Probability Maximization via Minkowski Functionals: Convex Representations and Tractable Resolution

I. E. Bardakci, A. Jalilzadeh, C. Lagoa, and U. V. Shanbhag[*]

March 11, 2022

Abstract

In this paper, we consider the maximization of a probability $\mathbb{P}\{\zeta \mid \zeta \in \mathbf{K}(\mathbf{x})\}$ over a closed and convex set $\mathcal{X}$, a special case of the chance-constrained optimization problem. We define $\mathbf{K}(\mathbf{x})$ as $\mathbf{K}(\mathbf{x}) \triangleq \{\zeta \in \mathcal{K} \mid c(\mathbf{x},\zeta) \geq 0\}$ where ζ is uniformly distributed on a convex and compact set $\mathcal{K}$ and $c(\mathbf{x},\zeta)$ is defined as either $c(\mathbf{x},\zeta) \triangleq 1-|\zeta^T\mathbf{x}|^m$, $m \geq 0$ (Setting A) or $c(\mathbf{x},\zeta) \triangleq T\mathbf{x}-\zeta$ (Setting B). We show that in either setting, by leveraging recent findings in the context of non-Gaussian integrals of positively homogeneous functions, $\mathbb{P}\{\zeta \mid \zeta \in \mathbf{K}(\mathbf{x})\}$ can be expressed as the expectation of a suitably defined continuous function $F(\bullet,\xi)$ with respect to an appropriately defined Gaussian density (or its variant), i.e. $\mathbb{E}_{\tilde{p}}[F(\mathbf{x},\xi)]$. Aided by a recent observation in convex analysis, we then develop a convex representation of the original problem requiring the minimization of $g(\mathbb{E}[F(\mathbf{x},\xi)])$ over $\mathcal{X}$ where g is an appropriately defined smooth convex function. Traditional stochastic approximation schemes cannot contend with the minimization of $g(\mathbb{E}[F(\bullet,\xi)])$ over $\mathcal{X}$, since conditionally unbiased sampled gradients are unavailable. We then develop a regularized variance-reduced stochastic approximation (**r-VRSA**) scheme that obviates the need for such unbiasedness by combining iterative regularization with variance-reduction. Notably, (**r-VRSA**) is characterized by both almost-sure convergence guarantees, a convergence rate of $\mathcal{O}(1/k^{1/2-a})$ in expected sub-optimality where $a > 0$, and a sample complexity of $\mathcal{O}(1/\epsilon^{6+\delta})$ where $\delta > 0$. To the best of our knowledge, this may be the first such scheme for probability maximization problems with convergence and rate guarantees. Preliminary numerics on a portfolio selection problem (Setting A) and a vehicle routing problem (Setting B) suggest that the scheme competes well with naive mini-batch SA schemes as well as integer programming approximation methods.

1 Introduction

This paper concerns the probability maximization problem (**PM**), defined as

$$\max_{\mathbf{x}\in\mathcal{X}} \; f(\mathbf{x}) \;\triangleq\; \mathbb{P}\{\zeta \mid \zeta \in \mathbf{K}(\mathbf{x})\}, \qquad \textbf{(PM)}$$

where f is a probability distribution function parametrized by a decision vector $\mathbf{x}$, $\mathcal{X} \subseteq \mathbb{R}^n$ denotes a closed and convex feasibility set, $\mathbf{K}(\mathbf{x}) \triangleq \{\zeta \in \mathcal{K} \mid c(\mathbf{x},\zeta) \geq 0\}$, $\mathcal{K}$ is a compact and convex set in $\mathbb{R}^n$, $c : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$. Here, $\zeta : \Omega \to \mathbb{R}^d$ is a $d-$dimensional random vector with a prescribed distribution $\mathbb{P}$. Problems of the form (**PM**) fall within the umbrella of chance-constrained optimization problems.

[*] The first author reachable at ibardakci@bartin.edu.tr, and the second author is reachable at afrooz@arizona.edu while third, and fourth authors are contactable at cml18,udaybag@psu.edu. The authors would like to acknowledge support from NSF CMMI-1538605, EPCN-1808266, DOE ARPA-E award DE-AR0001076, NIH R01-HL142732, and the Gary and Sheila Bello chair funds. Preliminary efforts at studying Setting A were carried out in [14].

1.1 Background on chance-constrained optimization

Chance-constrained optimization originates from the probabilistic scheduling of heating oil production by Charnes, Cooper, and Symonds [22]. A more formal description of chance-constrained programming as an avenue for optimization under uncertainty appeared in the eponymously titled paper by Charnes and Cooper [21]. Such avenues have assumed relevance in hydro reservoir management [8,64], portfolio management [54,70], power systems operation [15,26,36,37], routing [23], structural failure [66], and inventory and supply-chain management [35,77].

(a) *Analysis.* The analysis of optimization problems with probability functions has focused on questions of continuity, differentiability, and convexity. Of these, continuity and differentiability (and its generalized variants) are of particular relevance when developing algorithmic techniques. Convexity guarantees are important in their own right, allowing for certifying a stationary point as a global maximizer. Consider a probability function ψ, defined as $\psi(\mathbf{x}) = \mathbb{P}[\zeta \in A(\mathbf{x})]$ and $A : \mathbb{R}^n \to \mathbb{R}^m$ is a set-valued map. Under suitable convex-valuedness and continuity properties on A and an appropriate measure zero requirement on ζ, ψ is continuous [38, Th. 2.1]. In fact, if A is defined as $A(\mathbf{x}) \triangleq \{\zeta \mid c_i(\mathbf{x}, \zeta) \leq 0, i = 1, \cdots, m\}$, then continuity of ψ is implied by continuity of c_i in both its arguments for every i and a suitable regularity assumption [39].

Differentiability of ψ is a more subtle question. As eloquently described by van Ackooij [2], such results can be partitioned in two categories: (i) Under mild distributional requirements on ζ and differentiability of c_i for every i, differentiability of ψ can be concluded under a set of assumptions on $\nabla_\zeta c_i$, amongst others (cf. [72,73]); (ii) Alternately, by choosing the distribution, more refined statements are available. For instance, when the distribution of ζ belongs to the family of eliptically symmetric distributions, examples being multivariate Gaussian, Student, and logistic, under suitable differentiability and convexity properties of c_i (in the second argument) and additional assumptions, ψ can be proven to be locally Lipschitz and a characterization of its Clarke subdifferential may be provided [3, Th. 1]. In addition, if a suitable constraint qualification holds, ψ can be shown to be differentiable and its gradient can be analytically characterized [3, Cor. 1].

Finally, convexity of ψ can be claimed under various conditions [17,60,61]; for instance, joint quasi-convexity of c_i in both arguments and α-concavity of the distribution $\mathbb{P}$ implies convexity of ψ on a suitably defined set. More generally, convexity of joint chance constraints has also been studied [25,49], while more recent forays in this area have considered when the probability level is sufficiently high. Referred to as "eventual convexity", this avenue has been studied in the context of structured chance constraints involving copulae [1].

(b) *Computation.* We now discuss the main algorithmic thrusts for resolution of chance-constrained optimization.

(i) *Nonlinear programming and bundle-based approaches.* Amongst the earliest efforts for resolving chance-constrained optimization applied the penalization framework captured by the "SUMT" framework, first presented by Fiacco and McCormick [31,32], to the probabilistically constrained setting [60]. Naturally, any such effort requires having deriving gradients of probability functions, as seen in the context of nonlinear probabilistic constraints with nonconvex quadratic forms [9] as well as when contending with Gaussian distributions (and their variants) [6,7] (also see [57,71,72]). One challenge that has been observed in early efforts is the scourge of ill-conditioning in penalizaton efforts [62], leading to the development of bundle-based approaches, which have proven quite powerful [4,5,10].

(ii) *Convex characterizations and approximations*. Convexity of (Chance-Opt) can often be claimed. For instance, [63] proved that when the distribution function of ζ is logarithmically concave (or log-concave) and the functions $g_1(x,\zeta)$, $g_2(x,\zeta)$, $\ldots$, $g_r(x,\zeta)$ are quasi-concave, the function $G(x) = \mathbb{P}(\zeta : g_1(x,\zeta) \geq 0, g_2(x,\zeta) \geq 0, \ldots, g_r(x,\zeta) \geq 0)$ is a log-concave function. More recently, Lagoa et al. [42] showed that a set given by $\mathbb{P}\{a^T x \leq b\} \geq (1-\epsilon)$ is convex if (a,b) has a symmetric log-concave density and $\epsilon < 1/2$. Unfortunately, the convexity of the feasibility set does not directly allow for efficient computation, motivating a scenario-based approach. Consider (P) and its scenario-based approximation (P_N), defined next.

$$\left\{\begin{array}{ll}\min\limits_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x}) & \\ \text{s.t. } \ \mathbb{P}\left\{\zeta : c(\mathbf{x},\zeta) \geq 0\right\}. & (P)\end{array}\right\} \xrightarrow[\text{Scen. approx.}]{} \left\{\begin{array}{ll}\min\limits_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x}) & \\ \text{ s.t. } c(\mathbf{x},\zeta_j) \geq 0, j = 1,\cdots,N & (P_N)\end{array}\right\}$$

In [52], the authors examined how large N should be so that with probability $(1-\delta)$, the optimal solution of (P_N) is *feasible* with respect to (P) by developing conservative convex approximations. A related approach was considered by [20].

(iii) *Sample-average approximation and integer programming approaches.* Under suitable concavity assumptions on $c(\bullet,\zeta)$ and convexity requirements on h, (P_N) can be efficiently resolved. However, this is often quite conservative. Instead, the following integer program ($\mathrm{P}_N^{\mathrm{int}}$) can serve as an approximation [11, 47].

$$\min_{\mathbf{x}\in\mathcal{X},\mathbf{z}\in\{0,1\}^N} \left\{ h(\mathbf{x}) \mid \sum_{j=1}^{N} z_j \leq \gamma N, \quad -c(\mathbf{x},\zeta_j) \leq M_j z_j, \ j = 1,\cdots,N \right\}, \qquad (\mathrm{P}_N^{\mathrm{int}})$$

where $M_j \triangleq \max_{\mathbf{x}\in\mathcal{X}}(-c(\mathbf{x},\zeta_j))$ and $\gamma \in (0,1)$. We observe that the second constraint ensures that only $\gamma\%$ of the N scenarios are satisfied. In fact, when $\gamma = \epsilon$ where ϵ is the parameter in (P), $\hat{v}_N$ and $\hat{\mathbf{x}}_N$, the optimal value and solution of ($\mathrm{P}_N^{\mathrm{int}}$), converge almost surely to $v^*(\epsilon)$ and $\mathcal{X}^*(\epsilon)$ as $N \to \infty$ [55]. Naturally, if $c(\cdot,\zeta)$ is a nonlinear function, ($\mathrm{P}_N^{\mathrm{int}}$) is a mixed-binary nonlinear program, a challenging instance of a discrete optimization problem.

(iv) *Smoothing-enabled Monte-Carlo sampling techniques.* Amongst the earliest approaches proposed by Norkin [53] utilized the characteristic function χ_C of a set C, defined as $\chi_C(\zeta) = 1$ if $\zeta \in C$ and 0 otherwise. This allowed for expressing f as $f(\mathbf{x}) = \int_{\mathbf{K}(x)} d\mathbb{P}(\zeta) = \int_{\mathbb{R}^d} \chi_{\mathbf{K}(\mathbf{x})}(\zeta) d\mathbb{P}(\zeta)$. Unfortunately, the function $\chi_C(\bullet)$ is discontinuous at the boundary of C, motivating the "smoothing" the characteristic function by using Steklov-Sobolev smoothing [53] (also referred to as convolution-based smoothing). Specifically, $\chi_C(\bullet)$ is approximated by its smoothing $\chi_C^{\epsilon}(\bullet)$, defined as $\chi_C^{\epsilon}(\zeta) = \int_{-\infty}^{\infty} \chi_C(\zeta + \epsilon\tau) p_s(\tau) d\tau$, where $p_s(\cdot)$ is a symmetric density. The resulting approximation f^{ϵ} is defined as $f^{\epsilon}(\mathbf{x}) = \int_{\mathbb{R}^d} \int_{-\infty}^{\infty} \chi_{\mathbf{K}(\mathbf{x})}^{\epsilon}(\zeta + \tau\epsilon) p_s(\tau) d\tau d\mathbb{P}(\zeta)$. Under suitably log-concavity assumptions, Norkin [53] developed a stochastic approximation framework for maximizing the approximation f^{ϵ}; However, there are no bounds relating the approximation and its true counterpart. An alternate simulation-based approach reliant on difference-of-convex programming [40] has been recently proposed for chance-constrained optimization. An alternate framework [20] uses a sampling and rejection framework in developing estimators convergent to feasible solutions. More recent efforts have focused on obtaining stationary points of the smoothed problem [29, 56]. More refined statements deriving convergence claims to Clarke stationary points have been provided in [28] by conducting a variational analysis of affine chance-constrained programs.

1.2 Applications

(1) *Robust portfolio selection problem.* Portfolio selection problems consider the specification of portfolio weights while maximizing a suitable risk/reward metric while meeting risk/reward requirements. Much of the research in this area emerges from the seminal work by Markowitz in the 50s [48]. Consider a portfolio with n risky assets, whose random returns are denoted by a random variable $\zeta = [\zeta_1, \zeta_2, \ldots, \zeta_n]^T$ with mean returns $\mu = [\mu_1, \mu_2, \cdots, \mu_n]^T$ and covariance Σ. Let the proportion of the portfolio to be invested in each asset be represented by $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$. It is assumed that no assets will be shorted and, hence, without loss of generality, the set of all possible portfolio allocations is given by $\mathcal{X} \triangleq \left\{\mathbf{x} : \mathbf{1}^T\mathbf{x} = 1 \text{ and } \mathbf{x} \geq 0\right\}$. We consider the robust portfolio selection problem (RPS) where the distribution of asset returns is not known but some of its properties available. Given a threshold α and an allocation $\mathbf{x}$, the distributionally robust risk associated with portfolio weights $\mathbf{x}$ is defined as $f_\alpha(\mathbf{x}) \triangleq \sup_{\upsilon \in \mathcal{H}} \mathbb{P}_\upsilon \left\{\zeta : \zeta^T\mathbf{x} \leq -\alpha\right\}$ where $\mathbb{P}_\upsilon$ denotes probability computed using the probability density function υ belonging to admissible class of density functions $\mathcal{H}$. This class of distributions contains all distributions whose density is radially decreasing and have level sets of the form $\{\delta : \delta^T\Sigma^{-1}\delta = r\}$ for some r. Consider the following distributionally robust portfolio selection problem with $\gamma > 0$.

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sup_{\upsilon\in\mathcal{H}} \mathbb{P}_\upsilon \left\{\zeta : \zeta^T\mathbf{x} \leq -\alpha\right\} \mid \mu^T\mathbf{x} = \gamma \right\}. \tag{RPS}$$

In prior work [13], it was shown via the following lemma that the supremum in (RPS) is achieved when υ is a uniform density over an ellipsoid.

Lemma 1. *Let the random vector ζ is of the form $\zeta = \boldsymbol{\mu} + \Delta$ where the distribution h for Δ is taken to be unknown but assumed to belong to the class $\mathcal{H}$ and $\boldsymbol{\mu} \in \{\zeta : \zeta^T\mathbf{x} \geq -\alpha\}$. Then $\sup_{\upsilon\in\mathcal{H}} Prob_\upsilon\{\zeta : \zeta^T\mathbf{x} \leq -\alpha\}$ is achieved when Δ has a uniform distribution over the set*

$$\mathcal{R} = \{\Delta \in \mathbb{R}^n : \Delta^T\Sigma^{-1}\Delta \leq r_{\max}\}, \tag{1}$$

where $r_{\max}$ represents the uniform bound on the support of the distribution of Δ.

Consequently, the robust portfolio selection problem (RPS) reduces to the probability minimization problem.

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{\mathbb{P}_\upsilon \left\{\zeta : \zeta^T\mathbf{x} \leq -\alpha\right\} \mid \mu^T\mathbf{x} = \gamma\right\},$$

where $\mathbb{P}_\upsilon$ denotes the uniform distribution over an ellipsoid. This motivates the consideration of such a portfolio selection problem in the numerics.

(2) *Set covering problems.* Consider a set covering problem [69] (closely related to a vehicle routing problem)

$$\max \ \{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}, \text{ where } f(\mathbf{x}) \ \triangleq \ \mathbb{P}\{\zeta \in \mathcal{K} \mid T\mathbf{x} \geq \zeta\}, \tag{2}$$

where $\mathcal{X} \triangleq \left\{\mathbf{x} \mid c^T\mathbf{x} \leq \beta, \mathbf{x} \geq 0\right\}$, and $T \in \mathbb{R}^{d\times n}$. The incidence matrix T represents a network with d arcs and n routes with ijth component denoted by t_{ij} where $t_{ij} \triangleq 1$ if route j contains arc i and is zero otherwise. Furthermore, ζ_i denotes the random demand on arc i where $\zeta \triangleq (\zeta_1, \cdots, \zeta_d)^T \in \mathcal{K}$, c_j represents the non-negative cost of operating route j, and β represents a given cost threshold.

1.3 Gaps, contributions, and outline

Gaps. The optimization of distribution functions remains a challenging problem. To the best of our knowledge, *there exist no efficient schemes equipped with asymptotic convergence or rate guarantees for probability maximization problems or their generalizations, i.e. chance-constrained problems.* This is both a testament to the difficulty of such a problem as well as a motivation for the present work which intends provides precisely such schemes with suitable convergence and rate guarantees for a subclass of problems. The key contributions of this work are as follows.
(I) Representation of (**PM**) *as a convex program.* By leveraging recent findings on non-Gaussian integrals of positively homogeneous functions (PHFs) [43, 50], we consider regimes where $c(\mathbf{x},\zeta) = 1 - |\zeta^T\mathbf{x}|^2$ (Setting A, Section 2.1) or $c(\mathbf{x},\zeta) = T\mathbf{x} - \zeta$ (Setting B, Section 2.2) where $\mathcal{K}$ is a compact and convex set, symmetric about the origin. In both settings, we show that $\mathbb{P}\{\zeta \mid \zeta \in \mathbf{K}(\mathbf{x})\}$ can be expressed as the expectation of a suitably defined Clarke regular integrand $F(\mathbf{x},\xi)$, i.e. $\mathbb{E}_{\tilde{p}}[F(\mathbf{x},\xi)]$ where $\tilde{p}(\xi)$ is either a suitably defined Gaussian density or its variant. We then proceed to show that the original problem is equivalent to the minimization of a convex function $g(\mathbb{E}_{\tilde{p}}[F(\bullet,\xi)])$ over a closed and convex set $\mathcal{X}$ where g is a suitably defined convex and smooth function. We refine these relationships when $\mathcal{K}$ is either an ℓ_p ball or an ellipsoid centered at the origin (Setting A) or loses symmetry (Setting B).
(II) Regularized variance-reduced stochastic approximation (**r-VRSA**) *scheme.* The resulting convex program is an instance of a compositional stochastic optimization problem where unbiased first-order oracles are unavailable. In Section 3, we present a regularized variance-reduced SA scheme that combines variance-reduction (to accommodate bias) with regularization. The resulting scheme is characterized by a rate of convergence of $\mathcal{O}(1/k^{(1/2-a)})$ and $\mathcal{O}(1/K^{1/2})$ for diminishing and constant steplengths (the latter requiring the specification of a simulation length K) while the sample-complexity to achieve $\epsilon-$optimality, i.e. $\mathbb{E}[h(\mathbf{x}^*) - h(\mathbf{x}_K)] \leq \epsilon$ is $\mathcal{O}(1/\epsilon^{6+\delta})$ and $\mathcal{O}(1/\epsilon^6)$, respectively. While the rate of convergence matches the optimal rate for subgradient methods for convex programs, the sample-complexity is worse than the canonical $\mathcal{O}(1/\epsilon^2)$. The latter is unsurprising since we do not have access to unbiased oracles. It is worth emphasizing that this appears to be one of the first schemes with asymptotic convergence and rate guarantees for a class of non-trivial probability maximization problems. Apart from this, we provide some background in Section 1.1 and numerical experiments are discussed in Section 4.

Notation. We conclude this section with a review of *notation.* The sets of real numbers, non-negative real numbers, non-negative integers, and positive integers are denoted by $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{N}$, and $\mathbb{Z}$, respectively. The Euclidean norm of column vectors $\mathbf{x} \in \mathbb{R}^n$ is denoted by $\|\mathbf{x}\|$, while the spectral norm of $\mathbf{A} \in \mathbb{R}^{m\times n}$ is given by $\|\mathbf{A}\| = \max\{\|\mathbf{A}\mathbf{x}\| \colon \|\mathbf{x}\| \leq 1\}$. The n-by-n identity matrix is written as $\mathbf{I}_n$, and the m-by-n zero matrix as $\mathbf{0}_{m\times n}$. The projection onto the set X is denoted by Π_X, that is, $\Pi_X(y) = \operatorname{argmin}_{x\in X}\|x - y\|$. Finally, unless mentioned otherwise, any missing proofs are provided in the appendix.

2 An expectation-valued convex framework

Throughout this paper, we consider (**PM**) where

$$\mathbf{K}(\mathbf{x}) \triangleq \{\zeta \in \mathcal{K} \colon c(\mathbf{x},\zeta) \geq 0\}\,. \tag{3}$$

We consider two sets of regimes based on the choice of $\mathcal{K}$ and $c(\mathbf{x},\zeta)$. In addition, we impose an assumption on $\mathcal{K}$ and a distributional assumption on ζ.

Assumption 1 (Assumptions on $\mathcal{K}$ and $\mathcal{X}$)**.** *The random variable ζ is uniformly distributed on the set $\mathcal{K}$ where $\mathcal{K}$ is a compact and convex set in $\mathbb{R}^n$ and is symmetric about the origin. The set $\mathcal{X}$ is closed, convex, and bounded.*

Setting: A

The constraint $c(\mathbf{x}, \zeta)$ in (3) is defined as

$$c(\mathbf{x}, \zeta) \triangleq 1 - |\zeta^\mathsf{T} \mathbf{x}|^m, \tag{4}$$

where $m \in \mathbb{R}_+$ and $\zeta \in \mathbb{R}^n$ is a random variable. □

Setting: B

The constraint $c(\mathbf{x}, \zeta)$ in (3) is defined as

$$c(\mathbf{x}, \zeta) \triangleq T\mathbf{x} - \zeta, \tag{5}$$

where $T \in \mathbb{R}^{d \times n}$ and $\zeta \in \mathbb{R}^d$ is a random variable. □

We observe that Setting A can capture problems such as the portfolio optimization problem described in Section 1.2. Without loss of generality, $m = 2$ in this paper. Further, in such problems, $f(\mathbf{0}) = 1$ and $\lim_{\|\mathbf{x}\| \to \infty} f(\mathbf{x}) = 0$, where $f(\mathbf{x}) \triangleq \mathbb{P}(\{\zeta \mid \zeta \in \mathbf{K}(\mathbf{x})\}$. Setting B assumes relevance when considering chance constraints defined using polyhedral constraints with uncertain right-hand sides. An instance of such a problem is the set covering problem described in Section 1.2. We qualify problems in Settings A and B as ($\mathbf{PM}_A$) and ($\mathbf{PM}_B$), respectively. Before proceeding, we recall two definitions of relevance.

Definition 1 (Log-concavity, positive homogeneity)**.** *A function $f : \mathbb{R}^d \to [0, \infty)$ is log-concave if for any $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$, $f((1-\lambda)x + \lambda y) \geq [f(x)]^{1-\lambda}[f(y)]^\lambda$. A continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is called positively homogeneous function of degree $p \in \mathbb{R}$ if $f(\alpha x) = \alpha^p f(x)$ for all $\alpha > 0$ and all $x \in \mathbb{R}^d$.*

Definition 2 (Minkowski Functional)**.** *Let the set $\mathcal{K} \subset \mathbb{R}^n$. Then, the Minkowski functional associated with the set $\mathcal{K}$, denoted by $\|\cdot\|_\mathcal{K}$, is defined as $\|\zeta\|_\mathcal{K} \triangleq \inf\{t > 0 : \zeta/t \in \mathcal{K}\}$ for all $\zeta \in \mathbb{R}^n$.*

Note that $\|\cdot\|_\mathcal{K}$ defines a norm when $\mathcal{K}$ is compact, convex and symmetric. For instance, if $\mathcal{K}$ is the unit ball in $\mathbb{R}^n$, then the Minkowski functional reduces to $\|\cdot\|_2$ in $\mathbb{R}^n$ i.e. $\|\zeta\|_\mathcal{K} = \|\zeta\|_2$. In the remainder of this section, after providing some background in Section 1.1, we proceed to show that the function f, defined in (**PM**), is equivalent to the expectation of a nonsmooth integrand. In fact, we prove that this integrand is Clarke regular and the reciprocal of f is a convex function for setting A (Section 2.1) while the negative log-transformation of f is a convex function in setting B (Section 2.2).

2.1 Expectation-valued convex representations for Setting A.

We begin by recalling that $\mathbb{P}\{\zeta \mid \zeta \in \mathbf{K}(\mathbf{x})\}$ can be rewritten as

$$f(\mathbf{x}) = \mathbb{P}\{\zeta \mid \zeta \in \mathbf{K}(\mathbf{x})\} = \frac{1}{\mathrm{Vol}(\mathcal{K})} \int_{\mathbf{K}(\mathbf{x})} 1 \, d\zeta, \tag{6}$$

where the last equality follows from Assumption 1 and $\mathrm{Vol}(\mathcal{K})$ denotes the volume of the set $\mathcal{K}$. We now show that (6) can be expressed as an expectation with respect to prescribed probability measure. Recall that a function r is a PHF of degree m where $r(\zeta) = \max\{r_1(\zeta), \ldots, r_\ell(\zeta)\}$ if $r_1, \ldots, r_\ell$ are PHFs of degree m. Lemma 2 provides conditions under which the integral of a PHF over a suitable set is equal to another integral which is an expectation over a suitably defined measure.

Lemma 2. [43, Cor. 2.3] *Let h be a positively homogenous function of degree p and let $r_1, \ldots, r_\ell$ be positively homogeneous functions (PHFs) of degree $0 \neq t \in \mathbb{R}$. Let Λ be a bounded set defined as $\Lambda \triangleq \{\zeta : r_k(\zeta) \leq 1, k = 1, \ldots, \ell\}$. If $\int_{\mathbb{R}^n} |h(\xi)| e^{-\max\{r_1(\xi), \ldots, r_\ell(\xi)\}}\ d\xi < \infty$, then the following holds.*

$$\int_\Lambda h(\zeta)\ d\zeta = \frac{1}{\Gamma(1 + (n+p)/t)} \int_{\mathbb{R}^n} h(\xi) e^{-\max\{r_1(\xi), \ldots, r_\ell(\xi)\}}\ d\xi.$$

We now show that f is given by the expectation with respect to $\tilde{p}(\xi)$, the density function of a suitably defined random variable dependent on the choice of $\mathcal{K}$.

Theorem 1 (Representation of f as expectation for general $\mathcal{K}$)**.** *Consider* (**PM**)*. Suppose Assumption 1 holds, where c is defined as (4) and $\mathbf{K}$ is defined as (3). Then the following equality holds.*

$$\mathbb{P}\left\{\zeta : \zeta \in \mathbf{K}(\mathbf{x})\right\} = \mathbb{E}_{\tilde{p}(\xi)}[F(\mathbf{x}, \xi)] = \int_{\mathbb{R}^n} F(\mathbf{x}, \xi) \tilde{p}(\xi) d\xi, \ \ \textit{where} \tag{7}$$

$$F(\mathbf{x}, \xi) \triangleq \mathcal{C}_{\mathcal{K}} (2\pi)^{n/2} e^{-\max\{|\xi^\mathsf{T}\mathbf{x}|^2, \|\xi\|_{\mathcal{K}}^2\} + \frac{\|\xi\|_{\mathcal{K}}^2}{2}}, \mathcal{C} \triangleq \tfrac{1}{\mathrm{Vol}(\mathcal{K})} \tfrac{1}{\Gamma(1+n/2)}, \tag{8}$$

$$\tilde{p}(\xi) \triangleq \tfrac{1}{(2\pi)^{n/2} D_{\mathcal{K}}} e^{\frac{-\|\xi\|_{\mathcal{K}}^2}{2}}, \tag{9}$$

$D_{\mathcal{K}}$ is a positive scalar such that $\int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2} D_{\mathcal{K}}} e^{\frac{-\|\xi\|_{\mathcal{K}}^2}{2}} = 1$, and $\mathcal{C}_{\mathcal{K}} \triangleq \mathcal{C} D_{\mathcal{K}}$.

Proof. We begin by noting that $\mathbf{K}(\mathbf{x})$ can be expressed as:

$$\mathbf{K}(\mathbf{x}) = \{\zeta : \zeta \in \mathcal{K}\} \cap \{\zeta : |\zeta^\mathsf{T}\mathbf{x}| \leq 1\}$$

Since the set $\mathcal{K}$ is convex, compact, and symmetric, the Minkowski functional of $\mathcal{K}$ defines a norm, and hence, it is a PHF. Moreover, by the definition of the Minkowski functional we have $\zeta \in \mathcal{K} \Leftrightarrow \|\zeta\|_{\mathcal{K}} \leq 1$. By using this definition, we may rewrite $\mathbf{K}(x)$ as follows.

$$\mathbf{K}(\mathbf{x}) = \{\zeta : |\zeta^\mathsf{T}\mathbf{x}|^2 \leq 1\} \cap \{\zeta \colon \|\zeta\|_{\mathcal{K}}^2 \leq 1\} = \left\{\zeta : \max\{|\zeta^\mathsf{T}\mathbf{x}|^2, \|\zeta\|_{\mathcal{K}}^2\} \leq 1\right\}.$$

Since $|\zeta^\mathsf{T}\mathbf{x}|^2$ and $\|\zeta\|_{\mathcal{K}}^2$ are both PHFs of degree 2, then $g(\mathbf{x}, \bullet)$ is also a PHF of degree 2 where $g(\mathbf{x}, \zeta)$ is defined as $g(\mathbf{x}, \zeta) \triangleq \max\{|\zeta^\mathsf{T}\mathbf{x}|^2, \|\zeta\|_{\mathcal{K}}^2\}$. By selecting $h(\zeta) = 1$ and $\Lambda = \mathbf{K}(\mathbf{x})$, we may invoke Lemma 2, leading to the following equality.

$$f(\mathbf{x}) = \frac{1}{\mathrm{Vol}(\mathcal{K})} \int_{\mathbf{K}(x)} 1\ d\zeta = \frac{1}{\mathrm{Vol}(\mathcal{K})} \frac{1}{\Gamma(1 + n/2)} \int_{\mathbb{R}^n} e^{-g(\mathbf{x}, \xi)}\ d\xi, \tag{10}$$

whenever $\int_{\mathbb{R}^n} e^{-g(\mathbf{x}, \xi)}\ d\xi$ is finite. In fact, the expression (10) can be written as

$$f(\mathbf{x})\ = \mathcal{C} \int_{\mathbb{R}^n} \left((2\pi)^{n/2} e^{-\max\{|\xi^\mathsf{T}\mathbf{x}|^2, \|\xi\|_{\mathcal{K}}^2\} + \frac{\|\xi\|_{\mathcal{K}}^2}{2}} \right) \left(\tfrac{1}{(2\pi)^{n/2}} e^{\frac{-\|\xi\|_{\mathcal{K}}^2}{2}} \right)\ d\xi$$

$$= \int_{\mathbb{R}^n} \underbrace{\left(\mathcal{C}_{\mathcal{K}}(2\pi)^{n/2} e^{-\max\{|\xi^{\mathsf{T}}\mathbf{x}|^2, \|\xi\|_{\mathcal{K}}^2\} + \frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right)}_{\triangleq F(\mathbf{x},\xi)} \underbrace{\left(\frac{1}{D_{\mathcal{K}}(2\pi)^{n/2}} e^{\frac{-\|\xi\|_{\mathcal{K}}^2}{2}}\right)}_{\triangleq \tilde{p}(\xi)} \, d\xi$$

$$= \int_{\mathbb{R}^n} F(\mathbf{x},\xi)\ \tilde{p}(\xi)\ d\xi = \mathcal{C}\ \mathbb{E}_{\tilde{p}}[F(\mathbf{x},\xi)],\ \text{where } \mathcal{C} \triangleq \tfrac{1}{\mathrm{Vol}(\mathcal{K})} \tfrac{1}{\Gamma(1+n/2)},$$

$\tilde{p}(\xi)$ denotes the density, $D_{\mathcal{K}}$ is such that $\int_{\mathbb{R}^n} \tilde{p}(\xi) d\xi = 1$, and $\mathcal{C}_{\mathcal{K}} \triangleq \mathcal{C} D_{\mathcal{K}}$. □

Next, we examine the convexity properties of a related problem, given by (11).

$$\min_{\mathbf{x} \in \mathcal{X}}\ h(\mathbf{x})\ \triangleq\ \frac{1}{f(\mathbf{x})}. \tag{11}$$

Crucial to this claim is the leveraging of a result provided in [16] which allows for claiming that the reciprocal of $\mathbb{P}\{\zeta \in \mathbf{K}(\mathbf{x})\}$ is a convex function in $\mathbf{x}$ when ζ satisfies a suitable requirement.

Theorem 2 (Transformation of ($\mathbf{PM}_A$) to convex program)**.** *Suppose the function f, $\mathbf{K}(\mathbf{x})$, and $\mathcal{X}$ are as defined in* ($\mathbf{PM}$). *Suppose $f(\mathbf{x}) \in [\epsilon, 1]$ for $\mathbf{x} \in \mathcal{X}$ and $\epsilon > 0$ and h is defined such that $h(\mathbf{x}) = 1/f(\mathbf{x})$. Then the following hold.*

(a) h is convex in $\mathbf{x}$ over $\mathcal{X}$ where $h(\mathbf{x}) \triangleq 1/f(\mathbf{x})$.

(b) A global maximizer of ($\mathbf{PM}$) *is a global minimizer of* (11)*.*

Before proceeding, we provide a lemma for computing the maximal value of $u^c e^{-u}$ where c is a positive integer.

Lemma 3. *Consider the function $u^c e^{-u}$ defined on $u \in \mathbb{R}_+$ where $c \geq 1$ and $c \in \mathbb{Z}_+$. Then we have that $\max\limits_{u \geq 0} u^c e^{-u} = \frac{c^c}{e^c}$ and $\arg\max\limits_{u \geq 0} u^c e^{-u} = c$.*

Note that since $F(\bullet, \xi)$ is not necessarily convex, we cannot employ subdifferentials of F. Instead, we begin by recalling some key elements of Clarke's nonsmooth calculus and start by providing the definition of the Clarke generalized gradient of a function h by leveraging its directional derivatives.

Definition 3 (**Directional derivatives and Clarke generalized gradient [27]**)**.** The directional derivative of h at $\mathbf{x}$ in a direction v is defined as

$$h^\circ(\mathbf{x}, v) \triangleq \limsup_{\mathbf{y} \to \mathbf{x}, t \downarrow 0} \left(\frac{h(\mathbf{y} + tv) - h(\mathbf{y})}{t}\right). \tag{12}$$

The Clarke generalized gradient at $\mathbf{x}$ can then be defined as

$$\partial h(\mathbf{x}) \triangleq \left\{\xi \in \mathbb{R}^n \mid h^\circ(\mathbf{x}, v) \geq \xi^T v, \quad \forall v \in \mathbb{R}^n\right\}. \tag{13}$$

In other words, $h^\circ(\mathbf{x}, v) = \sup\limits_{g \in \partial h(\mathbf{x})} g^T v$.

If h is C^1 at $\mathbf{x}$, the Clarke generalized gradient reduces to the standard gradient, i.e. $\partial h(\mathbf{x})$ is a singleton at $\nabla_{\mathbf{x}} h(\mathbf{x})$. We now review some properties of $\partial h(\mathbf{x})$. In particular, if h is locally Lipschitz on an open set $\mathcal{C}$ containing $\mathcal{X}$, then h is differentiable almost everywhere on $\mathcal{C}$ by Rademacher's theorem [27]. Suppose $\mathcal{C}_h$ denotes the set of points where h is not differentiable. We now provide some properties of the Clarke generalized gradient.

Proposition 1 (**Properties of Clarke generalized gradients [27]**)**.** Suppose h is L_0-Lipschitz continuous on $\mathbb{R}^n$. Then the following hold.

(i) $\partial h(\mathbf{x})$ is a nonempty, convex, and compact set and $\|u\| \leq L_0$ for any $u \in \partial h(\mathbf{x})$.

(ii) h is differentiable almost everywhere.

(iii) ∂h is an upper semicontinuous map defined as

$$\partial h(\mathbf{x}) \triangleq \text{conv}\left\{u \mid u = \lim_{k\to\infty} \nabla_{\mathbf{x}} h(\mathbf{x}_k), \mathcal{C}_h \not\ni \mathbf{x}_k \to \mathbf{x}\right\}.$$

To employ the Clarke generalized gradient, we require that the function be at least locally Lipschitz. We proceed to prove that $F(\bullet, \xi)$ satisfies this requirement. We further show that this result paves the way for showing that we can interchange the Clarke subdifferential and the expectation operator.

Lemma 4. *Consider the function $F(\bullet, \xi)$ defined as*

$$F(\mathbf{x}, \xi) = \begin{cases} \left(\mathcal{C}_{\mathcal{K}}(2\pi)^{n/2} e^{-|\xi^{\mathsf{T}}\mathbf{x}|^2 + \frac{\|\xi\|^2_{\mathcal{K}}}{2}}\right) & \xi \in \Xi_1(\mathbf{x}) \triangleq \left\{\xi \mid |\xi^{\mathsf{T}}\mathbf{x}|^2 > \|\xi\|^2_{\mathcal{K}}\right\} \\ \left(\mathcal{C}_{\mathcal{K}}(2\pi)^{n/2} e^{-\max\{|\xi^{\mathsf{T}}\mathbf{x}|^2, \|\xi\|^2_{\mathcal{K}}\} + \frac{\|\xi\|^2_{\mathcal{K}}}{2}}\right) & \xi \in \Xi_0(\mathbf{x}) \triangleq \left\{\xi \mid |\xi^{\mathsf{T}}\mathbf{x}|^2 = \|\xi\|^2_{\mathcal{K}}\right\} \\ \left(\mathcal{C}_{\mathcal{K}}(2\pi)^{n/2} e^{-\frac{\|\xi\|^2_{\mathcal{K}}}{2}}\right). & \xi \in \Xi_2(\mathbf{x}) \triangleq \left\{\xi \mid |\xi^{\mathsf{T}}\mathbf{x}|^2 < \|\xi\|^2_{\mathcal{K}}\right\} \end{cases}$$

Then the following hold.
(a) $F(\bullet, \xi)$ is locally Lipschitz for every ξ.
(b) $F(\bullet, \xi)$ is a Clarke regular function for almost every $\xi \in \mathbb{R}^n$.
(c) For any $\mathbf{x} \in \mathbb{R}^n$, $\partial\mathbb{E}[F(\mathbf{x}, \xi)] = \mathbb{E}[\partial F(\mathbf{x}, \xi)]$.

Proof. (a) This follows by observing that $F(\bullet, \xi)$ is C^1 when $\xi \in \Xi_1(\mathbf{x}) \cup \Xi_2(\mathbf{x})$ and piecewise C^1 if $\xi \in \Xi_0(\mathbf{x})$. Therefore $F(\bullet, \xi)$ is locally Lipschitz for every $\xi \in \Xi$ [67, Cor. 4.1.1.].
(b) Since $\Xi_0(\mathbf{x})$ is a lower-dimensional set in $\mathbb{R}^n$, we have that $\Xi_1(\mathbf{x}) \cup \Xi_2(\mathbf{x}) = \mathbb{R}^n$. Therefore for almost every $\xi \in \mathbb{R}^n$, we have that $F(\bullet, \xi)$ is C^1. Consequently, $F(\bullet, \xi)$ is a Clarke regular function for almost every ξ.
(c) Since $F(\bullet, \xi)$ is Clarke regular for almost every $\xi \in \mathbb{R}^n$, by [18, Theorem 3.4.], we have that $\partial\mathbb{E}[F(\mathbf{x}, \xi)] = \mathbb{E}[\partial F(\mathbf{x}, \xi)]$. □

Computational schemes, particularly via stochastic approximation, rely on boundedness of $F(\bullet, \xi)$ and $G(\mathbf{x}, \xi)$ where we denote an element of $\partial F(\mathbf{x}, \xi)$ by $G(\mathbf{x}, \xi)$, i.e. $G(\mathbf{x}, \xi) \in \partial F(\mathbf{x}, \xi)$.

Proposition 2 (Properties of $F(\mathbf{x}, \xi)$ and $G(\mathbf{x}, \xi)$ under general $\mathcal{K}$)**.** *Consider the function f in* (**PM**) *and suppose Assumption 1 holds. Suppose $c(\mathbf{x}, \zeta)$, $F(\mathbf{x}, \xi)$, and $\tilde{p}(\xi)$ are defined as (4),* (8), *and* (9), *respectively and $G(\mathbf{x}, \xi) \in \partial F(\mathbf{x}, \xi)$ for any $\mathbf{x} \in \mathcal{X}$. Then the following hold.*
(a) For any $\mathbf{x} \in \mathcal{X}$, $|F(\mathbf{x}, \xi)|^2 \leq \mathcal{C}^2_{\mathcal{K}}(2\pi)^n$ for every $\xi \in \mathbb{R}^n$.
(b) For any $\mathbf{x} \in \mathcal{X}$, $\mathbb{E}[\|G(\mathbf{x}, \xi)\|^2] \leq \frac{\mathcal{C}^2_{\mathcal{K}}(2\pi)^n}{e}\mathbb{E}_{\tilde{p}}[\|\xi\|^2]$.

We may specialize this representation and the bounds to regimes where $\mathcal{K}$ is an ℓ_p-ball in $\mathbb{R}^n$ via the following Lemma. In addition, we recall that the density of a multivariate Gaussian with independent components, each with mean zero and variance σ^2, has a density given by $\tilde{p}(\xi)$ defined as

$$\tilde{p}(\xi) \triangleq \frac{e^{-\frac{\|\xi\|_2^2}{2\sigma^2}}}{(2\pi\sigma)^{n/2}}, \text{ where } \int_{\mathbb{R}^n} \tilde{p}(\xi) d\xi = 1.$$

Consequently, $D_{\mathcal{K}} = 1$ and $\mathcal{C}_{\mathcal{K}} = C$. We rely on the following standard lemma.

Lemma 5. *Let the ℓ_p-norm of a vector $x \in \mathbb{R}^n$ be defined as $\|x\|_p \triangleq (\sum_{i=1}^n \mid x_i \mid^p)^{1/p}$. For any $1 \leq a < b$, there exists a scalar $\beta \triangleq n^{(1/a-1/b)}$ such that for every $x \in \mathbb{R}^n$, $\|x\|_b \leq \|x\|_a \leq \beta \|x\|_b$.*

Proposition 3 (Representation and boundedness of $F(\mathbf{x}, \xi)$ and $G(\mathbf{x}, \xi)$ when $\mathcal{K}$ is an ℓ_p ball)**.** *Consider the function f in* (**PM**)*. Suppose Assumption 1 holds and $c(\mathbf{x}, \xi)$ is defined as (4). Suppose $\mathcal{K}$ is an ℓ_p-ball in $\mathbb{R}^n$ where $p \geq 1$. Then the following hold.*
(a) $f(\mathbf{x}) = \mathbb{E}_{\tilde{p}}[F(\mathbf{x}, \xi)]$, where

$$F(\mathbf{x}, \xi) \triangleq \left(\mathcal{C}(2\pi\sigma^2)^{n/2} e^{-\max\{|\xi^\intercal \mathbf{x}|^2, \|\xi\|_p^2\} + \frac{\|\xi\|_2^2}{2\sigma^2}} \right), \mathcal{C} \triangleq \frac{1}{\text{Vol}(\mathcal{K})} \frac{1}{\Gamma(1+n/2)}, \text{ and,}$$

$$\tilde{p}(\xi) \triangleq \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}}, \quad \text{where } \sigma^2 \triangleq \begin{cases} n^{1/2-1/p}, & p \geq 2 \\ 1. & 1 \leq p < 2 \end{cases}$$

(b) For any $\mathbf{x} \in \mathcal{X}$, $|F(\mathbf{x}, \xi)|^2 \leq \mathcal{C}^2 (2\pi\sigma^2)^n$ for every $\xi \in \mathbb{R}^n$.
(c) For every $\mathbf{x} \in \mathcal{X}$, $\mathbb{E}[\|G(\mathbf{x}, \xi)\|^2] \leq e^{-1} \mathcal{C}^2 (2\pi\sigma^2)^n \mathbb{E}_{\tilde{p}}[\|\xi\|^2]$, where $G(\mathbf{x}, \xi) \in \partial F(\mathbf{x}, \xi)$ for any $\mathbf{x} \in \mathcal{X}$ and $\xi \in \mathbb{R}^n$.

Next, we consider the regime where $\mathcal{K} \triangleq \mathcal{K}_{\mathcal{E}}$ is an ellipsoid in $\mathbb{R}^n$, defined as

$$\mathcal{K}_{\mathcal{E}} \triangleq \left\{ \zeta \in \mathbb{R}^n \mid \zeta^\intercal U^\intercal \Sigma^{-1} U \zeta \leq 1 \right\}, \tag{14}$$

where $U \in \mathbb{R}^{n \times n}$ is an orthogonal matrix whose columns represent unit vectors along the principal axes of the ellipsoid and Σ^{-1} is a positive diagonal matrix with the ith diagonal element denoted by $1/\sigma_{ii}^2$. By defining $\eta = \Sigma^{-1/2} U \zeta$ or $U^\intercal \Sigma^{1/2} \eta = \zeta$, we may observe that $\mathcal{K}_2 = \{\eta : \|\eta\|_2^2 \leq 1\}$, i.e. $\mathcal{K}_{\mathcal{E}}$ can be transformed to $\mathcal{K}_2$. We now prove that there is an equivalence between ($\mathbf{PM}_A^{\mathcal{E}}$) and its transformed counterpart ($\mathbf{PM}_A^2$).

Lemma 6 (Equivalence between ($\mathbf{PM}_A^{\mathcal{E}}$) and ($\mathbf{PM}_A^2$))**.** *Consider the function f in* (**PM**)*. Suppose Assumption 1 holds and c is defined as (4). Suppose $\mathcal{K}_{\mathcal{E}}$ is an ellipsoid in $\mathbb{R}^n$, defined in* (14)*. Then $\mathbf{x}$ is a solution of* ($\mathbf{PM}_A^{\mathcal{E}}$)

$$\max_{\mathbf{x} \in \mathcal{X}} \quad f(\mathbf{x}) \triangleq \mathbb{P} \left\{ \zeta \mid \zeta \in \mathcal{K}_{\mathcal{E}}, |\zeta^\intercal \mathbf{x}| \leq 1 \right\} \tag{$\mathbf{PM}_A^{\mathcal{E}}$}$$

if and only if $\mathbf{x}$ is a solution of ($\mathbf{PM}_A^2$)*, defined as*

$$\max_{\mathbf{x} \in \mathcal{X}} \quad g(\mathbf{x}) \triangleq \mathbb{P} \left\{ \eta \mid \eta \in \mathcal{K}_2, |\eta^\intercal \Sigma^{1/2} U \mathbf{x}| \leq 1 \right\}. \tag{$\mathbf{PM}_A^2$}$$

We may then solve ($\mathbf{PM}_A^2$) by adding a new variable $\mathbf{y}$, defined using the linear constraint $\mathbf{y} = \Sigma^{1/2} U \mathbf{x}$, and restate ($\mathbf{PM}_A^2$) as follows.

$$\max_{\mathbf{x} \in \mathcal{X}, \mathbf{y} = \Sigma^{1/2} U \mathbf{x}} \mathbb{P} \left\{ \eta \mid \eta \in \mathcal{K}_2, |\eta^\intercal \mathbf{y}| \leq 1 \right\}. \tag{15}$$

By our representation theorem (Theorem 1), we may then claim the following from Lemma 6 (proof omitted).

Proposition 4 (Representation of ($\mathbf{PM}_A^{\mathcal{E}}$) via Gaussian transformation)**.** *Consider the function f in* (**PM**)*. Suppose Assumption 1 holds and $c(\mathbf{x}, \zeta)$ is defined as (4). Suppose $\mathcal{K}$ is an ellipsoid in $\mathbb{R}^n$, defined in* (14)*. Suppose* $(\mathbf{x}, \mathbf{y})$ *is a solution of* (15)*. Then $\mathbf{x}$ is a solution to* ($\mathbf{PM}_A^{\mathcal{E}}$)*.*

2.2 Expectation-valued convex representations for Setting (B)

In this section, we consider the regime where $c(\mathbf{x}, \zeta) = T\mathbf{x} - \zeta$, $T \in \mathbb{R}^{d \times n}$, and $\zeta \in \mathbb{R}^d$. We denote the ith component of c by c_i while the ith row of T is denoted by $T_{i,\bullet}$. The following proposition articulates a convex counterpart of (**PM**).

Theorem 3 (Transformation of ($\mathbf{PM}_B$) to convex program)**.** *Consider the problem* (**PM**)*. Suppose Assumptions 1 holds and* $c(\mathbf{x}, \zeta)$ *is defined as* (5)*. Suppose* $h(\mathbf{x}) \triangleq -\log\left(f(\mathbf{x})\right).$ *Then the following hold.*

(a) h *is convex on* $\mathcal{X}$*.*

(b) $\mathbf{x}^*$ *minimizes* h *over* $\mathcal{X}$ *if and only if* $\mathbf{x}^*$ *is a maximizer of* f *over* $\mathcal{X}$*.*

Proof. (a) We observe that $f(\mathbf{x}) = \mathbb{P}\left(\zeta \mid T\mathbf{x} \geq \zeta\right) = H_\zeta(T\mathbf{x})$, where H_ζ is the probability distribution function of the random vector ζ. If $T\mathbf{x} \in \mathbb{R}^d$, H_ζ is a log-concave function [63, Theorem 4.2.4.], implying that $\log(f)$ is a concave function. Hence, it follows that h is a convex function where $h(\mathbf{x}) = -\log(f(\mathbf{x}))$.
(b) Suppose $\mathbf{x}^*$ is a maximizer of f over $\mathcal{X}$. Then $f(\mathbf{x}) \leq f(\mathbf{x}^*)$ for all $\mathbf{x} \in \mathcal{X}$. Since $h = -\log(f)$ is a monotonically decreasing function for $f(\mathbf{x}) > 0$ we have that $f(\mathbf{x}) \leq f(\mathbf{x}^*)$ if and only if $-\log(f(\mathbf{x})) \geq -\log(f(\mathbf{x}^*))$ for all $\mathbf{x} \in \mathcal{X}$. Consequently, $\mathbf{x}^*$ is a minimizer of h over $\mathcal{X}$. □

Proposition 5 (Representation of f as expectation for symmetric $\mathcal{K}$)**.** *Consider the problem* (**PM**)*. Suppose Assumptions 1 holds,* $c(\mathbf{x}, \zeta)$ *is defined as* (5)*, and for all* $\mathbf{x} \in \mathcal{X}$*,* $T_{i,\bullet}\mathbf{x} \geq \delta$ *for every* i *and for some* $\delta > 0$*. Then the following equality holds.*

$$\mathbb{P}\left\{\zeta : \zeta \in \mathbf{K}(\mathbf{x})\right\} = \int_{\mathbb{R}^d} F(\mathbf{x}, \xi)\tilde{p}(\xi)d\xi = \mathbb{E}_{\tilde{p}(\xi)}[F(\mathbf{x}, \xi)], \quad \textit{where} \tag{16}$$

$$F(\mathbf{x}, \xi) \triangleq \mathcal{C}_{\mathcal{K}}(2\pi)^{d/2} e^{-g(\mathbf{x},\xi) + \frac{\|\xi\|_{\mathcal{K}}^2}{2}}, \mathcal{C} \triangleq \frac{1}{\mathrm{Vol}(\mathcal{K})}\frac{1}{\Gamma(1+d/2)}, \tilde{p}(\xi) \triangleq \frac{1}{(2\pi)^{d/2} D_{\mathcal{K}}} e^{\frac{-\|\xi\|_{\mathcal{K}}^2}{2}},$$

$$\textit{where } g(\mathbf{x}, \xi) \triangleq \max\left\{\left(\frac{\max(\xi_1, 0)}{T_{1,\bullet}\mathbf{x}}\right)^2, \cdots, \left(\frac{\max(\xi_d, 0)}{T_{d,\bullet}\mathbf{x}}\right)^2, \|\xi\|_{\mathcal{K}}^2\right\}, \textit{ and}$$

$D_{\mathcal{K}}$ *is a positive scalar such that* $\int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2} D_{\mathcal{K}}} e^{\frac{-\|\xi\|_{\mathcal{K}}^2}{2}} = 1$*, and* $\mathcal{C}_{\mathcal{K}} \triangleq \mathcal{C} D_{\mathcal{K}}$*.*

Proof. In this instance, the set $\mathbf{K}(\mathbf{x})$ is defined as following:

$$\mathbf{K}(\mathbf{x}) \triangleq \left\{\zeta : \zeta \in \mathcal{K}\right\} \cap \left\{\zeta : \zeta \leq T\mathbf{x}\right\}.$$

Since the set $\mathcal{K}$ is convex, compact, and symmetric, the Minkowski functional of $\mathcal{K}$ defines a norm, and hence, it is a PHF. Moreover, by the definition of the Minkowski functional we have $\zeta \in \mathcal{K} \Leftrightarrow \|\zeta\|_{\mathcal{K}} \leq 1$. By using this definition, we may rewrite $\mathbf{K}(x)$ as follows, where by assumption $T_{i,\bullet}\mathbf{x} \geq \delta > 0$ for all $\mathbf{x}$ implying that $T_{i,\bullet}\mathbf{x} > 0$ for $i = 1, \cdots, d$.

$$\begin{aligned}
\mathbf{K}(\mathbf{x}) &= \left\{\zeta : \|\zeta\|_{\mathcal{K}} \leq 1\right\} \bigcap \left\{\zeta : \bigcap_{i=1}^{d} \frac{\max\left\{\zeta_i, 0\right\}}{T_{i,\bullet}\mathbf{x}} \leq 1\right\} \\
&= \left\{\zeta : \|\zeta\|_{\mathcal{K}}^2 \leq 1\right\} \bigcap \left\{\zeta : \bigcap_{i=1}^{d} \left(\frac{\max\left\{\zeta_i, 0\right\}}{T_{i,\bullet}\mathbf{x}}\right)^2 \leq 1\right\}
\end{aligned}$$

$$= \left\{ \zeta : \max\left\{ \|\zeta\|_{\mathcal{K}}^2, \left(\frac{\max\{\zeta_1,0\}}{T_{1,\bullet}\mathbf{x}}\right)^2, \cdots, \left(\frac{\max\{\zeta_d,0\}}{T_{d,\bullet}\mathbf{x}}\right)^2 \right\} \le 1 \right\},$$

where the squared expression is employed to obtain a PHF of degree 2. Since $g_i(\mathbf{x},\zeta) \triangleq \left(\frac{\max\{\zeta_i,0\}}{T_{i,\bullet}\mathbf{x}}\right)^2$ for $i = 1,\ldots,d$ and $g_{d+1}(\mathbf{x},\zeta) \triangleq \|\zeta\|_{\mathcal{K}}^2$ are PHFs of degree 2, then $g(\mathbf{x},\zeta) \triangleq \max\{g_1(\mathbf{x},\zeta),\ldots,g_{d+1}(\mathbf{x},\zeta)\}$ is positively homogeneous of degree 2. By selecting $h(\zeta) = 1$ and $\Lambda = \mathbf{K}(\mathbf{x})$, we may invoke Lemma 2, leading to the following equality.

$$f(\mathbf{x}) = \frac{1}{\mathrm{Vol}(\mathcal{K})} \int_{\mathbf{K}(\mathbf{x})} 1 \; d\xi = \frac{1}{\mathrm{Vol}(\mathcal{K})} \frac{1}{\Gamma(1+n/2)} \int_{\mathbb{R}^d} e^{-g(\mathbf{x},\xi)} d\xi. \tag{17}$$

In fact the expression (17) can be restated as

$$\begin{aligned} f(\mathbf{x}) &= \int_{\mathbb{R}^d} \underbrace{\left(\mathcal{C}_{\mathcal{K}}(2\pi)^{d/2} e^{-g(\mathbf{x},\xi)+\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right)}_{\triangleq F(\mathbf{x},\xi)} \underbrace{\left(\frac{1}{D_{\mathcal{K}}(2\pi)^{d/2}} e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right)}_{\triangleq \tilde{p}(\xi)} d\xi \\ &= \int_{\mathbb{R}^d} F(\mathbf{x},\xi)\ \tilde{p}(\xi)\ d\xi = \mathcal{C}\ \mathbb{E}_{\tilde{p}(\xi)}[F(\mathbf{x},\xi)], \text{ where } \mathcal{C} \triangleq \tfrac{1}{\mathrm{Vol}(\mathcal{K})} \tfrac{1}{\Gamma(1+d/2)}, \end{aligned}$$

$\tilde{p}(\xi)$ denotes the density, $D_{\mathcal{K}}$ is such that $\int_{\mathbb{R}^d} \tilde{p}(\xi) d\xi = 1$, and $\mathcal{C}_{\mathcal{K}} \triangleq \mathcal{C}D_{\mathcal{K}}$. □

Next, we provide an analog of Lemma 4 for this integrand $F(\mathbf{x},\xi)$ but omit the proof. Note that $g_i(\bullet,\xi)$ is differentiable for every ξ and $i = 1,\cdots,d+1$.

Lemma 7. *Consider the function $F(\bullet,\xi)$ defined as*

$$F(\mathbf{x},\xi) = \begin{cases} \left(\mathcal{C}_{\mathcal{K}}(2\pi)^{d/2} e^{-g_i(\mathbf{x},\xi)+\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right), & \xi \in \Xi_i(\mathbf{x}) \\ \left(\mathcal{C}_{\mathcal{K}}(2\pi)^{d/2} e^{-g(\mathbf{x},\xi)+\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right), & \xi \in \Xi_0(\mathbf{x}), \quad \textit{where} \end{cases}$$

$$\begin{aligned} \Xi_i(\mathbf{x}) &\triangleq \{\xi \mid g_i(\mathbf{x},\xi) > g_j(\mathbf{x},\xi) \quad \forall j \neq i\}, i = 1,\cdots,d+1 \\ \Xi_0(\mathbf{x}) &\triangleq \{\xi \mid g_i(\mathbf{x},\xi) = g_j(\mathbf{x},\xi), \quad \forall (i,j) \in \mathcal{I}, \\ &\qquad\quad g_i(\mathbf{x},\xi) > g_l(\mathbf{x},\xi) \quad \forall l \notin \mathcal{I}, \quad \mathcal{I} \subseteq \{1,\cdots,d+1\}\}. \end{aligned}$$

Then the following hold.
(a) $F(\bullet,\xi)$ is locally Lipschitz for every ξ.
(b) $F(\bullet,\xi)$ is a Clarke regular function for almost every $\xi \in \mathbb{R}^d$.
(c) For any $\mathbf{x} \in \mathbb{R}^n$, $\partial\mathbb{E}[F(\mathbf{x},\xi)] = \mathbb{E}[\partial F(\mathbf{x},\xi)]$.

We now analyze $F(\mathbf{x},\xi)$ and $G(\mathbf{x},\xi)$ where $G(\mathbf{x},\xi) \in \partial F(\mathbf{x},\xi)$.

Lemma 8 (Properties of $F(\mathbf{x},\xi)$ and $G(\mathbf{x},\xi)$ under symmetric $\mathcal{K}$)**.** *Consider the problem* (**PM**)*. Suppose Assumption 1 holds, $c(\mathbf{x},\zeta)$ is defined as (5), and for all $\mathbf{x} \in \mathcal{X}$, $T_{i,\bullet}\mathbf{x} \ge \delta$ for every i for some $\delta > 0$. Suppose $f(\mathbf{x})$ is defined as in Prop. 5. Then the following hold.*
(a) For any $\mathbf{x} \in \mathcal{X}$, $|F(\mathbf{x},\xi)|^2 \le \mathcal{C}_{\mathcal{K}}^2(2\pi)^d$ for every $\xi \in \mathbb{R}^d$.
(b) Given an $\mathbf{x} \in \mathcal{X}$ such that $T_{i,\bullet}\mathbf{x} \ge \delta$ for every i and for some $\delta > 0$, and $G(\mathbf{x},\xi) \in \partial F(\mathbf{x},\xi)$, then it holds that $\mathbb{E}[\|G(\mathbf{x},\xi)\|^2] \le 16\mathcal{C}_{\mathcal{K}}^2(2\pi)^d \frac{\|T_{i,\bullet}\|^2}{\delta^2 e^2}$

Proof. Recall the definition of $F(\mathbf{x},\xi)$ from the statement of Lemma 7. Suppose $\Xi_{i,0}(\mathbf{x}) \triangleq \{\xi : g_i(\mathbf{x},\xi) \geq g_j(\mathbf{x},\xi), j \neq i\}$ for $i = 1,\ldots,d+1$. It can be seen that $\cup_{i=1}^{d+1}\Xi_{i,0}(\mathbf{x}) = \mathbb{R}^d$. We prove (a) by considering two cases. Case (i): $\xi \in \Xi_{i,0}(\mathbf{x})$ for $i = 1,\cdots,d$. It follows that

$$|F(\mathbf{x},\xi)|^2 = \mathcal{C}_{\mathcal{K}}^2\left((2\pi)^d e^{-2g(\mathbf{x},\xi)+\|\xi\|_{\mathcal{K}}^2}\right) \leq \mathcal{C}_{\mathcal{K}}^2\left((2\pi)^d e^{-2g_i(\mathbf{x},\xi)+g_i(\mathbf{x},\xi)}\right) \leq \mathcal{C}_{\mathcal{K}}^2(2\pi)^d.$$

Case (ii): $\xi \in \Xi_{d+1,0}(\mathbf{x})$. Proceeding similarly, we obtain that

$$|F(\mathbf{x},\xi)|^2 \leq \mathcal{C}_{\mathcal{K}}^2\left((2\pi)^d e^{-2g(\mathbf{x},\xi)+\|\xi\|_{\mathcal{K}}^2}\right) \leq \mathcal{C}_{\mathcal{K}}^2\left((2\pi)^d e^{-2\|\xi\|_{\mathcal{K}}^2+\|\xi\|_{\mathcal{K}}^2}\right) \leq \mathcal{C}_{\mathcal{K}}^2(2\pi)^d.$$

Consequently, $|F(\mathbf{x},\xi)|^2 \leq \mathcal{C}_{\mathcal{K}}^2(2\pi)^d$ for $\xi \in \mathbb{R}^d$.
(b). We observe that $\partial F(\mathbf{x},\xi)$ is defined as follows.

$$\partial F(\mathbf{x},\xi) = \begin{cases} \left(\mathcal{C}_{\mathcal{K}}(2\pi)^{d/2}\frac{2(\max\{\xi_i,0\})^2 T_{i,\bullet}^T}{(T_{i,\bullet}\mathbf{x})^3}e^{-g_i(\mathbf{x},\xi)+\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right), & \xi \in \Xi_i(\mathbf{x}), i = 1,\ldots,d \\ \left(-\mathcal{C}_{\mathcal{K}}(2\pi)^{d/2}e^{-g(\mathbf{x},\xi)+\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right)H(\mathbf{x},\xi), & \xi \in \Xi_0(\mathbf{x}) \\ \mathbf{0}. & \xi \in \Xi_{d+1}(\mathbf{x}), \end{cases}$$

where $H(\mathbf{x},\xi)$ denotes the Clarke generalized gradient of $g(\mathbf{x},\xi)$, defined as

$$H(\mathbf{x},\xi) = \left\{\sum_{\ell\in\mathcal{I}}\alpha_\ell\beta_\ell \mid \alpha_\ell \geq 0, \sum_{\ell\in\mathcal{I}}\alpha_\ell = 1, \beta_\ell = \nabla_{\mathbf{x}}g_\ell(\mathbf{x},\xi)\right\}. \tag{18}$$

Proceeding as in Prop 2, we have that $\mathbb{E}_{\tilde{p}}\left[\|G(\mathbf{x},\xi)\|^2\right]$ can be expressed as follows.

$$\begin{aligned} \mathbb{E}_{\tilde{p}}\left[\|G(\mathbf{x},\xi)\|^2\right] &= \int_{\mathbb{R}^n}\|G(\mathbf{x},\xi)\|^2\tilde{p}(\xi)d\xi \\ &= \sum_{i=1}^d\int_{\Xi_i(\mathbf{x})}\|G(\mathbf{x},\xi)\|^2\tilde{p}(\xi)d\xi + \int_{\Xi_{d+1}(\mathbf{x})}\|\underbrace{G(\mathbf{x},\xi)}_{=\mathbf{0}}\|^2\tilde{p}(\xi)d\xi \\ &+ \int_{\Xi_0(\mathbf{x})}\|G(\mathbf{x},\xi)\|^2\tilde{p}(\xi)d\xi \qquad (19) \\ &= \sum_{i=1}^d\int_{\Xi_i(\mathbf{x})}\|G(\mathbf{x},\xi)\|^2\tilde{p}(\xi)d\xi, \end{aligned}$$

where the last equality follows from observing that $G(\mathbf{x},\xi) = 0$ for $\xi \in \Xi_{d+1}(\mathbf{x})$ and the integral in (19) is zero because $\Xi_0(\mathbf{x})$ is a measure zero set. It follows that $\mathbb{E}[\|G(\mathbf{x},\xi)\|^2]$ can be bounded as follows:

$$\begin{aligned} &\mathbb{E}[\|G(x,\xi)\|^2] \\ &= \sum_{i=1}^d\int_{\Xi_i(x)}4\mathcal{C}_{\mathcal{K}}^2(2\pi)^d\left(\frac{\xi_i^2 T_{i,\bullet}^{\intercal}}{(T_{i,\bullet}\mathbf{x})^3}\right)^T\left(\frac{\xi_i^2 T_{i,\bullet}^{\intercal}}{(T_{i\bullet}\mathbf{x})^3}\right)e^{-\frac{2(\xi_i)^2}{(T_{i,\bullet}\mathbf{x})^2}+\|\xi\|_{\mathcal{K}}^2}\left(\frac{1}{D_{\mathcal{K}}(2\pi)^{d/2}}e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right)d\xi, \\ &= \sum_{i=1}^d\int_{\Xi_i(x)}4\mathcal{C}_{\mathcal{K}}^2(2\pi)^d\frac{\|T_{i,\bullet}\|^2}{(T_{i,\bullet}\mathbf{x})^2}\left(\frac{\xi_i}{T_{i,\bullet}\mathbf{x}}\right)^4 e^{-\frac{2(\xi_i)^2}{(T_{i,\bullet}\mathbf{x})^2}+\|\xi\|_{\mathcal{K}}^2}\left(\frac{1}{D_{\mathcal{K}}(2\pi)^{d/2}}e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right)d\xi, \\ &\leq \sum_{i=1}^d\int_{\Xi_i(x)}4\mathcal{C}_{\mathcal{K}}^2(2\pi)^d\frac{\|T_{i,\bullet}\|^2}{\delta^2}\left(\frac{\xi_i}{T_{i,\bullet}\mathbf{x}}\right)^4 e^{-\frac{(\xi_i)^2}{(T_{i,\bullet}\mathbf{x})^2}}\left(\frac{1}{D_{\mathcal{K}}(2\pi)^{d/2}}e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right)d\xi \end{aligned}$$

where the inequality follows from $\xi \in \Xi_i(\mathbf{x})$ for $i = 1, \ldots, d$ and $T_{i,\bullet}\mathbf{x} \geq \delta$ for all i. Next, we consider the expression $\left(\frac{\xi_i}{T_{i,\bullet}\mathbf{x}}\right)^4 e^{-\left(\frac{\xi_i}{T_{i,\bullet}\mathbf{x}}\right)^2}$ or $u^2 e^{-u}$. We note that by Lemma 3, $u^* = 2$ is a maximizer with value $4e^{-2}$. Hence, we have that

$$\begin{aligned}
\mathbb{E}[\|G(\mathbf{x},\xi)\|^2] &\leq \sum_{i=1}^{d} \int_{\Xi_i(x)} 4\mathcal{C}_{\mathcal{K}}^2 (2\pi)^d \frac{\|T_{i,\bullet}\|^2}{\delta^2} \frac{4}{e^2} \left(\frac{1}{D_{\mathcal{K}}(2\pi)^{d/2}} e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}} \right) d\xi \\
&= 16\mathcal{C}_{\mathcal{K}}^2 (2\pi)^d \sum_{i=1}^{d} \frac{\|T_{i,\bullet}\|^2}{\delta^2 e^2} \int_{\Xi_i(\mathbf{x})} \tfrac{1}{D_{\mathcal{K}}(2\pi)^{d/2}} e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}} d\xi \\
&= 16\mathcal{C}_{\mathcal{K}}^2 (2\pi)^d \sum_{i=1}^{d} \frac{\|T_{i,\bullet}\|^2}{\delta^2 e^2}.
\end{aligned}$$

□

We now specialize these results to regimes where $\mathcal{K}$ is an ℓ_p-ball in $\mathbb{R}^n$ and not necessarily symmetric about the origin via the following Proposition.

Assumption 2. *The random variable ζ is uniformly distributed on the set $\mathcal{K} \subset \mathbb{R}^d$ where $\mathcal{K} \triangleq \{\zeta : \|\zeta - \mu\|_p \leq \alpha\}$. The set $\mathcal{X}$ is closed, convex, and bounded.*

Proposition 6 (Representation and boundedness under asymmetric $\mathcal{K}$)**.** *Consider the problem* (**PM**)*. Suppose Assumption 2 holds and there exists $\delta > 0$ such that $T_{i,\bullet}\mathbf{x} - \mu_i \geq \delta$ for $i = 1, \cdots, d$. Then the following hold.*
(a) $f(\mathbf{x}) \triangleq \mathbb{E}_{\tilde{p}(\xi)}[F(\mathbf{x},\xi)]$, where $\sigma^2 = \alpha^2$,
$F(\mathbf{x},\xi) \triangleq \mathcal{C}(2\pi\sigma^2)^{d/2} e^{-g(\mathbf{x},\xi) + \frac{\|\xi\|_2^2}{2\sigma^2}}$, $\mathcal{C} \triangleq \frac{1}{\mathrm{Vol}(\mathcal{K})}\frac{1}{\Gamma(1+d/2)}$, $\tilde{p}(\xi) \triangleq \frac{1}{(2\pi\sigma^2)^{d/2}} e^{\frac{-\|\xi\|_2^2}{2\sigma^2}}$, and, $g(\mathbf{x},\xi) \triangleq \max\left\{ \frac{1}{\alpha^2}\|\xi\|_p^2, \left(\frac{\max(\xi_1,0)}{T_{1,\bullet}\mathbf{x} - \mu_1}\right)^2, \cdots, \left(\frac{\max(\xi_d,0)}{T_{d,\bullet}\mathbf{x} - \mu_d}\right)^2 \right\}$.
(b) Given an $\mathbf{x} \in \mathcal{X}$ and $T\mathbf{x} - \mu \geq \delta e$, $F(\mathbf{x},\xi) \geq 0$ and $|F(\mathbf{x},\xi)|^2 \leq \mathcal{C}^2(2\pi\sigma^2)^d$ for every $\xi \in \mathbb{R}^d$.
(c) Given an $\mathbf{x} \in \mathcal{X}$ and $T\mathbf{x} - \mu \geq \delta e$, and $G(\mathbf{x},\xi) \in \partial F(\mathbf{x},\xi)$, then it holds that

$$\mathbb{E}[\|G(\mathbf{x},\xi)\|^2] \leq 16\mathcal{C}^2(2\pi\sigma^2)^d \frac{\|T_{i,\bullet}\|^2}{\delta^2 e^2}. \tag{20}$$

Before concluding, we comment on the assumptions employed in this section. (a) *Assumptions on ζ.* We assume that ζ is uniformly distributed on the set $\mathcal{K}$ which is a compact and convex set, symmetric about the origin. The second requirement is that $\mathcal{K}$ be convex, compact, and symmetric (with corollories provided for specializing these results to an ellipsoid). This property allows us to claim that the Minkowski function of $\mathcal{K}$ is a norm, a key step in the analysis. However, we do develop an extension to non-symmetric regimes where $\mathcal{K} \triangleq \{\zeta \mid \|\zeta - \mu\|_p \leq \alpha\}$. The requirement that ζ is uniformly distributed may be weakened to log-concave measures and this will be the focus of future work, as noted in the concluding section.

(b) *Definition of $c(\mathbf{x},\zeta)$ in Settings A and B.* We have adopted two distinct choices for c; i.e. $c(\mathbf{x},\zeta) = 1 - |\zeta^T\mathbf{x}|^m$ (Setting A) and $T\mathbf{x} - \zeta$ (Setting B). Extensions to this are also possible where $\zeta^T A\mathbf{x}$ is employed in Setting A. This can be easily addressed by adding variables. More general extensions will be considered as part of future work.

3 An efficient stochastic approximation framework

In the prior section, we observed that the function f could be recast as an expectation of $F(\mathbf{x},\xi)$ with respect to a suitable density function. In Section 3.1, we cast the stochastic optimization of problem as a *convex compositional stochastic optimization problem* and comment on why available schemes do not suffice. We then provide some background and define the algorithmic framework in Section 3.2. Finally, convergence and rate analysis are provided in Section 3.3.

3.1 Convex compositional stochastic optimization problem

The optimization problem of interest, denoted by **(PM)**, can be cast as the following convex compositional optimization problem.

$$\min_{\mathbf{x}\in\mathcal{X}} \ h(\mathbf{x}), \ \text{where } h(\mathbf{x}) \triangleq \psi(\mathbb{E}[F(\mathbf{x},\xi)]) \text{ and } \psi(\mathbf{y}) \triangleq \begin{cases} \frac{1}{y} & (\mathbf{PM}_A) \\ -\log(y). & (\mathbf{PM}_B) \end{cases} \tag{21}$$

Before proceeding, we provide a brief review of SA schemes and their variance-reduced and compositional counterparts.

(a) *Stochastic approximation (SA) schemes.* SA schemes represent a class of techniques rooted in the seminal work by Robbins and Monro [65]. In the last several decades, there has been a tremendous amount of research in stochastic approximation applied to minimizing a convex function h, defined as $h(\mathbf{x}) \triangleq \mathbb{E}[H(\mathbf{x},\omega)]$ over a closed and convex set $\mathcal{X}$. Noteworthy amongst these being the long-step averaging framework by [58] and [59]. In fact, in [51] the authors developed a *robust* stochastic approximation framework for convex stochastic optimization in which a constant steplength of prescribed size was employed over a pre-selected number of SA steps. Such a scheme admits the optimal rate of convergence of $\mathbb{E}[h(\bar{\mathbf{x}}_K) - h^*] \leq \mathcal{O}(1/\sqrt{K})$ where K represented the number of steps and $\bar{\mathbf{x}}_K$ denotes the iterate average over K steps.

(b) *Variance-reduced schemes.* A key shortcoming of SA schemes is the gap in the convergence rate between the deterministic schemes and their SA analogs. This gap is particularly irksome in the presence of complicated constraints, since the projection operation is computationally expensive and in such cases, deterministic rates of convergence have profound benefits. For instance, to compute an ϵ-solution for a smooth convex expectation-valued problem, traditional SA schemes require at most $\mathcal{O}(1/\epsilon^2)$ while the variance-reduced counterparts require $\mathcal{O}(1/\epsilon)$. For instance, if $\epsilon = 1\text{e-}3$, standard SA schemes require $\mathcal{O}(1\text{e}6)$ projection steps while variance-reduced counterparts require $\mathcal{O}(1\text{e}3)$ steps, a significant difference. When considering sample-complexity, we note that in some instances such as [41], one may be able to (nearly) match the sample complexity of $\mathcal{O}(1/\epsilon^2)$. These improved rates are achieved by either utilizing an increasing batch-size of gradients or by solving a sequence of stochastic subproblems to increasing degrees of inexactness. Such avenues have derived deterministic rates of convergence in smooth strongly convex [19,68,76], smooth convex [33], nonsmooth convex [41], and nonconvex regimes [33,44]. Notably, in many of these settings, the schemes admit optimal or near-optimal sample complexities [41,44,68,76].

(c) Compositional stochastic optimization. The earliest efforts on compositional optimization appear to be the almost-sure convergence guarantees provided by Ermoliev [30] for two-level problems. Rate statements [74,75] and variance-reduction (in finite sample-space regimes) [45] has been studied while multi-level settings were first considered by [78]. Optimal sample-complexity in nonconvex regimes was shown for two-level [34] and multi-level [12,24] regimes. However, when the inner function is nonsmooth (as in this setting), the best known rate has been provided in [74] where a rate of $\mathcal{O}(k^{-1/4})$ has been derived. We note that in the present setting, sample complexity is of less relevance since ξ is a Gausian random variable and sampling is cheap with no explicit limitations on data (unlike in finite-sum machine learning problems). Instead, in this setting, we argue that iteration complexity is of more relevance.

(d) *Gaps and shortcomings in existing SA and compositional SA schemes.* A prototypical SA scheme for minimizing a convex function h, defined as $h(\mathbf{x}) \triangleq \mathbb{E}[H(\mathbf{x},\xi)]$ and given $\mathbf{x}_0 \in \mathcal{X}$, generates a sequence $\{\mathbf{x}_k\}$ as follows.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k d(\mathbf{x}_k,\xi_k)\right], \qquad k \geq 0 \tag{22}$$

where $d(\mathbf{x}_k,\xi_k)$ is assumed to be a sampled subgradient, the interchange between the expectation and subdifferential operator is assumed to hold for any $\mathbf{x}$, i.e. $\partial\mathbb{E}[H(\mathbf{x},\xi)] = \mathbb{E}[\partial H(\mathbf{x},\xi)]$, and $\mathbb{E}[d(\mathbf{x}_k,\xi) \mid \mathbf{x}_k] \in \partial_{\mathbf{x}}\mathbb{E}[H(\mathbf{x}_k,\xi)]$. When contending with $\psi(\mathbb{E}[F(\mathbf{x},\xi)])$, by the chain rule [27], we have that

$$\begin{aligned}\partial_{\mathbf{x}}\psi(\mathbb{E}[F(\mathbf{x},\xi)]) &= \partial_{\mathbf{x}}[\mathbb{E}[F(\mathbf{x},\xi)]]\psi'(\mathbb{E}[F(\mathbf{x},\xi)]) \\ &= \mathbb{E}[\partial_{\mathbf{x}}F(\mathbf{x},\xi)]\psi'(\mathbb{E}[F(\mathbf{x},\xi)]),\end{aligned} \tag{23}$$

where the second equality is a consequence of invoking Lemma 4. Consequently, an unbiased subgradient of $\psi(\mathbb{E}[F(\mathbf{x},\xi)])$ is given by $G(\mathbf{x},\xi)\psi'(\mathbb{E}[F(\mathbf{x},\xi)])$ and requires access to $\psi'(\mathbb{E}[F(\mathbf{x},\xi)])$; however, the latter cannot be accessed and therefore an unbiased subgradient cannot be tractably evaluated and standard SA schemes cannot be adopted.

(e) Related numerical schemes.

(i) *SA and Mini-batch SA schemes.* In the context of stochastic optimization with conditionally unbiased gradients being available, single-sample SA schemes are characterized by an optimal rate of $\mathcal{O}(k^{-1/2})$ while mini-batch variants employ a gradient estimator with reduced bias. However, in the current regime, such estimators are complicated by bias. Note that these schemes are not equipped by either asymptotic convergence or non-asymptotic rate guarantees. Yet, given that such schemes enjoy an optimal rate in unbiased regimes, SA schemes and mini-batch variants of SA provide a useful benchmark of comparison.

(ii) *Compositional SA schemes.* The presence of bias arising from the presence of a compositional structure has been addressed by compositional stochastic approximation schemes by adding a parallel updating scheme [74]. In nonsmooth regimes, such an avenue is characterized by a convergence rate of $\mathcal{O}(k^{-1/4})$ while our proposed scheme achieves a rate of approximately $\mathcal{O}(k^{-1/2})$. Note that sample (or oracle) complexity is less relevant here since sampling is (relatively) cheap and data is not limited by any means. Given the significant difference in rates, we have not provided an additional comparison with compositional SA schemes in the current manuscript.

(iii) *Sample-average approximation via integer programming.* Finally, the other competing approach for computing global minimizers of chance-constrained problems is via sample-average approximation (SAA) where the SAA problem is resolved via integer programming [55]. We introduce this comparison to demonstrate the difference in scalability and from the standpoint that this avenue also provides an additional certification that our proposed (r-VRSA) scheme is indeed finding near-global minimizers.

3.2 Background and Algorithm definition

We observe that the problem of interest is $\min_{\mathbf{x}\in\mathcal{X}} \psi(\mathbb{E}[F(\mathbf{x},\xi)])$. We first provide a result that allows us to claim that $\partial_{\mathbf{x}}[\psi(\mathbb{E}[F(\mathbf{x},\xi)])] = \psi'(\mathbb{E}[F(\mathbf{x},\xi)]\partial_{\mathbf{x}}[F(\mathbf{x},\xi)]$ indeed holds.

Lemma 9. *Suppose $F(\bullet,\xi)$ is a Clarke regular function for every $\xi \in \Xi$, ψ is a continuously differentiable function, and $\mathcal{X}$ is a nonempty, compact, and convex set in $\mathbb{R}^n$. Then the following hold.*
(a) $F(\bullet,\xi)$ is Lipschitz continuous on $\mathcal{X}$ with a Lipschitz constant $L(\xi)$ where $\mathbb{E}[L(\xi)] \leq \tilde{L}$.
(c) Suppose $f(\mathbf{x}) = \mathbb{E}[F(\mathbf{x},\xi)]$ and $f(\bar{\mathbf{x}})$ is finite for some $\bar{\mathbf{x}} \in \mathcal{X}$. Then f is Clarke-regular and for any $\mathbf{x} \in \mathcal{X}$, $\partial_{\mathbf{x}} f(\mathbf{x}) = \mathbb{E}[\partial_{\mathbf{x}} F(\mathbf{x},\xi)]$.
(d) Suppose $\psi : \mathbb{R}^+ \to \mathbb{R}$ is a continuously differentiable function. Then for any $\mathbf{x} \in \mathcal{X}$ such that $f(\mathbf{x}) = \mathbb{E}[F(\mathbf{x},\xi)] > 0$, $\partial_{\mathbf{x}}[\psi(f(\mathbf{x})] = \psi'(\mathbb{E}[F(\mathbf{x},\xi)])\mathbb{E}[\partial_{\mathbf{x}}[F(\mathbf{x},\xi)]]$.

Proof. (a) We present the proof for Setting A. We observe that $F(\bullet,\xi)$ is a piecewise-smooth function for every ξ. Then for any $\mathbf{x} \in \mathcal{X}$, $\|\nabla_{\mathbf{x}} F(\mathbf{x},\xi)\|$ is bounded as follows at points where $\mathbf{x}$ is smooth:

$$\begin{aligned}\|\nabla_{\mathbf{x}} F(\mathbf{x},\xi)\| &= \|C_{\mathcal{K}}(2\pi)^{n/2}(2\xi\mathbf{x}^T\mathbf{x})e^{-\max\{\xi^T x,\|\xi\|_{\mathcal{K}}^2\}+\|\xi\|_{\mathcal{K}}^2/2}\| \\ &\leq C_{\mathcal{K}}(2\pi)^{n/2}(\|\xi\|^2 + (\xi^T\mathbf{x})^2).\end{aligned}$$

Consequently, $F(\bullet,\xi)$ is a Lipschitz continuous function with $L(\xi) = C_{\mathcal{K}}(2\pi)^{n/2}(\|\xi\|^2 + (\xi^T\mathbf{x})^2)$. Furthermore, we have that $\mathbb{E}[L(\xi)] \leq C_{\mathcal{K}}(2\pi)^{n/2}(\mathbb{E}[\|\xi\|^2] + \frac{(\mathbb{E}[\|\xi\|^2]+\|\mathbf{x}\|^2])}{2} < \tilde{L}$, a consequence of the boundedness of the second moment and the compactness of $\mathcal{X}$.

(b) By definition, $f(\mathbf{x}) = \mathbb{E}[F(\mathbf{x},\xi)]$. Therefore, f is Lipschitz continuous with constant $\tilde{L}$ by utilizing convexity of the norm and Jensen's inequality as well as part (a).

(c) Since $F(\bullet,\xi)$ is Clarke-regular on $\mathcal{X}$, f is Lipschitz continuous on $\mathcal{X}$, f is defined at some $\bar{\mathbf{x}} \in \mathcal{X}$, we have that f is Clarke regular on $\mathcal{X}$ and for any $\mathbf{x} \in \mathcal{X}$, $\partial_{\mathbf{x}} f(\mathbf{x}) = \mathbb{E}[\partial_{\mathbf{x}} F(\mathbf{x},\xi)]$ [27, Th. 2.7.2].

(d) This follows from noting that by recalling that $\psi : \mathbb{R}_+ \to \mathbb{R}$ is continuously differentiable and f is Lipschitz continuous on $\mathcal{X}$, and then invoking [27, Cor. 2.6.6], we have that $\partial_{\mathbf{x}}[\psi(\mathbb{E}[F(\mathbf{x},\xi)]] = \psi'(\mathbb{E}[F(\mathbf{x},\xi)])\partial_{\mathbf{x}}[\mathbb{E}[F(\mathbf{x},\xi)]]$.

□

Consequently, an unbiased stochastic subgradient of h is given by a measurable selection $h'(\mathbb{E}[F(\mathbf{x},\xi)]G(\mathbf{x},\xi)$ where $G(\mathbf{x},\xi) \in \partial_{\mathbf{x}} F(\mathbf{x},\xi)$. However, such a selection cannot be efficiently evaluated since it requires $\mathbb{E}[F(\mathbf{x},\xi)]$ which is unavailable. Instead, we employ a biased variance-reduced counterpart given by the following.

$$D_k = \psi'_\epsilon\left(\frac{\sum_{j=1}^{N_k} F(\mathbf{x}_k,\xi_{k,j})}{N_k}\right)\frac{\sum_{j=1}^{N_k} G(\mathbf{x}_k,\xi_{j,k})}{N_k}. \tag{24}$$

We observe that the bias in defining the estimator D_k arises from approximating $\psi'(\mathbb{E}[F(\mathbf{x}_k,\xi)])$ by $\psi'_\epsilon\left(\frac{\sum_{j=1}^{N_k} F(\mathbf{x}_k,\xi_{k,j})}{N_k}\right)$ where ψ'_ϵ is suitably defined approximation of ψ with parameter ϵ. Consequently, we propose the following regularized variance-reduced stochastic subgradient scheme for minimizing $h(\mathbf{x})$ in either $(\mathbf{PM}_A)$ or $(\mathbf{PM}_B)$. We define the variance-reduced sampled gradient D_k as follows for each of these settings.

$$D_k \triangleq \begin{cases} \frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k)+\bar{w}_{f,k})^2+\epsilon_k}, & \text{(Setting A)} \\ -\frac{(G_j+\bar{w}_{G,k})}{(f(\mathbf{x}_k)+\bar{w}_{f,j,k})+\epsilon_k}, & \text{(Setting B)} \end{cases} \tag{25}$$

where $\bar{w}_k \triangleq d_k - D_k$, $d_k \in \partial_{\mathbf{x}} h(\mathbf{x}_k)$, $\bar{w}_{f,k}$ and $\bar{w}_{G,k}$ are defined as

$$\bar{w}_{f,k} \triangleq \frac{\sum_{j=1}^{N_k} F(\mathbf{x}_k,\xi)-f(\mathbf{x}_k)}{N_k} \text{ and } \bar{w}_{G,k} \triangleq \frac{\sum_{j=1}^{N_k} G(\mathbf{x}_k,\xi_j)-\mathbb{E}[G(\mathbf{x}_k,\xi)]}{N_k}, \tag{26}$$

respectively. We begin by assuming the existence of the following stochastic oracles, crucial for the development of the proposed first-order schemes.

Assumption 3 (Stochastic zeroth and first-order oracles)**.** *There exist a stochastic zeroth-order oracle and a stochastic first-order oracle that given* $\mathbf{x}$*, produce independent samples* $F(\mathbf{x},\xi)$ *and* $G(\mathbf{x},\xi) \in \partial F(\mathbf{x},\xi)$ *in Settings A and B.*

We now define the σ-algebra $\mathcal{F}_k$ for Setting B (Setting A is defined analogously).

$$\mathcal{F}_{f,k} \triangleq \left\{ \{F(\mathbf{x}_0,\xi_j)\}_{j=1}^{N_0}, \{F(\mathbf{x}_1,\xi_j)\}_{j=1}^{N_1}, \cdots, \{F(\mathbf{x}_k,\xi_j)\}_{j=1}^{N_k} \right\}, \tag{27}$$

$$\mathcal{F}_{G,k} \triangleq \left\{ \{G(\mathbf{x}_0,\xi_j)\}_{j=1}^{N_0}, \{G(\mathbf{x}_1,\xi_j)\}_{j=1}^{N_1}, \cdots, \{G(\mathbf{x}_k,\xi_j)\}_{j=1}^{N_k} \right\}, \tag{28}$$

$$\mathcal{F}_k \triangleq \mathcal{F}_{f,k} \cup \mathcal{F}_{G,k} \cup \{\mathbf{x}_0\}. \tag{29}$$

Suppose $\{\mathbf{x}_k\}$ is a sequence in $\mathcal{X}$. Then the following result holds.

Lemma 10. *For any* $\mathbf{x}_k \in \mathcal{X}$*, suppose* $\bar{w}_{f,k}$ *and* $\bar{w}_{G,k}$ *are defined as in* (26)*. Then for all* $k \geq 0$*,* $\mathbb{E}[\|\bar{w}_{f,k}\|^2 \mid \mathcal{F}_k] \leq \frac{\nu_f^2}{N_k}$ *and* $\mathbb{E}[\|\bar{w}_{G,k}\|^2 \mid \mathcal{F}_k] \leq \frac{\nu_G^2}{N_k}$ *where*

$$\nu_f^2 \triangleq 2(\mathcal{C}_{\mathcal{K}}^2(2\pi)^n + 1), \nu_G^2 \triangleq \frac{\mathcal{C}_{\mathcal{K}}^2(2\pi)^n \mathbb{E}_{\tilde{p}}[\|\xi\|^2]}{e}, \tag{Setting A}$$

$$\nu_f^2 \triangleq \mathcal{C}^2(2\pi\sigma^2)^d \text{ and } \nu_G^2 \triangleq 16\mathcal{C}_{\mathcal{K}}^2(2\pi)^d \sum_{i=1}^{d} \frac{\|T_{i,\bullet}\|^2}{\delta^2 e^2}. \tag{Setting B}$$

Algorithm 1 Regularized VR stochastic approximation (r-VRSA)

```
(0) Given $\mathbf{x}_0 \in \mathcal{X}$ and positive sequences $\{\gamma_k, \epsilon_k, N_k\}$; set $k := 1$.
(1) $\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}[\mathbf{x}_k - \gamma_k D_k]$, where $D_k$ is defined in (25)
(2) If $k > K$, then stop; else $k := k + 1$; return to (1).
```

Assumption 4. *There exists an* ϵ_f *such that* $f(\mathbf{x}_k) \geq \epsilon_f$ *and for any* $\mathbf{x}_k \in \mathcal{X}$*. For any* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$*,* $\|\mathbf{x} - \mathbf{y}\|^2 \leq B^2$.

Lemma 11. *Suppose Assumptions 3 and 4 hold. Consider any* $\mathbf{x}_k \in \mathcal{X}$*. Suppose* $N_k \in \mathbb{Z}_+$ *and* $\epsilon_k \triangleq \frac{1}{N_k^{1/4}}$*. Suppose* $\bar{w}_k \triangleq D_k - d_k$*, where* D_k *is defined in* (25) *and* $d_k \in \partial h(\mathbf{x}_k)$*. Suppose*

$\mathbb{E}[\|G(\mathbf{x}_k,\xi)\|^2 \mid \mathcal{F}_k] \leq M_G^2$ *and* $|F(\mathbf{x},\xi)| \leq M_F$ *for any* $\mathbf{x},\xi$. *Then* $\mathbb{E}[\|\bar{w}_k\|^2] \leq \frac{\nu^2}{\sqrt{N_k}}$*, where* $\nu^2 = \sum_{j=1}^d \nu_j^2$ *in Setting B,*

$$\nu^2 \triangleq \frac{3\nu_G^2}{\epsilon_f^2} + M_G^2 \frac{24\nu_f^2}{\epsilon_f^4} + \frac{6({M_F}^2+1)\nu_f^2}{\epsilon_f^4} \qquad \text{(Setting A)}$$

$$\nu^2 \triangleq \left(3\frac{\nu_G^2}{\epsilon_f^2} + 3M_G^2\frac{\nu_f^2}{\epsilon_f^2} + 3\left(\frac{M_G^2}{\epsilon_f^4}\right)\right). \qquad \text{(Setting B)}$$

and the constants $\nu_f, \nu_{f,j}, \nu_G$*, and* $\nu_{G,j}$ *are as specified in Lemma 10.*

3.3 Convergence Analysis

Proposition 7. *Suppose* h*, defined as in* (21)*, is a convex function on an open set containing* $\mathcal{X}$*. Suppose Assumption 1 holds and either Assumption 1 or Assumption 2 holds under Setting B. In addition, suppose Assumptions 3 and 4 hold. Consider the iterates generated by Algorithm 1. If* $\bar{x}_{\widehat{K},K} \triangleq \frac{\sum_{k=\widehat{K}}^{K-1}\gamma_k x_k}{\sum_{k=\widehat{K}}^{K-1}\gamma_k}$*, then for all* $K > 0$ *and* $\widehat{K}$ *satisfying* $0 \leq \widehat{K} < K-1$*,*

$$\mathbb{E}\left[h(\bar{x}_{\widehat{K},K}) - h(\mathbf{x}^*)\right] \leq \frac{\mathbb{E}[\|x_{\widehat{K}}-x^*\|^2]+\sum_{k=\widehat{K}}^{K-1}\gamma_k^2(M_G^2+B^2)+\sum_{k=\widehat{K}}^{K-1}\frac{\nu^2}{\sqrt{N_k}}}{\sum_{k=\widehat{K}}^{K-1}2\gamma_k}. \tag{30}$$

We now present a rate statement for diminishing and constant steplengths.

Theorem 4 (Rate statement for diminishing and constant steplengths). *Suppose* h*, defined as in* (21)*, is a convex function on an open set containing* $\mathcal{X}$*. Suppose Assumption 1 holds and either Assumption 1 or Assumption 2 holds under Setting B. In addition, suppose Assumptions 3 and 4 hold. Consider the iterates generated by Algorithm 1.*
(a) Suppose $\gamma_k = \frac{1}{k^{1/2+a}}$ *and* $N_k \triangleq \lceil 1/\gamma_k^4 \rceil$ *for all* k *where* $a < 1/2$*. If* $\widehat{K} \triangleq \lfloor K/2 \rfloor$*, then the following holds for every integer* $K \geq 2$*.*

$$\mathbb{E}\left[(h(\bar{\mathbf{x}}_{\widehat{K},K}) - h(\mathbf{x}^*))\right] \leq (1/2-a)\frac{B^2+\frac{1}{2a}(M_G^2+B^2+\nu^2)}{2(1-1/2^{1/2-a})}\frac{1}{K^{1/2-a}}. \tag{31}$$

(b) Given a positive integer K*, suppose* $\gamma_k \triangleq \sqrt{\frac{B^2}{(B^2+M_G^2+\nu^2)K}}$ *and* $N_k \triangleq \lceil 1/\gamma_k^4 \rceil$ *for all* k*. Then the following holds.*

$$\mathbb{E}\left[(h(\bar{\mathbf{x}}_K) - h(\mathbf{x}^*))\right] \leq \sqrt{\frac{(B^2+M_G^2+\nu^2)}{B^2K}}. \tag{32}$$

Proof. (a) Suppose $\widehat{K} = \lfloor K/2 \rfloor$ and $\gamma_k = \frac{\gamma_0}{k^{1/2+a}}$ for any $k \geq 0$. Then we have that

$$\begin{aligned}\sum_{k=\widehat{K}}^{K-1}\gamma_k &\geq \int_{\widehat{K}-1}^{K}\frac{1}{x^{1/2+a}}dx = \frac{K^{1/2-a}-(\widehat{K}-1)^{1/2-a}}{1/2-a} \geq \frac{K^{1/2-a}-(K/2)^{1/2-a}}{1/2-a} \\ \text{and } \sum_{k=\widehat{K}}^{K-1}\gamma_k^2 &\leq \int_{\widehat{K}-1}^{K}\frac{1}{x^{1+2a}}dx = \frac{K^{-2a}-(\widehat{K}-1)^{-2a}}{-2a} \leq \frac{(\widehat{K}-1)^{-2a}}{2a} \leq \frac{1}{2a}.\end{aligned}$$

It follows that

$$\mathbb{E}\left[(h(\bar{\mathbf{x}}_{\widehat{K},K}) - h(\mathbf{x}^*))\right] \leq \frac{B^2}{\sum_{k=\widehat{K}}^{K-1}2\gamma_k} + \frac{\sum_{k=\widehat{K}}^{K-1}\gamma_k^2(M_G^2+B^2)}{\sum_{k=\widehat{K}}^{K-1}2\gamma_k} + \frac{\sum_{k=\widehat{K}}^{K-1}\frac{\nu^2}{\sqrt{N_k}}}{\sum_{k=\widehat{K}}^{K-1}2\gamma_k}$$

$$\leq (1/2-a)\frac{B^2+\frac{1}{2a}(M_G^2+B^2+\nu^2)}{2(1-(1/2)^{1/2-a})}\frac{1}{K^{1/2-a}}.$$

(b) Suppose $\widehat{K}=0$ and $\gamma_k=\gamma$ for all k. Then we obtain the following bound.

$$\mathbb{E}\left[(h(\bar{\mathbf{x}}_K)-h(\mathbf{x}^*))\right] \leq \tfrac{B^2}{2K\gamma}+\tfrac{(M_G^2+B^2+\nu^2)K\gamma^2}{2K\gamma} = \tfrac{B^2}{2K\gamma}+\tfrac{(M_G^2+B^2+\nu^2)\gamma}{2}.$$

By minimizing the right hand side, which is a convex function in γ, we obtain

$$-\tfrac{B^2}{2K\gamma^2}+\tfrac{(M_G^2+B^2+\nu^2)}{2}=0 \implies \gamma^*=\sqrt{\tfrac{B^2}{(B^2+M_G^2+\nu^2)K}}.$$

The resulting bound on the expected sub-optimality is

$$\mathbb{E}_{\tilde{p}}\left[(h(\bar{\mathbf{x}}_K)-h(\mathbf{x}^*))\right] \leq \sqrt{\tfrac{(B^2+M_G^2+\nu^2)}{B^2K}}. \qquad \square$$

$\square$

We now employ the aforementioned rate to compute the sample (or oracle) complexity of computing a random $\mathbf{x}_K$ such that $\mathbb{E}[h(\mathbf{x}_K)-h(\mathbf{x}^*)] \leq \epsilon$.

Proposition 8 (**Oracle complexity for diminishing & constant steplengths**)**.** *Suppose h, defined as in* (21)*, is a convex function on an open set containing $\mathcal{X}$. Suppose Assumption 1 holds and either Assumption 1 or Assumption 2 holds under Setting B. In addition, suppose Assumptions 3 and 4 hold. Consider the iterates generated by Algorithm 1.*
(a) Suppose $\gamma_k=\frac{1}{k^{1/2+a}}$ and $N_k \triangleq \lceil 1/\gamma_k^4 \rceil$ for all k where $a<1/2$. If $\widehat{K} \triangleq \lfloor K/2 \rfloor$, then the following holds for every integer $K \geq 2$. Let $K(\epsilon)$ be any positive integer, $K(\epsilon) \geq 2$, such that $\mathbb{E}_{\tilde{p}}[h(\mathbf{x}_{K(\epsilon)})-h(\mathbf{x}^*)] \leq \epsilon$. Then $\sum_{k=0}^{K(\epsilon)} N_k \leq \mathcal{O}(1/\epsilon^{(6+8a)/(1-a)})$.
(b) Given a positive integer K, suppose $\gamma_k \triangleq \sqrt{\frac{B^2}{(B^2+M_G^2+\nu^2)K}}$ and $N_k \triangleq \lceil 1/\gamma_k^4 \rceil$ for all k. Let K_ϵ be any positive integer $K(\epsilon) \geq 2$ such that $\mathbb{E}_{\tilde{p}}[h(\mathbf{x}_{K(\epsilon)})-h(\mathbf{x}^*)] \leq \epsilon$. Then $\sum_{k=0}^{K(\epsilon)} N_k \leq \mathcal{O}(1/\epsilon^6)$.

Proof. (a). By utilizing Theorem 4(a), we have that $K(\epsilon)=\lceil \frac{\widehat{D}}{\epsilon^{2/(1-a)}} \rceil$, where $\hat{D}>0$. Consequently, we have that

$$\begin{aligned}\sum_{k=0}^{K(\epsilon)} N_k \leq \sum_{k=0}^{K(\epsilon)} (k+1)^{2+4a} &= \sum_{t=1}^{K(\epsilon)+1} t^{2+4a} \leq \int_1^{K(\epsilon)+2} x^{2+4a}dx \\ &\leq \frac{(K(\epsilon)+2)^{3+4a}}{3+4a} \leq \frac{2^{3+4a}\widehat{D}^{3+4a}}{(3+4a)\epsilon^{(6+8a)/(1-a)}}.\end{aligned}$$

(b) By utilizing Theorem 4(b), we have that $K(\epsilon)=\lceil \frac{\widehat{D}}{\epsilon^2} \rceil$, implying that $\gamma_k=\gamma=\frac{\tilde{c}}{\sqrt{K}}$ where $\tilde{c} \triangleq \sqrt{\frac{B^2}{(B^2+M_G^2+\nu^2)}}$. It follows that $N_k=N=\lceil \frac{K_\epsilon^2}{\tilde{c}^4} \rceil$. The oracle complexity may then be bounded as $\sum_{k=0}^{K(\epsilon)} N_k \leq \frac{8\widehat{D}^3}{(3\tilde{c}^4)\epsilon^6}$. $\square$

Next, we prove a.s. convergence of the sequence to a solution $\mathbf{x}^*$.

Theorem 5 (Almost sure convergence)**.** *Suppose h, defined as in* (21)*, is a convex function on an open set containing $\mathcal{X}$. Suppose Assumption 1 holds and either Assumption 1 or Assumption 2 holds under Setting B. In addition, suppose Assumptions 3 and 4 hold. Consider the iterates generated by Algorithm 1, where $\sum_{k=0}^\infty \gamma_k=\infty$, $\sum_{k=0}^\infty \gamma_k^2<\infty$, and $\sum_{k=0}^\infty \frac{1}{\sqrt{N_k}}<\infty$. Then $\mathbf{x}_k \xrightarrow[a.s.]{k\to\infty} \mathcal{X}^*$.*

Proof. We resume our argument utilizing the following inequality.

$$\begin{aligned}
\frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &\leq \frac{1}{2}\|\mathbf{x}_k - \gamma_k(d_k + \bar{w}_k) - x^*\|^2 \\
&= \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \frac{1}{2}\gamma_k^2\|d_k + \bar{w}_k\|^2 - \gamma_k(x_k - x^*)^T(d_k + \bar{w}_k) \\
&\leq \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \frac{1}{2}\gamma_k^2\|d_k + \bar{w}_k\|^2 - \gamma_k(x_k - x^*)^T(d_k) \\
&+ \gamma_k^2\|x_k - x^*\|^2 + \|\bar{w}_k\|^2.
\end{aligned}$$

This implies the following inequality holds.

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \gamma_k^2\|d_k + \bar{w}_k\|^2 - 2\gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*)) \\
&+ \gamma_k^2\|x_k - x^*\|^2 + \|\bar{w}_k\|^2 \\
&= (1 + \gamma_k^2)\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \gamma_k^2\|d_k + \bar{w}_k\|^2 \\
&- 2\gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*)) + \|\bar{w}_k\|^2.
\end{aligned}$$

By taking expectations conditioned on $\mathcal{F}_k$, we have the following inequality.

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathcal{F}_k] \leq (1 + \gamma_k^2)\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \gamma_k^2 M_G^2 - 2\gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*)) + \frac{\nu^2}{\sqrt{N_k}}.$$

Since $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$, $\sum_{k=0}^{\infty} \frac{1}{\sqrt{N_k}} < \infty$, it follows that $\{\|\mathbf{x}_k - \mathbf{x}^*\|^2\}$ is a convergent sequence in an a.s. sense and $\sum_{k=0}^{\infty} \gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*)) < \infty$ a.s. Consequently, $\{\mathbf{x}_k\}$ is bounded a.s. and has a convergent subsequence, indexed by $\mathcal{I}$. Since $\sum_{k=0}^{\infty} \gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*)) < \infty$ a.s. and $\sum_{k=0}^{\infty} \gamma_k = \infty$, it follows that $\liminf_{k\to\infty, k\in\mathcal{I}} h(\mathbf{x}_k) = h(\mathbf{x}^*)$ a.s. Consequently, there exists a subsequence of $\{\mathbf{x}_k\}$ that converges to the solution set $\mathcal{X}^*$ almost surely. But we have that $\{\|\mathbf{x}_k - \mathbf{x}^*\|^2\}$ is convergent a.s. and converges to zero along some subsequence. Consequently, the entire sequence $\{\|\mathbf{x}_k - \mathbf{x}^*\|^2\}$ converges to zero a.s. and the result holds. □

4 Numerical Results

In this section, we compare the performance of our scheme with standard stochastic approximation and integer programming approaches on two sets of examples. In all instances, the components of $x_0 \in \mathbb{R}^n$ are chosen randomly from the standard uniform distribution. In standard (**SA**) and (**batch-SA**) algorithms, the step length sequence is $\{1/\sqrt{k}\}$, while in (**batch-SA**), we compute the approximate subgradients using batch size of 100 samples. In **(r-VRSA)**, the step length sequence is $\{\frac{\gamma_0}{k^{1/2+a}}\}$. The parameters γ_0 and a are chosen as $\gamma_0 = 20,\ a = 0$ in Example 1 and $\gamma_0 = 1,\ a = 0$ in Example 2. In all instances, the components of $x_0 \in \mathbb{R}^n$ are chosen randomly from the standard uniform distribution.

Example 1. Set Covering. Demand is assumed to be uniformly distributed on $\mathcal{K} = \{\zeta \mid \|\zeta - \alpha\| \leq \alpha\}$ while the cost of operating a vehicle on route j is given by c_j while β is a cost threshold. By Prop 6, we may rewrite the problem (2) as

$$\min_{\mathbf{x}} h(\mathbf{x}) \ \triangleq\ -\log \mathbb{E}_{\tilde{p}(\xi)}[F(\mathbf{x}, \xi)], \quad \text{s. t. } \ c^T\mathbf{x} \leq \beta, \qquad \mathbf{x} \geq 0 \text{ where}$$

$F(\mathbf{x}, \xi) \ \triangleq\ \mathcal{C}(2\pi\sigma^2)^{d/2} e^{-g(\mathbf{x},\xi) + \frac{\|\xi\|_2^2}{2\sigma^2}}$, $\mathcal{C} \ \triangleq\ \frac{1}{\mathrm{Vol}(\mathcal{K})}\frac{1}{\Gamma(1+d/2)}$, $\tilde{p}(\xi) \ \triangleq\ \frac{1}{(2\pi\sigma^2)^{d/2}} e^{\frac{-\|\xi\|_2^2}{2\sigma^2}}$, and, $g(\mathbf{x}, \xi) \ \triangleq\ \max\left\{\frac{1}{\alpha^2}\|\xi\|_p^2, \left(\frac{\max(\xi_1,0)}{T_{1,\bullet}\mathbf{x} - \mu_1}\right)^2, \cdots, \left(\frac{\max(\xi_d,0)}{T_{d,\bullet}\mathbf{x} - \mu_d}\right)^2\right\}$.

Note that $\xi' s$ are normally distributed with zero mean and standard deviation σ where $\sigma^2 = \alpha^2$. We compare the performance of these algorithms for different setups. In these problems, the vehicle routing network is randomly generated and corresponding incidence matrix $T \in \mathbb{R}^{d\times n}$ is obtained. The elements of cost vector $c \in \mathbb{R}^n$ are randomly chosen from the uniform distribution on $[0, c_{\max}]$. We compare the performance and quality of the solutions with those obtained via an integer programming approximation as proposed in [47]. This avenue employs a sample average approximation approach facilitated by integer programming (denoted by (SAA-IP)), defined as follows:

$$\begin{aligned} &\max_{\mathbf{x}\in\mathcal{X},\mathbf{z}\in\{0,1\}^N} \frac{1}{N}\sum_{j=1}^{N} z_j \\ &\text{subject to } T_{i,\bullet}\mathbf{x} \geq v_i, \quad i = 1,\ldots,d \\ &\qquad v_i \geq \zeta_i^j z_j, \quad i = 1,\ldots,d,\ j = 1,\ldots,N \\ &\qquad v_i \geq 0. \quad i = 1,\ldots,d \end{aligned} \tag{SAA-IP$_N$}$$

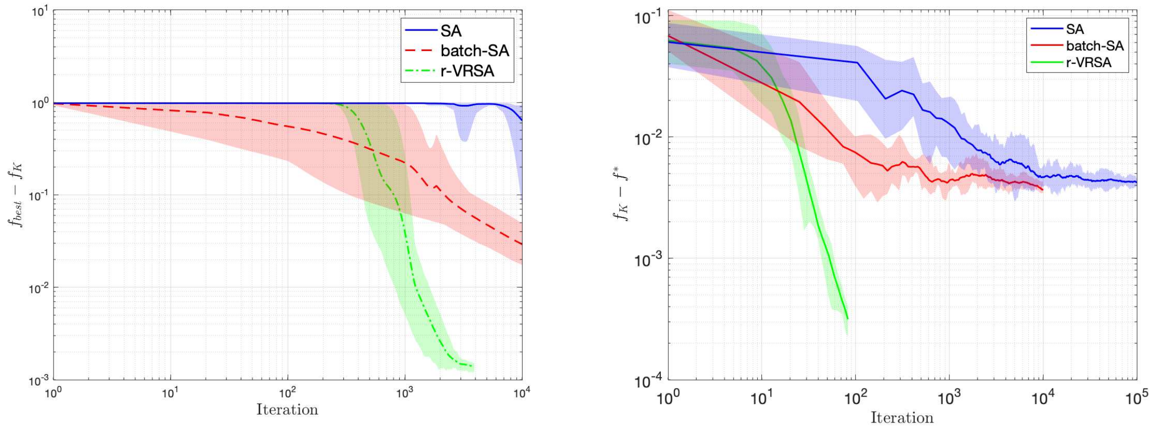
In this formulation, auxiliary variables v_i for $i = 1,\ldots,d$ represent $T_{i,\bullet}\mathbf{x}$ and $z_j = 1$ (or 0), then the constraints $T_{i,\bullet}\mathbf{x} \geq \zeta_i^j$ for $i = 1,\ldots,d$ corresponding to the realization j in the sample are enforced (or not enforced). We solve this problem using Gurobi MIP solver. The sample size for SAA-IP scheme is $N = 1e4$.

To compare across the solutions of various schemes including SA, batch-SA, r-VRSA, and SAA-IP, we first generate samples of demand vector ζ from the set $\mathcal{K} = \{\zeta \mid \|\zeta - \alpha\| \leq \alpha\}$. We then use Monte Carlo simulation to estimate $f(\mathbf{x}) \triangleq \mathbb{P}\{\zeta \in \mathcal{K} \mid T\mathbf{x} \geq \zeta\}$ for the solution $\mathbf{x}$ of each scheme. In Table 1, the first column prescribes problem parameters as follows: (Problem #, d, n, α, $c_{\max}$, γ). In SAA-IP scheme, the *Gap* refers to the reported gap between upper and lower bounds. Note that the table shows the probability being maximized.

Problem	**r-VRSA**						**SAA-IP**	
	B=1e6		**B=1e7**		**B=1e8**		**B=1e4**	
	f(x)	**Time**	**f(x)**	**Time**	**f(x)**	**Time**	**f(x)**	**Gap**
(1, 10, 9, 10, 5, 170)	0.9840	24s	0.9847	154s	0.9852	821s	0.9852	%0
(2, 14, 16, 8, 3, 46)	0.8341	26s	0.8356	158s	0.8357	1248s	0.8346	%0.4
(3, 18, 23, 16, 7, 250)	0.9325	32s	0.9327	172s	0.9328	1335s	0.9317	%0.2
(4, 23, 54, 40, 20, 530)	0.8255	33s	0.8767	177s	0.8768	1391s	0.8759	%0.8

Table 1: Set covering problem. (SAA-IP algorithm is terminated after 10000s.)

Example 2. Robust portfolio selection problem. We now consider the robust portfolio selection problem described in Section 1.2. We compare the proposed approach with the quadratic minimization (QM) framework [13] through which exact solutions are available. The portfolio weights are restricted to lie in the set $\mathcal{X}$ where $\mathcal{X} \triangleq \left\{\mathbf{x} : \mathbf{1}^T\mathbf{x} = 1 \text{ and } \mathbf{x} \geq 0\right\}$. The parameter α is set as ($\alpha = 0$). Given a threshold α and an allocation $\mathbf{x}$, we use the proposed framework to estimate the probability of a loss being less or equal than α as $f_\alpha(\mathbf{x}) \triangleq \mathbb{P}\left\{\tilde{\zeta} : \tilde{\zeta}^T\mathbf{x} \leq -\alpha\right\}$ where $\tilde{\zeta} = \zeta + \boldsymbol{\mu}$. In our simulations, given the number of assets n, mean μ (randomly generated), and covariance of random returns Σ, ζ, the returns, are assumed to be uniformly distributed over the set $\mathcal{K}_\epsilon = \{\zeta \in \mathbb{R}^n : \zeta^T \Sigma^{-1} \zeta \leq 1\}$. In Table 2, Problem column corresponds to (Problem no., number of assets n, γ).



(a) Example 1 (Vehicle Routing) (b) Example 2 (Portfolio Optimization)

Figure 1: Comparison of algorithms.

In Figure 1(a), the budget is $1e7$ and f_{best} (an approx. of f^*) is obtained by running the **(r-VRSA)** with a budget of $1e11$. In Figure 1(b), the budget is $1e6$. In both figures, the standard (SA) algorithm is terminated after $1e5$ iterations.

Problem	**batch-SA**		**r-VRSA**				**QM**
	B=1e5	**B=1e6**	**B=1e5**	**B=1e6**	**B=1e7**	**B=1e8**	**(f^*)**
(1, 4, 0.25)	0.3730	0.3723	0.3715	0.3712	0.3712	0.3711	0.3710
(2, 16, 0.2)	0.3000	0.2993	0.2976	0.2964	0.2963	0.2961	0.2961
(3, 64, 0.05)	0.3818	0.3752	0.3894	0.3787	0.3751	0.3743	0.3743
(4, 128, 0.15)	0.0899	0.0872	0.0992	0.0886	0.0869	0.0868	0.0867
(5, 256, 0.1)	0.1321	0.0991	0.1340	0.1031	0.0975	0.0972	0.0966

Table 2: Portfolio selection

Comments. Several observations can be made. (i) In Example 1, **(r-VRSA)** obtains near-optimal solutions within 1-2% of the time taken by **(SAA-IP)**, an integer programming approach. (ii) While **(batch-SA)** performs reasonably in Setting A, it tends to degenerate in Setting B. Further, convergence theory is unavailable for this scheme. Such schemes perform less favorably in comparison to **(r-VRSA)**. (iii) (**SAA-IP**) produces solutions of inferior gap as dimension grows and cannot accommodate growing number of samples, impacting solution quality.

5 Concluding remarks and future work

Traditional approaches for contending with chance-constrained optimimzation problems have relied on resolving convex approximations or computing stationary points. We concentrate our efforts on a subclass of such problems that require maximization the probability of a suitably specified event. By leveraging a recent result on non-Gaussian integrals of PHFs, we show that the probability of interest is an expectation of a possibly nonsmooth integrand. It is then shown that the composition of this expectation with a suitably specified smooth convex function leads to a convex program. However, a direct application of SA schemes is impeded by the inability to generate unbiased samples of the gradient. This motivated the development of a regularized variance-reduced SA scheme

(r-VRSA) that matches the optimal rate of subgradient methods for nonsmooth convex optimization problems but has somewhat poorer sample complexity, a consequence of the unavailability of conditionally unbiased gradients. We believe that this set of contributions represents amongst the first avenues (to the best of our knowedge) for tractably resolving probability maximization problems and is a crucial first step in examining more intricate problems in chance-constrained optimization.

This framework will provide the cornerstone for at least two key generalizations in our future work.

(i) *Extensions to log-concave measures.* First, this avenue may be extended to symmetric log-concave measures, subsuming Guassian, Laplace, Subbotin, amongst others. Pathways being exploited include alternative representations of probability density functions such as the so-called layer cake representation; e.g., see [46].

(ii) *Extensions to chance-constrained regimes.* This framework also allows for contending with constrained regimes. Consider the following chance-constrained problem and its expectation-valued counterpart.

$$\left\{\begin{array}{ll} \displaystyle\min_{\mathbf{x}\in\mathcal{X}} & f(\mathbf{x}) \\ \text{subject to} & \mathbb{P}\{\zeta \mid \zeta \in \mathbf{K}(\mathbf{x})\} \geq \bar{p} \end{array}\right\} \equiv \left\{\begin{array}{ll} \displaystyle\min_{\mathbf{x}\in\mathcal{X}} & f(\mathbf{x}) \\ \text{subject to} & \psi(\mathbb{E}[F(\mathbf{x},\xi)]) \leq \bar{c}, \end{array}\right\}$$

where $\bar{c}$ is related to $\bar{p}$. Assuming that f is a convex function on $\mathcal{X}$, we observe that the techniques in this paper allow for recasting the chance constrained problem can be recast as a convex optimization problem with nonsmooth compositional expectation-valued constraints.

Extensions and generalizations captured in (i) and (ii), while challenging, remain the focus of future work.

6 Appendix

Proof of Theorem 2: (a) When considering uniform distributions over a compact and convex set $\mathcal{K}$, the density is constant in this set and zero outside the set. It can then be concluded that ζ has a log-concave density. Furthermore, ζ has a symmetric density about the origin since $\mathcal{K}$ is a symmetric set about the origin. Hence by Lemma 6.2 in [16], h is convex where $h(\mathbf{x}) \triangleq 1/f(\mathbf{x})$.

(b) Since (11) is a convex program, any solution $\mathbf{x}^*$ satisfies $h(\mathbf{x}^*) \leq h(\mathbf{x})$, $\forall \mathbf{x} \in \mathcal{X}$. From the positivity of f over $\mathcal{X}$, $\frac{1}{f(\mathbf{x}^*)} \leq \frac{1}{f(\mathbf{x})}$ for every $\mathbf{x} \in \mathcal{X}$ implying that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$. Consequently, $\mathbf{x}^*$ is a global maximizer of (11).

□

Proof of Lemma 3: We prove this result by showing the unimodality of f on $\mathbb{R}_+$ where $f(u) = u^c e^{-u}$, implying that $f'(u) = cu^{c-1}e^{-u} - u^c e^{-u} = 0$ if $u = c$. Furthermore, $f'(u) > 0$ when $u < c$ and $f'(u) < 0$ when $u > c$. Finally, $f(0) = 0$. It follows that $u^* = c$ is a maximizer of $u^c e^{-u}$ on $[0, \infty)$ where $f(c) = \frac{c^c}{e^c}$.

□

Proof of Proposition 2: Recall the definition of $F(\mathbf{x}, \xi)$ from the statement of Lemma 4. We prove (a) by considering two cases. Case (i): $\xi \in \Xi_1(\mathbf{x}) \cup \Xi_0(\mathbf{x})$. It follows that

$$|F(\mathbf{x},\xi)|^2 = \mathcal{C}_{\mathcal{K}}^2 \left((2\pi)^n e^{-2|\xi^T\mathbf{x}|^2 + \|\xi\|_{\mathcal{K}}^2}\right) \leq \mathcal{C}_{\mathcal{K}}^2 \left((2\pi)^n e^{-2|\xi^T\mathbf{x}|^2 + |\xi^T\mathbf{x}|^2}\right) \leq \mathcal{C}_{\mathcal{K}}^2 (2\pi)^n.$$

Case (ii): $\xi \in \Xi_2(\mathbf{x})$. Proceeding similarly, we obtain that

$$|F(\mathbf{x},\xi)|^2 \leq \mathcal{C}_{\mathcal{K}}^2 \left((2\pi)^n e^{-2\|\xi\|_{\mathcal{K}}^2+\|\xi\|_{\mathcal{K}}^2}\right) \leq \mathcal{C}_{\mathcal{K}}^2 \left((2\pi)^n e^{-2\|\xi\|_{\mathcal{K}}^2+\|\xi\|_{\mathcal{K}}^2}\right) \leq \mathcal{C}_{\mathcal{K}}^2(2\pi)^n.$$

Consequently, $|F(\mathbf{x},\xi)|^2 \leq \mathcal{C}_{\mathcal{K}}^2(2\pi)^n$ for every $\xi \in \mathbb{R}^n$.
(b) We observe that $\partial F(\mathbf{x},\xi)$ is defined as follows.

$$\partial F(\mathbf{x},\xi) = \begin{cases} \left(\mathcal{C}_{\mathcal{K}}(2\pi)^{n/2}(-2\xi\xi^T\mathbf{x})e^{-|\xi^\mathsf{T}\mathbf{x}|^2+\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right), & \xi \in \Xi_1(\mathbf{x}) \triangleq \{\xi \mid |\xi^\mathsf{T}\mathbf{x}|^2 > \|\xi\|_{\mathcal{K}}^2\} \\ \left(-\mathcal{C}_{\mathcal{K}}(2\pi)^{n/2}e^{-\max\{|\xi^\mathsf{T}\mathbf{x}|^2,\|\xi\|_{\mathcal{K}}^2\}+\frac{\|\xi\|_{\mathcal{K}}^2}{2}}\right)\left[0, 2\xi(\xi^T\mathbf{x})\right], & \xi \in \Xi_0(\mathbf{x}) \triangleq \{\xi \mid |\xi^\mathsf{T}\mathbf{x}|^2 = \|\xi\|_{\mathcal{K}}^2\} \\ \mathbf{0}. & \xi \in \Xi_2(\mathbf{x}) \triangleq \{\xi \mid |\xi^\mathsf{T}\mathbf{x}|^2 < \|\xi\|_{\mathcal{K}}^2\} \end{cases}$$

Consequently, it follows that $\mathbb{E}_{\tilde{p}}\left[\|G(\mathbf{x},\xi)\|^2\right]$ is bounded as follows.

$$\begin{aligned} \mathbb{E}\left[\|G(\mathbf{x},\xi)\|^2\right] &= \int_{\Xi} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi \\ = \int_{\Xi_1(\mathbf{x})} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi &+ \int_{\Xi_2(\mathbf{x})} \|\underbrace{G(\mathbf{x},\xi)}_{=\,\mathbf{0}}\|^2 \tilde{p}(\xi) d\xi \\ + \int_{\Xi_0(\mathbf{x})} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi & \qquad (33) \\ = \int_{\Xi_1(\mathbf{x})} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi, & \end{aligned}$$

where the last equality follows from observing that $G(\mathbf{x},\xi) = 0$ for $\xi \in \Xi_2(\mathbf{x})$ and the integral in (33) is zero because $\Xi_0(\mathbf{x})$ is a measure zero set. It follows that $\mathbb{E}[\|G(\mathbf{x},\xi)\|^2]$ can be bounded as follows:

$$\begin{aligned} &\mathbb{E}[\|G(\mathbf{x},\xi)\|^2] \\ &= \int_{\Xi_1(\mathbf{x})} \left(\mathcal{C}_{\mathcal{K}}^2(2\pi)^n(4\|\xi\|_2^2(\xi^T\mathbf{x})^2)e^{-2(\xi^T\mathbf{x})^2+\|\xi\|_{\mathcal{K}}^2}\right) \frac{1}{D_{\mathcal{K}}}(2\pi)^{-\frac{n}{2}} e^{\frac{\|\xi\|_{\mathcal{K}}^2}{2}} d\xi && (34) \\ &\leq \int_{\Xi_1(\mathbf{x})} \left(\mathcal{C}_{\mathcal{K}}^2(2\pi)^n(4\|\xi\|_2^2(\xi^T\mathbf{x})^2)e^{-(\xi^T\mathbf{x})^2}\right) \frac{1}{D_{\mathcal{K}}}(2\pi)^{-\frac{n}{2}} e^{\frac{\|\xi\|_{\mathcal{K}}^2}{2}} d\xi, && (35) \end{aligned}$$

where the inequality follows from $\xi \in \Xi_1(\mathbf{x}, u)$. Next, we consider the expression $(\xi^T\mathbf{x})^2 e^{-(\xi^T\mathbf{x})^2}$ or ue^{-u}. We note that by Lemma 3, ue^{-u} is a unimodal function and $u^* = 1$ is a maximizer with value e^{-1}. Consequently, we have that

$$\max_{\{(\xi^T\mathbf{x}) | \xi \in \Xi(\mathbf{x})\}} (\xi^T\mathbf{x})^2 e^{-(\xi^T\mathbf{x})^2} \leq \max_{u \in \mathbb{R}_+} ue^{-u} \overset{\text{lemma 3}}{\leq} \tfrac{1}{e},$$

implying that

$$\begin{aligned} \mathbb{E}[\|G(\mathbf{x},\xi)\|^2] &\leq \int_{\Xi_1(\mathbf{x})} \left(\mathcal{C}_{\mathcal{K}}^2(2\pi)^n(\|\xi\|_2^2(\xi^T\mathbf{x})^2)e^{-(\xi^T\mathbf{x})^2}\right) \tfrac{1}{D_{\mathcal{K}}}(2\pi)^{-\frac{n}{2}} e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}} d\xi \\ \leq e^{-1}\mathcal{C}_{\mathcal{K}}^2(2\pi)^n &\int_{\Xi_1(\mathbf{x})} \|\xi\|_2^2 \tfrac{1}{D_{\mathcal{K}}}(2\pi)^{-\frac{n}{2}} e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}} d\xi \\ \leq e^{-1}\mathcal{C}_{\mathcal{K}}^2(2\pi)^n &\int_{\mathbb{R}^n} \|\xi\|_2^2 \tfrac{1}{D_{\mathcal{K}}}(2\pi)^{-\frac{n}{2}} e^{-\frac{\|\xi\|_{\mathcal{K}}^2}{2}} d\xi = e^{-1}\mathcal{C}_{\mathcal{K}}^2(2\pi)^n \mathbb{E}_{\tilde{p}}[\|\xi\|_2^2]. \end{aligned}$$

□

Proof of Proposition 3: (a) Since $\|\xi\|_{\mathcal{K}}^2 = \|\xi\|_p^2$, it follows from Theorem 1 that

$$
\begin{aligned}
f(\mathbf{x}) &= \int_{\mathbb{R}^n} \left(\mathcal{C}(2\pi\sigma^2)^{\frac{n}{2}} e^{-\max\{|\xi^T\mathbf{x}|^2, \|\xi\|_p^2\}}\right) d\xi \\
&= \int_{\mathbb{R}^n} \underbrace{\left(\mathcal{C}(2\pi\sigma^2)^{\frac{n}{2}} e^{-\max\{|\xi^T\mathbf{x}|^2, \|\xi\|_p^2\}+\frac{\|\xi\|_2^2}{2\sigma^2}}\right)}_{\triangleq F(\mathbf{x},\xi)} \underbrace{(2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}}}_{\triangleq \tilde{p}(\xi)} d\xi,
\end{aligned}
$$

(b) Omitted (similar to proof of Proposition 2(a).
(c) Next, we derive a bound on the second moment of $\|G(\mathbf{x},\xi)\|$ akin to Prop. 2(b). We observe that $\partial F(\mathbf{x},\xi)$ is defined as

$$
\partial F(\mathbf{x},\xi) = \begin{cases}
\left(\mathcal{C}(2\pi\sigma^2)^{n/2}(-2\xi\xi^T\mathbf{x})e^{-|\xi^\mathsf{T}\mathbf{x}|^2+\frac{\|\xi\|_2^2}{2\sigma^2}}\right), & \xi \in \Xi_1(\mathbf{x}) \triangleq \{\xi \mid |\xi^\mathsf{T}\mathbf{x}|^2 > \|\xi\|_p^2\} \\
\left(-\mathcal{C}(2\pi\sigma^2)^{n/2}e^{-\max\{|\xi^\mathsf{T}\mathbf{x}|^2, \|\xi\|_p^2\}+\frac{\|\xi\|_2^2}{2\sigma^2}}\right)\left[0, 2\xi(\xi^T\mathbf{x})\right], & \xi \in \Xi_0(\mathbf{x}) \triangleq \{\xi \mid |\xi^\mathsf{T}\mathbf{x}|^2 = \|\xi\|_p^2\} \\
\mathbf{0}. & \xi \in \Xi_2(\mathbf{x}) \triangleq \{\xi \mid |\xi^\mathsf{T}\mathbf{x}|^2 < \|\xi\|_p^2\}
\end{cases}
$$

Consequently, $\mathbb{E}[\|G(\mathbf{x},\xi)\|^2]$ can be bounded as follows.

$$
\begin{aligned}
\mathbb{E}\left[\|G(\mathbf{x},\xi)\|^2\right] &= \int_{\Xi} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi \\
= \int_{\Xi_1(\mathbf{x})} &\|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi + \int_{\Xi_2(\mathbf{x})} \|\underbrace{G(\mathbf{x},\xi)}_{=\mathbf{0}}\|^2 \tilde{p}(\xi) d\xi \\
+ \int_{\Xi_0(\mathbf{x})} &\|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi \qquad (36) \\
= \int_{\Xi_1(\mathbf{x})} &\|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi,
\end{aligned}
$$

where the last equality follows from observing that $G(\mathbf{x},\xi) = 0$ for $\xi \in \Xi_2(\mathbf{x})$ and the integral in (36) is zero because $\Xi_0(\mathbf{x})$ is a measure zero set. It follows that

$$
\begin{aligned}
\mathbb{E}[\|G(\mathbf{x},\xi)\|^2] &= \int_{\Xi_1(\mathbf{x})} \left(\mathcal{C}^2(2\pi\sigma^2)^n(4\|\xi\|_2^2(\xi^T\mathbf{x})^2)e^{-2(\xi^T\mathbf{x})^2+\frac{\|\xi\|_2^2}{\sigma^2}}\right)(2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{-\|\xi\|_2^2}{2\sigma^2}} d\xi \\
\leq \int_{\Xi_1(\mathbf{x})} &\left(\mathcal{C}^2(2\pi\sigma^2)^n(4\|\xi\|_2^2(\xi^T\mathbf{x})^2)e^{-2(\xi^T\mathbf{x})^2+\|\xi\|_p^2}\right)(2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{-\|\xi\|_2^2}{2\sigma^2}} d\xi \qquad (37) \\
\leq \int_{\Xi_1(\mathbf{x})} &\left(\mathcal{C}^2(2\pi\sigma^2)^n(4\|\xi\|_2^2(\xi^T\mathbf{x})^2)e^{-(\xi^T\mathbf{x})^2}\right)(2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{-\|\xi\|_2^2}{2\sigma^2}} d\xi, \qquad (38)
\end{aligned}
$$

where (38) follows from $\xi \in \Xi_1(\mathbf{x})$ and (37) follows from

$$
\frac{\|\xi\|_2^2}{\sigma^2} \leq \|\xi\|_p^2, \text{ where } \sigma^2 = \begin{cases} n^{1/2-1/p}, & p \geq 2 \\ 1. & 1 \leq p < 2 \end{cases}
$$

We may then conclude that

$$
\mathbb{E}[\|G(\mathbf{x},\xi)\|^2] \leq \int_{\Xi_1(\mathbf{x})} \left(\mathcal{C}^2(2\pi\sigma^2)^n(\|\xi\|_2^2(\xi^T\mathbf{x})^2)e^{-(\xi^T\mathbf{x})^2}\right)(2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}} d\xi
$$

$$\begin{aligned}
&\leq e^{-1}\mathcal{C}^2(2\pi\sigma^2)^n \int_{\Xi_1(\mathbf{x})} \left(\|\xi\|^2\right)(2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}} d\xi \qquad (39)\\
&\leq e^{-1}\mathcal{C}^2(2\pi\sigma^2)^n \int_{\mathbb{R}^n} \left(\|\xi\|_2^2\right)(2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}} d\xi = e^{-1}\mathcal{C}^2(2\pi\sigma^2)^n\mathbb{E}[\|\xi\|^2].
\end{aligned}$$

where (39) follows from Lemma 3. □

Proof of Lemma 6: Suppose $(\mathbf{x},\mathbf{y})$ is feasible with respect to $(\mathbf{PM}^2_{A,\mathrm{ext}})$. Then $\mathbf{x}\in\mathcal{X}$ and is therefore feasible with $(\mathbf{PM}^{\mathcal{E}}_A)$. In addition,

$$\begin{aligned}
f(\mathbf{x}) &\triangleq \mathbb{P}\left\{\zeta\in\mathbb{R}^n \mid \zeta\in\mathcal{K}_{\mathcal{E}}, \mid \zeta^\intercal\mathbf{x}\mid \leq 1\right\} = \mathbb{P}\left\{\zeta \mid \zeta^\intercal U^\intercal\Sigma^{-1}U\zeta\leq 1, \mid \zeta^\intercal\mathbf{x}\mid\leq 1\right\}\\
&= \mathbb{P}\left\{\zeta\in\mathbb{R}^n \mid \|\Sigma^{-1/2}U\zeta\|_2^2\leq 1, \mid \zeta^\intercal\mathbf{x}\mid\leq 1\right\}\\
&= \mathbb{P}\left\{U^\intercal\Sigma^{1/2}\eta\in\mathbb{R}^n \mid \|\eta\|_2^2\leq 1, \mid (U^\intercal\Sigma^{1/2}\eta)^\intercal\mathbf{x}\mid\leq 1\right\}\\
&= \mathbb{P}\left\{U^\intercal\Sigma^{1/2}\eta\in\mathbb{R}^n \mid \|\eta\|_2^2\leq 1, \mid \eta^\intercal\Sigma^{1/2}U\mathbf{x}\mid\leq 1\right\}\\
&= \mathbb{P}\left\{\eta\in\mathbb{R}^n \mid \eta\in\mathcal{K}_2, \mid \eta^\intercal\Sigma^{1/2}U\mathbf{x}\mid\leq 1\right\} \triangleq g(\mathbf{x}).
\end{aligned}$$

□

Proof of Proposition 6: (a) The result follows by a transformation argument. We define a new variable $\tilde{\zeta}\in\tilde{\mathcal{K}}$ such that $\tilde{\zeta}\triangleq\zeta-\mu$ where $\tilde{\mathcal{K}}\triangleq\{\tilde{\zeta}:\|\tilde{\zeta}\|_p\leq\alpha\}$. The set $\mathbf{K}(\mathbf{x})$ can be defined as the following

$$\tilde{\mathbf{K}}(\mathbf{x}) = \left\{\tilde{\zeta}:\tilde{\zeta}\in\tilde{\mathcal{K}}\right\}\cap\left\{\tilde{\zeta}:\tilde{\zeta}\leq T\mathbf{x}-\mu\right\}.$$

We first show that $\zeta\in\mathbf{K}(\mathbf{x})$ if and only if $\tilde{\zeta}\in\tilde{\mathbf{K}}(\mathbf{x})$. Suppose $\zeta\in\mathbf{K}(\mathbf{x})$. Then $\zeta\in\mathcal{K}$ and $c(\mathbf{x},\zeta)=T\mathbf{x}-\zeta\geq 0$. If $\zeta\in\mathcal{K}$, then $\|\zeta-\mu\|_p\leq\alpha$ or $\|\tilde{\zeta}\|_p\leq\alpha$ where $\tilde{\zeta}=\zeta-\mu$. Furthermore, $T\mathbf{x}\geq\zeta$ can be rewritten as $T\mathbf{x}-\mu\geq\zeta-\mu$ or $T\mathbf{x}-\mu\geq\tilde{\zeta}$. It follows that

$$\tilde{\zeta}\in\tilde{\mathbf{K}}(\mathbf{x}) = \left\{\tilde{\zeta}\mid\tilde{\zeta}\in\tilde{\mathcal{K}}\right\}\cap\left\{\tilde{\zeta}\mid T\mathbf{x}-\mu\geq\tilde{\zeta}\right\}.$$

The reverse direction follows similarly. Consequently, $\mathbb{P}\left\{\zeta\mid\zeta\in\mathbf{K}(\mathbf{x})\right\} = \mathbb{P}\left\{\tilde{\zeta}\mid\tilde{\zeta}\in\tilde{\mathbf{K}}(\mathbf{x})\right\}$. We now proceed analyze the latter probability. It may be observed that the Minkowski functional associated with $\tilde{\mathcal{K}}$ is given by $\|\tilde{\zeta}\|_{\tilde{\mathcal{K}}}=\frac{1}{\alpha}\|\tilde{\zeta}\|_p$. Since $T_{i,\bullet}\mathbf{x}-\mu_i\geq\delta>0$ for $i=1,\ldots,d$, it follows that

$$\begin{aligned}
\tilde{\mathbf{K}}(\mathbf{x}) &= \left\{\tilde{\zeta}:\tfrac{1}{\alpha}\|\tilde{\zeta}\|_p\leq 1\right\}\bigcap\left\{\tilde{\zeta}:\bigcap_{i=1}^{d}\frac{\max\{\tilde{\zeta}_i,0\}}{T_{i,\bullet}\mathbf{x}-\mu_i}\leq 1\right\}\\
&= \left\{\tilde{\zeta}:\tfrac{1}{\alpha^2}\|\tilde{\zeta}\|_p^2\leq 1\right\}\bigcap\left\{\tilde{\zeta}:\bigcap_{i=1}^{d}\left(\frac{\max\{\tilde{\zeta}_i,0\}}{T_{i,\bullet}\mathbf{x}-\mu_i}\right)^2\leq 1\right\}\\
&= \left\{\tilde{\zeta}:\max\left\{\frac{1}{\alpha^2}\|\tilde{\zeta}\|_p^2,\left(\frac{\max\{\tilde{\zeta}_1,0\}}{T_{1,\bullet}\mathbf{x}-\mu_1}\right)^2,\cdots,\left(\frac{\max\{\tilde{\zeta}_d,0\}}{T_{d,\bullet}\mathbf{x}-\mu_d}\right)^2\right\}\leq 1\right\}.
\end{aligned}$$

Since $g_i(\mathbf{x},\tilde{\zeta}) \triangleq \left(\frac{\max\{\tilde{\zeta}_i,0\}}{T_{i,\bullet}\mathbf{x}-\mu_i}\right)^2$ for $i=1,\ldots,d$ and $g_{d+1}(\mathbf{x},\tilde{\zeta})\triangleq\frac{1}{\alpha^2}\|\tilde{\zeta}\|_p^2$ are PHFs with degree 2, then $g(\mathbf{x},\tilde{\zeta})\triangleq\max\{g_1(\mathbf{x},\tilde{\zeta}),\ldots,g_{d+1}(\mathbf{x},\tilde{\zeta})\}$ is positively homogeneous with degree 2. By selecting

$h(\zeta) = 1$ and $\Lambda = \tilde{\mathbf{K}}(\mathbf{x})$, we may invoke Lemma 2, leading to the following equality.

$$f(\mathbf{x}) = \int_{\tilde{\mathbf{K}}(\mathbf{x})} 1 \; d\tilde{\zeta} = \frac{1}{\mathrm{Vol}(\mathcal{K})} \frac{1}{\Gamma(1+d/2)} \int_{\mathbb{R}^d} e^{-g(\mathbf{x},\xi)} \; d\xi. \tag{40}$$

The equation (40) can be rewritten as

$$\begin{aligned} f(\mathbf{x}) &= \int_{\mathbb{R}^d} \underbrace{\left(\mathcal{C}(2\pi\sigma^2)^{d/2} e^{-g(\mathbf{x},\xi)+\frac{\|\xi\|_2^2}{2\sigma^2}} \right)}_{\triangleq F(\mathbf{x},\xi)} \underbrace{\left(\tfrac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}} \right)}_{\triangleq p(\xi)} \; d\xi \\ &= \int_{\mathbb{R}^d} F(\mathbf{x},\xi) \; \tilde{p}(\xi) \; d\xi = \mathcal{C} \; \mathbb{E}_{\tilde{p}(\xi)}[F(\mathbf{x},\xi)], \text{ where } \mathcal{C} \triangleq \tfrac{1}{\mathrm{Vol}(\mathcal{K})} \; \tfrac{1}{\Gamma(1+d/2)}, \end{aligned}$$

(b) Omitted (similar to proof of Lemma 8 (a)).
(c) When $\mathcal{K}$ satisfies Assumption 2, the proof of Lemma 8(b) requires slight modification. Suppose $F(\mathbf{x},\xi)$ and $p(\xi)$ are defined as in (a). Then we may define $\partial F(\mathbf{x},\xi)$ as

$$\partial F(\mathbf{x},\xi) = \begin{cases} \left(\mathcal{C}(2\pi\sigma^2)^{d/2} \frac{2(\max\{\xi_i,0\})^2 T_{i,\bullet}^T}{(T_{i,\bullet}\mathbf{x}-\mu_i)^3} e^{-g_i(\mathbf{x},\xi)+\frac{\|\xi\|_2^2}{2\sigma^2}} \right), & \xi \in \Xi_i(\mathbf{x}), i = 1, \cdots, d \\ \left(-\mathcal{C}(2\pi\sigma^2)^{d/2} e^{-g(\mathbf{x},\xi)+\frac{\|\xi\|_2^2}{2\sigma^2}} \right) H(\mathbf{x},\xi), & \xi \in \Xi_0(\mathbf{x}) \\ \mathbf{0}. & \xi \in \Xi_{d+1}(\mathbf{x}), \end{cases}$$

where $H(\mathbf{x},\xi)$ denotes the Clarke generalized gradient of $g(\mathbf{x},\xi)$, defined as in (18). Consequently, it follows that $\mathbb{E}\left[\|G(\mathbf{x},\xi)\|^2\right]$ is bounded as follows.

$$\begin{aligned} &\mathbb{E}\left[\|G(\mathbf{x},\xi)\|^2\right] = \int_{\mathbb{R}^d} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi \\ &= \sum_{i=1}^d \int_{\Xi_i(\mathbf{x})} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi + \int_{\Xi_{d+1}(\mathbf{x})} \| \underbrace{G(\mathbf{x},\xi)}_{=\,\mathbf{0}} \|^2 \tilde{p}(\xi) d\xi \\ &+ \int_{\Xi_0(\mathbf{x})} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi \qquad (41) \\ &= \sum_{i=1}^d \int_{\Xi_i(\mathbf{x})} \|G(\mathbf{x},\xi)\|^2 \tilde{p}(\xi) d\xi, \end{aligned}$$

where the last equality follows from observing that $G(\mathbf{x},\xi) = 0$ for $\xi \in \Xi_{d+1}(\mathbf{x})$ and the integral in (41) is zero because $\Xi_0(\mathbf{x})$ is a measure zero set. It follows that

$$\begin{aligned} &\mathbb{E}[\|G(x,\xi)\|^2] \\ &= \sum_{i=1}^d \int_{\Xi_i(x)} 4\mathcal{C}^2 (2\pi\sigma^2)^d \frac{\|T_{i,\bullet}\|^2}{(T_{i,\bullet}\mathbf{x}-\mu_i)^2} \left(\frac{\xi_k}{T_{i,\bullet}\mathbf{x}-\mu_i} \right)^4 e^{-\frac{2(\xi_i)^2}{(T_{i,\bullet}\mathbf{x}-\mu_i)^2}+\frac{\|\xi\|_2^2}{\sigma^2}} \left(\frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}} \right) d\xi, \\ &\leq \sum_{i=1}^d \int_{\Xi_i(x)} 4\mathcal{C}^2 (2\pi\sigma^2)^d \frac{\|T_{i,\bullet}\|^2}{\delta^2} \left(\frac{\xi_k}{T_{i,\bullet}\mathbf{x}-\mu_i} \right)^4 e^{-\frac{2(\xi_i)^2}{(T_{i,\bullet}\mathbf{x}-\mu_i)^2}+\frac{\|\xi\|_2^2}{\sigma^2}} \left(\frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}} \right) d\xi \\ &\leq \sum_{i=1}^d \int_{\Xi_i(x)} 4\mathcal{C}^2 (2\pi\sigma^2)^d \frac{\|T_{i,\bullet}\|^2}{\delta^2} \left(\frac{\xi_i}{T_{i,\bullet}\mathbf{x}-\mu_i} \right)^4 e^{-\left(2-\frac{\alpha^2}{\sigma^2}\right)\frac{(\xi_i)^2}{(T_{i,\bullet}\mathbf{x}-\mu_i)^2}} \left(\frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|\xi\|_2^2}{2\sigma^2}} \right) d\xi \end{aligned}$$

where the first inequality follows from $T_{i,\bullet}\mathbf{x} - \mu_i \geq \delta > 0$ for all i, and the second inequality follows from $\xi \in \Xi_i(\mathbf{x})$. It follows from Lemma 3 that given any α, choosing the variance σ^2 of ξ such that $\sigma^2 = \alpha^2$ leads to the bound $\mathbb{E}[\|G(\mathbf{x},\xi)\|^2] \leq 16\mathcal{C}^2(2\pi\sigma^2)^d\sum_{i=1}^d \frac{\|T_{i,\bullet}\|^2}{\delta^2 e^2}$. □

Proof of Lemma 10: If $\tilde{G}(\mathbf{x}_k,\xi) \triangleq G(\mathbf{x}_k,\xi) - \mathbb{E}[G(\mathbf{x}_k,\xi)]$, by the conditional independence of $\tilde{G}(\mathbf{x}_k,\xi_j)$ and $\tilde{G}(\mathbf{x}_k,\xi_\ell)$ for $j \neq \ell$, we have

$$\begin{aligned}
&\mathbb{E}[\|\bar{w}_{G,k}\|^2 \mid \mathcal{F}_k] = \frac{1}{N_k^2}\mathbb{E}\left[\left\|\sum_{j=1}^{N_k}\tilde{G}(\mathbf{x}_k,\xi_j)\right\|^2 \mid \mathcal{F}_k\right] \\
&= \frac{1}{N_k^2}\mathbb{E}\left[\left[\sum_{j=1}^{N_k}\|\tilde{G}(\mathbf{x}_k,\xi_j)\|^2 + \sum_{\ell\neq j} 2\tilde{G}(\mathbf{x}_k,\xi_\ell)^T\tilde{G}(\mathbf{x}_k,\xi_j)\right] \mid \mathcal{F}_k\right] \\
&= \frac{1}{N_k}\Big(\mathbb{E}\left[\|G(\mathbf{x}_k,\xi)\|^2 \mid \mathcal{F}_k\right] + \|\mathbb{E}[G(\mathbf{x}_k,\xi) \mid \mathcal{F}_k]\|^2 - 2\mathbb{E}\left[G(\mathbf{x}_k,\xi) \mid \mathcal{F}_k\right]^T\mathbb{E}[G(\mathbf{x}_k,\xi) \mid \mathcal{F}_k]\Big) \\
&\leq \frac{1}{N_k}\mathbb{E}\left[\|G(\mathbf{x}_k,\xi)\|^2 \mid \mathcal{F}_k\right]. \qquad (42)
\end{aligned}$$

By (42) and Prop. 2, $\mathbb{E}[\|\bar{w}_{G,k}\|^2 \mid \mathcal{F}_k] \leq \frac{\mathcal{C}_K^2(2\pi)^n}{eN_k}\mathbb{E}_{\tilde{p}}[\|\xi\|^2]$ for Setting A. Similarly, for Setting B, by Lemma. 8,

$$\mathbb{E}[\|\bar{w}_{G,k}\|^2 \mid \mathcal{F}_k] \leq 16\mathcal{C}^2(2\pi\sigma^2)^d\sum_{i=1}^d\frac{\|T_{i,\bullet}\|^2}{\delta^2e^2N_k}.$$

In addition, for Setting A, $\mathbb{E}[\|\bar{w}_{f,k}\|^2 \mid \mathcal{F}_k] \leq \frac{2(\mathcal{C}_K^2(2\pi)^n+1)}{N_k}$ and $\mathbb{E}[\|\bar{w}_{f,k}\|^2 \mid \mathcal{F}_k] \leq \frac{\mathcal{C}^2(2\pi\sigma^2)^d}{N_k}$. □

Proof of Lemma 11: (Setting A) Consider $\bar{w}_k$, defined as $\bar{w}_k \triangleq \frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k)+\bar{w}_{f,k})^2+\epsilon_k} - \frac{-G_k}{(f(\mathbf{x}_k))^2}$. We have that

$$\begin{aligned}
\|\bar{w}_k\|^2 &= \left\|\frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k)+\bar{w}_{f,k})^2+\epsilon_k} - \frac{-G_k}{(f(\mathbf{x}_k))^2}\right\|^2 \\
&= \left\|\frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k)+\bar{w}_{f,k})^2+\epsilon_k} - \frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k))^2+\epsilon_k} + \frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k))^2+\epsilon_k}\right. \\
&\left. - \frac{-G_k}{(f(\mathbf{x}_k))^2+\epsilon_k} + \frac{-G_k}{(f(\mathbf{x}_k))^2+\epsilon_k} - \frac{-G_k}{(f(\mathbf{x}_k))^2}\right\|^2 \\
&\leq 3\left\|G_k - G_k + \bar{w}_{G,k}\right\|^2\frac{1}{((f(\mathbf{x}_k))^2+\epsilon_k)^2} \\
&+ 3\left\|G_k + \bar{w}_{G,k}\right\|^2\left\|\frac{1}{(f(\mathbf{x}_k))^2+\epsilon_k} - \frac{1}{(f(\mathbf{x}_k)+\bar{w}_{f,k})^2+\epsilon_k}\right\|^2 \\
&+ 3\left\|G_k\right\|^2\left\|\frac{1}{(f(\mathbf{x}_k))^2} - \frac{1}{(f(\mathbf{x}_k))^2+\epsilon_k}\right\|^2 \\
&\leq 3\left\|\bar{w}_{G,k}\right\|^2\frac{1}{((f(\mathbf{x}_k))^2+\epsilon_k)^2} \\
&+ 3\left\|G_k + \bar{w}_{G,k}\right\|^2\left\|\frac{(2f(\mathbf{x}_k)+\bar{w}_{f,k})\bar{w}_{f,k}}{((f(\mathbf{x}_k))^2+\epsilon_k)((f(\mathbf{x}_k)+\bar{w}_{f,k})^2+\epsilon_k)}\right\|^2 \\
&+ 3\left\|G_k\right\|^2\left\|\frac{\epsilon_k}{(f(\mathbf{x}_k))^2((f(\mathbf{x}_k))^2+\epsilon_k)}\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 3\left\|\bar{w}_{G,k}\right\|^2 \frac{1}{\epsilon_f^4} + 3\left\|G_k + \bar{w}_{G,k}\right\|^2 \left\| \frac{(2f(\mathbf{x}_k) + \bar{w}_{f,k})}{\epsilon_f^2 \epsilon_k} \right\|^2 \|\bar{w}_{f,k}\|^2 + 3\left\|G_k\right\|^2 \left(\frac{\epsilon_k^2}{\epsilon_f^8}\right) \\
&\leq 3\left\|\bar{w}_{G,k}\right\|^2 \frac{1}{\epsilon_f^4} + 3\left\|G_k + \bar{w}_{G,k}\right\|^2 \frac{(8f^2(\mathbf{x}_k)\|\bar{w}_{f,k}\|^2 + 2\|\bar{w}_{f,k}\|^4)}{\epsilon_f^4\epsilon_k^2} + \left\|G_k\right\|^2 \left(\frac{\epsilon_k^2}{\epsilon_f^8}\right),
\end{aligned}
$$

where $f(\mathbf{x}_k) \geq \epsilon_f$ for every $\mathbf{x}_k \in \mathcal{X}$. Taking conditional expectations and recalling the independence of $\bar{w}_{f,k}$ and $\bar{w}_{G,k}$ conditional on $\mathcal{F}_k$, the following bound emerges.

$$
\begin{aligned}
&\mathbb{E}[\|\bar{w}_k\|^2 \mid \mathcal{F}_k] \leq 3\mathbb{E}\left[\|\bar{w}_{G,k}\|^2 \mid \mathcal{F}_k\right] \frac{1}{\epsilon_f^2} \\
&+ 3\mathbb{E}\left[\|G_k + \bar{w}_{G,k}\|^2 \frac{(8f^2(\mathbf{x}_k)\|\bar{w}_{f,k}\|^2 + 2\|\bar{w}_{f,k}\|^4)}{\epsilon_f^4\epsilon_k^2} \mid \mathcal{F}_k\right] + 3\mathbb{E}\left[\|G_k\|^2 \left(\frac{\epsilon_k^2}{\epsilon_f^8}\right) \mid \mathcal{F}_k\right] \\
&\leq 3\frac{\nu_G^2}{\epsilon_f^2 N_k} + 3\mathbb{E}\left[\|G_k + \bar{w}_{G,k}\|^2 \mid \mathcal{F}_k\right] \mathbb{E}\left[\frac{(8f^2(\mathbf{x}_k)\|\bar{w}_{f,k}\|^2 + 2\|\bar{w}_{f,k}\|^4)}{\epsilon_f^4\epsilon_k^2} \mid \mathcal{F}_k\right] + \left(\frac{3\epsilon_k^2 M_G^2}{\epsilon_f^8}\right) \\
&\leq 3\frac{\nu_G^2}{\epsilon_f^2 N_k} + 3M_G^2 \frac{8f^2(\mathbf{x}_k)\nu_f^2}{\epsilon_f^4\epsilon_k^2 N_k} + 3M_G^2 \mathbb{E}\left[\frac{\|\bar{w}_{f,k}\|^4}{\epsilon_f^4\epsilon_k^2} \mid \mathcal{F}_k\right] + 3\left(\frac{\epsilon_k^2 M_G^2}{\epsilon_f^8}\right),
\end{aligned}
$$

where $\|G_k\|^2 = \|\mathbb{E}[G(\mathbf{x}_k,\xi) \mid \mathcal{F}_k]\|^2 \leq \mathbb{E}[\|G(\mathbf{x}_k,\xi)\|^2 \mid \mathcal{F}_k] \leq M_G^2$ by Jensen's inequality. From Prop. 2(b,c), $|F(\mathbf{x},\xi)| \leq M_F$ for any $\mathbf{x}, \xi$, implying that

$$
\|\bar{w}_{f,k}\|^2 = \left\|\frac{\sum_{j=1}^{N_k} F(\mathbf{x}_k,\xi_j)}{N_k} - f(\mathbf{x}_k)\right\|^2 \leq 2\left\|\frac{\sum_{j=1}^{N_k} F(\mathbf{x}_k,\xi_j)}{N_k}\right\|^2 + 2f^2(\mathbf{x}_k) \leq 2({M_F}^2 + 1).
$$

Consequently, by recalling that $\epsilon_k = 1/N_k^{1/4}$, the following holds a.s.

$$
\begin{aligned}
\mathbb{E}[\|\bar{w}_k\|^2 \mid \mathcal{F}_k] &\leq \frac{3\nu_G^2}{\epsilon_f^2 N_k} + 24M_G^2 \frac{f^2(\mathbf{x}_k)\nu_f^2}{\epsilon_f^4\epsilon_k^2 N_k} + 3\mathbb{E}\left[\frac{\|\bar{w}_{f,k}\|^4}{\epsilon_f^4\epsilon_k^2} \mid \mathbf{x}_k\right] + \left(\frac{\epsilon_k^2 M_G^2}{\epsilon_f^8}\right) \\
&\leq \frac{\nu_G^2}{\epsilon_f^2 N_k} + 24M_G^2 \frac{f^2(\mathbf{x}_k)\nu_f^2}{\epsilon_f^4\epsilon_k^2 N_k} + \frac{6({M_F}^2+1)M_G^2\nu_f^2}{\epsilon_f^4\epsilon_k^2 N_k} \\
&\leq \frac{3\nu_G^2}{\epsilon_f^2\sqrt{N_k}} + M_G^2 \frac{24f^2(\mathbf{x}_k)\nu_f^2}{\epsilon_f^4\sqrt{N_k}} + \frac{6(M_F^2+1)\nu_f^2}{\epsilon_f^4\sqrt{N_k}} + \left(\frac{3M_G^2}{\epsilon_f^8\sqrt{N_k}}\right) \\
&\triangleq \frac{\nu^2}{\sqrt{N_k}}, \text{ where } \nu^2 \triangleq \frac{3\nu_G^2}{\epsilon_f^2} + M_G^2\frac{24\nu_f^2}{\epsilon_f^4} + \frac{6({M_F}^2+1)\nu_f^2}{\epsilon_f^4} + \left(\frac{3M_G^2}{\epsilon_f^8}\right).
\end{aligned}
$$

(Setting B) Since $\bar{w}_k \triangleq \frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k)+\bar{w}_{f,k})+\epsilon_k} + \frac{G_k}{f(\mathbf{x}_k)}$ and

$$
\begin{aligned}
\|\bar{w}_k\|^2 &= \left\|\frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k)+\bar{w}_{f,k})+\epsilon_k} - \frac{-G_k}{(f(\mathbf{x}_k))}\right\|^2 \\
&= \left\|\frac{-(G_k+\bar{w}_{G,k})}{(f(\mathbf{x}_k)+\bar{w}_{f,k})+\epsilon_k} - \frac{-(G_k+\bar{w}_{G,k})}{f(\mathbf{x}_k)+\epsilon_k} + \frac{-(G_k+\bar{w}_{G,k})}{f(\mathbf{x}_k)+\epsilon_k}\right. \\
&\left. - \frac{-G_k}{f(\mathbf{x}_k)+\epsilon_k} + \frac{-G_k}{f(\mathbf{x}_k)+\epsilon_k} - \frac{-G_k}{f(\mathbf{x}_k)}\right\|^2
\end{aligned}
$$

$$\begin{aligned}
&\le 3\left\|G_k - G_k + \bar{w}_{G,k}\right\|^2 \frac{1}{(f(\mathbf{x}_k)+\epsilon_k)^2}\\
&+ 3\left\|G_k + \bar{w}_{G,k}\right\|^2 \left\|\frac{1}{f(\mathbf{x}_k)+\epsilon_k} - \frac{1}{(f(\mathbf{x}_k)+\bar{w}_{f,k})+\epsilon_k}\right\|^2\\
&+ 3\left\|G_k\right\|^2 \left\|\frac{1}{f(\mathbf{x}_k)} - \frac{1}{f(\mathbf{x}_k)+\epsilon_k}\right\|^2\\
&\le 3\left\|\bar{w}_{G,k}\right\|^2 \frac{1}{(f(\mathbf{x}_k)+\epsilon_k)^2} + 3\left\|G_k + \bar{w}_{f,k}\right\|^2 \left\|\frac{\bar{w}_{f,k}}{(f(\mathbf{x}_k)+\epsilon_k)(\underbrace{(f(\mathbf{x}_k)+\bar{w}_{f,k})}_{\ge 0, F(\mathbf{x}_k,\xi)\ge 0}+\epsilon_k)}\right\|^2\\
&+ 3\left\|G_k\right\|^2 \left\|\frac{\epsilon_k}{f(\mathbf{x}_k)(f(\mathbf{x}_k)+\epsilon_k)}\right\|^2\\
&\le 3\left\|\bar{w}_{G,k}\right\|^2 \frac{1}{\epsilon_f^2} + 3\left\|G_k + \bar{w}_{G,k}\right\|^2 \left\|\frac{1}{\epsilon_f\epsilon_k}\right\|^2 \left\|\bar{w}_{f,k}\right\|^2 + 3\left\|G_k\right\|^2 \left(\frac{\epsilon_k^2}{\epsilon_f^4}\right)\\
&\le 3\left\|\bar{w}_{G,k}\right\|^2 \frac{1}{\epsilon_f^2} + 3\left\|G_k + \bar{w}_{G,k}\right\|^2 \frac{\left\|\bar{w}_{f,k}\right\|^2}{\epsilon_f^2\epsilon_k^2} + \left\|G_k\right\|^2 \left(\frac{\epsilon_k^2}{\epsilon_f^4}\right),
\end{aligned}$$

where $f(\mathbf{x}_k) \ge \epsilon_f$ and for every $\mathbf{x}_k \in \mathcal{X}$. Taking expectations conditioned on $\mathcal{F}_k$ and recalling the independence of $\bar{w}_{f,k}$ and $\bar{w}_{G,k}$ conditional on $\mathcal{F}_k$, we have the following bound.

$$\begin{aligned}
&\mathbb{E}[\|\bar{w}_k\|^2 \mid \mathcal{F}_k]\\
&\le \left(3\mathbb{E}\left[\|\bar{w}_{G,k}\|^2 \mid \mathcal{F}_k\right]\frac{1}{\epsilon_f^2} + 3\mathbb{E}\left[\|G_k + \bar{w}_{G,k}\|^2 \frac{\|\bar{w}_{f,k}\|^2}{\epsilon_f^2\epsilon_k^2} \mid \mathcal{F}_k\right] + 3\mathbb{E}\left[\|G_k\|^2\left(\frac{\epsilon_k^2}{\epsilon_f^4}\right) \mid \mathcal{F}_k\right]\right)\\
&\le \left(3\frac{\nu_G^2}{\epsilon_f^2 N_k} + 3\mathbb{E}\left[\|G_k + \bar{w}_{G,k}\|^2 \mid \mathcal{F}_k\right]\mathbb{E}\left[\frac{\|\bar{w}_{f,k}\|^2}{\epsilon_f^2\epsilon_k^2} \mid \mathcal{F}_k\right] + 3\left(\frac{\epsilon_k^2 M_G^2}{\epsilon_f^4}\right)\right)\\
&\le \left(3\frac{\nu_G^2}{\epsilon_f^2 N_k} + 3M_G^2\frac{\nu_f^2}{\epsilon_f^2\epsilon_k^2 N_k} + 3\left(\frac{\epsilon_k^2 M_G^2}{\epsilon_f^4}\right)\right).
\end{aligned}$$

By selecting $\epsilon_k = 1/N_k^{1/4}$, we have that

$$\mathbb{E}[\|\bar{w}_k\|^2 \mid \mathcal{F}_k] \le \frac{\nu^2}{\sqrt{N_k}}, \text{ where } \nu^2 \triangleq \left(3\frac{\nu_G^2}{\epsilon_f^2} + 3M_G^2\frac{\nu_f^2}{\epsilon_f^2} + 3\left(\frac{M_G^2}{\epsilon_f^4}\right)\right).$$

Proof of Proposition 7: (i) Using the update rule of $\mathbf{x}_{k+1}$ and the fact that $\mathbf{x}^* = \Pi_{\mathcal{X}}[\mathbf{x}^*]$, for any $d_k + \bar{w}_k$ where $d_k \in \partial h(\mathbf{x}_k)$ and $k \ge 1$,

$$\begin{aligned}
\frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \frac{1}{2}\|\Pi_{\mathcal{X}}(\mathbf{x}_k - \gamma_k(d_k + \bar{w}_k)) - \Pi_{\mathcal{X}}(\mathbf{x}^*))\|^2\\
&\le \frac{1}{2}\|\mathbf{x}_k - \gamma_k(d_k + \bar{w}_k) - \mathbf{x}^*\|^2\\
&= \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \frac{1}{2}\gamma_k^2\|d_k + \bar{w}_k\|^2 - \gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T(d_k + \bar{w}_k),
\end{aligned}$$

where in the second inequality, we employ the non-expansivity of projection operator. Now by using the convexity of h, we obtain:

$$2\gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*)) \le \left(\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right) + \|d_k + \bar{w}_k\|^2\gamma_k^2$$

$$\begin{aligned}&- 2\gamma_k \bar{w}_k^T(\mathbf{x}_k - \mathbf{x}^*)\\ &\leq \left(\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right) + \|d_k + \bar{w}_k\|^2\gamma_k^2\\ &+ \gamma_k^2\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \|\bar{w}_k\|^2,\end{aligned}$$

where we use $a^Tb \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$. Now by summing from $k = \widehat{K}$ to $K-1$, where $\widehat{K}$ is an integer satisfying $0 \leq \widehat{K} < K - 1$, we obtain the next inequality.

$$\sum_{k=\widehat{K}}^{K-1} 2\gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*)) \leq \|\mathbf{x}_{\widehat{K}} - \mathbf{x}^*\|^2 + \sum_{k=\widehat{K}}^{K-1} \gamma_k^2(\|d_k + \bar{w}_k\|^2 + \|\mathbf{x}_k - \mathbf{x}^*\|^2) + \|\bar{w}_k\|^2.$$

Dividing both sides by $2\sum_{k=\widehat{K}}^{K-1}\gamma_k$, taking expectations on both sides, and invoking Lemma 11 which leads to $\mathbb{E}[\|\bar{w}_k \mid \mathcal{F}_k\|]^2 \leq \frac{\nu^2}{\sqrt{N_k}}$ and the bound of the subgradient, i.e., $\mathbb{E}[\|d_k + \bar{w}_k\|^2] \leq M_G^2$, we obtain the following bound.

$$\begin{aligned}&\mathbb{E}\left[\frac{\sum_{k=\widehat{K}}^{K-1} 2\gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*))}{\sum_{k=\widehat{K}}^{K-1} 2\gamma_k}\right]\\ &\leq \mathbb{E}\left[\frac{\|\mathbf{x}_{\widehat{K}} - \mathbf{x}^*\|^2 + \sum_{k=\widehat{K}}^{K-1}\gamma_k^2\|d_k + \bar{w}_k\|^2 + \sum_{k=\widehat{K}}^{K-1}\gamma_k^2\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \sum_{k=\widehat{K}}^{K-1}\|\bar{w}_k\|^2}{\sum_{k=\widehat{K}}^{K-1} 2\gamma_k}\right] &(43)\\ &\leq \frac{\mathbb{E}[\|x_{\widehat{K}} - x^*\|^2]}{\sum_{k=\widehat{K}}^{K-1} 2\gamma_k} + \frac{\sum_{k=\widehat{K}}^{K-1}\gamma_k^2(M_G^2 + B^2)}{\sum_{k=\widehat{K}}^{K-1} 2\gamma_k} + \frac{\sum_{k=\widehat{K}}^{K-1}\frac{\nu^2}{\sqrt{N_k}}}{\sum_{k=\widehat{K}}^{K-1} 2\gamma_k}. &(44)\end{aligned}$$

By utilizing Jensen's inequality, we obtain that

$$\mathbb{E}\left[(h(\bar{x}_{\widehat{K},K} - h(\mathbf{x}^*))\right] \leq \mathbb{E}\left[\frac{\sum_{k=\widehat{K}}^{K-1} 2\gamma_k(h(\mathbf{x}_k) - h(\mathbf{x}^*))}{\sum_{k=\widehat{K}}^{K-1} 2\gamma_k}\right],$$

where $\bar{x}_{\widehat{K},K} \triangleq \frac{\sum_{k=\widehat{K}}^{K-1}\gamma_k x_k}{\sum_{k=\widehat{K}}^{K-1}\gamma_k}$, which when combined with (44) leads to (30). □

References

[1] van Ackooij, W.: Eventual convexity of chance constrained feasible sets. Optimization **64**(5), 1263–1284 (2015)

[2] van Ackooij, W.: A discussion of probability functions and constraints from a variational perspective. Set-Valued Var. Anal. **28**(4), 585–609 (2020). DOI 10.1007/s11228-020-00552-2. URL `https://doi.org/10.1007/s11228-020-00552-2`

[3] van Ackooij, W., Aleksovska, I., Munoz-Zuniga, M.: (Sub-)differentiability of probability functions with elliptical distributions. Set-Valued Var. Anal. **26**(4), 887–910 (2018). DOI 10.1007/s11228-017-0454-3. URL `https://doi.org/10.1007/s11228-017-0454-3`

[4] van Ackooij, W., Berge, V., de Oliveira, W., Sagastizábal, C.: Probabilistic optimization via approximate *p*-efficient points and bundle methods. Comput. Oper. Res. **77**, 177–193 (2017). DOI 10.1016/j.cor.2016.08.002. URL `https://doi.org/10.1016/j.cor.2016.08.002`

[5] van Ackooij, W., Demassey, S., Javal, P., Morais, H., de Oliveira, W., Swaminathan, B.: A bundle method for nonsmooth DC programming with application to chance-constrained problems. Comput. Optim. Appl. **78**(2), 451–490 (2021). DOI 10.1007/s10589-020-00241-8. URL `https://doi.org/10.1007/s10589-020-00241-8`

[6] van Ackooij, W., Henrion, R.: Gradient formulae for nonlinear probabilistic constraints with Gaussian and Gaussian-like distributions. SIAM J. Optim. **24**(4), 1864–1889 (2014). DOI 10.1137/130922689. URL `https://doi.org/10.1137/130922689`

[7] van Ackooij, W., Henrion, R.: (Sub-)gradient formulae for probability functions of random inequality systems under Gaussian distribution. SIAM/ASA J. Uncertain. Quantif. **5**(1), 63–87 (2017). DOI 10.1137/16M1061308. URL `https://doi.org/10.1137/16M1061308`

[8] van Ackooij, W., Henrion, R., Möller, A., Zorgati, R.: Joint chance constrained programming for hydro reservoir management. Optim. Eng. **15**(2), 509–531 (2014). DOI 10.1007/s11081-013-9236-4. URL `https://doi.org/10.1007/s11081-013-9236-4`

[9] van Ackooij, W., Pérez-Aros, P.: Gradient formulae for nonlinear probabilistic constraints with non-convex quadratic forms. J. Optim. Theory Appl. **185**(1), 239–269 (2020). DOI 10.1007/s10957-020-01634-9. URL `https://doi.org/10.1007/s10957-020-01634-9`

[10] van Ackooij, W., Sagastizábal, C.: Constrained bundle methods for upper inexact oracles with application to joint chance constrained energy problems. SIAM J. Optim. **24**(2), 733–765 (2014). DOI 10.1137/120903099. URL `https://doi.org/10.1137/120903099`

[11] Ahmed, S., Luedtke, J., Song, Y., Xie, W.: Nonanticipative duality, relaxations, and formulations for chance-constrained stochastic programs. Math. Program. **162**(1-2, Ser. A), 51–81 (2017)

[12] Balasubramanian, K., Ghadimi, S., Nguyen, A.: Stochastic multi-level composition optimization algorithms with level-independent convergence rates. arXiv preprint arXiv:2008.10526 (2020)

[13] Bardakci, I., Lagoa, C.M.: Distributionally robust portfolio optimization. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 1526–1531. IEEE (2019)

[14] Bardakci, I.E., Lagoa, C., Shanbhag, U.V.: Probability maximization with random linear inequalities: Alternative formulations and stochastic approximation schemes. In: 2018 Annual American Control Conference, ACC 2018, Milwaukee, WI, USA, June 27-29, 2018, pp. 1396–1401. IEEE (2018)

[15] Bienstock, D., Chertkov, M., Harnett, S.: Chance-constrained optimal power flow: Risk-aware network control under uncertainty. SIAM Review **56**(3), 461–495 (2014)

[16] Bobkov, S.G.: Convex bodies and norms associated to convex measures. Probability theory and related fields **147**(1-2), 303–332 (2010)

[17] Brascamp, H.J., Lieb, E.H.: On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. J. Functional Analysis **22**(4), 366–389 (1976). DOI 10.1016/0022-1236(76)90004-5. URL `https://doi.org/10.1016/0022-1236(76)90004-5`

[18] Burke, J.V., Chen, X., Sun, H.: The subdifferential of measurable composite max integrands and smoothing approximation. Math. Program. **181**(2, Ser. B), 229–264 (2020). DOI 10.1007/s10107-019-01441-9. URL `https://doi.org/10.1007/s10107-019-01441-9`

[19] Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Math. Program. **134**(1), 127–155 (2012)

[20] Campi, M.C., Garatti, S.: A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. J. Optim. Theory Appl. **148**(2), 257–280 (2011)

[21] Charnes, A., Cooper, W.W.: Chance-constrained programming. Management Sci. **6**, 73–79 (1959/1960)

[22] Charnes, A., Cooper, W.W., Symonds, G.H.: Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. Management Science **4**(3), 235–263 (1958). URL `https://EconPapers.repec.org/RePEc:inm:ormnsc:v:4:y:1958:i:3:p:235-263`

[23] Chen, L.: An approximation-based approach for chance-constrained vehicle routing and air traffic control problems. In: Large scale optimization in supply chains and smart manufacturing, *Springer Optim. Appl.*, vol. 149, pp. 183–239. Springer, Cham (2019)

[24] Chen, T., Sun, Y., Yin, W.: Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. IEEE Transactions on Signal Processing **69**, 4937–4948 (2021)

[25] Chen, W., Sim, M., Sun, J., Teo, C.P.: From cvar to uncertainty set: Implications in joint chance-constrained optimization. Operations research **58**(2), 470–485 (2010)

[26] Cheng, J., Chen, R.L.Y., Najm, H.N., Pinar, A., Safta, C., Watson, J.P.: Chance-constrained economic dispatch with renewable energy and storage. Comput. Optim. Appl. **70**(2), 479–502 (2018). DOI 10.1007/s10589-018-0006-2. URL `https://doi.org/10.1007/s10589-018-0006-2`

[27] Clarke, F.H.: Optimization and nonsmooth analysis, *Classics in Applied Mathematics*, vol. 5, second edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1990). DOI 10.1137/1.9781611971309. URL `https://doi.org/10.1137/1.9781611971309`

[28] Cui, Y., Liu, J., Pang, J.S.: Nonconvex and nonsmooth approaches for affine chance constrained stochastic programs. submitted (2020)

[29] Curtis, F.E., Wächter, A., Zavala, V.M.: A sequential algorithm for solving nonlinear optimization problems with chance constraints. SIAM J. Optim. **28**(1), 930–958 (2018)

[30] Ermoliev, Y.: Methods of Stochastic Programming. Monographs in Optimization and OR, Nauka, Moscow (1976)

[31] Fiacco, A.V., McCormick, G.P.: The sequential maximization technique (SUMT) without parameters. Operations Res. **15**, 820–827 (1967). DOI 10.1287/opre.15.5.820. URL `https://doi.org/10.1287/opre.15.5.820`

[32] Fiacco, A.V., McCormick, G.P.: Nonlinear programming: Sequential unconstrained minimization techniques. John Wiley and Sons, Inc., New York-London-Sydney (1968)

[33] Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming **156**(1-2), 59–99 (2016)

[34] Ghadimi, S., Ruszczynski, A., Wang, M.: A single timescale stochastic approximation method for nested stochastic optimization. SIAM Journal on Optimization **30**(1), 960–979 (2020)

[35] Gicquel, C., Cheng, J.: A joint chance-constrained programming approach for the single-item capacitated lot-sizing problem with stochastic demand. Ann. Oper. Res. **264**(1-2), 123–155 (2018). DOI 10.1007/s10479-017-2662-5. URL `https://doi.org/10.1007/s10479-017-2662-5`

[36] Göttlich, S., Kolb, O., Lux, K.: Chance-constrained optimal inflow control in hyperbolic supply systems with uncertain demand. Optimal Control Appl. Methods **42**(2), 566–589 (2021). DOI 10.1002/oca.2689. URL `https://doi.org/10.1002/oca.2689`

[37] Guo, G., Zephyr, L., Morillo, J., Wang, Z., Anderson, C.L.: Chance constrained unit commitment approximation under stochastic wind energy. Comput. Oper. Res. **134**, Paper No. 105398, 13 (2021). DOI 10.1016/j.cor.2021.105398. URL `https://doi.org/10.1016/j.cor.2021.105398`

[38] Guo, S., Xu, H., Zhang, L.: Convergence analysis for mathematical programs with distributionally robust chance constraint. SIAM J. Optim. **27**(2), 784–816 (2017). DOI 10.1137/15M1036592. URL `https://doi.org/10.1137/15M1036592`

[39] Henrion, R.: Optimierungsprobleme mit wahrscheinlichkeitsrestriktionen: Modelle, struktur, numerik. Lecture notes p. 43 (2010)

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

[41] Jalilzadeh, A., Shanbhag, U.V., Blanchet, J.H., Glynn, P.W.: Optimal smoothed variable sample-size accelerated proximal methods for structured nonsmooth stochastic convex programs. arXiv preprint arXiv:1803.00718 (2018)

[42] Lagoa, C.M., Li, X., Sznaier, M.: Probabilistically constrained linear programs and risk-adjusted controller design. SIAM Journal on Optimization **15**(3), 938–951 (2005)

[43] Lasserre, J.B.: Level sets and nongaussian integrals of positively homogeneous functions. IGTR **17**(1) (2015)

[44] Lei, J., Shanbhag, U.V.: Asynchronous variance-reduced block schemes for composite nonconvex stochastic optimization: block-specific steplengths and adapted batch-sizes. Optimization Methods and Software **0**(0), 1–31 (2020)

[45] Lian, X., Wang, M., Liu, J.: Finite-sum composition optimization via variance reduced gradient descent. In: Artificial Intelligence and Statistics, pp. 1159–1167. PMLR (2017)

[46] Lieb, E., Loss, M.: Analysis. Crm Proceedings & Lecture Notes. American Mathematical Society (2001). URL `https://books.google.com/books?id=Eb_7oRorXJgC`

[47] Luedtke, J., Ahmed, S.: A sample approximation approach for optimization with probabilistic constraints. SIAM J. Optim. **19**(2), 674–699 (2008)

[48] Markowitz, H.: Portfolio selection. The Journal of Finance **7**(1), 77–91 (1952)

[49] Miller, B.L., Wagner, H.M.: Chance constrained programming with joint constraints. Operations Research **13**(6), 930–945 (1965)

[50] Morozov, A., Shakirov, S.: Introduction to integral discriminants. Journal of High Energy Physics **2009**(12), 002 (2009)

[51] Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on optimization **19**(4), 1574–1609 (2009)

[52] Nemirovski, A., Shapiro, A.: Convex approximations of chance constrained programs. SIAM Journal on Optimization **17**(4), 969–996 (2006)

[53] Norkin, V.I.: The analysis and optimization of probability functions (1993)

[54] Pagnoncelli, B.K., Ahmed, S., Shapiro, A.: Sample average approximation method for chance constrained programming: theory and applications. J. Optim. Theory Appl. **142**(2), 399–416 (2009). DOI 10.1007/s10957-009-9523-6. URL `https://doi.org/10.1007/s10957-009-9523-6`

[55] Pagnoncelli, B.K., Ahmed, S., Shapiro, A.: Sample average approximation method for chance constrained programming: theory and applications. J. Optim. Theory Appl. **142**(2), 399–416 (2009)

[56] Peña-Ordieres, A., Luedtke, J.R., Wächter, A.: Solving chance-constrained problems via a smooth sample-based nonlinear approximation. `https://arxiv.org/abs/1905.07377` (2019)

[57] Pflug, G.C., Weisshaupt, H.: Probability gradient estimation by set-valued calculus and applications in network design. SIAM J. Optim. **15**(3), 898–914 (2005). DOI 10.1137/S1052623403431639. URL `https://doi.org/10.1137/S1052623403431639`

[58] Polyak, B.T.: New stochastic approximation type procedures. Automat. i Telemekh **7**(98-107), 2 (1990)

[59] Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization **30**(4), 838–855 (1992)

[60] Prékopa, A.: A class of stochastic programming decision problems. Math. Operationsforsch. Statist. **3**(5), 349–354 (1972). DOI 10.1080/02331937208842107. URL `https://doi.org/10.1080/02331937208842107`

[61] Prékopa, A.: On logarithmic concave measures and functions. Acta Scientiarum Mathematicarum **34**, 335–343 (1973)

[62] Prékopa, A.: Probabilistic programming. In: Stochastic programming, *Handbooks Oper. Res. Management Sci.*, vol. 10, pp. 267–351. Elsevier Sci. B. V., Amsterdam (2003). DOI 10.1016/S0927-0507(03)10005-9. URL `https://doi.org/10.1016/S0927-0507(03)10005-9`

[63] Prékopa, A.: Stochastic programming, vol. 324. Springer Science & Business Media (2013)

[64] Prékopa, A., Szántai, T.: Flood control reservoir system design using stochastic programming. In: Mathematical programming in use, pp. 138–151. Springer (1978)

[65] Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics pp. 400–407 (1951)

[66] Royset, J.O., Polak, E.: Extensions of stochastic optimization results to problems with system failure probability functions. J. Optim. Theory Appl. **133**(1), 1–18 (2007). DOI 10.1007/s10957-007-9178-0. URL `https://doi.org/10.1007/s10957-007-9178-0`

[67] Scholtes, S.: Introduction to piecewise differentiable equations. Springer Science & Business Media (2012)

[68] Shanbhag, U.V., Blanchet, J.H.: Budget-constrained stochastic approximation. In: Proceedings of the 2015 Winter Simulation Conference, Huntington Beach, CA, USA, December 6-9, 2015, pp. 368–379 (2015)

[69] Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on stochastic programming: modeling and theory. SIAM (2009)

[70] Sun, Y., Aw, G., Loxton, R., Teo, K.L.: Chance-constrained optimization for pension fund portfolios in the presence of default risk. European J. Oper. Res. **256**(1), 205–214 (2017). DOI 10.1016/j.ejor.2016.06.019. URL `https://doi.org/10.1016/j.ejor.2016.06.019`

[71] Uryasev, S.: Derivatives of probability functions and integrals over sets given by inequalities. pp. 197–223 (1994). DOI 10.1016/0377-0427(94)90388-3. URL `https://doi.org/10.1016/0377-0427(94)90388-3`. Stochastic programming: stability, numerical methods and applications (Gosen, 1992)

[72] Uryasev, S.: Derivatives of probability functions and some applications. pp. 287–311 (1995). DOI 10.1007/BF02031712. URL `https://doi.org/10.1007/BF02031712`. Stochastic programming (Udine, 1992)

[73] Uryasev, S.: Derivatives of probability functions and some applications. pp. 287–311 (1995). DOI 10.1007/BF02031712. URL `https://doi.org/10.1007/BF02031712`. Stochastic programming (Udine, 1992)

[74] Wang, M., Fang, E.X., Liu, H.: Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. Mathematical Programming **161**(1-2), 419–449 (2017)

[75] Wang, M., Liu, J., Fang, E.X.: Accelerating stochastic composition optimization. The Journal of Machine Learning Research **18**(1), 3721–3743 (2017)

[76] Xie, Y., Shanbhag, U.V.: SI-ADMM: A stochastic inexact ADMM framework for stochastic convex programs. IEEE Trans. Autom. Control. **65**(6), 2355–2370 (2020)

[77] Yadollahi, E., Aghezzaf, E.H., Raa, B.: Managing inventory and service levels in a safety stock-based inventory routing system with stochastic retailer demands. Appl. Stoch. Models Bus. Ind. **33**(4), 369–381 (2017). DOI 10.1002/asmb.2241. URL `https://doi.org/10.1002/asmb.2241`

[78] Yang, S., Wang, M., Fang, E.X.: Multilevel stochastic gradient methods for nested composition optimization. SIAM Journal on Optimization **29**(1), 616–659 (2019)