Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Continuous stochastic gradient and spherical radial decomposition

Daniela Bernhard[1], Holger Heitsch[2], René Henrion[2], Frauke Liers[1], Michael Stingl[1],

Andrian Uihlein[1], Viktor Zipf[1]

submitted: Dec 15, 2025

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Department of Data Science
Cauerstr. 11
91058 Erlangen
Germany
E-Mail: daniela.db.bernhard@fau.de
frauke.liers@fau.de
michael.stingl@fau.de
andrian.uihlein@fau.de
viktor.v.zipf@fau.de

[2] Weierstrass Institute
Anton-Wilhelm-Amo Str. 39
10117 Berlin
Germany
E-Mail: holger.heitsch@wias-berlin.de
rene.henrion@wias-berlin.de

No. 3245
Berlin 2025

2020 *Mathematics Subject Classification.* 90B15, 90C15, 90C26.

Key words and phrases. Chance constraints, continuous stochastic gradient, spheric-radial decomposition.

This work is supported by the German Research Foundation (DFG) within the projects B04 and B06 of CRC/Transregio 154 and by the FMJH Program Gaspard Monge in optimization and operations research including the support by EDF.

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

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Continuous stochastic gradient and spherical radial decomposition

Daniela Bernhard, Holger Heitsch, René Henrion, Frauke Liers, Michael Stingl, Andrian Uihlein, Viktor Zipf

Abstract

In this paper, a new method is presented for solving chance-constrained optimization problems. The method combines the well-established Spherical-Radial Decomposition approach with the Continuous Stochastic Gradient method. While the Continuous Stochastic Gradient method has been successfully applied to chance-constrained problems in the past, only the combination with the Spherical-Radial Decomposition allows to avoid smoothing of the integrand. In this chapter, we prove this fact for a relevant class of chance-constrained problems and apply the resulting method to the capacity maximization problem for gas networks.

1 Introduction

We consider optimization problems with a *probabilistic constraint* or *chance constraint* of the following form:

$$\begin{aligned} \min_{x} \quad & g_0(x), \\ \text{s.t.} \quad & \varphi(x) := \mathbb{P}_\zeta(g(x,\zeta) \leq 0) \geq p. \end{aligned} \tag{P}$$

Here, the decision x belongs to some finite- or infinite dimensional space, g_0 is an objective function to be minimized, g is a random constraint function, ζ is a random vector defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $p \in (0,1]$ is a probability level. The interpretation of the chance constraint is as follows: the decision x is defined to be feasible, whenever the random inequality $g(x,\zeta) \leq 0$ is satisfied with probability at least p. Formally, a chance constraint is a deterministic inequality constraint of the form $\varphi(x) \geq p$, where φ denotes the probability function assigning to each x the probability of satisfying the random inequality. The challenge posed by chance constraints relies on the fact that usually no explicit formula for φ is given. This is even more prevalent for gradients $\nabla\varphi$, if they exist at all. For a comprehensive classical introduction to chance constraints, we refer to the monograph [21] and to the more recent presentation in [27].

The concept of chance constraints to model stochastic uncertainties in the data was first introduced by [8]. As already mentioned, φ is often not given explicitly in the case of continuous distributions. Therefore, much research focuses on approximating the chance constraint with easier functions. One possibility is to replace the continuous distribution by a finite sample. Here, scenario approximations (e.g., [7], [17]) or sample approximation techniques (e.g., [1], [16], [19]) are often the methods of choice. The approximations result in classical mixed-integer nonlinear programming problems (MINLPs) that can be solved by available approaches from this field. Nevertheless, good discrete approximations of the original distribution often need too many samples to solve the resulting MINLPs efficiently in practice. So, there also exist various different approximations of the function φ dealing directly with the continuous distributions. Common methods are for example the Bernstein approximation [18], conditional Value-at-Risk approximations [24] or robust approximations (e.g., [2], [3], [29]).

DOI 10.20347/WIAS.PREPRINT.3245 Berlin 2025

Concentrating on special kinds of distributions, there also exist results to directly deal with the chance-constrained function φ. For example, in the case of elliptically symmetric (e.g., Gaussian) random vectors, a successful workhorse for the numerical solution of optimization problems (P) is the so-called *Spherical-Radial Decomposition* (SRD), which provides strong regularity guarantees for φ. Nevertheless, evaluating the probability function with respect to a multivariate Gaussian distribution is still necessary. So far, quasi Monte-Carlo methods are often used for this task. In this work, we propose a different approach, combining the *Continuous Stochastic Gradient* (CSG) method [20] with SRD.

To avoid a numerically expensive (quasi) Monte Carlo integration for the probability function in each iteration, CSG follows ideas from traditional stochastic sample-based optimization like the *Stochastic Gradient* (SG) method [23] or similar related approaches [6]. Thus, in each iteration, rather than being fully evaluated, the probability function is sampled for a small number of realizations of the random variable (mini-batch). In contrast to most other stochastic sample-based optimization schemes, these gradient samples are not discarded after the iteration but instead stored to accumulate information during the optimization process. By adaptively recombining old and new gradient information, CSG constructs a gradient approximation with vanishing approximation error [11, 20], resulting in an overall reduction in the required number of gradient evaluations.

To apply the CSG method to the chance-constrained problem, we penalize constraint violations in the objective function. We note that the article [4] already uses the CSG method to solve chance-constrained problems. However, convergence results for CSG are restricted to probability functions with strong regularity assumptions [11]. To be precise, in [11], φ is assumed to be of the form

$$\varphi(x) = \mathbb{E}_\zeta\big[\phi(x,\zeta)\big],$$

with globally Lipschitz $\nabla_x\phi$. Thus, an additional smoothing approximation to obtain the necessary regularity of the underlying functions is carried out in [4]. By combining SRD and CSG, we circumvent this smoothing operation in a nontrivial fashion. Nonetheless, the SRD reformulation of (P) is still not sufficiently regular to apply known convergence results of CSG. Therefore, we present an extension of CSG convergence theory that covers our numerical examples and more general classes of objectives.

The efficiency of the proposed method is numerically demonstrated for small analytical problems as well as for the capacity maximization problem in gas networks.

2 Basic components

2.1 Spherical-radial decomposition

The concept of *Spherical-Radial Decomposition* (SRD) relates to random vectors having an elliptically-symmetric distribution (e.g., Gaussian, Student, Laplace, etc.). Such distributions can be decomposed into a uniform distribution over the unit sphere and some one-dimensional radial distribution (e.g., Chi-distribution for the Gaussian case). For a basic monograph on this topic, we refer to [9]. Exploiting this property for elliptically-symmetric distributions leads to several benefits like variance reduction in the estimation of probabilities or representation formulae for gradients or subdifferentials of probability functions. SRD has been successfully applied to the numerical solution of many problems of operations research and optimal control involving probabilistic or chance constraints (e.g., [10, 12, 14]). For an introduction to SRD, we refer to [13].

2.2 Continuous stochastic gradient method

The *Continuous Stochastic Gradient* (CSG) method [20] is an optimization scheme designed for problems containing expected values. Consider a set of admissible designs $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, a set of uncertain parameters $\zeta \in \mathcal{Z}^{\mathsf{ad}} \subseteq \mathbb{R}^{d_\zeta}$ as well as a Borel probability measure μ on $\mathcal{Z}^{\mathsf{ad}}$ with $\zeta \sim \mu$. Then, we consider CSG's application to the problem

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & g_0(x), \\ \text{s.t.} \quad & \varphi(x) := \mathbb{E}_\zeta\big[\phi(x,\zeta)\big] = \int_{\mathcal{Z}^{\mathsf{ad}}} \phi(x,\zeta)\mu(\mathrm{d}\zeta) \geq p. \end{aligned} \tag{1}$$

Note that this includes problem (P) from the introduction, by reformulating the probability $\mathbb{P}$ as the expected value of an appropriate indicator function. In such cases, gradient-based methods require the calculation of

$$\nabla\varphi(x) = \mathbb{E}_\zeta\big[\nabla_x\phi(x,\zeta)\big]$$

in each iteration, which may be numerically intractable. Thus, following the concepts of the *Stochastic Gradient* (SG) method in [23], the current true gradient $\nabla\varphi(x_n)$ is replaced by a stochastic direction $\hat{G}_n$, which is obtained by sampling $\nabla_x\phi(x_n,\cdot)$ at a small number N_B (batch size) of independent and identically distributed random samples $\{\zeta_n^{(i)}\}_{i=1,\dots,N_B}$. In SG, this direction is constructed by simply setting

$$\hat{G}_n^{\mathsf{SG}} := \frac{1}{N_B}\sum_{i=1}^{N_B} \nabla_x\phi\big(x_n,\zeta_n^{(i)}\big),$$

which yields an *unbiased* gradient estimator, i.e.,

$$\mathbb{E}_{\{\zeta_n^{(i)}\}}\big[\hat{G}_n^{\mathsf{SG}}\big] = \nabla\varphi(x_n).$$

Unfortunately, this approach has two major drawbacks we want to avoid in this contribution:

1 If evaluating $\nabla_x\phi(x,\cdot)$ is time consuming, e.g., because it requires the simulation of a highly complex system, discarding gradient samples from previous iterations is wasting precious information.

2 The constraint appearing in (1) will later on be handled by a penalty approach. However, the resulting penalty term

$$P_\lambda(x) = \frac{\lambda}{2}\max\big\{0, p-\varphi(x)\big\}^2 \tag{2}$$

depends on φ in a nonlinear fashion. Thus, by the chain rule, we cannot construct an unbiased estimator to $\nabla P_\lambda(x_n)$, as this would require knowledge of $\varphi(x_n)$. Therefore, the resulting optimization problem is out of scope for SG.

To deal with both of these drawbacks simultaneously, CSG stores old gradient information. Then, in each iteration, old and new gradient samples are used to build a *biased* gradient approximation

$$\hat{G}_n^{\mathsf{CSG}} := \sum_{i,k} \alpha_{k,i}\nabla_x\phi\big(x_k,\zeta_k^{(i)}\big), \quad i=1,\dots,N_B,\ k=1,\dots,n.$$

The parameters $\alpha_{k,i} \geq 0$ are called *integration weights* and correspond to an efficient on-the-fly integration of the nearest neighbor surrogate model

$$\nabla_x\phi(x,\zeta) \approx \nabla_x\phi\big(x_k,\zeta_k^{(i)}\big), \quad (k,i) \in \operatorname*{arg\,min}_{(u,v)} \left\|(x,\zeta) - \big(x_u,\zeta_u^{(v)}\big)\right\|.$$

Several numerical schemes for the calculation of integration weights have been proposed in [11]. Therein, a key result for CSG convergence theory, the *approximation property* ([11, Lemma 4.6]), was shown as well. It states that, assuming some regularity of ϕ and $\nabla_x \phi$, we have

$$\left\| \hat{G}_n^{\mathsf{CSG}} - \nabla \varphi(x_n) \right\| + \left| \hat{J}_n^{\mathsf{CSG}} - \varphi(x_n) \right| \xrightarrow{a.s.} 0 \quad (n \to \infty),$$

where $\hat{J}_n^{\mathsf{CSG}}$ corresponds to the stochastic CSG approximation of $\varphi(x_n)$, i.e.,

$$\hat{J}_n^{\mathsf{CSG}} = \sum_{i,k} \alpha_{k,i}\, \phi\big(x_k, \zeta_k^{(i)}\big).$$

This result now implies the following properties of CSG:

1 As the gradient (and function value) approximation error vanishes over the course of iterations, the method asymptotically behaves like an exact full gradient scheme.

2 Using the function value approximations $\hat{J}_n^{\mathsf{CSG}}$, the CSG method is capable of handling the penalty term P_λ, see (2), allowing us to solve the chance-constrained optimization problem (P).

Nonetheless, there are some drawbacks we need to address before combining CSG with SRD. First off, the integration weight calculation requires additional effort, but can be seen as fixed costs for the optimization, added on top of the gradient and function evaluations. Especially in settings where evaluating $\nabla_x \phi(x, \cdot)$ is numerically expensive, the amount of time spent calculating integration weights is negligible in contrast to the time gained by performing fewer evaluations overall. However, if evaluations come basically for free, it is likely that CSG will be outperformed by other approaches.

Moreover, one of the prerequisites for the approximation property is global Lipschitz continuity of ϕ and $\nabla_x \phi$ in all arguments. As we will see, this assumption is violated when reformulating φ via SRD. Thus, we will extend the approximation property to cases with less regularity in Section 4.

2.3 Combination of CSG and SRD

Since a general chance constraint is typically not continuous, the authors in [4] smoothen the indicator function with a continuously differentiable one, which is a common technique in chance-constrained optimization (see, e.g., [15], [22], [25] and [26]). Reformulating the chance constraint via SRD, this challenge no longer occurs, as the probability function is continuous in this case. Therefore, we get rid of the necessity to smooth the expressions. Similarly to [4], we penalize a violation of the chance constraint in the objective function, since the CSG method is only able to handle constraints directly by projecting onto the feasible set, which is typically impossible for the chance constraint. We apply a quadratic penalty term to keep the differentiability properties of the probability function φ. Concretely, the approximation to the chance-constrained problem reads for a fixed penalty parameter $\lambda > 0$

$$\min_x \; g_0(x) + \frac{1}{2}\lambda \max\{0, p - \varphi(x)\}^2.$$

We note that we only consider finite values for λ instead of $\lambda \to \infty$ in our numerical tests since appropriately chosen values already produce nearly feasible solutions. With the extended convergence analysis for the CSG method in Section 4 this approximation problem can be directly solved with the CSG method for a large number of applications.

3 Application to analytic model problems

After presenting the general framework for combining the CSG method with SRD, we discuss its performance on small analytic examples with two different chance constraints. The first one contains a norm constraint that defines a ball, and the second one is a joint chance constraint whose inner constraints define a box. Before presenting the numerical results in Section 3.2, we briefly introduce the two problem classes.

3.1 Problem formulations

Our first analytic example contains a single chance constraint and is formally given by

$$\begin{aligned} \min_x \quad & \sum_{i=1}^m x_i \\ s.t. \quad & \varphi_{\text{ball}}(x) := \mathbb{P}_\zeta(\|\zeta - x\|^2 \leq R^2) \geq p, \end{aligned}$$

where $R > 0$ is a given radius, $p \in (0,1)$ is the desired probability level, $\|\cdot\|$ denotes the Euclidean norm and ζ follows a standard multivariate Gaussian distribution in $\mathbb{R}^m$. Owing to SRD, we know that $\varphi_{\text{ball}}(x) \geq p$ is equivalent to

$$\|x\| \leq \sqrt{q_{p,m}},$$

where $q_{p,m}$ is the solution of the equation $F_{\chi^2}^{m,q}(R^2) = p$ in q. We note that the existence of $q_{p,m}$ is guaranteed only for sufficiently large $R > 0$ and sufficiently small $p \in (0,1)$, which is the case in our following numerical examples. Here, $F_{\chi^2}^{m,q}$ denotes the distribution function of the non-central χ^2-distribution with m degrees of freedom and non-centrality parameter q^2. Evidently, the unique globally optimal solution to the chance-constrained problem is given by the analytic formula

$$x^* = -\left(\sqrt{\frac{q_{p,m}}{m}}, \ldots, \sqrt{\frac{q_{p,m}}{m}}\right)^T$$

due to the symmetry of the objective function.

The second example extends the first setting to a joint chance constraint, where multiple conditions have to hold simultaneously with probability p. Concretely, the conditions in the chance constraint are given by a box. Considering again a linear objective function, we obtain the optimization problem

$$\begin{aligned} \min_x \quad & \sum_{i=1}^m x_i \\ s.t. \quad & \varphi_{\text{box}}(x) := \mathbb{P}_\zeta(-R \leq \zeta_i - x_i \leq R \text{ for all } i = 1, \ldots, m) \geq p, \end{aligned}$$

where $R > 0$ is the given size of the box, $p \in (0,1)$ is again the desired probability level, and ζ follows a standard multivariate Gaussian distribution. Compared to the first setting, a globally optimal solution can no longer be computed analytically. Since the components of ζ are independent, the chance constraint is equivalent to

$$p \leq \varphi_{\text{box}}(x) = \prod_{i=1}^m [\Phi(x_i + R) - \Phi(x_i - R)],$$

where Φ is the distribution function of the one-dimensional standard Gaussian distribution. Since the objective function is again symmetric, finding a globally optimal solution can be reduced to computing a scalar t with

$$\Phi(t + R) - \Phi(t - R) = \sqrt[m]{p}.$$

Although there is no analytical solution to this equation, high-quality solutions can be computed numerically. Therefore, it is possible to compare the results of our method to this solution.

Coming back to the example with a ball constraint, choosing appropriate combinations of the radius R and the probability level p, the gradient of the respective penalized objective function is globally Lipschitz continuous, which implies that the convergence theory for the CSG method holds. Nevertheless, there also exist parameter combinations, where this property is violated. For the box constraints global Lipschitz continuity is always violated.

3.2 Numerical results

Although the convergence criteria for the CSG method are only fulfilled in special settings with the ball constraint, we numerically test our combination of SRD and CSG on both problem formulations. Concretely, we choose instances with dimension $m = 3$ for both chance constraints. The radius of the ball and the size of the box are chosen as $R = 5$ and the penalty parameter for our formulation is given by $\lambda = 10^4$. As mentioned above, choosing $p \in (0, 1)$ large enough guarantees Lipschitz continuity of $\nabla\varphi_{\text{ball}}$. Loosely speaking, in this case, the SRD rays are viewed from a point within the ball and the length of these rays depends smoothly on the direction. In our numerical experiments, we found $p = 0.8$ to be sufficiently large. Note that similar regularity cannot be expected for $\nabla\varphi_{\text{box}}$, due to edges and corners (see also Section 4).

To obtain examples, where SRD rays are constructed from points outside the ball or box, we use the probability levels $p = 0.4$ (ball) and $p = 0.1$ (box). In these cases, the Lipschitz condition is violated. We note that we choose these values to obtain different settings, where the convergence criteria for the CSG method are sometimes satisfied and sometimes not, to compare the method's performance for these cases, but the concrete values do not matter for these small analytical examples. We stop the CSG method after 4000 iterations and evaluate the feasibility for the original chance constraint with a Quasi-Monte-Carlo (QMC) approximation with 50000 sample points. To overcome the issue of randomness we perform 500 random runs of the CSG method with the zero-vector as starting point. Furthermore, we choose the decreasing step size $\tau_n = 0.2/n^{0.7}$. We note that these parameters have been chosen with some preliminary computational experiments, but do not represent optimal choices.

The algorithm is implemented in Matlab R2024b and all computations are executed on a machine with an Intel Xeon E3-1240 v6 CPU (3.7 gigahertz base frequency), 2 cores and 32 gigabyte RAM.

We start by numerically evaluating the convergence of the CSG method for the four tested problems. To this end, we study the original and penalized objective functions over the iterations. We note that there is no significant difference between the penalized and the original version, and therefore, we only show the medians and some quantile values of the original one in Figure 1.

On the left-hand side we see the development over all iterations, whereas on the right-hand side the y-axis is shrunk to have a better view on the quantiles. We observe that the CSG method converges for all four instances within the allowed number of iterations. Studying the quantiles we observe larger quantiles for boxes with $p = 0.1$ than for the other instances. Considering the maximal deviation, given by the fraction of the maximal difference between the objective function values over all runs and the median of the objective function values, we obtain for the final iteration deviations between 0.02 and 0.04 for all four instances and there is no significant difference between boxes and balls. Therefore, the larger quantiles result from the larger objective function values for $p = 0.1$ and the CSG method behaves similarly in all four tests.

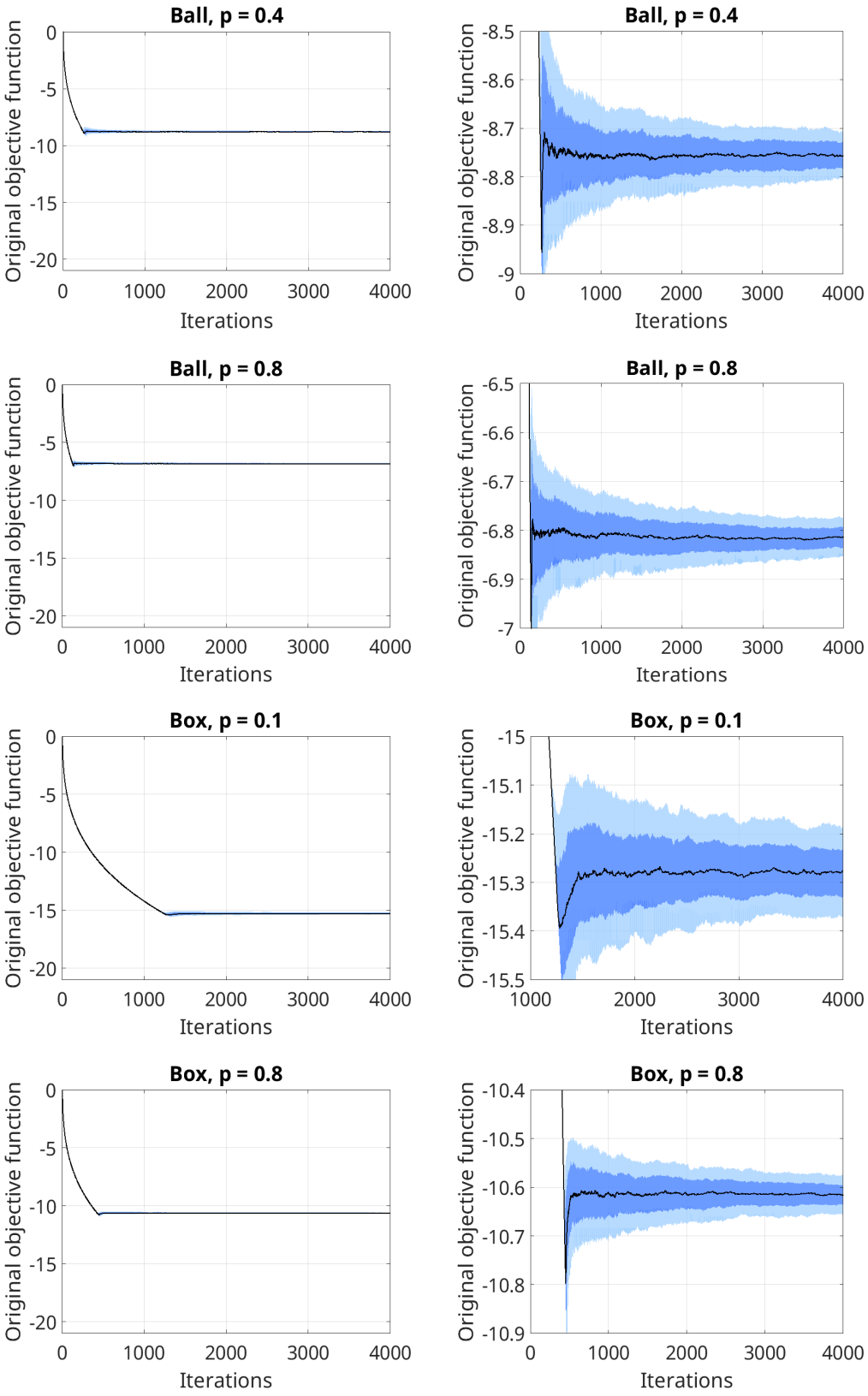

Figure 1: Objective function values over iterations for the median (black line), shaded areas for quantiles $P_{0.25,0.75}$ (dark) and $P_{0.1,0.9}$ (light)

To analyze the feasibility of the returned solutions with respect to the original problem, we study the QMC approximation of the chance constraint as well as the returned violation by the CSG method for the quality of the penalty approximation. Since the maximal returned violation is 0.004 for all instances, we conclude that the penalty parameter is chosen large enough. With respect to the QMC approximations we see that the median probability always meets exactly the desired probability level and the minimal and maximal values deviate up to 2% in each direction from the desired level p. Furthermore, we do not notice a significant difference between the tested examples.

Since we have access to the globally optimal solution for the ball problems and to a high-quality approximation for the boxes, we furthermore compare our solutions to the exact ones in the left part of Figure 2.

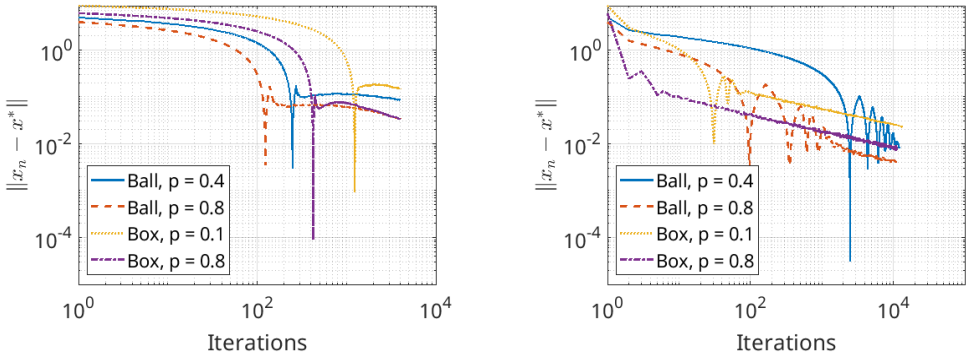

Figure 2: Median distance of the CSG iterates (left) and the SCGD iterates (right) to the optimal solution x^* in every iteration

We see that the CSG method reaches for all four instances an accuracy of about 10^{-1} to 10^{-2}. Nevertheless, the CSG method seems to be more accurate for $p = 0.8$ than for the smaller probability levels. Furthermore, we notice in all four cases one sudden improvement before the difference to the optimal solution becomes larger again. At this point, the CSG iteration is closest to the analytical solution over all iterations. We note that we start for all four instances with a feasible point and the CSG method reduces the objective function values until the penalty term for the chance constraint becomes active. Since we deal with an inexact penalty term, the solution returned by the CSG method is always slightly infeasible and the peak occurs exactly where the CSG iterations cross the analytical solutions, but due to the inexact penalty term, the CSG method reduces the objective function even more before reaching the optimum of the penalty problem. Therefore, the distance to the analytical solution becomes larger again after the peak. Nevertheless, we conclude that our combined approach produces high-quality solutions for all four instances although the theoretical convergence criteria of the CSG method are sometimes violated. For this reason, the next section deals with an extension of the convergence theory to these cases.

Before stating these theoretical results, we briefly compare the CSG method with another stochastic gradient method from the literature, namely the Stochastic compositional gradient descent algorithm (SCGD) introduced by Wang et al. [28]. Although this method is not as flexible as the CSG method, it is designed to handle two nested expectations, and is therefore applicable to our penalty formulation of the chance-constrained problem. SCGD needs the same number of evaluations per iteration than the CSG method, but instead of reusing information from the last iterations SCGD decomposes the approximation of the two nested functions by separately calculating efficient step sizes. Therefore, the additional effort for calculating the integration weights in CSG no longer occurs, but SCGD typically needs much more iterations to converge. To compare the two approaches, we allow a running time of

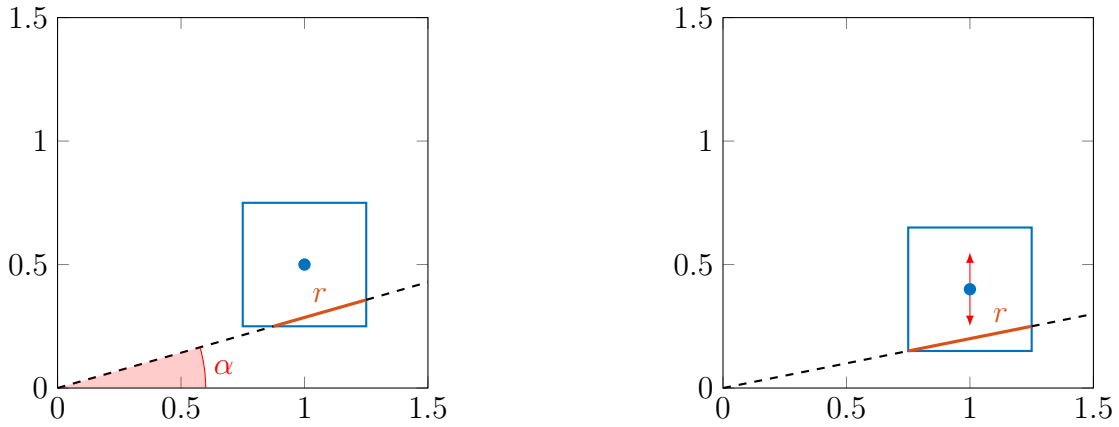

Figure 3: Illustration of the setup in the academic example using φ_{box}. Utilizing the SRD reformulation, the probability of a region with center $\bar{x} \in \mathcal{X}$ (blue dot) is obtained by integrating the (weighted) intersection length r (red) over all possible directions $v \in \mathbb{S}^1$. The value of r might change non-smoothly when varying the direction v (left) or moving $\bar{x}$ (right).

70 seconds for SCGD since all performed runs of the CSG method need between 60 and 70 seconds. The respective distance with respect to the exact solution over the iterations is shown in the right part of Figure 2. We note that the maximal difference to the optimal solution for the vectors created with SCGD is typically one decimal better than for CSG. But if we compare the results for both methods with the same number of iterations, namely 4000, CSG outperforms SCGD in the minimal, maximal and median values.

In general, we conclude that for these small analytic examples, both methods work quite well, but SCGD needs way more iterations to approximate the optimal solution well. Therefore, CSG should be preferred if the function or gradient evaluations are time-consuming or complicated to avoid a huge number of iterations, and for easy evaluation as in this example, SCGD is an alternative to the CSG approach. Nevertheless, we will not focus on SCGD in the following, but concentrate on an extension of the convergence theory for CSG.

4 Extension of the CSG convergence theory beyond problems that are globally Lipschitz

As stated in Section 2.2, convergence results for CSG require the inner gradient function $\nabla_x \phi$ to be Lipschitz in both arguments. However, the numerical results presented above indicate that this assumption can be weakened to cover a more general setting. In fact, both academic examples violate this assumption, as can be seen in Figure 3 and Figure 4.

In this section, we will extend the convergence theory of CSG to cover more general classes of problems and show that, due to the SRD reformulation of φ, the examples presented above fall within this category. An important observation is that for fixed $x \in \mathbb{R}^{d_x}$, discontinuities of $\nabla_x \phi$ appear precisely in directions $\zeta \in \mathbb{R}^{d_\zeta}$ for which the ray $\{\lambda\zeta \ : \ \lambda \geq 0\}$ is tangent to the boundary of the box or passes through a corner. In other words, the discontinuity set of $\nabla_x \phi$ is directly correlated to geometric properties of the set $\mathcal{Z}^{\text{ad}}$. With this in mind, we construct our general set of assumptions.

Assumption 1. *(Regularity of $\mathcal{X}, \mathcal{Z}^{\text{ad}}$ and μ) The set $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ is compact and convex. The set $\mathcal{Z}^{\text{ad}} \subseteq \mathbb{R}^{d_\zeta}$ is bounded with $\operatorname{supp}(\mu) \subseteq \mathcal{Z}^{\text{ad}}$.*

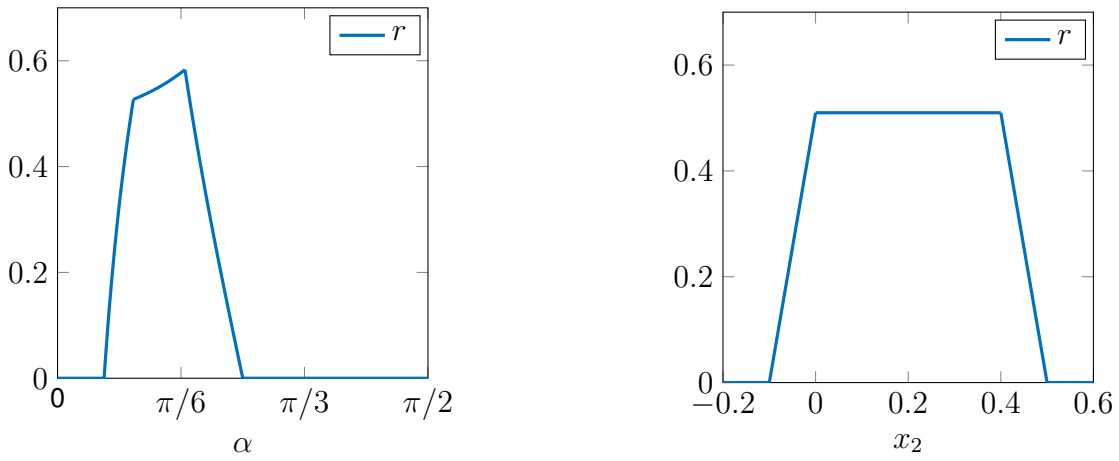

Figure 4: Left: Intersection length r (see Figure 3) for fixed midpoint $\bar{x}$ and different directions, parametrized by the angle α.
Right: Value of r for fixed direction v and varying $\bar{x} = (1, x_2)$, see Figure 3.
In both cases, we observe that kinks in the graph of r correspond to cases in which the considered ray $\{\lambda v \,:\, \lambda \geq 0\}$ passes through one of the four corners of the square.

Assumption 2. *(Regularity of φ and ϕ)*

1 *The function $\nabla\varphi : \mathcal{X} \to \mathbb{R}$ is Lipschitz continuous.*

2 *For all $\varepsilon > 0$ and for all $x \in \mathcal{X}$ exists a closed subset $\mathcal{M}(\varepsilon, x) \subseteq \mathcal{Z}^{ad}$ and a $\delta(\varepsilon, x) > 0$ such that*

$$\mu\left(\mathcal{M}(\varepsilon, x)\right) < \varepsilon \quad \text{and} \quad \phi(\cdot, \zeta) \in C^1\left(\mathcal{B}_{\delta(\varepsilon,x)}(x)\right) \quad \text{for every } \zeta \in \left(\mathcal{M}(\varepsilon, x)\right)^{\mathrm{c}}.$$

In particular, the set

$$\mathcal{V}(x) :- \left\{\zeta \in \mathcal{Z}^{ad} : \phi(\cdot, \zeta) \text{ is not continuously differentiable in } x\right\}$$

has zero measure for all $x \in \mathcal{X}$.

3 *The gradient $\nabla_x\phi$ is uniformly piecewise Lipschitz continuous, i.e., there exists $N \in \mathbb{N}$ such that for each $\varepsilon > 0$ and all $x \in \mathcal{X}$ there exists a partition of $\mathcal{M}(\varepsilon, x)^{\mathrm{c}}$ consisting of open sets $\{\mathcal{O}_j(\varepsilon, x)\}_{j=1,\dots,N}$ with the property that $\nabla_x\phi$ is uniformly Lipschitz continuous on the sets $\mathcal{B}_{\delta(\varepsilon,x)}(x) \times \mathcal{O}_j(\varepsilon, x)$.*
Hence, there exists a constant $L \in \mathbb{R}_{>0}$, independent of $\varepsilon > 0$ and $x \in \mathcal{X}$, such that

$$\|\nabla_x\phi(x_1, \zeta_1) - \nabla_x\phi(x_2, \zeta_2)\| \leq L\left(\|x_1 - x_2\|_{\mathcal{X}} + \|\zeta_1 - \zeta_2\|_{\mathcal{Z}^{ad}}\right)$$

for all $(x_1, \zeta_1), (x_2, \zeta_2) \in \mathcal{B}_{\delta(\varepsilon,x)}(x) \times \mathcal{O}_j(\varepsilon, x), j = 1, ..., N.$

Assumption 2 allows us to decompose the product space $\mathcal{X} \times \mathcal{Z}^{\mathrm{ad}}$ into smaller regions on which $\nabla_x\phi$ is well-behaved. An illustration is provided in Figure 5. The key idea to show that the approximation property [11, Lemma 4.6] remains valid under these milder assumptions is to “cut outßmall regions around the discontinuities of $\nabla_x\phi$. By Assumption 2, these regions can be chosen small enough such that their contribution to the value of $\nabla\varphi$ is arbitrary small. On the remaining regions, we can proceed as in the proof of [11, Lemma 4.6].

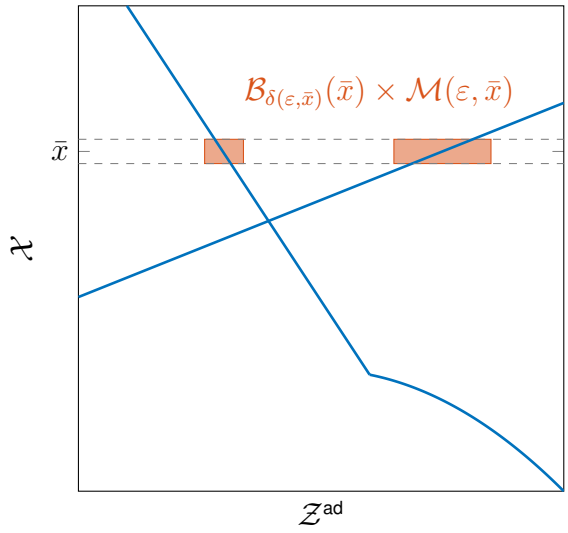

Figure 5: Product space $\mathcal{X} \times \mathcal{Z}^{\text{ad}}$ with possible discontinuity set (blue). Given $\bar{x} \in \mathcal{X}$ and $\varepsilon > 0$, we can construct a set $\mathcal{B}_{\delta(\varepsilon,\bar{x})}(\bar{x}) \times \mathcal{M}(\varepsilon,\bar{x})$ (red) that locally contains these discontinuities.

To simplify notation in the upcoming convergence result, we only consider a batch size of $N_B = 1$ and define the nearest neighbor index of $\zeta \in \mathcal{Z}^{\text{ad}}$ in iteration n as

$$k^n(\zeta) := \min\left\{ i \in \{1,\dots,n\} \,:\, i \in \operatorname*{arg\,min}_{k=1,\dots,n} \left\|(x_n,\zeta) - (x_k,\zeta_k)\right\| \right\}.$$

Thus, $k^n(\zeta)$ denotes the index of the sample point with minimal distance to (x_n,ζ). If the minimum is attained at multiple indices, $k^n(\zeta)$ is chosen as the smallest one. With $k^n(\zeta)$, the CSG gradient approximation can can be expressed as

$$\hat{G}_n = \sum_{k=1}^{n} \alpha_k \nabla_x \phi(x_k,\zeta_k) = \int_{\mathcal{Z}^{\text{ad}}} \nabla_x \phi(x_{k^n(\zeta)},\zeta_{k^n(\zeta)}) \mu(\mathrm{d}\zeta).$$

Proposition 4.1. *(Approximation Result for $\nabla\varphi$) Under Assumptions 1 and 2, the approximation error for $\nabla\varphi$ almost surely vanishes in the course of iterations, i.e., $\|\nabla\varphi(x_n) - \hat{G}_n\| \xrightarrow{a.s.} 0$ for $n \to \infty$.*

Proof. Let $\varepsilon, \tilde{\varepsilon} > 0$ and $0 < \tilde{\varepsilon} \leq \frac{\varepsilon}{6C}$, with $C > 0$ to be defined later. Utilizing the compactness of $\mathcal{X}$, we find a finite cover $\left(\mathcal{B}_{\delta_i^{\tilde{\varepsilon}}}(\bar{x}_i)\right)_{i=1,\dots,K}$ of $\mathcal{X}$ consisting of $K \in \mathbb{N}$ balls with radius $\delta_i^{\tilde{\varepsilon}} := \delta(\tilde{\varepsilon},\bar{x}_i)$ centered at points $\bar{x}_i \in \mathcal{X}$. In addition, we choose compact subsets $K_j(\tilde{\varepsilon},\bar{x}_i) \subseteq \mathcal{O}_j(\tilde{\varepsilon},\bar{x}_i)$ and a radius $r(\tilde{\varepsilon}) > 0$ such that

$$\sum_{j=1}^{N} \mu\left(\mathcal{O}_j(\tilde{\varepsilon},\bar{x}_i) \setminus K_j(\tilde{\varepsilon},\bar{x}_i)\right) < \tilde{\varepsilon} \quad \text{and} \quad K_j(\tilde{\varepsilon},\bar{x}_i) + \mathcal{B}_{r(\tilde{\varepsilon})}(0) \subseteq \mathcal{O}_j(\tilde{\varepsilon},\bar{x}_i)$$

for $i = 1,\dots,K$ and $j = 1,\dots,N$. Now, by applying [11, Lemma 4.5], we find $N(\varepsilon) \in \mathbb{N}$ such that for all $n \geq N(\varepsilon)$ and all $\zeta \in \mathcal{Z}^{\text{ad}}$ we have

$$Z_n(\zeta) := \|x_n - x_{k^n(\zeta)}\|_{\mathcal{X}} + \|\zeta - \zeta_{k^n(\zeta)}\|_{\mathcal{X}} < \min\left\{ r(\tilde{\varepsilon}), \frac{\lambda(\tilde{\varepsilon})}{2}, \frac{\varepsilon}{3L} \right\},$$

with probability 1. Here, $\lambda(\tilde{\varepsilon}) > 0$ denotes the corresponding Lebesgue number of the finite cover. In particular, for each $n \geq N(\varepsilon)$ this implies the existence of an index $i_n \in \{1, ..., K\}$ with the property $x_n, x_{k^n(\zeta)} \in \mathcal{B}_{\delta^{\tilde{\varepsilon}}_{i_n}}(\bar{x}_{i_n})$ and $\zeta_{k^n(\zeta)} \in \mathcal{O}_j(\tilde{\varepsilon}, \bar{x}_i)$ for all $\zeta \in K_j(\tilde{\varepsilon}, \bar{x}_i)$. Now, we decompose $\mathcal{Z}^{\mathsf{ad}}$ as illustrated in Figure 5. We obtain for all $n \geq N(\varepsilon)$

$$
\begin{aligned}
\left\|\nabla\varphi(x_n) - \hat{G}_n\right\| \leq & \left\| \int_{\mathcal{Z}^{\mathsf{ad}}} \nabla_x \phi(x_n, \zeta) \mu(\mathrm{d}\zeta) - \int_{\mathcal{M}(\tilde{\varepsilon}, \bar{x}_{i_n})^{\mathrm{c}}} \nabla_x \phi(x_n, \zeta) \mu(\mathrm{d}\zeta) \right\| \\
& + \left\| \int_{\mathcal{M}(\tilde{\varepsilon}, \bar{x}_{i_n})^{\mathrm{c}}} \nabla_x \phi(x_n, \zeta) \mu(\mathrm{d}\zeta) - \int_{\mathcal{Z}^{\mathsf{ad}}} \nabla_x \phi(x_{k^n(\zeta)}, \zeta_{k^n(\zeta)}) \mu(\mathrm{d}\zeta) \right\| \\
\leq & 2C(\bar{x}_{i_n}) \mu\left(\mathcal{M}\left(\tilde{\varepsilon}, \bar{x}_{i_n}\right)\right) \\
& + \sum_{j=1}^{N} \int_{\mathcal{O}_j(\tilde{\varepsilon}, \bar{x}_{i_n})} \left\| \nabla_x \phi(x_n, \zeta) - \nabla_x \phi(x_{k^n(\zeta)}, \zeta_{k^n(\zeta)}) \right\| \mu(\mathrm{d}\zeta) \\
\leq & 2C(\bar{x}_{i_n}) \tilde{\varepsilon} + 2C(\bar{x}_{i_n}) \sum_{j=1}^{N} \mu\left(\mathcal{O}_j\left(\tilde{\varepsilon}, \bar{x}_{i_n}\right) \setminus K_j\left(\tilde{\varepsilon}, \bar{x}_{i_n}\right)\right) \\
& + \sum_{j=1}^{N} \int_{K_j(\tilde{\varepsilon}, \bar{x}_{i_n})} \left\| \nabla_x \phi(x_n, \zeta) - \nabla_x \phi(x_{k^n(\zeta)}, \zeta_{k^n(\zeta)}) \right\| \mu(\mathrm{d}\zeta) \\
\leq & 4C(\bar{x}_{i_n}) \tilde{\varepsilon} + \sum_{j=1}^{N} \int_{K_j(\tilde{\varepsilon}, \bar{x}_{i_n})} L Z_n(\zeta) \mu(\mathrm{d}\zeta).
\end{aligned}
$$

Note that, by the assumed regularity of $\nabla\varphi$ and compactness of $\mathcal{X}$, the constants $C(\bar{x}_{i_n})$ with

$$
\left\| \int_{\mathcal{M}(\tilde{\varepsilon}, \bar{x}_{i_n})} \nabla_x \phi(x_n, \zeta)\, \mu(\mathrm{d}\zeta) \right\| \leq 2C(\bar{x}_{i_n}) \mu\left(\mathcal{M}\left(\tilde{\varepsilon}, \bar{x}_{i_n}\right)\right)
$$

can be chosen independent of x_n and $\bar{x}_{i_n}$. Thus, setting $\tilde{\varepsilon} \leq \frac{\varepsilon}{6C}$ yields

$$
\left\|\nabla\varphi(x_n) - \hat{G}_n\right\| < \varepsilon,
$$

finishing the proof. □

4.1 Convergence for numerical examples

By Proposition 4.1, we know that the CSG method converges for our numerical examples, as long as the regularity assumptions (Assumption 2) are satisfied. To show that this is indeed the case, we note that both settings can be captured by the following general framework:

Assumption 3. *Given $x \in \mathbb{R}^{d_x}$, we denote*

$$
\mathcal{Z}^{\mathsf{ad}}_x := \left\{ \zeta \in \mathbb{R}^{d_\zeta} \,:\, g(x, \zeta) \leq 0 \right\}.
$$

We assume that there exists a set $\mathcal{Z}^{\mathsf{ad}} \subset \mathbb{R}^{d_\zeta}$ with the following properties:

1 *$\mathcal{Z}^{\mathsf{ad}}$ is compact and convex.*

2 *There exists* $\tilde{g} \in C^1(\mathbb{R}^{d_\zeta}; \mathbb{R}^m)$ *such that* $\mathcal{Z}^{ad} = \left\{\zeta \in \mathbb{R}^{d_\zeta} \,:\, \tilde{g}(\zeta) \leq 0\right\}$. *Moreover, linear independence constraint qualification LICQ is satisfied for all* $\zeta \in \mathcal{Z}^{ad}$.

3 *There exist functions* $R \in C^{1,1}\left(\mathbb{R}^{d_x}; \mathrm{SO}(d_\zeta)\right)$ *and* $T \in C^{1,1}(\mathbb{R}^{d_x}; \mathbb{R}^{d_\zeta})$ *such that*

$$\mathcal{Z}^{ad}_x = R(x)\mathcal{Z}^{ad} + T(x) \quad \text{for all } x \in \mathbb{R}^{d_x},$$

i.e., $\mathcal{Z}^{ad}_x$ *is obtained by rotations and translations of* $\mathcal{Z}^{ad}$. *Here,* $\mathrm{SO}(d_\zeta)$ *denotes the special orthogonal group.*

4 *There exists a vector* $\bar{\zeta} \in \mathcal{Z}^{ad}$ *satisfying the Slater condition* $\tilde{g}(\bar{\zeta}) < 0$.

Note that, by construction, both examples from Section 3 satisfy these assumptions. The remainder of this section is dedicated to prove that Assumption 1 and Assumption 3 imply Assumption 2, i.e., that the generalized convergence result (Proposition 4.1) is applicable for our numerical examples.

First off, we show that these assumptions imply the Lipschitz continuity of $\nabla\varphi$ on the set $\mathcal{X}$. Since this latter set is compact by Assumption 1, it is sufficient to verify the local Lipschitz continuity of the globally defined probability function $\nabla\varphi$ in (P). The following lemma provides a verifiable condition on the density of the random vecor ζ which is in particular satisfied for the Gaussian distribution.

Lemma 4.1. *Let* $A \subseteq \mathbb{R}^d$ *be a compact set. Let* $T : \mathbb{R}^n \to \mathbb{R}^d$ *and* $R : \mathbb{R}^n \to \mathbb{R}^{d\times d}$ *be mappings with locally Lipschitz continuous derivatives such that* $R(x)$ *is a regular* $d \times d-$ *matrix for all* $x \in \mathbb{R}^n$. *Finally, let* ξ *be a* $d-$ *dimensional random vector having a density* f_ξ *with locally Lipschitz continuous derivative (e.g., multivariate Gaussian distribution). Then, the probability function*

$$\varphi(x) := \mathbb{P}(\xi \in T(x) + R(x)A) \quad (x \in \mathbb{R}^n)$$

is continuously differentiable and its derivative is locally Lipschitz.

Proof. By assumption, we have that, thanks to regularity of $R(x)$,

$$\varphi(x) = \int\limits_{T(x)+R(x)A} f_\xi(z)dz = \int\limits_A \underbrace{f_\xi(T(x) + R(x)y) \det R(x)}_{h(x,y)}\, dy \quad (x \in \mathbb{R}^n).$$

Fix an arbitrary $\bar{x} \in \mathbb{R}^n$. Our assumptions ensure that h is continuously differentiable. In particular, thanks to the compactness of A, there exists a constant M such that for all x close to $\bar{x}$, for all $y \in A$ and all $j = 1, \ldots, n$, the estimate $\left|\frac{\partial h}{\partial x_j}(x, y)\right| \leq M$ holds true. Next, fix an arbitrary $j \in \{1, \ldots, n\}$ and denote by e_j the corresponding standard unit vector in $\mathbb{R}^n$. For an arbitrary sequence $t_k \to 0$, we obtain that

$$\frac{\varphi(\bar{x} + t_k e_j) - \varphi(\bar{x})}{t_k} = \int\limits_A \frac{h(\bar{x} + t_k e_j, y) - h(\bar{x}, y)}{t_k} dy$$

By the Mean Value Theorem, for each $y \in A$, there exists some $\tau_k(y)$ with $|\tau_k(y)| \leq |t_k|$ such that

$$\frac{h(\bar{x} + t_k e_j, y) - h(\bar{x}, y)}{t_k} = \frac{\partial h}{\partial x_j}(\bar{x} + \tau_k(y)e_j, y) \to \frac{\partial h}{\partial x_j}(\bar{x}, y).$$

Moreover, $|\frac{\partial h}{\partial x_j}(\bar{x} + \tau_k(y)e_j, y)| \leq M$ for all k sufficiently large and all $y \in A$. Now, since A has finite Lebesgue measure as a compact set, the Bounded Convergence Theorem yields that

$$\frac{\varphi(\bar{x} + t_k e_j) - \varphi(\bar{x})}{t_k} \to \int\limits_A \frac{\partial h}{\partial x_j}(\bar{x}, y)dy = \frac{\partial \varphi}{\partial x_j}(\bar{x})$$

because $t_k \to 0$ was arbitrary. Since $\bar{x} \in \mathbb{R}^n$ and $j \in \{1, \dots, n\}$ were arbitrary, we have that all partial derivatives of φ exist and are given by

$$\frac{\partial \varphi}{\partial x_j}(x) = \int_A \frac{\partial h}{\partial x_j}(x, y) dy \quad (j = 1, \dots, n;\ x \in \mathbb{R}^n).$$

Now, let again $\bar{x} \in \mathbb{R}^n$ and $j \in \{1, \dots, n\}$ be arbitrary and consider an arbitrary converging sequence $x^{(k)} \to \bar{x}$. Then, repeating the argument of the Bounded Convergence Theorem already employed above, we check that

$$\frac{\partial h}{\partial x_j}(x^{(k)}, y) \to \frac{\partial h}{\partial x_j}(\bar{x}, y); \quad \left| \frac{\partial h}{\partial x_j}(x^{(k)}, y) \right| \leq M \ \forall k \geq k_0,$$

whence

$$\frac{\partial \varphi}{\partial x_j}(x^{(k)}) \to \int_A \frac{\partial h}{\partial x_j}(\bar{x}, y) dy = \frac{\partial \varphi}{\partial x_j}(\bar{x}).$$

Therefore, φ is continuously differentiable. It remains to show that the partial derivatives are locally Lipschitz. To this aim, let once more $\bar{x} \in \mathbb{R}^n$ and $j \in \{1, \dots, n\}$ be arbitrary. Observe that our assumptions imply that $\frac{\partial h}{\partial x_j}$ is locally Lipschitz. As a consequence of A being compact, there exists an open neighborhood $\mathcal{U}(\bar{x})$ of $\bar{x}$ and a modulus $L > 0$ such that

$$\left| \frac{\partial h}{\partial x_j}(x^a, y) - \frac{\partial h}{\partial x_j}(x^b, y) \right| \leq L \|x^a - x^b\| \quad \forall x^a, x^b \in \mathcal{U}(\bar{x}) \ \forall y \in A.$$

This then entails, with c denoting the Lebesgue measure of A, that

$$\left| \frac{\partial \varphi}{\partial x_j}(x^a) - \frac{\partial \varphi}{\partial x_j}(x^b) \right| \leq Lc \|x^a - x^b\| \quad \forall x^a, x^b \in \mathcal{U}(\bar{x}).$$

Hence, the derivative of φ is locally Lipschitz. □

For the other properties listed in Assumption 2 we have to distinguish two cases: tangent directions and directions passing through an edge of $\mathcal{Z}^{\text{ad}}$. We begin with the latter.

Edges

Let $\mathcal{P}_{\mathbb{S}^{d_\zeta - 1}}(\zeta) := \frac{\zeta}{\|\zeta\|}$ for $\zeta \in \mathbb{R}^{d_\zeta} \setminus \{0\}$ denote the projection of $\zeta \in \mathbb{R}^{d_\zeta}$ on the unit sphere $\mathbb{S}^{d_\zeta - 1}$. Moreover, for $\mathcal{I} \subseteq \{1, \dots, m\}$, we define $\mathcal{I}^{\text{c}} := \{1, \dots, m\} \setminus \mathcal{I}$ and

$$\mathcal{E}(\mathcal{I}) := \big\{ \zeta \in \mathcal{Z}^{\text{ad}} \,:\, \tilde{g}_{\mathcal{I}^{\text{c}}}(\zeta) < 0,\ \tilde{g}_{\mathcal{I}}(\zeta) = 0 \big\}.$$

By Assumption 3, it holds

$$\partial \mathcal{Z}^{\text{ad}} = \bigcup_{|\mathcal{I}| \geq 1} \mathcal{E}(\mathcal{I}).$$

Furthermore, the edges of the set $\mathcal{Z}^{\text{ad}}$ correspond to points of $\mathcal{Z}^{\text{ad}}$ at which at least two inequalities are active, i.e., they are given by

$$\bigcup_{|\mathcal{I}| \geq 2} \mathcal{E}(\mathcal{I}).$$

We show that the corresponding set of directions $v \in \mathbb{S}^{d_\zeta - 1}$ passing through an edge is of Hausdorff measure zero, as our regularity assumptions imply that $\mathcal{Z}^{\text{ad}}$ has a finite amount of edges and each individual edge is of low dimension.

Lemma 4.2. *Under Assumption 3, it holds that* $\mathcal{H}^{d_\zeta-1}_{\mathbb{S}^{d_\zeta-1}}\left(\mathcal{P}_{\mathbb{S}^{d_\zeta-1}}\left(\bigcup_{|\mathcal{I}|\geq 2}\mathcal{E}(\mathcal{I})\right)\right)=0.$

Proof. Let $k := |\mathcal{I}|$. Utilizing LICQ and the preimage theorem, we conclude that $\mathcal{E}(\mathcal{I})$ is a $(d_\zeta - k)$-dimensional submanifold. Hence, for every $p \in \mathcal{E}(\mathcal{I})$ there exists a C^1-diffeomorphism Φ_p on an open neighborhood V_p of p such that

$$\Phi_p\left(\mathcal{E}(\mathcal{I})\cap V_p\right)=\Phi_p\left(V_p\right)\cap\left(\mathbb{R}^{d_\zeta-k}\times\{0\}^k\right).$$

Then, $\psi_p :\equiv \Phi_p^{-1}(\cdot,0)$ defines a C^1-mapping on the open set

$$U_p:=\left\{\zeta\in\mathbb{R}^{d_\zeta-1}\mid(\zeta,0)\in\Phi_p\left(V_p\right)\right\}\subseteq\mathbb{R}^{d_\zeta-1}.$$

Now, let $W_p := \left\{\zeta \in U_p \mid \zeta_{d_\zeta-j} = 0 \text{ for all } j < k\right\}$ and note that $\psi_p(W_p) = \mathcal{E}(\mathcal{I}) \cap V_p$ as well as $\lambda_{d_\zeta-1}(W_p) = 0$. We consider φ_p to be a local parametrization of the sphere $\mathbb{S}^{d_\zeta-1}$ such that $\mathcal{P}_{\mathbb{S}^{d_\zeta-1}}\left(\mathcal{E}(\mathcal{I})\cap V_p\right)\subseteq\operatorname{im}(\varphi_p)$. Since

$$\varphi_p^{-1}\circ\mathcal{P}_{\mathbb{S}^{d_\zeta-1}}\circ\psi_p\in C^1\left(U_p;\mathbb{R}^{d_\zeta-1}\right),$$

we deduce

$$\lambda_{d_\zeta-1}\left(\varphi_p^{-1}\left(\mathcal{P}_{\mathbb{S}^{d_\zeta-1}}\left(\mathcal{E}(\mathcal{I})\cap V_p\right)\right)\right)=\lambda_{d_\zeta-1}\left(\left(\varphi_p^{-1}\circ\mathcal{P}_{\mathbb{S}^{d_\zeta-1}}\circ\psi_p\right)\left(W_p\right)\right)=0$$

and the claim follows from the compactness of $\mathcal{Z}^{\mathsf{ad}}$. □

Tangent directions

By the assumed convexity of $\mathcal{Z}^{\mathsf{ad}}$, if $0 \in \operatorname{int}(\mathcal{Z}^{\mathsf{ad}})$, there can be no directions $v \in \mathbb{S}^{d_\zeta-1}$ that are tangent to $\mathcal{Z}^{\mathsf{ad}}$ and therefore nothing has to be shown in this case. Thus, we always assume $0 \notin \operatorname{int}(\mathcal{Z}^{\mathsf{ad}})$ in the following, without mentioning it explicitly every time.

Definition 4.1. *(Tangential Points). We say $\zeta \in \mathcal{Z}^{ad}$ is a tangential point of the set $\mathcal{Z}^{ad}$, if there exists an outer normal vector $n_\zeta \in \mathbb{R}^{d_\zeta}\backslash\{0\}$ and a corresponding half space $\mathcal{H}_\zeta := \left\{x \in \mathbb{R}^{d_\zeta} \mid \langle n_\zeta, x\rangle \leq \langle n_\zeta, \zeta\rangle\right\}$ such that $\mathcal{Z}^{ad} \subseteq \mathcal{H}_\zeta$ and $0 \in (int\,(\mathcal{H}_\zeta))^{\mathrm{c}}$. We denote the set of tangential points of $\mathcal{Z}^{ad}$ by $\mathcal{T}$.*

Definition 4.2. *(Tangential Directions). Let $\mathcal{T}_0 := \{\zeta \in \mathcal{T} \mid \langle n_\zeta, \zeta\rangle = 0\} \setminus \{0\}$. We define the set of tangential directions of $\mathcal{Z}^{ad}$ as $\mathcal{V} := \mathcal{P}_{\mathbb{S}^{d_\zeta-1}}(\mathcal{T}_0)$.*

In the following, we denote the closed and convex conical hull of the set $\mathcal{Z}^{\mathsf{ad}}$ as

$$\mathcal{K}^*:=\operatorname{cone}(\mathcal{Z}^{\mathsf{ad}})=\left\{\lambda\zeta\mid\zeta\in\mathcal{Z}^{\mathsf{ad}},\lambda\geq 0\right\}.$$

Note that if $0 \in \partial\mathcal{Z}^{\mathsf{ad}}$, $\mathcal{K}^*$ is identical to the usual tangential cone of $\mathcal{Z}^{\mathsf{ad}}$ at 0.

Lemma 4.3. *Under Assumption 3, it holds that* $\mathcal{V}=\partial_{\mathbb{S}^{d_\zeta-1}}\left(\mathbb{S}^{d_\zeta-1}\cap\mathcal{K}^*\right)=\mathbb{S}^{d_\zeta-1}\cap\partial\mathcal{K}^*.$

Proof. Let $v \in \mathcal{V}$. Then, there exists $\lambda > 0$ such that $v \in \lambda^{-1}\mathcal{T}_0 \subseteq \lambda^{-1}\partial\mathcal{Z}^{\mathsf{ad}} \subseteq \mathcal{K}^*$. Hence, we find an outer normal vector $n_{\lambda v} \in \mathbb{R}^{d_\zeta} \setminus \{0\}$ such that $\langle n_{\lambda v}, v\rangle = 0$ and $\mathcal{Z}^{\mathsf{ad}} \subseteq \mathcal{H}_{\lambda v}$. This implies $\mathcal{K}^* \subseteq \mathcal{H}_{\lambda v}$ and for every $\delta > 0$, there exists $\alpha_\delta > 0$ such that

$$\mathcal{P}_{\mathbb{S}^{d_\zeta-1}}\left(\lambda v+\alpha_\delta n_{\lambda v}\right)\in\mathbb{S}^{d_\zeta-1}\cap\left(\mathcal{K}^*\right)^{\mathrm{c}}\cap\mathcal{B}_\delta(v).$$

Thus, $v \notin \text{int}_{\mathbb{S}^{d_\zeta-1}}\left(\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^*\right)$, which implies $v \in \partial_{\mathbb{S}^{d_\zeta-1}}\left(\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^*\right)$.
Now, let $v \in \partial_{\mathbb{S}^{d_\zeta-1}}\left(\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^*\right)$. Then, $v \in \partial\mathcal{K}^* \subseteq \mathcal{K}^*$ and we find $\mu > 0$ such that $\mu v \in \mathcal{Z}^{\text{ad}}$. The convexity of $\mathcal{K}^*$ implies the existence of $n_{\mu v} \in \mathbb{R}^{d_\zeta} \setminus \{0\}$ such that $\langle n_{\mu v}, k - \mu v\rangle \leq 0$ for all $k \in \mathcal{K}^*$. Choosing $\lambda > 1$, we obtain

$$0 \geq \langle n_{\mu v}, \lambda\mu v - \mu v\rangle = (\lambda - 1)\langle n_{\mu v}, \mu v\rangle \geq 0.$$

Hence, $\langle n_{\mu v}, \mu v\rangle = 0$ and the claim follows. □

Next, we show that $\mathcal{V}$ is of Hausdorff measure zero and has open neighborhoods of arbitrary small Hausdorff measure. For $\varepsilon > 0$, let $(\mathcal{V})_\varepsilon := \left\{v \in \mathbb{S}^{d_\zeta-1} \mid \text{dist}\,(v, \mathcal{V}) \leq \varepsilon\right\}$.

Lemma 4.4. *Under Assumption 3, it holds that* $\mathcal{H}^{d_\zeta-1}_{\mathbb{S}^{d_\zeta-1}}(\mathcal{V}) = 0$. *Furthermore, for every* $\delta > 0$ *there exists* $\varepsilon_\delta > 0$ *such that* $\mathcal{H}^{d_\zeta-1}_{\mathbb{S}^{d_\zeta-1}}\left((\mathcal{V})_{\varepsilon_\delta}\right) < \delta$.

Proof. First, assume that $0 \notin \mathcal{Z}^{\text{ad}}$. From the convexity of $\mathcal{Z}^{\text{ad}}$, it follows that $\mathcal{K}^*$ is a convex and pointed cone. Thus, the set $\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^*$ is spherically convex and we can choose $w \in \text{int}_{\mathbb{S}^{d_\zeta-1}}\left(\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^*\right)$ such that

$$\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^* \subseteq S^+_w := \left\{v \in \mathbb{S}^{d_\zeta-1} \mid \langle w, v\rangle > 0\right\}.$$

Now, let $\mathbb{R}^{d_\zeta}_w := \left\{\zeta \in \mathbb{R}^{d_\zeta} \mid \langle w, \zeta\rangle = 0\right\}$ and denote the corresponding gnomonic projection as $g_w : S^+_w \to \mathbb{R}^{d_\zeta}_w, v \mapsto \langle w, v\rangle^{-1}v - w$. The function g_w is a bijection with inverse $g_w^{-1} :\equiv \mathcal{P}_{\mathbb{S}^{d_\zeta-1}}(\cdot + w)$ and it maps closed and spherically convex subsets in S^+_w to convex sets in $\mathbb{R}^{d_\zeta}_w$, see [5]. In the following, we proceed in a similar way to Lemma 4.3. By choosing a parametrization φ of the sphere $\mathbb{S}^{d_\zeta-1}$ with the property $\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^* \subseteq \text{im}(\varphi)$, rotating the hyperplane $\mathbb{R}^{d_\zeta}_w$ by a linear mapping $R \in \text{SO}(d_\zeta)$ such that $R(\mathbb{R}^{d_\zeta}_w) \subseteq \mathbb{R}^{d_\zeta-1} \times \{0\}$, then applying the canonical embedding E from $\mathbb{R}^{d_\zeta-1}$ to $\mathbb{R}^{d_\zeta}$ and denoting $\phi :\equiv E^{-1} \circ R \circ g_w$, we obtain

$$\lambda_{d_\zeta-1}\left(\varphi^{-1}(\mathcal{V})\right) \leq \lambda_{d_\zeta-1}\left(\left(\varphi^{-1} \circ \phi^{-1}\right)\left(\partial\left(\phi\left(\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^*\right)\right)\right)\right) = 0.$$

Here, we utilized that the set $\partial\left(\phi\left(\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^*\right)\right)$ has zero measure as the boundary of the convex set $\phi\left(\mathbb{S}^{d_\zeta-1} \cap \mathcal{K}^*\right) \subseteq \mathbb{R}^{d_\zeta-1}$.

If $0 \in \partial\mathcal{Z}^{\text{ad}}$, $\mathcal{K}^*$ is either a convex pointed cone, or a half-space, due to the convexity of $\mathcal{Z}^{\text{ad}}$. In the first case, we can proceed as above, while in the second case, $\mathcal{H}^{d_\zeta-1}_{\mathbb{S}^{d_\zeta-1}}(\mathcal{V}) = 0$ follows directly from $\mathcal{H}^{d_\zeta-1}_{\mathbb{S}^{d_\zeta-1}}\left(\mathbb{S}^{d_\zeta-2}\right) = 0$.

We now prove the second assertion. Since $\mathcal{V}$ is compact, we have $\mathcal{V} = \bigcap_{i=1}^{\infty}(\mathcal{V})_{\varepsilon_i}$ for any monotonically decreasing null sequence $\{\varepsilon_i\}_{i\in\mathbb{N}}$ and this implies

$$0 = \mathcal{H}^{d_\zeta-1}_{\mathbb{S}^{d_\zeta-1}}(\mathcal{V}) = \mathcal{H}^{d_\zeta-1}_{\mathbb{S}^{d_\zeta-1}}\left(\bigcap_{i=1}^{\infty}(\mathcal{V})_{\varepsilon_i}\right) = \lim_{i\to\infty}\mathcal{H}^{d_\zeta-1}_{\mathbb{S}^{d_\zeta-1}}\left((\mathcal{V})_{\varepsilon_i}\right).$$

□

To show that part 2 of Assumption 2 is indeed satisfied, we need to make sure that the set $\mathcal{V}$ does not change too rapidly for varying designs $x \in \mathcal{X}$.

Lemma 4.5. *For all $\varepsilon > 0$ and all $x \in \mathbb{R}^{d_\zeta}$, there exists $\delta(\varepsilon, x) > 0$ and a set $\mathcal{M}(\varepsilon, x) \supseteq \mathcal{V}(x)$ such that $\mathcal{H}^{d_\zeta - 1}_{\mathbb{S}^{d_\zeta - 1}}\big(\mathcal{M}(\varepsilon, x)\big) < \varepsilon$ and*

$$\mathcal{V}(y) \subseteq \mathcal{M}(\varepsilon, x) \quad \textit{for all } y \in \mathcal{B}_{\delta(\varepsilon,x)}(x).$$

Proof. Let $\varepsilon > 0$ and $x \in \mathbb{R}^{d_\zeta}$. First, assume that $0 \notin \partial \mathcal{Z}^{\mathsf{ad}}_x$. Due to the Slater condition, there exists a $\zeta^\star \in \operatorname{int}(\mathcal{Z}^{\mathsf{ad}}_x)$. Define the functions $f^-_\phi : \mathbb{R}^n \to \mathbb{R}^n$ and $f^+_\phi : \mathbb{R}^n \to \mathbb{R}^n$ as

$$f^-_\phi : z \mapsto (z - \zeta^\star)(1 - \phi) + \zeta^\star \quad \text{and} \quad f^+_\phi : z \mapsto (z - \zeta^\star)(1 + \phi) + \zeta^\star.$$

Moreover, let $\mathcal{Z}^-(\phi, x) := f^-_\phi\left(\mathcal{Z}^{\mathsf{ad}}_x\right), \mathcal{Z}^+(\phi, x) := f^+_\phi\left(\mathcal{Z}^{\mathsf{ad}}_x\right)$ and let $\mathcal{K}^-(\phi, x)$ and $\mathcal{K}^+(\phi, x)$ be the conical hulls of the sets $\mathcal{Z}^-(\phi, x)$ and $\mathcal{Z}^+(\phi, x)$. Set $\mathcal{K}^\pm(\phi, x) := \mathcal{K}^+(\phi, x) \setminus \mathcal{K}^-(\phi, x)$. There exists a $\phi(\varepsilon) > 0$ such that

$$\mathbb{S}^{d_\zeta - 1} \cap \mathcal{K}^\pm(\phi(\varepsilon), x) \subseteq \mathcal{V}_\varepsilon(x).$$

There exists a $\delta(\varepsilon, x) > 0$ such that the inclusion $\partial \mathcal{Z}^{\mathsf{ad}}_y \subseteq \mathcal{Z}^+(\phi, x) \setminus \mathcal{Z}^-(\phi, x)$ holds true for all $y \in \mathcal{B}_{\delta(\varepsilon,x)}(x)$. This implies $\partial \mathcal{K}^*(y) \subseteq \mathcal{K}^\pm(\phi(\varepsilon), x)$ for all $y \in \mathcal{B}_{\delta(\varepsilon,x)}(x)$. In total, we obtain

$$\mathcal{V}(y) = \mathbb{S}^{d_\zeta - 1} \cap \partial \mathcal{K}^*(y) \subseteq \mathbb{S}^{d_\zeta - 1} \cap \mathcal{K}^\pm(\phi(\varepsilon), x) \subseteq \mathcal{V}_\varepsilon(x)$$

for all $y \in \mathcal{B}_{\delta(\varepsilon,x)}(x)$, i.e., we can simply choose $\mathcal{M}(\varepsilon, x) = \mathcal{V}_\varepsilon(x)$.

In the case that $0 \in \partial \mathcal{Z}^{\mathsf{ad}}_x$, we choose

$$\mathcal{M}(\varepsilon, x) := \left\{\zeta \in \mathbb{R}^{d_\zeta} \,:\, \zeta \in \mathcal{V}_{\varepsilon/2}(x) \text{ or } -\zeta \in \mathcal{V}_{\varepsilon/2}(x)\right\}$$

and consider $\delta(\frac{\varepsilon}{2}, x)$ instead. Then, the same steps as above yield the claim. □

Remark 4.1. *In the proof of Lemma 4.5, the distinction whether or not $0 \in \partial \mathcal{Z}^{ad}_x$ has a simple geometric reason: If 0 is a corner of $\mathcal{Z}^{ad}_x$, the set of tangential directions rapidly changes, even for minor perturbations of x. To see this, assume that $x = 1$ and*

$$\mathcal{Z}^{ad}_x = \{\zeta \in \mathbb{R}^2 \,:\, \|\zeta - (1, x)^T\|_\infty \leq 1\}.$$

Then, the associated tangent directions are given by $\mathcal{V}(1) = \left\{(1, 0)^T, (0, 1)^T\right\}$. However, for any $\varepsilon > 0$, we have $\mathcal{V}(1 - \varepsilon) = \left\{(0, 1)^T, (0, -1)^T\right\}$. In general, if 0 is a corner of $\mathcal{Z}^{ad}_x$ and $v \in \mathcal{V}(x)$ a tangent direction, translating $\mathcal{Z}^{ad}_x$ in direction $-v$ will immediately result in $-v \in \mathcal{V}(\tilde{x})$, even if $-v \notin \mathcal{V}(x)$.

Note that the same steps as in the proof of Lemma 4.5 yield an analogous result for directions passing through an edge of $\mathcal{Z}^{\mathsf{ad}}_x$. Thus, we see that, under Assumptions 1 and 3, we indeed satisfy Assumption 2 and are able to apply Proposition 4.1 to our numerical examples.

5 Applications to capacity maximization

In this section we want to solve the problem of capacity maximization in the setting of stationary gas flow for a small example instance, where the random demand is given by a (multivariate) Gaussian

distribution. The pick up the problem formulation that has been given in [12]. We recall that the capacity problem in the stationary case can be reformulated as a classical probabilistic program

$$\min \left\{ \langle c, u \rangle \;\middle|\; \mathbb{P}\big(g_{k,\ell}(u,\xi) \leq 0 \quad k,\ell = 0,\dots,N \big) \geq \mathrm{p} \right\}$$

with a system of random inequalities defined by

$$\begin{aligned} g_{k,\ell}(u,\xi) \quad := \quad & \left(p_k^{\min}\right)^2 + \sum_{e \in \Pi(k) \setminus \Pi(\ell)} \Phi_e \left(\sum_{n \succeq \tau(e)} (\xi_n + u_n) \right)^2 \\ & - \left(p_\ell^{\max}\right)^2 - \sum_{e \in \Pi(\ell) \setminus \Pi(k)} \Phi_e \left(\sum_{n \succeq \tau(e)} \xi_n \right)^2 \leq 0. \end{aligned}$$

The inequalities turn out to be quadratic in both the control and the random variable. This structure allows for an easy computation of the partial derivatives and the radius function needed when applying the SRD method. For more details we refer to [12], where in addition to the algorithmic aspects also a theoretical justification for the existence of involved gradients of the constraint function is given.

For our numerical study, we consider the capacity problem example of medium size represented in [12, Section 6]. Although it is not clear whether the requirements for the convergence of the CSG method hold for the capacity maximization problem, we apply the CSG method to it. Concretely, we consider the case of equal pressure bounds, where the number of inequalities to calculate the constraint function can be significantly decreased, and the case of general pressure bounds. We choose the probability level $p = 0.9$ and the penalty parameter $\lambda = 10^7$. Furthermore, we choose the all-ones vector for the linear costs and start the CSG method with the zero-vector. In total, we perform 8000 iterations with the CSG method, but since the weight calculations become very time-consuming for this number of iterations, we restrict the maximal number of old sample points used for the gradient approximation to 4000 and discard the samples that are farthest away from the actual point based on the nearest neighbor computation for the weights. Furthermore, we use an adaptive step size rule, where we start each iteration with a constant step size $\tau_0 = 5$ and increase or decrease it if the gradient steps get very small or very large. Furthermore, after the first 5000 iterations, we additionally decrease this adaptively chosen step size by the factor $1/(n-5000)^{0.9}$ in iteration n. To evaluate the probabilities with respect to the original problem we perform a QMC-approximation with 100000 sample points. Furthermore, we perform 500 random runs for both test cases to overcome the issue of randomness in the CSG method.

First, we study again the convergence of the CSG method by looking at the objective function values over the iterations. Comparing the original objective function and the penalized version, the behavior is very similar, but due to the huge penalty parameter there is more variation for the penalized version and the values are typically larger, but the differences are small, and so, we only show the median and some quantile values for the original objective function in Figure 6. Since the behavior is also very similar for the case of equal and general pressure bounds we restrict to a visualization of the general case.

We see that the medians and quantiles converge. The maximal deviations are given by 0.16 for both problem formulations. Therefore, the CSG method indeed converges, providing numerical evidence that our theoretical requirements are satisfied. Furthermore, the penalty formulation works well since the maximal returned violation by the CSG method is for both settings given by 0.01.

For the feasibility with respect to the original chance constraint we consider the QMC-approximations. We note that in the median the estimated probability always equals the desired value 0.9, but single runs can be slightly infeasible and the probabilities range from 0.88 to 0.91. Nevertheless, we

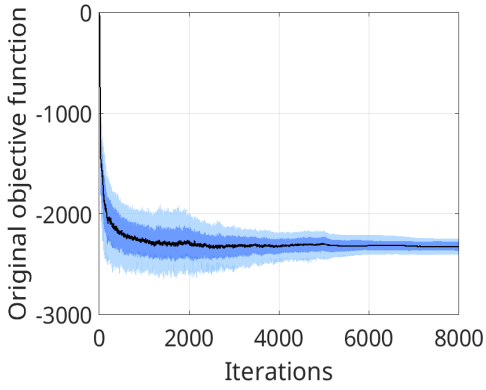

Figure 6: Objective function values over iterations for the median (black line), shaded areas for quantiles $P_{0.25,0.75}$ (dark) and $P_{0.1,0.9}$ (light) for general pressure bounds

can conclude that the combination of CSG and SRD is able to solve this difficult class of problems. Furthermore, the running times for both settings lie in the range of 240 to 270 seconds. We see no essential difference between the two settings since the number of total function evaluations is quite small, namely 8000, whereas Monte-Carlo based methods that require multiple function evaluations in every gradient iteration need significantly more time for the setting with general pressure bounds. Therefore, the CSG method should be preferred if the function or gradient evaluations are time-consuming and the stochastic dimension is small enough. Solving the example problem with SRD and by the built-in SQP solver of Matlab for the computation times in both settings we obtain approximately 40 and 120 seconds, respectively. Even if the SQP method in combination with SRD works faster in this example, we notice an increase of computation time of factor 3 in the general setting that involves substantial more random constraints. The CSG method turns out to be more self-contained if the effort of computing the probability function increases, which emphasizes the potential of this method.

6 Conclusion and outlook

In this paper, a new method for the solution of chance-constrained problems based on a combination of the SRD technique with the CSG method was presented. The convergence behavior of the CSG-SRD method was investigated using a class of academic test problems. Although standard assumptions required for the CSG convergence theory to hold were violated, our numerical studies demonstrated the correct behavior of the CSG method applied to the SRD formulation of the chance-constrained problem. This motivated us to establish a generalized convergence theory of the CSG method under weaker assumptions. It was also demonstrated that these weaker assumptions hold for a relevant class of problems covering the analytic test problems used in our numerical experiments. Beyond this, the CSG method using SRD has been applied to capacity maximization problems for gas networks. While for this problem it is open, if our new convergence framework applies, again convincing numerical results could be obtained. A comparison to classical Monte-Carlo based approaches shows that our method needs far fewer function evaluations, and is therefore especially useful if function evaluations are time-consuming. As an extension of the presented capacity maximization problems with chance constraints relying on the algebraic gas model, our proposed method should be very efficient when moving to models including ODEs or PDEs in the case of considering transient gas flow models.

References

[1] Shabbir Ahmed and Alexander Shapiro. Solving chance-constrained stochastic programs via sampling and integer programming. In *State-of-the-art decision-making tools in the information-intensive age*, pages 261–269. Informs, 2008.

[2] Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of operations research*, 23(4):769–805, 1998.

[3] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical programming*, 88:411–424, 2000.

[4] Daniela Bernhard, Frauke Liers, Michael Stingl, and Andrian Uihlein. A gradient-based method for joint chance-constrained optimization with continuous distributions. *Optimization Methods and Software*, page 1–44, 2025.

[5] Florian Besau. *Binary operations and floating bodies in spherical convexity*. PhD thesis, TU Wien, 2014.

[6] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[7] Giuseppe Calafiore and Marco C Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102:25–46, 2005.

[8] Abraham Charnes, William W Cooper, and Gifford H Symonds. Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management science*, 4(3):235–263, 1958.

[9] Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC, 2018.

[10] Caroline Geiersbach, René Henrion, and Pedro Pérez-Aros. Numerical solution of an optimal control problem with probabilistic and almost sure state constraints. *Journal of Optimization Theory and Applications*, 204(1), 2024.

[11] Max Grieshammer, Lukas Pflug, Michael Stingl, and Andrian Uihlein. The continuous stochastic gradient method: part i–convergence theory. *Computational Optimization and Applications*, 87(3):935–976, 2023.

[12] Holger Heitsch. On probabilistic capacity maximization in a stationary gas network. *Optimization*, 69(3):575–604, 2019.

[13] René Henrion and Wim Stefanus van Ackooij. (Sub-) Gradient formulae for probability functions of random inequality systems under gaussian distribution. *SIAM/ASA J. Uncertainty Quantification*, 5:63–87, 2017.

[14] René Henrion, Georg Stadler, and Florian Wechsung. Optimal control under uncertainty with joint chance state constraints: Almost-everywhere bounds, variance reduction, and application to (bi)linear elliptic PDEs. *SIAM/ASA Journal on Uncertainty Quantification*, 13(3):1028–1053, 2025.

[15] L. Jeff Hong, Yi Yang, and Liwei Zhang. Sequential convex approximations to joint chance constrained programs: A monte carlo approach. *Operations Research*, 59(3):617–630, 2011.

[16] James Luedtke and Shabbir Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699, 2008.

[17] Arkadi Nemirovski and Alexander Shapiro. Scenario approximations of chance constraints. *Probabilistic and randomized methods for design under uncertainty*, pages 3–47, 2006.

[18] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

[19] Bernardo K Pagnoncelli, Shabbir Ahmed, and Alexander Shapiro. Sample average approximation method for chance constrained programming: Theory and applications. *Journal of optimization theory and applications*, 142(2):399–416, 2009.

[20] Lukas Pflug, Niklas Bernhardt, Max Grieshammer, and Michael Stingl. CSG: A new stochastic gradient method for the efficient solution of structural optimization problems with infinitely many states. *Structural and Multidisciplinary Optimization*, 61(6):2595–2611, 2020.

[21] András Prékopa. *Stochastic Programming*. Mathematics and Its Applications. Springer Netherlands, 1995.

[22] Yong H. Ren, Ying Xiong, Yu H. Yan, and Jian Gu. A smooth approximation approach for optimization with probabilistic constraints based on sigmoid function. *Journal of Inequalities and Applications*, 2022(1), 2022.

[23] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[24] R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000.

[25] F. Shan, X. T. Xiao, and L. W. Zhang. Convergence analysis on a smoothing approach to joint chance constrained programs. *Optimization*, 65(12):2171–2193, 2016.

[26] Feng Shan, Liwei Zhang, and Xiantao Xiao. A smoothing function approach to joint chance-constrained programs. *Journal of Optimization Theory and Applications*, 163(1):181–199, 2014.

[27] Wim Stefanus van Ackooij and Welington Luis de Oliveira. *Methods of Nonsmooth Optimization in Stochastic Programming: From Conceptual Algorithms to Real-World Applications*. Springer Nature Switzerland, 2025.

[28] Mengdi Wang, Ethan X. Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1–2):419–449, 2016.

[29] Yuan Yuan, Zukui Li, and Biao Huang. Robust optimization approximation for joint chance constrained optimization problem. *Journal of Global Optimization*, 67:805–827, 2017.