Ann Oper Res (2010) 177: 115–125
DOI 10.1007/s10479-009-0598-0

# Lipschitz and differentiability properties of quasi-concave and singular normal distribution functions

**René Henrion · Werner Römisch**

Published online: 25 August 2009


**Abstract** The paper provides a condition for differentiability as well as an equivalent criterion for Lipschitz continuity of singular normal distributions. Such distributions are of interest, for instance, in stochastic optimization problems with probabilistic constraints, where a comparatively small (nondegenerate-) normally distributed random vector induces a large number of linear inequality constraints (e.g. networks with stochastic demands). The criterion for Lipschitz continuity is established for the class of quasi-concave distributions which the singular normal distribution belongs to.

**Keywords** Quasi-concave measures · Singular normal distributions · Lipschitz continuity · Differentiability · Stochastic optimization · Probabilistic constraints

## 1 Introduction

An $m$-dimensional random vector $\eta$ is said to have a singular normal distribution if there exists some $s$-dimensional random vector $\xi$ having a nondegenerate normal distribution such that

$$\eta = A\xi + b,$$

where $A$ is an $(m, s)$-matrix with rank smaller than $m$ and $b$ is an $m$-vector. In particular, one may choose $A = 0$ to see that the Dirac measure, placing mass one at the point $b$, has a singular normal distribution. More generally, singular normal distributions are those normal distributions whose covariance matrix has a rank strictly smaller than the dimension of the random vector.

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

R. Henrion (✉)
Weierstrass Institute Berlin, 10117 Berlin, Germany
e-mail: henrion@wias-berlin.de

W. Römisch
Institute of Mathematics, Humboldt-University Berlin, 10099 Berlin, Germany

Such seemingly artificial distributions arise in a natural way in problems of stochastic optimization, where a relatively small (nondegenerate-) normally distributed random vector induces a large number of linear inequality constraints. As an example, consider the problem of optimal capacity expansion in a network with stochastic demands (see Prékopa 1995, p. 453). Let the random vector $\xi$ represent the demands in the nodes of the network and let $x$ be a vector of capacities for the arcs in the network. The costs of installing these capacities are to be minimized as a function of $x$ under the constraint that there exists a flow through the network which is feasible at high probability, i.e., which satisfies both the capacity restrictions along the arcs and the random demands in the nodes (at high probability). Using the *Gale-Hoffman* theorem, feasibility can be modeled as a linear relation

$$A\xi \leq Bx.$$

Taking into account the random character of $\xi$, it makes sense to require feasibility in a probabilistic sense:

$$P\,(A\xi \leq Bx) \geq p,$$

where $P$ denotes probability and $p \in [0, 1]$ is some chosen level of reliability. In general, the sizes of $A$ and $B$ can be drastically reduced by eliminating redundancy etc. Nevertheless, even the reduced systems may contain a number of inequalities which is considerably larger than the dimension of $\xi$ (number of nodes). Passing to the transformed random vector $\eta = A\xi$, the probabilistic constraint obtained above can be rewritten as

$$\Phi(Bx) \geq p,$$

where $\Phi$ is the distribution function of $\eta$. However, since $A$ may have more rows than columns, we have to expect that $\eta$ has a singular normal distribution even though $\xi$ had a regular normal distribution.

The example shows that, in order to cope with certain types of probabilistic constraints, it is important to be able to calculate values and gradients of singular normal distribution functions. As the latter need not exist in general, it is of interest to characterize differentiability of such functions. If differentiability fails to hold, one could rely on more general tools from nonsmooth optimization (both for algorithmic purposes and optimality conditions). In such constellation, local or global Lipschitz continuity is a favorable property. Whether a singular normal distribution function is discontinuous or not does not depend on the rank of the covariance matrix. Figure 1 shows (from the left to the right) the distribution functions of 2-dimensional normal distributions with zero mean and covariance matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

all of which have rank one. Note that, in the first case, the distribution function is discontinuous whereas it is Lipschitz continuous (piecewise selection of smooth functions of min- and max-type, respectively) in the remaining cases.

The paper provides a condition for differentiability as well as an equivalent criterion for Lipschitz continuity of singular normal distribution functions. The criterion for Lipschitz continuity can be obtained for the general class of quasi-concave distributions which singular normal distributions belong to.

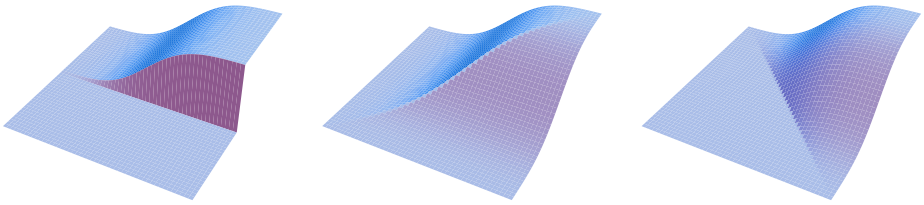

**Fig. 1** Distribution functions of 2-dimensional singular normal distributions with covariance matrix having rank one (see text)

## 2 Lipschitz continuity of quasi-concave distributions

We start this section by introducing the class of quasi-concave probability measures (see Prékopa 1995). By $\mathcal{P}(\mathbb{R}^s)$ we denote the set of probability measures on $\mathbb{R}^s$.

**Definition 2.1** A probability measure $\mu \in \mathcal{P}(\mathbb{R}^s)$ is called quasi-concave whenever

$$\mu(\lambda A + (1-\lambda)B) \geq \min\{\mu(A), \mu(B)\}$$

holds true for all convex and Borel measurable subsets $A, B \subseteq \mathbb{R}^s$ and all $\lambda \in [0,1]$ such that $\lambda A + (1-\lambda)B$ is Borel measurable.

It is well known that a large class of prominent multivariate distributions shares the property of being quasi-concave. Among those are the multivariate normal distribution (nondegenerate or singular), the Dirichlet-, Pareto-, Gamma-, Log-normal distributions (possibly with a restricted range of parameters) as well as uniform distributions over compact, convex subsets of $\mathbb{R}^s$ (see Prékopa 1995; Borell 1975). Consequently, all future statements in this section apply in particular to singular normal distributions.

For the proof of our Lipschitz criterion, we shall make use of the following three propositions:

**Proposition 2.1** *A quasiconcave measure $\mu \in \mathcal{P}(\mathbb{R})$ has either a density or coincides with some Dirac measure, i.e. $\mu = \delta_x$ for some $x \in \mathbb{R}$.*

*Proof* Follows immediately from Theorem 3.2 in Borell (1975). □

**Proposition 2.2** *If for all marginal distributions $\mu_i$ of $\mu \in \mathcal{P}(\mathbb{R}^s)$ there exist bounded densities on $\mathbb{R}$, then the distribution function $F_\mu$ of $\mu$ is Lipschitz continuous.*

*Proof* See Proposition 3.8 in Römisch and Schultz (1993). □

**Proposition 2.3** *If $\mu \in \mathcal{P}(\mathbb{R})$ is a quasiconcave measure with density $f_\mu$, then $f_\mu$ is bounded.*

*Proof* According to Theorem 3.2 in Borell (1975), the possibly extended-valued function $1/f_\mu$ is convex and the support of $\mu$ is a convex subset of $\mathbb{R}$. Assuming that $f_\mu$ is unbounded, there exists a sequence $\{x_n\} \subseteq \mathbb{R}$ such that $f_\mu(x_n) \geq n$. If $\{x_n\}$ is unbounded, then, without loss of generality, it is increasing, hence $[x_1, \infty) \subseteq \operatorname{supp}\mu$ and $\{1/f_\mu(x_n)\}$ is decreasing. Since $1/f_\mu$ is convex, it follows that $1/f_\mu$ is decreasing on $[x_1, \infty)$. Therefore,

$f_\mu$ is increasing on $[x_1, \infty)$ which contradicts the fact that $f_\mu$ is a density. Now, assume that $\{x_n\}$ is bounded, hence $x_n \to \bar{x}$ upon passing to some subsequence. Then, $1/f_\mu(\bar{x}) = 0$. Indeed, this follows in case of $\bar{x} \in \operatorname{int} \operatorname{supp} \mu$ from the continuity of the convex function $1/f_\mu$ on the interior of its domain. In case that $\bar{x}$ belongs to the boundary of $\operatorname{supp} \mu$, we may redefine $f_\mu(\bar{x}) := \infty$ without changing the measure $\mu$ and without affecting the convexity of $1/f_\mu$ (due to $1/f_\mu(x_n) \to 0$). Now, from $1/f_\mu \geq 0$ being convex and satisfying $1/f_\mu(\bar{x}) = 0$, it follows that $1/f_\mu(\bar{x} + h)$ is nondecreasing for $h > 0$ and that the difference quotients

$$h \mapsto h^{-1}(1/f_\mu(\bar{x} + h) - 1/f_\mu(\bar{x}))$$

are nondecreasing in $h$. Consequently, one has for $h_2 \geq h_1 > 0$

$$f_\mu(\bar{x} + h_1) \geq f_\mu(\bar{x} + h_2), \tag{1}$$

$$f_\mu(\bar{x} + h_1)h_1 \geq f_\mu(\bar{x} + h_2)h_2. \tag{2}$$

We assume that either $\bar{x} \in \operatorname{int} \operatorname{supp} \mu$ or that $\bar{x}$ belongs to the left boundary of $\operatorname{supp} \mu$ (the proof running analogously in case that $\bar{x}$ belongs to the right boundary of $\operatorname{supp} \mu$). In both cases there exists some $\delta > 0$ such that $f_\mu(\bar{x} + \delta) > 0$. It follows for arbitrary $n \in \mathbb{N}$ that

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f_\mu(x)dx \geq \int_{\bar{x}+2^{-n}\delta}^{\bar{x}+\delta} f_\mu(x)dx = \sum_{j=0}^{n-1} \int_{\bar{x}+2^{-(j+1)}\delta}^{\bar{x}+2^{-j}\delta} f_\mu(x)dx \\
&\geq \sum_{j=0}^{n-1} f_\mu(\bar{x} + 2^{-j}\delta)2^{-j}\frac{\delta}{2} \quad \text{(by (1))} \\
&\geq \sum_{j=0}^{n-1} f_\mu(\bar{x} + \delta)\frac{\delta}{2} \quad \text{(by (2))} \\
&= n\frac{\delta}{2} f_\mu(\bar{x} + \delta).
\end{aligned}$$

This, however, is a contradiction to

$$n\frac{\delta}{2} f_\mu(\bar{x} + \delta) \to_n \infty. \qquad \square$$

For the narrower class of log-concave measures, Proposition 2.3 is a (1-dimensional) special case of a Theorem by Barndorff-Nielsen (1978).

**Definition 2.2** We call a subset $H \subseteq \mathbb{R}^s$ a canonical hyperplane if there exist $t \in \mathbb{R}$ and $i \in \{1, \ldots, s\}$ such that

$$H = \overset{1}{\mathbb{R}} \times \cdots \times \overset{i-1}{\mathbb{R}} \times \{t\} \times \overset{i+1}{\mathbb{R}} \times \cdots \times \overset{s}{\mathbb{R}}.$$

Now, we are in a position to formulate the desired criterion for Lipschitz continuity of distribution functions in the considered class of distributions:

**Theorem 2.1** *A quasiconcave probability measure $\mu \in \mathcal{P}(\mathbb{R}^s)$ has a Lipschitz continuous distribution function $F_\mu$ if and only if the support of $\mu$ is not contained in a canonical hyperplane of $\mathbb{R}^s$.*

*Proof* We denote by $\mu_i \in \mathcal{P}(\mathbb{R})$ the $i$th marginal distribution of $\mu$. Clearly, the $\mu_i$ are quasi-concave on $\mathbb{R}$. With $T$ being the support of $\mu$ and $\delta_t$ referring to the one-dimensional Dirac measure placed at $t \in \mathbb{R}$, the following chain of equivalences results:

$$\begin{aligned}
&T \text{ is contained in a canonical hyperplane of } \mathbb{R}^s \\
\iff\ & \exists t \in \mathbb{R}\ \exists\, i \in \{1,\dots,s\} : \mu(\overset{1}{\mathbb{R}} \times \cdots \times \overset{i-1}{\mathbb{R}} \times \overset{i}{\{t\}} \times \overset{i+1}{\mathbb{R}} \times \cdots \times \overset{s}{\mathbb{R}}) = 1 \\
\iff\ & \exists t \in \mathbb{R}\ \exists\, i \in \{1,\dots,s\} : \mu_i(\{t\}) = 1 \\
\iff\ & \exists t \in \mathbb{R}\ \exists\, i \in \{1,\dots,s\} : \mu_i = \delta_t \\
\iff\ & \exists\, i \in \{1,\dots,s\} : \mu_i \text{ doesn't have a density.}
\end{aligned}$$

Here, the last equivalence is implied by Proposition 2.1. Contraposition gives the following chain of implications with the second and third one following from Propositions 2.3 and 2.2, respectively.

$$\begin{aligned}
&T \text{ is not contained in any canonical hyperplane of } \mathbb{R}^s \\
\implies\ & \mu_i \text{ has a density } f_{\mu_i} \text{ for all } i \in \{1,\dots,s\} \\
\implies\ & f_{\mu_i} \text{ is bounded for all } i \in \{1,\dots,s\} \\
\implies\ & F_\mu \text{ is globally Lipschitzian.}
\end{aligned}$$

Now, this chain of implications proves the 'if'-part of the theorem. For the reverse direction, assume that $T$ is contained in a canonical hyperplane of $\mathbb{R}^s$. Then, the above chain of equivalences shows that

$$\mu(\overset{1}{\mathbb{R}} \times \cdots \times \overset{i-1}{\mathbb{R}} \times \overset{i}{\{t\}} \times \overset{i+1}{\mathbb{R}} \times \cdots \times \overset{s}{\mathbb{R}}) = 1 \quad \text{for some } t \in \mathbb{R} \text{ and } i \in \{1,\dots,s\}.$$

Consequently, one may choose some $\tau \in \mathbb{R}$ large enough such that

$$F_\mu(\tau,\dots,\tau,t,\tau,\dots,\tau) > 0.$$

On the other hand, $F_\mu(\tau,\dots,\tau,t',\tau,\dots,\tau) = 0$ for any $t' < t$, hence $F_\mu$ is not continuous (much less it is Lipschitz continuous). □

The last argument in the proof of Theorem 2.1 shows that the failure of Lipschitz continuity entails the failure of continuity, so we get the following useful observation:

**Corollary 2.1** *The distribution function of some quasiconcave probability measure is Lipschitz continuous if and only if it is continuous. In particular, the distribution function of some quasiconcave probability measure with density is Lipschitz continuous.*

Concerning the second statement of the last corollary, we emphasize that in general even the existence of a bounded and continuous density does not imply the Lipschitz continuity of the distribution function (for a counterexample see Henrion and Römisch 1999, Ex. 9). A slightly more illustrative reformulation of Theorem 2.1 is:

**Theorem 2.2** *Let $\xi$ be an $s$-dimensional random vector with quasi-concave distribution $\mu \in \mathcal{P}(\mathbb{R}^s)$. Then, the distribution function of $\xi$ is Lipschitz continuous if and only if none of the components $\xi_i$ has zero variance.*

As an application of Theorem 2.2 we come back to the singular normal distributions with the three covariance matrices mentioned in the introduction. The first covariance matrix contains a zero diagonal element whereas the second and third ones do not. This explains why the first distribution function depicted in Fig. 1 is discontinuous whereas the second and third ones are Lipschitz continuous.

At the end of this section, we consider an application to probability functions

$$\varphi(x) = P(A\xi \leq h(x)), \tag{3}$$

where, $\xi$ is an $s$-dimensional random vector, $A$ is an $(m, s)$-matrix, $x \in \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^m$. Recall that such type of probability functions arises in the context of probabilistic constraints $\varphi(x) \geq p$ as presented in the introduction.

**Corollary 2.2** *In* (3), *assume that $h$ is locally Lipschitzian and that $\xi$ has a quasi-concave distribution with some covariance matrix $\Sigma$. Then, $\varphi$ is locally Lipschitzian under the condition*

$$a_i \notin \operatorname{Ker} \Sigma, \quad \forall i \in \{1, \ldots, m\}, \tag{4}$$

*where the $a_i$ denote the rows of $A$.*

*Proof* The transformed random vector $\eta := A\xi$ inherits a quasi-concave distribution from that of $\xi$. With $F_\eta$ being the distribution function of $\eta$, one may write $\varphi = F_\eta \circ h$. The $i$th component of $\eta$ has variance $a_i^T \Sigma a_i$. Since this variance is larger than zero according to (4), Theorem 2.2 provides that $F_\eta$ is Lipschitz continuous. Hence, $\varphi$ is locally Lipschitzian as a composition of two such mappings. □

## 3 Differentiability of singular normal distribution functions

Although the 3 examples of singular normal distribution functions presented in the introduction and depicted in Fig. 1 fail to be differentiable in a global sense, they are differentiable almost everywhere. In order to establish a condition for differentiability, we shall introduce some concepts related with systems of linear inequalities. More precisely, let $A$ be an $(m, s)$-matrix and $b \in \mathbb{R}^m$. We shall briefly speak of the system $(A, b)$ to refer to the system $Az \leq b$ of linear inequalities in $\mathbb{R}^s$. For an index set $I \subseteq \{1, \ldots, m\}$, we shall denote by $A^I$ the submatrix of $A$ which is built up from those rows of $A$ which are indexed by $I$. Accordingly, $b^I$ will be the subvector of $b$ consisting of the components indexed by $I$. Furthermore, we shall use the short-hand notation '$u < v$' for vectors $u, v$ to mean a strict inequality for all their components.

With the system $(A, b)$ we associate a family of index sets defined by

$$I(A, b) := \{I \subseteq \{1, \ldots, m\} | \exists z \in \mathbb{R}^s : A^I z = b^I, A^{\{1,\ldots,m\}\setminus I} z < b^{\{1,\ldots,m\}\setminus I}\}.$$

The system $(A, b)$ is said to be *nondegenerate*, if $\operatorname{rank} A^I = \#I$ for all $I \in I(A, b)$. In the language of optimization theory, the system $(A, b)$ is nondegenerate if and only if it satisfies the *Linear Independence Constraint Qualification* (LICQ). We shall need a rather obvious result of technical nature:

**Proposition 3.1** *Suppose that the system $(A, b)$ is nondegenerate. Then, there exists a neighborhood $U$ of $b$ such that for all $b' \in U$ the systems $(A, b')$ are nondegenerate too and $I(A, b') = I(A, b)$.*

*Proof* According to the definition of nondegeneracy, the first assertion is an immediate consequence of the second one. We show first that there is a neighborhood $U$ of $b$ such that $I(A,b) \subseteq I(A,b')$ for all $b' \in U$. Let $I \in I(A,b)$ be arbitrary. By definition, there is some $z \in \mathbb{R}^s$ with $A^I z = b^I$ and $A^{\{1,\dots,m\}\setminus I} z < b^{\{1,\dots,m\}\setminus I}$. Let $U, V$ be neighborhoods of $b$ and $z$, respectively, such that $A^{\{1,\dots,m\}\setminus I} z' < (b')^{\{1,\dots,m\}\setminus I}$ for all $z' \in V$ and $b' \in U$. Due to the nondegeneracy of the system $(A,b)$, $A^I$ has full rank. Hence, choosing $U$ small enough, for all $b' \in U$ there are $z' \in V$ with $A^I z' = (b')^I$. Consequently, for all $b' \in U$, there exists some $z'$ satisfying $A^I z' = (b')^I$ and $A^{\{1,\dots,m\}\setminus I} z' < (b')^{\{1,\dots,m\}\setminus I}$. This amounts to $I \in I(A,b')$, whence the desired inclusion. Now, we show that there is a neighborhood $U$ of $b$ such that

$$I(A,b') \subseteq I(A,b), \quad \forall b' \in U. \tag{5}$$

Choosing the intersection of this neighborhood $U$ with the one found above for the reverse inclusion will prove the assertion of the proposition. It is well-known (see, e.g., Bank et al. 1982, Theorem 3.4.1) that the multifunction $M$ which assigns to each $b'$ the solution of the system $(A,b')$, can be decomposed as $M(b') = K(b') + U$, where $K$ is a Hausdorff-continuous multifunction such that the $K(b')$ are convex, compact polyhedra for all $b'$, and where $U = \{u \mid Au \leq 0\}$. Now, negating (5) and using a subsequence argument, one would derive the existence of sequences $x_k$ and $b^{(k)} \to b$ as well as of an index set $I \subseteq \{1,\dots,m\}$ with $I \notin I(A,b)$ such that $A^I x_k = (b^{(k)})^I$ and $A^{\{1,\dots,m\}\setminus I} x_k < (b^{(k)})^{\{1,\dots,m\}\setminus I}$. Clearly, $x_k \in M(b^{(k)})$, hence there are sequences $y_k \in K(b^{(k)})$ and $u_k \in U$ with $y_k = x_k - u_k$. By the Hausdorff continuity of $K$ and the compactness of $K(b)$ it follows that $y_k$ is bounded. Therefore, without loss of generality, we may assume that $y_k \to \bar{y}$ for some $\bar{y} \in K(b)$ (again by Hausdorff continuity of $K$). Consequently,

$$A^I \bar{y} = \lim_k \left(A^I x_k - A^I u_k\right) \geq \lim_k \left(b^{(k)}\right)^I = b^I.$$

On the other hand, since $0 \in U$, we know that $\bar{y} \in M(b)$, whence $A^I \bar{y} \leq b^I$. Summarizing, $A^I \bar{y} = b^I$. Since $\bar{y}$ solves the system $(A,b)$, there is some index set $I' \supseteq I$ such that $A^{I'} \bar{y} = b^{I'}$ and $A^{\{1,\dots,m\}\setminus I'} \bar{y} < b^{\{1,\dots,m\}\setminus I'}$. In other words, $I' \in I(A,b)$. Invoking once more the nondegeneracy of the system $(A,b)$, we see that $A^{I'}$ has full rank. As a consequence, there exists some $h$ such that $A^I h = 0$ and $A^{I'\setminus I} h = -\mathbf{1}$, where $\mathbf{1} := (1,\dots,1)$. Now, for small enough $t > 0$, one gets that

$$\begin{aligned} A^I (\bar{y} + th) &= b^I, \\ A^{I'\setminus I} (\bar{y} + th) &= b^{I'\setminus I} - t\mathbf{1} < b^{I'\setminus I}, \\ A^{\{1,\dots,m\}\setminus I'} (\bar{y} + th) &< b^{\{1,\dots,m\}\setminus I'}. \end{aligned}$$

This amounts to the contradiction $I \in I(A,b)$. □

Our differentiability result will basically rely on the following inclusion-exclusion formula for the probability of polyhedra proved in Naiman and Wynn (1997) by means of the so-called abstract-tube theory (a recent proof based on more elementary arguments like duality of linear programming can be found in Bukszár et al. 2004):

**Theorem 3.1** *Let $\xi$ be an $s$-dimensional random vector. If the system $(A,b)$ is nondegenerate, then the probability of the polyhedron induced by $(A,b)$ equals*

$$P(A\xi \leq b) = \sum_{I \in I(A,b)} (-1)^{\#I} P\left(A^I \xi > b^I\right).$$

We note that the assumed nondegeneracy implies $\emptyset \in I(A,b)$. In this case, the corresponding term in the sum above is equal to one just by formal argumentation:

$$(-1)^{\#\emptyset} P(\langle a_i, \xi\rangle > b_i \ (i \in \emptyset)) = P(\mathbb{R}^s) = 1.$$

Recall from the introduction that a singular normal distribution can always be obtained as a linear transformation of some nondegenerate normal distribution. If this linear transformation is not explicitly given but just the covariance matrix $\Xi$ and the mean vector $\gamma$ of the singular normal distribution are known, this transformation can be found as follows: First decompose the (possibly degenerate) covariance matrix $\Xi$ as $\Xi = AA^T$ such that $A$ has full rank. Let $\xi$ be a random vector whose dimension coincides with the number of columns of $A$ and which has independent normally distributed components with zero mean. Then, the transformation $A\xi + \gamma$ generates a random vector with covariance matrix $AA^T = \Xi$ and mean $\gamma$, i.e., $A$ and $\gamma$ define the desired linear transformation. Now, we state the main result of this section.

**Theorem 3.2** *Let $\xi$ have an $s$-dimensional nondegenerate normal distribution. Denote by $\Phi_\eta$ the distribution function of the linearly transformed random vector $\eta = A\xi + b$, where $A$ is an $(m,s)$-matrix and $b \in \mathbb{R}^m$. Then, $\Phi_\eta$ is smooth (infinitely many times differentiable) at any point $\bar{x} \in \mathbb{R}^m$ for which the system $(A, \bar{x} - b)$ is nondegenerate.*

*Proof* By Proposition 3.1, there exists a neibhborhood $U$ of $\bar{x}$ such that the system $(A, x - b)$ is nondegenerate for all $x \in U$. By definition, one has that

$$\Phi_\eta(x) = P(\eta \le x) = P(A\xi \le x - b).$$

Application of Theorem 3.1 to the systems $(A, x - b)$ yields that, for all $x \in U$:

$$\Phi_\eta(x) = \sum_{I \in I(A, x-b)} (-1)^{\#I} P(A^I \xi > x^I - b^I).$$

We note that in the last relation, one may pass to a non-strict inequality. Indeed, since all the $A^I$ have full rank by nondegeneracy, the set of $\xi$ satisfying $A^I\xi \ge x^I - b^I$ but violating $A^I\xi > x^I - b^I$ has Lebesgue measure zero. Since $\xi$ has a density, passing to non-strict inequalities will not change the probability:

$$\Phi_\eta(x) = \sum_{I \in I(A, x-b)} (-1)^{\#I} P(A^I \xi \ge x^I - b^I).$$

For each $I \subseteq \{1, \ldots, m\}$, define random vectors $\eta^I := -A^I\xi$. Then, one has for all $x \in U$ that

$$\Phi_\eta(x) = \sum_{I \in I(A, x-b)} (-1)^{\#I} P\left(\eta^I \le b^I - x^I\right) = \sum_{I \in I(A, x-b)} (-1)^{\#I} F^I\left(b^I - x^I\right). \quad (6)$$

Here, $F^I$ refers to the distribution function of $\eta^I$. Obviously, $\eta^I$ has a normal distribution with covariance matrix $A^I \Sigma (A^I)^T$, where $\Sigma$ denotes the positive definite (by assumption) covariance matrix of $\xi$. Due to nondegeneracy of the systems $(A, x - b)$, we know that $A^I$ has full rank for all $I \in I(A, x - b)$ and all $x \in U$. Consequently, $A^I \Sigma (A^I)^T$ is positive definite too, which means that all the $\eta^I$ have nondegenerate normal distributions. Therefore,

all distribution functions $F^I$ are (globally) smooth. We are tempted now, to differentiate the sum in (6) all terms of which are differentiable. This would imply the desired smoothness of $\Phi_\eta$ at $\bar{x}$. However, care has to be taken since the number of terms in the sum, which is given by the cardinality of $I(A, x-b)$, does formally depend on $x$. Hence, certain terms could suddenly disappear or appear, when moving away from $\bar{x}$. Fortunately, we know from Proposition 3.1 that $I(A, x-b) = I(A, \bar{x}-b)$ for all $x \in U$. This allows to write $\Phi_\eta$ locally around $\bar{x}$ as a sum of a *fixed* number of smooth functions:

$$\Phi_\eta(x) = \sum_{I \in I(A,\bar{x}-b)} (-1)^{\#I} F^I\left(b^I - x^I\right), \quad \forall x \in U. \tag{7}$$

This implies smoothness of $\Phi_\eta$ at $\bar{x}$. □

Note that Theorem 3.2 does not just make a theoretical statement on smoothness of singular normal distribution functions, but even provides a formula how to calculate their derivatives. Indeed, one may use (7) in order to calculate the gradient (or higher order derivatives) of $\Phi_\eta$ on the basis of the same objects for nondegenerate (!) normal distribution functions (the $F^I$). As first and higher order derivatives of nondegenerate normal distribution functions can be analytically reduced to functional values themselves (see, e.g., Prékopa 1995), everything boils down to the mere calculation of nondegenerate normal distribution functions. This can be carried out by several existing algorithms (see, e.g., Gassmann et al. 2002; Genz 1992; Szántai 2000).

We want to illustrate Theorem 3.2 by applying it to the singular normal distribution with zero mean vector and covariance matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = AA^T \quad \text{with } A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

(see second picture in Fig. 1). Such distribution is realized by a random vector $\eta = A\xi$, where $\xi$ has a one-dimensional standard normal distribution (compare remark in front of Theorem 3.2). We have to check, for which vectors $x \in \mathbb{R}^2$ the system $(A, x)$ is nondegenerate. Concerning the calculation of the index family $I(A, x)$, one has to distinguish three cases:

$$\begin{aligned} x_1 < x_2 \quad &\Longrightarrow \quad I(A, x) = \{\emptyset, \{1\}\}, \\ x_1 > x_2 \quad &\Longrightarrow \quad I(A, x) = \{\emptyset, \{2\}\}, \\ x_1 = x_2 \quad &\Longrightarrow \quad I(A, x) = \{\emptyset, \{1, 2\}\}. \end{aligned}$$

Obviously, nondegeneracy holds true in the first two cases since both 'rows' of $A$ (which reduce to real numbers here) are different from zero. Consequently, Theorem 3.2 guarantees differentiability of the distribution function of $\eta$ whenever $x_1 \neq x_2$ (this can be verified from Fig. 1). On the other hand, the two 'rows' of $A$ cannot be linearly independent, hence nondegeneracy is lost in case of $x_1 = x_2$. This harmonizes with the fact that the distribution function of $\eta$ is not differentiable on the bisectrix $x_1 = x_2$ (see Fig. 1). Now, using formula (7), we may also calculate the gradient of $\Phi_\eta$ at points $x$ where it exists, e.g., where $x_1 < x_2$. We obtain:

$$\Phi_\eta(x) = F^{\emptyset}(-x^{\emptyset}) - F^{\{1\}}(-x^{\{1\}}) = 1 - F^{\{1\}}(-x^{\{1\}}), \quad \forall x \in U,$$

where we used that the first probability term referring to the empty index set is equal to one by formal reasons (see remark below Theorem 3.1). Moreover, by definition, $F^{\{1\}}$ is the distribution function of $\eta^{\{1\}} = A^{\{1\}}\xi = \xi$, hence $F^{\{1\}}$coincides with the one-dimensional standard normal distribution function $\Phi$, whence

$$\Phi_\eta(x) = 1 - \Phi(-x_1) = \Phi(x_1), \quad \forall x \in U.$$

Derivation at $\bar{x}$ now yields $\nabla\Phi_\eta(\bar{x}) = (\Phi'(\bar{x}_1), 0)$. Similarly, for $x_1 > x_2$, one obtains that $\nabla\Phi_\eta(\bar{x}) = (0, \Phi'(\bar{x}_2))$.

Finally, we note, that the smoothness result of Theorem 3.2 allows to calculate derivatives of probability functions

$$\varphi(x) = P(A\xi \leq h(x))$$

as they occured in (3), with the additional assumption that $h$ be smooth (a particular instance is given by the case $h(x) = Bx$ considered in the introduction). More precisely, under the assumption that the system $(A, \bar{x})$ is nondegenerate, one arrives at

$$\nabla\varphi(\bar{x}) = \sum_{I \in I(A,\bar{x})} (-1)^{\#I+1} \nabla F^I\left((-h(\bar{x}))^I\right)(Dh(\bar{x}))^I.$$

Of course, for many practical applications, it would be interesting to derive analogous results in the case that not only the righ-hand side but also the matrix depends on the decision $x$, i.e.:

$$\varphi(x) = P(A(x)\xi \leq h(x)).$$

In this situation, by using a slight generalization of Proposition 3.1, which takes into account perturbations of $A$ as well, one could still derive the representation formula (7), but now the random vector $\eta$ would no longer be fixed but depend on $x$: $\eta(x) = A(x)\xi$. As a consequence, the nondegenerate multivariate normal distribution functions $F^I$ in (7) would also depend on $x$ in that their covariance matrices $A(x)^I(A^I(x))^T$ depend on $x$. Therefore, calculating the desired gradients of $F^I$ requires to compute the partial derivatives the $F^I$ with respect to the entries of the covariance matrix, which may be very difficult, although the differentiability result itself might hold true.

## References


Bank, B., Guddat, J., Klatte, D., Kummer, B., & Tammer, K. (1982). *Non-linear parametric optimization*. Berlin: Akademie-Verlag.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. Chichester: Wiley.

Borell, C. (1975). Convex sets in d-space. *Periodica Mathematica Hungarica*, *6*, 111–136.

Bukszár, J., Henrion, R., Hujter, M., & Szántai, T. (2004). *Polyhedral inclusion-exclusion*. Weierstrass Institute Berlin, Preprint No. 913.

Gassmann, H. I., Deák, I., & Szántai, T. (2002). Computing multivariate normal probabilities: A new look. *Journal of Computational and Graphical Statistics*, *11*, 920–949.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*, 141–149.

Henrion, R., & Römisch, W. (1999). Metric regularity and quantitative stability in stochastic programs with probabilistic constraints. *Mathematical Programming*, *84*, 55–88.

Naiman, D. Q., & Wynn, H. P. (1997). Abstract tubes, improved inclusion-exclusion identities and inequalities and importance sampling. *Annals of Statistics*, *25*, 1954–1983.

Prékopa, A. (1995). *Stochastic programming*. Dordrecht: Kluwer.

Römisch, W., & Schultz, R. (1993). Stability of solutions for stochastic programs with complete recourse. *Mathematics of Operations Research*, *18*, 590–609.

Szántai, T. (2000). Improved bounds and simulation procedures on the value of the multivariate normal probability distribution function. *Annals of Operations Research*, *100*, 85–101.