Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 0946 – 8633

A gradient formula for linear chance constraints under Gaussian distribution

René Henrion, Andris Möller

submitted: February 14, 2012

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: rene.henrion@wias-berlin.de
andris.moeller@wias-berlin.de

No. 1687
Berlin 2012

2010 *Mathematics Subject Classification.* 90C15.

Key words and phrases. Chance constraints, probabilistic constraints, derivative of singular normal distribution, derivative of Gaussian probability for polyhedra.

This work was supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 2044975
E-Mail: `preprint@wias-berlin.de`
World Wide Web: `http://www.wias-berlin.de/`

ABSTRACT. We provide an explicit gradient formula for linear chance constraints under a (possibly singular) multivariate Gaussian distribution. This formula allows one to reduce the calculus of gradients to the calculus of values of the same type of chance constraints (in smaller dimension and with different distribution parameters). This is an important aspect for the numerical solution of stochastic optimization problems because existing efficient codes for e.g., calculating singular Gaussian distributions or regular Gaussian probabilities of polyhedra can be employed to calculate gradients at the same time. Moreover, the precision of gradients can be controlled by that of function values which is a great advantage over using finite difference approximations. Finally, higher order derivatives are easily derived explicitly. The use of the obtained formula is illustrated for an example of a transportation network with stochastic demands.

1. INTRODUCTION

A chance constraint (or probabilistic constraint) is an inequality of the type

$$\mathbb{P}\left(g(z,\xi)\leq 0\right)\geq p, \tag{1}$$

where g is a mapping defining a (random) inequality system and ξ is an s- dimensional random vector defined on some probability space $(\Omega,\mathcal{A},\mathbb{P})$. The chance constraint expresses the requirement that a decision vector z is feasible if and only if the random inequality system $g(z,\xi)\leq 0$ is satisfied at least with probability $p\in[0,1]$. The use of chance constraints is highly relevant for engineering problems involving uncertain data. Among its numerous applications one may find topics like water management, telecommunications, electricity network expansion, mineral blending, chemical engineering etc. For a comprehensive overview on the theory, numerics and applications of chance constrained programming, we refer to, e.g., [15], [16], [17].

From a formal viewpoint, a chance constraint is a conventional constraint $\alpha(z)\geq p$ with $\alpha(z):=\mathbb{P}\left(g(z,\xi)\leq 0\right)$ on the decision vector (because the dependence on ξ vanishes by taking the probability). However, the major difficulty imposed by chance constraints arises from the fact that typically no analytical expression is available for α. All one can hope for, in general, are efficient tools for numerically approximating α. On the other hand, calculating just functional values of α is not enough for employing optimization algorithms in reasonable dimension, one also has to have access to gradients of α. The need to calculate gradients of probability functions has been recognized a long time ago and has given rise to many papers on representing such gradients (e.g., [11], [20], [10], [14], [5]). The resulting formulae can be used to approximate $\nabla\alpha$ via Monte Carlo methods similar to α itself.

On the other hand, for special cases much more efficient methods than Monte Carlo may exist for numerical approximation. For instance, if in (1) the random vector is separated, i.e., $g(z,\xi)=\xi-h(z)$, then

$$\mathbb{P}\left(g(z,\xi)\leq 0\right)=\mathbb{P}\left(\xi\leq h(z)\right)=F_\xi(h(z)), \tag{2}$$

where F_ξ denotes the (multivariate) distribution function of ξ. We note that for many prominent multivariate distributions (like Gaussian, t-, Gamma, Dirichlet, Exponential, log-normal, truncated normal) there exist methods for calculating the corresponding distribution function that clearly outperform a crude Monte Carlo approach (see, e.g., [7], [19], [18], [8], [13]). When it comes to calculating gradients of such distribution functions in the context of applying some optimization algorithm, then, of course, it would be desirable to carry out this calculus in a similarly efficient way as it was done for the values themselves. In some special cases it is possible indeed to analytically reduce the calculus of gradients to the calculus of function values of the same distribution. This is true, for instance, for the Dirichlet (see [8], p. 195) and for the Gaussian distribution. We cite here the corresponding result for the Gaussian distribution which will be the starting point for the investigations in this paper. We shall adopt the usual notation $\xi\sim\mathcal{N}\left(\mu,\Sigma\right)$ to characterize an s-dimensional random vector having a normal distribution with expected value μ and covariance matrix Σ.

Theorem 1.1 ([15], p. 204)**.** *Let $\xi \sim \mathcal{N}(\mu, \Sigma)$ with some positive definite covariance matrix $\Sigma = (\sigma_{ij})$ of order (s, s). Then, the distribution function F_ξ is continuously differentiable at any $z \in \mathbb{R}^s$ and*

$$\frac{\partial F_\xi}{\partial z_j}(z) = f_{\xi_j}(z_j) \cdot F_{\tilde{\xi}(z_j)}(z_1, \ldots, z_{j-1}, z_{j+1} \ldots, z_s) \quad (j = 1, \ldots, m).$$

Here, f_{ξ_j} denotes the one-dimensional Gaussian density of the component ξ_j, $\tilde{\xi}(z_j)$ is an (s-1)-dimensional Gaussian random vector distributed according to $\tilde{\xi}(z_j) \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$, $\hat{\mu}$ results from the vector $\mu + \sigma_{jj}^{-1}(z_j - \mu_j)\sigma_j$ by deleting component j and $\hat{\Sigma}$ results from the matrix $\Sigma - \sigma_{jj}^{-1}\sigma_j\sigma_j^T$ by deleting row j and column j, where σ_j refers to column j of Σ. Moreover, $\hat{\Sigma}$ is positive definite.

An important consequence of this theorem is that the same powerful tool used to calculate values of multivariate Gaussian distribution functions (e.g. the MVN code by Genz, [7]) can be used at the same time to calculate the gradient of such distribution function. All one has to do is to adjust the distribution parameters according to the rule specified in the theorem. The purpose of this paper is to generalize this idea to a setting where the Gaussian random vector is not separated as in (2) but subject to a possibly nonregular linear transformation which has important applications in engineering.

2. Linear chance constraints with Gaussian distribution

We are interested in linear chance constraints of the type

$$\mathbb{P}(A\xi \le z) \ge p, \tag{3}$$

where $z \in \mathbb{R}^m$ is a decision vector, A denotes a matrix of order (m, s) and ξ is a s- dimensional Gaussian random vector distributed according to $\xi \sim \mathcal{N}(\mu, \Sigma)$. We shall assume that ξ has a regular Gaussian distribution, i.e., Σ is positive definite. Applications of linear chance constraints of type (3) are abundant in engineering and finance (e.g., water reservoir management [1] or cash matching problem [3]). For applying algorithms to solve optimization problems involving a constraint like (3) we are interested in calculating values and gradients of the function

$$\beta(z) := \mathbb{P}(A\xi \le z). \tag{4}$$

When passing to the linearly transformed random vector $\eta := A\xi$, it is well known that $\eta \sim \mathcal{N}\big(A\mu, A\Sigma A^T\big)$, i.e., η has a Gaussian distribution too and one knows exactly how to derive the parameters of this distribution from those of ξ. This allows then to rewrite β in the form

$$\beta(z) = \mathbb{P}(\eta \le z) = F_\eta(z).$$

In other words, β is the distribution function of some Gaussian distribution with well known parameters. At this point, care has to be taken with respect to the transformation matrix A. In the most favorable situation the rank of A equals m, i.e., the rows of A are linearly independent. Then, the covariance matrix $A\Sigma A^T$ of η is positive definite (of order (m, m)) because so was Σ by assumption. In other words, F_η is again a **regular** multivariate Gaussian distribution function and so one is completely led back to the situation discussed in the introduction: One may calculate F_η using appropriate codes and one may also compute ∇F_η via Theorem 1.1 upon respecting the transformed parameters $A\mu$ and $A\Sigma A^T$. Hence, there is no substantial impact of the linear transformation $A\xi$ in this case. A situation like this arises, for instance, in reservoir problems, where the cumulative amount of time dependent random inflows enters the description of the chance constraint. Accumulation of components can be described by a regular lower triangular matrix A.

In many other applications however (e.g., network optimization with random demands or avoidance of polyhedral random obstacles in robotics), A has typically more rows than columns ($m > s$) so that definitely rank $A < m$. In this case, the covariance matrix $A\Sigma A^T$ becomes necessarily singular and, hence, F_η is a **singular** multivariate Gaussian distribution function. In particular Theorem 1.1 does not apply (and cannot apply because F_η is not differentiable in general). Nevertheless, values of F_η can still be calculated efficiently in moderate dimension. One possibility is to employ an algorithm specially designed for singular Gaussian distribution functions (see Genz and Kwong [6]). A second possibility is to use Deák's method for calculating Gaussian probabilities of convex sets which applies of course to the polyhedron $A\xi \leq z$ (see [2]). Now the important question arises, whether in the vein of Theorem 1.1 it is again possible to analytically reduce the calculus of gradients of F_η to that of values of F_η and thus to benefit from the aforementioned algorithmic approaches in order to obtain sufficiently precise approximations for the gradients with reasonable effort. The answer given in this paper is affirmative, and, in the main result proved in the following section, we shall present a generalization of Theorem 1.1 to the singular case (of course under an additional assumption guaranteeing differentiability). Apart from the just mentioned important algorithmic aspect, our gradient formula has further impact on numerics in that it allows to control the precision of the gradient by that of function values (and thus promises much better results than by using finite difference approximations which are prone to noise) and to explicitly calculate higher order derivatives. These issues are discussed in detail in section 4. The relation with existing gradient formulae as they were mentioned in the beginning of the introduction is also addressed in this same section. Finally, section 5 presents an application to network capacity optimization illustrating the numerical use of the gradient formula.

3. Main result

We start by introducing the family of active index sets associated with the polyhedron $Ax \leq z$ given A and z as introduced in (3) (with a_i^T denoting the rows of A):

(5)

$$\mathcal{I}(A,z) := \{I \subseteq \{1,\ldots,m\} | \exists x \in \mathbb{R}^s : a_i^T x = z_i \quad (i \in I), \quad a_i^T x < z_i \quad (i \in \{1,\ldots,m\} \backslash I)\}.$$

Definition 3.1. *The linear inequality system $Ax \leq z$ is called nondegenerate if*

$$\operatorname{rank} \{a_i\}_{i \in I} = \# I \quad \forall I \in \mathcal{I}(A,z).$$

In the language of optimization theory, nondegeneracy means that the inequality system $Ax \leq z$ satisfies the Linear Independence Constraint Qualification. Observe that, if the linear inequality system $Ax \leq z$ is nondegenerate and has a solution at all then the set of solutions has nonempty interior, whence $\emptyset \in \mathcal{I}(A,z)$ (see Corollary 6.1).

The following theorem is a translation of a result by Naiman and Wynn ([12], Th. 2) to our notation and our setting (see also Th. 3.2. in [9]):

Theorem 3.1. *Let z be such that the system $Ax \leq z$ is nondegenerate. Furthermore, let ξ be an s-dimensional random vector distributed according to $\xi \sim \mathcal{N}(\mu, \Sigma)$ with some positive definite Σ. Then, the distribution function associated with $\eta := A\xi$ satisfies*

$$F_\eta(z) = \sum_{I \in \mathcal{I}(A,z)} (-1)^{\# I} F_{-\eta^I}(-z^I), \tag{6}$$

where η^I and z^I are subvectors of η and z, respectively, according to the index set I. In (6), the corresponding term for $I := \emptyset$ is defined to take value 1. Moreover, for $I \neq \emptyset$, the random vectors $-\eta^I$ have a regular Gaussian distribution according to

$$-\eta^I \sim \mathcal{N}\left(-A^I \mu, A^I \Sigma (A^I)^T\right), \tag{7}$$

where A^I is the submatrix of A defined by selecting rows according to the index set I.

Theorem 3.1 allows one to reduce the calculus of a possibly singular Gaussian distribution function F_η to the calculus of (possibly many) regular Gaussian distribution functions $F_{-\eta^I}(I \in \mathcal{I}(A,z))$. An important consequence of the theorem is that it provides us with a tool for calculating the gradient of a singular Gaussian distribution function (under the nondegeneracy assumption made) because the terms on the right hand side of (6) do have gradients as regular Gaussian distribution functions (recall Th. 1.1). More precisely, we have the following Theorem, where the meaning of superscript index sets is as in Theorem 3.1:

Theorem 3.2. *Under the assumptions of Theorem 3.1, F_η is continuously differentiable and it holds that*

$$\frac{\partial F_\eta}{\partial z_j}(z) = -f_j(z_j) \sum_{I \in \mathcal{I}(A,z): j \in I} (-1)^{\# I} F_{\tilde{\eta}(I,j)}(-z^{I \setminus \{j\}}) \quad (j = 1, \ldots, m). \tag{8}$$

Here, f_j denotes the one-dimensional Gaussian density of the component $\eta_j \sim \mathcal{N}(a_j^T \mu, a_j^T \Sigma a_j)$ and the $\tilde{\eta}(I,j)$ are Gaussian random vectors of dimension $\# I - 1$ with distribution $\tilde{\eta}(I,j) \sim \mathcal{N}(\mu(I,j), \Sigma(I,j))$, where

$$\mu(I,j) := -A^{I \setminus \{j\}} \left(\mu + \frac{z_j - a_j^T \mu}{a_j^T \Sigma a_j} \Sigma a_j \right) \tag{9}$$

$$\Sigma(I,j) := A^{I \setminus \{j\}} \left(\Sigma - \frac{1}{a_j^T \Sigma a_j} \Sigma a_j a_j^T \Sigma \right) \left(A^{I \setminus \{j\}} \right)^T. \tag{10}$$

Moreover, the $\Sigma(I,j)$ are positive definite.

Proof. Fix an arbitrary differentiation index $j \in \{1, \ldots, m\}$. According to Prop. 3.1 in [9], the nondegeneracy assumption on the inequality system $Ax \le z$ implies that $\mathcal{I}(A, z') = \mathcal{I}(A, z)$ for all z' close to z. As a consequence, the index sets $I \in \mathcal{I}(A,z)$ in (6) do not change under small perturbations of z and, hence, we are allowed to differentiate $F_\eta(z)$ in (6) term by term with respect to z_j. Doing so first for index sets I with $j \notin I$, we obviously get

$$\frac{\partial F_{-\eta^I}}{\partial z_j}(-z^I) = 0.$$

Therefore, differentiation of (6) yields

$$\frac{\partial F_\eta}{\partial z_j}(z) = \sum_{I \in \mathcal{I}(A,z): j \in I} (-1)^{\# I} \frac{\partial F_{-\eta^I}}{\partial z_j}(-z^I). \tag{11}$$

Now, for the remaining terms one has $j \in I$ and, since by Theorem 3.1 the $-\eta^I$ have a regular Gaussian distribution according to (7), we may apply Theorem 1.1 to see that

$$\frac{\partial F_{-\eta^I}}{\partial z_j}(-z^I) = -f_{-\eta_j}(-z_j) F_{\tilde{\eta}(I,j)}(-z^{I \setminus \{j\}}) = -f_{\eta_j}(z_j) F_{\tilde{\eta}(I,j)}(-z^{I \setminus \{j\}}), \tag{12}$$

where $\tilde{\eta}(I,j) \sim \mathcal{N}(\mu(I,j), \Sigma(I,j))$ for certain mean vectors and covariance matrices to be determined according to the rules of Theorem 1.1. Combination of (11) and (12) yields (8), hence it remains to verify (9) and (10). Observe first that the diagonal element of the matrix $A^I \Sigma (A^I)^T$ corresponding to index $j \in I$ equals $a_j^T \Sigma a_j$. Note that $a_j^T \Sigma a_j \neq 0$ because Σ is positive definite and $a_j \neq 0$ (see Corollary 6.1). Moreover, the column of the matrix $A^I \Sigma (A^I)^T$ corresponding to index $j \in I$ equals $A^I \Sigma a_j$. Therefore, applying Theorem 1.1 to the parameters of (7), $\mu(I,j)$ results from the vector

$$-A^I \mu + \frac{1}{a_j^T \Sigma a_j} \left(-z_j + a_j^T \mu \right) A^I \Sigma a_j$$

by deleting the component corresponding to index j. This, of course, yields (9). Similarly, $\Sigma(I,j)$ results from the matrix

$$A^I \Sigma (A^I)^T - \frac{1}{a_j^T \Sigma a_j} A^I \Sigma a_j a_j^T \Sigma (A^I)^T = A^I \left(\Sigma - \frac{1}{a_j^T \Sigma a_j} \Sigma a_j a_j^T \Sigma \right) (A^I)^T$$

by deleting the row and column corresponding to index j. This yields (10). That the $\Sigma(I,j)$ are positive definite, follows from the corresponding last statement of Theorem 1.1. □

In principle, Theorem 3.2 already comes close to our intentions: it represents the gradient ∇F_η in terms of values of regular Gaussian distribution functions $F_{\tilde{\eta}(I,j)}$ which can be efficiently calculated. However, the practical use of the derived formula is limited because the number of terms in the alternating sum (8) may become extremely large. Nonetheless, Theorem 3.2 is crucial for proving our main result which provides a practicable representation of gradients. The following lemma compiles some elementary statements needed further on.

Lemma 3.1. *For the following expressions occuring in* (9) *and* (10)*,*

$$S^{(j)} := \Sigma - \frac{1}{a_j^T \Sigma a_j} \Sigma a_j a_j^T \Sigma, \quad w^{(j)} := \mu + \frac{z_j - a_j^T \mu}{a_j^T \Sigma a_j} \Sigma a_j \quad (j = 1, \ldots, m),$$

one has that:

1 $S^{(j)}$ *is symmetric and positive semidefinite.*
2 *ker* $S^{(j)} = \mathbb{R}\{a_j\}$.
3 *There exists a factorization* $S^{(j)} = L^{(j)} L^{(j)^T}$*, where* $L^{(j)}$ *is of order* $(s, s-1)$ *and rank* $L^{(j)} = s - 1$.
4 $a_j^T L^{(j)} = 0$.
5 $a_j^T w^{(j)} = z_j$.

Proof. Symmetry of $S^{(j)}$ is evident. With the Cholesky decomposition $\Sigma = PP^T$ of the positive definite and symmetric matrix Σ, the Cauchy-Schwarz inequality yields that

$$\begin{aligned} v^T S^{(j)} v &= v^T \Sigma v - \left(a_j^T \Sigma a_j\right)^{-1} \left(v^T \Sigma a_j\right)^2 = \|Pv\|^2 - \|Pa_j\|^{-2} \langle Pv, Pa_j \rangle^2 \\ &\geq \|Pv\|^2 - \|Pa_j\|^{-2} \|Pv\|^2 \|Pa_j\|^2 = 0 \end{aligned}$$

for all $v \in \mathbb{R}^s$. Hence, $S^{(j)}$ is positive semidefinite. Evidently, $a_j \in$ ker $S^{(j)}$, whence $\mathbb{R}\{a_j\} \subseteq$ ker $S^{(j)}$. Conversely, $v \in$ ker $S^{(j)}$ implies

$$\Sigma \left(v - \frac{a_j^T \Sigma v}{a_j^T \Sigma a_j} a_j \right) = 0.$$

Since Σ is regular, one derives that $v = \left(a_j^T \Sigma a_j\right)^{-1} \left(a_j^T \Sigma v\right) a_j$, whence $v \in \mathbb{R}\{a_j\}$. Therefore, ker $S^{(j)} = \mathbb{R}\{a_j\}$ and, consequently, rank $S^{(j)} = s - 1$. Since $S^{(j)}$ is also symmetric and positive semidefinite, there exist an orthogonal matrix $V^{(j)}$ (of eigenvectors) and a diagonal matrix $\Lambda^{(j)} :=$ diag $\left[\lambda_1^{(j)}, \ldots, \lambda_{s-1}^{(j)}, 0\right]$ (of eigenvalues) with $\lambda_1^{(j)} > 0, \ldots, \lambda_{s-1}^{(j)} > 0$ and $S^{(j)} = V^{(j)} \Lambda^{(j)} V^{(j)^T} = L^{(j)} L^{(j)^T}$, where $L^{(j)} := V^{(j)} \Lambda^{(j)^{1/2}}$. Clearly, rank $L^{(j)} = s - 1$. Finally, $\left\|a_j^T L^{(j)}\right\|^2 = a_j^T S^{(j)} a_j = 0$ (see (*ii*)), whence assertion (*iv*) holds true. Assertion (*v*) is obvious from the definition of $w^{(j)}$. □

Now, we are in a position to prove our main result:

Theorem 3.3. *Let* $z \in \mathbb{R}^m$ *be such that the system* $Ax \leq z$ *is nondegenerate, where* A *is of order* (m, s)*. Furthermore, let* $\xi \sim \mathcal{N}(\mu, \Sigma)$ *with* $\mu \in \mathbb{R}^s$ *and positive definite* Σ *of order* (s, s)*. Then, for* $j = 1, \ldots, m$*, one has the formula*

$$\frac{\partial}{\partial z_j} \mathbb{P}\left(A\xi \leq z\right) = \begin{cases} 0 & \textit{if } \{j\} \notin \mathcal{I}(A,z) \\ f_j(z_j) \mathbb{P}\left(A^{(j)} L^{(j)} \xi^{(j)} \leq z^{(j)} - A^{(j)} w^{(j)}\right) & \textit{if } \{j\} \in \mathcal{I}(A,z) \end{cases},$$

where $\xi^{(j)} \sim \mathcal{N}(0, I_{s-1})$*,* $A^{(j)}$ *results from* A *by deleting row* j*,* $z^{(j)}$ *results from* z *by deleting component* j*,* $\mathcal{I}(A,z)$*,* $L^{(j)}$ *and* $w^{(j)}$ *are defined in* (5) *and Lemma 3.1, respectively, and* f_j *is the one-dimensional Gaussian density with mean value* $a_j^T \mu$ *and variance* $a_j^T \Sigma a_j$*. Moreover, the inequality system*

(13)
$$A^{(j)} L^{(j)} y \leq z^{(j)} - A^{(j)} w^{(j)}$$

occuring in the second case of the formula is nondegenerate.

Proof. Consider first the case $\{j\} \notin \mathcal{I}(A,z)$. Assume that there exists some x such that $Ax \leq z$ and $a_j^T x = z_j$. Putting

$$I := \{i \in \{1,\ldots,m\} | a_i^T x = z_i\},$$

we see that $I \in \mathcal{I}(A,z)$ and $\{j\} \subseteq I$, whence the contradiction $\{j\} \in \mathcal{I}(A,z)$ from Lemma 6.1 which is proved in the appendix. Hence, $a_j^T x < z_j$ for all x with $Ax \leq z$. In other words, the inequality $a_j^T x \leq z_j$ is redundant for the system $Ax \leq z$ and this situation is stable under small perturbations of z_j. Hence the solution set of $Ax \leq z$ is locally constant with respect to variations of z_j and we obtain the (trivial) first part of our derivative formula.

From now on, we assume that $\{j\} \in \mathcal{I}(A,z)$ is arbitrarily fixed. To unburden the notation, we assume without loss of generality that $j = m$. Now, Proposition 6.1 (with $j = m$) proved in the appendix yields the identity

$$\mathcal{I}^{(m)} = \{I \backslash \{m\} | I \in \mathcal{I}(A,z),\, m \in I\}\,, \tag{14}$$

where $\mathcal{I}^{(m)}$ is introduced in (27) as the family of active indices of the inequality system (13) (for $j = m$). Now, let $\hat{I} \in \mathcal{I}^{(m)}$ be arbitrarily given. Then, (14) yields the existence of some index set $I \in \mathcal{I}(A,z)$ such that $m \in I$ and $\hat{I} = I \backslash \{m\}$. From (10) in Theorem 3.2 and (*iii*) in Lemma 3.1, we infer that the matrix

$$A^{\hat{I}} L^{(m)} L^{(m)^T} \left(A^{\hat{I}}\right)^T$$

is positive definite. Consequently, $\operatorname{rank} A^{\hat{I}} L^{(m)} = \#\hat{I}$, which proves that the inequality system (13) is nondegenerate (for $j = m$) in the sense of Definition 3.1.

Now, let some Gaussian random vector $\xi^{(m)} \sim \mathcal{N}(0, I_{s-1})$ be given. The just shown nondegeneracy of the inequality system (13) for $j = m$, allows us to put $\hat{\eta} := A^{(m)} L^{(m)} \xi^{(m)}$ and to apply Theorem 3.1:

$$\mathbb{P}\left(A^{(m)} L^{(m)} \xi^{(m)} \leq z^{(m)} - A^{(m)} w^{(m)}\right) = F_{\hat{\eta}}\left(z^{(m)} - A^{(m)} w^{(m)}\right) \tag{15}$$

$$= \sum_{\hat{I} \in \mathcal{I}^{(m)}} (-1)^{\#\hat{I}} F_{-\hat{\eta}^{\hat{I}}}\left(-\left(z^{(m)} - A^{(m)} w^{(m)}\right)^{\hat{I}}\right). \tag{16}$$

Here, we have taken into account the above mentioned fact that $\mathcal{I}^{(m)}$ is the family of active indices of (13) (for $j = m$). By definition in the statement of this theorem, $z^{(m)}$and $A^{(m)} w^{(m)}$ result from the vectors z and $Aw^{(m)}$ by deleting the respective component m. Moreover, the upper index set $\hat{I}$ indicates that only components with indices from $\hat{I}$ have to be retained in the given vector (see statement of Theorem 3.1). Furthermore, (14) implies that $m \notin \hat{I}$ for $\hat{I} \in \mathcal{I}^{(m)}$. Therefore, we may conclude that

$$\left(z^{(m)} - A^{(m)} w^{(m)}\right)^{\hat{I}} = \left(z^{(m)}\right)^{\hat{I}} - \left(A^{(m)}\right)^{\hat{I}} w^{(m)} = z^{\hat{I}} - A^{\hat{I}} w^{(m)} \quad \forall \hat{I} \in \mathcal{I}^{(m)}.$$

Similarly,

$$\hat{\eta}^{\hat{I}} = \left(A^{(m)} L^{(m)} \xi^{(m)}\right)^{\hat{I}} = A^{\hat{I}} L^{(m)} \xi^{(m)} \quad \forall \hat{I} \in \mathcal{I}^{(m)}.$$

This allows us, upon taking into account (14) again, to continue (16) as

$$\begin{aligned} & \mathbb{P}\left(A^{(m)} L^{(m)} \xi^{(m)} \leq z^{(m)} - A^{(m)} w^{(m)}\right) \\ = & \sum_{I \in \mathcal{I}(A,z),\, m \in I} (-1)^{\#(I \backslash \{m\})} F_{-\hat{\eta}^{I \backslash \{m\}}}\left(A^{I \backslash \{m\}} w^{(m)} - z^{I \backslash \{m\}}\right) \\ = & \sum_{I \in \mathcal{I}(A,z): m \in I} (-1)^{\#(I \backslash \{m\})} \mathbb{P}\left(-A^{I \backslash \{m\}} L^{(m)} \xi^{(m)} \leq A^{I \backslash \{m\}} w^{(m)} - z^{I \backslash \{m\}}\right). \end{aligned} \tag{17}$$

From $\xi^{(m)} \sim \mathcal{N}(0, I_{s-1})$, (10) and the definition of $S^{(m)}$ and $L^{(m)}$ in Lemma 3.1, we infer that for any index set $I \in \mathcal{I}(A,z)$ with $m \in I$:

$$-A^{I\backslash\{m\}}L^{(m)}\xi^{(m)} \sim \mathcal{N}\left(0, A^{I\backslash\{m\}}L^{(m)}L^{(m)^T}A^{I\backslash\{m\}^T}\right) = \mathcal{N}(0, \Sigma\left(I,m\right)).$$

Consequently, by (9) and the definition of $w^{(m)}$ in Lemma 3.1,

$$-A^{I\backslash\{m\}}L^{(m)}\xi^{(m)} - A^{I\backslash\{m\}}w^{(m)} \sim \mathcal{N}(-A^{I\backslash\{m\}}w^{(m)}, \Sigma\left(I,m\right)) = \mathcal{N}(\mu\left(I,m\right), \Sigma\left(I,m\right)),$$

and, hence, the random vectors $\tilde{\eta}\left(I,m\right)$ from Theorem 3.2 (for $j = m$) have the same distribution as the random vectors $-A^{I\backslash\{m\}}L^{(m)}\xi^{(m)} - A^{I\backslash\{m\}}w^{(m)}$. Then, (17) may be continued as

$$\begin{aligned}\mathbb{P}\left(A^{(m)}L^{(m)}\xi^{(m)} \leq z^{(m)} - A^{(m)}w^{(m)}\right) &= \sum_{I\in\mathcal{I}(A,z):m\in I} (-1)^{\#(I\backslash\{m\})}\,\mathbb{P}\left(\tilde{\eta}\left(I,m\right) \leq -z^{I\backslash\{m\}}\right) \\ &= -\sum_{I\in\mathcal{I}(A,z):m\in I} (-1)^{\#I}\,F_{\tilde{\eta}(I,m)}\left(-z^{I\backslash\{m\}}\right).\end{aligned}$$

Now, Theorem 3.2 (for $j = m$ and with $\eta := A\xi$) yields that

$$f_m(z_m)\mathbb{P}\left(A^{(m)}L^{(m)}\xi^{(m)} \leq z^{(m)} - A^{(m)}w^{(m)}\right) = \frac{\partial F_\eta}{\partial z_m}\left(z\right) = \frac{\partial}{\partial z_m}\mathbb{P}\left(A\xi \leq z\right).$$

This, however, is the asserted formula for $j = m$. □

4. Discussion of the result

4.1. **Reduction of gradients to function values.** The importance of Theorem 3.3 relies on the fact that it reduces the computation of gradients to Gaussian probabilities of polyhedra to the computation of objects of the same type, namely Gaussian probabilities of polyhedra (in different dimension, with different parameters). Hence, one may employ, for instance, Deák's method [2] in order to calculate both objects (function values and gradients) by means of the same efficient code. But there also exists an alternative numerical approach to dealing with the chance constraint (3) offered by the same theorem: according to Section 2, the value of $\mathbb{P}\left(A\xi \leq z\right)$ can be interpreted as the value $F_\eta(z)$ of the possibly singular Gaussian distribution function of the random vector $\eta := A\xi$. As already mentioned before, singular Gaussian distribution functions can be calculated by means of an algorithm described in [6]. Now, when it comes to gradients $\nabla F_\eta(z)$, one would be interested of course in a similar representation in terms of objects of the same nature, namely singular Gaussian distribution functions (in different dimension, with different parameters). Such conclusion can be indeed drawn from Theorem 3.3 as shown in the next section.

4.2. **A gradient formula for singular Gaussian distribution functions.** The following Theorem is a direct generalization of the classical Theorem 1.1 by substantially weakening the assumption of positive definiteness for the covariance matrix made there, in other words it generalizes the gradient formula for regular Gaussian distribution functions to singular ones.

Theorem 4.1. *Let $\xi \sim \mathcal{N}(\mu, \Sigma)$ with some (possibly singular) covariance matrix $\Sigma = (\sigma_{ij})$ of order (s,s). Denote by $\Sigma = AA^T$ any factorization of the positive semidefinite matrix Σ (see, e.g., (iii) in Lemma 3.1). Then, the distribution function F_ξ is continuously differentiable at any $z \in \mathbb{R}^s$ for which the inequality system $Ax \leq z - \mu$ is nondegenerate and it holds that*

$$\frac{\partial F_\xi}{\partial z_j}(z) = \begin{cases} 0 & \text{if } \{j\} \notin \mathcal{I}(A, z-\mu) \\ f_{\xi_j}\left(z_j\right)\cdot F_{\tilde{\xi}(z_j)}\left(z_1,\ldots,z_{j-1},z_{j+1}\ldots,z_s\right) \quad (j=1,\ldots,m) & \text{if } \{j\} \in \mathcal{I}(A, z-\mu) \end{cases}.$$

Here, f_{ξ_j} denotes the one-dimensional Gaussian density of the component ξ_j, $\tilde{\xi}(z_j)$ is an (s-1)-dimensional (possibly singular) Gaussian random vector distributed according to $\tilde{\xi}(z_j) \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$, $\hat{\mu}$ results from the vector $\mu+\sigma_{jj}^{-1}\left(z_j-\mu_j\right)\sigma_j$ by deleting component j and $\hat{\Sigma}$ results from the matrix $\Sigma-\sigma_{jj}^{-1}\sigma_j\sigma_j^T$

by deleting row j and column j, where σ_j refers to column j of Σ. Moreover, $\mathcal{I}(A, z-\mu)$ refers to the family of active indices for the inequality system $Ax \leq z-\mu$ (see (5)).

Proof. Let η be a t-dimensional (with t being the number of columns of the matrix A) Gaussian random vector distributed according to $\eta \sim \mathcal{N}(0, I_t)$. Then, the transformed random vector $A\eta + \mu \sim \mathcal{N}(\mu, AI_tA^T) = \mathcal{N}(\mu, \Sigma)$ has the same distribution as ξ. Therefore,

$$\frac{\partial F_\xi}{\partial z_j}(z) = \frac{\partial}{\partial z_j}\mathbb{P}(\xi \leq z) = \frac{\partial}{\partial z_j}\mathbb{P}(A\eta \leq z-\mu).$$

Theorem 3.3 (applied to η rather than ξ and to right-hand side $z-\mu$ rather than just z) then yields that

$$(18)\quad \frac{\partial F_\xi}{\partial z_j}(z) = \begin{cases} 0 & \text{if } \{j\} \notin \mathcal{I}(A, z-\mu) \\ f_j(z_j-\mu_j)\mathbb{P}\left(A^{(j)}L^{(j)}\eta^{(j)} \leq z^{(j)} - \mu^{(j)} - A^{(j)}w^{(j)}\right) & \text{if } \{j\} \in \mathcal{I}(A, z-\mu) \end{cases},$$

where $\eta^{(j)} \sim \mathcal{N}(0, I_{t-1})$, $A^{(j)}$ results from A by deleting row j, $z^{(j)}$ and $\mu^{(j)}$ result from z and μ, respectively, by deleting component j, $L^{(j)}$ and $w^{(j)}$ are defined in (5) and Lemma 3.1, respectively, (but applied to the distribution parameters of η rather than ξ) and f_j is the one-dimensional Gaussian density with mean value 0 and variance $a_j^T I_t a_j = \|a_j\|^2$. Since the first case of this formula is already compatible with the corresponding case in the asserted formula, we may continue with the second case where $\{j\} \in \mathcal{I}(A, z-\mu)$. First observe that, by assumption, component j of ξ is distributed according to $\xi_j \sim \mathcal{N}(\mu_j, \sigma_{jj})$. Hence, $\xi_j - \mu_j \sim \mathcal{N}(0, \sigma_{jj}) = \mathcal{N}(0, \|a_j\|^2)$ by $\Sigma = AA^T$. It follows that the density f_j coincides with the density $f_{\xi_j-\mu_j}$ and we obtain that

$$f_{\xi_j}(z_j) = f_{\xi_j-\mu_j}(z_j-\mu_j) = f_j(z_j-\mu_j).$$

Next, introduce the random vector

$$(19)\quad \tilde{\xi}(z_j) := A^{(j)}L^{(j)}\eta^{(j)} + \mu^{(j)} + A^{(j)}w^{(j)}.$$

Then, the second case of (18) may be written as

$$(20)\quad \frac{\partial F_\xi}{\partial z_j}(z) = f_{\xi_j}(z_j)\, F_{\tilde{\xi}(z_j)}(z^{(j)}) = f_{\xi_j}(z_j)\, F_{\tilde{\xi}(z_j)}(z_1, \ldots, z_{j-1}, z_{j+1} \ldots, z_s),$$

where $F_{\tilde{\xi}(z_j)}$ refers to the distribution function of $\tilde{\xi}(z_j)$. Since $\eta^{(j)} \sim \mathcal{N}(0, I_{t-1})$, we derive from (19) that

$$\tilde{\xi}(z_j) \sim \mathcal{N}(\mu^{(j)} + A^{(j)}w^{(j)}, A^{(j)}L^{(j)}L^{(j)T}A^{(j)T}).$$

In view of (20), the Theorem will be proved, once we have checked that the above parameters of $\tilde{\xi}(z_j)$ coincide with those asserted in the statement of the theorem, i.e., we have to show that

$$(21)\quad \hat{\mu} = \mu^{(j)} + A^{(j)}w^{(j)}$$

$$(22)\quad \hat{\Sigma} = A^{(j)}L^{(j)}L^{(j)T}A^{(j)T}$$

As far as (22) is concerned, recall that $L^{(j)}L^{(j)T} = S^{(j)}$ by definition of $L^{(j)}$ in Lemma 3.1 (iii), where $S^{(j)}$ calculates according to its definition in Lemma 3.1 but with the covariance matrix of η (which is I_t). Accordingly,

$$S^{(j)} = I_t - \|a_j\|^{-2}\, a_j a_j^T.$$

Recalling that $\sigma_{jj} = \|a_j\|^2$, we arrive at

$$A^{(j)}L^{(j)}L^{(j)T}A^{(j)T} = A^{(j)}A^{(j)T} - \sigma_{jj}^{-1}A^{(j)}a_j a_j^T A^{(j)T}.$$

Since $A^{(j)}$ results from A by deleting row j, it follows that the matrix $A^{(j)}L^{(j)}L^{(j)T}A^{(j)T}$ results from the matrix

$$AA^T - \sigma_{jj}^{-1}Aa_j a_j^T A^T = \Sigma - \sigma_{jj}^{-1}\sigma_j\sigma_j^T$$

by deleting row j and column j. This proves (22). Addressing now (21), we calculate first $w^{(j)}$ from its definition in Lemma 3.1 but with the mean vector and covariance matrix of η (which are 0 and I_t, respectively) and with the argument $z_j - \mu_j$ rather than z_j (see remark before (18)). Accordingly,

$$w^{(j)} = \|a_j\|^{-2}(z_j - \mu_j)a_j.$$

It follows from the definition of $\mu^{(j)}$ and $A^{(j)}$ that $\mu^{(j)} + A^{(j)}w^{(j)}$ results from the vector

$$\mu + Aw^{(j)} = \mu + \sigma_{jj}^{-1}(z_j - \mu_j)\sigma_j$$

by deleting component j. This proves (21). □

4.3. **Control of precision for gradients.** Usually, the absolute error in calculating probabilities $\mathbb{P}(A\xi \leq z)$ can be controlled in the application of numerical methods. Let us assume that the discrepancy between theoretical and computed values is bounded by some $\varepsilon > 0$. Then, according to Theorem 3.3, the absolute error in the computation of partial derivatives can be estimated by

$$\begin{aligned}&\left|\frac{\partial}{\partial z_j}\mathbb{P}\left(A\xi \leq z\right) - \left(\frac{\partial}{\partial z_j}\mathbb{P}\left(A\xi \leq z\right)\right)^{\mathrm{comp}}\right| \\ &= f_j(z_j)\left|\mathbb{P}\left(\hat{A}^{(j)}\xi^{(j)} \leq \hat{z}^{(j)}\right) - \left(\mathbb{P}\left(\hat{A}^{(j)}\xi^{(j)} \leq \hat{z}^{(j)}\right)\right)^{\mathrm{comp}}\right| \leq f_j(z_j)\varepsilon,\end{aligned} \tag{23}$$

where $\hat{A}^{(j)} = A^{(j)}L^{(j)}$ and $\hat{z}^{(j)} = z^{(j)} - A^{(j)}w^{(j)}$. Hence, the absolute error in the computation of partial derivatives can be controlled by that of function values. This information, however, is of limited use because already the nominal values of partial derivatives are typically small. Moreover, for numerical optimization (e.g., cutting plane method), the direction of a gradient is more important than its norm. Therefore, one should be more interested in controlling the precision of normed gradients. Using the maximum norm and applying first the triangle inequality and then (23), one gets that

$$\begin{aligned}&\left\|\frac{\nabla\mathbb{P}\left(A\xi \leq z\right)}{\left\|\nabla\mathbb{P}\left(A\xi \leq z\right)\right\|_\infty} - \frac{\left(\nabla\mathbb{P}\left(A\xi \leq z\right)\right)^{\mathrm{comp}}}{\left\|\left(\nabla\mathbb{P}\left(A\xi \leq z\right)\right)^{\mathrm{comp}}\right\|_\infty}\right\|_\infty \leq 2\frac{\left\|\nabla\mathbb{P}\left(A\xi \leq z\right) - \left(\nabla\mathbb{P}\left(A\xi \leq z\right)\right)^{\mathrm{comp}}\right\|_\infty}{\left\|\left(\nabla\mathbb{P}\left(A\xi \leq z\right)\right)^{\mathrm{comp}}\right\|_\infty} \\ &= 2\frac{\max\limits_j\left|\frac{\partial}{\partial z_j}\mathbb{P}\left(A\xi \leq z\right) - \left(\frac{\partial}{\partial z_j}\mathbb{P}\left(A\xi \leq z\right)\right)^{\mathrm{comp}}\right|}{\left\|\left(\nabla\mathbb{P}\left(A\xi \leq z\right)\right)^{\mathrm{comp}}\right\|_\infty} \leq 2\varepsilon\frac{\max\limits_j f_j(z_j)}{\left\|\left(\nabla\mathbb{P}\left(A\xi \leq z\right)\right)^{\mathrm{comp}}\right\|_\infty}.\end{aligned}$$

Since all quantities on the right-hand side are available at any given z, it is possible in this way to estimate the precision of the normed computed gradient from the chosen precision of the absolute error for function values without knowing explicitly the theoretical gradient $\nabla\mathbb{P}\left(A\xi \leq z\right)$ at z.

4.4. **Higher order derivatives.** Another important feature of Theorem 3.3 is its inductive nature: if the original inequality system $Ax \leq z$ happens to be nondegenerate, then so does the reduced inequality system $\hat{A}y \leq \hat{z}$ occuring in the derivative formula of Theorem 3.3. This means that the reduced inequality system fulfills the assumptions of the same Theorem again, so its consecutive application allows one to calculate derivatives of any order. In other words, at such arguments z (satisfying nondegeneracy), the given probability function is of class $\mathcal{C}^\infty$. In particular, as a consequence of Theorem 4.1, singular Gaussian distribution functions are of class $\mathcal{C}^\infty$ at any points z satisfying the nondegeneracy condition of that theorem. Though an explicit formula for second order derivatives could be given on the basis of Theorem 3.3, it seems to be more elegant to recursively apply the result in a numerical context. We do not have any experience so far, to judge whether or not the effort to calculate second order derivatives would pay in the context of solving a chance constrained optimization problem of the given type.

4.5. **A numerical solution approach for optimization problems with chance constraint (3).** Let us consider the following optimization problem:

$$\min\{c^T z | \mathbb{P}\left(A\xi \leq z\right) \geq p\}, \tag{24}$$

where $z \in \mathbb{R}^m$ is a decision vector, $c \in \mathbb{R}^m$ is a cost vector, A denotes a matrix of order (m,s), $p \in [0,1]$ is a probability level and ξ is an s- dimensional Gaussian random vector distributed according to $\xi \sim \mathcal{N}(\mu, \Sigma)$ with Σ positive definite. The first important observation concerning the solution of (24) is that the feasible set defined by the chance constraint happens to be convex. Indeed this is an immediate consequence of the theory of log-concave measures by Prékopa [15]: the Gaussian distribution is log-concave and so is any linear transformation of it. This implies the mapping $z \mapsto \log \mathbb{P}\left(A\xi \leq z\right)$ to be concave which in turn shows that the feasible set defined by the equivalent logarithmized chance constraint is convex. As a consequence, (24) may be solved by classical methods of convex optimization, for instance by a cutting plane method. This latter approach requires the following components: determination of a Slater point, evaluation of values and gradients (for defining cuts) of the function $z \mapsto \mathbb{P}\left(A\xi \leq z\right)$, solution of a linear program defined by the polyhedral outer approximation of the feasible set. Existence of a Slater point is guaranteed if $p < 1$ (which is typically the case) and such point can be easily determined by driving the components of z uniformly to infinity and thus pushing the probability of $\mathbb{P}\left(A\xi \leq z\right)$ towards one. As already noted before, values of the given probability function can be approximated by existing efficient codes (e.g., [2],[7]), and thanks to Theorem 3.3 (or Theorem 4.1, respectively), the same codes can be employed for computing gradients. There is one additional effort needed at each iteration, however, in order to decide on the case distinction $\{j\} \in \mathcal{I}(A,z)$ in the gradient formulae. This problem can be recasted as a linear program at feasible points z. More precisely, one has the following:

Proposition 4.1. *Let $p > 0$ and z such that the system $Ax \leq z$ is nondegenerate and that it satisfies the chance constraint $\mathbb{P}\left(A\xi \leq z\right) \geq p$. Then the linear program*

$$\text{(P)} \qquad \min\{u_j \mid Ax + u = z,\, u \geq 0\}$$

in variables x, u is solvable for all $j = 1, \ldots, m$ and it holds that $\{j\} \in \mathcal{I}(A,z)$ if and only if the optimal value of (P) equals zero.

Proof. Since $p > 0$ and z satisfies the chance constraint, the system $Ax \leq z$ has a solution. Therefore, the feasible set of (P) is nonempty. On the other hand, the objective of (P) is bounded below by zero on this feasible set. Consequently, (P) is solvable. Let $(\bar{x}, \bar{u})$ be an optimal solution of (P). If the optimal value of (P) happens to be zero, i.e., $\bar{u}_j = 0$, then $A\bar{x} \leq z$ and $a_j^T x = z_j$. Define

$$I := \{i \in \{1, \ldots, m\} \mid a_i^T \bar{x} = z_i\}$$

Then, $\{j\} \subset I$ and from the definition of $\mathcal{I}(A,z)$ one gets that $I \in \mathcal{I}(A,z)$. Now, Lemma 6.1 guarantees that $\{j\} \in \mathcal{I}(A,z)$. Conversely, if $\{j\} \in \mathcal{I}(A,z)$, then, by definition, there exists a point x such that $a_j^T x = z_j$ and $a_i^T x < z_i$ for $i \neq j$. With $u := z - Ax$, we see that (x,u) is feasible for (P) and that the objective in that point equals zero. It follows that the optimal value of (P) equals zero. □

4.6. **The case of rectangle probabilities.** Many chance constrained optimization problems are of two-sided type, where the chance constraint is given by

$$\delta(x) := \mathbb{P}(a(x) \leq \xi \leq b(x)) \geq p$$

with certain mappings a and b acting on the decision vector x (see, e.g., the hydro reservoir problem considered in [1]). With

$$\gamma(z_1, z_2) := \mathbb{P}(\xi \leq z_1, -\xi \leq z_2)$$

one may represent the gradient of δ as

$$\nabla\delta(x) = \nabla_{z_1}\gamma(b(x), a(x)) \circ Db(x) + \nabla_{z_2}\gamma(b(x), a(x)) \circ Da(x).$$

As a and b are usually given by analytical formulae, the interesting part here is represented by the gradient of γ. Clearly, γ is a special case of the function β in (4) with $A = (I, -I)^T$. Hence, one could apply

Theorem 3.3 in order to derive a gradient formula for Gaussian probabilities of rectangles boiling down to Gaussian probabilities of rectangles again (in one dimension less and with new distribution parameters). Due to the simple structure of rectangles there is no need, however, to rely on Theorem 3.3, because the mentioned formula can be derived in a direct and elementary manner then (see [1, Theorem 1]).

4.7. **Truncated Gaussian distribution.** The Gaussian distribution may not be the right characterization of random vectors taking only positive values by their physical nature. One possible alternative then is to model the random vector by means of a truncated Gaussian distribution. More precisely, let $[a, b] \subset \mathbb{R}^s$ be a nondegenerate generalized rectangle, i.e., $a < b$ componentwise and components equal to $\pm\infty$ are allowed. Then, the random vector ξ is said to have a truncated Gaussian distribution $\xi \sim \mathcal{TN}(\mu, \Sigma, a, b)$ if its density is given by

$$f_\xi(z) := \begin{cases} f_\eta(z)/\mathbb{P}(\eta \in [a,b]) & \text{if } z \in [a,b] \\ 0 & \text{else} \end{cases},$$

where $\eta \sim \mathcal{N}(\mu, \Sigma)$ with positive definite Σ and with density f_η. Then,

$$\begin{aligned} \mathbb{P}(A\xi \le z) &= \mathbb{P}(A\xi \le z, \xi \in [a,b]) = \mathbb{P}(A\eta \le z, \eta \in [a,b])/\mathbb{P}(\eta \in [a,b]) \\ &= [\mathbb{P}(\eta \in [a,b])]^{-1} \cdot \mathbb{P}\left(\begin{pmatrix} A \\ I \\ -I \end{pmatrix} \eta \le \begin{pmatrix} z \\ b \\ -a \end{pmatrix} \right) \end{aligned}$$

Now, the arisen inequality system is basically of the form (4) because η has a regular Gaussian distribution. The fact that part of the right-hand side of this inequality system is fixed (in contrast to (4)) does not matter if partial derivatives with respect to z shall be computed because then all remaining components of z are fixed anyway. Consequently, Theorem 3.3 can also be employed to derive gradients of chance constraints (3) in case of truncated Gaussian random vectors by leading this issue back to the case of a standard Gaussian distribution.

4.8. **Relation with existing general derivative formulae.** At this point, one may ask how the gradient formulae of Theorems 3.3 and 4.1 relate to the general derivative formulae mentioned in the introduction. Specializing, for instance, Theorem 1 in [14] or Theorem 2.4 in [10] to the setting which is of interest here, we have the following result:

Theorem 4.2. *Let η be a random vector with continuous density g. Assume that, for a given z, the linear system $Ax \le z$ is nondegenerate (see Def. 3.1) and has a compact solution set. Then,*

$$\frac{\partial}{\partial z_j}\mathbb{P}\left(A\eta \le z\right) = \frac{1}{\|a_j\|} \int\limits_{Ax \le z, a_j^T x = z_j} g(x) do_j(x),$$

where $do_j(x)$ refers to the surface measure on the hyperplane defined by $a_j^T x = z_j$.

Evidently, this formula is not explicit yet because it requires the computation of a surface integral depending on the distribution of η. If η is Gaussian as in our case, it is likely that this computation can be carried out in a way that it leads to a result as in Theorem 3.3. Note, however, that the formula is justified only under the assumption that the polyhedron $Ax \le z$ is compact which is not the case in many applications. This assumption is already violated if $A = I$, i.e., when $\mathbb{P}\left(A\eta \le z\right)$ is the distribution function of η. Indeed the Theorem is false, in general, when dropping the compactness assumption because distribution functions need not be differentiable even if the underlying density g is continuous (as required in the Theorem) or even if g is continuous and bounded along with all its marginal densities. The reason that we are able to prove the gradient formula in Theorem 3.3 without compactness is that we exploit the Gaussian character of the distribution: the compactness issue is already part of the classical regular derivative formula in Theorem 1.1 which is the basis of our result. Note that the main tool for deriving our gradient formula is the alternating representation of singular Gaussian distribution functions in terms of regular ones in Theorem 3.1, which does not rely on any compactness assumption.

5. An Example from network capacity optimization under random demands

In order to illustrate a possible application of the gradient formula obtained in Theorem 3.3, we consider a problem of network capacity optimization under random demand as introduced in [15], p.452. Assume we are given an electricity network with a set of node and arcs and that at each node there exists a random demand of electricity following a joint Gaussian distribution (according to Section 4.7, nonnegativity can be easily taken care of by truncation). The demand of electricity may be covered by production facilities at the nodes as well as by transmission of electricity along the arcs. As in [15], we will assume that the power flow satisfies Kirchhoff's first law only, i.e., it is handled as a linear transportation flow. We will also assume that the network topology is given (of course, in general, this is just a subproblem of a general network design problem). In a planning phase, one may be concerned with installing production and transmission capacities at minimum cost such that future random demand patterns can be covered at a specified probability by directing a suitable flow through the network satisfying the capacity constraints. The question, whether for a specific realization of the demand vector and for given vectors of production and transmission capacities there exists such a feasible flow can be answered by the Gale-Hoffman inequalities stating that for each subset of nodes the total net demand (sum of demands minus production capacities in the nodes of this subset) should be not greater than the total transmission capacity of arcs joining nodes of the considered subset with nodes of its complement. In formal terms, if ξ_i and x_i denote the demand and production capacity at node i and y_j refers to the transmission capacity of arc j then, the following linear inequality system is equivalent with the existence of a feasible node:

$$\sum_{i\in S}(\xi_i - x_i) \leq \sum_{j\in A(S,\bar{S})} y_j \quad \forall S, \tag{25}$$

where S runs through all subsets of nodes and $A(S,\bar{S})$ is the set of arcs joining nodes from S with nodes from $\bar{S}$. We will write the system of linear inequalities in the more compact form of $A\xi \leq Ax+By$, where ξ, x, y refer to the vectors composed of ξ_i, x_i, y_j and the matrices A and B depend on the concrete network topology. The optimization problem can now be formulated as

$$\min\left\{c^Tx + d^Ty|\mathbb{P}\left(A\xi \leq Ax+By\right) \geq p\right\}.$$

Here, c, d are cost vectors for installing production and transmission capacities, respectively, and the chance constraint expresses the fact that in a later operational phase the power demand can be met at probability at least p. Of course, additional explicit constraints (e.g., simple bounds) on the decision variables x, y can be also included. Rewriting the optimization problem in the equivalent form

$$\min\left\{c^Tx + d^Ty|\mathbb{P}\left(A\xi \leq z\right) \geq p,\ z = Ax+By\right\}, \tag{26}$$

we see that the chance constraint is of type (3). According to (25), the number of inequalities equals 2^s if s equals the number of nodes in the network (because the set S in (25) runs through all subsets of $\{1,\ldots,s\}$). Hence, formally, the matrix A occuring in (26) is of order $(2^s, s)$, so that the transformed random vector $A\xi$ has a highly singular Gaussian distribution. Fortunately, it turns out that many inequalities in the huge system $A\xi \leq z$ are redundant (which can be checked by linear programming, for instance). Sometimes, exploiting additional information on possible bounds for the demand, one may be lucky to further reduce the inequality system until the number of finally remaining inequalities is actually smaller than the number of nodes. Then, one is usually back to the regular Gaussian case for which the classical gradient formula from Theorem 1.1 can be employed. Such instance is described in [15] (Section 14.4). However, there is no guarantee to arrive at such comfortable situation, in particular not, if information on efficient bounds for demands is missing. Then, even after deleting redundant inequalities, their number may still substantially exceed the number of nodes (i.e., the dimension of the random vector). As a consequence, Theorem 1.1 is no longer applicable but one may exploit Theorems 3.3 and 4.1 then and embed it in the numerical solution scheme sketched in Section 4.5.

For the purpose of illustration we consider the network depicted in Figure 1 a) consisting of 13 nodes and 13 arcs. The demands at the nodes are assumed to be Gaussian with expected values proportional to

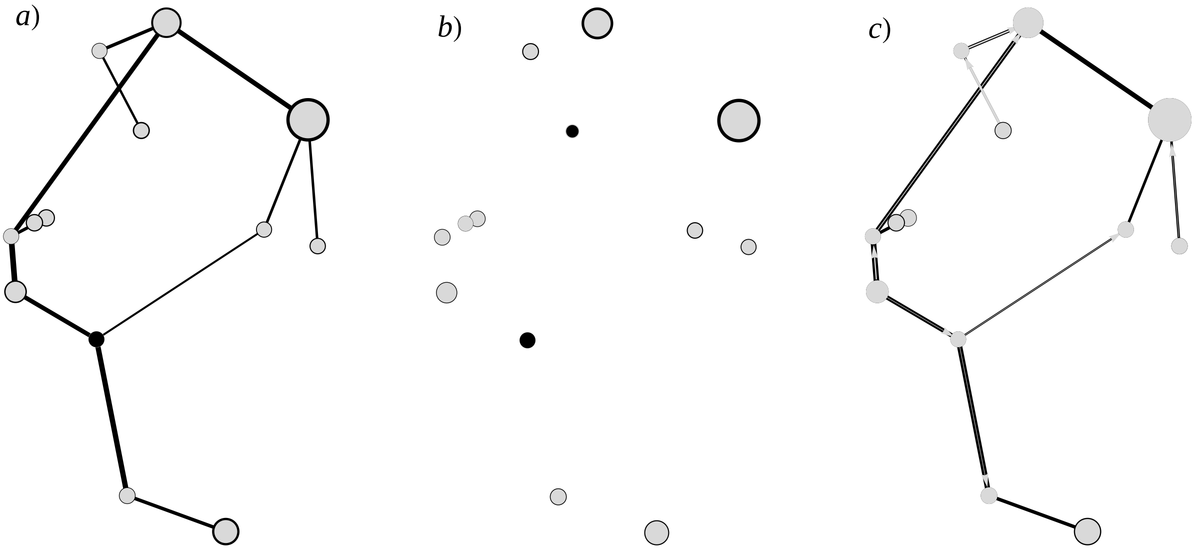

FIGURE 1. Illustration of the solution for a probabilistic network capacity problem. For details see text.

the areas of the gray shaded discs. The covariance matrix was set up as follows: the relative standard deviation (w.r.t. mean values) at each node was chosen to be 20% and between different nodes a common correlation coefficient of 0.3 was assumed. Constant cost coefficients c_j were considered for the installation of production capacities whereas cost coefficients d_j for the installation of transmission capacities were assumed to be proportional to the arc lengths. A probability level of $p = 0.99$ was required for the chance constraint. Given the number of $s = 13$ nodes, one ends up at a number of $2^s = 8192$ Gale-Hoffman inequalities according to (25). After a redundancy check, these could be reduced to 439 inequalities, still substantially exceeding the dimension s of the random vector. According to our previous remarks, the chance constraint in (26) can be either understood as defined by the probability of a rectangle with 439 faces with respect to a 13 dimensional regular Gaussian random vector or as defined by the value of the distribution function of a 439-dimensional (highly singular) Gaussian random vector. The solution of of the problem is shown in Figure 1 a). Optimal transmission capacities y_j are represented by proportional thicknesses of the joining line segments. Optimal production capacities are represented as black discs at the nodes with proportional areas such that the black disc remains in the background if the corresponding capacity exceeds the expected demand (all but one node) and comes into the foreground otherwise (one node). In order to check a posteriori the validity of the obtained solution, we simulated 100 different demand patterns according to the Gaussian distribution specified above. One of these scenarios is illustrated as an example in Figure 1 b), where expected values are gray shaded as in Figure 1 a) and the simulated demand vector is represented by black discs with the same background-foreground rule as before. According to the calculated optimal solution, we should expect that 99 out of 100 scenarios are feasible in the sense of the chance constraint (of course this would only hold true on the average when repeating a simulation of 100 scenarios; in our concrete case, all 100 scenarios turned out to be feasible). Note that feasibility here means that the demands at all nodes for the given scenario can be satisfied by a flow through the network which respects the capacity limits obtained for transmission and production. For the concrete scenario of Figure 1 b) a possible (directed) flow is illustrated in Figure 1 c). The concrete directed transmission is represented by gray arrows of corresponding thickness (all of which fit into the thickness of the capacity line). The needed operational production is represented by gray discs (all of which fit into the black capacity discs).

6. Appendix

The following technical proposition is needed in the proof of Theorem 3.3. First, for each $j \in \{1, \dots, m\}$, we associate with the inequality system (13) the family of active indices $\mathcal{I}^{(j)}$ in the same spirit as $\mathcal{I}(A, z)$ was associated with the originally given inequality system $Ax \leq z$ via (5). Taking into account the quantities defined in the statement of Theorem 3.3, this yields:

$$\begin{aligned} \mathcal{I}^{(j)} = \{I \subseteq \{1, \dots, m\} \backslash \{j\} | \exists y \in \mathbb{R}^{s-1} : \ & a_i^T L^{(j)} y = z_i - a_i^T w^{(j)} \quad (i \in I) \\ & a_i^T L^{(j)} y < z_i - a_i^T w^{(j)} \quad (i \in \{1, \dots, m\} \backslash (I \cup \{j\})\}. \end{aligned} \tag{27}$$

Proposition 6.1. *Let z be such that the system $Ax \leq z$ is nondegenerate and assume in addition that $\{j\} \in \mathcal{I}(A, z)$. Then, for arbitrarily fixed index $j \in \{1, \dots, m\}$, the following identity holds true:*

$$\{I \backslash \{j\} | I \in \mathcal{I}(A, z), \, j \in I\} = \mathcal{I}^{(j)}.$$

Proof. Let j be arbitrarily fixed. To prove the inclusion '$\subseteq$', let $I \in \mathcal{I}(A, z)$ with $j \in I$ be arbitrary. We have to show that $I \backslash \{j\} \in \mathcal{I}^{(j)}$. By definition of $\mathcal{I}(A, z)$, there exists some $\bar{x}$ such that

$$a_i^T \bar{x} = z_i \quad (i \in I), \quad a_i^T \bar{x} < z_i \quad (i \in \{1, \dots, m\} \backslash I). \tag{28}$$

Referring to Theorem 3.2 and to Lemma 3.1, $\Sigma(I, j) = A^{I \backslash \{j\}} S^{(j)} \left(A^{I \backslash \{j\}}\right)^T$ is seen to be positive definite, hence, it follows from $S^{(j)} = L^{(j)} L^{(j)^T}$ that $A^{I \backslash \{j\}} L^{(j)}$ is a matrix of order $(\#I - 1, s - 1)$ whose rows are linearly independent. Recall that $\#I \leq s$ as a consequence of the assumed nondegeneracy of the system $Ax \leq z$. Therefore, there exists a matrix B of order $(s - \#I, s - 1)$ such that the completion

$$\begin{pmatrix} A^{I \backslash \{j\}} L^{(j)} \\ B \end{pmatrix}$$

is of order $(s-1, s-1)$ and invertible. Moreover, since rank $L^{(j)} = s - 1$ by assertion (*iii*) of Lemma 3.1, $L^{(j)^T} L^{(j)}$ is of order $(s-1, s-1)$ and invertible too. Therefore, the matrix $CL^{(j)}$ with

$$C := \begin{pmatrix} A^{I \backslash \{j\}} \\ B \left(L^{(j)^T} L^{(j)}\right)^{-1} L^{(j)^T} \end{pmatrix}$$

is invertible. Now, with $w^{(j)}$ from Lemma 3.1, define

$$\bar{y} := \left(CL^{(j)}\right)^{-1} C \left(\bar{x} - w^{(j)}\right). \tag{29}$$

Fix an arbitrary $k \in \{1, \dots, m\}$ and put

$$u := \left(\left(CL^{(j)}\right)^{-1}\right)^T L^{(j)^T} a_k.$$

Then, $L^{(j)^T} C^T u = L^{(j)^T} a_k$ and it follows from assertion (*iv*) of Lemma 3.1 that

$$C^T u - a_k = \lambda a_j$$

for some $\lambda \in \mathbb{R}$. Therefore, (29) entails that

$$a_k^T L^{(j)} \bar{y} = u^T C \left(\bar{x} - w^{(j)}\right) = (a_k + \lambda a_j)^T \left(\bar{x} - w^{(j)}\right).$$

Now, since $j \in I$, the first relation of (28) shows that $a_j^T \bar{x} = z_j$. Exploiting also assertion (*v*) of Lemma 3.1, we may continue as

$$a_k^T L^{(j)} \bar{y} = a_k^T \left(\bar{x} - w^{(j)}\right).$$

Since $k \in \{1, \dots, m\}$ was arbitrary, (28) yields that

$$a_k^T L^{(j)} \bar{y} = z_k - a_k^T w^{(j)} \quad (k \in I), \quad a_k^T L^{(j)} \bar{y} < z_k - a_k^T w^{(j)} \quad (k \in \{1, \dots, m\} \backslash I).$$

Now, the asserted relation $I \backslash \{j\} \in \mathcal{I}^{(j)}$ follows from (27) upon recalling that $j \in I$.

Conversely, let $\hat{I} \in \mathcal{I}^{(j)}$ be arbitrarily given. By definition, $\hat{I} \subseteq \{1, \ldots, m\} \backslash \{j\}$ and there exists some $y \in \mathbb{R}^{s-1}$ such that

$$a_i^T L^{(j)} y = z_i - a_i^T w^{(j)} \quad (i \in \hat{I}), \quad a_i^T L^{(j)} y < z_i - a_i^T w^{(j)} \quad (i \in \{1, \ldots, m\} \backslash (\hat{I} \cup \{j\}).$$

Putting $\bar{x} := L^{(j)} y + w^{(j)}$, this yields that

$$a_i^T \bar{x} = z_i \quad \left(i \in \hat{I}\right), \quad a_i^T \bar{x} < z_i \quad \left(i \in \{1, \ldots, m\} \backslash (\hat{I} \cup \{j\})\right). \tag{30}$$

Furthermore, from assertions (*iv*) and (*v*) of Lemma 3.1, it follows that

$$a_j^T \bar{x} = a_j^T \left(L^{(j)} y + w^{(j)}\right) = z_j. \tag{31}$$

By definition (5) of $\mathcal{I}(A, z)$, (30) and (31) provide that $I := \hat{I} \cup \{j\} \in \mathcal{I}(A, z)$. Since $j \notin \hat{I}$, it follows that $\hat{I} = I \backslash \{j\}$. Consequently, $\hat{I}$ belongs to the set

$$\{I \backslash \{j\} | I \in \mathcal{I}(A, z), \, j \in I\}$$

as was to be shown. □

Lemma 6.1. *Let $z \in \mathbb{R}^m$ be such that the system $Ax \leq z$ is nondegenerate. Then for every $I \in \mathcal{I}(A, z)$ and every $J \subseteq I$ one has that $J \in \mathcal{I}(A, z)$.*

Proof. Let $I \in \mathcal{I}(A, z)$ and $J \subseteq I$ be arbitrary. By definition, there is some $x \in \mathbb{R}^s$ such that

$$a_i^T x = z_i \quad (i \in I), \quad a_i^T x < z_i \quad (i \in \{1, \ldots, m\} \backslash I).$$

By the nondegeneracy assumption, $\operatorname{rank} \{a_i\}_{i \in I} = \# I$. Therefore, there exists a solution $\bar{h}$ to the linear equations

$$a_i^T h = 0 \quad (i \in J), \quad a_i^T h = -1 \quad (i \in I \backslash J).$$

Then, for $\bar{x} := x + t\bar{h}$ with $t > 0$ small enough, one has that

$$a_i^T \bar{x} = z_i \quad (i \in J), \quad a_i^T \bar{x} < z_i \quad (i \in \{1, \ldots, m\} \backslash J).$$

This entails $J \in \mathcal{I}(A, z)$. □

Corollary 6.1. *Under the assumptions of Lemma 6.1, if $Ax \leq z$ has a solution at all, then $\emptyset \in \mathcal{I}(A, z)$ and $a_k \neq 0$ for all k with $k \in I$ for some $I \in \mathcal{I}(A, z)$.*

Proof. Let $\bar{x}$ be a solution of $Ax \leq z$ and put

$$I := \{i \in \{1, \ldots, m\} | a_i^T \bar{x} = z_i\}.$$

Then, $I \in \mathcal{I}(A, z)$, whence $\emptyset \in \mathcal{I}(A, z)$ by Lemma 6.1. Now, let $k \in I$ for some $I \in \mathcal{I}(A, z)$. Then, the same argument shows that $\{k\} \in \mathcal{I}(A, z)$, whence rank $\{a_k\} = 1$ by the nondegeneracy assumption. In other words, $a_k \neq 0$. □

References

[1] W. v. Ackooij, R. Henrion, A. Möller and R. Zorgati, *On probabilistic constraints induced by rectangular sets and multivariate normal distributions*, Math. Meth. Oper. Res. **71** (2010), 535-549.

[2] I. Deák, *Computing probabilities of rectangles in case of multinormal distribution*, J. Stat. Comput. Simul. **26** (1986), 101-114.

[3] D. Dentcheva, B. Lai and A. Ruszczyński, *Dual methods for probabilistic optimization problems*, Math. Meth. Oper. Res. **60** (2004), 331-346.

[4] K. Fukuda, *cdd and cddplus Homepage* at:
http://www.ifor.math.ethz.ch/~fukuda/cdd_home/cdd.html

[5] J. Garnier, A. Omrane and Y. Rouchdy, *Asymptotic formulas for the derivatives of probability functions and their Monte Carlo estimations*, European J. Oper. Res. **198** (2009), 848-858.

[6] A. Genz and K.S. Kwong, *Numerical evaluation of singular multivariate normal distributions*, J. Stat. Comput. Simul. **68** (2000), 1-21.

[7] A. Genz and F. Bretz, *Computation of Multivariate Normal and t Probabilities*, Lecture Notes in Statistics, vol. 195, Springer, Heidelberg, 2009.

[8] A.A. Gouda and T. Szántai, *On numerical calculation of probabilities according to Dirichlet distribution*, Ann. Oper. Res. **177** (2010), 185-200.

[9] R. Henrion and W. Römisch, *Lipschitz and differentiability properties of quasi-concave and singular normal distribution functions*, Ann. Oper. Res. **177** (2010), 115-125.

[10] A.I. Kibzun and S. Uryas'ev, *Differentiability of Probability function*, Stoch. Anal. Appl. **16** (1998), 1101-1128.

[11] K. Marti, *Differentiation of probability functions: The transformation method*, Computers Math. Applic. **30** (1995), 361-382.

[12] D.Q. Naiman and H.P. Wynn, *Abstract tubes, improved inclusion-exclusion identities and inequalities and importance sampling*, Ann. Statist. **25** (1997), 1954-1983.

[13] N. J. Olieman and B. van Putten, *Estimation method of multivariate exponential probabilities based on a simplex coordinates transform*, J. Stat. Comput. Simul. **80** (2010), 355–361.

[14] G. Pflug and H. Weisshaupt, *Probability gradient estimation by set-valued calculus and applications in network design*, SIAM J. Optim. **15** (2005), 898-914.

[15] A. Prékopa, *Stochastic Programming*, Kluwer, Dordrecht, 1995.

[16] A. Prékopa, *Probabilistic Programming*, Stochastic Programming (A. Ruszczyński and A. Shapiro, eds.), Handbooks in Operations Research and Management Science, Vol. 10, Elsevier, Amsterdam, 2003.

[17] A. Shapiro, D. Dentcheva and A. Ruszczyński, *Lectures on Stochastic Programming*, MPS-SIAM series on optimization **9**, 2009.

[18] T. Szántai, *Evaluation of a special multivariate gamma distribution function*, Mathematical Programming Studies **27** (1986), 1-16.

[19] T. Szántai, *Improved bounds and simulation procedures on the value of the multivariate normal probability distribution function*, Ann. Oper. Res. **100** (2001), 85-101.

[20] S. Uryas'ev, *Derivatives of probability functions and integrals over sets given by inequalities*, J. Comp. Appl. Math. **56** (1995), 197-223.