A randomized method for smooth convex minimization, motivated by probability maximization

Csaba I. Fábián* Tamás Szántai†

Abstract

We propose a randomized gradient method – or a randomized cutting-plane method from a dual viewpoint. From the primal viewpoint, our method bears a resemblance to the stochastic approximation family. But in contrast to stochastic approximation, the present method builds a model problem.

Keywords. Convex optimization, stochastic optimization, probabilistic problems.

1 Introduction

We deal with approximate methods for the solution of the problem

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

where $\boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{b} \in \mathbb{R}^r$, and the matrices T and A are of sizes $n \times m$ and $r \times m$, respectively. For the sake of simplicity we assume that the feasible domain is not empty and is bounded.

The motivation for the above form are the classic probability maximization and probabilistic constrained problems, where $\phi(\boldsymbol{z}) = -\log F(\boldsymbol{z})$ with a logconcave distribution function $F(\boldsymbol{z})$. In [7], an inner approximation was proposed for the probabilistic function. This was motivated by the concept of p-efficient points introduced by Prékopa [16]. The approach proved easy to implement and invulnerable to noise in gradient computation. – Noisy gradient estimates may yield iterates that do not improve much on our current model. But we retain a true inner approximation, provided objective values at new iterates are evaluated with appropriate accuracy. – In this paper we present the inner approximation approach in an idealized setting:

Assumption 1 *The function $\phi : \mathbb{R}^n \to \mathbb{R}$ is convex, twice continuously differentiable, and the Hessian matrix $\nabla^2\phi(\boldsymbol{z})$ satisfies*

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

with $\alpha, \omega \in \mathbb{R}$ $(0 < \alpha \leq \omega)$. Here I is the identity matrix, and the relation $U \preceq V$ between matrices means that $V - U$ is positive semidefinite.

We'll present a dual view of the method proposed in [7]. Then we'll propose a randomized version of that method: a randomized gradient method from a primal viewpoint – or a randomized cutting-plane method from a dual viewpoint. But first let us briefly overview a couple of closely related probabilistic programming approaches. – For a broader survey, see [7] and references therein.

Given a distribution and a number p $(0 < p < 1)$, a probabilistic constraint confines search to the level set $\mathcal{L}(F, p) = \{\, \boldsymbol{z} \mid F(\boldsymbol{z}) \geq p \,\}$ of the distribution function $F(\boldsymbol{z})$. Prékopa [16] initiated a novel solution approach

*Department of Informatics, Faculty of Engineering and Computer Science, Kecskemét College, Pallasz Athéné University. Izsáki út 10, 6000 Kecskemét, Hungary. Email: fabian.csaba@gamf.kefo.hu.

†Department of Differential Equations, Faculty of Natural Sciences, Budapest University of Technology and Economics. Műegyetem rkp 3-9, 1111 Budapest, Hungary. Email: szantai@math.bme.hu.

by introducing the concept of p-efficient points. $\boldsymbol{z}$ is p-efficient if $F(\boldsymbol{z}) \geq p$ and there exists no $\boldsymbol{z}'$ such that $\boldsymbol{z}' \leq \boldsymbol{z}$, $\boldsymbol{z}' \neq \boldsymbol{z}$, $F(\boldsymbol{z}') \geq p$. Prékopa, Vizvári, and Badics [18] consider problems with random parameters having a discrete finite distribution. They first enumerate p-efficient points, and based on these, build a convex relaxation of the problem.

Dentcheva, Prékopa, and Ruszczyński [6] formulate the probabilistic constraint in a split form: $T\boldsymbol{x} = \boldsymbol{z}$, where $\boldsymbol{z}$ belongs to the level set $\mathcal{L}(F, p)$; and construct a Lagrangian dual by relaxing the constraint $T\boldsymbol{x} = \boldsymbol{z}$. The dual functional is the sum of two functionals that are respective optimal objective value functions of two simpler problems. The first auxiliary problem is a linear programming problem, and the second one is the minimization of a linear function over the level set $\mathcal{L}(F, p)$. Based on this decomposition, the authors develop a method, called cone generation, that finds new p-efficient points in course of the optimization process.

[7] focusses on probability maximization. A polyhedral approximation is constructed to the epigraph of the probabilistic function. This is analogous to the use of p-efficient points. The dual problem is constructed and decomposed in the manner of [6], but the nonlinear subproblem is easier. In [6], finding a new p-efficient point amounts to minimization over the level set $\mathcal{L}(F, p)$. In contrast, a new approximation point in [7] is found by unconstrained minimization. Moreover, a practical approximation scheme was developed in [7]: instead of exactly solving an unconstrained subproblem occurring during the process, just a single line search was made in each case. Implementation based on this approximation scheme proved quite robust, and a theoretic explanation for this behavior was also found.

The main result of the present paper is extending the approach of [7] to handling gradient estimates.

In Sections 2 and 3 we present a brisk overview of the models and the column generation approach of [7].

In Section 4 we present the column-generation approach from a dual viewpoint, as a cutting-plane method. The advantage of the dual viewpoint is that the cutting-plane method can be regularized. Actual means of regularization are not discussed in this paper, but we present the method in a form that admits of iterate selection rules other than plain cutting-plane method.

The cutting-plane model of the dual approach – like the inner approximation of the primal one – is invulnerable to gradient computation errors. We retain a true inner approximation of the primal objective – or a true outer approximation of the dual objective –, provided objective values at new iterates are evaluated with appropriate accuracy.

This feature facilitates the use of gradient estimates. In Section 5 we extend the method in this direction. The motivation for applying gradient estimates was our computational experience reported in [7]: most of the computation effort was spent in computing gradients. – In that computational study we solved classic probability maximization problems; namely, we had $\phi(\boldsymbol{z}) = -\log F(\boldsymbol{z})$ with a multivariate normal distribution function $F(\boldsymbol{z})$. Given an n-dimensional normal distribution, a component of the gradient $\nabla F(\boldsymbol{z})$ was obtained as the product of an appropriate $(n-1)$-dimensional normal distribution function value and a univariate normal density function value (see the formula in Section 6.6.4 of Prékopa's book [17]). The numerical computation of multivariate normal distribution function values was performed by Genz's subroutine [10]. In our study, most of the computation time was spent in the Genz subroutine. Most demanding were the gradient computations, each requiring n calls to the Genz subroutine. – We conclude that easily computable estimates for the gradients are well worth using, even if the iteration count increases due to estimation errors.

For the estimation of function values and gradients in case of a probabilistic objective, we present a variance reduction Monte Carlo simulation procedure in Section 6. This procedure is applicable to gradient estimation in case of normal, Dirichlet, and t-distributions.

2 Problem and model formulation

In this section we formulate the dual problem and construct polyhedral models of the primal and dual problems. We follow the construction in [7], details can be found in that paper. Though in [7], we exploited

monotonicity of the probabilistic objective and variable splitting was based on $\boldsymbol{z} \leq T\boldsymbol{x}$. In the present paper, we apply the traditional form of variable splitting: Problem (1) will be written as

$$\min\ \phi(\boldsymbol{z}) \quad \text{subject to} \quad A\boldsymbol{x} - \boldsymbol{b} \leq \boldsymbol{0},\ \ \boldsymbol{z} - T\boldsymbol{x} = \boldsymbol{0}. \tag{2}$$

This problem has an optimal solution due to our assumption that on feasible domain of (1). Introducing the multiplier vector $-\boldsymbol{y} \in \mathbb{R}^r, -\boldsymbol{y} \geq \boldsymbol{0}$ to the constraint $A\boldsymbol{x} - \boldsymbol{b} \leq \boldsymbol{0}$, and $-\boldsymbol{u} \in \mathbb{R}^n$ to the constraint $\boldsymbol{z} - T\boldsymbol{x} = \boldsymbol{0}$, the Lagrangian dual of (2) can be written as

$$\max\ \{\boldsymbol{y}^T\boldsymbol{b} - \phi^\star(\boldsymbol{u})\} \quad \text{subject to} \quad (\boldsymbol{y}, \boldsymbol{u}) \in \mathcal{D}, \tag{3}$$

where

$$\mathcal{D} := \ \left\{\ (\boldsymbol{y}, \boldsymbol{u}) \in \mathbb{R}^{r+n}\ \mid\ \boldsymbol{y} \leq \boldsymbol{0},\ \ T^T\boldsymbol{u} = A^T\boldsymbol{y}\ \right\}. \tag{4}$$

According to the theory of convex duality, this problem has an optimal solution.

2.1 Polyhedral models

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

$$\phi_k(\boldsymbol{z}) = \min \sum_{i=0}^{k} \lambda_i \phi_i \quad \text{such that} \quad \lambda_i \geq 0\ (i = 0, \ldots, k),\ \ \sum_{i=0}^{k} \lambda_i = 1, \quad \sum_{i=0}^{k} \lambda_i \boldsymbol{z}_i = \boldsymbol{z}. \tag{5}$$

If $\boldsymbol{z} \notin \mathrm{Conv}(\boldsymbol{z}_0, \ldots, \boldsymbol{z}_k)$, then let $\phi_k(\boldsymbol{z}) := +\infty$. A polyhedral model of Problem (2) is

$$\min\ \phi_k(\boldsymbol{z}) \quad \text{subject to} \quad A\boldsymbol{x} - \boldsymbol{b} \leq \boldsymbol{0},\ \boldsymbol{z} - T\boldsymbol{x} = \boldsymbol{0}. \tag{6}$$

We assume that (6) 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

$$\phi_k^\star(\boldsymbol{u}) = \max_{0 \leq i \leq k} \{\boldsymbol{u}^T\boldsymbol{z}_i - \phi_i\}. \tag{7}$$

As $\phi_k^\star(.)$ is a cutting-plane model of $\phi^\star(.)$, the following problem is a polyhedral model of Problem (3):

$$\max\ \{\boldsymbol{y}^T\boldsymbol{b} - \phi_k^\star(\boldsymbol{u})\} \quad \text{subject to} \quad (\boldsymbol{y}, \boldsymbol{u}) \in \mathcal{D}. \tag{8}$$

2.2 Linear programming formulations

The primal model problem (5)-(6) will be formulated as

$$\begin{array}{lllll}
\min & \displaystyle\sum_{i=0}^{k} \phi_i \lambda_i & & & \\
\text{such that} & \lambda_i \geq 0 & & (i = 0, \ldots, k), & \\
& \displaystyle\sum_{i=0}^{k} \lambda_i & & = 1, & \\
& \displaystyle\sum_{i=0}^{k} \lambda_i \boldsymbol{z}_i & -T\boldsymbol{x} & = \boldsymbol{0}, & \\
& & A\boldsymbol{x} & \leq \boldsymbol{b}. &
\end{array} \tag{9}$$

The dual model problem (7)-(8), formulated as a linear programming problem, is just the LP dual of (9):

$$\begin{array}{lllll} \max & \vartheta & + & \boldsymbol{b}^T\boldsymbol{y} & \\ \text{such that} & & & \boldsymbol{y} & \leq \mathbf{0}, \\ & \vartheta & + \boldsymbol{z}_i^T\boldsymbol{u} & \leq \phi_i & (i=0,\ldots,k), \\ & & -T^T\boldsymbol{u} \quad + A^T\boldsymbol{y} & = \mathbf{0}. & \end{array} \tag{10}$$

Let $(\overline{\lambda}_0,\ldots,\overline{\lambda}_k,\overline{\boldsymbol{x}})$ and $(\overline{\vartheta},\overline{\boldsymbol{u}},\overline{\boldsymbol{y}})$ denote respective optimal solutions of the problems (9) and (10) – both existing due to our assumption concerning the feasibility of (6) and hence (9). Let moreover

$$\overline{\boldsymbol{z}} = \sum_{i=0}^{k} \overline{\lambda}_i \boldsymbol{z}_i. \tag{11}$$

Observation 2 *We have* $\phi_k(\overline{\boldsymbol{z}}) = \sum_{i=0}^{k} \phi_i\overline{\lambda}_i = \overline{\vartheta} + \overline{\boldsymbol{u}}^T\overline{\boldsymbol{z}}$.

The first equality follows from the equivalence of (9) on the one hand, and (5)-(6) on the other hand. The second equality is a straight consequence of complementarity.

Observation 3 *We have* $\overline{\vartheta} = -\phi_k^\star(\overline{\boldsymbol{u}})$.

This follows from the equivalence between (10) on the one hand and (7)-(8) on the other hand.

Remark 4 *A consequence of Observations 2 and 3 is* $\phi_k(\overline{\boldsymbol{z}}) + \phi_k^\star(\overline{\boldsymbol{u}}) = \overline{\boldsymbol{u}}^T\overline{\boldsymbol{z}}$. *This is Fenchel's equality between* $\overline{\boldsymbol{u}}$ *and* $\overline{\boldsymbol{z}}$, *with respect to the model function* $\phi_k(.)$.

3 Primal viewpoint: column generation

In [7], the probability maximization problem is solved by iteratively adding improving columns to the primal model. In this section we give a brisk overview of the practical approximation scheme proposed in that paper.

An optimal dual solution (i.e., shadow price vector) of the current model problem is $(\overline{\vartheta},\overline{\boldsymbol{u}},\overline{\boldsymbol{y}})$. Given a vector $\boldsymbol{z}\in\mathbb{R}^n$, we can add a new column in (9), corresponding to $\boldsymbol{z}_{k+1}=\boldsymbol{z}$. This is an improving column if its reduced cost

$$\overline{\rho}(\boldsymbol{z}) := \overline{\vartheta} + \overline{\boldsymbol{u}}^T\boldsymbol{z} - \phi(\boldsymbol{z}) \tag{12}$$

is positive. – It is easily seen that the reduced cost of $\overline{\boldsymbol{z}}$ is non-negative. Indeed,

$$\overline{\rho}(\overline{\boldsymbol{z}}) \geq \overline{\vartheta} + \overline{\boldsymbol{u}}^T\overline{\boldsymbol{z}} - \phi_k(\overline{\boldsymbol{z}}) = 0 \tag{13}$$

follows from $\phi_k(.)\geq\phi(.)$ and Observation 2.

Let $\overline{\mathcal{R}} := \max_{\boldsymbol{z}}\overline{\rho}(\boldsymbol{z})$. We have $\overline{\mathcal{R}} = \overline{\vartheta} + \phi^\star(\overline{\boldsymbol{u}})$ by the definition of the conjugate function. If $\overline{\mathcal{R}}$ is small, then $(\overline{\boldsymbol{x}},\overline{\boldsymbol{z}})$ is a near-optimal solution to (2). Otherwise an improving column can be constructed to (9). The column with the largest reduced cost can, in theory, be found by a steepest descent method applied to the function $-\overline{\rho}(\boldsymbol{z})$. Though finding a near-optimal solution proved rather time-consuming in the computational study of [7]. As a practical alternative, only a single line search was performed, starting from $\overline{\boldsymbol{z}}$. This simple method proved effective and robust. Moreover, a theoretical explanation was also found for the efficiency of the approach, based on the following well-known theorem:

Theorem 5 *Let Assumption 1 hold for the function $f : \mathbb{R}^n \to \mathbb{R}$. Let us 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*

$$f\left(\boldsymbol{z}^j\right) - \mathcal{F} \;\leq\; \left(1 - \frac{\alpha}{\omega}\right)^j \left[\, f\left(\boldsymbol{z}^0\right) - \mathcal{F} \,\right], \tag{14}$$

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

Proof of this theorem can be found in, e.g., [20], [13]. The following corollary was obtained in [7]:

Corollary 6 *Let β $(0 < \beta \ll 1)$ be given. A finite (and moderate) number of steps with the steepest descent method results a vector $\widehat{\boldsymbol{z}}$ such that*

$$\overline{\rho}\,(\widehat{\boldsymbol{z}}) \;\geq\; (1 - \beta)\,\overline{\mathcal{R}}.$$

In the context of the simplex method, the Markowitz rule is a well-known and often-used rule of column selection. The Markowitz rule always selects the vector with the largest reduced cost. In this view, the vector $\widehat{\boldsymbol{z}}$ in Corollary 6 is a fairly good improving vector in the column generation scheme.

In the computational study of [7], just a single line search was performed in each reduced cost maximization; i.e., $j = 1$ according to the notation of Theorem 5. (Even this single line search was inexact, making a limited number of steps in the direction of steepest ascent.) Our implementation proved reliable even with this simple procedure.

In case of probabilistic functions, Assumption 1 does not hold for every $\boldsymbol{z} \in \mathbb{R}^n$. Our computational experience in [7] was, however, that the probabilistic objectives were well conditioned over certain domains. The iterates obtained by the above approximative procedure always remained in the respective safe domains.

4 Dual viewpoint: cutting planes

The simplex method can be viewed as a cutting-plane method. This fact has been part of the professional folklore ever since the simplex method became widely known. Simplex and cutting-plane methods are parallelly discussed in Section 3.4 of Prékopa's book [17]. A closer description of the present situation can be found in [8], Section 4.

4.1 Dimension reduction

For the sake of simplicity, let us make

Assumption 7 *The inequality system $A\boldsymbol{x} \leq \boldsymbol{b}$ contains box constraints in the form of $\underline{\boldsymbol{b}} \leq \boldsymbol{x} \leq \overline{\boldsymbol{b}}$, where $\underline{\boldsymbol{b}}, \overline{\boldsymbol{b}} \in \mathbb{R}^r$ are given vectors $(\underline{\boldsymbol{b}} \leq \overline{\boldsymbol{b}})$.*

I.e., we have

$$A = \begin{pmatrix} A' \\ I \\ -I \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}' \\ \overline{\boldsymbol{b}} \\ -\underline{\boldsymbol{b}} \end{pmatrix}. \tag{15}$$

Let

$$\mu(\boldsymbol{u}) \;\; = \min \{\, -\boldsymbol{b}^T\boldsymbol{y} \mid (\boldsymbol{y}, \boldsymbol{u}) \in \mathcal{D} \,\}. \tag{16}$$

This function is defined for every $\boldsymbol{u}$ since $T^T\boldsymbol{u} = A^T\boldsymbol{y}$ is solvable in $\boldsymbol{y}$ $(\boldsymbol{y} \leq \boldsymbol{0})$ due to A having the form (15). We will formulate the dual problems using the function $\mu(\boldsymbol{u})$. For convenience, we transform the problems into minimization forms. The original dual problem (3) assumes the form

$$\min \{\phi^\star(\boldsymbol{u}) + \mu(\boldsymbol{u})\}, \tag{17}$$

and the model problem (8) assumes the form

$$\min \{\phi_k^\star(\boldsymbol{u}) + \mu(\boldsymbol{u})\}. \tag{18}$$

4.2 Cut generation

Given $\widehat{\boldsymbol{u}} \in \mathbb{R}^n$, we are going to compute an approximate support function to $\phi^\star(\boldsymbol{u})$. This will be of the form

$$\hat{\ell}(\boldsymbol{u}) := \boldsymbol{u}^T \widehat{\boldsymbol{z}} - \phi\left(\widehat{\boldsymbol{z}}\right), \tag{19}$$

with an appropriate vector $\widehat{\boldsymbol{z}}$. We have $\hat{\ell}(\boldsymbol{u}) \leq \phi^\star(\boldsymbol{u})$ for any $\boldsymbol{u}$ by the definition of $\phi^\star(\boldsymbol{u})$. We are going to compute $\widehat{\boldsymbol{z}}$ such that the gap $\phi^\star(\widehat{\boldsymbol{u}}) - \hat{\ell}(\widehat{\boldsymbol{u}})$ be relatively small.

Using support functions of the above form, a cutting-plane scheme for the problem (17) is easily implemented. We build a polyhedral model $\phi_k^\star(\boldsymbol{u})$ of $\phi^\star(\boldsymbol{u})$ – always adding the appropriate $\widehat{\boldsymbol{z}}$ vector to the dual model as $\boldsymbol{z}_{k+1}$. On the other hand, we will work with $\mu(\boldsymbol{u})$ as a polyhedral model of itself. This is a workable setup; the current model function can be minimized by just solving the linear programming problem (10) – take into account Observation 3.

Let $\Psi^\star$ denote the minimum of (17). Moreover let $\overline{\boldsymbol{u}}$ be a minimizer of (18), in accordance with our notation in former sections. Let us further introduce the notation

$$\underline{\Psi} := \phi_k^\star(\overline{\boldsymbol{u}}) + \mu(\overline{\boldsymbol{u}}) \qquad \text{and} \qquad \widehat{\Psi} := \phi^\star(\widehat{\boldsymbol{u}}) + \mu(\widehat{\boldsymbol{u}}). \tag{20}$$

Obviously we have $\underline{\Psi} \leq \Psi^\star \leq \widehat{\Psi}$.

Coming back to the construction of the support function (19), a deep cut at $\widehat{\boldsymbol{u}}$ – relative to the model function $\phi_k^\star(\boldsymbol{u})$ – can be obtained by setting $\widehat{\boldsymbol{z}}$ to be a near-maximizer of

$$\widehat{\rho}(\boldsymbol{z}) \;=\; \widehat{\boldsymbol{u}}^T \boldsymbol{z} - \phi(\boldsymbol{z}) \; - \phi_k^\star(\widehat{\boldsymbol{u}}). \tag{21}$$

Let $\widehat{\mathcal{R}} := \max_{\boldsymbol{z}} \widehat{\rho}(\boldsymbol{z})$. We have

$$\widehat{\mathcal{R}} = \phi^\star(\widehat{\boldsymbol{u}}) - \phi_k^\star(\widehat{\boldsymbol{u}}) \tag{22}$$

by the definition of the conjugate function. As for the gap between the upper and the lower bound, we have

$$\begin{aligned} \widehat{\Psi} - \underline{\Psi} \;&=\; \phi^\star(\widehat{\boldsymbol{u}}) - \phi_k^\star(\widehat{\boldsymbol{u}}) \; + \phi_k^\star(\widehat{\boldsymbol{u}}) - \phi_k^\star(\overline{\boldsymbol{u}}) \; + \big(\mu(\widehat{\boldsymbol{u}}) - \mu(\overline{\boldsymbol{u}})\big) \\ &=\; \widehat{\mathcal{R}} \; + \big(\phi_k^\star(\widehat{\boldsymbol{u}}) - \phi_k^\star(\overline{\boldsymbol{u}})\big) \; + \big(\mu(\widehat{\boldsymbol{u}}) - \mu(\overline{\boldsymbol{u}})\big), \end{aligned} \tag{23}$$

from (20) and (22). We are going to work with inexact cuts. Let us first consider a corollary of Theorem 5.

Corollary 8 *Let β $(0 < \beta \ll 1)$ be given. A finite (and moderate) number of steps with the steepest descent method results a vector $\widehat{\boldsymbol{z}}$ such that*

$$\widehat{\rho}\left(\widehat{\boldsymbol{z}}\right) \;\geq\; (1-\beta)\, \widehat{\mathcal{R}} + \beta \widehat{\rho}\left(\overline{\boldsymbol{z}}\right). \tag{24}$$

Proof. Substituting $f(\boldsymbol{z}) = -\widehat{\rho}(\boldsymbol{z})$ and $\boldsymbol{z}^0 = \overline{\boldsymbol{z}}$ in (14), and introducing the notation $\varrho = 1 - \alpha/\omega$, we get

$$\widehat{\mathcal{R}} - \widehat{\rho}\left(\boldsymbol{z}^j\right) \;\leq\; \varrho^j \left[\, \widehat{\mathcal{R}} - \widehat{\rho}\left(\overline{\boldsymbol{z}}\right) \,\right].$$

(We have $\mathcal{F} = -\widehat{\mathcal{R}}$ by definition.) Selecting j such that $\varrho^j \leq \beta$ yields an appropriate $\widehat{\boldsymbol{z}} = \boldsymbol{z}^j$. □

Proposition 9 *Let β $(0 < \beta \ll 1)$ be given. We can construct a linear function $\hat{\ell}(\boldsymbol{u})$ such that*

$$\hat{\ell}(\boldsymbol{u}) \leq \phi^\star(\boldsymbol{u}) \quad \textit{holds for any } \boldsymbol{u}, \textit{ and}$$

$$\hat{\ell}\left(\widehat{\boldsymbol{u}}\right) \geq \phi^\star\left(\widehat{\boldsymbol{u}}\right) - \beta \left[\, \widehat{\mathcal{R}} - \widehat{\rho}(\overline{\boldsymbol{z}}) \,\right].$$

Proof. According to Corollary 8, we can construct a vector $\widehat{\boldsymbol{z}}$ such that (24) holds. Using this $\widehat{\boldsymbol{z}}$, let us define $\hat{\ell}(\boldsymbol{u})$ according to (19). Then we have

$$\begin{aligned} \hat{\ell}\,(\widehat{\boldsymbol{u}}) &= \widehat{\rho}\,(\widehat{\boldsymbol{z}}) + \phi_k^\star\,(\widehat{\boldsymbol{u}}) && \text{due to (21)} \\ &\geq (1-\beta)\,\widehat{\mathcal{R}} + \beta\widehat{\rho}(\overline{\boldsymbol{z}}) \; + \phi_k^\star\,(\widehat{\boldsymbol{u}}) && \text{due to (24)} \\ &= \phi^\star\,(\widehat{\boldsymbol{u}}) - \beta\,\widehat{\mathcal{R}} + \beta\widehat{\rho}(\overline{\boldsymbol{z}}) && \text{due to (22).} \end{aligned} \tag{25}$$

4.3 A special case

In the remaining part of this paper we consider a special case:

Assumption 10 *The next iterate $\widehat{u}$ is always selected in such a way that*

$$\widehat{\rho}(\overline{\boldsymbol{z}}) \;=\; \widehat{\boldsymbol{u}}^T\overline{\boldsymbol{z}} - \phi(\overline{\boldsymbol{z}}) - \phi_k^\star(\widehat{\boldsymbol{u}}) \;\geq 0. \tag{26}$$

In words, (26) means that adding the support function (19: $\widehat{\boldsymbol{z}} = \overline{\boldsymbol{z}}$) would improve the model function $\phi_k^\star(\boldsymbol{u})$ at $\widehat{\boldsymbol{u}}$.

According to (13), this assumption holds in case of the plain cutting-plane method, where the next iterate is simply the minimizer of the model problem (18); namely, we have $\widehat{\boldsymbol{u}} = \overline{\boldsymbol{u}}$ and hence $\widehat{\rho}(\overline{\boldsymbol{z}}) = \overline{\rho}(\overline{\boldsymbol{z}})$.

Under Assumption 10, we can simplify Corollary 8 by deleting the term $\beta\,\widehat{\rho}\,(\overline{\boldsymbol{z}})$ in the right-hand side of (24). In a similar manner, Proposition 9 will assume the form

Proposition 11 *Let β $(0 < \beta \ll 1)$ be given. We can construct a linear function $\hat{\ell}(\boldsymbol{u})$ such that*

$$\hat{\ell}(\boldsymbol{u}) \leq \phi^\star(\boldsymbol{u}) \quad \textit{holds for any } \boldsymbol{u}, \;\; \textit{and}$$

$$\hat{\ell}\,(\widehat{\boldsymbol{u}}) \geq \phi^\star\,(\widehat{\boldsymbol{u}}) - \beta\,\widehat{\mathcal{R}}.$$

In words: $\hat{\ell}(\boldsymbol{u})$ is an approximate support function to $\phi^\star(\boldsymbol{u})$ at $\widehat{\boldsymbol{u}}$. The difference between the function values at the current iterate is bounded by the portion $\beta\,\widehat{\mathcal{R}}$ of the gap.

5 Working with gradient estimates

In this section we show that the column generation scheme of [7] (sketched in Section 3), and the cutting-plane scheme of Section 4 can be implemented as a randomized method using gradient estimates.

We wish to minimize $-\widehat{\rho}(\boldsymbol{z})$ over $\mathbb{R}^n$. Given $\boldsymbol{z}^\circ \in \mathbb{R}^n$, let $\boldsymbol{g}^\circ = -\nabla\overline{\rho}(\boldsymbol{z}^\circ)$.

Assumption 12 *Let $\sigma > 0$ be given. We can construct a realization of a random vector $\boldsymbol{G}^\circ$ satisfying*

$$E(\boldsymbol{G}^\circ) = \boldsymbol{g}^\circ \qquad \textit{and} \qquad E\left(\left\|\boldsymbol{G}^\circ - \boldsymbol{g}^\circ\right\|^2\right) \;\leq\; \sigma\left\|\boldsymbol{g}^\circ\right\|^2. \tag{27}$$

From (27) follows

$$\mathrm{E}\left(\left\|\boldsymbol{G}^\circ\right\|^2\right) \;=\; \mathrm{E}\left(\left\|\boldsymbol{G}^\circ - \boldsymbol{g}^\circ\right\|^2\right) + \left\|\boldsymbol{g}^\circ\right\|^2 \;\leq\; (\sigma+1)\,\left\|\boldsymbol{g}^\circ\right\|^2. \tag{28}$$

Let us consider the following randomized form of Theorem 5:

Theorem 13 *Let Assumption 1 hold for the function $f : \mathbb{R}^n \to \mathbb{R}$. Let us minimize $f(\boldsymbol{z})$ over $\mathbb{R}^n$. We perform a steepest descent method using gradient estimates. (Given an iterate $\boldsymbol{z}^\circ$, a gradient estimate $\boldsymbol{G}^\circ$ is generated and a line search is performed in that direction.) We assume that gradient estimates at the respective iterates are generated independently, and (27) - (28) hold for each of them.*

Having started from the point $\boldsymbol{z}^0$, and having performed j line searches, let $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^j$ denote the respective iterates. Then we have

$$E\left[f\left(\boldsymbol{z}^j\right)\right] - \mathcal{F} \;\leq\; \left(1 - \frac{\alpha}{\omega(\sigma+1)}\right)^j \left(f\left(\boldsymbol{z}^0\right) \;- \mathcal{F}\right), \tag{29}$$

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

Proof. Let $\boldsymbol{G}^0, \ldots, \boldsymbol{G}^{j-1}$ denote the respective gradient estimates for the iterates $\boldsymbol{z}^0, \ldots, \boldsymbol{z}^{j-1}$.

To begin with, let us focus on the first line search whose starting point is $\boldsymbol{z}^\circ = \boldsymbol{z}^0$. Here $\boldsymbol{z}^\circ$ is a given (not random) vector. We are going to adopt the usual proof of Theorem 5 to employing the gradient estimate $\boldsymbol{G}^\circ$ instead of the gradient $\boldsymbol{g}^\circ$. From $\nabla^2 f(\boldsymbol{z}) \preceq \omega I$, it follows that

$$f\left(\boldsymbol{z}^\circ - t\boldsymbol{G}^\circ\right) \;\leq\; f\left(\boldsymbol{z}^\circ\right) \;-\; t\,\boldsymbol{g}^{\circ\,T}\boldsymbol{G}^\circ \;+\; \frac{\omega}{2}t^2\,\boldsymbol{G}^{\circ\,T}\boldsymbol{G}^\circ$$

holds for any $t \in \mathbb{R}$. Considering expectations in both sides, we get

$$\begin{aligned}\mathrm{E}\left[f\left(\boldsymbol{z}^\circ - t\boldsymbol{G}^\circ\right)\right] \;&\leq\; f\left(\boldsymbol{z}^\circ\right) \;-\; t\,\left\|\boldsymbol{g}^\circ\right\|^2 \;+\; \tfrac{\omega}{2}t^2\,\mathrm{E}\left(\left\|\boldsymbol{G}^\circ\right\|^2\right) \\ &\leq\; f\left(\boldsymbol{z}^\circ\right) \;-\; t\,\left\|\boldsymbol{g}^\circ\right\|^2 \;+\; \tfrac{\omega}{2}t^2\,(\sigma+1)\,\left\|\boldsymbol{g}^\circ\right\|^2\end{aligned}$$

according to (28). Let us consider the respective minima in t separately of the two sides. The right-hand side is a quadratic expression, yielding minimum at $t = \frac{1}{\omega(\sigma+1)}$. Inequality is inherited to minima, hence

$$\min_t \mathrm{E}\left[f\left(\boldsymbol{z}^\circ - t\boldsymbol{G}^\circ\right)\right] \;\leq\; f\left(\boldsymbol{z}^\circ\right) \;-\; \frac{1}{2\omega(\sigma+1)}\,\left\|\boldsymbol{g}^\circ\right\|^2. \tag{30}$$

For the left-hand side, we obviously have

$$\mathrm{E}\left[\,\min_t\, f\left(\boldsymbol{z}^\circ - t\boldsymbol{G}^\circ\right)\right] \;\leq\; \min_t \mathrm{E}\left[f\left(\boldsymbol{z}^\circ - t\boldsymbol{G}^\circ\right)\right]. \tag{31}$$

(This is analogous to the basic inequality comparing the wait-and-see and the here-and-now approaches for classic two-stage stochastic programing problems.)

Let $\boldsymbol{z}'$ denote the minimizer of the line search in the left-hand side of (31), i.e., $f\left(\boldsymbol{z}'\right) = \min_t\, f\left(\boldsymbol{z}^\circ - t\boldsymbol{G}^\circ\right)$. (Of course $\boldsymbol{z}'$ is a random vector since it depends on $\boldsymbol{G}^\circ$.) Substituting this in (31) and comparing with (30), we get

$$\mathrm{E}\left[f\left(\boldsymbol{z}'\right)\right] \;\leq\; f\left(\boldsymbol{z}^\circ\right) \;-\; \frac{1}{2\omega(\sigma+1)}\,\left\|\boldsymbol{g}^\circ\right\|^2.$$

Subtracting $\mathcal{F}$ from both sides results

$$\mathrm{E}\left[f\left(\boldsymbol{z}'\right)\right] - \mathcal{F} \;\leq\; f\left(\boldsymbol{z}^\circ\right) - \mathcal{F} \;-\; \frac{1}{2\omega(\sigma+1)}\,\left\|\boldsymbol{g}^\circ\right\|^2. \tag{32}$$

Coming to the lower bound, a well-known consequence of $\alpha I \preceq \nabla^2 f(\boldsymbol{z})$ is

$$\left\|\boldsymbol{g}^\circ\right\|^2 \;\geq\; 2\alpha\,\left(f\left(\boldsymbol{z}^\circ\right) \;- \mathcal{F}\right)$$

(see the classic proof of Theorem 5). Combining this with (32), we get

$$\mathrm{E}\left[f\left(\boldsymbol{z}'\right)\right] - \mathcal{F} \;\leq\; f\left(\boldsymbol{z}^\circ\right) - \mathcal{F} \;-\; \frac{\alpha}{\omega(\sigma+1)}\,\left(f\left(\boldsymbol{z}^\circ\right) \;- \mathcal{F}\right) \;=\; \left(1 - \frac{\alpha}{\omega(\sigma+1)}\right)\left(f\left(\boldsymbol{z}^\circ\right) \;- \mathcal{F}\right). \tag{33}$$

As we have assumed that $\boldsymbol{z}^\circ$ is a given (not random) vector, the right-hand side of (33) is deterministic, and the expectation in the left-hand side is considered according to the distribution of $\boldsymbol{G}^\circ$.

Let us now examine the $(l+1)$th line search (for $1 \leq l \leq j-1$) where the starting point is $\boldsymbol{z}^\circ = \boldsymbol{z}^l$ and the minimizer is $\boldsymbol{z}' = \boldsymbol{z}^{l+1}$. Of course (33) holds with these objects also, but now both sides are random variables, depending on the vectors $\boldsymbol{G}^0, \ldots, \boldsymbol{G}^{l-1}$. (The expectation in the left-hand side is a conditional expectation.) Let us consider the respective expectations of the two sides, according to the joint distribution of $\boldsymbol{G}^0, \ldots, \boldsymbol{G}^{l-1}$. As the random gradient vectors were generated independently, we get

$$\mathrm{E}\left[f\left(\boldsymbol{z}^{l+1}\right)\right] - \mathcal{F} \;\leq\; \left(1 - \frac{\alpha}{\omega(\sigma+1)}\right)\left(\mathrm{E}\left[f\left(\boldsymbol{z}^{l}\right)\right] \;- \mathcal{F}\right), \tag{34}$$

where the left-hand expectation is now taken according to the joint distribution of $\boldsymbol{G}^0, \ldots, \boldsymbol{G}^l$. – This technique of proof is well known in the context of stochastic gradient schemes.

Finally, (29) follows from the iterative application of (34). □

Under Assumption 10, we have

Corollary 14 *Let a tolerance β $(0 < \beta \ll 1)$ and a probability p $(0 < p \ll 1)$ be given. A finite (and moderate) number of steps with the above randomized steepest descent method results a vector $\widehat{\boldsymbol{z}}$ such that*

$$P\Big(\;\widehat{\rho}\,(\widehat{\boldsymbol{z}}\;) \;\geq\; (1-\beta)\widehat{\mathcal{R}}\Big) \;\geq\; 1-p.$$

I.e., with a high probability, $\widehat{\boldsymbol{z}}$ is a fairly good improving vector in the column generation scheme.

Proof. Substituting $f(\boldsymbol{z}) = -\widehat{\rho}(\boldsymbol{z})$ and $\boldsymbol{z}^0 = \overline{\boldsymbol{z}}$ in (29) and taking into account Assumption 10, we get

$$\mathrm{E}\left[\widehat{\rho}\,(\boldsymbol{z}^j)\right] \;\geq\; \left(1-\varrho^j\right)\,\widehat{\mathcal{R}}$$

with $\varrho = 1 - \frac{\alpha}{\omega(\sigma+1)}$. We have $\widehat{\mathcal{R}} \geq 0$ by (22). In case $\widehat{\mathcal{R}} = 0$, the starting iterate $\boldsymbol{z}^0 = \overline{\boldsymbol{z}}$ of the steepest descent method was already optimal, due to Assumption 10. In what follows we assume $\widehat{\mathcal{R}} > 0$. A trivial transformation results

$$\mathrm{E}\left[1 - \frac{\widehat{\rho}\,(\boldsymbol{z}^j)}{\widehat{\mathcal{R}}}\right] \;\leq\; \varrho^j.$$

By Markov's inequality, we get

$$\mathrm{P}\left(\;1 - \frac{\widehat{\rho}\,(\boldsymbol{z}^j)}{\widehat{\mathcal{R}}} \geq \beta\right) \;\leq\; \frac{\varrho^j}{\beta},$$

and a trivial transformation yields

$$\mathrm{P}\Big(\;\widehat{\rho}\,(\boldsymbol{z}^j\;) \;\leq\; (1-\beta)\widehat{\mathcal{R}}\Big) \;\leq\; \frac{1}{\beta}\,\varrho^j.$$

Hence

$$\mathrm{P}\Big(\;\widehat{\rho}\,(\boldsymbol{z}^j\;) \;>\; (1-\beta)\widehat{\mathcal{R}}\Big) \;\geq\; 1 - \frac{1}{\beta}\,\varrho^j.$$

Selecting j such that $\varrho^j \leq \beta p$ yields an appropriate $\widehat{\boldsymbol{z}} = \boldsymbol{z}^j$. □

Remark 15 *We work with gradients of $f(\boldsymbol{z}) = -\widehat{\rho}(\boldsymbol{z})$. By the definition of the latter function, we have*

$$\nabla f(\boldsymbol{z}) = \nabla\phi(\boldsymbol{z}) - \overline{\boldsymbol{u}}.$$

As the procedure progresses, the difference $\nabla\phi(\boldsymbol{z}^j) - \overline{\boldsymbol{u}}$ gets small. To satisfy the requirement (27) on variance, better and better estimates are needed.

5.1 Random cutting planes

Under Assumption 10 we can apply random cutting planes in the cutting-plane scheme of Section 4:

Proposition 16 *Let a tolerance β $(0 < \beta \ll 1)$ and a probability p $(0 < p \ll 1)$ be given. We can construct a random linear function $\hat{\ell}(\boldsymbol{u})$ such that*

$$\hat{\ell}(\boldsymbol{u}) \leq \phi^{\star}(\boldsymbol{u}) \quad \textit{holds for any } \boldsymbol{u}, \quad \textit{and}$$

$$P\Big(\hat{\ell}\,(\overline{\boldsymbol{u}}) \;\geq\; \phi^{\star}\,(\overline{\boldsymbol{u}}) - \beta\,\widehat{\mathcal{R}} \Big) \;\geq\; 1 - p.$$

The proof is the same as those of Propositions 9 and 11, but use Corollary 14 instead of Corollary 8.

6 Easily computable estimates of function values and gradients

In this section we present a variance reduction Monte Carlo simulation procedure for the estimation of multivariate probability distribution function values. The procedure was proposed in Szántai's thesis [21], and presented in Section 6.5 of Prékopa's book [17].

This procedure can be applied to any multivariate probability distribution function. The only condition is that we have to be able to calculate one- and two-dimensional marginal probability distribution function values. Accuracy is easily controlled by sample size. The procedure can be applied to evaluate objective values at new iterates with appropriate accuracy.

All these algorithms can be applied to gradient estimation if the multivariate probability distribution at issue has conditional probability distributions of its own type. E.g., the multivariate normal, the multivariate t-distribution and the Dirichlet distribution have this property. In such cases way we can effectively construct gradient estimates satisfying Assumption 12.

As we have

$$F(x_1, \ldots, x_n) = P\left(\xi_1 < x_1, \ldots, \xi_n < x_n\right) = 1 - P\left(\{\xi_1 \geq x_1\} + \cdots + \{\xi_n \geq x_n\}\right) = 1 - P(\overline{A}_1 + \cdots + \overline{A}_n),$$

where

$$\overline{A}_i = \{\xi_i \geq x_i\}\,, i = 1, \ldots, n,$$

we can apply bounding and simulation results for the probability of union of events.

6.1 Crude Monte Carlo simulation

If μ denotes the number of those events which occur out of the events $\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_n$, then the random variable

$$\nu_0 = \begin{cases} 0, & \text{if } \mu = 0 \\ 1, & \text{if } \mu \geq 1 \end{cases}$$

obviously has expected value $\overline{P} = P(\overline{A}_1 + \overline{A}_2 + \cdots + \overline{A}_n)$.

6.2 Monte Carlo simulation of the differences between the true probability value and its Boole–Bonferroni bounds

If we take L_2 as the second order Boole–Bonferroni lower bound of the probability of union of events and U_2 as the second order Boole–Bonferroni upper bound of the probability of union of events

$$L_2 = \frac{2}{k^* + 1}\overline{S}_1 - \frac{2}{k^*(k^* + 1)}\overline{S}_2 \leq \overline{P} \leq \overline{S}_1 - \frac{2}{n}\overline{S}_2 = U_2,$$

then

$$\overline{P} - L_2 = \overline{S}_1 - \overline{S}_2 + ... + (-1)^{n-1}\overline{S}_n - \frac{2}{k^*+1}\overline{S}_1 + \frac{2}{k^*(k^*+1)}\overline{S}_2,$$

and

$$\overline{P} - U_2 = -\overline{S}_2 + ... + (-1)^{n-1}\overline{S}_n + \frac{2}{n}\overline{S}_2.$$

If μ denotes the number of those events which occur out of the events $\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_n$, then

$$\nu_{L_2} = \begin{cases} 0, & \text{if } \mu \leq 1 \\ \sum\limits_{i=1}^{n}(-1)^{i-1}\binom{\mu}{i} - \frac{2}{k^*+1}\binom{\mu}{1} + \frac{2}{k^*(k^*+1)}\binom{\mu}{2} \; = \frac{1}{k^*(k^*+1)}(\mu - k^*)(\mu - k^* - 1), & \text{if } \mu \geq 2 \end{cases}$$

and

$$\nu_{U_2} = \begin{cases} 0, & \text{if } \mu \leq 1 \\ \sum\limits_{i=2}^{n}(-1)^{i-1}\binom{\mu}{i} - \frac{2}{n}\binom{\mu}{2} = \frac{1}{n}\,(\mu+n)\,(1-\mu)\,, & \text{if } \mu \geq 2 \end{cases}$$

have expected values $\overline{P} - L_2$ and $\overline{P} - U_2$. This way the transformed random variables $\nu_{L_2} + L_2$ and $\nu_{U_2} + U_2$ also have expected value $\overline{P}$.

6.3 Monte Carlo simulation of the differences between the true probability value and its Hunter–Worsley bound

Let

$$\overline{P} \leq \overline{S}_1 - \sum_{(i,j)\in T^*} P(\overline{A}_i\overline{A}_j) = U_{HW},$$

then

$$\overline{P} - U_{HW} = -\overline{S}_2 + \overline{S}_3 - ... + (-1)^{n-1}\overline{S}_n + \sum_{(i,j)\in T^*} P(\overline{A}_i\overline{A}_j).$$

If μ denotes the number of those events which occur out of the events $\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_n$, further λ denotes the number of those $\overline{A}_i\overline{A}_j$, $(i,j) \in T^*$ events which occur in a random trial, then the random variable

$$\nu_{HW} = \begin{cases} 0, & \text{if } \mu \leq 1 \\ \sum\limits_{i=2}^{n}(-1)^{i-1}\binom{\mu}{i} + \lambda = 1 - \mu + \lambda, & \text{if } \mu \geq 2 \end{cases}$$

has expected value $\overline{P} - U_{HW}$ and so the transformed random variable $\nu_{HW} + U_{HW}$ also has expected value $\overline{P}$.

6.4 Determination of the final estimation with minimal variance

Let us chose the random variables ν_0, $\nu_{HW} + U_{HW}$ and $\nu_{L_2} + L_2$ and denote $\widehat{\overline{P}}_0, \widehat{\overline{P}}_1, \widehat{\overline{P}}_2$ the three different estimations based on these. All of these are unbiased estimations of the probability value $\overline{P}$. Let us estimate the empirical covariances of these estimations in a simulation procedure:

$$\widehat{C} = \begin{pmatrix} \widehat{c}_{00} & \widehat{c}_{01} & \widehat{c}_{02} \\ \widehat{c}_{01} & \widehat{c}_{11} & \widehat{c}_{12} \\ \widehat{c}_{02} & \widehat{c}_{12} & \widehat{c}_{22} \end{pmatrix}$$

If we introduce the new estimation

$$\widehat{\overline{P}} = w_0\widehat{\overline{P}}_0 + w_1\widehat{\overline{P}}_1 + w_2\widehat{\overline{P}}_2$$

where $w_0 + w_1 + w_2 = 1$, then it will be also an unbiased estimation of the probability value $\overline{P}$. As $\widehat{\overline{P}}$ has variance $w^T\widehat{C}w$, where $w^T = (w_0, w_1, w_2)$, therefore the coefficients w_0, w_1, w_2 resulting the minimal variance estimation can be determined by the solution of the Lagrangian problem:

$$\min_{w_0+w_1+w_2=1} w^T\widehat{C}w.$$

As the gradient of $\mathbf{w}^T\widehat{C}\mathbf{w}$ equals to $2\mathbf{w}^T\widehat{C}$ it is easy to see that the unknown values of w_1, w_2, w_3, λ can be determined by the solution of the following linear equation system:

$$\begin{array}{ccccccccc} \widehat{c}_{00}w_0 & + & \widehat{c}_{01}w_1 & + & \widehat{c}_{02}w_2 & -\lambda & = 0, \\ \widehat{c}_{01}w_0 & + & \widehat{c}_{11}w_1 & + & \widehat{c}_{12}w_2 & -\lambda & = 0, \\ \widehat{c}_{02}w_0 & + & \widehat{c}_{12}w_1 & + & \widehat{c}_{22}w_2 & -\lambda & = 0, \\ w_0 & + & w_1 & + & w_2 & & = 1. \end{array} \tag{35}$$

6.5 Further Monte Carlo simulation algorithms

For the case of multivariate normal probability distribution there are other known Monte Carlo simulation algorithms, see Deák [3], [4] and Ambartzumian et al [1]. Gassmann [9] combined Szántai's general algorithm and Deák's algorithm into a hybrid algorithm. The efficiency of this algorithm was explored by Deák, Gassmann and Szántai in [5].

7 Conclusion and discussion

In this paper we present the column-generation approach of [7] from a dual viewpoint, as a cutting-plane method. Moreover we propose a randomized version of this method. There is an important contrast between direct cutting-plane methods and the present approach. Direct cutting-plane methods for probabilistic functions are difficult to implement due to noise in gradient computation. Even a fairly accurate gradient may result a cut cutting into the epigraph of the objective function (especially in regions farther away from the current iterate). One either needs sophisticated tolerance handling to avoid cutting into the epigraph – see, e.g., [22], [14], [2] –, or else one needs a sophisticated convex optimization method that can handle cuts cutting into the epigraph – see [23]. (Yet another alternative, developed for a different type of problem, is perpetual adjustment of the existing model to information revealed in course of the process; see [11].)

The present models are invulnerable to gradient computation errors. Noisy gradient estimates may yield iterates that do not improve much on our current models. But we retain a true inner approximation of the primal objective – or a true outer approximation of the dual objective –, provided objective values at new iterates are evaluated with appropriate accuracy. This feature facilitates the use of gradient estimates. Our randomized method bears a resemblance to the stochastic approximation family that goes back to [19] (see [15], [12] for recent forms).

The use of gradient estimates may substantially decrease total computational effort, even though a certain (moderate) accuracy *is* demanded in objective values. Computing a single component of a gradient vector will involve an effort comparable to that of computing an objective value, e.g., in case of probability maximization under multivariate normal distribution of the random parameters.

The variance reduction Monte Carlo simulation procedure described in Section 6 was successfully applied in the solution of jointly probabilistic constrained stochastic programming problems, see [22]. The situation was similar to the present one; as the procedure progressed, higher and higher accuracy became necessary.

Acknowledgement

This research is supported by EFOP-3.6.1-16-2016-00006 "The development and enhancement of the research potential at Pallas Athena University" project. The Project is supported by the Hungarian Government and co-financed by the European Social Fund.

References

[1] R. Ambartzumian, A. Der Kiureghian, V. Ohanian, and H. Sukiasian. Multinormal probability by sequential conditioned importance sampling: Theory and applications. *Probabilistic Engineering Mechanics*, 13:299–308, 1998.

[2] T. Arnold, R. Henrion, A. Möller, and S. Vigerske. A mixed-integer stochastic nonlinear optimization problem with joint probabilistic constraints. *Pacific Journal of Optimization*, 10:5–20, 2014.

[3] I. Deák. Three digit accurate multiple normal probabilities. *Numerische Mathematik*, 35:369–380, 1980.

[4] I. Deák. Computing probabilities of rectangles in case of multinormal distributions. *Journal of Statistical Computation and Simulation*, 26:101–114, 1986.

[5] I. Deák, H. Gassmann, and T. Szántai. Computing multivariate normal probabilities: a new look. *Journal of Statistical Computation and Simulation*, 11:920–949, 2002.

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

[7] C.I. Fábián, E. Csizmás, R. Drenyovszki, W. van Ackooij, T. Vajnai, L. Kovács, and T. Szántai. Probability maximization by inner approximation. *Submitted to Acta Polytechnica Hungarica*, 2017.

[8] C.I. Fábián, O. Papp, and K. Eretnek. Implementing the simplex method as a cutting-plane method, with a view to regularization. *Computational Optimization and Applications*, 56:343–368, 2013.

[9] H. Gassmann. Conditional probability and conditional expectation of a random vector. In Y.M. Ermoliev and R.J.-B. Wets, editors, *Numerical Techniques for Stochastic Optimization*, pages 237–254. Springer-Verlag, Berlin, 1988.

[10] A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–150, 1992.

[11] J.L. Higle and S. Sen. *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming*, volume 8 of *Nonconvex Optimization and Its Applications*. Springer, 1996.

[12] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:365–397, 2012.

[13] D.G. Luenberger and Y. Ye. *Linear and Nonlinear Programming*. International Series in Operations Research and Management Science. Springer, 2008.

[14] J. Mayer. *Stochastic Linear Programming Algorithms: A Comparison Based on a Model Management System*. Gordon and Breach Science Publishers, 1998.

[15] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

[16] A. Prékopa. 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, 1990.

[17] A. Prékopa. *Stochastic Programming*. Kluwer Academic Publishers, Dordrecht, 1995.

[18] A. Prékopa, B. Vizvári, and T. Badics. Programming under probabilistic constraint with discrete random variable. In F. Giannesi, T. Rapcsák, and S. Komlósi, editors, *New Trends in Mathematical Programming*, pages 235–255. Kluwer, Dordrecht, 1998.

[19] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[20] A. Ruszczyński. *Nonlinear Optmization*. Princeton University Press, 2006.

[21] T. Szántai. *Numerical Evaluation of Probabilities Concerning Multidimensional Probability Distributions, Thesis*. Hungarian Academy of Sciences, Budapest, 1985.

[22] T. Szántai. A computer code for solution of probabilistic-constrained stochastic programming problems. In Y.M. Ermoliev and R.J.-B. Wets, editors, *Numerical Techniques for Stochastic Optimization*, pages 229–235. Springer-Verlag, Berlin, 1988.

[23] W. van Ackooij and C. Sagastizábal. Constrained bundle methods for upper inexact oracles with application to joint chance constrained energy problems. *Siam Journal on Optimization*, 24:733–765, 2014.