mathcal{A}W - \eta\|^2. \tag{2.14}$$

One can show via standard saddle-point arguments (v. [29, Thm 37.6]) that each maximizing argument (W^*, η^*) of ψ (these are not necessarily unique) yields the optimal solution of (2.10) via

$$\hat{y}^* = y_{W^*,\eta^*} \quad \text{for each} \quad (W^*, \eta^*) \in \operatorname{Arg}\max_{\widehat{\mathcal{W}}\times\mathbf{R}^m_+} \psi \neq \emptyset. \tag{2.15}$$

In the algorithm there is no need to solve (2.10) exactly. In fact, already a very rough approximation (W^+, η^+) to (W^*, η^*) of (2.15) gives rise to a candidate $y^+ = y_{W^+,\eta^+}$ that ensures convergence. We determine this approximation (W^+, η^+) in a Gauss-Seidel fashion as follows. First fix a multiplier $\hat{\eta} \in \mathbb{R}^m_+$ and find

$$W^+ \in \operatorname{Argmax}\left\{ \psi(W, \hat{\eta}) : W \in \widehat{\mathcal{W}} \right\} \tag{2.16}$$

via an interior point algorithm (v. §5), and then maximize $\psi(W^+, \cdot)$ over $\eta \in \mathbb{R}^m_+$. By (2.14) the latter problem is separable concave, i.e., it can be solved for each coordinate of η separately,

$$\eta^+ := \operatorname{argmax}\left\{ \psi(W^+, \eta) : \eta \geq 0 \right\} = \max\left\{ 0, -u\hat{y} + b - \mathcal{A}W^+ \right\}. \tag{2.17}$$

Inserting the right hand side of (2.17) in the right hand side of (2.13), it is not difficult to check that the corresponding primal point y_{W^+,η^+} is both feasible and complementary:

$$y^+ := y_{W^+,\eta^+} = \max\left\{ \hat{y} - \tfrac{1}{u}(b - \mathcal{A}W^+), 0 \right\} \geq 0, \tag{2.18}$$

$$\left\langle \eta^+, y^+ \right\rangle = 0. \tag{2.19}$$

Remark 2.1 *The two "coordinatewise" maximization steps may also be interpreted as two separate steps in a proximal bundle approach. Indeed, in Lem. 3.1 we show that the first step* (2.16) *corresponds to solving the usual bundle subproblem for the cutting surface model* $f_{\widehat{\mathcal{W}},\hat{\eta}}$ *(v.* (2.8)*), yielding an intermediate candidate (v.* (2.13)*)*

$$y^{+\frac{1}{2}} = y_{W^+,\hat{\eta}}. \tag{2.20}$$

In the second step the cutting surface model is updated to $f_{W^+,Y}$ *(v.* (2.8)*). This model includes the previous optimizing cutting plane* $f_{W^+,\hat{\eta}}$ *(v.* (2.7)*) and numerous new minorants of* f *that improve the model in* $y^{+\frac{1}{2}}$ *if* $y^{+\frac{1}{2}} \notin \mathbb{R}^m_+$*. The optimal* η^+ *of* (2.17) *gives rise to the solution* y^+ *(v.* (2.18)*) of the bundle subproblem over this second cutting surface model (v. Lem. 3.1).*

For practical reasons we state the algorithm with an inner loop allowing for several repetitions of the two "coordinatewise" maximization steps (2.16) and (2.17).

Algorithm 2.2
Input: $y^0 \in \mathbb{R}^m_+$, $\varepsilon_{\rm opt} \geq 0$, $\kappa_{\rm M} \in (0,\infty]$, $\kappa \in (0,1)$, a weight $u > 0$.

Step 0 (Initialization). Set $k = 0$, $\hat{y}^0 = y^0$, $\eta^0 = 0$, $f(\hat{y}^0)$ and $\widehat{\mathcal{W}}^0$ (v. §5).

Step 1 (Trial point finding). Set $\hat{y} = \hat{y}^k$, $\widehat{\mathcal{W}} = \widehat{\mathcal{W}}^k$, $\hat{\eta} = \eta^k$.

(a) Find $W^+ \in \mathrm{Argmax}_{W\in\widehat{\mathcal{W}}}\, \psi(W,\hat{\eta})$ (v. (2.14)) and set $y^{+\frac{1}{2}} = y_{W^+,\hat{\eta}}$ (v. (2.13)).

(b) Set $\eta^+ = \mathrm{argmax}_{\eta\geq 0}\, \psi(W^+,\eta)$ (v. (2.17)) and $y^+ = y_{W^+,\eta^+}$ (feasible by (2.18)).

(c) (Stopping criterion) If $f(\hat{y}) - f_{W^+,\eta^+}(y^+) \leq \varepsilon_{\rm opt}(|f(\hat{y})|+1)$, then STOP.

(d) If $f_{\widehat{\mathcal{W}},0}(y^+) - f_{W^+,\eta^+}(y^+) > \kappa_{\rm M}[f(\hat{y}) - f_{W^+,\eta^+}(y^+)]$, then set $\hat{\eta} = \eta^+$ and go to (a).

(e) Set $y^{k+1} = y^+$, $W^{k+1} = W^+$, and $\eta^{k+1} = \eta^+$.

Step 2 (Evaluation). Find $W_S^{k+1} \in \mathrm{Argmax}_{\mathcal{W}} \left\langle C - \mathcal{A}^T y^{k+1}, \cdot \right\rangle$ and $f(y^{k+1})$ (v. (2.4),(2.1)).

Step 3 (Descent test). If $f(\hat{y}^k) - f(y^{k+1}) \geq \kappa[f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(y^{k+1})]$ then set $\hat{y}^{k+1} = y^{k+1}$ (descent step); otherwise set $\hat{y}^{k+1} = \hat{y}^k$ (null step).

Step 4 (Model updating). Choose a closed convex $\widehat{\mathcal{W}}^{k+1} \supset \left\{W^{k+1}, W_S^{k+1}\right\}$ (v. §5).

Step 5. Increase k by 1 and go to Step 2.

Remarks 2.3 (i) The stopping criterion in Step 1(c) is motivated as follows. The gradient of the affine minorant f_{W^+,η^+} (v. (2.7)) of f satisfies (v. (2.18), (2.13))

$$\nabla f_{W^+,\eta^+} = b - \eta^+ - \mathcal{A}W^+ = u(\hat{y} - y^+). \tag{2.21}$$

Therefore the values of f_{W^+,η^+} at $\hat{y}$ and y^+ are related by

$$f_{W^+,\eta^+}(y^+) = f_{W^+,\eta^+}(\hat{y}) + \left\langle \nabla f_{W^+,\eta^+}, y^+ - \hat{y} \right\rangle = f_{W^+,\eta^+}(\hat{y}) - u\|y^+ - \hat{y}\|^2.$$

With $f_{W^+,\eta^+} \leq f$ (v. (2.7)) we obtain

$$f(\hat{y}) - f_{W^+,\eta^+}(y^+) = f(\hat{y}) - f_{W^+,\eta^+}(\hat{y}) + u\|y^+ - \hat{y}\|^2 \geq f(\hat{y}) - f_{W^+,\eta^+}(\hat{y}) \geq 0, \quad (2.22)$$

$$f(\cdot) \geq f_{W^+,\eta^+}(\hat{y}) + \langle \nabla f_{W^+,\eta^+}, \cdot - \hat{y}\rangle = f(\hat{y}) - [\, f(\hat{y}) - f_{W^+,\eta^+}(\hat{y})\,] + u\left\langle \hat{y} - y^+, \cdot - \hat{y}\right\rangle. \quad (2.23)$$

Thus, a small value of $f(\hat{y}) - f_{W^+,\eta^+}(y^+)$ and (2.22) imply that $f(\hat{y}) - f_{W^+,\eta^+}(\hat{y})$ and $\|y^+ - \hat{y}\|$ are both small and so $\hat{y}$ is approximately optimal by (2.23).

(ii) We show in Lemma 3.2(c) below that the inner iterations of Step 1 make y^+ approach the solution $\hat{y}^*$ of the bundle subproblem (v. (2.10)). The criterion of step 1(d) for executing additional inner iterations is based on the following considerations. Since the candidate y^+ is feasible (v. (2.18)), the value of the cutting surface model satisfies $f_{\widehat{\mathcal{W}},Y}(y^+) = f_{\widehat{\mathcal{W}},0}(y^+)$ (v. (2.8)). The affine function f_{W^+,η^+} minorizes $f_{\widehat{\mathcal{W}},Y}$ (use $(W^+,\eta^+) \in \widehat{\mathcal{W}} \times \mathbb{R}^n_+$ in (2.8)) and together with $f_{W^+,0}(y^+) = f_{W^+,\eta^+}(y^+)$ ($\langle \eta^+, y^+\rangle = 0$ by (2.18)) we obtain

$$f_{\widehat{\mathcal{W}},Y}(y^+) = f_{\widehat{\mathcal{W}},0}(y^+) \geq f_{W^+,0}(y^+) = f_{W^+,\eta^+}(y^+).$$

If equality holds then (y^+, W^+, η^+) is a saddle-point of L (v. (2.15)) and, thus, $y^+ = \hat{y}^*$ (v. (2.10)). In general, however, W^+ will not be the maximizing element of $\widehat{\mathcal{W}}$ for $f_{\widehat{\mathcal{W}},\eta^+}(y^+)$ (cf. (2.8)). If the gap $f_{\widehat{\mathcal{W}},0}(y^+) - f_{W^+,\eta^+}(y^+)$ between cutting plane model and its approximation is "large" relative to the gap $f(\hat{y}) - f_{W^+,\eta^+}(y^+)$ then we consider the approximate solution (W^+,η^+) as too inaccurate and improve its quality by further inner iterations. The parameter κ_M allows to specify what "large" means, but even for arbitrarily large gaps convergence is guaranteed. Indeed, for $\kappa_{\mathrm{M}} = \infty$ just one inner iteration is performed, but convergence is ensured via Lemma 3.2(a).

(iii) Evaluation step 2 may be implemented by computing the maximum eigenvalue $\lambda_{\max}(C - \mathcal{A}^T y^{k+1})$ and a corresponding (normalized) eigenvector v, e.g., by the Lanczos method (see, e.g., [9]). Then $W_S^{k+1} = vv^T$ is an appropriate choice; we use subscript S to indicate that W_S gives rise to a new subgradient of f in y^+ (v. (2.9),(2.4)). Since it would be difficult to compute an exact eigenvector v, we show in §4.1 that a modification of the algorithm allows the use of approximations.

3 Convergence

We first exhibit useful saddle-point relations at steps 1(a) and 1(b). In terms of cutting surface models they state that $y^{+\frac{1}{2}}$ of (2.20) is the optimal solution of the bundle subproblem for $f_{\widehat{\mathcal{W}},\hat{\eta}}$ (and W^+ is an optimal multiplier in $\widehat{\mathcal{W}}$) and that y^+ of (2.18) is the optimal solution of the bundle subproblem for $f_{W^+,Y}$ (and η^+ is the optimal multiplier in $\mathbb{R}^m_+$).

Lemma 3.1 *For some fixed $\hat{\eta} \in \mathbb{R}^m_+$ and closed convex $\widehat{\mathcal{W}} \subset \mathcal{W}$ (v. (2.2)) let W^+, η^+, y^+, and $y^{+\frac{1}{2}}$ be determined by* (2.16), (2.17), (2.18), *and* (2.20), *respectively. Then for L of* (2.11) *the following relations hold*:

$$L(y^{+\frac{1}{2}}; W, \hat{\eta}) \leq L(y^{+\frac{1}{2}}; W^+, \hat{\eta}) \leq L(y; W^+, \hat{\eta}) \quad \forall y \in \mathbb{R}^m, W \in \widehat{\mathcal{W}}, \quad (3.1a)$$

$$L(y^+; W^+, \eta) \leq L(y^+; W^+, \eta^+) \leq L(y; W^+, \eta^+) \quad \forall y \in \mathbb{R}^m, \eta \in \mathbb{R}^m_+. \quad (3.1b)$$

Proof. In order to simplify notation, let $L_{+\frac{1}{2}}(\cdot;\cdot) := L(\cdot;\cdot,\hat{\eta})$ on $\mathbb{R}^m \times \widehat{\mathcal{W}}$. $L_{+\frac{1}{2}}$ is convex-concave, $\widehat{\mathcal{W}}$ is compact convex, and $\forall W \in \widehat{\mathcal{W}}$, $L(y;W,\hat{\eta}) \to \infty$ when $\|y\| \to \infty$. Hence $L_{+\frac{1}{2}}$ has a saddle-point $(\bar{y}, \overline{W})$ [18, Thm VII.4.3.1]. Since (v. Step 1(a),(2.14)) $W^+ \in \operatorname{Argmax}_{\widehat{\mathcal{W}}} \min_y L(y;\cdot)$, $(\bar{y}, W^+)$ is a saddle-point as well [18, Thm VII.4.2.5]. Then $L_{+\frac{1}{2}}(\bar{y};W^+) \le L_{+\frac{1}{2}}(y;W^+)$ $\forall y$ yields $\bar{y} = y_{W^+,\hat{\eta}} = y^{+\frac{1}{2}}$ (v. (2.13),(2.20)), so (3.1a) holds. (3.1b) follows from standard convex quadratic programming duality. □

An abstract algorithmic framework for minimizing a convex function f is the proximal point algorithm (see, e.g., [28]); it moves iteratively from $\hat{y}$ to the so called proximal point $\bar{y}^* := \operatorname{argmin} f(\cdot) + \|\cdot - \hat{y}\|^2$. As observed in [19], skipping the descent step test of the proximal bundle method yields an infinite sequence of null steps whose iterates converge to the proximal point, while simultaneously the values of the cutting plane models at the iterates converge to the function value in the proximal point; this setting is also referred to as an ideal bundle prox iteration. Our algorithm builds on the proximal bundle method and exhibits the same behavior. The following lemma states that for an infinite sequence of null steps the y^k converge to the proximal point of f, whereas for an infinite sequence of inner iterations within the same outer iteration the inner iterates converge to the proximal point of the model $f_{\widehat{\mathcal{W}},Y}$. In the latter case, the proximal point is $\hat{y}$ and, at the same time, an optimal solution of the minimization problem.

Lemma 3.2 *Suppose that, starting from some iteration $\hat{k} \ge 0$, the descent test of Step 3 is omitted and only null steps $\hat{y}^{k+1} = \hat{y}^{\hat{k}} =: \hat{y}$ can be made for $k \ge \hat{k}$. Then:*

(a) *If $\kappa_{\mathrm{M}} = \infty$ in the test of Step 1(d) and $k \to \infty$, then*

$$\epsilon^k := f(y^k) - f_{W^k,\eta^k}(y^k) \to 0, \tag{3.2}$$

$$y^k \to y^\infty := \operatorname{argmin}\left\{ f(\cdot) + \tfrac{u}{2}\|\cdot - \hat{y}\|^2 \right\}. \tag{3.3}$$

(b) *If $\kappa_{\mathrm{M}} < \infty$ and $k \to \infty$, then (3.2)–(3.3) hold.*

(c) *If an infinite loop occurs in Step 1(a–d) for $k = \hat{k}$, then at Step 1(d)*

$$\epsilon^+ := f_{\widehat{\mathcal{W}}^k,Y}(y^+) - f_{W^+,\eta^+}(y^+) \to 0, \tag{3.4}$$

$$y^+ \to \hat{y} = \operatorname{argmin}\left\{ f_{\widehat{\mathcal{W}}^k,Y}(\cdot) + \tfrac{u}{2}\|\cdot - \hat{y}\|^2 \right\}. \tag{3.5}$$

Further, in this case $\hat{y}^k = \hat{y} \in \operatorname{Argmin} f$ and $f(\hat{y}) - f_{W^+,\eta^+}(y^+) \to 0$.

Proof. (a) For $k > \hat{k}$, just one inner iteration occurs at Step 1 (since $\kappa_{\mathrm{M}} = \infty$). At Step 1(e) of iteration k we have $\eta^k = \hat{\eta}$, $W^{k+1} = W^+$ (v. (2.16)), an intermediate $y^{k+\frac{1}{2}} := y^{+\frac{1}{2}} = y_{W^{k+1},\eta^k}$, the new multiplier $\eta^{k+1} = \eta^+$ (v. (2.17)), and the new feasible candidate $y^{k+1} = y^+ = y_{W^{k+1},\eta^{k+1}}$ (v. (2.18)). We first trace the development of L (v. (2.11)) as the algorithm "moves" from (y^k, W^k, η^k) to $(y^{k+\frac{1}{2}}, W^{k+1}, \eta^k)$ to $(y^{k+1}, W^{k+1}, \eta^{k+1})$. To this end observe that

$$L(\cdot;W,\eta) = L(y_{W,\eta};W,\eta) + \frac{u}{2}\|\cdot - y_{W,\eta}\|^2, \tag{3.6}$$

because $L(\cdot; W, \eta)$ is convex quadratic in y and $y_{W,\eta}$ is its minimizer. We obtain

$$\begin{aligned}
L(y^{k+\frac{1}{2}}; W^{k+1}, \eta^k) &= \max_{W \in \widehat{\mathcal{W}}^k} L(y^{k+\frac{1}{2}}; W, \eta^k) && [(3.1a)] \\
&\geq L(y^{k+\frac{1}{2}}; W^k, \eta^k) && [W^k \in \widehat{\mathcal{W}}^k \text{ by Step 4}] \qquad (3.7a) \\
&= L(y^k; W^k, \eta^k) + \tfrac{u}{2}\|y^{k+\frac{1}{2}} - y^k\|^2 && [y^k = y_{W^k,\eta^k} \text{ and } (3.6)] \\
L(y^{k+1}; W^{k+1}, \eta^{k+1}) &= \max_{\eta \in \mathbb{R}^m_+} L(y^{k+1}; W^{k+1}, \eta) && [(3.1b)] \\
&\geq L(y^{k+1}; W^{k+1}, \eta^k) && [\eta^k \in \mathbb{R}^m_+] \qquad (3.7b) \\
&= L(y^{k+\frac{1}{2}}; W^{k+1}, \eta^k) + \tfrac{u}{2}\|y^{k+1} - y^{k+\frac{1}{2}}\|^2 && [(3.6)].
\end{aligned}$$

So in each intermediate step L increases by at least $\frac{u}{2}$ times the squared norm of the change of y. But, for all $k \geq \hat{k}$, $L(y^k; W^k, \eta^k)$ is bounded above by $f(\hat{y})$, because

$$\begin{aligned}
L(y^k; W^k, \eta^k) &\leq L(\hat{y}; W^k, \eta^k) && [y^k = y_{W^k,\eta^k} \text{ minimizes } L(\cdot; W^k, \eta^k), \text{ v. } (2.13)] \\
&= f_{W^k,\eta^k}(\hat{y}) && [\text{definition } (2.11) \text{ of } L] \\
&\leq f(\hat{y}) && [\text{by } (2.7)].
\end{aligned}$$

Therefore, there exists $\xi \leq f(\hat{y})$ such that

$$L(y^k; W^k, \eta^k), L(y^{k+\frac{1}{2}}; W^{k+1}, \eta^k) \to \xi \quad \text{and} \quad y^{k+\frac{1}{2}} - y^k, y^{k+1} - y^{k+\frac{1}{2}} \to 0. \qquad (3.8)$$

Informally, the next step of the proof has the following rational. Because the changes in y go to zero, the subgradient added in Step 2/Step 4 provides "enough local support" to force $f_{W^k,\eta^k}(y^k)$ towards $f(y^k)$ (this works, because the subgradients employed have bounded norm by construction). We make this mathematically precise.

In Step 2 we determine W^k_S satisfying

$$f(y^k) = f_{W^k_S,0}(y^k) = f_{W^k_S,\eta^k}(y^k) \geq f_{W^k,\eta^k}(y^k), \qquad (3.9)$$

where the first equation is due to (2.4) and the definition (2.7) of $f_{W,\eta}$, the second equation follows from $\left\langle \eta^k, y^k \right\rangle = 0$ (v. (2.18)), and the inequality stems from the fact that f_{W^k,η^k} minorizes f (v. (2.7)). Since $f(y^k) = f_{W^k_S,\eta^k}(y^k)$, the gradient of the affine function $f_{W^k_S,\eta^k}$ is a subgradient of f at y^k (v. (2.9)); we denote it by

$$g^k := \nabla f_{W^k_S,\eta^k} = b - \eta^k - \mathcal{A} W^k_S. \qquad (3.10)$$

By Step 4, W^k_S is contained in $\widehat{\mathcal{W}}^k$; thus, the cutting plane $f_{W^k_S,\eta^k}$ minorizes the intermediate model $f_{\widehat{\mathcal{W}}^k,\eta^k}$ (v. (2.8), (3.1a) with $\hat{\eta} = \eta^k$),

$$f_{W^k_S,\eta^k}(y^{k+\frac{1}{2}}) \leq \max_{W \in \widehat{\mathcal{W}}^k} f_{W,\eta^k}(y^{k+\frac{1}{2}}) = f_{W^{k+1},\eta^k}(y^{k+\frac{1}{2}}). \qquad (3.11)$$

This allows to bound ϵ^k from above (≥ 0 follows from (3.9)),

$$\begin{aligned}
\epsilon^k &= f_{W^k_S,\eta^k}(y^k) - f_{W^k,\eta^k}(y^k) && [\text{def. } (3.2) \text{ of } \epsilon^k, (3.9)] \\
&= f_{W^k_S,\eta^k}(y^{k+\frac{1}{2}}) - f_{W^k,\eta^k}(y^k) + \left\langle g^k, y^k - y^{k+\frac{1}{2}} \right\rangle && [f_{W^k_S,\eta^k} \text{ is affine}, (3.10)] \\
&\leq f_{W^{k+1},\eta^k}(y^{k+\frac{1}{2}}) - f_{W^k,\eta^k}(y^k) + \|g^k\| \, \|y^{k+\frac{1}{2}} - y^k\| && [(3.11), \text{Cauchy-Schwarz}] \qquad (3.12) \\
&= L(y^{k+\frac{1}{2}}; W^{k+1}, \eta^k) - L(y^k; W^k, \eta^k) - \tfrac{u}{2}\|y^{k+\frac{1}{2}} - \hat{y}\|^2 + \tfrac{u}{2}\|y^k - \hat{y}\|^2 \\
&\qquad + \|g^k\| \|y^{k+\frac{1}{2}} - y^k\| && [\text{def. } (2.11) \text{ of } L].
\end{aligned}$$

Now, $W_S^k, W^k \in \mathcal{W}$ are bounded (and so is $\mathcal{W}$, v. (2.2)); so by (2.17) $\|\eta^k\| \leq \|u\hat{y} - b + \mathcal{A}W^k\|$ and $g^k = b - \mathcal{A}W_S^k - \eta^k$ are bounded, and therefore (v. (2.13)) $y^k = y_{W^k,\eta^k}$ and $y^{k+\frac{1}{2}} = y_{W^{k+1},\eta^k}$ are bounded, as well. Hence, (3.8) and the last line of (3.12) yield $\epsilon^k \to 0$.

It remains to verify (3.3). Let y^∞ be the (unique) minimizer of $f + \frac{u}{2}\|\cdot - \hat{y}\|^2$; we have to show $y^k \to y^\infty$. For this we exploit $f_{W^k,\eta^k}(y^k) \to f(y^k)$,

$$\begin{aligned} f(y^\infty) + \tfrac{u}{2}\|y^\infty - \hat{y}\|^2 &\geq L(y^\infty; W^k, \eta^k) && [f_{W^k,\eta^k} \leq f (2.7), (2.11)] \\ &\geq \min L(\cdot; W^k, \eta^k) \\ &= L(y^k, W^k, \eta^k) && [y^k = y_{W^k,\eta^k} (2.18), (2.13)] \\ &= f(y^k) + \tfrac{u}{2}\|y^k - \hat{y}\|^2 - \epsilon^k, && [(3.2), (2.11)] \\ &\geq f(y^\infty) + \tfrac{u}{2}\|y^\infty - \hat{y}\|^2 + \tfrac{u}{2}\|y^k - y^\infty\|^2 - \epsilon^k, \end{aligned} \tag{3.13}$$

where the last inequality follows from y^∞ being the minimizer of $f + \frac{u}{2}\|\cdot - \hat{y}\|^2$ with f convex. The first and last line of (3.13) imply $\frac{u}{2}\|y^k - y^\infty\|^2 \leq \epsilon^k$. By (3.2) $\epsilon^k \to 0$, i.e., (3.3) holds. Thus (a) has been proved.

(b) We reduce this case to case (a) by interpreting an inner iteration as an outer iteration with the evaluation Step 2 restricted to optimizing over $\widehat{\mathcal{W}}^k$ instead of $\mathcal{W}$ and with the updating Step 4 setting $\widehat{\mathcal{W}}^{k+1} = \widehat{\mathcal{W}}^k$; in doing so we have to take care that the subsequence of iterates corresponding to these restricted evaluations is not included in the argument for (3.2) and (3.3). To make this precise, consider the following modifications of Algorithm 2.2. At Step 0, set $K = \emptyset$. At Step 1(d), instead of returning to Step 1(a), set $K := K \cup \{k+1\}$ and go to Step 1(e), then at Step 2 find $W_S^{k+1} \in \operatorname{Argmax}_{\widehat{\mathcal{W}}^k} \left\langle C - \mathcal{A}^T y^{k+1}, \cdot \right\rangle$, at Step 3 set $\hat{y}^{k+1} = \hat{y}^k$ without evaluating f, and at Step 4 set $\widehat{\mathcal{W}}^{k+1} = \widehat{\mathcal{W}}^k$. This modification only renumbers iterations, so that K indexes the subsequence of additional inner iterations of Step 1, if any. It is not difficult to check that (3.8) remains valid, whereas for $k \in K' := \{1, 2, \ldots\} \setminus K$, (3.12)–(3.13) give (3.2)–(3.3) with "$\to$" replaced by "$\xrightarrow{K'}$", i.e., (3.2)–(3.3) for the original version of the algorithm.

(c) Arguing as in (b), suppose $K = \{\hat{k}, \hat{k}+1, \ldots\}$ indexes an uninterrupted infinite sequence of inner iterations (the number of outer iterations $|K'|$ is finite). For $k \geq \hat{k}$, denote by $\widehat{\mathcal{W}} := \widehat{\mathcal{W}}^{\hat{k}} = \widehat{\mathcal{W}}^k$, let $\epsilon^k := f_{\widehat{\mathcal{W}},Y}(y^k) - f_{W^k,\eta^k}(y^k)$. For (3.12) use $f_{W_S^k,\eta^k}(y^k) = f_{\widehat{\mathcal{W}},Y}(y^k)$ (y^k is feasible, (2.18)) and $f_{W^k,\eta^k} \leq f_{\widehat{\mathcal{W}},Y}$; in (3.13) replace f by $f_{\widehat{\mathcal{W}},Y}$. We obtain (3.2)–(3.3) with f replaced by $f_{\widehat{\mathcal{W}},\eta}$. This yields (3.4)–(3.5) except for $y^\infty = \hat{y}$.

It remains to show $y^\infty = \hat{y}$ and $\hat{y} \in \operatorname{Argmin} f$ when the condition of Step 1(d) is satisfied in each iteration $k \geq \hat{k}$ of the modified algorithm,

$$\epsilon^{k+1} = f_{\widehat{\mathcal{W}},Y}(y^{k+1}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) > \kappa_{\mathrm{M}} \left[f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) \right] \quad \text{for } k \geq \hat{k}. \tag{3.14}$$

Since $\epsilon^{k+1} \to 0$ and $\kappa_{\mathrm{M}} > 0$, relation (3.14) gives $f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) \to 0$; this allows to prove optimality of $\hat{y}$. Indeed, replacing in (2.22)–(2.23) the superscript "+" by "$k+1$" yields

$$f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) = f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(\hat{y}) + u\|y^{k+1} - \hat{y}\|^2 \geq 0, \tag{3.15}$$

$$f(\cdot) \geq f(\hat{y}) - \left[f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(\hat{y}) \right] + u \left\langle \hat{y} - y^{k+1}, \cdot - \hat{y} \right\rangle. \tag{3.16}$$

Now, $f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) \to 0$ and (3.15) imply $y^{k+1} \to \hat{y}$ and $f_{W^{k+1},\eta^{k+1}}(\hat{y}) \to f(\hat{y})$ and together with (3.16) this proves $f(\cdot) \geq f(\hat{y})$. But $y^k \to y^\infty$ and $\hat{y} \in \mathbb{R}^m_+$, so $y^\infty = \hat{y} \in \operatorname{Argmin} f$, as required. □

Remark 3.3 *In fact, the first part of the proof of Lem. 3.2(c) shows that the y^+ of successive inner iterations (within the same outer iteration k) converge to $\hat{y}^*$ of (2.10). An infinite inner loop corresponds to an ideal bundle prox iteration for the cutting surface model.*

If k remains finite, then either the algorithm stops in step 1(c) after a finite number of iterations (this case is discussed in Rem. 2.3(i)) or an infinite loop of inner iterations occurs. The latter case is dealt with by Lem. 3.2(c), so we assume from now on that $k \to \infty$.

Next, we finish the case of finitely many descent steps. The basic idea is quickly explained. Since $k \to \infty$ the algorithm must end with an infinite sequence of null steps whose iterates converge, by Lem. 3.2(a), to the proximal point y^∞. But if $f(\hat{y}) > f(y^\infty)$ then by (3.2) and Step 3, a descent step would be unavoidable eventually. Thus, $f(\hat{y}) = f(y^\infty)$ and, by the uniqueness of the proximal point, $\hat{y} = y^\infty$ will turn out to be an optimal solution of the minimization problem.

Lemma 3.4 *Suppose that, starting from some iteration $\hat{k} \geq 0$, only null steps are made. Then $\hat{y}^{\hat{k}} \in \operatorname{Argmin} f$.*

Proof. By Lem. 3.2, $\epsilon^{k+1} = f(y^{k+1}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) \to 0$. For $k \geq \hat{k}$, due to the null step in Step 3, we have $\hat{y}^k = \hat{y}^{\hat{k}} = \hat{y}$ (v. Step 1) and

$$f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) - \epsilon^{k+1} = f(\hat{y}) - f(y^{k+1}) < \kappa \left[f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) \right],$$

so $\epsilon^{k+1} > (1-\kappa)[f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1})]$. This is like (3.14) with κ_{M} replaced by $1-\kappa > 0$ (since $\kappa < 1$, v. Input of Algorithm 2.2); hence, the argument following (3.14) yields $\hat{y} \in \operatorname{Argmin} f$. □

In the case of infinitely many descent steps we have to make sure that there is no danger of false convergence, i.e., the function values of the algorithm cannot converge to a value strictly greater than the optimal solution value. To this end, consider the following condition,

$$f(\hat{y}^k) \geq f(\tilde{y}) \quad \text{for some fixed } \tilde{y} \in Y \text{ and all } k. \tag{3.17}$$

This condition holds if optimal solutions exist ($\operatorname{Argmin} f \neq \emptyset$) or if $\tilde{y}$ is a cluster point of $\{\hat{y}^k\}$. Indeed, any cluster point $\tilde{y}$ of $\{\hat{y}^k\}$ must be feasible (all $\hat{y}^k$ are feasible and $\mathbb{R}^m_+$ is closed), and since $f(\hat{y}^k) \geq f(\hat{y}^{k+1})$ and f is continuous on $\mathbb{R}^m_+$, the objective value $f(\tilde{y})$ must satisfy (3.17).

The following Lemma asserts that, whenever condition (3.17) holds, the iterates $\{\hat{y}^k\}$ converge to an optimal solution of the minimization problem.

Lemma 3.5 *If (3.17) holds, then $\hat{y}^k \to \bar{y}$ for some $\bar{y} \in \operatorname{Argmin} f$, and $f(\hat{y}^k) \downarrow f(\bar{y})$.*

Proof. In view of Lem. 3.4, we may assume that the set of descent iterations $D := \{k : \hat{y}^{k+1} = y^{k+1}\}$ is infinite. At Step 3, each iteration $k \in D$ satisfies the descent step criterion and therefore $0 \le f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) \le \frac{1}{\kappa}[f(\hat{y}^k) - f(\hat{y}^{k+1})]$ with $\kappa > 0$. So by condition (3.17)

$$\sum_{k\in D} \left[f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) \right] \le \tfrac{1}{\kappa} \sum_{k\in D} \left[f(\hat{y}^k) - f(\hat{y}^{k+1}) \right] \le \tfrac{1}{\kappa} \left[f(\hat{y}^0) - f(\tilde{y}) \right] < \infty. \tag{3.18}$$

Next we show that $\|\hat{y}^k - \tilde{y}\|^2$ increases by at most $\frac{2}{u}\left[f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) \right]$ in each descent step (the distance remains the same for null steps $k \notin D$, because $\hat{y}^{k+1} = \hat{y}^k$ in Step 3); together with (3.18) this will show the boundedness of $\{\hat{y}^k\}$ and thus the existence of an accumulation point.

For $k \in D$, the gradient of $f_{W^{k+1},\eta^{k+1}}$ may be expressed by (2.21) as (replace "+" by "k+1" and use $\hat{y}^{k+1} = y^{k+1}$)

$$\nabla f_{W^{k+1},\eta^{k+1}} = u(\hat{y}^k - \hat{y}^{k+1}). \tag{3.19}$$

Since $f_{W^{k+1},\eta^{k+1}} \le f$ is affine (v. (2.7)), condition (3.17) implies

$$\begin{aligned} f(\hat{y}^k) \ge f(\tilde{y}) &\ge f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) + \left\langle \nabla f_{W^{k+1},\eta^{k+1}}, \tilde{y} - \hat{y}^{k+1} \right\rangle \\ &= f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) + u \left\langle \hat{y}^k - \hat{y}^{k+1}, \tilde{y} - \hat{y}^{k+1} \right\rangle. \end{aligned}$$

Using this inequality, we deduce that for each $k \in D$

$$\begin{aligned} \|\tilde{y} - \hat{y}^{k+1}\|^2 &\le \|\tilde{y} - \hat{y}^k + \hat{y}^k - \hat{y}^{k+1}\|^2 + \|\hat{y}^k - \hat{y}^{k+1}\|^2 \\ &= \|\tilde{y} - \hat{y}^k\|^2 + 2\left\langle \tilde{y} - \hat{y}^k, \hat{y}^k - \hat{y}^{k+1} \right\rangle + 2\left\langle \hat{y}^k - \hat{y}^{k+1}, \hat{y}^k - \hat{y}^{k+1} \right\rangle \\ &= \|\tilde{y} - \hat{y}^k\|^2 + 2\left\langle \tilde{y} - \hat{y}^{k+1}, \hat{y}^k - \hat{y}^{k+1} \right\rangle \\ &\le \|\tilde{y} - \hat{y}^k\|^2 + \tfrac{2}{u}\left[f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) \right]. \end{aligned}$$

Applying this bound recursively down to some $i < k$ and using (3.18) yields

$$\|\tilde{y} - \hat{y}^k\|^2 \le \|\tilde{y} - \hat{y}^i\|^2 + \tfrac{2}{u} \sum_{j\ge i, j\in D} \left[f(\hat{y}^j) - f_{W^{j+1},\eta^{j+1}}(\hat{y}^{j+1}) \right] < \infty \quad \text{if } k > i. \tag{3.20}$$

Observe, that this also holds for null step iterations $k \notin D$ because in this case $\hat{y}^{k+1} = \hat{y}^k$ by Step 3. Hence, $\{\hat{y}^k\}$ is bounded and has a cluster point $\bar{y}$ which is in $\mathbb{R}^m_+$ because $\hat{y}^k \in \mathbb{R}^m_+$. In order to show convergence of the iterates to this point set $\tilde{y} = \bar{y}$ (we may do so because condition (3.17) holds for any cluster point of $\{\hat{y}^k\}$, see the paragraph following (3.17)). For any $\epsilon > 0$ we can choose i such that both terms in the right-hand side of (3.20) are at most ϵ; thus $\hat{y}^k \to \bar{y}$. It remains to show that $\bar{y}$ is an optimal solution of our optimization problem.

Since $\hat{y}^k \to \bar{y}$, (3.19) yields $\nabla f_{W^{k+1},\eta^{k+1}} = u(\hat{y}^k - \hat{y}^{k+1}) \xrightarrow{D} 0$ and the continuity of f on $\mathbb{R}^m_+$ implies $f(\hat{y}^k) \to f(\bar{y})$. From (3.18) we obtain $f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) \xrightarrow{D} 0$, therefore $f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) \xrightarrow{D} f(\bar{y})$. Hence $f(y) \ge f_{W^{k+1},\eta^{k+1}}(y) = f_{W^{k+1},\eta^{k+1}}(\hat{y}^{k+1}) + \left\langle \nabla f_{W^{k+1},\eta^{k+1}}, y - \hat{y}^{k+1} \right\rangle$ gives in the limit $f(y) \ge f(\bar{y})\ \forall y$. □

The remaining case, that condition (3.17) does not hold, is quickly settled; the algorithm exhibits a sequence $\{\hat{y}^k\}$ of feasible points with $f(\hat{y}^k) \to -\infty$ which is in this case $\inf f$. Such a sequence must satisfy $\|\hat{y}^k\| \to \infty$, because for a cluster point condition (3.17) would hold. We may now state our principal result.

Theorem 3.6 *Either* $\hat{y}^k \to \bar{y} \in \operatorname{Argmin} f$, *or* $\operatorname{Argmin} f = \emptyset$ *and* $\|\hat{y}^k\| \to \infty$. *In both cases* $f(\hat{y}^k) \downarrow \inf f$.

Proof. If (3.17) holds, then by Lem. 3.5, $\hat{y}^k \to \bar{y} \in \operatorname{Argmin} f$ and $f(\hat{y}^k) \downarrow f(\bar{y}) = f(\tilde{y})$, so $\tilde{y} \in \operatorname{Argmin} f$. Otherwise, $\|\hat{y}^k\| \to \infty$ and, for $k \to \infty$, $f(\hat{y}^k)$ is unbounded from below; so the definition of $\inf f$ yields the desired conclusion. □

4 Modifications

4.1 Inexact null steps

A driving force in the proof of Lem. 3.2 is that the cutting plane $f_{W_S^k,\eta^k}$ improves the cutting surface model in y^k to a value above the null step threshold. We will show below that the additional property of $f_{W_S^k,\eta^k}$ of being maximal in this respect, i.e., that $f_{W_S^k,\eta^k}$ supports f in y^k is not really needed. In fact, determining the exact maximizer may be theoretically and computationally a difficult task. Therefore we now introduce *inexact null steps*. In an inexact null step a newly generated cutting plane has to improve the cutting surface model in y^k by a certain minimal amount relative to the predicted progress $f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(y^{k+1})$ but may well fall short of being supporting. If no such cutting plane can be found we need to compute $f(y^k)$ exactly and perform a descent step.

More formally, fixing $\bar{\kappa} \in [\kappa, 1)$, suppose Steps 2 and 3 of Algorithm 2.2 are replaced by

Step 2' (Descent test). Find $W_S^{k+1} \in \mathcal{W}$ such that either

(a) $f(\hat{y}^k) - f_{W_S^{k+1},0}(y^{k+1}) \le \bar{\kappa}[f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(y^{k+1})]$, or

(b) $f_{W_S^{k+1},0}(y^{k+1}) = f(y^{k+1})$ and $f(\hat{y}^k) - f(y^{k+1}) \ge \kappa[f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(y^{k+1})]$.

In case (a), set $\hat{y}^{k+1} = \hat{y}^k$ (null step), otherwise set $\hat{y}^{k+1} = y^{k+1}$ (descent step).

Observe, that the regions for accepting null steps and accepting descent steps have some overlap. If an oracle returns a supporting cutting plane that intersects this overlapping region one may decide either way without harm. In fact, for $\bar{\kappa} \in (\kappa, 1)$ this may allow to establish finite convergence properties for certain oracles, but we will not delve into this here.

We establish convergence of the above modification. If a null step occurs at Step 2', then, since $f_{W_S^{k+1},\eta^{k+1}}(y^{k+1}) = f_{W_S^{k+1},0}(y^{k+1})$ (v. (2.7),(2.19)), we have

$$\epsilon^{k+1} := f_{W_S^{k+1},\eta^{k+1}}(y^{k+1}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) \ge (1-\bar{\kappa})\left[f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1})\right]. \quad (4.1)$$

Thus $\epsilon^{k+1} \geq 0$, since $1-\bar{\kappa} > 0$ and $f(\hat{y}) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) \geq 0$ (v. (2.22)). Hence, an easy modification of the proof of Lem. 3.2 shows that Lem. 3.2 holds with (3.3) omitted; in particular, $\epsilon^k \to 0$ in cases (a) and (b). Then in the proof of Lem. 3.4 one may use (4.1), which is like (3.14) with κ_{M} replaced by $1-\bar{\kappa} > 0$. Since for descent steps the modified method behaves like the original one, the proof of Lem. 3.5 is not affected, and Thm 3.6 remains valid.

In the current context the modifications have both theoretical and practical relevance. The original Step 2 needs $\lambda = \lambda_{\max}(C - \mathcal{A}^T y^{k+1})$ and its normalized eigenvector $v = v^{k+1}$ to compute $f(y^{k+1}) = \lambda + \left\langle b, y^{k+1} \right\rangle$ and $W_S^{k+1} = vv^T$. Now, λ is an algebraic but in general irrational number and cannot be computed exactly, even less so an eigenvector v. In practice we use a Lanczos process (see, e.g., [9]) to iteratively generate better v and corresponding lower bounds on λ; convergence of λ is typically fast, whereas the error bound relies on the convergence of v which is usually much slower. The new Step 2' allows to stop the Lanczos process for computing $\lambda_{\max}(C - \mathcal{A}^T y^{k+1})$ as soon as a normalized vector v is found so that

$$v^T(C - \mathcal{A}^T y^{k+1})v + \left\langle b, y^{k+1} \right\rangle = f_{vv^T,0}(y^{k+1}) \geq f(\hat{y}^k) - \bar{\kappa}\left[f(\hat{y}^k) - f_{W^{k+1},\eta^{k+1}}(y^{k+1}) \right]; \quad (4.2)$$

then $W_S^{k+1} = vv^T$ meets the null-step criterion of Step 2'. There is no need to wait for convergence and this often saves much work in practice.

4.2 Weight updating

For efficiency, it is crucial to use a variable weight $u = u_k$. The convergence results of §3 extend easily to the weight updates of [19]. Consider the following *level modification* of these updates. Set $\bar{f}_0^* = -\infty$ at Step 0, and $\bar{f}_{k+1}^* = \max\{\bar{f}_k^*, f_{W^{k+1},\eta^{k+1}}(y^{k+1})\}$ at Step 1(e) of Algorithm 2.2. Insert the following after (d) of Procedure 2.2 of [19]:

(d') If $u = u_k$ and $f(\hat{y}^{k+1}) < \bar{f}_{k+1}^*$ then set $u = u_k/2$.

Before exiting in (e) of Procedure 2.2 of [19], set $\bar{f}_{k+1}^* = -\infty$ if $u_{k+1} < u_k$.

This modification is motivated as follows. We would like $f_{W^{k+1},\eta^{k+1}}(y^{k+1})$ to be close to the optimal value f^* of (2.1); if it is significantly higher (i.e., u_k is too large), then slow progress occurs. Introduce an indicator l by setting $l = 1$ at Step 0, and $l = k+1$ if $u_{k+1} < u_k$. Then $\bar{f}_{k+1}^* = \max_{j=l}^k f_{W^{j+1},\eta^{j+1}}(y^{j+1})$ is the highest objective level aimed for since the latest decrease of u_k. Thus $\bar{f}_{k+1}^* > f(\hat{y}^{k+1})$ implies $\bar{f}_{k+1}^* > f^*$ and hence that u_k is too large. In effect, this modification provides an additional mechanism for decreasing u_k, even if no successive descent steps occur, as required in the original version.

If $u_k \neq u_{k-1}$ or $\hat{y}^k \neq \hat{y}^{k-1}$, then Step 1 may use a better initial multiplier $\hat{\eta} = \max\{0, -u\hat{y} + b - \mathcal{A}W^k\}$ (v. (2.17)); again, the results of §3 are not affected.

5 Implementation

Although Step 4 imposes only general requirements on the next *working set* $\widehat{\mathcal{W}}^{k+1}$, its actual choice is of utmost importance in practice. Following [16], our implementation uses

$$\widehat{\mathcal{W}} = \widehat{\mathcal{W}}^k = \left\{ PVP^T + \alpha\overline{W} : \operatorname{tr} V + \alpha = 1, V \in \mathcal{S}_+^r, \alpha \geq 0 \right\}, \quad (5.1)$$

with $P = P^k \in \mathbb{R}^{n\times r}$ and $\overline{W} = \overline{W}^k \in \mathcal{S}^n_+$ such that $P^TP = I_r$ and $\operatorname{tr}\overline{W} = 1$.

By (2.14) and (5.1), (2.16) is equivalent to the quadratic semidefinite program

$$\begin{array}{ll} \min & \frac{1}{2u}\|b - \mathcal{A}(PVP^T + \alpha\overline{W}) - \hat{\eta}\|^2 - \left\langle PVP^T + \alpha\overline{W}, C - \mathcal{A}^T\hat{y}\right\rangle - \langle b - \hat{\eta}, \hat{y}\rangle\,, \\ \text{s.t.} & \operatorname{tr} V + \alpha = 1, \\ & V \succeq 0, \alpha \geq 0. \end{array} \tag{5.2}$$

Its optimal solution (V_*, α_*) yields $W^+ = PV_*P^T + \alpha_*\overline{W}$. In updating $\widehat{\mathcal{W}}$ we exploit information in V^* and α^* like in [16]. Due to several small differences we state the entire process. Using the spectral decomposition $V_* = Q\Lambda Q^T$ with $\Lambda_{11} \geq \ldots \geq \Lambda_{rr}$, let $\hat{P} := PQ$. Let $\hat{\jmath}$ be the largest j satisfying $\min\{n_{\min}, r\} \leq j \leq \min\{n_K, r\}$ with $\Lambda_{jj} \geq t_a\Lambda_{11}$ if $j > n_{\min}$, where n_K and $n_{\min} = \min\{5, n_K\}$ are the maximum and minimum number of kept columns of $\hat{P}$ and $t_a \in [0, 1)$ is an aggregation tolerance. To ensure that $W^+ \in \widehat{\mathcal{W}}^{k+1}$ (v. Step 4), we let

$$\overline{W}^{k+1} = \left(\alpha_*\overline{W} + \textstyle\sum_{j=\hat{\jmath}+1}^{r} \Lambda_{jj}\hat{P}_j\hat{P}_j^T\right) / \left(\alpha_* + \textstyle\sum_{j=\hat{\jmath}+1}^{r} \Lambda_{jj}\right), \tag{5.3}$$

where $\hat{P}_j$ is column j of $\hat{P}$. Next, Step 2' (v. §4.1) delivers a matrix $L \in \mathbb{R}^{n\times l}$ of Lanczos vectors of $C - \mathcal{A}^T y^{k+1}$ such that $W_S^{k+1} = L_1L_1^T$. Then P^{k+1} is obtained via QR factorization as an orthonormal basis of $\{\hat{P}_j\}_{j=1}^{\hat{\jmath}} \cup \{L_j\}_{j=1}^{\min\{l, n_A\}}$, where n_A is the maximum number of added columns of L. Thus $W_S^{k+1} \in \widehat{\mathcal{W}}^{k+1}$. Since aggregation is not necessary initially, we use (5.1) with $\alpha \equiv 0$ until $\hat{\jmath} < r$, in which case $\overline{W}^{k+1}$ is initialized via (5.3) with $\alpha_* = 0$. P^0 is set to the first $\min\{n_K + n_A, l\}$ columns of L found while evaluating $\lambda_{\max}(C - \mathcal{A}^T y^0)$.

The special structure (5.1) of $\widehat{\mathcal{W}}$ allows us to evaluate $f_{\widehat{\mathcal{W}},0}$ (v. (2.8)) directly:

$$f_{\widehat{\mathcal{W}},0}(y^+) = \max\left\{\lambda_{\max}(P^T(C - \mathcal{A}^T y^+)P), \left\langle\overline{W}, C - \mathcal{A}^T y^+\right\rangle\right\} + \left\langle b, y^+\right\rangle. \tag{5.4}$$

The eigenvalue computation involved in (5.4) is cheap, since the argument is in $\mathcal{S}^r$. Therefore Step 1(d) of Algorithm 2.2 can be executed efficiently.

Subproblem (5.2) is solved for each update of $\hat{\eta}$ in the inner loop or, if $\kappa_{\mathrm{M}} = \infty$, at least once per iteration. Thus the overall efficiency hinges on the speed of this computation. We now briefly explain the most important issues arising in this context.

Using the svec-operator described in [31] to expand symmetric matrices from $\mathcal{S}^r$ into column vectors of length $\binom{r+1}{2}$ and by ignoring all constants, (5.2) can be brought into the following form (note that $\langle A, B\rangle = (\operatorname{svec} A)^T \operatorname{svec} B$ for $A, B \in \mathcal{S}^r$ and $\operatorname{tr} V = \langle I, V\rangle$)

$$\begin{array}{ll} \min & \frac{1}{2}(\operatorname{svec} V)^T Q_{11} \operatorname{svec} V + \alpha q_{12}^T \operatorname{svec} V + \frac{1}{2}q_{22}\alpha^2 + c_1^T \operatorname{svec} V + c_2\alpha, \\ \text{s.t.} & \alpha + s_I^T \operatorname{svec} V = 1, \\ & \alpha \geq 0, V \succeq 0, \end{array} \tag{5.5}$$

where

$$Q_{11} = \tfrac{1}{u}\sum_{i=1}^{m} \left(\operatorname{svec} P^T A_i P\right)\left(\operatorname{svec} P^T A_i P\right)^T,$$

$$q_{12} = \tfrac{1}{u}\operatorname{svec} P^T\mathcal{A}^T(\mathcal{A}\overline{W})P,$$

$$q_{22} = \tfrac{1}{u}\left\langle \mathcal{A}\overline{W}, \mathcal{A}\overline{W}\right\rangle,$$
$$c_1 = -\operatorname{svec} P^T\left[\tfrac{1}{u}\mathcal{A}^T(b-\hat{\eta}) + C - \mathcal{A}^T\hat{y}\right]P,$$
$$c_2 = -\left\langle \tfrac{1}{u}(b-\hat{\eta}) - \hat{y}, \mathcal{A}\overline{W}\right\rangle - \left\langle C, \overline{W}\right\rangle,$$
$$s_I = \operatorname{svec} I_r.$$

This problem with $\binom{r+1}{2}+1$ variables can be solved quite efficiently by a feasible primal-dual interior point method if r is not too large, say, at most 40. Thus our algorithm can be run for problems with a huge number of constraints m.

Note, that it suffices to store and update, instead of $\overline{W}$, the m-vector $\mathcal{A}\overline{W}$ and the scalar $\left\langle C, \overline{W}\right\rangle$. The projected matrices $P^TA_iP \in \mathcal{S}^r$ are computed at most once for each instance of (5.5), and by accumulating the values we need only one of them at a time. The accumulation of Q_{11} requires $O(mr^4)$ operations. Fortunately, changes in $\hat{\eta}$ only affect the linear cost coefficients c_1 and c_2; these are easy to update. If the inner iteration of Algorithm 2.2 yields only small changes in $\hat{\eta}$, then the optimal solution will also change only slightly. We exploit this in a restarting heuristic, that is described in detail in [13].

In order to support our claim that solving (2.15) over $(W,\eta) \in \widehat{\mathcal{W}} \times \mathbb{R}^m_+$ directly by an interior point code would be too expensive, we outline its structural disadvantages. Adding variables $\eta \geq 0$, the quadratic cost matrix of (5.5) is increased to

$$Q = \begin{bmatrix} Q_{11} & q_{12} & Q_{13} \\ q_{12}^T & q_{22} & q_{23}^T \\ Q_{13}^T & q_{23} & I_m \end{bmatrix}$$

where

$$Q_{13} = \frac{1}{u}\left[\operatorname{svec}(P^TA_1P)\ldots\operatorname{svec}(P^TA_mP)\right] \quad \text{and} \quad q_{23} = \frac{1}{u}(\mathcal{A}\overline{W}).$$

In an interior point code the special structure of the coefficients of η allows to eliminate the step $\Delta\eta$ from the Newton system more or less in advance. Yet, in each iteration of an interior point code the new right hand side and the new step direction $\Delta\eta$ have to be computed by matrix vector multiplications of the form

$$\begin{bmatrix} Q_{13} \\ q_{23} \end{bmatrix}\eta \qquad \text{and} \qquad \begin{bmatrix} Q_{13}^T & q_{23}^T \end{bmatrix}\begin{bmatrix} \operatorname{svec} V \\ \alpha \end{bmatrix}$$

These cannot be avoided and so the running time of the subproblem is no longer independent of m. Indeed, the size of the matrix Q_{13} is $\binom{r+1}{2} \times m$ and that of q_{13} is $1 \times m$. Matrix q_{13} is available, but the dense Q_{13} would have to be stored for efficiency reasons. So in comparison to the suggested Lagrangian approach one would need $O(m\binom{r+1}{2})$ additional storage and each Newton step would require several additional $O(m\binom{r+1}{2})$ operations. Since a design criterion for the spectral bundle method is to be applicable to problems with $m > 100000$ (e.g., the recent DIMACS challenge instances [27] include semidefinite relaxations of frequency assignment problems with more than one million constraints) and r up to at least 40, solving (2.15) by interior point methods directly is in our opinion not a realistic alternative.

6 Computational results

Our implementation is a revised and extended version of the one described in [16]. Our experiments were carried out on a SUN Ultra 10 with a 299 MHz SUNW, UltraSPARC-IIi CPU and 576 MB main memory (but none of our examples required more than 100 MB).

In bundle methods, bundle sizes and parameter settings may have considerable influence on computational results. It seems practically impossible to provide one set of parameters that reliably performs well for all problem instances. Yet tuning parameters for each instance separately is even less acceptable and so we prefer to present results with the same parameter settings for all instances. In particular, the weight u_k is updated as described in §4.2, we use inexact null steps throughout (in comparison to exact null step evaluation we observed a reduction of the average computation time by roughly 30% on examples G_1–G_{42} of [16]). The parameter values are $\kappa_R = 0.5$, $\bar{\kappa} = \kappa = 0.1$, $\kappa_M = 0.6$. For consistency with [16], the bundle update employs $n_K = 20$, $n_A = 5$, $t_a = 0.01$ (v. §5). We also impose a time limit of one hour; this limit is checked together with the usual stopping criterion in Step 1(c).

The spectral bundle method is designed for large scale problems, but to the best of our knowledge there is no test set of suitable large scale instances with known optimal solutions. Therefore we generated a set of rather small max-cut relaxations of the graphs G_1–G_{21} with 800 nodes of [16] so that our feasible primal-dual interior point code of [15] could still solve the instances to optimality (unfortunately, it is not worth-while to report its computation times).

The relaxations were obtained as follows. Let A be the (weighted) adjacency matrix of a graph and let $C = \frac{1}{4}(\mathrm{Diag}(Ae) - A)$, where e denotes the vector of all ones and Diag transforms a vector into a diagonal matrix. For the solution X_* of the semidefinite relaxation $\max\{\, \langle C, X\rangle : \mathrm{diag}(X) = e, X \succeq 0 \,\}$ (computed by the code of [15]) we employed the separation procedure described in [15] to find 1600 violated triangle inequalities (these are facet defining cutting planes of the form $\left\langle bb^T, X \right\rangle \geq 1$ with $b \in \{-1, 0, 1\}^n$ having exactly three nonzero elements). Writing these inequalities as $\mathcal{A}X \geq e$ the primal formulation of our max-cut relaxation is

$$\max \langle C, X\rangle \quad \text{s.t.} \quad \mathrm{diag}(X) = e,\ \mathcal{A}X \geq e,\ X \succeq 0. \tag{6.1}$$

In spite of the rather severe size restriction imposed by the interior point code we consider our setting to be both realistic and relevant for the following reasons. Max-cut and its semidefinite relaxation are intimately connected to semidefinite relaxations of constrained quadratic 0-1 programming problems (see, e.g., [21]) and are thus at the heart of many combinatorial optimization problems. For these problems the cutting plane approach of improving an initial relaxation by separating violated inequalities has repeatedly proven to be highly effective. In fact, within the integer programming community the acceptance of a new method depends strongly on its ability to solve such problems efficiently. So it makes sense to look at this type of problems.

Since $\langle \mathrm{diag}(C), \mathrm{diag}(X)\rangle = e^T \mathrm{diag}(C)$ is constant for all feasible X, replacing the diagonal elements of C by the average over the diagonal, $e^T \mathrm{diag}(C)/n$, does not change the problem. With $\hat{C} := C + \mathrm{Diag}(\frac{e^T \mathrm{diag}(C)}{n} e - \mathrm{diag}(C))$, $\bar{\mathcal{I}} = \{1, \ldots, n\}$, and $\mathcal{I} =$

Table 6.1: Max-cut on graphs with 800 nodes and 1600 triangle inequalities

Problem	f^*	rel_acc	time	λ_%	k	inner	descent	$\|\nabla f_+\|$
G_1	12049.28	$2.0{\cdot}10^{-5}$	3600	32	297	321	31	0.08
G_2	12053.88	$1.0{\cdot}10^{-5}$	3603	34	290	313	27	0.08
G_3	12048.46	$1.8{\cdot}10^{-5}$	3605	34	284	313	28	0.09
G_4	12071.50	$1.3{\cdot}10^{-5}$	2797	35	220	235	27	0.12
G_5	12059.66	$1.7{\cdot}10^{-5}$	3613	33	295	316	29	0.11
G_6	2615.542	$8.1{\cdot}10^{-5}$	3601	27	306	350	27	0.08
G_7	2451.593	$2.9{\cdot}10^{-5}$	3605	29	294	353	30	0.07
G_8	2463.678	$3.8{\cdot}10^{-5}$	3609	30	297	339	37	0.06
G_9	2493.495	$5.4{\cdot}10^{-5}$	3609	29	301	341	34	0.08
G_{10}	2447.766	$1.6{\cdot}10^{-5}$	3423	27	249	410	29	0.04
G_{11}	623.4910	$3.7{\cdot}10^{-5}$	3602	12	436	2354	50	0.10
G_{12}	613.0154	$2.3{\cdot}10^{-5}$	3600	19	636	3067	52	0.05
G_{13}	636.4489	$1.4{\cdot}10^{-5}$	3600	22	578	3158	58	0.05
G_{14}	3181.331	$9.6{\cdot}10^{-6}$	3086	33	265	294	43	0.09
G_{15}	3161.874	$1.1{\cdot}10^{-5}$	2392	32	210	251	41	0.12
G_{16}	3164.930	$1.2{\cdot}10^{-5}$	2981	33	260	283	41	0.09
G_{17}	3161.765	$1.1{\cdot}10^{-5}$	3304	31	287	330	39	0.08
G_{18}	1147.605	$9.9{\cdot}10^{-6}$	2677	27	150	475	41	0.03
G_{19}	1064.577	$8.7{\cdot}10^{-6}$	2614	27	142	487	40	0.04
G_{20}	1095.452	$1.4{\cdot}10^{-5}$	1819	32	111	322	40	0.05
G_{21}	1087.888	$1.3{\cdot}10^{-5}$	1193	40	89	228	40	0.07

$\{n+1,\ldots,3n\}$, the eigenvalue problem (1.1) corresponding to the dual of (6.1) reads

$$\min_{y\in\mathbb{R}^{3n},y_{\mathcal{I}}\leq 0} n\lambda_{\max}(\hat{C}-\mathrm{Diag}(y_{\bar{\mathcal{I}}})-\mathcal{A}^T y_{\mathcal{I}})+e^T y.$$

This is the problem we actually solve. In each case our starting point is $y_0=0$.

The results for $\varepsilon_{\mathrm{opt}}=10^{-5}$ are given in Table 6.1 and for $\varepsilon_{\mathrm{opt}}=10^{-3}$ in Table 6.2. The first column in the tables is the problem name. The second column (f^*) displays our reference values for judging a posteriori the quality of the final solutions. These values are lower bounds obtained from primal objective values of the interior point code [15]; they should match the true optima in the first 5 digits. The third column (*rel_acc*) gives the relative precision $[f(\hat{y}^k)-f^*]/(|f^*|+1)$ of the final objective value $f(\hat{y}^k)$ with respect to the reference value f^*. The fourth and fifth columns (*time* and λ_%) show the total time in CPU-seconds and its percentage spent in the Lanczos code for computing maximal eigenvalues and Lanczos vectors. The sixth column (k) lists the final iteration numbers. Column *inner* gives the number of executions of Step 1(a); the difference between *inner* and $k+1$ accounts for additional inner loops due to Step 1(d). The last but one column (*descent*) displays the number of descent steps. The final column ($\|\nabla f_+\|$) gives the norm of the gradient of the linear model f_{W^+,η^+} at termination.

The results of Table 6.1 show that the numbers of additional inner iterations for updating η are usually acceptable. They are rather high only for examples G_{10}–G_{12} (these are instances of grid graphs). A closer inspection of these problems reveals that about 900 out of the 1600 primal inequality constraints are active (have $\eta_i>0$) for instances G_1–G_{10},

Table 6.2: Instances of Table 6.1 for $\varepsilon_{\text{opt}} = 10^{-3}$

Problem	f^*	rel_acc	time	λ_%	k	inner	descent	$\|\nabla f_+\|$
G_1	12049.28	$6.9{\cdot}10^{-4}$	104	54	13	14	10	2.60
G_2	12053.88	$7.7{\cdot}10^{-4}$	176	46	19	20	9	1.71
G_3	12048.46	$1.0{\cdot}10^{-3}$	96	52	13	14	9	2.79
G_4	12071.50	$8.4{\cdot}10^{-4}$	111	48	14	15	8	2.16
G_5	12059.66	$8.1{\cdot}10^{-4}$	190	44	20	23	10	1.43
G_6	2615.542	$8.6{\cdot}10^{-4}$	540	31	53	58	15	0.54
G_7	2451.593	$9.2{\cdot}10^{-4}$	538	32	53	56	17	0.42
G_8	2463.678	$8.8{\cdot}10^{-4}$	618	32	59	62	20	0.42
G_9	2493.495	$8.5{\cdot}10^{-4}$	548	33	53	56	21	0.42
G_{10}	2447.766	$7.8{\cdot}10^{-4}$	502	31	50	54	16	0.39
G_{11}	623.4910	$9.6{\cdot}10^{-4}$	496	17	85	162	30	1.03
G_{12}	613.0154	$8.5{\cdot}10^{-4}$	474	31	127	182	31	0.68
G_{13}	636.4489	$9.2{\cdot}10^{-4}$	408	33	120	163	32	0.61
G_{14}	3181.331	$8.6{\cdot}10^{-4}$	192	41	35	36	22	1.85
G_{15}	3161.874	$9.5{\cdot}10^{-4}$	203	41	36	37	23	1.84
G_{16}	3164.930	$7.7{\cdot}10^{-4}$	174	41	32	33	21	2.08
G_{17}	3161.765	$6.8{\cdot}10^{-4}$	185	38	34	36	21	1.86
G_{18}	1147.605	$8.0{\cdot}10^{-4}$	322	36	47	51	23	0.77
G_{19}	1064.577	$9.1{\cdot}10^{-4}$	301	42	46	54	23	0.88
G_{20}	1095.452	$8.6{\cdot}10^{-4}$	322	34	48	58	23	0.54
G_{21}	1087.888	$6.2{\cdot}10^{-4}$	268	39	44	52	24	0.66

around 1300 are active for G_{11}–G_{12}, and roughly 1000 are active for G_{13}–G_{21}. Since inner iterations approximate an optimal η in a sequence of bundle steps, the difficulty of this process strongly depends on the number of positive coordinates of η.

The spectral bundle method slows down as it approaches the optimal solution. This, however, need not be an obstacle in applications where one is not interested in the exact optimum, but prefers getting a rough estimate rapidly. In particular, this applies to relaxations of combinatorial optimization problems and, even more so, to cutting plane methods, where much faster progress can be achieved by adding new inequalities at an early stage. Within this framework it makes sense to look at the results presented in Table 6.2 that were obtained for a less stringent optimality tolerance $\varepsilon_{\text{opt}} = 10^{-3}$.

Acknowledgment. We thank Franz Rendl for many fruitful discussions and two anonymous referees for their constructive criticism.

References

[1] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton. Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results. *SIAM J. Optim.*, 8(3):746–768, Aug. 1998.

[2] S. Benson, Y. Ye, and X. Zhang. Mixed linear and semidefinite programming for combinatorial and quadratic optimization. *Optimization Methods and Software*, 11&12:515–544, 1999.

[3] S. Benson, Y. Ye, and X. Zhang. Solving large-scale sparse semidefinite programs for combinatorial optimization. *SIAM J. Optim.*, 10(2):443–461, 2000.

[4] S. Burer, R. D. Monteiro, and Y. Zhang. Interior-point algorithms for semidefinite programming based on a nonlinear programming formulation. Technical report, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005, Nov. 1999.

[5] S. Burer, R. D. Monteiro, and Y. Zhang. Solving semidefinite programs via nonlinear programming, part I: Transformations and derivatives. Technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, Sept. 1999.

[6] S. Burer, R. D. Monteiro, and Y. Zhang. Solving semidefinite programs via nonlinear programming, part II: Interior point methods for a subclass of sdps. Technical report, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005, Sept. 1999.

[7] S. Burer, R. D. C. Monteiro, and Y. Zhang. Rank-two relaxation heuristics for max-cut and other binary quadratic programs. Technical Report TR00-33, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005, Nov. 2000. To appear in SIAM J. Optim.

[8] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42:1115–1145, 1995.

[9] G. H. Golub and C. F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, 2^{nd} edition, 1989.

[10] M. Grötschel, L. Lovász, and A. Schrijver. Polynomial algorithms for perfect graphs. *Annals of Discrete Mathematics*, 21:325–356, 1984.

[11] C. Helmberg. Semidefinite programming for combinatorial optimization. Habilitationsschrift TU Berlin, Jan. 2000; ZIB-Report ZR 00-34, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Takustraße 7, 14195 Berlin, Germany, Oct. 2000.

[12] C. Helmberg. `SBmethod` — a C++ implementation of the spectral bundle method. Manual to Version 1.1, ZIB-Report ZR 00-35, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Takustraße 7, 14195 Berlin, Germany, Oct. 2000. URL: `http://www.zib.de/helmberg/SBmethod`.

[13] C. Helmberg, K. C. Kiwiel, and F. Rendl. Incorporating inequality constraints in the spectral bundle method. In R. E. Bixby, E. A. Boyd, and R. Z. Ríos-Mercado, editors, *Integer Programming and Combinatorial Optimization*, volume 1412 of *Lecture Notes in Computer Science*, pages 423–435. Springer, 1998.

[14] C. Helmberg and F. Oustry. Bundle methods to minimize the maximum eigenvalue function. In H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors, *Handbook of Semidefinite Programming*, pages 307–337. Kluwer Academic Publishers, Boston/Dordrecht/London, 2000.

[15] C. Helmberg and F. Rendl. Solving quadratic (0,1)-problems by semidefinite programs and cutting planes. *Math. Programming*, 82(3):291–315, Aug. 1998.

[16] C. Helmberg and F. Rendl. A spectral bundle method for semidefinite programming. *SIAM J. Optim.*, 10(3):673–696, 2000.

[17] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz. An interior–point method for semidefinite programming. *SIAM J. Optim.*, 6(2):342–361, May 1996.

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

[19] K. C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Math. Programming*, 46:105–122, 1990.

[20] M. Kojima, S. Shindoh, and S. Hara. Interior–point methods for the monotone linear complementarity problem in symmetric matrices. *SIAM J. Optim.*, 7(1):86–125, Feb. 1997.

[21] M. Laurent, S. Poljak, and F. Rendl. Connections between semidefinite relaxations of the max-cut and stable set problems. *Math. Programming*, 77(2):225–246, 1997.

[22] L. Lovász. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory*, IT-25(1):1–7, Jan. 1979.

[23] L. Lovász and A. Schrijver. Cones of matrices and set-functions and 0-1 optimization. *SIAM J. Optim.*, 1(2):166–190, May 1991.

[24] M. V. Nayakkankuppam. *Optimization Over Symmetric Cones*. Phd-thesis, Department of Computer Science, Graduate School of Arts and Science, New York University, Sept. 1999.

[25] Y. Nesterov and M. J. Todd. Self-scaled barriers and interior-point methods for convex programming. *Math. Oper. Res.*, 22:1–42, 1997.

[26] M. L. Overton. Large–scale optimization of eigenvalues. *SIAM J. Optim.*, 2(1):88–120, Feb. 1992.

[27] G. Pataki and S. Schmieta. The DIMACS library of mixed semidefinite-quadratic-linear programs. URL: `http://dimacs.rutgers.edu/Challenges/Seventh/Instances/`, 1999.

[28] R. T. Rockafella. Monotone operators and the proximal point algorithm. *SIAM J. on Control and Optimization*, 14:877–898, 1976.

[29] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

[30] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. *SIAM J. Optim.*, 2:121–152, 1992.

[31] M. J. Todd, K. C. Toh, and R. H. Tütüncü. On the Nesterov-Todd direction in semidefinite programming. *SIAM J. Optim.*, 8(3):769–796, Aug. 1998.

[32] R. Vanderbei and H. Y. Benson. On formulating semidefinite programming problems as smooth convex nonlinear optimization problems. Technical Report ORFE 99-01, Dept. of Operations Research and Financial Engineering, Princeton University, Princeton NJ, Nov. 1999.

A Notation

$\mathbb{R}$, $\mathbb{R}^m$	real numbers, real column vector of dimension m
$\mathbb{R}^m_+, Y$	nonnegative real column vector of dimension m
$\mathcal{S}^n$	$n \times n$ symmetric real matrices
$\mathcal{S}^n_+$, $A \succeq 0$	$n \times n$ symmetric positive semidefinite matrices
I, I_n	identity of appropriate size or of size n
e	vector of all ones of appropriate dimension
e_i	i-th column of I
A_j	j-th column of A
A^T	transpose of A
$\mathrm{svec}(A)$	vector obtained by stacking the columns of the lower triangle of $A \in S_n$ with offdiagonals multiplied by $\sqrt{2}$
$\mathrm{Diag}(v)$	diagonal matrix with v on its main diagonal
$\mathrm{diag}(A)$	the diagonal of $A \in \mathbb{R}^{n\times n}$ as a column vector
$\mathrm{tr}(A)$	trace of $A \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(A) = \sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i(A)$
$\langle A, B\rangle$	inner product in $\mathbb{R}^{m\times n}$, $\langle A, B\rangle = \mathrm{tr}(B^T A)$
$\|v\|$	Euclidean norm of $v \in \mathbb{R}^m$, $\|v\| = \sqrt{\langle v, v\rangle}$
$\imath_Y(\cdot)$	indicator function, $\imath_Y(y) = 0$ for $y \in Y = \mathbb{R}^m_+$, ∞ otherwise
$\lambda_i(A)$	i-th eigenvalue of an $n \times n$ matrix A, usually $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$
$\lambda_{\min}(A)$, $\lambda_{\max}(A)$	minimal and maximal eigenvalue of A
Λ_A	diagonal matrix with $(\Lambda_A)_{ii} = \lambda_i(A)$
$\mathrm{conv}(S)$	convex hull of a set S
argmin /argmax	(unique) minimizing/maximizing argument of a function
Argmin /Argmax	set of minimizing/maximizing arguments of a function
y	design variables in $\mathbb{R}^m$, feasible for $y \in \mathbb{R}^m_+ = Y$
$\eta \in \mathbb{R}^m_+$	Lagrange multipliers for $y \geq 0$
$\mathcal{W}$, $W \in \mathcal{W}$	positive semidefinite matrices of trace 1, $\mathcal{W} = \{ W \succeq 0 : \mathrm{tr}\, W = 1 \}$
$\mathcal{A}, \mathcal{A}^T$	constraint matrix and adjoint, $[\mathcal{A}X]_i = \langle A_i, X\rangle$, $\mathcal{A}^T y = \sum_{i=1}^m y_i A_i$
$C - \mathcal{A}^T y$	affine matrix function, C is a given "cost matrix"
b	given "right hand side" vector
f	objective function, $f(y) = \lambda_{\max}(C - \mathcal{A}^T y) + b^T y + \imath_Y(y)$
$f_{W,\eta}$	affine minorant of f, $f_{W,\eta} = \langle C, W\rangle + \langle b - \eta - \mathcal{A}W, y\rangle$
$f_{\widehat{\mathcal{W}},\widehat{Y}}$	model of f for $\widehat{\mathcal{W}} \subseteq \mathcal{W}$, $\widehat{Y} \subseteq \mathbb{R}^m_+$, $f_{\widehat{\mathcal{W}},\widehat{Y}}(y) = \sup_{(W,\eta)\in\widehat{\mathcal{W}}\times\widehat{Y}} f_{W,\eta}(y)$
$\nabla f_{W,\eta}$	first derivative with respect y, $\nabla f_{W,\eta} = b - \eta - \mathcal{A}W$
$\hat{y}$	center of bundle subproblem $\min f_{\widehat{\mathcal{W}},Y}(\cdot) + \frac{u}{2}\|\cdot - \hat{y}\|^2$
$\hat{y}^*$	(unique) minimizer of bundle subproblem
$L(y; W, \eta)$	Lagrange function, $L(y; W, \eta) = f_{W,\eta}(y) + \frac{u}{2}\|y - \hat{y}\|^2$
$y_{W,\eta}$	(unique) minimizer of L for fixed (W, η), $y_{W,\eta} = \min L(\cdot; W, \eta)$
$\psi(W, \eta)$	dual function to the bundle subproblem, $\psi(W, \eta) = L(y_{W,\eta}; W, \eta)$
W^+ (W^{k+1})	maximizer (i.e. not unique) of $\psi(\cdot, \hat{\eta})$ over $\widehat{\mathcal{W}}$ for fixed $\hat{\eta} \in \mathbb{R}^m_+$
η^+ (η^{k+1})	(unique) maximizer of $\psi(W^+, \cdot)$ for fixed $W^+ \in \widehat{\mathcal{W}}$
$y^{+\frac{1}{2}}$, y^+ $(y^{k+\frac{1}{2}}, y^{k+1})$	minimizer $y_{W^+,\hat{\eta}}$ of $L(\cdot, W^+, \hat{\eta})$ and y_{W^+,η^+} of $L(\cdot, W^+, \eta^+)$
W_S^{k+1}	matrix $\in \mathcal{W}$ giving rise to a new subgradient in iteration k