Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

SAMPLE AVERAGE APPROXIMATION OVER FUNCTION SPACES: STATISTICAL CONSISTENCY AND RATE OF CONVERGENCE

Zihe Zhou
Harsha Honnappa

School of Industrial Engineering
Purdue University
150 N. University St
West Lafayette, IN 47907, USA

Raghu Pasupathy

Department of Statistics
Purdue University
150 N. University St
West Lafayette, IN, 47907
and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras
Chennai, 600036, INDIA

ABSTRACT

This paper considers sample average approximation (SAA) of a general class of stochastic optimization problems over a function space constraint set and driven by "regulated" Gaussian processes. We establish statistical consistency by proving equiconvergence of the SAA estimator via a sophisticated sample complexity result. Next, recognizing that implementation over such infinite-dimensional spaces is possible only if numerical optimization is performed over a finite-dimensional subspace of the constraint set, and if sample paths of the driving process can be generated over a finite grid, we identify the decay rate of the SAA estimator's expected optimality gap as a function of the optimization error, Monte Carlo sampling error, path generation approximation error, and subspace projection error.

1 INTRODUCTION

We consider infinite-dimensional stochastic optimization problems of the form

$$\begin{aligned} \min \quad & J(F) = \int_C \tilde{J}(x)d\pi_F^\Gamma(x) = \int_C \left(\tilde{J}\circ\Gamma\right)(F+z)d\pi_0(z) \\ \text{s.t.} \quad & F \in \mathscr{F}, \end{aligned} \tag{OPT}$$

where $\tilde{J}: C \to \mathbb{R}$ is some "cost" functional, C is the space of $\mathbb{R}$-valued continuous functions with domain $[0,T]$ and equipped with the supremum norm, $\mathscr{F} \subset C$ is a subspace of C, and $F \in \mathscr{F}$ is the "decision variable". The functional $\tilde{J}$ takes as argument "paths" $X^F := \Gamma(F+Z)$, where $\Gamma: C \to C$ is a continuous "regulator" map that confines $Z+F$ to a subdomain of C, Z is a C-valued Gaussian random variable that induces a measure π_0 on the Borel space $(C,\mathscr{C})$ and X^F induces the 'push-forward' measure π_F^Γ (see Definition 8 below).

1.1 Motivating Examples

Roughly speaking, the problem formulation in (OPT) asks for the extent to which Gaussian paths Z should be (additively) shifted so that the resulting cost $\tilde{J}\circ\Gamma(F+Z)$ is minimized in expectation. And, as the following examples suggest, (OPT) subsumes a multitude of problems in operations research, optimal

978-1-6654-7661-4/22/$31.00 ©2022 IEEE

control and machine learning, when formulated as stochastic optimization problems driven by Gaussian processes.

Example 1 Let $\mathscr{F} = W_0^{1,2}$ the Sobolev space consisting of $\mathbb{R}$-valued absolutely continuous functions with L^2-integrable derivatives and initial value 0. If $Z = \sigma B$, where B is a Wiener process with measure π_0 and $\sigma > 0$, then $(C, \mathscr{F}, \pi_0)$ is the classic Cameron-Martin-Wiener space. Let Γ be the so-called Skorokhod regulator map (Chen and Yao 2001, Ch. 5), which satisfies $\Gamma(x)(\cdot) = x(\cdot) + \sup_{0\le s\le \cdot} \max\{-(x(s)), 0\}$ for any function $x \in C$. Then, the random variable X^F is a so-called *reflected Brownian motion* (RBM) with drift F. Consider a cost functional over $x \in C$, $x \mapsto \tilde{J}(x) := a_1 \int_0^T g(x(s))ds + a_2 G(x(T))$, where $(a_1, a_2) \in \mathbb{R}^2$, and $g : \mathbb{R} \to \mathbb{R}$ and $G : \mathbb{R} \to \mathbb{R}$ are well-defined functions. The corresponding optimization problem represents a class of 'open-loop' optimal control problems over $W_0^{1,2}$, driven by an RBM. This class of problems arises in nonstationary queueing network control, scheduling and inventory control.

Example 2 Suppose that $X^F = F + Z$, $Z = \sigma B$ and $\mathscr{F} := \{F \in C : \mathscr{L}_F p = 0,\ p(0) = \delta_{x_0}\}$, where $\mathscr{L}_F \equiv \partial_t + F'(t)\partial_x - \frac{\sigma^2}{2}\partial_{xx}$ is the Fokker-Planck partial differential operator corresponding to X^F, $F'(t) = dF(t)/dt$ and $p(0) = \delta_{x_0}$ is the initial condition. The solution of this equation is the marginal density $p_F(t,\cdot)$ of $X^F(t)$. Consider the cost functional $\tilde{J}(x) := \log\left(p_F(T, x(T))/p_0(x(T))\right)$ where $p_0(\cdot)$ is a *reference* density function, and the optimization problem $\min_{F\in\mathscr{F}} \mathbb{E}[\tilde{J}\left(X^F\right)] = \int_{\mathbb{R}^d} \log\left(\frac{p_F(T,z)}{p_0(z)}\right) p_F(T,z)dx$. Roughly speaking, this problem computes arbitrary Gaussian approximations $p_F(T,\cdot)$ to $p_0(\cdot)$ by minimizing the Kullback-Leibler divergence between these densities. This formulation underlies the use of so-called stochastic normalizing flows for variational inference (VI) in probabilistic machine learning.

1.2 Method and Overview of Results

Analytical solutions to (OPT) are accessible only in a few special cases, and simulation optimization is natural and almost imperative. This paper explores the use of sample average approximation (SAA) toward solving (OPT). Specifically, recall that the SAA approximation of (OPT) is defined as

$$\begin{aligned} \min \quad & \left\{ J_N(F) := \frac{1}{N}\sum_{j=1}^{N} (\tilde{J}\circ\Gamma)(Z_j + F) \right\}, \quad Z_j \overset{\text{iid}}{\sim} \pi_0 \\ \text{s.t.} \quad & F \in \mathscr{F}, \end{aligned} \tag{MC-OPT}$$

where the random variables $\mathbf{Z} := (Z_1, \cdots, Z_N)$ are independent and identically distributed (iid). The main idea in SAA is the recognition that since (MC-OPT) is a deterministic optimization problem that in a sense approximates (OPT), a solution to (MC-OPT) might reasonably be expected to approximate a solution to (OPT). While this idea is sound in principle, the context raises a number of statistical questions that need resolution. Accordingly, this paper establishes the following two "first order" results.

1. **Asymptotic Consistency.** We first demonstrate that the optimal value and optimizers of (MC-OPT) are asymptotically consistent (in the number of samples N from π_0) by proving convergence in probability. Our approach to this first establishes a novel uniform equiconvergence result over function spaces by showing that the Gaussian complexity of the SAA estimator of the objective is inversely proportional to $\sqrt{N}$ (for every N), assuming the diameter of the constraint set $\mathscr{F}$ is bounded.
2. **Rate of Convergence.** The Gaussian paths $\{Z_j, j \ge 1\}$ from π_0 in (MC-OPT) cannot in general be sampled directly. For instance, if Z_1 is a Brownian motion, then sample paths may (only) be approximated using Euler-Maruyama or Euler-Milstein schemes (Asmussen and Glynn 2007). In other words, the problem in (MC-OPT) is "fictitious" from the standpoint of computation and a further approximation to (MC-OPT) is necessary for implementation. Furthermore, since $\mathscr{F}$ might be infinite dimensional, the solving of (MC-OPT) must be performed (only) over a finite-dimensional subspace of the constraint set $\mathscr{F}$ to allow computation using a method such as gradient descent. Our second main result is a convergence rate result that accounts for the above two sources of error, and quantifies the expected decay rate of the true

optimality gap of a solution obtained by executing *mirror descent* on a finite-dimensional approximation of (MC-OPT) generated using approximations to $\{Z_j, j \geq 1\}$. The convergence rate result clarifies the relationship between four sources of error: (i) numerical optimization error due to the use of an iterative scheme such as mirror descent; (ii) Monte Carlo sampling error; (iii) path approximation error due to "time" discretization; and (iv) projection error due to the use of a finite-dimensional subspace in lieu of $\mathscr{F}$.

1.3 Literature

The use of SAA methodology toward stochastic optimization in $\mathbb{R}^d$ has an extensive literature, as comprehensively surveyed in (Shapiro 2003; Shapiro et al. 2009; Kim et al. 2014; Pasupathy 2010; Banholzer et al. 2019). Corresponding results in the non-Euclidean setting appear in (Dupačová and Wets 1988; Robinson 1996)) where the feasible $\mathscr{F}$ is assumed to be finite dimensional. Especially relevant to what we present here is the extensive treatment of consistency and uniform rate properties of M-estimators (van de Geer 2000; Bose 1998) in normed spaces. (A solution to (OPT) is indeed an M-estimator.) However, we are not aware of infinite-dimensional SAA rate results in the typical context where the SAA estimator is not available in "closed form" but is computed using an iterative optimization technique. Nonetheless, as pointed out in the introduction, there are a number of problems that require optimization over function spaces, wherein SAA is a natural approximation to such problems.

2 PRELIMINARIES

In this section, we discuss mathematical preliminaries including key definitions, assumptions, and notation.

2.1 Key Defintions

In the definitions that follow the space $\mathscr{F}$ is a subspace of a *normed* space X over $\mathbb{R}$. Recall that a Banach space is a complete normed space.

Definition 1 (Linear Functionals) $x : \mathscr{F} \to \mathbb{R}$ is called a linear functional on the (real) normed space $\mathscr{F}$ if

$$x(\alpha F) = \alpha x(F), \alpha \in \mathbb{R}; \quad x(F_1 + F_2) = x(F_1) + x(F_2), F_1, F_2 \in \mathscr{F}.$$

A linear functional $x : \mathscr{F} \to \mathbb{R}$ is said to be a *bounded linear functional* if

$$\|x\| := \sup\{|x(F)| : \|F\| = 1, F \in \mathscr{F}\} < \infty.$$

It can be shown that $x : \mathscr{F} \to \mathbb{R}$ is a bounded linear functional if and only if x is continuous on $\mathscr{F}$, and that continuity of x at any point $F_0 \in \mathscr{F}$ implies boundedness of x. (It is important that $x : \mathscr{F} \to \mathbb{R}$ being bounded does not mean $\sup_{F \in \mathscr{F}} |x(F)| < \infty$; indeed, it is routinely the case that $\|x\| < \infty$ but $\sup_{F \in \mathscr{F}} |x(F)| = \infty$.)

Definition 2 (Dual Space, Adjoint Space, Conjugate Space) The space $\mathscr{F}^*$ of linear functionals on $\mathscr{F}$ is called the algebraic dual space of $\mathscr{F}$. $\mathscr{F}^*$ should be distinguished from the dual space $\mathscr{F}'$, which is the space of of *bounded* linear functionals on $\mathscr{F}$. $\mathscr{F}^*$ is sometimes also called the *adjoint space* or the *conjugate space* of $\mathscr{F}$.

Definition 3 (Dual Norm) The (operator) norm of the functional $T \in \mathscr{F}^*$ is called the *dual norm* or *conjugate norm* of T:

$$\begin{aligned} \|T\|_* &:= \sup\left\{\frac{|Tx|}{\|x\|} : x \in \mathscr{F}, x \neq 0\right\} \\ &= \sup\{|Tx| : x \in \mathscr{F}, \|x\| = 1\}. \end{aligned} \tag{1}$$

Definition 4 (Right and Left Directional Derivatives) The *right directional derivative* $J'_+(F,v)$ and the *left directional derivative* $J'_-(F,v)$ of the functional $J : \mathscr{F} \to \mathbb{R}$ at the point $F \in \mathscr{F}$ are defined as

$$J'_+(F,v) := \lim_{t\to 0^+} \frac{1}{t}\left(J(F+tv) - J(F)\right);$$
$$J'_-(F,v) := \lim_{t\to 0^-} \frac{1}{t}\left(J(F+tv) - J(F)\right).$$

Definition 5 (Gâteaux and Fréchet Differentiability) The functional $J : \mathscr{F} \to \mathbb{R}$ is *Gâteaux differentiable* at $F \in \mathscr{F}$ if the limit

$$S_J(F)(v) := \lim_{t\to 0} \frac{1}{t}\left(J(F+tv) - J(F)\right) \tag{2}$$

exists for each $v \in \mathscr{F}$, and $S_J(F) \in \mathscr{F}'$, that is, $S_J(F)(v) : \mathscr{F} \to \mathbb{R}$ is a bounded linear functional.

The functional $J : \mathscr{F} \to \mathbb{R}$ is *Fréchet differentiable* if the limit in (2) is uniform in v, that is,

$$|J(F+v) - (J(F) + S_J(F)(v))| = o(\|v\|), \quad v \in \mathscr{F}.$$

From Definition 4 and Defintion 5, we see that Fréchet differentiability $\Rightarrow$ Gâteaux differentiability $\Rightarrow$ Directional Derivative Existence. Also, if $\mathscr{F}$ is finite-dimensional and J is Lipschitz in some neighborhood of $F \in \mathscr{F}$, then J is Fréchet differentiable at F if and only if it is Gâteaux differentiable at F.

Definition 6 (Subgradient and Subdifferentials of a Convex Functional) The functional $J : \mathscr{F} \to \mathbb{R}$ is *convex* if for any $\alpha \in [0,1]$, $J(\alpha F_1 + (1-\alpha)F_2) \geq \alpha J(F_1) + (1-\alpha)J(F_2), \quad \forall F_1, F_2 \in \mathscr{F}$. $S_J(F_0) \in \mathscr{F}'$ is called a *subgradient* to J at $F_0 \in \mathscr{F}$ if

$$J(F) \geq J(F_0) + S_J(F_0)(F - F_0). \tag{3}$$

The set $\partial J(F_0)$ of subgradients to J at F_0 is called the *subdifferential* to J at F_0. Convex functionals have a subdifferential structure in the sense that if $J : \mathscr{F} \to \mathbb{R}$ is convex, then $\partial J(F_0) \neq \emptyset$ for each $F_0 \in \mathscr{F}^\circ$; conversely, if $\partial J(F) \neq \emptyset$ for each $F \in \mathscr{F}^\circ$, then J is necessarily a convex functional.

Definition 7 (Mirror Map) Suppose $\bar{\mathscr{D}} \supset \mathscr{F}$ and $\mathscr{D} \cap \mathscr{F} \neq \emptyset$. A map $\psi : \mathscr{D} \to \mathscr{R}$ is called a *mirror map* if it satisfies the following three conditions:

1. ψ is Fréchet differentiable and strongly convex in $\mathscr{D}$;
2. for each $y \in \mathscr{F}^*$, there exists $F \in \mathscr{F}$ such that $\nabla\psi(F) = y$; and
3. $\lim_{F\to\partial\mathscr{D}} \|\psi(F)\|_* \to +\infty$.

Definition 8 (Push-Forward Measure) This paper focuses on stochastic optimization problems defined with respect to *regulated* Gaussian processes, $X^F = \Gamma(Z+F)$, whose paths are confined to a subdomain of C. To define the measure corresponding to such a regulated process, define the shift operator $T_g(x) : (\mathscr{F}, C) \to C$ as $T_g(x) = g + x$, and the push-forward measure corresponding to the shift operator $\pi_g(A) := (T_g)_*(\pi_0)(A) = \pi_0(T_g^{-1}(A))$ for any $A \in \mathscr{C}$. Then, the push-forward measure corresponding to X^F is defined as $\pi_F^\Gamma(A) = \pi_F(\Gamma^{-1}(A))$, for any $A \in \mathscr{C}$.

2.2 Key Assumptions

We now list key assumptions on the cost functional $\tilde{J} : C \to \mathbb{R}$ to be invoked in the results that follow.

Assumption 1 The cost functional $\tilde{J} : C \to \mathbb{R}$ is Fréchet differentiable.

Assumption 2 The cost functional $\tilde{J} : C \to \mathbb{R}$ satisfies

$$|\tilde{J}(x+F_1) - \tilde{J}(x+F_2)| \leq K_x \|F_1 - F_2\|_\infty, \tag{4}$$

where $K_x > 0$ for every sample path $x \in C$, $F_1, F_2 \in \mathscr{F}$ and $\mathbb{E}[K_Z^p] < +\infty$ for some $2 \leq p < +\infty$.

Assumption 3 The cost functional $\tilde{J}$ is L-Lipschitz in $z \in C$, i.e., for any $F \in \mathscr{F}$ and $z, z' \in C$, we have

$$|\tilde{J}(z+F) - \tilde{J}(z'+F)| \leq \kappa \|z - z'\|_\infty. \tag{5}$$

Assumption 4 The cost functional $\tilde{J} : C \to \mathbb{R}$ is sufficiently regular such that the composite functional $\tilde{J} \circ \Gamma : C \to \mathbb{R}$ is integrable, that is, $\int_C \left|(\tilde{J} \circ \Gamma)(x+F)\right| d\pi_0(x) < +\infty$.

Assumption 5 The composition $\tilde{J} \circ \Gamma : C \times \mathscr{F} \to C$ is $L_{\Gamma,Z}$-Lipschitz in F, that is, for any $Z \in C$, $\|\tilde{J} \circ \Gamma(Z+F_1) - \tilde{J} \circ \Gamma(Z+F_2)\| \leq L_{\Gamma,Z} \|F_1 - F_2\|$, where $\mathbb{E}[L^2_{\Gamma,Z}] < \infty$.

This assumption is easily satisfied by the Skorokhod regulator which is 2-Lipschitz continuous in the space C.

We assume that the subspace $\mathscr{F} \subseteq C$ satisfies

Assumption 6 $\mathscr{F}$ has a finite diameter. That is, $diam(\mathscr{F}) := \sup_{F_1, F_2 \in \mathscr{F}} \|F_1 - F_2\|_\infty < +\infty$.

This is a reasonably strong assumption, that is nonetheless satisfied by many problems settings; for instance, if the function class $\mathscr{F}$ is parameterized by a compact set. We also believe that it should be possible to relax this condition, at the expense of more complicated computations.

3 EQUICONVERGENCE AND CONSISTENCY

Our approach to proving consistency is to first establish *equiconvergence* of the SAA functional over the subspace $\mathscr{F}$. For simplicity, at the outset let us assume that Γ is the identity map. We will subsequently observe that the forthcoming results extend to the reflected case. We prove equiconvergence by bounding the Gaussian complexity of the SAA, defined as

$$\mathscr{R}_N(\mathscr{F}) := \mathbb{E}_{g,\pi_0} \left[\sup_{F \in \mathscr{F}} \left\{ \frac{1}{N} \sum_{i=1}^N g_i \tilde{J}(Z_i + F) \right\} \right], \tag{6}$$

where the expectation is taken with respect to the Gaussian random vector $g \sim \mathscr{N}(0, I_{N \times N})$, which is independent of the iid samples $\mathbf{Z} := (Z_1, \cdots, Z_N)$.

Next, define the $\mathbb{R}^N$-valued random field $\mathscr{G}_\cdot(\cdot)$ as $F \mapsto \mathscr{G}_F(\mathbf{Z}) := \left(\tilde{J}(Z_1 + F), \cdots \tilde{J}(Z_N + F)\right)$. For each $\mathbf{z} \in C^N$ define the set $\mathscr{B} \equiv \mathscr{B}(\mathbf{z}) := \{\mathscr{G}_F(\mathbf{z}) : F \in \mathscr{F}\} \subseteq \mathbb{R}^N$, and the pseudometric $d : \mathscr{B} \times \mathscr{B} \to [0, \infty)$, given by

$$d(x,y) = \frac{1}{\sqrt{N}} \|K(\mathbf{z})\|_p \|F_x - F_y\|_\infty, \tag{7}$$

where $F_x, F_y \in \mathscr{F}$ correspond to x, y (respectively) through the map $\mathscr{G}_\cdot$, for any $x \in \mathbb{R}^N$ $\|x\|_p := \left(\sum_{i=1}^N |x_i|^p\right)^{1/p}$ and K_{z_i} are the Lipschitz variables defined in Assumption 2.

Let $\{Y_F(\mathbf{Z}) : F \in \mathscr{F}\}$ be the real-valued random field defined as $Y_F(\mathbf{Z}) := \frac{1}{\sqrt{N}} \sum_{i=1}^N g_i \tilde{J}(Z_i + F)$, where g is a N-dimensional standard Gaussian random vector, as before. The next lemma shows that $\{Y_F(\mathbf{Z}) : F \in \mathscr{F}\}$ satisfies a sub-Gaussian concentration inequality, conditioned on $\mathbf{Z}$.

Lemma 1 For any $F, G \in \mathscr{F}$ such that $\|F - G\|_\infty \neq 0$ we have $\mathbb{P}\left(|Y_F(\mathbf{z}) - Y_G(\mathbf{z})| > u | \mathbf{Z} = \mathbf{z}\right) \leq 2 \exp\left(-\frac{u^2}{2L^2 d(\mathscr{G}_F(\mathbf{z}), \mathscr{G}_G(\mathbf{z}))^2}\right)$, where $d(\cdot, \cdot)$ is defined in (7) and $L := \sup_{y \in \mathbb{R}^N : \|y\|_2 = 1} \|y\|_q = 1$ for $q \geq 2$.

Proof. Fix $F, G \in \mathscr{F}$ such that $F \neq G$. By Hölder's inequality we have $|Y_F(\mathbf{Z}) - Y_G(\mathbf{Z})| \leq \frac{1}{\sqrt{N}} \|g\|_q \|\mathscr{G}_F(\mathbf{Z}) - \mathscr{G}_G(\mathbf{Z})\|_p$, where $\frac{1}{p} + \frac{1}{q} = 1$ and $q \geq 2$. Next, following Assumption 2, we have

$$\|\mathscr{G}_F - \mathscr{G}_G\|_p = \left(\sum_{i=1}^N \left(\tilde{J}(Z_i + F) - \tilde{J}(Z_i + G)\right)^p \right)^{1/p} \leq \left(\sum_{i=1}^N |K_{Z_i}|^p \|F - G\|_\infty^p \right)^{1/p} = \|F - G\|_\infty \|K(\mathbf{Z})\|_p.$$

It follows that

$$\mathbb{P}\left(|Y_F(\mathbf{z}) - Y_G(\mathbf{z})| > u \middle| \mathbf{Z} = \mathbf{z}\right) \leq \mathbb{P}\left(\|g\|_q > \frac{u\sqrt{N}}{\|K(\mathbf{z})\|_p \|F - G\|_\infty} \middle| \mathbf{Z} = \mathbf{z}\right). \tag{8}$$

It is straightfoward to see that $x \mapsto \|x\|_q$ is a Lipschitz function from $\mathbb{R}^N$ to $\mathbb{R}$. Then, by (Boucheron, Lugosi, and Massart 2013, Theorem 5.6), $\|g\|_q$ satisfies the sub-Gaussian concentration inequality $\mathbb{P}(\|g\|_q > \varepsilon) \leq 2\exp\left(-\frac{\varepsilon^2}{2L^2}\right)$, where L is defined above. Applying this to (8) completes the proof. □

Next, we show that an ε-cover under the pseudometric can be "translated" into a corresponding ε-cover under the supremum-norm.

Lemma 2 Fix $\varepsilon > 0$. Let $\mathbf{z} = (z_1, \cdots, z_N) \in C^N$ and suppose $B_1, \cdots, B_l \subset \mathbb{R}^N$ is an ε-cover of $\mathscr{B} = \{\mathscr{G}_F(\mathbf{z}) : F \in \mathscr{F}\}$ under the pseudometric (7). Then, there exist subsets $B'_1, \cdots, B'_l$ that form an ε'-cover of $\mathscr{F}$ under the supremum norm $\|\cdot\|_\infty$, with $\varepsilon' = \frac{\varepsilon\sqrt{N}}{\|K(\mathbf{z})\|_p}$.

Proof. By definition, $B_i = \{y \in \mathscr{B} : d(y_i, y) \leq \varepsilon\}$ for some $y_i \in \mathscr{B}$. Consider the set $\{F \in \mathscr{F} : \mathscr{G}_F(\mathbf{z}) \in B_i\} =: \tilde{B}_i$. For any $F \in \tilde{B}_i$, we have $d(y_i, \mathscr{G}_F(\mathbf{z})) = \frac{1}{\sqrt{N}}\|K(\mathbf{z})\|_p \|F_{y_i} - F\|_\infty \leq \varepsilon$. It follows that $\|F_{y_i} - F\|_\infty \leq \frac{\varepsilon\sqrt{N}}{\|K(\mathbf{z})\|_p} = \varepsilon'$.

Now, let $F' \in \mathscr{F} \setminus \cup_{i=1}^l \tilde{B}_i$. It follows that $\min_{1 \leq i \leq l} \|F_{y_i} - F'\| > \varepsilon'$, implying that $d(y_i, \mathscr{G}_{F'}(\mathbf{z})) > \varepsilon$. Therefore, $\mathscr{G}_{F'}(\mathbf{z}) \notin \cup_{i=1}^l B_i$. But, this is a contradiction since $B_1, \cdots, B_l$ is an ε-cover of $\mathscr{B}$, implying that $\mathscr{F} \setminus \cup_{i=1}^l \tilde{B}_i = \emptyset$. □

The proof of equiconvergence in Theorem 2 below follows as a consequence of Proposition 1 and Proposition 2 below.

Proposition 1 Suppose Assumption 1 and Assumption 6 hold. Furthermore, suppose that $\log N(\varepsilon, \mathscr{F}, \|\cdot\|_\infty) \leq \varepsilon^{-1/\alpha}$ for some $\alpha > 1$ and $\varepsilon > 0$. Then, for any $F_0 \in \mathscr{F}$, there exists a constant $0 < \mathtt{C} < +\infty$ such that

$$\mathbb{E}_g\left[\sup_{F \in \mathscr{F}} |Y_F - Y_{F_0}| \middle| \mathbf{Z} = \mathbf{z}\right] \leq \frac{\mathtt{C}\|K(\mathbf{z})\|_p}{\sqrt{N}} \left(\frac{1}{2} diam(\mathscr{F})\right)^{\frac{\alpha-1}{\alpha}}. \tag{9}$$

Proof. It is straightforward to see that $\{Y_F(\mathbf{Z}) : F \in \mathscr{F}\}$ is a separable random field. Further, by Lemma 1 $\{Y_F(\mathbf{Z}) : F \in \mathscr{F}\}$ is sub-Gaussian. By Assumption 6, and the definition of the pseudometric d, we have $D := \sup_{z_1, z_2 \in \mathscr{B}} d(z_1, z_2) < +\infty$. By Dudley's Theorem for separable random fields it follows that there exists a constant $0 < \mathtt{C}' < +\infty$ such that $\mathbb{E}_g\left[\sup_{F \in \mathscr{F}} |Y_F - Y_{F_0}| \middle| \mathbf{Z} = \mathbf{z}\right] \leq \mathtt{C}' \int_0^{D/2} \sqrt{\log N(\varepsilon, \mathscr{B}, d)} d\varepsilon$. By Lemma 2 it follows that $N(\varepsilon, \mathscr{B}, d) = N(\varepsilon', \mathscr{F}, \|\cdot\|_\infty)$, where $\varepsilon' = \varepsilon\frac{\sqrt{N}}{\|K(\mathbf{z})\|_p}$. Then, changing variables in the integral above to ε', we have $D' = D\sqrt{N}/\|K(\mathbf{z})\|_p = diam(\mathscr{F})$ and

$$\begin{aligned}
\int_0^{D/2} \sqrt{\log N(\varepsilon, \mathscr{B}, d)} d\varepsilon &= \frac{\|K(\mathbf{z})\|_p}{\sqrt{N}} \int_0^{D'/2} \sqrt{\log N(\varepsilon, \mathscr{F}, \|\cdot\|_\infty)} d\varepsilon \\
&\leq \frac{\|K(\mathbf{z})\|_p}{\sqrt{N}} \frac{\alpha}{\alpha - 1} \left(\frac{1}{2} diam(\mathscr{F})\right)^{\frac{\alpha-1}{\alpha}}.
\end{aligned}$$

Note that by Assumption 6 it follows that the right hand side above is finite. Setting $\mathtt{C} = \mathtt{C}'\frac{\alpha}{\alpha-1}$ completes the proof. □

Recall that the sub-Gaussian diameter for a metric probability space $(\mathscr{X}, d, \pi)$ with metric d and measure π is defined as $\Delta^2_{\text{SG}}(\mathscr{X}) := \sigma^*(Y)$ where $\sigma^*(Y)$ is the smallest σ that satisfies $\mathbb{E}\left[e^{\lambda Y}\right] \leq e^{\sigma^2\lambda^2/2}$, $\lambda \in \mathbb{R}$, $Y := \varepsilon d(X, X')$ is the symmetrized distance on the metric space $\mathscr{X}$, $\varepsilon = \pm 1$ with probability 1/2 and X, X' are $\mathscr{X}$-valued random variables with measure π. Consider the following generalization of McDiarmid's inequality.

Theorem 1 (Theorem 1 (Kontorovich 2014)) Let $(\mathscr{X}, d, \pi)$ be a metric space that satisfies $\Delta_{\text{SG}}(\mathscr{X}) < +\infty$, and $\varphi : \mathscr{X}^N \to \mathbb{R}$ is 1-Lipschitz, then $E_\pi[\varphi(\mathbf{Z})] < +\infty$, and $\pi\left(|\varphi(\mathbf{Z}) - \mathbb{E}_\pi[\varphi(\mathbf{Z})]| > t\right) \leq 2\exp\left(-\frac{t^2}{2N\Delta^2_{\text{SG}}(\mathscr{X})}\right)$, where $\mathbf{Z} = (Z_1, \cdots, Z_N)$ is an independent sample drawn from π.

Observe that this result significantly loosens the requirements in McDiarmid's inequality from boundedness to Lipschitz continuity.

Proposition 2 Let $\mathbf{Z} = (Z_1, \cdots, Z_N)$ be N i.i.d. random variables with measure π_0. Suppose the cost function satisfies Assumption 3. Suppose that the metric probability space $(C, \|\cdot\|_\infty, \pi_0)$ satisfies $\Delta_{\text{SG}}(C) < +\infty$. Then for any $F \in \mathscr{F}$ and $\delta > 0$, with probability at least $1 - \delta$, we have

$$J(F) \leq \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i + F) + \mathbb{E}\left[\sup_{G\in\mathscr{F}}\left\{J(G) - \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i + G)\right\}\right] + \left(\frac{2\kappa^2\Delta^2_{\text{SG}}(C)\log(1/\delta)}{N}\right)^{1/2}. \tag{10}$$

Remark: We note that the assumption that $\Delta_{\text{SG}}(C) < +\infty$ is reasonable – for instance, it is satisfied in the case where π_0 is the Wiener measure.

Proof. We start by considering the functional $\varphi : C^N \to \mathbb{R}$ defined as $\varphi(\mathbf{z}) := \sup_{F\in\mathscr{F}}\left\{J(F) - \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(z_i + F)\right\}$, for any $\mathbf{z} \in C^N$. Let $\mathbf{z} = (z_1, \cdots, z_N) \in C^N$, $\mathbf{z}' = (z'_1, \cdots, z'_N) \in C^N$; the metric distance between these vectors of functions is given by $\|\mathbf{z} - \mathbf{z}'\| = \sum_{i=1}^{N}\|z_i - z'_i\|_\infty$. Also, define the sequence of vectors $\mathbf{z}^1 = (z'_1, z_2, \cdots, z_N)$, $\mathbf{z}^2 = (z'_1, z'_2, z_3, \cdots, z_N)$, ..., $\mathbf{z}^N = (z'_1, z'_2, \ldots, z'_N) \equiv \mathbf{z}'$. Using the triangle inequality, it is straightforward to see that

$$\begin{aligned}|\varphi(\mathbf{z}) - \varphi(\mathbf{z}')| &= |\varphi(\mathbf{z}) - \varphi(\mathbf{z}^1) + \varphi(\mathbf{z}^1) - \varphi(\mathbf{z}^2) + \cdots + \varphi(\mathbf{z}^{N-1}) - \varphi(\mathbf{z}')| \\ &\leq |\varphi(\mathbf{z}) - \varphi(\mathbf{z}^1)| + |\varphi(\mathbf{z}^1) - \varphi(\mathbf{z}^2)| + \cdots + |\varphi(\mathbf{z}^{N-1}) - \varphi(\mathbf{z}')|,\end{aligned} \tag{11}$$

where each pair of $\mathbf{z}^{k-1}$ and $\mathbf{z}^k$ differs only by the kth element. Let $\mathbf{z}^k(i)$ represent the i^{th} element of the k^{th} vector and $F^* \in \mathscr{F}$ be the function that achieves the supremum in $\varphi(\mathbf{z})$. For any such pair of vectors, we have

$$\begin{aligned}\left|\varphi(\mathbf{z}^{k-1}) - \varphi(\mathbf{z}^k)\right| = \Bigg|\sup_{F\in\mathscr{F}}&\left\{J(F) - \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(\mathbf{z}^{k-1}(i) + F)\right\} \\ &- \sup_{F\in\mathscr{F}}\left\{\left(J(F) - \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(\mathbf{z}^{k-1}(i) + F)\right) + \frac{1}{N}\left(\tilde{J}(z_k + F) - \tilde{J}(z'_k + F)\right)\right\}\Bigg| \\ \leq \frac{1}{N}&\left|\left(\tilde{J}(z_k + F^*) - \tilde{J}(z'_k + F^*)\right)\right| \leq \frac{\kappa}{N}\|z_k - z'_k\|_\infty.\end{aligned}$$

where the last inequality follows from Assumption 3. Consequently, substituting this into (11) we have

$$\begin{aligned}|\varphi(\mathbf{z}) - \varphi(\mathbf{z}')| &\leq |\varphi(\mathbf{z}) - \varphi(\mathbf{z}^1)| + |\varphi(\mathbf{z}^1) - \varphi(\mathbf{z}^2)| + \cdots + |\varphi(\mathbf{z}^{N-1}) - \varphi(\mathbf{z}')| \\ &\leq \frac{\kappa}{N}\left(\|z_1 - z'_1\|_\infty + \|z_2 - z'_2\|_\infty + \cdots + \|z_N - z'_N\|_\infty\right) = \frac{\kappa}{N}\|\mathbf{z} - \mathbf{z}'\|.\end{aligned} \tag{12}$$

In other words, the functional φ is $\frac{L}{N}$-Lipschitz continuous.

Now, by hypothesis we have $\Delta^2_{\mathrm{SG}}(C) < +\infty$, and therefore applying Theorem 1 we have $\mathbb{P}\left(\varphi - \mathbb{E}(\varphi) > t\right) = \mathbb{P}\left(\frac{N}{L}\left(\varphi - \mathbb{E}(\varphi)\right) > \frac{N}{L}t\right) \le \exp\left(-\frac{Nt^2}{2\kappa^2\Delta^2_{\mathrm{SG}}(C)}\right)$. Now, for any $\delta > 0$, $\exp\left(-\frac{Nt^2}{2L^2\Delta^2_{\mathrm{SG}}(C)}\right) \le \delta$ implies that $t \ge \left(\frac{2\kappa^2\Delta^2_{\mathrm{SG}}(C)\log(1/\delta)}{N}\right)^{1/2}$. Hence, with probability at least $1-\delta$, we have $\varphi < \mathbb{E}(\varphi) + \left(\frac{2\kappa^2\Delta^2_{\mathrm{SG}}(C)\log(1/\delta)}{N}\right)^{1/2}$, which yields the final expression in (10). □

Now, our main sample complexity result follows by combining Proposition 1 and Proposition 2. By taking an expectation with respect to π_0 over (9) it follows that the Gaussian complexity of the function space $\mathscr{F}$ is

$$\mathscr{R}_N(\mathscr{F}) := \frac{\mathtt{C}\,\mathbb{E}\left[\|K(\mathbf{Z})\|_p\right]}{\sqrt{N}}\left(\frac{1}{2}diam(\mathscr{F})\right)^{\frac{\alpha-1}{\alpha}}, \tag{13}$$

provided $\mathbb{E}_{\pi_0}\left[\|K(\mathbf{Z})\|_p\right] < +\infty$; this is a consequence of Assumption 2.

Theorem 2 Let $\mathscr{F} \subseteq C$ satisfy Assumption 6 and suppose that $\log N(\varepsilon, \mathscr{F}, \|\cdot\|_\infty) \le \varepsilon^{-1/\alpha}$ for $\alpha \ge 1$ and $\varepsilon > 0$. Suppose the cost function $\tilde{J}$ satisfies Assumption 1, Assumption 2 (for some $1 \le p < +\infty$) and Assumption 3. Let $\mathbf{Z} = (Z_1, \cdots, Z_N)$ be an i.i.d. sample drawn from π_0. Then, for any $\delta > 0$ and some $1 \le p < +\infty$, with probability at least $1-\delta$, for any $F \in \mathscr{F}$ we have

$$J(F) \le \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i + F) + 2\mathscr{R}_N(\mathscr{F}) + O\left(\sqrt{\frac{\log(1/\delta)}{N}}\right).$$

Proof. We sketch the proof. By standard considerations (see (Bartlett and Mendelson 2002) for instance), it can be shown that $\mathbb{E}\left[\sup_{F\in\mathscr{F}}\left\{J(F) - \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i+F)\right\}\right] \le 2\mathscr{R}_N(\mathscr{F})$. The theorem follows by using this to bound the right hand side in (10). □

Recall from Assumption 5 that the composed functional $\tilde{J}\circ\Gamma$ is $L_{\Gamma,Z}$-Lipschitz continuous. Consequently, the consistency result proved in Theorem 2 holds for the composed functional as well. Theorem 2 yields a uniform convergence (or 'equiconvergence') result for $\tilde{J}$ and, in particular as an immediate consequence we have $|J^* - \tilde{J}^*_N| \xrightarrow{P} 0$ as $N \to \infty$, where $J^* := \inf_{F\in\mathscr{F}} J(F)$ and $\tilde{J}^*_N := \inf_{F\in\mathscr{F}} \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i+F)$. Furthermore, let $\Pi^*_n := \arg\inf_{F\in\mathscr{F}} \frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i+F)$ and $\pi^* := \arg\inf_{F\in\mathscr{F}} J(F)$. Consider two scenarios, $J^* > \tilde{J}^*_N$ and $J^* < \tilde{J}^*_N$. In the former case, $|J^* - \tilde{J}^*_N| < |J(\Pi^*_n) - \tilde{J}^*_N|$. In the latter case, $|J^* - \tilde{J}^*_N| < |\frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i+\pi^*) - J^*|$. Therefore,

$$|J^* - \tilde{J}^*_N| < \max\left\{|J(\Pi^*_n) - \tilde{J}^*_N|, |\frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i+\pi^*) - J^*|\right\} < |J(\Pi^*_n) - \tilde{J}^*_N| + |\frac{1}{N}\sum_{i=1}^{N}\tilde{J}(Z_i+\pi^*) - J^*| \xrightarrow{P} 0$$

as $N \to \infty$ by Theorem 2.

4 RATE OF CONVERGENCE

Let's introduce further notation to keep our exposition clear. Recall that $\mathscr{F}$ is a compact subspace of the space of continuous functions on $[0,T]$. Suppose $F \in \mathscr{F}$ and that we can generate N independent realizations of the process $\{(Z_h(t), t \in [0,T]\}$ with measure $\pi^{\Gamma}_{0,h}$ and having continuous paths and having possible non-differentiabilities over the partition points

$$0 = t_0 < t_1 < t_2 < \cdots < t_{n-1} = T,$$

where

$$h = h(n) := \max\{t_1 - t_0, t_2 - t_1, \ldots, t_n - t_{n-1}\}.$$

Let $\mathscr{F}_n$ denote an n-dimensional ($n < \infty$) closed subspace of $\mathscr{F}$ such that elements in $\mathscr{F}$ can be approached by a sequence of elements in $\mathscr{F}_n$, that is, for every $F \in \mathscr{F}$, there exists $\{F_n, n \geq 1\}, F_n \in \mathscr{F}_n$ such that $\|F_n - F\| \to 0$. An example of $\mathscr{F}_n$ is the span of the first n Legendre polynomials (Kreyszig 1989, pp. 176) on the interval $[0,T]$. More generally, $\mathscr{F}_n$ can be chosen as the span of the first n elements of any Schauder basis of $\mathscr{F}$. (Recall that a sequence $\{P_j, j \geq 1\}$ of vectors in a normed space $\mathscr{F}$ is called a *Schauder basis* of $\mathscr{F}$ if for every $F \in \mathscr{F}$ there is a unique sequence $\{a_j, j \geq 1\}$ of scalars such that $\|F - \sum_{j=1}^{n} a_j P_j\| \to 0$ as $n \to \infty$.) Consequently, we assume that

Assumption 7 The closed finite-dimensional function subspace $\mathscr{F}_n \subset \mathscr{F}$ is such that

$$\psi(n) := \sup_{F \in \mathscr{F}} \|F - \Pi_{\mathscr{F}_n}(F)\| = O(g(n)), \tag{14}$$

where $g(n) \to 0$ as $n \to \infty$.

With the above notation in place, the SAA problem (MC-n-OPT) approximating (OPT) is:

$$\begin{aligned} &\min. \left\{ J_{N,h}(F) := \frac{1}{N} \sum_{j=1}^{N} \tilde{J} \circ \Gamma(Z_{h,j} + F) \right\}, \quad Z_{h,j} \overset{\text{iid}}{\sim} \pi_{0,h} \\ &\text{s.t. } F \in \mathscr{F}_n \end{aligned} \tag{MC-n-OPT}$$

where the measure $\pi_{0,h}$ approximates the measure π_0. For brevity, we will write $Z_{h,j}$ as Z_j in the remainder of this section. To facilitate a basic result that quantifies the quality of the solution to (MC-n-OPT) as an estimator to the solution to (OPT) we assume that

Assumption 8 The random functional $F \mapsto \tilde{J} \circ \Gamma(Z + F)$ is convex in F.

Observe that the problem in (MC-n-OPT) is obtained by replacing the expectation appearing in (OPT) by a Monte Carlo sum obtained by generating N samples of a process $\{X_h^F(t), t \in [0,T]\}$ that approximates the process $\{X^F(t), t \in [0,T]\}$. We define the following optimal values and optimal solution (sets) corresponding to (OPT) and (MC-n-OPT), the existence of which will become evident.

$$\begin{aligned} J^* &:= \inf_{F \in \mathscr{F}} \{J(F)\}; \quad \mathscr{F}^* := \arg\inf_{F \in \mathscr{F}} \{J(F)\} \\ J^*_{N,n} &:= \inf_{F \in \mathscr{F}_n} \{J_{N,h}(F)\}; \quad \mathscr{F}^*_{N,n} := \arg\inf_{F \in \mathscr{F}_n} \{J_{N,h}(F)\}. \end{aligned} \tag{15}$$

It is important that the optimization in (MC-n-OPT) be performed over a finite-dimensional subspace $\mathscr{F}_n$ of $\mathscr{F}$ so as to allow computation using a method such as gradient descent (Nesterov 2004). Also, in (15), notice that we have suppressed the dependence of $J^*_{N,n}$ and $\mathscr{F}^*_{N,n}$ on the partition width h used to generate Monte Carlo samples from the measure $\pi_{0,h}$. A result we present shortly will imply that the sub-space dimension n and the partition width h bear a certain relationship that can be exploited to maximize the decay rate of the expected optimality gap $\mathbb{E}\left[J^*_{N,n} - J^*\right]$.

4.1 Consistency and Rate of the SAA Estimator

We call any solution $F^*_{N,n} \in \mathscr{F}^*_{N,n}$ to (MC-n-OPT) an SAA estimator of the solution to (OPT). An SAA estimator cannot be obtained in "closed form" in general. However, given that (MC-n-OPT) is a deterministic convex optimization problem over a closed finite-dimensional subspace, one of various existing iterative techniques, e.g., mirror descent (Bubeck 2015), can be used to generate a sequence $\{F^*_{N,n,k}, k \geq 1\} \subset \mathscr{F}_{N,n}$ that converges to a point in $\mathscr{F}^*_{N,n}$, that is, $F^*_{N,n,k} \to \mathscr{F}^*_{N,n}$ as $k \to \infty$ for fixed N, n, h. Before we present the main result that characterizes the accuracy of $F^*_{N,n,k}$, we state a lemma that will be invoked.

Lemma 3 Let Assumption 2 and Assumption 6 hold, and suppose there exists $F_0 \in \mathscr{F}$ such that

$$\sigma_0^2(h) := \mathrm{Var}(\tilde{J}(X_h^{F_0})) < \infty; \quad X_h^{F_0} \overset{\text{iid}}{\sim} \pi^{\Gamma}_{F_0,h}. \tag{16}$$

Then,

$$\sup_{F\in\mathscr{F}} \mathrm{Var}(\tilde{J}(X_h^F)) \leq \left(\sigma_0(h) + \mathrm{diam}(\mathscr{F})\sqrt{\mathbb{E}[L^2_{\Gamma,Z}]}\right)^2. \tag{17}$$

Proof. We can write

$$\tilde{J}(X_h^F) = \tilde{J}(X_h^{F_0}) + \tilde{J}(X_h^F) - \tilde{J}(X_h^{F_0}), \tag{18}$$

and due to Assumption 5,

$$\left|\tilde{J}(X_h^F) - \tilde{J}(X_h^{F_0})\right| \leq L_{\Gamma,Z}\,\mathrm{diam}(\mathscr{F}) \tag{19}$$

where $\mathbb{E}[L^2_{\Gamma,Z}] < \infty$. From (19) we see that

$$\mathrm{Var}(\tilde{J}(X_h^F) - \tilde{J}(X_h^{F_0})) \leq \mathbb{E}\left[L^2_{\Gamma,Z}\right]\mathrm{diam}^2(\mathscr{F}). \tag{20}$$

Use (18) and (20) along with (16) to conclude that the assertion of the lemma holds. □

We now present the main rate result governing the solution estimator $F^*_{N,n,k}$ of (MC-n-OPT).

Theorem 3 Let Assumptions 2, 6, 7, 8 hold, and suppose that the method used to generate paths $X_h^F \overset{\text{iid}}{\sim} \pi^{\Gamma}_{F,h}$ exhibits weak convergence order β, implying that there exists $\ell_1 < \infty$ such that

$$\sup_{F\in\mathscr{F}} \left|\mathbb{E}\left[\tilde{J}(X_h^F)\right] - J(F)\right| \leq \ell_1 h^{\beta}. \tag{21}$$

Furthermore, suppose mirror descent (Bubeck 2015, pp. 80) is executed for k steps on (MC-n-OPT):

$$\begin{aligned} F_{N,n,j+1} &= \sup_{x\in\mathscr{F}_n\cap\mathscr{D}} D_{\psi}(x, G_{N,n,j+1}); \\ \nabla\psi(G_{N,n,j+1}) &= \nabla\psi(F_{N,n,j}) - \eta S_{J_{N,h}}(F_{N,n,j}); \quad j = 0,1,\ldots,k-1 \\ F^*_{N,n,k} &:= \frac{1}{k}\sum_{j=1}^{k} F_{N,n,j}, \end{aligned} \tag{22}$$

where $\psi : \mathscr{D} \subset \mathscr{F} \to \mathbb{R}$ is a chosen ρ-strongly convex, mirror-map (see Definition 7) with $\mathscr{F}_n \cap \mathscr{D} \neq \emptyset$, the Bregman divergence

$$D_{\psi}(x,y) := \psi(x) - (\psi(y) + \langle\nabla\psi(y), x-y\rangle), \quad \forall x,y \in \mathscr{D},$$

and the step size

$$\eta = \eta_0 \frac{R}{\bar{K}}\sqrt{\frac{2\rho}{k}}, \quad \eta_0 \in (0,1)$$

where $R^2 := \sup_{x\in\mathscr{F}_n\cap\mathscr{D}} \psi(x) - \psi(F_{N,n,0})$, and $\bar{K} := N^{-1}\sum_{j=1}^{N} K_{Z_j}$ is the iid sample mean of Lipschitz constants $K_{Z_j}, j = 1,2,\ldots,N$ appearing in Assumption 2 satisfying

$$\sup_{F\in\mathscr{F}} \|S_{J_{N,h}}(F)\|_* \leq \bar{K}; \quad S_{J_{N,h}}(F) \in \partial J_{N,h}(F); \quad \mathbb{E}[K^2_{Z_j}] < \infty, \tag{23}$$

where $S_{J_{N,h}(F)}$ is a subgradient and $\partial J_{N,h}(F)$ the subdifferential of the convex functional $J_{N,h}$ at the point F. Then, for all $k \geq 1$,

$$0 \leq \mathbb{E}\left[J(F^*_{N,n,k}) - J(F^*)\right] \leq \frac{c_1}{\sqrt{k}} + \frac{c_2}{\sqrt{N}} + c_3 h^\beta + c_4 g(n), \tag{24}$$

where

$$\begin{aligned}
c_1 &:= \sqrt{\frac{2}{\rho}} \left(\mathbb{E}[R^2]\left(\frac{1}{k}\mathrm{Var}(K_Z) + \mathbb{E}[K_Z^2]\right)\right)^{1/2};\\
c_2 &:= \frac{3}{\sqrt{N}}\left(\mathrm{diam}(\mathscr{F})\sqrt{\mathbb{E}[L^2_{\Gamma,Z}]} + \sigma_0(h)\right);\\
c_3 &:= \ell_1; \quad \textit{and}\\
c_4 &:= \mathbb{E}\left[K_Z\right].
\end{aligned} \tag{25}$$

Proof. Observe that

$$\begin{aligned}
0 \leq J(F^*_{N,n,k}) - J(F^*) &= J(F^*_{N,n,k}) - J_{N,h}(F^*_{N,n,k}) + J_{N,h}(F^*_{N,n,k}) - J_{N,h}(F^*_{N,n})\\
&\qquad + J_{N,h}(F^*_{N,n}) - J(F^*_n) + J(F^*_n) - J(F^*)\\
&\leq J_{N,h}(F^*_{N,n,k}) - J_{N,h}(F^*_{N,n}) + \sum_{F\in\{F^*_{N,n,k},F^*_{N,n},F^*_n\}} |J_{N,h}(F) - J(F)| + J(F^*_n) - J(F^*)\\
&\leq \overbrace{J_{N,h}(F^*_{N,n,k}) - J_{N,h}(F^*_{N,n})}^{\text{opt. error}} + \sum_{F\in\{F^*_{N,n,k},F^*_{N,n},F^*_n\}} \overbrace{|J_{N,h}(F) - \mathbb{E}\left[J_{N,h}(F)\right]|}^{\text{sampling error}}\\
&\qquad + \sum_{F\in\{F^*_{N,n,k},F^*_{N,n},F^*_n\}} \underbrace{|\mathbb{E}\left[J_{N,h}(F)\right] - J(F)|}_{\text{approx. error}} - \underbrace{S_J(F^*_n)\,\|F^*_n - F^*\|}_{\text{proj. error}},
\end{aligned} \tag{26}$$

where the penultimate inequality in (26) follows from rearrangement of terms and the last inequality follows upon using the sub-gradient inequality (3) for the convex functional $J(\cdot)$. Now we quantify (in expectation) each of the error terms appearing on the right-hand side of (26). Applying mirror descent's complexity bound (Bubeck 2015, pp. 80) on the $\bar{K}$-smooth function $J_{N,h}(\cdot)$ and taking expectation, we get

$$0 \leq \mathbb{E}\left[J_{N,h}(F^*_{N,n,k}) - J_{N,h}(F^*_{N,n})\right] \leq \frac{1}{\sqrt{k}}\sqrt{\frac{2}{\rho}}\left(\mathbb{E}[R^2]\left(\frac{1}{k}\mathrm{Var}(K_Z) + \mathbb{E}[K_Z^2]\right)\right)^{1/2}. \tag{27}$$

Next, using Lemma 3 we get the bound on approximation error in (26):

$$\mathbb{E}\left[\sum_{F\in\{F^*_{N,n,k},F^*_{N,n},F^*_n\}} |J_{N,h}(F) - \mathbb{E}\left[J_{N,h}(F)\right]|\right] \leq \frac{3}{\sqrt{N}}\left(\mathrm{diam}(\mathscr{F})\sqrt{\mathbb{E}[L^2_{\Gamma,Z}]} + \sigma_0(h)\right). \tag{28}$$

Due to the assumption in (21), we have

$$\sup_{F\in\mathscr{F}} |\mathbb{E}\left[J_{N,h}(F)\right] - J(F)| \leq \ell_1 h^\beta. \tag{29}$$

And since J is convex, we see that

$$J(F^*_n) - J(F^*) \leq \|S_J(F^*_n)\|_* \|F^*_n - F^*\| \leq \sup_{F\in\mathscr{F}} \|S_J(F)\|_* \|F^*_n - F^*\| \leq \mathbb{E}\left[K_Z\right] g(n), \tag{30}$$

where the last inequality in (30) is from Assumption 2. Now use (27), (28), (29), and (30) to conclude. $\square$

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York, NY: Springer.

Banholzer, D., J. Fliege, and R. Werner. 2019. "On rates of convergence for sample average approximations in the almost sure sense and in mean". *Mathematical Programming* 191(1):307–345.

Bartlett, P. L., and S. Mendelson. 2002. "Rademacher and Gaussian complexities: Risk bounds and structural results". *Journal of Machine Learning Research* 3(Nov):463–482.

Bose, A. 1998. "Bahadur representation of M_m estimates". *The Annals of Statistics* 26(2):771–777.

Boucheron, S., G. Lugosi, and P. Massart. 2013. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bubeck, S. 2015. "Convex Optimization: Algorithms and Complexity". *Foundations and Trends in Machine Learning* 8(3–4):231–358.

Chen, H., and D. D. Yao. 2001. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, Volume 4. Springer.

Dupačová, J., and R. J. B. Wets. 1988. "Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems". *The Annals of Statistics* 16(4):1517–1549.

Kim, S., R. Pasupathy, and S. G. Henderson. 2014. "A Guide to SAA". In *Encyclopedia of Operations Research and Management Science*, edited by M. Fu, Hillier and Lieberman OR Series. Elsevier.

Kontorovich, A. 2014. "Concentration in unbounded metric spaces and algorithmic stability". In *Proceedings of the 31st International Conference on Machine Learning*. June 22$^{\text{nd}}$-24$^{\text{th}}$, Beijing, China, 28–36.

Kreyszig, E. 1989. *Introductory functional analysis with applications*. Wiley Classics Library ed. Wiley classics library. New York: Wiley.

Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization: A Basic Course*. New York, NY: Springer Science + Business Media, LLC.

Pasupathy, R. 2010. "On Choosing Parameters in Retrospective-Approximation Algorithms for Stochastic Root Finding and Simulation Optimization". *Operations Research* 58(4):889–901.

Robinson, S. 1996. "Analysis of Sample-path Optimization". *Mathematics of Operations Research* 21(3):513–528.

Shapiro, A. 2003. "Monte Carlo sampling methods". *Handbooks in operations research and management science* 10:353–425.

Shapiro, A., D. Dentcheva, and A. Ruszczynski. 2009. *Lectures on Stochastic Programming: Modeling and Theory*. 2nd ed. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

van de Geer, S. 2000. *Empirical Processes in M-estimation*. Cambridge, UK: Cambridge University Press.

AUTHOR BIOGRAPHY

ZIHE ZHOU is a graduate student in the School of Industrial Engineering at Purdue University. Her research interests lie broadly in applied probability and simulation. Her email address is zhou408@purdue.edu.

HARSHA HONNAPPA is Associate Professor in the School of Industrial Engineering at Purdue University. His research interests lie broadly in applied probability, stochastic optimization, simulation methodology and machine learning. His email address is honnappa@purdue.edu and his web page is http://engineering.purdue.edu/SSL.

RAGHU PASUPATHY is Professor of Statistics at Purdue University. His current research interests lie broadly in general simulation methodology, stochastic optimization, statistical inference and statistical computation. Raghu Pasupathy's email address is pasupath@purdue.edu, and his web page https://web.ics.purdue.edu/~pasupath contains links to papers, software codes, and other material.