Applying Random Coordinate Descent in a Probability Maximization Scheme

Edit Csizmás · Rajmund Drenyovszki · Tamás Szántai · Csaba I. Fábián

Received: date / Accepted: date

Abstract Gradient computation of multivariate distribution functions calls for considerable effort. A standard procedure is component-wise computation, hence coordinate descent is an attractive choice. This paper deals with constrained convex problems. We apply random coordinate descent in an approximation scheme that is an inexact cutting-plane method from a dual viewpoint. We show that a cutting-plane scheme will converge unless the information added by new cuts persistently deteriorates to zero. Our cuts are dual images of test points, each generated by a single step of random coordinate descent, and we can prove almost sure convergence. We present a computational study comparing gradient descent and coordinate descent applied in the approximation scheme. We found that even an unsophisticated implementation of coordinate descent progresses faster in the initial iterations.

Keywords Stochastic programming, probability maximization, approximation schemes, coordinate-descent methods, cutting-plane methods

1 Introduction

Let us consider a probability maximization problem of the form

$$\max \ \mathrm{P}(T\boldsymbol{x} \geq \boldsymbol{\zeta}) \quad \text{subject to} \quad A\boldsymbol{x} \leq \boldsymbol{b}, \tag{1}$$

where $\boldsymbol{\zeta}$ is a random vector having a nondegenerate normal distribution. Let $F(\boldsymbol{z})$ denote the distribution function. Introducing the probabilistic function

Edit Csizmás corresponding author · Rajmund Drenyovszki · Csaba I. Fábián
Department of Informatics, GAMF: Faculty of Engineering and Computer Science, John von Neumann University. Izsáki út 10, 6000 Kecskemét, Hungary.
E-mail: csizmas.edit@nje.hu, drenyovszki.rajmund@nje.hu, fabian.csaba@nje.hu

Tamás Szántai
Department of Analysis and Operations Research, Institute of Mathematics, Budapest University of Technology and Economics. Műegyetem rakpart 3-9, 1111 Budapest, Hungary.
E-mail: szantai@math.bme.hu

$\varphi(\boldsymbol{z}) = -\log F(\boldsymbol{z})$, the problem is transformed into a convex minimization form, due to the logconcavity of the normal distribution.

To solve this problem, we apply a scheme analogous to the inner approximation approach widely used to handle probabilistic constraints. This approach was initiated by Prékopa [16]. The idea is to approximate a level set of the probabilistic function. Main contributions are Prékopa, Vizvári and Badics [18], Dentcheva, Prékopa and Ruszczyński [4], Dentcheva, Lai and Ruszczyński [2], Dentcheva and Martinez [3], Van Ackooij, Berge, de Oliveira and Sagastizábal [21]. Looking at the proposed methods in chronological order, an increasing level of complexity is noticeable.

The probability maximization scheme proposed in our former paper [7] approximates the epigraph of the probabilistic function, instead of a level set. As new test points are generated by unconstrained optimization, the procedure is simple, easy to implement, and remarkably well tolerates noise in gradient computation. An efficient version was developed in [6], where improving test points are generated by randomized gradient descent, starting from an appropriate convex combination of existing test points. — There is an analogy with averaging in stochastic gradient methods. (The scheme can be extended to handle probabilistic constraints; a Newton-type approach was proposed in the latter paper.)

In the present paper, we modify this probability maximization scheme to generate new test points by random coordinate descent, using the ideas of Nesterov [13]. In order to prove the convergence of the modified scheme, we extend the convergence proof of [5] that we formerly worked out for a deterministic gradient descent scheme.

The test point-generation scheme is an inexact cutting-plane method from a dual viewpoint. The advantages of the dual viewpoint are that it presents a clear visual image, and we can apply classic results on convex functions. Figure 1 depicts the dual objective function and its current cutting-plane model. In the next iteration, we construct an approximate linear support function

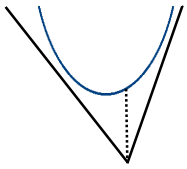

Fig. 1 The graph of the dual objective function and its current cutting-plane model.

at the minimizer of the model function. The new cut intersects the line of the dotted interval. The information content of the new cut is measured by the ratio between the lengths of the dotted sections below and above the

intersection. We are going to show that, unless the information content of the cuts persistently deteriorates to zero, the scheme will converge. In case the new test points (i.e., cuts) are generated by just a single step of random coordinate descent, there will almost surely exist a sequence of cuts each having an information content above a threshold.

In section 2 of this paper, we discuss the solution scheme in an idealized setting. In section 3, we adapt this scheme to probability maximization. Specifically, we show that Lipschitz constants for the axis-directional derivatives of the normal distribution function are easily computed. A computational study is presented in section 4.

2 Problem, Model and Solution Scheme in an Idealized Setting

We consider a problem of the form

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

where $\phi : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function. The 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. We assume that the feasible domain of (2) is nonempty and bounded, hence the problem has an optimal solution. — We are going to attribute further useful properties to $\phi(.)$.

Introducing the multiplier vectors $-\boldsymbol{u} \in \mathbb{R}^n$ and $-\boldsymbol{y} \in \mathbb{R}^r_+$ to the respective constraints, the Lagrangian dual of (2) can be written as

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

with an appropriate closed convex polyhedron $\mathcal{D} \subset \mathbb{R}^{r+n}$. Let us introduce the function

$$\nu(\boldsymbol{u}) = \inf_{(\boldsymbol{y},\boldsymbol{u})\in\mathcal{D}} -\boldsymbol{y}^T\boldsymbol{b} \qquad (\boldsymbol{u} \in \mathbb{R}^n). \tag{4}$$

We assume that $\nu(.) < \infty$. Moreover, we assume that the convex conjugate function $\phi^\star(.)$ is finite valued. Problem (3) can then be written in unconstrained form:

$$-\min\ \left\{\phi^\star(\boldsymbol{u}) + \nu(\boldsymbol{u})\right\}. \tag{5}$$

(Note that $\nu(.) > -\infty$ due to convex duality.)

2.1 Polyhedral Models

Suppose we have evaluated the function $\phi(\boldsymbol{z})$ at points $\boldsymbol{z}_i$ $(i = 0, 1, \ldots, k)$. An inner approximation of $\phi(.)$ is

$$\phi_k(\boldsymbol{z}) = \min \sum_{i=0}^{k} \lambda_i \phi(\boldsymbol{z}_i) \quad \text{such that} \quad \lambda_i \geq 0,\ \sum \lambda_i = 1, \quad \sum_{i=0}^{k} \lambda_i \boldsymbol{z}_i = \boldsymbol{z}. \tag{6}$$

A polyhedral model of problem (2) is

$$\min\ \phi_k(\boldsymbol{z}) \quad \text{subject to} \quad \boldsymbol{z} = T\boldsymbol{x},\ \ A\boldsymbol{x} \le \boldsymbol{b}. \tag{7}$$

We assume that the convex hull of the test points $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_k$ is wide enough to produce a finite minimum.

The convex conjugate of $\phi_k(\boldsymbol{z})$ is

$$\phi_k^\star(\boldsymbol{u}) = \max_{0 \le i \le k} \left\{ \boldsymbol{u}^T \boldsymbol{z}_i - \phi(\boldsymbol{z}_i) \right\}, \tag{8}$$

and the following problem is a polyhedral model of problem (3):

$$-\min \left\{ \phi_k^\star(\boldsymbol{u}) - \boldsymbol{y}^T \boldsymbol{b} \right\} \quad \text{subject to} \quad (\boldsymbol{y}, \boldsymbol{u}) \in \mathcal{D}. \tag{9}$$

Let $(\,\overline{\lambda}_0, \ldots, \overline{\lambda}_k, \overline{\boldsymbol{x}}\,)$ and $(\,\overline{\boldsymbol{u}}, \overline{\boldsymbol{y}}\,)$ denote respective optimal solutions of the problems (7) and (9) – both existing due to our assumption concerning the feasibility of (7). Let moreover

$$\overline{\boldsymbol{z}} = \sum_{i=0}^{k} \overline{\lambda}_i \boldsymbol{z}_i. \tag{10}$$

Complementarity between the optimal solutions of the primal-dual pair of linear programming problems (7) and (9) results in

Observation 2.1 *We have* $\phi_k(\overline{\boldsymbol{z}}) + \phi_k^\star\,(\overline{\boldsymbol{u}}) = \overline{\boldsymbol{u}}^T \overline{\boldsymbol{z}}$ *(hence* $\overline{\boldsymbol{u}}$ *is a subgradient of* $\phi_k(\boldsymbol{z})$ *at* $\overline{\boldsymbol{z}}$*).*

Details of linear programming formulations and corresponding observations with sketches of proofs can be found in [6].

2.2 Solution Scheme

Assume that the initial model (6) is based on $K+1$ test points, $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_K$. From a primal viewpoint, the solution scheme is a column-generation method that iteratively constructs new test points $\boldsymbol{z}_{K+1}, \boldsymbol{z}_{K+2}, \ldots$ in the primal model function (6).

Here we are going to use the dual viewpoint, considering the scheme as a cutting-plane method that approximates $\phi^\star(.)$ with an iteratively improving sequence of functions $\phi_k^\star(.)$ $(k = K, K+1, \ldots)$.

The unconstrained dual problem (5) can be solved by minimizing $d(\boldsymbol{u}) := \phi^\star(\boldsymbol{u}) + \nu(\boldsymbol{u})$. The dual model problem (9) can be solved by minimizing $d_k(\boldsymbol{u}) := \phi_k^\star(\boldsymbol{u}) + \nu(\boldsymbol{u})$.

An approximate support function to $d(\boldsymbol{u})$ at $\overline{\boldsymbol{u}}$ is constructed in the form

$$\ell(\boldsymbol{u}) := \ell'(\boldsymbol{u}) + \ell''(\boldsymbol{u}) \qquad (\boldsymbol{u} \in \mathbb{R}^n), \tag{11}$$

where the right-hand-side functions are separate support functions to $\phi^\star(\boldsymbol{u})$ and $\nu(\boldsymbol{u})$, respectively.

As for $\nu(\boldsymbol{u})$, we can construct an exact support function $\ell^{\backprime}(\boldsymbol{u})$ by solving a linear programming problem.

An approximate support function to $\phi^{\star}(\boldsymbol{u})$ at $\overline{\boldsymbol{u}}$ is constructed in the form

$$\ell^{\backprime}(\boldsymbol{u}) := \boldsymbol{u}^T \boldsymbol{z}^{\backprime} - \phi(\boldsymbol{z}^{\backprime}) \qquad (\boldsymbol{u} \in \mathbb{R}^n), \tag{12}$$

with an appropriate vector $\boldsymbol{z}^{\backprime}$. (Note that $\ell^{\backprime}(.) \leq \phi^{\star}(.)$ by the above definition.) An exact support function could be obtained by setting $\boldsymbol{z}^{\backprime}$ to be the exact maximizer of $\overline{\boldsymbol{u}}^T \boldsymbol{z} - \phi(\boldsymbol{z})$. — Instead, we are going to perform approximate minimization of the negative function

$$f(\boldsymbol{z}) := \phi(\boldsymbol{z}) - \overline{\boldsymbol{u}}^T \boldsymbol{z} \qquad (\boldsymbol{z} \in \mathbb{R}^n), \tag{13}$$

taking a single step of random coordinate descent. We apply a special form of the framework of Nesterov [13]: the domain space is decomposed to one-dimensional subspaces and Euclidean norms are used throughout. See also section 8.6 in Luenberger and Ye [12]; and Wright [22], Bottou, Curtis and Nocedal [1] for discussions on coordinate descent.

Assumption 2.2 *The directional derivatives of $\phi(\boldsymbol{z})$, taken along the coordinate axes, are Lipschitz continuous.*

I.e., for any $i \in \{1, \ldots, n\}$ there exists a constants $L_i > 0$ such that

$$\left| \boldsymbol{e}_i^T \nabla \phi(\boldsymbol{z} + t\boldsymbol{e}_i) - \boldsymbol{e}_i^T \nabla \phi(\boldsymbol{z}) \right| \leq L_i |t| \tag{14}$$

holds for any $\boldsymbol{z} \in \mathbb{R}^n$ and $t \in \mathbb{R}$.

Assumption 2.3 *The function $\phi(\boldsymbol{z})$ is strongly convex with the parameter $\alpha > 0$.*

Assumptions 2.2 and 2.3 are obviously inherited to the current objective function $f(\boldsymbol{z})$. In order to construct an approximate support function in the form (12), we are going to perform a single coordinate-descent step starting from $\overline{\boldsymbol{z}}$. Let $\overline{\boldsymbol{g}} := \nabla f(\overline{\boldsymbol{z}})$ denote the (unknown) gradient.

We randomly select a coordinate axis and will estimate the corresponding directional derivative of the objective function. Formally, $\overline{\boldsymbol{e}}$ will denote the randomly selected axis. This is a random vector taking values from among the unit vectors $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$, according to uniform distribution. The corresponding directional derivative is $\overline{q} := \overline{\boldsymbol{e}}^T \overline{\boldsymbol{g}}$. As $\overline{q}$ is a randomly selected component of the n-dimensional vector $\overline{\boldsymbol{g}}$, it follows that the event

$$\overline{q}^2 \geq \frac{1}{n} \|\overline{\boldsymbol{g}}\|^2 \tag{15}$$

occurs with a probability at least $\frac{1}{n}$.

Let $\overline{L}$ denote the Lipschitz constant along the selected axis, and let

$$\boldsymbol{z}^{\backprime} := \overline{\boldsymbol{z}} - \frac{1}{\overline{L}} \overline{q}\, \overline{\boldsymbol{e}}. \tag{16}$$

Theorem 2.4 *Assume that the directional derivative along the selected axis has at least average magnitude, i.e., (15) holds. Then we have*

$$f\left(\boldsymbol{z}^{\backprime}\right) - \mathcal{F} \;\leq\; \left(1 - \frac{\alpha}{n\, L_{\max}}\right)\left(f\left(\overline{\boldsymbol{z}}\right) \;- \mathcal{F}\right), \tag{17}$$

where $\mathcal{F} = \min_{\boldsymbol{z}} f(\boldsymbol{z})$ *and* $L_{\max} = \max_i L_i$.

This is a special and slightly modified form of theorem 2 of [13]. We omit the proof, as Nesterov's line of proof is directly applicable. Simple proofs can also be found in [12], [22], [1]. The referenced works also contain related results about deterministic methods. — In contrast to those of the referenced results that involve random methods, here we deal with function values instead of expectations, as this fits the present cutting-plane scheme. (Hence we need the assumption on the selected axis.)

Corollary 2.5 *Assume that the directional derivative along the selected axis has at least average magnitude, i.e., (15) holds. Then the new linear support function adds significant information to the model, specifically,*

$$\ell(\overline{\boldsymbol{u}}) \;\geq\; \theta d(\overline{\boldsymbol{u}}) \;+\; (1-\theta) d_k(\overline{\boldsymbol{u}}) \tag{18}$$

holds with the constant $\theta = \frac{\alpha}{n\, L_{\max}}$.

Proof. It is enough to show that

$$\ell^{\backprime}(\overline{\boldsymbol{u}}) \;\;\geq\;\; \theta \phi^{\star}(\overline{\boldsymbol{u}}) \;+\; (1-\theta)\phi_k^{\star}(\overline{\boldsymbol{u}}) \tag{19}$$

holds with $\ell^{\backprime}(\boldsymbol{u})$ constructed in the form (12), with $\boldsymbol{z}^{\backprime}$ defined in (16).

(19) is in turn obtained by substituting $f(\boldsymbol{z}) = \phi(\boldsymbol{z}) - \overline{\boldsymbol{u}}^T \boldsymbol{z}$ and $\mathcal{F} = -\phi^{\star}(\overline{\boldsymbol{u}})$ in (17), and applying the equality of Observation 2.1. □

2.3 Convergence

Keeping the notation $\phi_k(\boldsymbol{z})$ and $d_k(\boldsymbol{u})$ for the current polyhedral model functions, let $\overline{\boldsymbol{z}}_{k+1}$ and $\overline{\boldsymbol{u}}_{k+1}$ (instead of just $\overline{\boldsymbol{z}}$ and $\overline{\boldsymbol{u}}$) denote the optimal solutions of the current model problems. – The next test point in the primal scheme, that is the next cut in the dual scheme, is constructed by performing a single coordinate-descent step, along a randomly selected axis $\overline{\boldsymbol{e}}_{k+1}$.

Assumption 2.6 *We assume that all the dual iterates* $\overline{\boldsymbol{u}}_k$ *are contained in a convex compact set* $\mathcal{O}_{\boldsymbol{u}} \subset \mathbb{R}^n$.

(In order to see that this assumption is not overly demanding, imagine that the optimal solutions $\overline{\boldsymbol{z}}_k$ of the primal model problems are all contained in a compact set, and take Observation 2.1 into account.)

From a dual point of view, we perform a cutting-plane procedure to solve the convex problem

$$\min\; d(\boldsymbol{u}) \qquad \text{such that} \qquad \boldsymbol{u} \in \mathcal{O}_{\boldsymbol{u}}, \tag{20}$$

where the domain is compact. (We put the problem in technically constrained form in order to simplify the convergence proof.) The model function is $d_k(\boldsymbol{u})$. In the next iterate $\overline{\boldsymbol{u}}_{k+1}$, an approximate support function $\ell_{l+1}(\boldsymbol{u})$ is generated. This scheme converges unless the information contributed by the successive cuts persistently deteriorates to nil:

Theorem 2.7 *When solving (20) with an approximate cutting-plane method, assume that there exists an infinite sequence (k_i) of strictly increasing natural numbers such that*

$$\ell_{l+1}(\overline{\boldsymbol{u}}_{l+1}) \;\geq\; \theta d(\overline{\boldsymbol{u}}_{l+1}) \;+\; (1-\theta) d_l(\overline{\boldsymbol{u}}_{l+1}) \qquad (l = k_i,\; i = 1, 2, \ldots) \tag{21}$$

holds with some constant $\theta \in (0, 1]$. Then the method generates a sequence of models and iterates satisfying

$$\lim_{k\to\infty} D_k = D, \tag{22}$$

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

The proof can be found in appendix A. – We mention that $\lim_{i\to\infty} d(\overline{\boldsymbol{u}}_{k_i+1}) = D$ also holds, but in the present dual-type application, convergence of the minima is sufficient.

Corollary 2.8 *Let us select the axes independently in the course of the column generation process. Then*

$$\lim_{k\to\infty} \overline{\boldsymbol{z}}_k = \boldsymbol{z}^\star \quad \textit{almost surely,}$$

where $\boldsymbol{z}^\star$ denotes the $\boldsymbol{z}$-part of the optimal solution of the convex problem (2).

Proof. From the independent selection of the axes, it follows that, with probability 1, there exists a sequence $k_1, k_2, \ldots$ of natural numbers such that $K \leq k_1 < k_2 < \ldots$, and

$$\overline{q}_{l+1}^2 \geq \frac{1}{n}\|\overline{\boldsymbol{g}}_{l+1}\|^2 \qquad \text{holds for} \quad l = k_i, \;\; i = 1, 2, \ldots. \tag{23}$$

If such a sequence (k_i) indeed exists then, by corollary 2.5, (21) holds with $\theta = \frac{\alpha}{n\, L_{\max}}$. This makes theorem 2.7 applicable.

In our solution scheme, $\phi(\boldsymbol{z}^\star) = -D$ is the common optimum of the convex problem (2) and the convex dual problem (5). Moreover $\phi_k(\overline{\boldsymbol{z}}_{k+1}) = -D_k$ is the common optimum of the polyhedral model problems. Hence $D_k \to D$ translates to

$$\phi_k(\overline{\boldsymbol{z}}_{k+1}) \to \phi(\boldsymbol{z}^\star). \tag{24}$$

Taking into account that $\phi_k(\overline{\boldsymbol{z}}_{k+1}) \geq \phi(\overline{\boldsymbol{z}}_{k+1}) \geq \phi(\boldsymbol{z}^\star)$ holds for every k, we get $\phi(\overline{\boldsymbol{z}}_{k+1}) \to \phi(\boldsymbol{z}^\star)$. Strict convexity of $\phi(\boldsymbol{z})$ then yields $\overline{\boldsymbol{z}}_k \to \boldsymbol{z}^\star$. □

3 Adapting the Approach to Probability Maximization

Let $F(z_1, \ldots, z_n)$ denote the joint distribution function of the random variables $\zeta_1, \ldots, \zeta_n$. Assumed that the distribution function is differentiable, its axis-directional derivatives can be computed as

$$\frac{\partial F(z_1, \ldots, z_n)}{\partial z_i} \;=\; F(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \mid z_i)\; h_i(z_i) \qquad (i = 1, \ldots, n), \tag{25}$$

where $F(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \mid z_i)$ is the conditional distribution function of the random variables $\zeta_1, \ldots, \zeta_{i-1}, \zeta_{i+1}, \ldots, \zeta_n$ given that $\zeta_i = z_i$, and $h_i(z)$ is the probability density function of the random variable ζ_i – see, e.g., section 6.6.4 in Prékopa [17].

In solving probabilistic problems, a standard procedure is component-wise computation of the gradient, based on (25). In the case of high dimensions, gradient computation (or estimation) requires a major effort, therefore coordinate descent is an attractive choice. – We mention that another means of alleviating the difficulty of gradient computation in the case of normal distribution was proposed by Hantoute, Henrion and Pérez-Aros [9].

We assume that the feasible domain of the probability maximization problem (1) is contained in a bounded (closed) box $\mathcal{X} \subset \mathbb{R}^m$. Exploiting the monotonicity of the objective function, problem (1) can be written as

$$\min\; \varphi(\boldsymbol{z}) \quad \text{subject to} \quad \boldsymbol{z} \leq T\boldsymbol{x},\;\; A\boldsymbol{x} \leq \boldsymbol{b},\;\; \boldsymbol{x} \in \mathcal{X}. \tag{26}$$

We assume that a feasible starting point $\check{\boldsymbol{z}}$ is known such that $F(\check{\boldsymbol{z}}) \gg 0$.

A further speciality of the normal distribution function is the existence of a bounded (closed) box $\mathcal{Z}$ outside which the probability weight can be ignored. We are going to include $\boldsymbol{z} \in \mathcal{Z}$ among the constraints of the probability maximization problem. This allows the regularization of the objective function, in the form of

$$\phi(\boldsymbol{z}) = -\log F(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z}\|^2 \tag{27}$$

with some $\rho > 0$. In what follows, we assume that ρ is small enough, and the regularizing term makes no significant variation in the objective value of $\boldsymbol{z} \in \mathcal{Z}$. We are going to solve the approximating problem

$$\min\; \phi(\boldsymbol{z}) \quad \text{subject to} \quad \boldsymbol{z} \in \mathcal{Z},\;\; \boldsymbol{z} \leq T\boldsymbol{x},\;\; A\boldsymbol{x} \leq \boldsymbol{b},\; \boldsymbol{x} \in \mathcal{X}. \tag{28}$$

(Details on the quality of this approximation can be found in [6], section 5.1.)

By splitting the variables $\boldsymbol{z}$, we can fit problem (28) to the template of (2). Relaxing every constraint except $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{x} \in \mathcal{X}$, we then obtain a Lagrangian dual problem that can be transformed to the unconstrained form of (5). — A detailed description can be found in [5]. If the boxes $\mathcal{Z}$ and $\mathcal{X}$ are origin-centered, then the polyhedral function $\nu(.)$ can be formulated using L_1 norms, showing that $\nu(.) < \infty$.

Polyhedral models can be constructed in the manner described in section 2.1, and observation 2.1 is inherited to the present models.

3.1 Justification of the Assumptions of Section 2

Convexity of the function $\varphi(\boldsymbol{z}) = -\log F(\boldsymbol{z})$ follows from the theory of log-concave measures developed by Prékopa [14], [15]. The regularizing term in (27) ensures that the objective function is strongly convex with a parameter at least ρ. This justifies assumption 2.3. (Moreover, it ensures that the convex conjugate of the regularized objective function is finite valued.)

Let us include a feasible solution $\check{\boldsymbol{z}}$ having $\check{p} = F(\check{\boldsymbol{z}}) \gg 0$ among the initial test points of the polyhedral model (6). Then the level set $\mathcal{L}(\check{p}) = \{\, \boldsymbol{z} \in \mathbb{R}^n \,|\, F(\boldsymbol{z}) \geq \check{p} \,\}$ will contain every optimal solution $\overline{\boldsymbol{z}}$ of any model problem. Assumption 2.6 is justified by the following

Observation 3.1 *The polyhedral models can be initialized in such a manner that*

- *there exists a compact set $\mathcal{O}_{\boldsymbol{z}}$ that contains every optimal solution $\overline{\boldsymbol{z}}$ of any model problem ;*
- *there exists a finite upper bound $\Gamma_{\boldsymbol{u}}$ such that $\|\overline{\boldsymbol{u}}\| \leq \Gamma_{\boldsymbol{u}}$ holds for every subgradient $\overline{\boldsymbol{u}}$ of any model function $\phi_k(\boldsymbol{z})$ on $\mathcal{O}_{\boldsymbol{z}}$.*

The objects $\mathcal{O}_{\boldsymbol{z}}$ and $\Gamma_{\boldsymbol{u}}$ are determined by the initialization of the polyhedral models.

Details of such initialization of the models can be found in [5] (section 4.)

As for assumption 2.2, we are going to show that new iterates can be kept in a region where the directional derivatives of the regularized objective (27) are Lipschitz continuous. But first, let us examine the directional derivatives of the nondegenerate normal distribution function we work with.

Theorem 3.2 *Let $C \in \mathbb{R}^{n\times n}$ denote the covariance matrix of our nondegenerate normal distribution. The corresponding distribution function $F(\boldsymbol{z})$ has the following properties.*

(i) Bounded axis-directional derivatives: for $i = 1, \ldots, n$, we have

$$\boldsymbol{e}_i^T \nabla F(\boldsymbol{z}) \;\leq M_i \quad (\boldsymbol{z} \in \mathbb{R}^n) \qquad \textit{with} \qquad M_i = \frac{1}{\sigma_i \sqrt{2\pi}}, \tag{29}$$

where $\sigma_i > 0$ denotes the standard deviation of ζ_i, the ith component of the random vector.

(ii) Lipschitz-continuous axis-directional derivatives: for $i = 1, \ldots, n$, we have

$$\begin{gathered} \left|\, \boldsymbol{e}_i^T \nabla F(\boldsymbol{z} + t\boldsymbol{e}_i) - \boldsymbol{e}_i^T \nabla F(\boldsymbol{z}) \,\right| \;\leq L_i |t| \quad (\boldsymbol{z} \in \mathbb{R}^n,\; t \in \mathbb{R}) \\ \textit{with} \qquad L_i = \tfrac{1}{\sqrt{2e\pi}} \, \tfrac{|C_{-i}|}{|C|}, \end{gathered} \tag{30}$$

where $|C|$ denotes the determinant of the covariance matrix, and C_{-i} denotes the covariance matrix of the random vector $(\zeta_1, \ldots, \zeta_{i-1}, \zeta_{i+1}, \ldots, \zeta_n)$.

The proof of the above theorem and of the following corollary can be found in appendix B.

Corollary 3.3 *Given $p > 0$, the function $\varphi(\boldsymbol{z}) = -\log F(\boldsymbol{z})$ has Lipschitz-continuous axis-directional derivatives on the level set*

$$\mathcal{L}(p) = \{\, \boldsymbol{z} \in \mathbb{R}^n \,|\, F(\boldsymbol{z}) \geq p \,\}.$$

Namely, $L_i(p) := \frac{L_i}{p} + \frac{M_i^2}{p^2}$ $(i = 1, \ldots, n)$ are appropriate Lipschitz-constants.

Remark 3.4 *The directional derivatives of the regularized objective (27) inherit Lipschitz continuity on the level sets $\mathcal{L}(p)$, with ρ added to the constants.*

As we have included a feasible solution $\check{\boldsymbol{z}}$ having $\check{p} = F(\check{\boldsymbol{z}}) \gg 0$ among the initial test points, the level set $\mathcal{L}(\check{p})$ will contain an optimal solution $\overline{\boldsymbol{z}}$ of the current primal model problem. We find a new test point $\boldsymbol{z}^{\backprime}$ by performing a coordinate-descent step starting from $\overline{\boldsymbol{z}}$. If we should know that, e.g., $\boldsymbol{z}^{\backprime} \in \mathcal{L}(0.1)$ would hold, then (by corollary 3.3 and remark 3.4) we could set the Lipschitz constant $L_i(0.1) + \rho$ when moving along the selected axis i.

A theoretically convergent version can be worked out by taking a tentative step:

$$\boldsymbol{z}^{\backprime} \;=\; \overline{\boldsymbol{z}} - \frac{1}{L_i(\check{p}) + \rho} \left[\boldsymbol{e}_i^T \nabla\phi(\overline{\boldsymbol{z}})\right] \boldsymbol{e}_i.$$

We evaluate $p^{\backprime} = F(\boldsymbol{z}^{\backprime})$, and, in case $p^{\backprime} < 0.1$, we set the actual test point to be

$$\overline{\boldsymbol{z}} - \frac{1}{L_i(p^{\backprime}) + \rho} \left[\boldsymbol{e}_i^T \nabla\phi(\overline{\boldsymbol{z}})\right] \boldsymbol{e}_i.$$

Theoretical convergence proof of this scheme can be based on observation 3.1.

4 Computational Study

Our approximation scheme was implemented in MATLAB using the IBM ILOG CPLEX optimization toolbox. Multivariate normal distribution function values and axis-directional derivatives were computed by the QSIMVNV Matlab function implemented by Genz [8].

Our solver is based on the implementation used in our former paper [6]. The difference is in the manner of finding new test points in the approximation scheme, i.e., in the method of finding an approximate minimizer to the objective function (13) of the actual column-generation subproblem. In the former paper, we applied gradient descent (GD). Gradients were estimated component-wise, and a single approximate line search was performed in each iteration. In the present project, we apply coordinate descent (CD), performing a single step along a randomly selected axis.

In axis selection, we implemented the method proposed by Wright [22], where the axis selection is done by sampling without replacement. The procedure is divided into 'epochs' of n consecutive iterations. At the start of

each epoch, a random permutation of $\{1, 2, \ldots n\}$ is generated, and the axes are selected according to this permutation. Wright mentions that this kind of randomization proved in several contexts to be superior to the sampling-with-replacement scheme. Computational experiments are reported in Liu et al. [11].

We compared the GD and CD solution approaches on a test problem that originates from Dentcheva et al. [2] and Henrion [10]. This is a cash matching problem where random liabilities are modelled with a fifteen-variate normal distribution. The problem was originally formulated as cost minimization under a probabilistic constraint. We transformed the problem to probability maximization under a cost constraint and formulated three instances with different right-hand sides of this constraint. The respective optima of these problem instances are $0.8, 0.9$ and 0.99.

For each problem instance, we found starting feasible solutions $\check{\boldsymbol{z}}$ having $F(\check{\boldsymbol{z}}) > 0.3$. In our test runs, the regularization parameter ρ of (27) was set to 0. We found that our way of computing the coordinate descent step length is rather conservative. Hence, in the present project, we experimented with step lengths that were multiples of the step lengths calculated from our present Lipschitz constants, and decided to work with multiples of 12. — We think it possible to calculate Lipschitz constants tighter than those in theorem 3.2, though it will require more intricate formulas. Moreover, our objective function has a special characteristic: the directional derivatives tend to diminish in the course of the solution process, as we move into level sets $\mathcal{L}(p)$ belonging to higher and higher values of probability p. (We conjecture that the same holds for the second directional derivatives.) We may need to calculate different Lipschitz constants for different regions.

In our test runs, most of the computational effort was devoted to estimating distribution function values and gradient components or axis-directional derivatives. We found that a single GD iteration demands as much computational effort as 10 CD iterations. Though we solved problems with $n = 15$ dimensional random vectors, computing function values was necessary in both approaches. (To be precise, a GD iteration in our implementation includes a single approximate line search that generally requires several function value computations.) — We expect the ratio of GD : CD iteration effort will approach n for high dimensions.

For each problem instance, we performed 10 runs with both solution approaches. In figures 2, 3 and 4, blue lines show the progress of the GD approaches, and orange lines show the progress of the CD approaches. Each interval on the horizontal axes of the diagrams represents 1 iteration in the case of the GD runs, and 10 iterations in the case of the CD runs. (We depict 8 iterations for every GD run, and 80 iterations for every CD run.) — Differences in starting objective values belonging to the different instances are due to differences in the right-hand sides of the cost constraints. (Our initialization procedure is constraint-dependent.)

The figures show that, in the initial iterations, CD (orange lines) progresses faster. For the GD approach (blue lines), it takes 2-8 iterations to overcome CD.

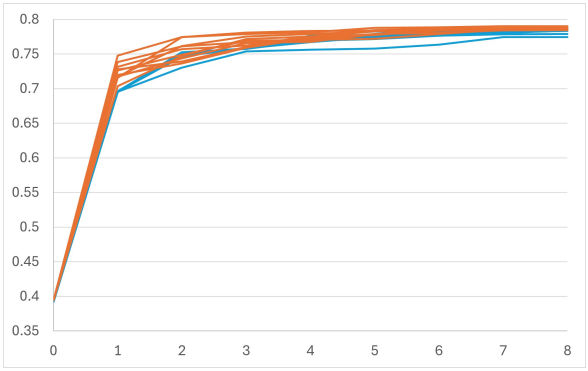

Fig. 2 Progress of the GD and CD approaches on the test instance having an optimum of 0.8

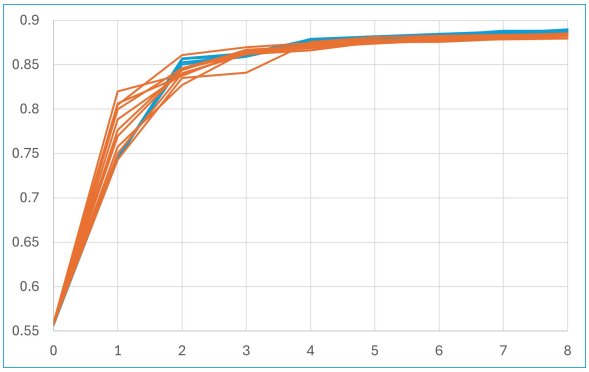

Fig. 3 Progress of the GD and CD approaches on the test instance having an optimum of 0.9

5 Conclusions

We compared a gradient descent and a coordinate descent approach in an approximation scheme for the solution of probability maximization problems. We conclude that even an unsophisticated implementation of the coordinate descent can be useful in finding good starting solutions.

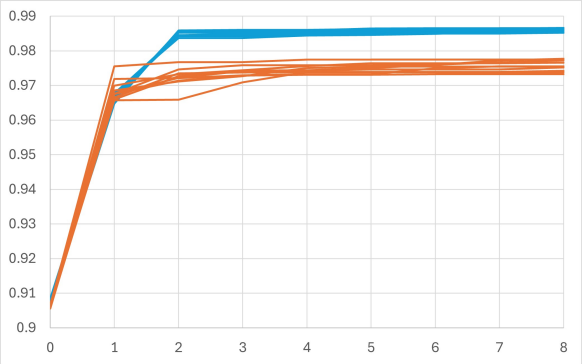

Fig. 4 Progress of the GD and CD approaches on the test instance having an optimum of 0.99

Potential improvements on our coordinate descent implementation include calculating tighter Lipschitz constants for the directional derivatives of the distribution function (possibly different constants for different regions of the domain.)

We mention potential applications for the model and the solution method presented in this paper:

In electricity networks, load levels must remain within safe limits. It is especially important if renewable sources and battery storage systems are used, and smart grid technology enables demand-side management.

The composition of cattle feed was one of the earliest examples of probabilistic models. (Nutrient contents of the different components may vary, and the aim is to provide a mixture containing the required nutrients.)

In urban traffic networks, signal timing can be used to avoid congestion. We can find an appropriate signal timing strategy that keeps the probability of congestion low.

A Convergence of the Inexact Cutting-Plane Method

The following proof is a generalized version of the proof of theorem 9 in [5]. In the present primal scheme, improving test points are found by random coordinate descent, hence we must allow for shallow cuts in the dual.

The proof applies some of the ideas from the convergence proof of the exact cutting-plane method in Ruszczyński [20], theorem 7.7. In order to handle inexactness, we need a well-known theorem from convex analysis.

Proof of theorem 2.7 As the sequence of the model functions is monotone increasing,

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

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

[19]. Let $D_\infty = \min d_\infty(\boldsymbol{u})$. By definition, we have

$$d_k(\overline{\boldsymbol{u}}_{k+1}) \;=\; D_k \;\le\; D_\infty \;\le\; d_\infty(\overline{\boldsymbol{u}}_{k+1}) \qquad (i = 1, 2, \ldots). \tag{31}$$

From the uniform convergence directly follows

$$\lim_{k\to\infty} D_k = D_\infty \qquad \text{and} \qquad \lim_{k\to\infty} d_\infty(\overline{\boldsymbol{u}}_{k+1}) = D_\infty. \tag{32}$$

In order to show that $D_\infty = D$, let us consider a subsequence of the sequence (k_i) whose existence has been assumed in the theorem. Specifically, let $l = k_{i_j},\ j = 1, 2, \ldots$ be such that $(\overline{\boldsymbol{u}}_{l+1})$ is convergent; and let $\widetilde{\boldsymbol{u}}$ denote the limit vector. (A convergent subsequence certainly exists due to the compactness of the domain $\mathcal{O}_{\boldsymbol{u}}$.) From the continuity of $d(\boldsymbol{u})$, it follows that

$$d(\overline{\boldsymbol{u}}_{l+1}) \to d(\widetilde{\boldsymbol{u}}) \qquad \text{as} \qquad l = k_{i_j},\ j \to \infty. \tag{33}$$

The function $d_\infty(.)$ tops any support function constructed in the process. Combining this with the assumption (21) in the theorem, we get

$$\begin{array}{ccccc} d_\infty(\overline{\boldsymbol{u}}_{l+1}) & \ge & \ell_{l+1}(\overline{\boldsymbol{u}}_{l+1}) & \ge & \theta\ d(\overline{\boldsymbol{u}}_{l+1}) \ + (1-\theta)\ d_l(\overline{\boldsymbol{u}}_{l+1}) \qquad (l = k_{i_j},\ j = 1, 2, \ldots). \\ \downarrow & & \downarrow & & \downarrow \\ D_\infty & & d(\widetilde{\boldsymbol{u}}) & & D_\infty \end{array} \tag{34}$$

The arrows represent convergence as $j \to \infty$. These are direct consequences of (32) and (33) — note that $d_l(\overline{\boldsymbol{u}}_{l+1}) = D_l$. Taking into account $\theta > 0$, we obtain $D_\infty \ge d(\widetilde{\boldsymbol{u}})$.

But $d(\widetilde{\boldsymbol{u}}) \ge D \ge D_\infty$ by definition. It follows that $D_\infty = D$, completing the proof. □

B Axis-Directional Derivatives of the Normal Distribution Function

Let $h(\boldsymbol{z})$ denote the probability density function of our n-variate normal distribution.

In the proof of theorem 3.2, we are going to consider directional derivatives along the nth axis only, i.e., assume $i = n$, for the sake of the simplicity of the formulas. The directional derivatives will be taken at $\overline{\boldsymbol{z}} = (\overline{z}_1, \ldots, \overline{z}_n) \in \mathbb{R}^n$. We'll first need

Lemma B.1 *The nth axis-directional derivative of $F(\boldsymbol{z})$ at $\overline{\boldsymbol{z}}$ is*

$$\boldsymbol{e}_n^T \nabla F(\overline{\boldsymbol{z}}) \;=\; \int\limits_{\overline{Z}'} h(z_1, \ldots, z_{n-1}, \overline{z}_n)\ d(z_1, \ldots, z_{n-1}), \tag{35}$$

where

$$\overline{Z}' = \left\{ (z_1, \ldots, z_{n-1}) \in \mathbb{R}^{n-1} \,\middle|\, z_i \le \overline{z}_i \ \ (i = 1, \ldots, n-1) \right\}.$$

Proof. By definition,

$$\boldsymbol{e}_n^T \nabla F(\overline{\boldsymbol{z}}) \;=\; \lim_{t\to 0} \frac{F(\overline{\boldsymbol{z}} + t\boldsymbol{e}_n) - F(\overline{\boldsymbol{z}})}{t}, \tag{36}$$

where the numerator is computed by integrating the density function $h(\boldsymbol{z})$ over the set $\overline{Z}' \times [\overline{z}_n, \overline{z}_n + t] \subset \mathbb{R}^n$. (We consider $t > 0$ for the sake of simplicity. The discussion is easily extended to negative values.)

Let $\mathcal{Z}$ denote a bounded (measurable) subset of $\overline{Z}'$. We are going to show that

$$\lim_{t\to 0} \frac{1}{t} \int\limits_{\mathcal{Z}\times[\overline{z}_n, \overline{z}_n + t]} h(\boldsymbol{z})\, d\boldsymbol{z} \;=\; \int\limits_{\mathcal{Z}} h(z_1, \ldots, z_{n-1}, \overline{z}_n)\ d(z_1, \ldots, z_{n-1}). \tag{37}$$

Concerning function values on the domain of the left-hand integral, we have

$$\Big|h(z_1,\dots,z_{n-1},\,\overline{z}_n) - h(z_1,\dots,z_{n-1},\,\overline{z}_n+s)\Big| \;\le\; t\max_{\boldsymbol{z}\in\mathbb{R}^n}\Big|\boldsymbol{e}_n\nabla h(\boldsymbol{z})\Big|$$

for any $(z_1,\dots,z_{n-1})\in\mathcal{Z}$ and $s\in[0,t]$, due to the mean-value theorem. (The maximum on the right-hand side is clearly finite.)

It follows that, when computing approximating sums of the left-hand integral in (37) for a given t, it is sufficient to consider only 'cylindric' partitions of the domain $\mathcal{Z}\times[\overline{z}_n,\overline{z}_n+t]$, i.e., partitions whose subsets are obtained in the form $Z\times[\overline{z}_n,\overline{z}_n+t]$ with some $Z\subset\mathcal{Z}$.

The equality (37) can then be extended from the bounded subset $\mathcal{Z}$ to $\overline{Z}'$. Combining this with (36), we get (35). □

Remark B.2 *Some well-known formulas for marginal and conditional density functions are easily verified with the help of the above lemma. E.g., substituting $\mathbb{R}^{n-1}$ for $\overline{Z}'$, it shows that the probability density function of the nth component of the random vector is*

$$h_n(\overline{z}_n) \;=\; \int\limits_{\mathbb{R}^{n-1}} h(z_1,\dots,z_{n-1},\,\overline{z}_n)\,d\,(z_1,\dots,z_{n-1}) \qquad (\overline{z}_n\in\mathbb{R}).$$

Moreover, given $\overline{z}_n\in\mathbb{R}$, the function $(z_1,\dots,z_{n-1})\mapsto\frac{1}{h_n(\overline{z}_n)}\,h(z_1,\dots,z_{n-1},\,\overline{z}_n)$ is the density function belonging to the conditional distribution function $F(z_1,\dots,z_{n-1}\,|\,\overline{z}_n)$.

Proof of theorem 3.2
(i). By lemma B.1, we get

$$\boldsymbol{e}_n^T\nabla F(\overline{\boldsymbol{z}}) \;\le\; \int\limits_{\mathbb{R}^{n-1}} h(z_1,\dots,z_{n-1},\,\overline{z}_n)\,d\,(z_1,\dots,z_{n-1}). \tag{38}$$

The right-hand expression is $h_n(\overline{z}_n)$, where $h_n(.)$ is the probability density function of the nth component of the random vector. Let $\sigma_n^2>0$ denote the variance. The value of the density function is maximized in the expectation, hence the maximum is $\frac{1}{\sigma_n\sqrt{2\pi}}$.

(ii). Applying lemma B.1, the difference $\big|\,\boldsymbol{e}_n^T\nabla F(\overline{\boldsymbol{z}}+t\boldsymbol{e}_n) - \boldsymbol{e}_n^T\nabla F(\overline{\boldsymbol{z}})\,\big|$ is bounded by

$$\int\limits_{\mathbb{R}^{n-1}}\Big|\,h(z_1,\dots,z_{n-1},\,\overline{z}_n+t) - h(z_1,\dots,z_{n-1},\,\overline{z}_n)\Big|\,d\,(z_1,\dots,z_{n-1}). \tag{39}$$

Due to the mean-value theorem, an upper bound on the integrand is

$$|t|\,\max_{z_n\in\mathbb{R}}\Big|\,\boldsymbol{e}_n^T\,\nabla h(z_1,\dots,z_{n-1},\,z_n)\,\Big|, \tag{40}$$

for any $(z_1,\dots,z_{n-1})\in\mathbb{R}^{n-1}$.

We start with computing the maximum in (40). As this maximum is considered over $\mathbb{R}$, and the integration domain in (39) is $\mathbb{R}^{n-1}$, we may assume that the expectation vector of our normal distribution is $\mathbf{0}$. As for the covariance matrix, let us compute its Cholesky decomposition in the form $C=AA^T$ where $A\in\mathbb{R}^{n\times n}$ is a lower triangular matrix. Considering inverses, we get $C^{-1}=L^TL$, where $L=A^{-1}$, a lower triangular matrix. Applying these in the formula of the normal density function, we get

$$\nabla h(\boldsymbol{z}) \;=\; -\frac{|L|}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\|L\boldsymbol{z}\|^2\right)L^TL\boldsymbol{z} \qquad (\boldsymbol{z}\in\mathbb{R}^n). \tag{41}$$

Let us partition $\boldsymbol{z}\in\mathbb{R}^n$ in the form $\boldsymbol{z}^T=\big(\boldsymbol{z}'^T,z_n\big)$, where $\boldsymbol{z}'\in\mathbb{R}^{n-1}$. Let us partition the lower triangular matrix accordingly:

$$L=\begin{bmatrix} L' & \mathbf{0}\\ \boldsymbol{l}'^T & l_{n,n}\end{bmatrix};\qquad\text{hence}\qquad L\boldsymbol{z}=\begin{bmatrix} L'\boldsymbol{z}'\\ \boldsymbol{l}'^T\boldsymbol{z}'+l_{n,n}z_n\end{bmatrix}, \tag{42}$$

where $L' \in \mathbb{R}^{(n-1)\times(n-1)}$, $\boldsymbol{l}', \mathbf{0} \in \mathbb{R}^{n-1}$ and $l_{n,n} \in \mathbb{R}$, $l_{n,n} > 0$.

We have $\|L\boldsymbol{z}\|^2 = \|L'\boldsymbol{z}'\|^2 + \left(\boldsymbol{l}'^T\boldsymbol{z}' + l_{n,n}z_n\right)^2$. Moreover, the nth component of $L^T L\boldsymbol{z}$ is $l_{n,n}\left(\boldsymbol{l}'^T\boldsymbol{z}' + l_{n,n}z_n\right)$. Using these, the absolute value of the nth component of the gradient (41) is

$$\frac{|L|\, l_{n,n}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|L'\boldsymbol{z}'\|^2\right) \underbrace{\exp\left(-\frac{1}{2}\left(\boldsymbol{l}'^T\boldsymbol{z}' + l_{n,n}z_n\right)^2\right) \left|\boldsymbol{l}'^T\boldsymbol{z}' + l_{n,n}z_n\right|}. \tag{43}$$

z_n occurs in the underbraced part only. Let us introduce the notation $\lambda := \left|\boldsymbol{l}'^T\boldsymbol{z}' + l_{n,n}z_n\right|$. For $\lambda > 0$, let us consider the logarithm of the underbraced expression as a function of λ. This is $-\frac{1}{2}\lambda^2 + \ln(\lambda)$, a differentiable concave function whose derivative disappears at $\lambda = 1$. Therefore the maximum of the underbraced expression, as a function of z_n, is $\frac{1}{\sqrt{e}}$. The corresponding maximum of (43) is

$$\mathcal{B}(\boldsymbol{z}') \;:=\; \frac{1}{\sqrt{e}}\, \frac{1}{(2\pi)^{n/2}}\, |L|\, l_{n,n}\, \exp\left(-\frac{1}{2}\|L'\boldsymbol{z}'\|^2\right).$$

Summing up, $\mathcal{B}(\boldsymbol{z}')$ is the maximum that appears in (40). Hence the integral (39) is bounded by

$$|t| \int\limits_{\mathbb{R}^{n-1}} \mathcal{B}(\boldsymbol{z}')\, d\,\boldsymbol{z}' \;=\; t\, \frac{1}{\sqrt{e}}\, \frac{1}{(2\pi)^{1/2}}\, \frac{|L|}{|L'|}\, l_{n,n} \int\limits_{\mathbb{R}^{n-1}} \frac{|L'|}{(2\pi)^{(n-1)/2}} \exp\left(-\frac{1}{2}\|L'\boldsymbol{z}'\|^2\right) d\,\boldsymbol{z}'. \tag{44}$$

The latter integrand is an $(n-1)$-variate normal density function, hence the integral is 1. It follows that

$$\frac{1}{\sqrt{2e\pi}}\, \frac{|L|}{|L'|}\, l_{n,n}$$

is an appropriate Lipschitz constant in *(ii)* of the theorem. In order to bring it to a more convenient form, note that L was constructed as the inverse of the lower triangular matrix A, which in turn was obtained by the Cholesky factorization of the covariance matrix, i.e., $C = AA^T$. Let C' denote the upper-left $(n-1)\times(n-1)$ submatrix of C, i.e., the covariance matrix of the first $(n-1)$ components of the random vector. Similarly, let A' denote the upper-left $(n-1)\times(n-1)$ submatrix of A. We have $C' = A'A'^T$ and $L' = A'^{-1}$. Hence

$$l_{n,n} = \frac{|L|}{|L'|} = \frac{|A'|}{|A|} \qquad \text{and} \qquad \frac{|A'|^2}{|A|^2} = \frac{|C'|}{|C|},$$

completing the proof. □

Proof of corollary 3.3
Let $\boldsymbol{z},\ \boldsymbol{z} + t\boldsymbol{e}_i \in \mathcal{L}(p)$. As $\varphi(\boldsymbol{z}) = -\log F(\boldsymbol{z})$, we have

$$\begin{aligned}
\left|\, \boldsymbol{e}_i^T\nabla\varphi(\boldsymbol{z} + t\boldsymbol{e}_i) - \boldsymbol{e}_i^T\nabla\varphi(\boldsymbol{z})\,\right| \;&=\; \left|\, \tfrac{1}{F(\boldsymbol{z}+t\boldsymbol{e}_i)}\boldsymbol{e}_i^T\nabla F(\boldsymbol{z} + t\boldsymbol{e}_i) - \tfrac{1}{F(\boldsymbol{z})}\boldsymbol{e}_i^T\nabla F(\boldsymbol{z})\,\right| \\
\leq\; \tfrac{1}{F(\boldsymbol{z}+t\boldsymbol{e}_i)}\left|\, \boldsymbol{e}_i^T\nabla F(\boldsymbol{z} + t\boldsymbol{e}_i) - \boldsymbol{e}_i^T\nabla F(\boldsymbol{z})\,\right| \;&+\; \left|\, \tfrac{1}{F(\boldsymbol{z}+t\boldsymbol{e}_i)} - \tfrac{1}{F(\boldsymbol{z})}\,\right| \left|\boldsymbol{e}_i^T\nabla F(\boldsymbol{z})\right| \qquad (45) \\
&\leq\; \tfrac{1}{p}L_i t \;+\; \tfrac{|F(\boldsymbol{z}+t\boldsymbol{e}_i) - F(\boldsymbol{z})|}{p^2} M_i,
\end{aligned}$$

where the last inequality is a direct consequence of theorem 3.2, *(ii)* and *(i)*. Another consequence of *(i)*, by the mean-value theorem, is

$$|\, F(\boldsymbol{z} + t\boldsymbol{e}_i) - F(\boldsymbol{z})\,| \;\leq M_i t. \tag{46}$$

The proof is completed by substituting this into the right-hand expression of (45). □

References

1. L. Bottou, F.E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60:223–311, 2018.
2. D. Dentcheva, B. Lai, and A. Ruszczyński. Dual methods for probabilistic optimization problems. *Mathematical Methods of Operations Research*, 60:331–346, 2004.
3. D. Dentcheva and G. Martinez. Regularization methods for optimization problems with probabilistic constraints. *Mathematical Programming*, 138:223–251, 2013.
4. D. Dentcheva, A. Prékopa, and A. Ruszczyński. Concavity and efficient points of discrete distributions in probabilistic programming. *Mathematical Programming*, 89:55–77, 2000.
5. C.I. Fábián. Gaining traction: on the convergence of an inner approximation scheme for probability maximization. *Central European Journal of Operations Research*, 29:491–519, 2021.
6. C.I. Fábián, E. Csizmás, R. Drenyovszki, T. Vajnai, L. Kovács, and T. Szántai. A randomized method for handling a difficult function in a convex optimization problem, motivated by probabilistic programming. *Annals of Operations Research*, 2019. 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.
7. C.I. Fábián, E. Csizmás, R. Drenyovszki, W. van Ackooij, T. Vajnai, L. Kovács, and T. Szántai. Probability maximization by inner approximation. *Acta Polytechnica Hungarica*, 15:105–125, 2018. Special issue dedicated to the memory of András Prékopa (editors: A. Bakó, I. Maros and T. Szántai).
8. A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–150, 1992.
9. A. Hantoute, R. Henrion, and P. Pérez-Aros. Subdifferential characterization of probability functions under Gaussian distribution. *Mathematical Programming*, 174:167–194, 2018.
10. R. Henrion. Introduction to chance constraint programming. Technical report, Weierstrass-Institut für Angewandte Analysis und Stochastik, 2004. www.wias-berlin.de/people/henrion/ccp.ps.
11. J. Liu, S.J. Wright, C. Ré, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *Journal of Machine Learning Research*, 16:285–322, 2015.
12. D.G. Luenberger and Y. Ye. *Linear and Nonlinear Programming*. International Series in Operations Research and Management Science. Springer, 2016.
13. Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22:341–362, 2012.
14. A. Prékopa. Logarithmic concave measures with applications to stochastic programming. *Acta Scientiarium Mathematicarum (Szeged)*, 32:301–316, 1971.
15. A. Prékopa. On logarithmic concave measures and functions. *Acta Scientiarium Mathematicarum (Szeged)*, 34:335–343, 1973.
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. R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
20. A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006.
21. W. van Ackooij, V. Berge, W. de Oliveira, and C. Sagastizábal. Probabilistic optimization via approximate p-efficient points and bundle methods. *Computers & Operations Research*, 77:177–193, 2017.
22. S.J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3–34, 2015.