Noname manuscript No.
(will be inserted by the editor)

Gaining traction — On the convergence of an inner approximation scheme for probability maximization

Csaba I. Fábián

Received: date / Accepted: date

Abstract We analyze an inner approximation scheme for probability maximization. The approach was proposed in Fábián, Csizmás, Drenyovszki, Van Ackooij, Vajnai, Kovács, Szántai (2018) Probability maximization by inner approximation, Acta Polytechnica Hungarica 15:105-125, as an analogue of a classic dual approach in the handling of probabilistic constraints.

Even a basic implementation of the maximization scheme proved usable and endured noise in gradient computations without any special effort. Moreover the speed of convergence was not affected by approximate computation of test points. This robustness was then explained in an idealized setting, considering a globally well-conditioned objective function. Here we work out convergence proofs for a logconcave distribution, specifically, for a normal distribution.

The main result of the present paper is that the procedure gains traction as an optimal solution is approached.

Keywords Convex optimization, stochastic optimization, probabilistic problems, cutting-plane method

1 Introduction

Let $F(\boldsymbol{z})$ denote an n-dimensional nondegenerate standard normal distribution function. Due to logconcavity of the normal distribution, the probabilistic function $\phi(\boldsymbol{z}) = -\log F(\boldsymbol{z})$ is convex. We discuss a probability maximization problem in the form

$$\min \phi(T\boldsymbol{x}) \quad \text{subject to} \quad A\boldsymbol{x} \leq \boldsymbol{b}, \tag{1}$$

where vectors are $\boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{b} \in \mathbb{R}^r$, and the matrices T and A are of sizes $n \times m$ and $r \times m$, respectively.

Our scheme builds an inner approximation of the epigraph of $\phi(\boldsymbol{z})$, based on function evaluations in certain test points. A master problem is formulated using this approximation. Further test points are selected in the course of the procedure, with a view of gradually improving the optimum of the master problem. This is analogous to the well-known dual approach in the handling of probabilistic constraints of the form

$$T\boldsymbol{x} \in \mathcal{L}_p = \{\, \boldsymbol{z} \,|\, F(\boldsymbol{z}) \geq p \,\} \tag{2}$$

where $p \gg 0$ is a prescribed probability. The classic dual approach applies an inner approximation of the level set $\mathcal{L}_p$. The approach has been initiated by Prékopa (1990), and the inner approximation was first applied to a

This research is supported by EFOP-3.6.1-16-2016-00006 project titled 'The development and enhancement of the research potential at John von Neumann University'. The Project is supported by the Hungarian Government and co-financed by the European Social Fund.

Department of Informatics, GAMF: Faculty of Engineering and Computer Science, John von Neumann University. Izsáki út 10, 6000 Kecskemét, Hungary. E-mail: fabian.csaba@gamf.uni-neumann.hu

probabilistic constraint by Prékopa, Vizvári, and Badics (1998). In Dentcheva, Prékopa, and Ruszczyński (2000), a cone generation scheme was developed for the continual improvement of the approximation. In this scheme, new test points are found by minimizing a tractable function over the level set $\mathcal{L}_p$. Since this minimization entails a substantial computational effort, the master part of the decomposition framework should succeed with as few test points as possible. Efficient solution methods were developed by Dentcheva, Lai, and Ruszczyński (2004) and Dentcheva and Martinez (2013), approximating the original distribution by a discrete one and applying regularization to the master problem. More recently, van Ackooij, Berge, de Oliveira, and Sagastizábal (2017) employed a special bundle-type method for the solution of the master problem, based on the on-demand accuracy approach of de Oliveira and Sagastizábal (2014).

In the classic scheme for probabilistic constraints, finding a new test point amounts to minimization over the level set $\mathcal{L}_p$. In contrast, a new approximation point in Fábián et al (2018) was found by unconstrained minimization, with considerably less computational effort. In the computational study of that paper, the unconstrained problems were solved by a simple gradient descent method. — Of course the probability maximization problem (1) is easier than the handling of a probabilistic constraint (2). A Newton-type scheme for the handling of the latter was proposed in Fábián, Csizmás, Drenyovszki, Vajnai, Kovács, and Szántai (2019). It requires the approximate solution of a short sequence of problems of the former type. (Initial problems in this sequence are solved with a large stopping tolerance, and the accuracy is gradually increased.)

Typically, gradient computations of probabilistic functions require a far greater effort than function evaluations. Computing a single non-zero component of a gradient vector will involve an effort comparable to that of computing a function value. (An alternative means of alleviating the difficulty of gradient computation in case of multivariate normal distribution has recently been proposed by Hantoute, Henrion, and Pérez-Aros (2018).)

In comparison with the outer approximation approach widely used in probabilistic programming, we mention that it requires a sophisticated implementation to deal with noise in gradient computation. Even a fairly accurate gradient may result in a cut cutting into the level set. In contrast, inner approximation results in a model that is easy to validate. A procedure that employs an inner approximation of the epigraph will endure noise in gradient computation without any special effort, provided function values are evaluated with appropriate accuracy. (Inherent stability of the model enables the application of randomized methods of simple structure. Such methods have been worked out in Fábián et al (2019).)

In the computational study of Fábián et al (2018), the speed of the convergence was not affected by approximate computation of test points. The number of the necessary test points did not increase significantly when we performed just a single line search in each gradient descent method. This robustness was then explained in an idealized setting, considering a globally well-conditioned objective function $f(\boldsymbol{z})$, with the following characteristics.

Assumption 1 *The function $f(\boldsymbol{z})$ is twice continuously differentiable, and real numbers α, ω $(0 < \alpha \leq \omega)$ exist such that*

$$\alpha I \preceq \nabla^2 f(\boldsymbol{z}) \preceq \omega I \qquad (\boldsymbol{z} \in \mathbb{R}^n).$$

Here $\nabla^2 f(\boldsymbol{z})$ is the Hessian matrix, I is the identity matrix, and the relation $U \preceq V$ between matrices means that $V - U$ is positive semidefinite.

In this paper we work out convergence proofs for a probabilistic function $\phi(\boldsymbol{z})$ derived from a logconcave distribution, specifically, from a normal distribution. Section 2 contains problem and model formulation, under mild additional assumptions. In Section 3, we discuss theoretical efficiency of the line search in finding new test points. We apply a local version of Assumption 1.

From a dual viewpoint, the epigraph approximation scheme is a cutting-plane method. We discuss this in Section 4, and give a theoretical convergence proof of the scheme. In Section 5, we show that the procedure gains traction as an optimal solution is approached. Finally, Section 6 contains comments from a practical perspective.

2 Problem and model formulation

We assume that the feasible domain of the probability maximization problem (1) is not empty, and is contained in the box $\mathcal{X} = \{\, \boldsymbol{x} \in \mathbb{R}^m \mid -1 \le x_j \le 1 \ (j = 1, \ldots, m) \,\}$. Exploiting the monotonicity of the objective function, problem (1) can be written as

$$\min\ \phi(\boldsymbol{z}) \quad \text{subject to} \quad \boldsymbol{z} - T\boldsymbol{x} \le \boldsymbol{0},\ \ A\boldsymbol{x} - \boldsymbol{b} \le \boldsymbol{0},\ \ \boldsymbol{x} \in \mathcal{X}. \tag{3}$$

Assumption 2 *A significantly high probability can be achieved in the probability maximization problem. Specifically, a feasible point $\check{\boldsymbol{z}}$ is known such that $F(\check{\boldsymbol{z}}) \ge 0.5$.*

A further speciality of the normal distribution function is the existence of a bounded box $\mathcal{Z}$ outside which the probability weight can be ignored. For the sake simplicity, we assume that this box has the form $\mathcal{Z} = \{\, \boldsymbol{z} \in \mathbb{R}^n \mid -1 \le z_j \le 1 \ (j = 1, \ldots, n) \,\}$. Including the constraint $\boldsymbol{z} \in \mathcal{Z}$ in (3) results in the approximating problem

$$\min\ \phi(\boldsymbol{z}) \quad \text{subject to} \quad \boldsymbol{z} - T\boldsymbol{x} \le \boldsymbol{0},\ \ A\boldsymbol{x} - \boldsymbol{b} \le \boldsymbol{0},\ \ \boldsymbol{z} \in \mathcal{Z},\ \boldsymbol{x} \in \mathcal{X}. \tag{4}$$

As observed in Fábián et al (2019), the difference between the respective optima of problems (3) and (4) is insignificant.

The bound $\boldsymbol{z} \in \mathcal{Z}$ in (4) allows regularization of the objective function, in the form of

$$\phi(\boldsymbol{z}) = -\log F(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z}\|^2 \tag{5}$$

with some $\rho > 0$. Substituting this regularized objective in (4) makes no significant variation in the objective value of $\boldsymbol{z} \in \mathcal{Z}$, provided ρ is small enough. On the other hand, the regularizing term improves the condition of the objective and ensures that the conjugate function $\phi^\star(.)$ is finite valued. In this paper, we are going to work with this regularized objective, with an appropriately set ρ.

Splitting the variables $\boldsymbol{z}$, we transform (4) into the equivalent form

$$\min\ \phi(\boldsymbol{z}) \quad \text{subject to} \quad \boldsymbol{z} - \boldsymbol{z}' = \boldsymbol{0},\ \ \boldsymbol{z}' - T\boldsymbol{x} \le \boldsymbol{0},\ \ A\boldsymbol{x} - \boldsymbol{b} \le \boldsymbol{0},\ \ \boldsymbol{z}' \in \mathcal{Z},\ \boldsymbol{x} \in \mathcal{X}. \tag{6}$$

Problem (6) has an optimal solution because the feasible domain is nonempty and bounded.

We relax the constraints $\boldsymbol{z} - \boldsymbol{z}' = \boldsymbol{0}$; $\boldsymbol{z} - T\boldsymbol{x} \le \boldsymbol{0}$ and $A\boldsymbol{x} - \boldsymbol{b} \le \boldsymbol{0}$ by introducing respective multiplier vectors $-\boldsymbol{u} \in \mathbb{R}^n$; $-\boldsymbol{v} \in \mathbb{R}^n, -\boldsymbol{v} \ge \boldsymbol{0}$ and $-\boldsymbol{y} \in \mathbb{R}^r, -\boldsymbol{y} \ge \boldsymbol{0}$. The Lagrangian is

$$L(\boldsymbol{z}, \boldsymbol{z}', \boldsymbol{x},\ -\boldsymbol{u}, -\boldsymbol{v}, -\boldsymbol{y}) = \ \phi(\boldsymbol{z}) \ - \ \boldsymbol{u}^T\boldsymbol{z} + \boldsymbol{u}^T\boldsymbol{z}' \ - \ \boldsymbol{v}^T\boldsymbol{z}' + \boldsymbol{v}^T T\boldsymbol{x} \ - \ \boldsymbol{y}^T A\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{b}. \tag{7}$$

The relaxed problem falls apart into three separate minimization problems:

$$\min_{\boldsymbol{z}} \Big\{ \phi(\boldsymbol{z}) - \boldsymbol{u}^T\boldsymbol{z} \Big\} \ + \ \min_{\boldsymbol{z}' \in \mathcal{Z}} (\boldsymbol{u} - \boldsymbol{v})^T \boldsymbol{z}' \ + \ \min_{\boldsymbol{x} \in \mathcal{X}} \Big(\boldsymbol{v}^T T - \boldsymbol{y}^T A \Big) \boldsymbol{x} \ + \ \boldsymbol{y}^T\boldsymbol{b}. \tag{8}$$

The first minimum is by definition the negative of the conjugate function value $\phi^\star(\boldsymbol{u})$. Due to the special form of the box $\mathcal{Z}$, the second minimum is $-\|\boldsymbol{u} - \boldsymbol{v}\|_1$. The third minimum can be computed in a similar manner, and the optimum of the relaxed problem is

$$-\phi^\star(\boldsymbol{u}) - \|\boldsymbol{u} - \boldsymbol{v}\|_1 - \Big\| T^T\boldsymbol{v} - A^T\boldsymbol{y} \Big\|_1 + \boldsymbol{b}^T\boldsymbol{y}. \tag{9}$$

Introducing the function

$$\nu(\boldsymbol{u}) := \ - \sup_{\boldsymbol{v}, \boldsymbol{y} \le \boldsymbol{0}} \Big\{ -\|\boldsymbol{u} - \boldsymbol{v}\|_1 - \Big\| T^T\boldsymbol{v} - A^T\boldsymbol{y} \Big\|_1 + \boldsymbol{b}^T\boldsymbol{y} \Big\} \qquad (\boldsymbol{u} \in \mathbb{R}^n), \tag{10}$$

the Lagrangian dual of (6) can be written as

$$\max_{\boldsymbol{u} \in \mathbb{R}^n} \big\{ -\phi^\star(\boldsymbol{u}) - \nu(\boldsymbol{u}) \big\} \ = \ -\min\ \phi^\star(\boldsymbol{u}) + \nu(\boldsymbol{u}). \tag{11}$$

According to the theory of convex duality, this problem has an optimal solution. — For a recent treatise on Lagrangian duality, see, e.g., Chapter 4 in the book Ruszczyński (2006).

Let $\boldsymbol{z}^\star$ be a partial optimal solution of the primal problem (6). This $\boldsymbol{z}^\star$ is unique due to the strict convexity of the regularized objective function $\phi(\boldsymbol{z})$. The objective function is also smooth, hence its conjugate is strictly convex (see Theorem 26.1 in Rockafellar (1970)). Hence the dual problem (11) has a unique optimal solution $\boldsymbol{u}^\star$.

Observation 3 *We have*

$$\boldsymbol{u}^\star = \nabla\phi(\boldsymbol{z}^\star). \tag{12}$$

Proof. Let us extend $\boldsymbol{z}^\star$ to be an optimal solution $(\boldsymbol{z}^\star, \boldsymbol{z}'^\star, \boldsymbol{x}^\star)$ of the primal problem (6). Given $\boldsymbol{u}^\star$, let $\boldsymbol{v}^\star, \boldsymbol{y}^\star$ denote an optimal solution of the supremum problem in (10). Then $\big((\boldsymbol{z}^\star, \boldsymbol{z}'^\star, \boldsymbol{x}^\star), (-\boldsymbol{u}^\star, -\boldsymbol{v}^\star, -\boldsymbol{y}^\star)\big)$ is a saddle point of the Lagrangian (7). Hence the Karush–Kuhn–Tucker conditions for the optimality of $(\boldsymbol{z}^\star, \boldsymbol{z}'^\star, \boldsymbol{x}^\star)$ in (6) hold with Lagrange multipliers $(-\boldsymbol{u}^\star, -\boldsymbol{v}^\star, -\boldsymbol{y}^\star)$, and (12) is part of the Karush–Kuhn–Tucker conditions. — Proofs of the cited statements can be found, e.g. in Ruszczyński (2006), Theorems 4.9, 4.7 and 3.34. □

2.1 Polyhedral models

Suppose we have evaluated the function $\phi(\boldsymbol{z})$ at points $\boldsymbol{z}_i$ $(i = 0, 1, \ldots, k)$; we introduce the notation $\phi_i = \phi(\boldsymbol{z}_i)$ for respective objective values. An inner approximation of $\phi(.)$ is

$$\begin{aligned} \phi_k(\boldsymbol{z}) = \min\ & \sum_{i=0}^{k} \lambda_i \phi_i \\ & \text{such that} \\ & \lambda_i \geq 0\ (i = 0, \ldots, k),\ \ \sum_{i=0}^{k} \lambda_i = 1, \quad \sum_{i=0}^{k} \lambda_i \boldsymbol{z}_i = \boldsymbol{z}. \end{aligned} \tag{13}$$

If $\boldsymbol{z} \notin \mathrm{Conv}(\boldsymbol{z}_0, \ldots, \boldsymbol{z}_k)$, then let $\phi_k(\boldsymbol{z}) := +\infty$. A polyhedral model of the equivalent problems (4)-(6) is

$$\min\ \phi_k(\boldsymbol{z}) \quad \text{subject to} \quad \boldsymbol{z} - T\boldsymbol{x} \leq \boldsymbol{0},\ \ A\boldsymbol{x} - \boldsymbol{b} \leq \boldsymbol{0},\ \ \boldsymbol{z} \in \mathcal{Z},\ \boldsymbol{x} \in \mathcal{X}. \tag{14}$$

We assume that (14) is feasible, i.e., its optimum is finite. This can be ensured by proper selection of the initial $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_k$ points. The convex conjugate of $\phi_k(\boldsymbol{z})$ is

$$\phi_k^\star(\boldsymbol{u}) = \max_{0 \leq i \leq k} \{\boldsymbol{u}^T \boldsymbol{z}_i - \phi_i\}. \tag{15}$$

As $\phi_k^\star(.)$ is a cutting-plane model of $\phi^\star(.)$, the following problem is a polyhedral model of problem (11):

$$\max_{\boldsymbol{u} \in \mathbb{R}^n} \{-\phi_k^\star(\boldsymbol{u}) - \nu(\boldsymbol{u})\} \ = \ -\min\ \phi_k^\star(\boldsymbol{u}) + \nu(\boldsymbol{u}). \tag{16}$$

At the same time, (16) is the linear programming dual of (14).

Let $\overline{\boldsymbol{z}}$ denote a partial optimal solution of the primal model problem (14), and let $\overline{\boldsymbol{u}}$ denote an optimal solution of the dual model problem (16) — existing due to our assumption concerning the feasibility of (14). The following observation was proved in Fábián et al (2018).

Observation 4 *We have* $\phi_k(\overline{\boldsymbol{z}}) + \phi_k^\star(\overline{\boldsymbol{u}}) = \overline{\boldsymbol{u}}^T \overline{\boldsymbol{z}}$ *and hence* $\overline{\boldsymbol{u}} \in \partial\phi_k(\overline{\boldsymbol{z}})$.

2.2 Initialization and successive improvement of the models

In Fábián et al (2018), we initialized the model function (13) with $n+1$ points (n being the dimension of the distribution). The 'most positive' vertex of $\mathcal{Z}$ was selected as $\boldsymbol{z}_0$, and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ were respective points from the edges adjoining $\boldsymbol{z}_0$. In this paper we follow this way of initialization, further adding the vector of Assumption 2, as $\boldsymbol{z}_{n+1} = \check{\boldsymbol{z}}$.

The probability maximization problem can be solved by a column generation procedure. New approximation points are found by unconstrained maximization. Specifically, let $\overline{\boldsymbol{u}}$ denote a partial dual optimal solution of the current model problem (14). A new approximation point $\boldsymbol{z}_{k+1}$ is found by maximizing the function $\overline{\boldsymbol{u}}^T\boldsymbol{z} - \phi(\boldsymbol{z})$ (this expression represents the reduced cost of $\boldsymbol{z}$ in the context of the simplex method). In Fábián et al (2018), we proposed approximate solutions of the column generation subproblems, always performing a single line search starting from the partial optimal solution $\overline{\boldsymbol{z}}$ of the current model problem.

Looking at the column-generation approach from a dual viewpoint we can see a cutting-plane method. This relationship between the primal and dual approaches is well known, see, e.g., Frangioni (2002, 2018), but it is immediately apparent in the present case: (16) is clearly a cutting-plane model of (11). Let $\overline{\boldsymbol{u}}$ denote the optimal solution of the current model problem, we construct an approximate support function to the objective at $\overline{\boldsymbol{u}}$. In case of $\nu(\boldsymbol{u})$, an exact support function can be computed through the solution of a linear programming problem. In case of $\phi_k^\star(\boldsymbol{u})$, we construct an approximate support function in the form

$$\ell'(\boldsymbol{u}) := \boldsymbol{u}^T\boldsymbol{z}' - \phi(\boldsymbol{z}'), \tag{17}$$

with an appropriate vector $\boldsymbol{z}'$. We have $\ell'(\boldsymbol{u}) \leq \phi^\star(\boldsymbol{u})$ for any $\boldsymbol{u}$ by the definition of $\phi^\star(\boldsymbol{u})$. We wish to construct $\ell'(\boldsymbol{u})$ whose graph cuts deeply into the epigraph of the model function $\phi_k^\star(\boldsymbol{u})$. Depth of cut being measured at $\overline{\boldsymbol{u}}$, this is achieved by approximately maximizing $\overline{\boldsymbol{u}}^T\boldsymbol{z} - \phi(\boldsymbol{z})$. Adding the appropriate $\boldsymbol{z}'$ vector as $\boldsymbol{z}_{k+1}$ to the test points means, from a dual point of view, adding the cut $\phi_{k+1}^\star(\boldsymbol{u}) \geq \boldsymbol{u}^T\boldsymbol{z}' - \phi(\boldsymbol{z}')$ in the next model function (15).

2.3 Bounds on optimal solutions

Our first aim is to construct a compact set that contains respective partial optimal solutions $\overline{\boldsymbol{z}}$ of each successive dual model problem (14). – An obvious choice would be the box $\mathcal{Z}$. But, for reasons that will become apparent shortly, we need a set well inside $\mathcal{Z}$.

As we include $\check{\boldsymbol{z}}$ of Assumption 2 among the initial test points of the model function, the following set contains all appropriate $\overline{\boldsymbol{z}}$ vectors:

$$\mathcal{O}_{\boldsymbol{z}} = \big(\{\, T\boldsymbol{x} \,|\, \boldsymbol{x} \in \mathcal{X} \,\} + \mathcal{N} \big) \cap \{\, \boldsymbol{z} \in \mathbb{R}^n \,|\, F(\boldsymbol{z}) \geq 0.5 \,\}, \tag{18}$$

where $\mathcal{N}$ denotes the negative (closed) orthant in $\mathbb{R}^n$. Moreover we have $\boldsymbol{z}^\star \in \mathcal{O}_{\boldsymbol{z}}$ with partial optimal solution $\boldsymbol{z}^\star$ of the primal problem (4). The set $\mathcal{O}_{\boldsymbol{z}}$ is obviously compact.

We are now going to construct a compact set that contains respective optimal solutions $\overline{\boldsymbol{u}}$ of each successive dual model problem (16). To this end, we are going to use Observation 4.

The model function (13) is initialized in the manner sketched in the previous section. I.e., the 'most positive' vertex of $\mathcal{Z}$ is selected as $\boldsymbol{z}_0$, and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are respective points from the edges adjoining $\boldsymbol{z}_0$. Let $\mathcal{S} = \mathrm{Conv}(\boldsymbol{z}_0, \ldots, \boldsymbol{z}_n)$ denote the convex hull of these $n+1$ test points. As the box $\mathcal{Z}$ can be extended arbitrarily, we may assume not only that $\mathcal{O}_{\boldsymbol{z}} \subset \mathcal{S}$ holds, but even that

$$\mathcal{O}_{\boldsymbol{z}} \subset \mathcal{S}_\eta \quad \text{holds with some } \eta \in (0,1), \tag{19}$$

where $\mathcal{S}_\eta$ denotes the simplex obtained by shrinking $\mathcal{S}$ towards its barycenter, the ratio of decrease being 1 to η. In what follows, we work with a fixed value η satisfying (19).

Observation 5 *Assume that the model functions have been initialized in such a way that (19) holds.*
Then there exists a finite upper bound $\Gamma_{\boldsymbol{u}}$ such that

$$\|\boldsymbol{u}\| \leq \Gamma_{\boldsymbol{u}} \;\; \text{holds for any} \;\; \boldsymbol{u} \in \partial\phi_k(\boldsymbol{z}) \qquad (\boldsymbol{z} \in \mathcal{O}_{\boldsymbol{z}}),$$

with any of the successive model functions $\phi_k(\boldsymbol{z})$.
Moreover, with the true objective function, we have $\|\nabla\phi(\boldsymbol{z})\| \leq \Gamma_{\boldsymbol{u}}$ $(\boldsymbol{z} \in \mathcal{O}_{\boldsymbol{z}})$ as well.

I include detailed proof in the appendix, section A.1.

3 Finding new test points

Let $\overline{\boldsymbol{z}}$ and $\overline{\boldsymbol{u}}$ be partial optimal solutions of the current primal and dual model problem, respectively. In the column generation scheme, the next test point $\boldsymbol{z}_{k+1} = \boldsymbol{z}'$ is obtained by approximate maximization of $\overline{\boldsymbol{u}}^T\boldsymbol{z} - \phi(\boldsymbol{z})$.

3.1 On the efficiency of line search

As motivation, let us first consider an idealized setting, with a globally well-conditioned function $f : \mathbb{R}^n \to \mathbb{R}$. By globally well-conditioned, I mean that Assumption 1 holds.

The following well-known theorem can be found e.g., in Chapter 8.6 of Luenberger and Ye (2008). (Ruszczyński (2006) in Chapter 5.3.5, Theorem 5.7 presents a slightly different form.)

Theorem 6 *Let Assumption 1 hold. We minimize $f(\boldsymbol{z})$ over $\mathbb{R}^n$ using a steepest descent method, starting from a point $\boldsymbol{z}^0$. Let $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^j, \ldots$ denote the iterates obtained by applying exact line search at each step. Then we have*

$$f\left(\boldsymbol{z}^j\right) - \mathcal{F} \;\leq\; \left(1 - \frac{\alpha}{\omega}\right)^j \left[\, f\left(\boldsymbol{z}^0\right) - \mathcal{F} \,\right], \tag{20}$$

where $\mathcal{F} = \min_{\boldsymbol{z}} f(\boldsymbol{z})$.

Our column generation subproblem consists in approximately minimizing the function $f(\boldsymbol{z}) = \phi(\boldsymbol{z}) - \overline{\boldsymbol{u}}^T\boldsymbol{z}$. We actually perform a single line search from $\overline{\boldsymbol{z}}$ in the opposite direction of the gradient $\overline{\boldsymbol{g}} = \nabla f(\overline{\boldsymbol{z}}) = \nabla\phi(\overline{\boldsymbol{z}}) - \overline{\boldsymbol{u}}$. We have $\|\overline{\boldsymbol{g}}\| \leq 2\Gamma_{\boldsymbol{u}}$ by Observation 5. Hence any line search is performed on a ray of the form

$$\left\{\, \overline{\boldsymbol{z}} - t\overline{\boldsymbol{g}} \;\middle|\; t \geq 0 \,\right\} \qquad \text{where} \quad \overline{\boldsymbol{z}} \in \mathcal{O}_{\boldsymbol{z}}, \;\; \|\overline{\boldsymbol{g}}\| \leq 2\Gamma_{\boldsymbol{u}}. \tag{21}$$

Due to the regularizing term in the objective (5), all eigenvalues of the Hessians are increased by ρ, hence $\nabla^2\phi(\boldsymbol{z}) \succeq \rho I$ $(\boldsymbol{z} \in \mathbb{R}^n)$ holds. Though there exists no finite upper bound on the eigenvalues, it turns out that a local version of Assumption 1 is sufficient.

Given $\boldsymbol{z} \in \mathbb{R}^n$, let $\omega(\boldsymbol{z})$ denote the largest eigenvalue of $\nabla^2 f(\boldsymbol{z})$. Continuity of this function is a straight consequence of a theorem of Ostrowski that I cite as Theorem 15 in Appendix B. Since the set $\mathcal{O}_{\boldsymbol{z}}$ is compact, it follows that

$$\Omega = \sup_{\boldsymbol{z} \in \mathcal{O}_{\boldsymbol{z}}} \omega(\boldsymbol{z})$$

is finite. With this, let

$$\overline{\mathcal{O}} = \left\{\, \overline{\boldsymbol{z}} - t\overline{\boldsymbol{g}} \;\middle|\; \overline{\boldsymbol{z}} \in \mathcal{O}_{\boldsymbol{z}}, \;\; \|\overline{\boldsymbol{g}}\| \leq 2\Gamma_{\boldsymbol{u}}, \;\; t \in [\,0,\, 1/\Omega\,] \,\right\}.$$

This again is a compact set, hence

$$\overline{\Omega} = \sup_{\boldsymbol{z} \in \overline{\mathcal{O}}} \omega(\boldsymbol{z})$$

is again finite. As $\overline{\mathcal{O}} \supseteq \mathcal{O}_{\boldsymbol{z}}$, it follows that $\overline{\Omega} \geq \Omega$.

Proposition 7 *With $\boldsymbol{z}'$ found by the line search, we have*

$$f\left(\boldsymbol{z}'\right) - \mathcal{F} \;\leq\; (1 - \rho/\overline{\Omega})\,[\, f\,(\overline{\boldsymbol{z}}) - \mathcal{F}\,]\,, \tag{22}$$

where $\mathcal{F} = \min_{\boldsymbol{z}} f(\boldsymbol{z})$.

In words, (22) means that the vector $\boldsymbol{z}'$ has a significant reduced cost in the master problem (14), hence we may expect a significant improvement from including $\boldsymbol{z}'$ as a new column. The proof will be an adaptation of that of Theorem 6, as related in Chapter 8.6 of Luenberger and Ye (2008). We restrict our investigation to $\overline{\mathcal{O}}$.
Proof of Observation 7. Due to the above construction, we have $\overline{\boldsymbol{z}} - t\overline{\boldsymbol{g}} \in \overline{\mathcal{O}}$ for t such that $0 \leq t \leq 1/\Omega$. Hence $\nabla^2\phi(\boldsymbol{z}) \preceq \overline{\Omega} I$ holds for $\boldsymbol{z} = \overline{\boldsymbol{z}} - t\overline{\boldsymbol{g}}$ with these t values. It follows that

$$f\,(\overline{\boldsymbol{z}} - t\overline{\boldsymbol{g}}) \;\leq\; f\,(\overline{\boldsymbol{z}}) \;-\; t\,\overline{\boldsymbol{g}}^T\overline{\boldsymbol{g}} \;+\; \frac{\overline{\Omega}}{2}t^2\,\overline{\boldsymbol{g}}^T\overline{\boldsymbol{g}} \qquad (0 \leq t \leq 1/\Omega)$$

holds (a consequence of Taylor's theorem). We consider the respective minima in $t \in \mathbb{R}$ separately of the two sides. The right-hand side is a quadratic expression, yielding minimum at $t = 1/\overline{\Omega}$. (Note that $1/\overline{\Omega} \leq 1/\Omega$.) It follows that

$$\min_{t\in\mathbb{R}} f\,(\overline{\boldsymbol{z}} - t\overline{\boldsymbol{g}}) \;\leq\; f\,(\overline{\boldsymbol{z}}) \;-\; \frac{1}{2\overline{\Omega}}\,\|\overline{\boldsymbol{g}}\|^2. \tag{23}$$

From $\nabla^2\phi(\boldsymbol{z}) \succeq \rho I \;\; (\boldsymbol{z} \in \mathbb{R}^n)$, it follows in a similar manner that

$$f(\boldsymbol{z}) \geq f(\overline{\boldsymbol{z}}) + \overline{\boldsymbol{g}}^T(\boldsymbol{z} - \overline{\boldsymbol{z}}) + \frac{\rho}{2}\|\boldsymbol{z} - \overline{\boldsymbol{z}}\|^2 \qquad (\boldsymbol{z} \in \mathbb{R}^n). \tag{24}$$

The right-hand side expression is a quadratic function of $\boldsymbol{z}$, which yields its minimum at $\boldsymbol{z} = \overline{\boldsymbol{z}} - \frac{1}{\rho}\overline{\boldsymbol{g}}$. Hence

$$\mathcal{F} \geq f(\overline{\boldsymbol{z}}) - \frac{1}{2\rho}\|\overline{\boldsymbol{g}}\|^2. \tag{25}$$

The proof is concluded by simple transformations of (23) and (25). The improving column $\boldsymbol{z}'$ is found by solving the minimization problem in the left-hand side of (23). Subtracting $\mathcal{F}$ from both sides of (23), and applying (25) to underestimate $\|\overline{\boldsymbol{g}}\|^2$, we get (22). □

3.2 Bounding improving columns

We always find an improving column by performing a single line search. Though exact optimal solutions of the column generation subproblems are never actually computed, we'll need a common bound that is independent of the master problem's iteration count.

Observation 8 *Let $\boldsymbol{z}^\circ$ denote the exact optimal solution of the current column generation subproblem $\max_{\boldsymbol{z}}\{\overline{\boldsymbol{u}}^T\boldsymbol{z} - \phi(\boldsymbol{z})\}$, where $\overline{\boldsymbol{u}}$ denotes a partial dual optimal solution of the current model problem (14).*

There exists a common finite bound on the norms $\|\boldsymbol{z}^\circ\|$ of the successive optimal columns.

I include a proof in the appendix, section A.2.

4 An approximate cutting-plane method

As mentioned before, the column generation scheme from a dual point of view results in a cutting-plane method. In this section, we interpret special features of our column generation method from a dual point of view.

Introducing the function $d(\boldsymbol{u}) := \; \phi^\star(\boldsymbol{u}) + \nu(\boldsymbol{u})$, the dual problem (11) can be solved in the form

$$\min_{\boldsymbol{u}\in\mathbb{R}^n} d(\boldsymbol{u}). \tag{26}$$

The kth model function is $d_k(\boldsymbol{u}) := \; \phi_k^\star(\boldsymbol{u}) + \nu(\boldsymbol{u})$, with $\phi_k^\star(\boldsymbol{u})$ defined in (15). The corresponding model problem is

$$\min_{\boldsymbol{u}\in\mathbb{R}^n} d_k(\boldsymbol{u}). \tag{27}$$

Moreover let

$$\mathcal{O}_{\boldsymbol{u}} = \{\ \boldsymbol{u} \in \mathbb{R}^n \ \big|\ \|\boldsymbol{u}\| \leq \Gamma_{\boldsymbol{u}}\ \}. \tag{28}$$

with $\Gamma_{\boldsymbol{u}}$ of Observation 5. This is a compact ball that contains respective optimal solutions $\overline{\boldsymbol{u}}$ of each of the successive model problems (27), as well as the optimal solution $\boldsymbol{u}^\star$ of the convex problem (26).

We perform a cutting-plane method, as sketched in section 2.2. Let $\overline{\boldsymbol{u}}$ denote the optimal solution of the current model problem, we construct an approximate support function to the objective at $\overline{\boldsymbol{u}}$. In case of $\nu(\boldsymbol{u})$, an exact support function can be computed through the solution of a linear programming problem. In case of $\phi^\star(\boldsymbol{u})$, we construct an approximate support function in the form of (17) that I copy here for convenience:

$$\ell'(\boldsymbol{u}) = \boldsymbol{u}^T\boldsymbol{z}' - \phi(\boldsymbol{z}'). \tag{29}$$

The appropriate $\boldsymbol{z}'$ vector is obtained by approximately maximizing $\overline{\boldsymbol{u}}^T\boldsymbol{z} - \phi(\boldsymbol{z})$, namely, by performing a single line search starting from $\overline{\boldsymbol{z}}$. (In the present dual viewpoint, $\overline{\boldsymbol{z}}$ is obtained as a partial dual solution of the current model problem.) Adding the appropriate $\boldsymbol{z}'$ vector as $\boldsymbol{z}_{k+1}$ to the test points means adding the cut $\phi^\star_{k+1}(\boldsymbol{u}) \geq \ell'(\boldsymbol{u})$ in the next model function (15).

Translating Proposition 7 to the present setting, we get

Corollary 9 *The support function (29), constructed by the above rule, satisfies*

$$\ell'(\overline{\boldsymbol{u}}) \ \geq\ \phi^\star(\overline{\boldsymbol{u}}) \ -\ (1-\theta)\Big[\ \phi^\star(\overline{\boldsymbol{u}}) - \phi^\star_k(\overline{\boldsymbol{u}})\ \Big], \tag{30}$$

with a constant θ $(0 < \theta \leq 1)$ independent of the iteration count k.

In words, (30) means that the new cut will significantly improve the model function value at the current iterate $\overline{\boldsymbol{u}}$. — Improvement is measured in terms of the true function value $\phi^\star(\overline{\boldsymbol{u}})$ and the current model function value $\phi^\star_k(\overline{\boldsymbol{u}})$.

Proof of Corollary 9. We start with formulating (22) in terms of $\phi(\boldsymbol{z})$ and $\phi^\star(\boldsymbol{u})$. By the definition $f(\boldsymbol{z}) = \phi(\boldsymbol{z}) - \overline{\boldsymbol{u}}^T\boldsymbol{z}$, we have $\mathcal{F} = -\phi^\star(\overline{\boldsymbol{u}})$. Let moreover $\theta = \rho/\overline{\Omega}$ with $\overline{\Omega}$ defined in section 3.1. Substituting these, we get

$$\phi^\star(\overline{\boldsymbol{u}}) + \phi(\boldsymbol{z}') - \overline{\boldsymbol{u}}^T\boldsymbol{z}' \ \leq\ (1-\theta)\big[\ \phi^\star(\overline{\boldsymbol{u}}) + \phi(\overline{\boldsymbol{z}}) - \overline{\boldsymbol{u}}^T\overline{\boldsymbol{z}}\ \big]. \tag{31}$$

A simple transformation results in

$$\ell'(\overline{\boldsymbol{u}}) \ \geq\ \phi^\star(\overline{\boldsymbol{u}}) \ -\ (1-\theta)\Big[\ \phi^\star(\overline{\boldsymbol{u}}) + \phi(\overline{\boldsymbol{z}}) - \overline{\boldsymbol{u}}^T\overline{\boldsymbol{z}}\ \Big]. \tag{32}$$

By Observation 4, we have $\overline{\boldsymbol{u}}^T\overline{\boldsymbol{z}} = \phi^\star_k(\overline{\boldsymbol{u}}) + \phi_k(\overline{\boldsymbol{z}})$. Substituting this in the bracketed expression, and taking into account $\phi(\boldsymbol{z}) \leq \phi_k(\boldsymbol{z})$ $(\boldsymbol{z} \in \mathbb{R}^n)$, we get $\phi^\star(\overline{\boldsymbol{u}}) + \phi(\overline{\boldsymbol{z}}) - \overline{\boldsymbol{u}}^T\overline{\boldsymbol{z}} \ \leq\ \phi^\star(\overline{\boldsymbol{u}}) - \phi^\star_k(\overline{\boldsymbol{u}})$, directly yielding (30). □

4.1 Convergence

A more detailed notation will be needed, including indexing of the iterates and support functions. Let $\overline{\boldsymbol{u}}_1, \ldots, \overline{\boldsymbol{u}}_k$ denote the known iterates. Let $\ell_i(\boldsymbol{u})$ $(i = 1, \ldots, k)$ denote approximate support functions at the respective iterates. These take the form $\ell_i(\boldsymbol{u}) = \ell'_i(\boldsymbol{u}) + \ell''_i(\boldsymbol{u})$, where $\ell'_i(\boldsymbol{u})$ is an approximate support function to $\phi^\star(\boldsymbol{u})$, and $\ell''_i(\boldsymbol{u})$ is an exact support function to $\nu(\boldsymbol{u})$, at $\overline{\boldsymbol{u}}_i$.

The model function $d_k(\boldsymbol{u})$ is the upper cover of the support functions $\ell_i(\boldsymbol{u})$ $(i = 1, \ldots, k)$, and the next iterate $\overline{\boldsymbol{u}}_{k+1}$ is a minimizer of $d_k(\boldsymbol{u})$. We then construct an approximate support function $\ell_{k+1}(\boldsymbol{u})$ at $\overline{\boldsymbol{u}}_{k+1}$. Due to the exactness of the ν-part, the inequality (30) of Corollary 9 is inherited with the same constant θ:

$$\ell_{k+1}(\overline{\boldsymbol{u}}_{k+1}) \ \geq\ \theta d(\overline{\boldsymbol{u}}_{k+1}) \ +\ (1-\theta)d_k(\overline{\boldsymbol{u}}_{k+1}). \tag{33}$$

Theorem 10 *Assume that there exists a compact set $\mathcal{O}_{\boldsymbol{u}}$ that contains all the iterates $\overline{\boldsymbol{u}}_k$. Assume moreover that there exists a constant $\theta > 0$ such that all the approximate supporting functions satisfy (33).*

Then the approximate cutting plane method generates a sequence of models and points satisfying

$$\lim_{k\to\infty} D_k = D \qquad \textit{and} \qquad \lim_{k\to\infty} d(\overline{\boldsymbol{u}}_k) = D, \tag{34}$$

where D_k and D are the minima of $d_k(\boldsymbol{u})$ and $d(\boldsymbol{u})$, respectively.

A nice elementary proof of the convergence of the exact cutting plane method can be found in Ruszczyński (2006), Theorem 7.7. The following proof uses some of the ideas presented there. To handle inexactness, I'll need a well-known though strong theorem from convex analysis. Using that, the proof becomes surprisingly simple.

Proof of Theorem 10. We have $d_1(\boldsymbol{u}) \leq d_2(\boldsymbol{u}) \leq \cdots \leq d(\boldsymbol{u})$ $(\boldsymbol{u} \in \mathbb{R}^n)$, hence

$$d_\infty(\boldsymbol{u}) = \lim_{k\to\infty} d_k(\boldsymbol{u}) \qquad (\boldsymbol{u} \in \mathbb{R}^n)$$

exists and is finite. $d_\infty(\boldsymbol{u})$ is a convex function and the sequence of the model functions converges uniformly on the compact set $\mathcal{O}_{\boldsymbol{u}}$ — see, e.g., Theorem 10.8 in Rockafellar's book Rockafellar (1970).

We have $D_k = \min_{\boldsymbol{u}\in\mathbb{R}^n} d_k(\boldsymbol{u}) = d_k(\overline{\boldsymbol{u}}_{k+1})$. Moreover, let $D_\infty = \min_{\boldsymbol{u}\in\mathbb{R}^n} d_\infty(\boldsymbol{u})$. These are finite and $D_k \leq D_\infty \leq D$. With any index k, we have

$$d_k(\overline{\boldsymbol{u}}_{k+1}) \;=\; D_k \;\leq\; D_\infty \;\leq\; d_\infty(\overline{\boldsymbol{u}}_{k+1}). \tag{35}$$

Let us now take into account the uniform convergence of the sequence of the model functions. Given $\epsilon > 0$, there exist a finite number K_ϵ such that $|d_k(\boldsymbol{u}) - d_\infty(\boldsymbol{u})| \leq \epsilon$ holds for $k \geq K_\epsilon$, $\boldsymbol{u} \in \mathcal{O}_{\boldsymbol{u}}$. Specifically, this holds with $\boldsymbol{u} = \overline{\boldsymbol{u}}_{k+1}$, showing that the difference between the left-hand side and the right-hand side of (35) is small for large enough k. It follows that

$$\lim_{k\to\infty} D_k \;=\; D_\infty \;=\; \lim_{k\to\infty} d_\infty(\overline{\boldsymbol{u}}_{k+1}). \tag{36}$$

Now let $(\overline{\boldsymbol{u}}_{k_i+1})$ be a convergent subsequence of $(\overline{\boldsymbol{u}}_{k+1})$, and let $\widetilde{\boldsymbol{u}} := \lim_{i\to\infty} \overline{\boldsymbol{u}}_{k_i+1}$. We have

$$d(\overline{\boldsymbol{u}}_{k_i+1}) \to d(\widetilde{\boldsymbol{u}}) \qquad \text{as} \qquad i \to \infty, \tag{37}$$

due to the continuity of $d(\boldsymbol{u})$.

With any index k, we have $d_\infty(\overline{\boldsymbol{u}}_{k+1}) \geq \ell_{k+1}(\overline{\boldsymbol{u}}_{k+1})$. Taking into account (33), we get

$$d_\infty(\overline{\boldsymbol{u}}_{k+1}) \;\geq\; \theta\, d(\overline{\boldsymbol{u}}_{k+1}) \;+ (1-\theta)\, D_k. \tag{38}$$

Here we have $d_\infty(\overline{\boldsymbol{u}}_{k+1}) \to D_\infty$ and $D_k \to D_\infty$ as $k \to \infty$, according to (36). As for the remaining expression $d(\overline{\boldsymbol{u}}_{k+1})$, we have (37). Taking limits in (38), we get $D_\infty \geq d(\widetilde{\boldsymbol{u}})$. But $d(\widetilde{\boldsymbol{u}}) \geq D \geq D_\infty$ by definition. It follows that $d(\widetilde{\boldsymbol{u}}) = D = D_\infty$ holds.

The proof can be completed in an indirect fashion. Assume that (34) does not hold, i.e., there exists some $\epsilon > 0$ and a subsequence $(\overline{\boldsymbol{u}}_{k_j})$ such that $|d(\overline{\boldsymbol{u}}_{k_j}) - D| > \epsilon$ for every j. But due to the compactness of the domain $\mathcal{O}_{\boldsymbol{u}}$, there exists a convergent subsequence of $(\overline{\boldsymbol{u}}_{k_j})$, which is a contradiction with the argument above. □

Using the terminology of section 2, partial optimal solutions of the true problem (4) and its dual (11) are denoted by $\boldsymbol{z}^\star$ and $\boldsymbol{u}^\star$, respectively.

Corollary 11 *We have $\overline{\boldsymbol{u}}_k \to \boldsymbol{u}^\star$ and $\overline{\boldsymbol{z}}_k \to \boldsymbol{z}^\star$.*

Proof. Strict convexity of $\phi^\star(\boldsymbol{u})$ results in strict convexity of $d(\boldsymbol{u})$. The latter, combined with the second statement of Theorem 10, yields $\overline{\boldsymbol{u}}_k \to \boldsymbol{u}^\star$.

Using the terminology of section 2, $-D_k$ is the maximum of the linear programming dual problem (16) which in turn equals the minimum of the linear programming problem (14), the latter minimum being $\phi_k(\overline{\boldsymbol{z}}_{k+1})$. Similarly, $-D$ is the maximum of the convex dual problem (11) which in turn equals the minimum of the convex problem (4), the latter minimum being $\phi(\boldsymbol{z}^\star)$. Hence the statement $D_k \to D$ of Theorem 10 translates to primal setting in the form

$$\phi_k(\overline{\boldsymbol{z}}_{k+1}) \to \phi(\boldsymbol{z}^\star). \tag{39}$$

As $\phi_k(\boldsymbol{z})$ is an upper approximation of $\phi(\boldsymbol{z})$, we have $\phi_k(\overline{\boldsymbol{z}}_{k+1}) \geq \phi(\overline{\boldsymbol{z}}_{k+1}) \geq \phi(\boldsymbol{z}^\star)$ for every k. Taking into account (39), we get $\phi(\overline{\boldsymbol{z}}_{k+1}) \to \phi(\boldsymbol{z}^\star)$. Strict convexity of $\phi(\boldsymbol{z})$ then yields $\overline{\boldsymbol{z}}_k \to \boldsymbol{z}^\star$. □

5 Gaining traction

For $k = 1, 2, \ldots$, in accordance with former notation, $\overline{\boldsymbol{z}}_k$ denotes a partial optimal solution of the $(k-1)$th primal model problem, and $\overline{\boldsymbol{u}}_k$ denotes an optimal solution of the $(k-1)$th dual model problem.

In the column generation scheme, an improving column is found by approximate minimization of the function $f_k(\boldsymbol{z}) = \phi(\boldsymbol{z}) - \overline{\boldsymbol{u}}_k^T \boldsymbol{z}$. Specifically, $\boldsymbol{z}'_k$ denotes the point found by performing a single line search from $\overline{\boldsymbol{u}}_k$, in the opposite direction of $\overline{\boldsymbol{g}}_k = \nabla f_k(\overline{\boldsymbol{z}}_k)$.

For technical reasons, some further notation is needed. Let $\boldsymbol{z}_k^\circ$ be the exact minimizer of the function $f_k(\boldsymbol{z})$. (This is never computed in the course of the column generation scheme.)

Observation 12 *We have* $\|\overline{\boldsymbol{g}}_k\| \to 0$ *and* $\boldsymbol{z}_k^\circ \to \boldsymbol{z}^\star$.

Proof. As for $\overline{\boldsymbol{g}}_k$, we have

$$\|\overline{\boldsymbol{g}}_k\| = \|\nabla\phi(\overline{\boldsymbol{z}}_k) - \overline{\boldsymbol{u}}_k\| = \|\nabla\phi(\overline{\boldsymbol{z}}_k) - \nabla\phi(\boldsymbol{z}^\star) + \boldsymbol{u}^\star - \overline{\boldsymbol{u}}_k\| \leq \|\nabla\phi(\overline{\boldsymbol{z}}_k) - \nabla\phi(\boldsymbol{z}^\star)\| + \|\boldsymbol{u}^\star - \overline{\boldsymbol{u}}_k\|.$$

The statement $\|\overline{\boldsymbol{g}}_k\| \to 0$ follows from Corollary 11, and $\phi(\boldsymbol{z})$ being continuously differentiable.

By definition, $\boldsymbol{z}_k^\circ$ is the minimizer of $f_k(\boldsymbol{z}) = \phi(\boldsymbol{z}) - \overline{\boldsymbol{u}}_k^T \boldsymbol{z}$, hence $\nabla\phi(\boldsymbol{z}_k^\circ) = \overline{\boldsymbol{u}}_k$. According to Observation 8, the norms $\|\boldsymbol{z}_k^\circ\|$ can be bounded by a bound independent of k. Hence there exists a convergent subsequence, let $\boldsymbol{z}_{k_i}^\circ \to \widehat{\boldsymbol{z}}$. Since $\phi(\boldsymbol{z})$ is continuously differentiable, and $\overline{\boldsymbol{u}}_k \to \boldsymbol{u}^\star$, it follows that $\nabla\phi(\widehat{\boldsymbol{z}}) = \boldsymbol{u}^\star$.

By Observation 3, we have $\boldsymbol{u}^\star = \nabla\phi(\boldsymbol{z}^\star)$, hence $\nabla\phi(\widehat{\boldsymbol{z}}) = \nabla\phi(\boldsymbol{z}^\star)$. From the strict convexity of $\phi(\boldsymbol{z})$, it follows that $\widehat{\boldsymbol{z}} = \boldsymbol{z}^\star$. It means that the sequence $(\boldsymbol{z}_k^\circ)$ has a single accumulation point $\boldsymbol{z}^\star$. □

In the following discussion I assume that $\phi(\boldsymbol{z})$ is well-conditioned in $\boldsymbol{z}^\star$, namely, $\alpha^\star/\omega^\star \gg 0$ holds, where $\omega^\star$ and $\alpha^\star$ denotes the smallest and the largest eigenvalue, respectively, of the Hessian matrix $\nabla^2\phi(\boldsymbol{z}^\star)$.

Let $\alpha(\boldsymbol{z})$ and $\omega(\boldsymbol{z})$ denote the smallest and the largest eigenvalue, respectively, of the Hessian matrix $\nabla^2\phi(\boldsymbol{z})$ as a function of $\boldsymbol{z} \in \mathbb{R}^n$. Continuity of both functions follows from a theorem of Ostrowski that I cite in Appendix B. (This theorem was already applied in section 3.1 to establish continuity of the largest eigenvalue.) It follows that $\phi(\boldsymbol{z})$ is well-conditioned on a neighbourhood of $\boldsymbol{z}^\star$. Specifically, there exists a ball $\mathcal{B}_r$ around $\boldsymbol{z}^\star$ with radius $r > 0$, such that

$$\alpha_r I \preceq \nabla^2\phi(\boldsymbol{z}) \preceq \omega_r I \quad (\boldsymbol{z} \in \mathcal{B}_r) \qquad \text{holds with } \alpha_r/\omega_r \gg 0. \tag{40}$$

As the sequences in Corollary 11 and Observation 12 converge, the line search gains traction.

Theorem 13 *Let (40) hold on a ball* $\mathcal{B}_r$*, drawn around* $\boldsymbol{z}^\star$ *with radius* $r > 0$*. In each iteration* k*, we perform a single line search starting from the current* $\overline{\boldsymbol{z}}_k$*. Let* $\boldsymbol{z}'_k$ *denote the vector found by this line search.*

There exists a natural number K_r *such that*

$$f_k\left(\boldsymbol{z}'_k\right) - \mathcal{F}_k \leq (1 - \alpha_r/\omega_r)\left[\, f_k\left(\overline{\boldsymbol{z}}_k\right) - \mathcal{F}_k \,\right] \qquad \textit{holds with each } k \geq K_r, \tag{41}$$

where $\mathcal{F}_k = \min_{\boldsymbol{z}} f_k(\boldsymbol{z})$.

Here (41) means that the vector $\boldsymbol{z}'_k$ has a substantial reduced cost in the $(k-1)$th master problem, hence we may expect a substantial improvement from including $\boldsymbol{z}'$ as a new column. The proof will be an adaptation of that of Theorem 6, as related in Chapter 8.6 of Luenberger and Ye (2008). It turns out that we only need $\nabla^2\phi(\boldsymbol{z}) \preceq \omega_r I$ for $\boldsymbol{z} = \overline{\boldsymbol{z}}_k - t\overline{\boldsymbol{g}}_k$ such that $0 \leq t \leq \frac{1}{\omega_r}$. Moreover, $\nabla^2\phi(\boldsymbol{z}) \succeq \alpha_r I$ is only needed for $\boldsymbol{z} \in [\overline{\boldsymbol{z}}_k,\, \boldsymbol{z}_k^\circ]$.

Proof of Theorem 13. According to Observation 12, there exists a natural number K_r such that

$$\|\overline{\boldsymbol{z}}_k - \boldsymbol{z}^\star\| \leq \frac{1}{2}\, r, \quad \|\overline{\boldsymbol{g}}_k\| \leq \frac{1}{2}\, r\, \omega_r, \quad \text{and} \quad \|\boldsymbol{z}_k^\circ - \boldsymbol{z}^\star\| \leq r \qquad \text{holds with each } k \geq K_r. \tag{42}$$

In the remaining part of the proof, we consider a fixed $k \geq K_r$. For the sake of simplicity, we are going to omit the lower index k.

Due to the first and the second assumption in (42), we have $\overline{\boldsymbol{z}}_k - t\overline{\boldsymbol{g}}_k \in \mathcal{B}_r$ for $t \in [0,\, 1/\omega_r]$. Let $T = 1/\omega_r$. Hence $\nabla^2\phi(\boldsymbol{z}) \preceq \omega_r I$ holds for $\boldsymbol{z} = \overline{\boldsymbol{z}}_k - t\overline{\boldsymbol{g}})k$ with $t \in [0,\, T]$. It follows that

$$f_k\left(\overline{\boldsymbol{z}}_k - t\overline{\boldsymbol{g}}_k\right) \leq f_k\left(\overline{\boldsymbol{z}}_k\right) - t\,\overline{\boldsymbol{g}}_k^T\overline{\boldsymbol{g}}_k + \frac{\omega_r}{2}t^2\,\overline{\boldsymbol{g}}_k^T\overline{\boldsymbol{g}}_k \qquad (0 \leq t \leq T)$$

holds (a consequence of Taylor's theorem). We consider the respective minima in $t \in \mathbb{R}$ separately of the two sides. The right-hand side is a quadratic expression, yielding minimum at $t = 1/\omega_r \in [0, T]$. It follows that

$$\min_{t \in \mathbb{R}} f_k\,(\overline{\boldsymbol{z}}_k - t\overline{\boldsymbol{g}}_k) \;\leq\; f_k\,(\overline{\boldsymbol{z}}_k) \;-\; \frac{1}{2\omega_r}\,\|\overline{\boldsymbol{g}}_k\|^2. \tag{43}$$

Coming to lower bounds, we have $[\overline{\boldsymbol{z}}_k,\, \boldsymbol{z}_k^\circ] \subset \mathcal{B}_r$ according to the first and the third assumption in (42). From $\nabla^2\phi(\boldsymbol{z}) \succeq \alpha_r I \;\; (\boldsymbol{z} \in [\overline{\boldsymbol{z}},\, \boldsymbol{z}^\circ])$, it follows by Taylor's theorem that

$$f_k(\boldsymbol{z}_k^\circ) \geq f_k(\overline{\boldsymbol{z}}_k) + \overline{\boldsymbol{g}}_k^T(\boldsymbol{z}_k^\circ - \overline{\boldsymbol{z}}_k) + \frac{\alpha_r}{2}\|\boldsymbol{z}_k^\circ - \overline{\boldsymbol{z}}_k\|^2. \tag{44}$$

The left-hand side is $\mathcal{F}_k$ by definition. A lower estimate of the right-hand side is obtained by taking its minimum in $\boldsymbol{z}_k^\circ$. The minimum is easily computed as the right-hand side expression is a quadratic function of $\boldsymbol{z}_k^\circ$. We get

$$\mathcal{F}_k \geq f_k(\overline{\boldsymbol{z}}_k) - \frac{1}{2\alpha_r}\|\overline{\boldsymbol{g}}_k\|^2. \tag{45}$$

The proof is concluded by simple transformations of (43) and (45). The improving column $\boldsymbol{z}_k'$ is found by solving the minimization problem in the left-hand side of (43). Subtracting $\mathcal{F}_k$ from both sides of (23), and applying (45) to underestimate $\|\overline{\boldsymbol{g}}_k\|^2$, we get (41). □

6 Discussion

In Section 5, we assumed that the objective function is well-conditioned in the optimal solution, i.e., $\alpha^\star/\omega^\star \gg 0$ holds with the smallest and largest eigenvalue, respectively, of the Hessian matrix $\nabla^2\phi(\boldsymbol{z}^\star)$. Let me make a case for this assumption by a simple illustration. Figure 1 shows contour lines of a two-dimensional standard normal distribution function, where the covariance between the marginals is 0.5. The contour lines tend to be straight as we move away from the diagonal $z_1 = z_2$. Given a probability maximization problem of the form (1), an optimal solution can be located on the curved part of a contour line, i.e., 'near' the diagonal. The two Figures 2 depict the smaller and the larger eigenvalue, respectively, of the Hessian matrix $\nabla^2[-\log F(\boldsymbol{z})]$ as a function of $\boldsymbol{z}$. In the respective areas not shaded, the smaller eigenvalue is above $1e-5$, and the larger eigenvalue is below 1.6. The function is well-conditioned on the area in which optimal solutions typically belong.

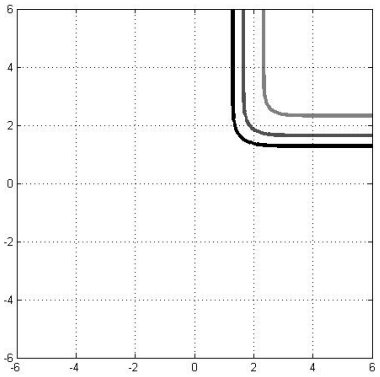

Fig. 1 Contour lines of a two-dimensional normal distribution function $F(\boldsymbol{z})$.

The column generation procedure sketched in Section 2.2 can be viewed as a simplex method applied to a linear programming problem having infinitely many columns. Observing this procedure from a dual viewpoint, we also described it as a cutting-plane method. Though the drawbacks of simplex and cutting-plane

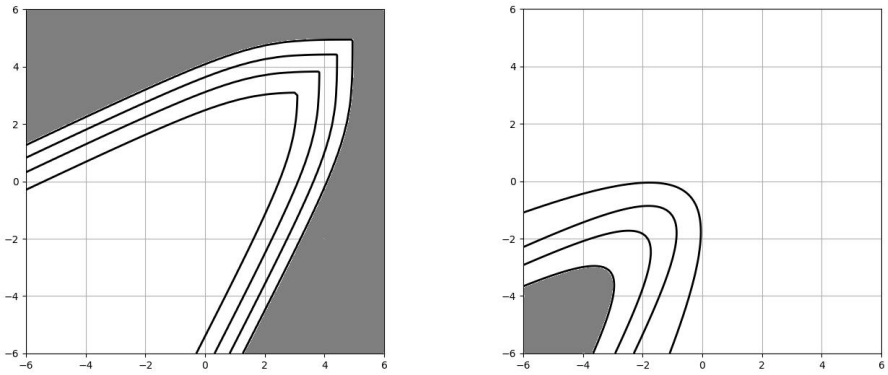

Fig. 2 Contour lines of the smaller and the larger eigenvalue, respectively, of $\nabla^2[-\log F(\boldsymbol{z})]$ as a function of $\boldsymbol{z}$, for a two-dimensional normal distribution function $F(\boldsymbol{z})$.

methods are well-known, they perform fairly well in most of the practical cases (even intriguingly well in many cases). Concerning the simplex method, Borgwardt (1987) and Spielman and Teng (2004) made substantial contributions to bridging the gap between practical effectiveness and theoretical complexity. Concerning the cutting-plane method, Dempster and Merkovsky (1995) present a geometrically convergent version for a continuously differentiable, strictly convex case.

In (5), we regularized the probabilistic objective function by adding the term $\frac{\rho}{2}\|\boldsymbol{z}\|^2$ with an appropriately fixed $\rho > 0$. From a practical point of view, it makes sense to start with a large ρ, and decrease it gradually as the optimal solution is approached.

A Technical proofs

A.1 Bounding the subgradients of the model functions

Before proving Observation 5, we need some preliminary observations.

It is easily seen that the shrunk simplex $\mathcal{S}_\eta$ of (19) can be obtained in the form

$$\mathcal{S}_\eta = \left\{ \sum_{i=0}^{n} \lambda_i \boldsymbol{z}_i \;\middle|\; \sum_{i=0}^{n} \lambda_i = 1,\;\; \lambda_i \geq \frac{1}{n+1}\,(1-\eta) \right\}. \tag{46}$$

Due to $\eta < 1$, the common lower bound of the weights is positive.

Let $\mathcal{R} := \{\, \boldsymbol{\varrho} \in \mathbb{R}^n \mid \|\boldsymbol{\varrho}\| = 1 \,\}$ denote the unit sphere.

Lemma 14 *There exists a positive constant Δ such that*

$$\max_{0 \leq i \leq n} \boldsymbol{\varrho}^T(\boldsymbol{z}_i - \boldsymbol{z}) \;\geq \Delta$$

holds for any $\boldsymbol{z} \in \mathcal{S}_\eta$ and $\boldsymbol{\varrho} \in \mathcal{R}$.

Proof. Let $\boldsymbol{z} = \sum_{i=0}^{n} \lambda_i \boldsymbol{z}_i$ be the representation according to (46). We have

$$\sum_{i=0}^{n} \lambda_i\, \boldsymbol{\varrho}^T(\boldsymbol{z}_i - \boldsymbol{z}) \;= 0, \tag{47}$$

because $\sum \lambda_i(\boldsymbol{z}_i - \boldsymbol{z}) = \boldsymbol{0}$ holds by the definition of the weights λ_i.

As the simplex $\mathcal{S}$ is non-degenerate, we cannot have $\boldsymbol{\varrho}^T(\boldsymbol{z}_i - \boldsymbol{z}) = 0$ for all i. Moreover, as the weights λ_i are all positive, no negative $\boldsymbol{\varrho}^T(\boldsymbol{z}_i - \boldsymbol{z})$ is zeroed out in (47). Hence there exists i such that $\boldsymbol{\varrho}^T(\boldsymbol{z}_i - \boldsymbol{z}) > 0$. It means that the function

$$(\boldsymbol{\varrho}, \boldsymbol{z}) \;\mapsto \max_{0 \leq i \leq n} \boldsymbol{\varrho}^T(\boldsymbol{z}_i - \boldsymbol{z})$$

is positive-valued on the compact set $\mathcal{R} \times \mathcal{S}_\eta$. Since the function is also continuous, it obtains its infimum by Weierstrass' extreme value theorem. Let $\Delta > 0$ denote the minimum. □

Proof of Observation 5. Given $\boldsymbol{z} \in \mathcal{O}_{\boldsymbol{Z}}$, we have $\boldsymbol{z} \in \mathcal{S}_\eta$ according to (19).

Let ϕ denote the current model function value at $\boldsymbol{z}$. Moreover, let $\boldsymbol{u}$ be a subgradient of the current model function at $\boldsymbol{z}$. We put it in the form $\boldsymbol{u} = t\boldsymbol{\varrho}$ with some $t \geq 0$ and $\boldsymbol{\varrho} \in \mathcal{R}$, and will find a bound on $t = \|\boldsymbol{u}\|$.

By the definition of the subgradient, we get

$$t\boldsymbol{\varrho}^T (\boldsymbol{z}_i - \boldsymbol{z}) \;\leq\; \phi_i - \phi \qquad (i = 0, \ldots, n), \tag{48}$$

where ϕ_i denotes the objective value at $\boldsymbol{z}_i$.

In the right-hand sides, we have

$$\phi \;\geq\; \phi(\boldsymbol{z}) \;\geq\; \min_{\boldsymbol{z} \in \mathcal{S}_\eta} \phi(\boldsymbol{z}). \tag{49}$$

The latter minimum exists and is finite as the objective function is continuous and $\mathcal{S}_\eta$ is compact. Hence all the right-hand sides in (48) are bounded from above by $\max_{0 \leq i \leq n} \phi_i - \min_{\boldsymbol{z} \in \mathcal{S}_\eta} \phi(\boldsymbol{z})$.

According to Lemma 14, we have $\boldsymbol{\varrho}^T (\boldsymbol{z}_i - \boldsymbol{z}) > \Delta$ for some i. It gives the upper bound

$$\Gamma_{\boldsymbol{u}} = \;\frac{1}{\Delta}\left(\max_{0 \leq i \leq n} \phi_i - \min_{\boldsymbol{z} \in \mathcal{S}_\eta} \phi(\boldsymbol{z})\right).$$

The statement with the true objective function, $\|\nabla\phi(\boldsymbol{z})\| \leq \Gamma_{\boldsymbol{u}}$ $(\boldsymbol{z} \in \mathcal{S}_\eta)$, can be proved in a similar manner. □

A.2 Bounding improving columns

Proof of Observation 8. A partial optimal solution $\overline{\boldsymbol{z}}$ of the current model problem serves as a starting solution for the column generation subproblem. Hence we have

$$\overline{\boldsymbol{u}}^T \overline{\boldsymbol{z}} - \phi(\overline{\boldsymbol{z}}) \;\leq\; \overline{\boldsymbol{u}}^T \boldsymbol{z}^\circ - \phi(\boldsymbol{z}^\circ). \tag{50}$$

We use the regularized objective (5). In the left-hand side of (50), we have $\phi(\overline{\boldsymbol{z}}) \leq \;\; 1 + \frac{\rho}{2}\|\overline{\boldsymbol{z}}\|^2$, taking into account $F(\overline{\boldsymbol{z}}) \geq 0.5$. In the right-hand side of (50), we have $\phi(\boldsymbol{z}^\circ) \geq \;\frac{\rho}{2}\|\boldsymbol{z}^\circ\|^2$. It follows that

$$\overline{\boldsymbol{u}}^T \overline{\boldsymbol{z}} - 1 - \frac{\rho}{2}\|\overline{\boldsymbol{z}}\|^2 \;\leq\; \overline{\boldsymbol{u}}^T \boldsymbol{z}^\circ - \frac{\rho}{2}\|\boldsymbol{z}^\circ\|^2.$$

Applying the Cauchy-Bunyakovsky-Schwarz inequality in both sides we get

$$-\|\overline{\boldsymbol{u}}\|\|\overline{\boldsymbol{z}}\| - 1 - \frac{\rho}{2}\|\overline{\boldsymbol{z}}\|^2 \;\leq\; \|\overline{\boldsymbol{u}}\|\|\boldsymbol{z}^\circ\| - \frac{\rho}{2}\|\boldsymbol{z}^\circ\|^2.$$

We have $\|\overline{\boldsymbol{u}}\| \leq \Gamma_{\boldsymbol{u}}$ by Observation 5. Moreover, we have $\overline{\boldsymbol{z}} \in \mathcal{O}_{\boldsymbol{Z}}$. The latter being a compact set, there exists a finite $\Gamma_{\boldsymbol{Z}}$ such that $\|\overline{\boldsymbol{z}}\| \leq \Gamma_{\boldsymbol{Z}}$. Hence we get

$$-\Gamma_{\boldsymbol{u}}\Gamma_{\boldsymbol{Z}} - 1 - \frac{\rho}{2}\Gamma_{\boldsymbol{Z}}^2 \;\leq\; \Gamma_{\boldsymbol{u}}\|\boldsymbol{z}^\circ\| - \frac{\rho}{2}\|\boldsymbol{z}^\circ\|^2.$$

This results in the existence of a finite upper bound on $\|\boldsymbol{z}^\circ\|$. □

B On the convergence of eigenvalues

The following theorem is due to Ostrowski. (I learned it from the textbook Szidarovszky (1974) that cited it in Chapter 7.4.)

Theorem 15 (from Appendix K in Ostrowski (1960)) *Given matrices $A, B \in \mathbb{R}^{n \times n}$ having components a_{ij} and b_{ij} $(1 \leq i, j \leq n)$, respectively, let*

$$M = \max_{i,j}\{|a_{ij}|, |b_{ij}|\} \qquad \textit{and} \qquad \delta = \frac{1}{nM}\sum_{i,j}|a_{ij} - b_{ij}|.$$

For any eigenvalue λ of A, there exists an eigenvalue μ of B such that

$$|\lambda - \mu| \leq (n+2)M\delta^{1/n}.$$

References

van Ackooij W, Berge V, de Oliveira W, Sagastizábal C (2017) Probabilistic optimization via approximate p-efficient points and bundle methods. Computers & Operations Research 77:177–193

Borgwardt KH (1987) The simplex method: a probabilistic analysis. Algorithms and Combinatorics, Springer-Verlag, New York

Dempster MAH, Merkovsky RR (1995) A practical geometrically convergent cutting plane algorithm. SIAM Journal on Numerical Analysis 32:631–644

Dentcheva D, Martinez G (2013) Regularization methods for optimization problems with probabilistic constraints. Mathematical Programming 138:223–251

Dentcheva D, Prékopa A, Ruszczyński A (2000) Concavity and efficient points of discrete distributions in probabilistic programming. Mathematical Programming 89:55–77

Dentcheva D, Lai B, Ruszczyński A (2004) Dual methods for probabilistic optimization problems. Mathematical Methods of Operations Research 60:331–346

Fábián CI, Csizmás E, Drenyovszki R, van Ackooij W, Vajnai T, Kovács L, Szántai T (2018) Probability maximization by inner approximation. Acta Polytechnica Hungarica 15:105–125, special issue dedicated to the memory of András Prékopa (editors: A. Bakó, I. Maros and T. Szántai)

Fábián CI, Csizmás E, Drenyovszki R, Vajnai T, Kovács L, Szántai T (2019) A randomized method for handling a difficult function in a convex optimization problem, motivated by probabilistic programming. Annals of Operations Research DOI: 10.1007/s10479-019-03143-z. To appear in S.I.: Stochastic Modeling and Optimization, in memory of András Prékopa (editors: E. Boros, M. Katehakis, A. Ruszczyński). Open access

Frangioni A (2002) Generalized bundle methods. SIAM Journal on Optimization 13:117–156

Frangioni A (2018) Standard bundle methods: untrusted models and duality. Tech. rep., Department of Informatics, University of Pisa, Italy, to appear in Numerical Nonsmooth Optimization, A. Bagirov, M. Gaudioso, N. Karmitsa and M.M. Mäkelä (eds.) Springer

Hantoute A, Henrion R, Pérez-Aros P (2018) Subdifferential characterization of probability functions under Gaussian distribution. Mathematical Programming 174:167–194

Luenberger DG, Ye Y (2008) Linear and Nonlinear Programming. International Series in Operations Research and Management Science, Springer

de Oliveira W, Sagastizábal C (2014) Level bundle methods for oracles with on-demand accuracy. Optimization Methods and Software 29:1180–1209, (Published in Optimization Online in 2012)

Ostrowski A (1960) Solution of Equations and Systems of Equations. Academic Press, New York and London

Prékopa A (1990) Dual method for a one-stage stochastic programming problem with random RHS obeying a discrete probability distribution. ZOR - Methods and Models of Operations Research 34:441–461

Prékopa A, Vizvári B, Badics T (1998) Programming under probabilistic constraint with discrete random variable. In: Giannesi F, Rapcsák T, Komlósi S (eds) New Trends in Mathematical Programming, Kluwer, Dordrecht, pp 235–255

Rockafellar RT (1970) Convex Analysis. Princeton University Press

Ruszczyński A (2006) Nonlinear Optmization. Princeton University Press

Spielman D, Teng SH (2004) Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. Journal of the ACM 51:385–463

Szidarovszky F (1974) Introduction to Numerical Methods (in Hungarian). Közgazdasági és Jogi Könyvkiadó, Budapest