SIAM J. OPTIM.
Vol. 30, No. 3, pp. 2103–2133

SAMPLE COMPLEXITY OF SAMPLE AVERAGE APPROXIMATION FOR CONDITIONAL STOCHASTIC OPTIMIZATION[*]

YIFAN HU[†], XIN CHEN[†], AND NIAO HE[†]

Abstract. In this paper, we study a class of stochastic optimization problems, referred to as the *conditional stochastic optimization* (CSO), in the form of $\min_{x\in\mathcal{X}} \mathbb{E}_\xi f_\xi(\mathbb{E}_{\eta|\xi}[g_\eta(x,\xi)])$, which finds a wide spectrum of applications including portfolio selection, reinforcement learning, robust learning, and causal inference. Assuming availability of samples from the distribution $\mathbb{P}(\xi)$ and samples from the conditional distribution $\mathbb{P}(\eta|\xi)$, we establish the sample complexity of the sample average approximation (SAA) for CSO, under a variety of structural assumptions, such as Lipschitz continuity, smoothness, and error bound conditions. We show that the total sample complexity improves from $\mathcal{O}(d/\epsilon^4)$ to $\mathcal{O}(d/\epsilon^3)$ when assuming smoothness of the outer function, and further to $\mathcal{O}(1/\epsilon^2)$ when the empirical function satisfies the quadratic growth condition. We also establish the sample complexity of a modified SAA when ξ and η are independent. Several numerical experiments further support our theoretical findings.

Key words. stochastic optimization, sample average approximation, large deviations theory

AMS subject classifications. 90C15, 90C30, 90C59

DOI. 10.1137/19M1284865

1. Introduction.

Decision-making in the presence of uncertainty has been a fundamental and long-standing challenge in many fields of science and engineering. In recent years, extensive research efforts have been devoted to the design and theory of efficient algorithms for solving the classical stochastic optimization (SO) in the form of

$$\min_{x\in\mathcal{X}} F(x) := \mathbb{E}_\xi[f(x,\xi)], \tag{1.1}$$

ranging from convex to nonconvex objectives, from first-order to second-order methods, and from sublinear to linear convergent algorithms; see, e.g., [5] and references therein for a comprehensive survey. Here $\mathcal{X} \subseteq \mathbb{R}^d$ is the decision set and $f(x,\xi)$ is some cost function dependent on the random vector ξ. In general, (1.1) cannot be computed analytically or solved exactly, even when the underlying distribution of the random vector ξ is known, and one has to resort to Monte Carlo sampling techniques.

An important Monte Carlo method—the sample average approximation (SAA), also known as the empirical risk minimization in the machine learning community—is widely used to solve (1.1), assuming availability of samples from the underlying distribution. SAA works by solving the approximation of the original problem:

$$\min_{x\in\mathcal{X}} \hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i), \tag{1.2}$$

where $\xi_1,\ldots,\xi_n$ are independent and identically distributed (i.i.d.) samples generated from the distribution of ξ. Note that $\hat{F}_n(x)$ converges pointwise to $F(x)$ with proba-

[*] Received by the editors September 3, 2019; accepted for publication (in revised form) May 26, 2020; published electronically August 11, 2020.
https://doi.org/10.1137/19M1284865
Funding: This work was supported by NSF awards 1755829 and 1934986.

[†] Department of Industrial and Enterprise Systems Engineering (ISE), University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801 (yifanhu3@illinois.edu, xinchen@illinois.edu, niaohe@illinois.edu).

bility 1 as n goes to infinity. Finite-sample convergence of SAA for SO has been well established. The seminal work by [20] proved that for general Lipschitz continuous objectives, SAA requires a sample complexity of $\mathcal{O}(d/\epsilon^2)$ to obtain an ϵ-optimal solution to the SO problem. The authors of [36] proved that for strongly convex and Lipschitz continuous objectives, the sample complexity of SAA is $\mathcal{O}(1/\epsilon)$. Detailed results can be found in the books [38] and [35].

More generally, SAA is also a popular computational tool for solving multistage stochastic programming (MSP) problems. In its general form, an MSP finds a sequence of decisions $\{x_t\}_{t=0}^T$ that minimizes the nested expectation in the following form:

$$\min_{x_0\in\mathcal{X}_0} f_0(x_0)+\mathbb{E}_{\xi_1}\left[\inf_{x_1} f_1(x_1,\xi_1)+\mathbb{E}_{\xi_2|\xi_1}\left[\cdots+\mathbb{E}_{\xi_T|\xi_{1:T-1}}\left[\inf_{x_T} f_T(x_T,\xi_T)\right]\right]\right], \tag{1.3}$$

where T is the number of decision periods, $\xi_1,\ldots,\xi_T$ can be considered as a random process, and the decision x_t is a function of the history of the process up to time t. Similarly, the SAA approach works by first generating a large scenario tree with conditional sampling and then processing with stage-based or scenario-based decomposition methods [31, 33, 34]. When extended to the multistage case, the finite-sample analysis indicates that the total number of samples, or scenarios, to achieve an ϵ-optimal solution to the original problem (1.3) grows exponentially as the number of stages increases [39, 38]. In particular, for general three-stage stochastic problems, the sample complexity of SAA cannot be smaller than $\mathcal{O}(d^2/\epsilon^4)$; this holds true even if the cost functions in all stages are linear and the random vectors are stagewise independent as discussed in [37].

In this paper, we study an intermediate class of problems, referred to as the *conditional stochastic optimization* (CSO), that sits in between the classical SO and the MSP. The problem of interest takes the following general form:

$$\min_{x\in\mathcal{X}} F(x):=\mathbb{E}_\xi\left[f_\xi\left(\mathbb{E}_{\eta|\xi}[g_\eta(x,\xi)]\right)\right]. \tag{1.4}$$

Here $\mathcal{X}$ is the domain of the decision variable $x\in\mathbb{R}^d$; $f_\xi(\cdot):\mathbb{R}^k\to\mathbb{R}$ is a continuous cost function dependent on the random vector ξ; and $g_\eta(\cdot,\xi):\mathbb{R}^d\to\mathbb{R}^k$ is a vector-valued continuous cost function dependent on both random vectors ξ and η. The inner expectation is with respect to η given ξ, and the outer expectation is with respect to ξ. Similar to the classical SO, we do not assume any knowledge of the underlying distribution of $\mathbb{P}(\xi)$ or the conditional distribution $\mathbb{P}(\eta|\xi)$. Instead, we assume availability of samples from the distribution $\mathbb{P}(\xi)$ and samples from the conditional distribution $\mathbb{P}(\eta|\xi)$ for any given ξ.

CSO is more general than the classical SO as it captures dynamic randomness and involves conditional expectation. It takes the SO as a special case when $g_\eta(x,\xi)$ is an identical function. On the other hand, it is less complicated than the MSP (in particular the three-stage case with $T=3$) as it seeks a static decision and is not subject to nonanticipativity constraints.

The goal of this paper is to analyze the sample complexity of SAA for solving CSO, which can be constructed as follows based on *conditional sampling*:

$$\min_{x\in\mathcal{X}} \hat{F}_{nm}(x):=\frac{1}{n}\sum_{i=1}^n f_{\xi_i}\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_{ij}}(x,\xi_i)\right), \tag{1.5}$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. samples generated from $\mathbb{P}(\xi)$ and $\{\eta_{ij}\}_{j=1}^m$ are i.i.d. samples generated from the conditional distribution $\mathbb{P}(\eta|\xi_i)$ for a given outer sample ξ_i. We would like to examine the total number of samples $T = nm+n$ required for SAA (1.5) to achieve an ϵ-optimal solution to the original CSO problem (1.4).

We also consider a special case of the CSO problem (1.4) when the random vectors ξ and η are independent:

$$\min_{x\in\mathcal{X}} F(x) := \mathbb{E}_\xi\Big[f_\xi\Big(\mathbb{E}_\eta[g_\eta(x,\xi)]\Big)\Big]. \tag{1.6}$$

One could still approximate (1.6) by the SAA (1.5), mimicking the conditional sampling scheme and using different samples $\{\eta_{i1},\ldots,\eta_{im}\}$ from the distribution of η for each ξ_i. However, since the inner expectation is no longer a conditional expectation, there is no necessity to estimate the inner expectation with different realizations of η for each ξ_i. Hence, an alternative way to approximate (1.6) is through a modified SAA:

$$\min_{x\in\mathcal{X}} \hat{F}_{nm}(x) := \frac{1}{n}\sum_{i=1}^{n} f_{\xi_i}\Bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta_j}(x,\xi_i)\Bigg), \tag{1.7}$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. samples generated from the distribution of ξ and $\{\eta_j\}_{j=1}^m$ are i.i.d. samples generated from the distribution of η. As a result, the component functions $f_{\xi_i}(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(x,\xi_i))$, $i=1,\ldots,n$, become dependent since they share the same $\{\eta_j\}_{j=1}^m$, making it very different from (1.5). In this case, the total number of samples becomes $T = n+m$. We refer to this sampling scheme as *independent sampling.*

1.1. Motivating applications. Notably, CSO can be used to model a variety of applications, including portfolio selection [16], robust supervised learning [7], reinforcement learning [7, 8], personalized medical treatment [45], and instrumental variable regression [27]. We discuss some of these examples in detail below.

Robust supervised learning. Incorporation of priors on invariance and robustness into the supervised learning procedures is crucial for computer vision and speech recognition [28, 3]. Taking image classification as an example, we would like to build a classifier that is both accurate and invariant to certain kinds of data transformation, such as rotation and perturbation. Let $\xi_1 = (a_1, b_1),\ldots,\xi_n = (a_n, b_n)$ be a set of input data, where a_i is the feature vector and b_i is the label. A plausible way to achieve such consistency is to consider the class of robust linear classifiers, say $f(x, x_0, \xi) = \mathbb{E}_{\eta|\xi\sim\mu(\sigma(a))}[x^T\eta+x_0]$ for given image data ξ, by averaging the prediction over all possible transformations $\sigma(a)$, and then finding the best fit by minimizing the expected risk:

$$\min_{(x,x_0)} \mathbb{E}_{\xi=(a,b)}\Big[\ell\big(b, \mathbb{E}_{\eta|\xi}[\eta^T x + x_0]\big)\Big] + \frac{\nu}{2}\|x\|_2^2.$$

Here $\ell(\cdot,\cdot)$ is some loss function, $\nu > 0$ is a regularization parameter, and $\mu(\cdot)$ is a given distribution (e.g., uniform) over the transformations. Clearly, such problems belong to the category of CSO.

Reinforcement learning. Policy evaluation is a fundamental task in Markov decision processes and reinforcement learning. Consider a discounted Markov decision process characterized by the tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P(s, a, s')$ represents the (unknown) state transition probability from state s to s' given action a, $r(s, a) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is a reward function, and $\gamma \in (0, 1)$ is a discount factor. Given a stochastic policy $\pi(a|s)$, the goal of the policy evaluation is to estimate the value function $V^\pi(s) := \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k r(s_k,a_k)\middle|\, s_0 = s\right]$ under the policy. It is well known that $V^\pi(\cdot)$ is a fixed point of the Bellman equation [1]

$$V^\pi(s) = \mathbb{E}_{s'|a,s}[r(s,a) + \gamma V^\pi(s')].$$

To estimate the value function $V^\pi(s)$, one could resort to minimizing the mean squared Bellman error [41, 7], namely

$$\min_{V(\cdot):\mathcal{S}\to\mathbb{R}} \mathbb{E}_{s\sim\mu(\cdot),a\sim\pi(\cdot|s)}\left[\left(r(s,a) - \mathbb{E}_{s'|a,s}[V(s) - \gamma V(s')]\right)^2\right].$$

Here $\mu(\cdot)$ is the stationary distribution. This minimization problem can be viewed as a special case of CSO. Recently, [8] showed that finding the optimal policy can also be formulated into an optimization problem in a similar form by exploiting the smoothed Bellman optimality equation. Again, the resulting problem falls under the category of CSO.

Uplift modeling. Uplift modeling aims at estimating individual treatment effects, and it has been widely studied in causal inference literature and used for personalized medicine treatment and targeted marketing [18, 45]. In an individual uplift model, the goal is to estimate the effect of a treatment on an individual with feature vector x, which could be represented by $u(x) := \mathbb{E}[y|x,t=1] - \mathbb{E}[y|x,t=-1]$. Here $t \in \{\pm 1\}$ represents whether a treatment has been given to an individual, and $y \in \mathcal{Y} \subseteq \mathbb{R}$ represents the outcome. In practice, obtaining joint labels (y,t) can be difficult, whereas obtaining one label (either t or y) of the individual is relatively easier. The authors of [45] considered an individual uplift model that assumes availability of only one label from the joint labels and estimates the unknown label with $p(y|x) = \sum_{t=\{\pm1\}} p(y|x,t)p(t|x)$. They showed that the individual uplift $u(x)$ is equivalent to the optimal solution to the following least-squares problem:

$$\min_{u\in L^2(p)} \mathbb{E}_{x\sim p(x)}\left[\left(\mathbb{E}_{w|x}[w]\cdot u(x) - 2\mathbb{E}_{z|x}[z]\right)^2\right],$$

where $L^2(p) = \{f : \mathcal{X} \to \mathbb{R} |\, \mathbb{E}_{x\sim p(x)}[f(x)^2] < \infty\}$ is a function space, and w and z are two auxiliary random variables whose conditional densities are given by $p\,(z = z_0|x) = \frac{1}{2}p\,(y = z_0|x) + \frac{1}{2}p\,(y = -z_0|x)$, $p\,(w = w_0|x) = \frac{1}{2}p\,(t = w_0|x) + \frac{1}{2}p\,(t = -w_0|x)$. If we further restrict $u(\cdot)$ to a finite dimensional parameterization, then the above problem becomes a special case of CSO.

For these applications, there are many settings in which samples can be generated according to our assumptions. For instance, in robust supervised learning and uplift modeling, there are multiple samples from $\mathbb{P}(\eta|\xi)$ available for any given ξ.

1.2. Related work. A closely related class of problems, called *stochastic composition optimization*, has been extensively studied in the literature; see, e.g., [46, 32, 12, 42], to name just a few. This class of problems takes the following form:

$$\min_{x\in\mathcal{X}} \mathbf{f}\circ\mathbf{g}(x) := \mathbb{E}_\xi\Big[f_\xi\Big(\mathbb{E}_\eta[g_\eta(x)]\Big)\Big], \tag{1.8}$$

where $\mathbf{f}(u) := \mathbb{E}_\xi[f_\xi(u)]$, and $\mathbf{g}(x) := \mathbb{E}_\eta[g_\eta(x)]$. Although the two problems, (1.8) and (1.4), share some similarities in that both objectives are represented by nested expectations, they are fundamentally different in two aspects: (i) the inner randomness η in (1.4) is conditionally dependent on the outer randomness ξ, while the inner expectation in (1.8) is taken over the marginal distribution of η; (ii) the inner random function $g_\eta(x,\xi)$ in (1.4) depends on both ξ and η. As a result, unlike (1.8), the CSO problem (1.4) cannot be formulated as a composition of two deterministic functions due to the dependence between the inner and outer functions. Another key distinction from (1.8) is that we assume availability of samples from the distributions $\mathbb{P}(\xi)$ and $\mathbb{P}(\eta|\xi)$, rather than samples from the joint distribution $\mathbb{P}(\xi,\eta)$. These two distinctions further lead to a drastic difference in the SAA construction and the sample complexity analysis of these two types of problems, as we will show in the rest of the paper.

When solving either (1.8) or (1.4), most of the existing work is devoted to developing stochastic oracle-based algorithms and their convergence analysis for solving these problems. Related work includes two-timescale [32, 46, 42, 43] and single-timescale [13] stochastic approximation algorithms for solving the problem (1.8), variance-reduced algorithms for solving the SAA counterpart of (1.8) [22, 17, 40], and a primal-dual functional stochastic approximation algorithm for solving the problem (1.4) [7]. These methods usually require convexity of the objective in order to obtain an ϵ-optimal solution. Our work differs from the works listed above in that we mainly focus on establishing the sample complexity of SAA itself rather than designing efficient algorithms to solve the resulting SAA.

We point out that our paper is in the same vein as a series of papers [20, 39, 37, 30, 12, 9, 2, 23], centered at the sample average approximation approach for stochastic programs. In particular, [9] derived a central limit theorem result for the SAA of the stochastic composition optimization problem (1.8), and [12, 29] established the rate of convergence. Despite these developments, the study of the basic SAA approach and its finite-sample complexity analysis remains unexplored for solving the general CSO problem (1.4) and even the special case (1.6). We aim to close this gap in this paper.

1.3. Contributions.

In this paper, we formally analyze the sample complexity of the corresponding SAA approach for solving CSO. Our contributions are summarized as follows and in Table 1.1.

(a) We establish the first sample complexity results of the SAA in (1.5) for the CSO problem (1.4) under several structural assumptions:

  (i) Both f_ξ and g_η are Lipschitz continuous.
  (ii) In addition to (i), f_ξ is Lipschitz smooth.
  (iii) In addition to (i), the empirical function satisfies the Hölderian error bound condition.
  (iv) In addition to (i), f_ξ is Lipschitz smooth, and the empirical function satisfies the Hölderian error bound condition.

None of these assumptions require convexity[1] of the underlying objective function. Note that the Hölderian error bound condition [4], which includes the quadratic growth (QG) condition [19] as a special case, is a much weaker assumption than strong convexity and holds for many nonconvex problems in machine learning applications [6]. We show that, for general Lipschitz continuous problems, the sample complexity of SAA improves from $\mathcal{O}(d/\epsilon^4)$ to $\mathcal{O}(d/\epsilon^3)$ when assuming smoothness; for problems satisfying the QG condition, the sample complexity of SAA improves from $\mathcal{O}(1/\epsilon^3)$ to $\mathcal{O}(1/\epsilon^2)$ when assuming smoothness. This is very different from the classical results on the SO and the MSP, where Lipschitz smoothness plays no essential role in the sample complexity [20, 37]. Our results are built on the

[1] However, when solving the SAA problem itself, convexity conditions are often necessary for obtaining a global minimizer.

Table 1.1
Sample complexity of SAA methods.

Problem	Assumptions		Sample complexity	
	$f_\xi(\cdot)$	$\hat{F}_n$ *or* $\hat{F}_{nm}$	*Conditional*	*Independent*
SO [20]	-	-	$\mathcal{O}(d/\epsilon^2)$	-
SO [36]	-	*strongly convex*	$\mathcal{O}(1/\epsilon)$	-
MSP ($T=3$) [37]	-	-	$\mathcal{O}(d^2/\epsilon^4)$	$\mathcal{O}(d^2/\epsilon^4)$
CSO	-	-	$\mathcal{O}(d/\epsilon^4)$	$\mathcal{O}(d/\epsilon^2)$
CSO	*smooth*	-	$\mathcal{O}(d/\epsilon^3)$	
CSO	-	*quadratic growth*	$\mathcal{O}(1/\epsilon^3)$	$\mathcal{O}(d/\epsilon^2)$
CSO	*smooth*	*quadratic growth*	$\mathcal{O}(1/\epsilon^2)$	

$\hat{F}_n$ or $\hat{F}_{nm}$ = empirical objective; ϵ = accuracy; d = dimension;
Conditional = conditional sampling; Independent = independent sampling

traditional large deviation theory and stability arguments, while leveraging several bias-variance decomposition techniques, in order to fully exploit the specific structure of CSO and other structural assumptions.

(b) We analyze the sample complexity of the modified SAA in (1.7) for the special case (1.6), where ξ and η are independent. We show that the total sample complexity of the modified SAA is $\mathcal{O}(d/\epsilon^2)$ for the general Lipschitz continuous problems. The existence of the QG condition only improves the complexity of the outer samples from $\mathcal{O}(d/\epsilon^2)$ to $\mathcal{O}(1/\epsilon)$, yet the overall complexity is dominated by the complexity of the inner samples, which is $\mathcal{O}(d/\epsilon^2)$. Our complexity result matches with the asymptotic rate established in [9], even without assuming smoothness of outer and inner functions, and is unimprovable.

(c) We conduct some simulations of the SAA approach on several examples, including the logistic regression, least absolute value (LAV) regression, and its smoothed counterpart, under some modifications. Our simulation results indicate that solving the nonsmooth LAV regression requires more samples than solving its smooth counterpart to achieve the same accuracy. We also observe that when the variance of the inner randomness is relatively large, for a fixed budget T, setting $n = O(\sqrt{T})$ samples seems to perform best for logistic regression, which matches with our theory. Although both conditional sampling and independent sampling schemes can be applied to solving the special case (1.6), with nearly matching sample complexity in situation (iv) (see the last row in Table 1.1), our simulations show that using the independent sampling scheme exhibits better performance in practice.

1.4. Paper organization.

The remainder of this paper is organized as follows. In section 2, we introduce some notation and preliminaries. In section 3, we give the basic assumptions and analyze the mean squared error (MSE) of the Monte Carlo estimation. In section 4, we present the main results on the sample complexity of SAA for CSO under different structural assumptions. In section 5, we provide results for the special case when ξ and η are independent. Numerical results are given in section 6.

2. Preliminaries.

For convenience, we collect here some notation that will be used throughout the paper. We also introduce some mathematical tools and propositions that are necessary for future discussion. For simplicity, we restrict our attention to the l_2-norm, denoted as $||\cdot||_2$. Similar results on sample complexity with respect to different norms can be obtained with minor modification of the analysis.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the decision set. We say $\mathcal{X}$ has a finite diameter $D_\mathcal{X}$ if $||x_1 - x_2||_2 \leq D_\mathcal{X}$ $\forall x_1, x_2 \in \mathcal{X}$. For $\upsilon \in (0,1)$, $\{x_l\}_{l=1}^Q$ is said to be a υ-net of $\mathcal{X}$ if $x_l \in \mathcal{X}$ $\forall l = 1, \dots, Q$, and the following holds: $\forall x \in \mathcal{X}, \exists l(x) \in \{1, \dots, Q\}$ such that $||x - x_{l(x)}||_2 \leq \upsilon$. If $\mathcal{X}$ has a finite diameter $D_\mathcal{X}$, for any $\upsilon \in (0,1)$, there exists a υ-net of $\mathcal{X}$, and the size of the υ-net is bounded, $Q \leq \mathcal{O}((D_\mathcal{X}/\upsilon)^d)$ [38].

A function $f : \mathcal{X} \to \mathbb{R}$ is said to be L-Lipschitz continuous if there exists a constant $L > 0$ such that $|f(x_1) - f(x_2)| \leq L||x_1 - x_2||_2 \, \forall x_1, x_2 \in \mathcal{X}$. The function $f : \mathcal{X} \to \mathbb{R}$ is said to be S-Lipschitz smooth if it is continuously differentiable and its gradient is S-Lipschitz continuous. This also implies that $\forall x_1, x_2 \in \mathcal{X} : |f(x_1) - f(x_2) - \nabla f(x_2)^\top (x_1 - x_2)| \leq \frac{S}{2}||x_1 - x_2||_2^2$. If a continuously differentiable function $f : \mathcal{X} \to \mathbb{R}$ satisfies that $\forall x_1, x_2 \in \mathcal{X}$, $f(x_1) - f(x_2) - \nabla f(x_2)^\top (x_1 - x_2) \geq \frac{\mu}{2}||x_1 - x_2||_2^2$, then f is called μ-strongly convex when $\mu > 0$, convex when $\mu = 0$, and μ-weakly convex when $\mu < 0$.

Definition 2.1 (Hölderian error bound condition). *Let $f : \mathcal{X} \to \mathbb{R}$ be a function with compact domain $\mathcal{X}$ and the optimal solution set $\mathcal{X}^*$ is nonempty. $f(\cdot)$ satisfies the (μ, δ)-Hölderian error bound condition if there exist $\delta \geq 0$ and $\mu > 0$ such that*

$$\forall x \in \mathcal{X}, \quad f(x) - \min_{x \in \mathcal{X}} f(x) \geq \mu \inf_{z \in \mathcal{X}^*} ||x - z||_2^{1+\delta}.$$

In particular, when $\delta = 1$, we say f satisfies the quadratic growth (QG) condition.

The Hölderian error bound condition is also known as the Łojasiewicz inequality [4]. When $\delta = 1$, the condition implies a quadratic growth of the function value near any local minima. The QG condition is a weaker assumption than strong convexity and does not need to be convex. When $f(\cdot)$ is convex, the QG condition is also referred to as optimal strong convexity [24] and semistrong convexity [14].

Cramér's large deviation theorem will be frequently used, so we list it as a lemma below based on the result in [20]. We extend the result to random vectors and provide the proof in Appendix A.

Lemma 2.1. *Let $X_1, \dots, X_n$ be i.i.d. samples of zero-mean random variable X with finite variance σ^2. For any $\epsilon > 0$, it holds that*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right) \leq \exp(-nI(\epsilon)),$$

where $I(\epsilon) := \sup_{t \in \mathbb{R}}\{t\epsilon - \log M(t)\}$ is the rate function of random variable X, and $M(t) := \mathbb{E}e^{tX}$ is the moment generating function of X. For any $\delta > 0$, there exists $\epsilon_1 > 0$, for any $\epsilon \in (0, \epsilon_1)$, $I(\epsilon) \geq \frac{\epsilon^2}{(2+\delta)\sigma^2}$. If X is a zero-mean sub-Gaussian, then $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon) \leq \exp(-\frac{n\epsilon^2}{2\sigma^2})$ $\forall \epsilon > 0$.

If X is a zero-mean random vector in $\mathbb{R}^k$ such that $\mathbb{E}\|X\|_2^2 = \sigma^2 < \infty$, then for any $\delta > 0$, there exists $\epsilon_1 > 0$, for any $\epsilon \in (0, \epsilon_1)$,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_2 \geq \epsilon\right) \leq 2k \exp\left(-\frac{n\epsilon^2}{(2+\delta)\sigma^2}\right).$$

We will also use the simple fact that for any random variables Y and Z, if random variable $W \leq X := Y + Z$, then for any $\epsilon > 0$, $\mathbb{P}(W > \epsilon) \leq \mathbb{P}(X > \epsilon) \leq \mathbb{P}(Y >$

$\frac{\epsilon}{2}) + \mathbb{P}(Z > \frac{\epsilon}{2})$. Lastly, throughout the paper, we call $x_\epsilon \in \mathcal{X}$ an ϵ-optimal solution to the problem $\min_{x\in\mathcal{X}} F(x)$ if $F(x_\epsilon) - \min_{x\in\mathcal{X}} F(x) \le \epsilon$.

3. Mean squared error of SAA estimator for CSO.

In this section, we make the basic assumptions and analyze the MSE of the Monte Carlo estimate of the function value $f(x)$ at a given point.

3.1. Problem formulation and assumptions.

Recall the problem (1.4),

$$\min_{x\in\mathcal{X}} F(x) := \mathbb{E}_\xi\Big[f_\xi\Big(\mathbb{E}_{\eta|\xi}[g_\eta(x,\xi)]\Big)\Big],$$

where $f_\xi(\cdot) : \mathbb{R}^k \to \mathbb{R}$, $g_\eta(\cdot,\xi) : \mathbb{R}^d \to \mathbb{R}^k$ are random functions. Recall its SAA counterpart (1.5):

$$\min_{x\in\mathcal{X}} \hat{F}_{nm}(x) := \frac{1}{n}\sum_{i=1}^n f_{\xi_i}\Bigg(\frac{1}{m}\sum_{j=1}^m g_{\eta_{ij}}(x,\xi_i)\Bigg).$$

We denote x^* and $\hat{x}_{nm}$ the optimal solutions to the CSO and the SAA problems, respectively. We are interested in estimating the probability of $\hat{x}_{nm}$ being an ϵ-optimal solution to the CSO problem, namely $\mathbb{P}\left(F(\hat{x}_{nm}) - F(x^*) \le \epsilon\right)$, for an arbitrary accuracy $\epsilon > 0$.

Throughout the paper, we assume availability of i.i.d. samples generated from distribution $\mathbb{P}(\xi)$ and conditional distribution $\mathbb{P}(\eta\,|\,\xi)$ for any given ξ, and we make the following basic assumptions.

Assumption 3.1. *We assume the following:*

(a) *The decision set $\mathcal{X} \subseteq \mathbb{R}^d$ has a finite diameter $D_\mathcal{X} > 0$.*

(b) *$f_\xi(\cdot)$ is L_f-Lipschitz continuous and $g_\eta(\cdot,\xi)$ is L_g-Lipschitz continuous for any given ξ and η.*

(c) *For all $x \in \mathcal{X}$, $f(x,\xi)$ is Borel measurable in ξ and $g_\eta(x,\xi)$ is Borel measurable in η for all ξ.*

(d) *$\sigma_f^2 := \max_{x\in\mathcal{X}} \mathbb{V}_\xi\left(f_\xi(\mathbb{E}_{\eta|\xi}[g_\eta(x,\xi)])\right) < \infty$.*

(e) *$\sigma_g^2 := \max_{x\in\mathcal{X},\xi} \mathbb{E}_{\eta|\xi}||g_\eta(x,\xi) - \mathbb{E}_{\eta|\xi} g_\eta(x,\xi)||_2^2 < \infty$.*

(f) *$|f_\xi(\cdot)| \le M_f$, $\|g_\eta(\cdot,\xi)\|_2 \le M_g$ for any ξ and η.*

The assumption (f) on the boundedness of function values is implied by assumptions (a) and (b). The assumptions (d) and (e) on boundedness of variances are commonly used for sample complexity analysis in the literature. The assumptions (b) and (c) together suggest that the functions f_ξ and $g_\eta(x,\xi)$ are Carathéodory functions [21]. Although the parameters L_f, L_g, σ_f, and σ_g could depend on dimensions d and k, we treat these parameters as given constants throughout the paper.

3.2. Mean squared error of SAA objective.

In this subsection, we analyze the MSE of the estimator $\hat{F}_{nm}(x)$, i.e., the SAA objective (or the empirical objective), for estimating the true objective function $F(x)$, at a given x. The MSE can be decomposed into the sum of squared bias and variance of the estimator:

$$\text{MSE}(\hat{F}_{nm}(x)) := \mathbb{E}|\hat{F}_{nm}(x) - F(x)|^2 = (\mathbb{E}\hat{F}_{nm}(x) - F(x))^2 + \mathbb{V}(\hat{F}_{nm}(x)). \tag{3.1}$$

We have the following lemmas on bounding the bias and variance.

Lemma 3.1. *Let $\{\eta_j\}_{j=1}^m$ be conditional samples from $P(\eta|\xi)$ given $\xi \sim P(\xi)$. Under Assumption* 3.1, *for any fixed $x \in \mathcal{X}$ that is independent of ξ and $\{\eta_j\}_{j=1}^m$, it* *holds that*

$$\left|\mathbb{E}_{\{\xi,\{\eta_j\}_{j=1}^m\}}\left[f_\xi\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(x,\xi)\right)-f_\xi\big(\mathbb{E}_{\eta|\xi}g_\eta(x,\xi)\big)\right]\right|\le\frac{L_f\sigma_g}{\sqrt{m}}. \tag{3.2}$$

If, additionally, $f_\xi(\cdot)$ *is S-Lipschitz smooth, we have*

$$\left|\mathbb{E}_{\{\xi,\{\eta_j\}_{j=1}^m\}}\left[f_\xi\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(x,\xi)\right)-f_\xi\big(\mathbb{E}_{\eta|\xi}g_\eta(x,\xi)\big)\right]\right|\le\frac{S\sigma_g^2}{2m}. \tag{3.3}$$

Proof. Define $X_j := g_{\eta_j}(x,\xi)-\mathbb{E}_{\eta|\xi}g_\eta(x,\xi)$ and $\bar{X} := \sum_{j=1}^m X_j/m$. It follows that $\mathbb{E}_{\{\eta_j\}_{j=1}^m|\xi}[\bar{X}] = 0$ by definition, and that $\mathbb{E}_{\{\eta_j\}_{j=1}^m|\xi}[\|\bar{X}\|_2^2] \le \sigma_g^2/m$ by Assumption 3.1(d). $\mathbb{E}_{\{\eta_j\}_{j=1}^m|\xi}\nabla f_\xi\big(\mathbb{E}_{\eta|\xi}g_\eta(x,\xi)\big)^\top\big(\frac{1}{m}\sum_{j=1}^m X_j(x)\big) = 0$ since x is independent of $\{\eta_j\}_{j=1}^m$. The results then follow directly by invoking the Lipschitz continuity or the Lipschitz smoothness and taking expectations. □

LEMMA 3.2. *Under Assumption* 3.1, *it holds that* $\mathbb{V}(\hat{F}_{nm}(x)) \le \frac{\sigma_f^2}{n}+\frac{4M_fL_f\sigma_g}{n\sqrt{m}}$.

Proof. We first introduce $\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^n f_{\xi_i}\big(\mathbb{E}_{\eta|\xi_i}[g_\eta(x,\xi_i)]\big)$. It follows from the independence among $\{\xi_i\}_{i=1}^n$ that $\mathbb{V}(\hat{F}_n(x)) \le \frac{\sigma_f^2}{n}$. By definition we have

$$\begin{aligned}
&\mathbb{V}\Big(\hat{F}_{nm}(x)\Big)-\mathbb{V}\Big(\hat{F}_n(x)\Big)\\
=&\frac{1}{n}\left[\mathbb{E}(\hat{F}_{1m}(x)^2)-(\mathbb{E}\hat{F}_{1m}(x))^2\right]-\frac{1}{n}\left[\mathbb{E}(\hat{F}_1(x)^2)-(\mathbb{E}\hat{F}_1(x))^2\right]\\
=&\frac{1}{n}\left[\mathbb{E}(\hat{F}_{1m}(x)^2)-\mathbb{E}(\hat{F}_1(x)^2)\right]+\frac{1}{n}\left[(\mathbb{E}\hat{F}_1(x))^2-(\mathbb{E}\hat{F}_{1m}(x))^2\right],
\end{aligned}$$

where $\hat{F}_{1m}(x) := f_{\xi_1}\big(\frac{1}{m}\sum_{j=1}^m g_{\eta_{1j}}(x,\xi_1)\big)$ and $\hat{F}_1(x) := f_{\xi_1}\big(\mathbb{E}_{\eta|\xi_1}g_\eta(x,\xi_1)\big)$. From Assumption 3.1(b) and Lemma 3.1, we have $\mathbb{E}(\hat{F}_{1m}(x)^2)-\mathbb{E}(\hat{F}_1(x)^2) \le 2M_f\mathbb{E}|\hat{F}_{1m}(x)-\hat{F}_1(x)| \le 2M_fL_f\sigma_g/\sqrt{m}$. In addition, $(\mathbb{E}\hat{F}_1(x))^2-(\mathbb{E}\hat{F}_{1m}(x))^2 \le 2M_fL_f\sigma_g/\sqrt{m}$. Hence, we obtain the desired result. □

The following result on the MSE follows naturally by (3.1).

THEOREM 3.1. *Under Assumption* 3.1, *we have*

$$MSE(\hat{F}_{nm}(x)) \le \frac{L_f^2\sigma_g^2}{m}+\frac{1}{n}\left(\sigma_f^2+\frac{4M_fL_f\sigma_g}{\sqrt{m}}\right). \tag{3.4}$$

If, additionally, $f_\xi(\cdot)$ *is S-Lipschitz smooth, the MSE is further bounded by*

$$MSE(\hat{F}_{nm}(x)) \le \frac{S^2\sigma_g^4}{4m^2}+\frac{1}{n}\left(\sigma_f^2+\frac{4M_fL_f\sigma_g}{\sqrt{m}}\right). \tag{3.5}$$

Unlike the classical stochastic optimization, the SAA objective of CSO is no longer unbiased. The estimation error of the SAA objective therefore comes from both bias and variance. A key observation from Theorem 3.1 is that Lipschitz smoothness of $f_\xi(\cdot)$ is essential to reduce the bias and can be potentially exploited to improve the sample complexity of SAA.

We point out that in [15], the authors also consider the estimation problem of the expected value of a nonlinear function on a conditional expectation, i.e., $\mathbb{E}[f(\mathbb{E}[\zeta|\xi])]$.

Their setting is slightly different from ours as they restrict f to be one-dimensional and assume f contains a finite number of discontinuous or nondifferential points and is thrice differentiable with finite derivatives on all continuous points. They provide an asymptotic bound $\mathcal{O}(1/m^2 + 1/n)$ of the MSE for their nested estimator based on Taylor expansion. Here we focus on a general continuous outer function $f_\xi(\cdot)$ and show that Lipschitz smoothness of $f_\xi(\cdot)$ is sufficient to achieve a similar error bound with finite samples.

Remark 3.1. When the outer function $f_\xi(\cdot)$ is linear, $\hat{F}_{nm}(x)$ is an unbiased estimator of $F(x)$. By setting $S = 0$ in (3.5) and applying a similar analysis, we have

$$MSE(\hat{F}_{nm}(x)) = \mathbb{V}(\hat{F}_{nm}(x)) \leq \frac{1}{n}\bigg(\frac{L_f^2\sigma_g^2}{m} + \sigma_f^2\bigg).$$

Note that in this special case, the error of $\hat{F}_{nm}(x)$ is dominated by the number of outer samples used, which is quite different from the general case.

4. Sample complexity of SAA for conditional stochastic optimization.

In this section, we analyze the number of samples required for the solution to the SAA (1.5) to be ϵ-optimal of the CSO problem (1.4), with high probability.

We consider two general cases: (i) when the objective is Lipschitz continuous and (ii) when the empirical objective satisfies the Hölderian error bound condition. In the former case, we establish a uniform convergence analysis based on concentration inequalities to bound $\mathbb{P}(F(\hat{x}_{nm}) - F(x^*) \geq \epsilon)$, and in the latter case, we provide a stability analysis. In both cases, we further take into account two scenarios, with and without the Lipschitz smoothness assumption of the outer function $f_\xi(\cdot)$.

4.1. Sample complexity for general Lipschitz continuous functions.

We first consider the case when the objective is Lipschitz continuous and prove the uniform convergence.

THEOREM 4.1 (uniform convergence). *Under Assumption* 3.1*, for any* $\delta > 0$*, there exists* $\epsilon_1 > 0$ *such that for* $\epsilon \in (0, \epsilon_1)$*, when* $m \geq L_f^2\sigma_g^2/\epsilon^2$*, we have*

(4.1)
$$\mathbb{P}\bigg(\sup_{x\in\mathcal{X}} |\hat{F}_{nm}(x) - F(x)| > \epsilon\bigg) \leq \mathcal{O}(1)\bigg(\frac{4L_fL_gD_\mathcal{X}}{\epsilon}\bigg)^d \exp\bigg(-\frac{n\epsilon^2}{16(2+\delta)(\sigma_f^2 + 4M_fL_f\sigma_g)}\bigg).$$

If, additionally, $f_\xi(\cdot)$ *is S-Lipschitz smooth, then* (4.1) *holds as long as* $m \geq 2S\sigma_g^2/\epsilon$*.*

Proof. We construct a υ-net to get rid of the supreme over x and use a concentration inequality to bound the probability. First, we pick a υ-net $\{x_l\}_{l=1}^Q$ on the decision set $\mathcal{X}$, such that $L_fL_g\upsilon = \epsilon/4$; thus $Q \leq \mathcal{O}(1)(\frac{4L_gL_fD_\mathcal{X}}{\epsilon})^d$. Note that $\{x_l\}_{l=1}^Q$ has no randomness. By definition of the υ-net, we have $\forall x \in \mathcal{X}$, $\exists\ l(x) \in \{1, 2, \ldots, Q\}$, s.t. $||x - x_{l(x)}||_2 \leq \upsilon = \epsilon/4L_fL_g$. Invoking Lipschitz continuity of f_ξ and g_η, we obtain

$$|\hat{F}_{nm}(x) - \hat{F}_{nm}(x_{l(x)})| \leq \frac{\epsilon}{4}, \quad |F(x) - F(x_{l(x)})| \leq \frac{\epsilon}{4}.$$

Hence, for any $x \in \mathcal{X}$,

$$\begin{aligned}
&|\hat{F}_{nm}(x) - F(x)| \\
\leq\ &|\hat{F}_{nm}(x) - \hat{F}_{nm}(x_{l(x)})| + |\hat{F}_{nm}(x_{l(x)}) - F(x_{l(x)})| + |F(x_{l(x)}) - F(x)| \\
\leq\ &\frac{\epsilon}{2} + |\hat{F}_{nm}(x_{l(x)}) - F(x_{l(x)})| \leq \frac{\epsilon}{2} + \max_{l\in\{1,2,\cdots,Q\}} |\hat{F}_{nm}(x_l) - F(x_l)|.
\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\hat{F}_{nm}(x)-F(x)|>\epsilon\Big) \leq&\mathbb{P}\Big(\max_{l\in\{1,2,\cdots,Q\}}|\hat{F}_{nm}(x_l)-F(x_l)|>\frac{\epsilon}{2}\Big)\\ &\leq\sum_{l=1}^{Q}\mathbb{P}\Big(|\hat{F}_{nm}(x_l)-F(x_l)|>\frac{\epsilon}{2}\Big).\end{aligned}\tag{4.2}$$

Define $Z_i(l):=f_{\xi_i}(\frac{1}{m}\sum_{j=1}^{m}g_{\eta_{ij}}(x_l,\xi_i))-F(x_l)$; then $Z_1(l),Z_2(l),\ldots,Z_n(l)$ are i.i.d. random variables. Denote their expectation as $\mathbb{E}Z(l)$. Then $Z_i(l)-\mathbb{E}Z(l)$ is a zero-mean random variable.

If $\max_l\mathbb{E}Z(l)\leq\epsilon/4$, by Lemma 2.1, we have

$$\begin{aligned}&\mathbb{P}\Big(\hat{F}_{nm}(x_l)-F(x_l)>\frac{\epsilon}{2}\Big)\leq\mathbb{P}\Big(\hat{F}_{nm}(x_l)-F(x_l)>\frac{\epsilon}{4}+\mathbb{E}Z(l)\Big)\\ =&\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}[Z_i(l)-\mathbb{E}Z(l)]>\frac{\epsilon}{4}\Big)\leq\exp\Big(-\frac{n\epsilon^2}{16(\delta+2)\mathbb{V}(Z(l))}\Big).\end{aligned}\tag{4.3}$$

Similarly, we could show that if $\max_l\mathbb{E}Z(l)\geq-\epsilon/4$,

$$\mathbb{P}\Big(F(x_l)-\hat{F}_{nm}(x_l)>\frac{\epsilon}{2}\Big)\leq\exp\Big(-\frac{n\epsilon^2}{16(\delta+2)\mathbb{V}(Z(l))}\Big).\tag{4.4}$$

Based on Lemma 3.1, we have, for Lipschitz continuous $f_\xi(\cdot)$, $|\mathbb{E}Z(l)|\leq L_f\sigma_g/\sqrt{m}$ $\forall l=1,\ldots,Q$; for Lipschitz smooth $f_\xi(\cdot)$, $|\mathbb{E}Z(l)|\leq S\sigma_g^2/2m$ $\forall l=1,\ldots,Q$. Thus, $\max_l\mathbb{E}Z(l)\leq\epsilon/4$ is satisfied when m is sufficiently large. By analysis of Lemma 3.2, we know that $\mathbb{V}(Z(l))\leq\sigma_f^2+4M_fL_f\sigma_g/\sqrt{m}\leq\sigma_f^2+4M_fL_f\sigma_g$. Plugging this into (4.2) with $Q\leq\mathcal{O}(1)(\frac{4L_gL_fD_{\mathcal{X}}}{\epsilon})^d$, we obtain the desired result. □

Since $\hat{F}_{nm}(\hat{x}_{nm})-\hat{F}_{nm}(x^*)\leq 0$, we have

$$\begin{aligned}&\mathbb{P}\left(F(\hat{x}_{nm})-F(x^*)\geq\epsilon\right)\\ =&\ \mathbb{P}\Big([F(\hat{x}_{nm})-\hat{F}_{nm}(\hat{x}_{nm})]+[\hat{F}_{nm}(\hat{x}_{nm})-\hat{F}_{nm}(x^*)]+[\hat{F}_{nm}(x^*)-F(x^*)]\geq\epsilon\Big)\\ \leq&\ \mathbb{P}\Big(F(\hat{x}_{nm})-\hat{F}_{nm}(\hat{x}_{nm})\geq\epsilon/2\Big)+\mathbb{P}\Big(\hat{F}_{nm}(x^*)-F(x^*)\geq\epsilon/2\Big).\end{aligned}\tag{4.5}$$

Invoking Theorem 4.1, we immediately have the following result.

COROLLARY 4.1 (SAA under general Lipschitz continuous condition). *Under Assumption* 3.1, *for any $\delta>0$, there exists $\epsilon_1>0$ such that for $\epsilon\in(0,\epsilon_1)$, when $m\geq L_f^2\sigma_g^2/\epsilon^2$,*

$$\mathbb{P}\Big(F(\hat{x}_{nm})-F(x^*)>\epsilon\Big)\leq\mathcal{O}(1)\Big(\frac{8L_fL_gD_{\mathcal{X}}}{\epsilon}\Big)^d\exp\Big(-\frac{n\epsilon^2}{64(2+\delta)(\sigma_f^2+4M_fL_f\sigma_g)}\Big).\tag{4.6}$$

If, additionally, $f_\xi(\cdot)$ is S-Lipschitz smooth, then (4.6) *holds as long as $m\geq 2S\sigma_g^2/\epsilon$.*

This further implies the following sample complexity result.

COROLLARY 4.2. *With probability at least $1-\alpha$, the solution to the SAA problem is ϵ-optimal to the original CSO problem if the sample sizes n and m satisfy that*

$$n \geq \mathcal{O}(1)\frac{\sigma_f^2 + 4M_f L_f \sigma_g}{\epsilon^2}\left[d\log\left(\frac{8L_f L_g D_{\mathcal{X}}}{\epsilon}\right) + \log\left(\frac{1}{\alpha}\right)\right],$$

$$m \geq \begin{cases} \frac{L_f^2\sigma_g^2}{\epsilon^2}, & \text{under Assumption } 3.1,\\ \frac{2S\sigma_g^2}{\epsilon}, & f_\xi(\cdot) \text{ is also Lipschitz smooth.}\end{cases}$$

Ignoring the log factors, under Assumption 3.1*, the total sample complexity of SAA for achieving an ϵ-optimal solution is $T = mn + n = \mathcal{O}(d/\epsilon^4)$; when $f_\xi(\cdot)$ is Lipschitz smooth, the total sample complexity reduces to $T = mn + n = \mathcal{O}(d/\epsilon^3)$.*

The above result indicates that in general, the sample complexity of the SAA for the CSO problem is $\mathcal{O}(d/\epsilon^4)$ when assuming only Lipschitz continuity of the functions f_ξ and g_η. The sample complexity drops to $\mathcal{O}(d/\epsilon^3)$ assuming additionally Lipschitz smoothness of the outer function f_ξ. Notice that the complexity depends only linearly on the dimension of the decision set. This is quite different from three-stage stochastic optimization. In [37], for three-stage stochastic programming, the authors showed the sample sizes for estimating the second and the third stages need to be at least $\mathcal{O}(d/\epsilon^2)$, leading to a total of $\mathcal{O}(d^2/\epsilon^4)$ samples, to guarantee uniform convergence even for stagewise independent random variables.

Remark 4.1. In the special case when the outer function $f_\xi(\cdot)$ is linear, by a similar analysis, one could show that for any fixed m, $n \geq \mathcal{O}(d(L_f^2\sigma_g^2/m+\sigma_f^2)/\epsilon^{-2})$ guarantees that the optimal solution of $\hat{F}_{nm}(x)$ is ϵ-optimal to $F(x)$ with high probability. In this case, it makes sense to simply set $m = 1$, and the total sample complexity becomes $\mathcal{O}(d/\epsilon^2)$.

4.2. Sample complexity under error bound conditions.

In this subsection, we consider the case when the empirical function satisfies the Hölderian error bound condition, which includes the QG condition and strong convexity as special cases. The Hölderian error bound condition has been widely studied recently in the context of (stochastic) oracle-based algorithm for faster convergence; see, e.g., [19, 11, 44] and references therein. To the best of our knowledge, very few papers have exploited the Hölderian error bound condition for the SAA approach and analyzed the sample complexity under such a condition. We show that the CSO problem under the Hölderian error bound condition yields smaller orders of sample complexity for the SAA approach. We make the following two assumptions throughout this subsection.

ASSUMPTION 4.1. *The empirical function $\hat{F}_{nm}(x)$ satisfies the (μ,δ)-Hölderian error bound condition with $\mu > 0, \delta \geq 0$; i.e., it holds that*

$$\forall x \in \mathcal{X},\ \hat{F}_{nm}(x) - \min_{x\in\mathcal{X}} \hat{F}_{nm}(x) \geq \mu \inf_{z\in\mathcal{X}_{nm}^*} ||x-z||_2^{1+\delta},$$

where n, m are any positive integers, and $\mathcal{X}_{nm}^*$ is the optimal solution set of the empirical objective function $\hat{F}_{nm}(x)$ over $\mathcal{X}$.

ASSUMPTION 4.2. *The empirical function $\hat{F}_{nm}$ has a unique minimizer $\hat{x}_{nm}$ on $\mathcal{X}$ for any n and m.*

An interesting special case of Assumption 4.1 is the quadratic growth (QG) condition when $\delta = 1$. The QG condition is actually satisfied by a wide spectrum of objectives, such as strongly convex functions, general strongly convex functions composed with piecewise linear functions, and general piecewise convex quadratic functions. There are also many other specific examples arising in machine learning applications that satisfy the QG condition, including logistic loss composed with linear functions and neural networks with linear activation functions; see [6, 19] and references therein. Another interesting case is the polyhedral error bound condition when $\delta = 0$, which is known to hold true for many piecewise linear loss functions [4]. For both cases, these functions are not necessarily strongly convex or convex. Relevant problems with SAA objective $\hat{F}_{nm}$ satisfying the QG condition are discussed in Appendix D.

Assumption 4.2 could be restrictive and less straightforward to verify. In general, for a nonstrictly convex empirical objective function, the optimal solution is not necessarily unique. Yet, it is not exclusive to strictly convex functions. We illustrate one such example below. Finally, we point out that when $\hat{F}_{nm}(x)$ is strongly convex, for example, l_2 regularized convex empirical objective, the above assumptions hold naturally. In the following, we give some examples when $\hat{F}_{nm}(x)$ satisfies the QG condition.

Example 1. Consider the following one-dimensional function:

$$F(x) = \mathbb{E}_\xi\left[(\mathbb{E}_{\eta|\xi}[\eta]x)^2 + 3\sin^2(\mathbb{E}_{\eta|\xi}[\eta]x)\right],$$

where ξ and η can be any random vectors that satisfy $\eta|\xi \geq \sqrt{\mu}$ with probability 1. Denote $\bar{\eta}_i = \frac{1}{m}\sum_{j=1}^m \eta_{ij}$; the empirical function is given by

$$\hat{F}_{nm}(x) = \frac{1}{n}\sum_{i=1}^n \bar{\eta}_i^2 x^2 + \frac{3}{n}\sum_{i=1}^n \sin^2(\bar{\eta}_i x).$$

It can be easily verified that $\hat{F}_{nm}(x)$ satisfies the QG condition with parameter $\mu > 0$. Moreover, the empirical function $\hat{F}_{nm}(x)$ has a unique minimizer $x^* = 0$ for any m, n.

Example 2. Consider the robust logistic regression problem with the objective

$$F(x) = \mathbb{E}_{\xi=(a,b)}[\log(1 + \exp(-b\mathbb{E}_{\eta|\xi}[\eta]^T x))], \tag{4.7}$$

where $a \in \mathbb{R}^d$ is a random feature vector and $b \in \{1, -1\}$ is the label; $\eta = a + \mathcal{N}(0, \sigma^2 I_d)$ is a perturbed noisy observation of the input feature vector a. The empirical objective function $\hat{F}_{nm}(x)$ is given by

$$\hat{F}_{nm}(x) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp\left(-b_i\frac{1}{m}\sum_{j=1}^m \eta_{ij}^\top x\right)\right). \tag{4.8}$$

$\hat{F}_{nm}(x)$ satisfies the QG condition on any compact convex set in Appendix D. Note that the minimizer of a general empirical objective function is not necessarily always unique. However, the Hessian of $\hat{F}_{nm}(x)$ shows that $\hat{F}_{nm}(x)$ is strictly convex if $\frac{1}{m}\sum_{j=1}^m \eta_{ij}^\top \neq 0$ for all i, which is satisfied with high probability. Thus, $\hat{F}_{nm}(x)$ has a unique minimizer with high probability.

Next, we present our main result on the sample complexity of SAA.

THEOREM 4.2 (SAA under error bound condition). *Under Assumptions* 3.1, 4.1, *and* 4.2, *for any $\epsilon > 0$, we have*

$$\mathbb{P}(F(\hat{x}_{nm}) - F(x^*) \geq \epsilon) \leq \frac{1}{\epsilon}\left(L_f L_g\left(\frac{2L_f L_g}{\mu n}\right)^{1/\delta} + \frac{2L_f\sigma_g}{\sqrt{m}}\right). \tag{4.9}$$

If, additionally, $f_\xi(\cdot)$ is S-Lipschitz smooth, then we further have

$$\mathbb{P}(F(\hat{x}_{nm}) - F(x^*) \geq \epsilon) \leq \frac{1}{\epsilon}\left(L_f L_g\left(\frac{2L_fL_g}{\mu n}\right)^{1/\delta} + \frac{S\sigma_g^2}{m}\right). \tag{4.10}$$

Differently from the previous section, we use a stability argument to exploit the error bound condition. As shown in Lemma 3.1, the empirical function is a biased estimator of the original function due to the composition of $f_\xi(\cdot)$ and $g_\eta(\cdot,\xi)$. Introducing a perturbed set of samples could reduce some dependence in randomness. We define a bias term which will be used later in the proof:

$$\Delta(m) := \begin{cases} \frac{L_f\sigma_g}{\sqrt{m}}, & f_\xi(\cdot) \text{ is } L_f\text{-Lipschitz continuous,} \\ \frac{S\sigma_g^2}{2m}, & f_\xi(\cdot) \text{ is additionally } S\text{-Lipschitz smooth.} \end{cases} \tag{4.11}$$

Below we provide the detailed proof of Theorem 4.2.

Proof. Recall that x^* and $\hat{x}_{nm}$ are the minimizers of $F(x)$ and $\hat{F}_{nm}(x)$, respectively. It is clear that x^* has no randomness, and $\hat{x}_{nm}$ is a function of $\{\xi_i\}_{i=1}^n, \{\eta_{ij}\}_{j=1}^m$. We decompose the error $F(\hat{x}_{nm}) - F(x^*)$ into three terms and analyze each term below:

$$F(\hat{x}_{nm}) - F(x^*) = \underbrace{F(\hat{x}_{nm}) - \hat{F}_{nm}(\hat{x}_{nm})}_{:=\mathcal{E}_1} + \underbrace{\hat{F}_{nm}(\hat{x}_{nm}) - \hat{F}_{nm}(x^*)}_{:=\mathcal{E}_2} + \underbrace{\hat{F}_{nm}(x^*) - F(x^*)}_{:=\mathcal{E}_3}.$$

First, we use a stability argument and Lemma 3.1 to bound $\mathbb{E}\mathcal{E}_1 = \mathbb{E}[F(\hat{x}_{nm}) - \hat{F}_{nm}(\hat{x}_{nm})]$. Define

$$\hat{F}_{nm}^{(k)}(x) := \frac{1}{n}\sum_{i\neq k}^{n} f_{\xi_i}\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_{ij}}(x,\xi_i)\right) + \frac{1}{n} f_{\xi_k'}\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_{kj}'}(x,\xi_k')\right) \tag{4.12}$$

as the empirical function by replacing the kth outer sample ξ_k with another i.i.d. outer sample ξ_k', and replacing the corresponding inner samples $\{\eta_{kj}\}_{j=1}^m$ with $\{\eta_{kj}'\}_{j=1}^m$, which are sampled from the conditional distribution of $\mathbb{P}(\eta|\xi_k')$ for a given sample ξ_k'. Denote $\hat{x}_{nm}^{(k)} := \operatorname{argmin}_{x\in\mathcal{X}} \hat{F}_{nm}^{(k)}(x)$. We decompose $\mathbb{E}\mathcal{E}_1 = \mathbb{E}[F(\hat{x}_{nm}) - \hat{F}_{nm}(\hat{x}_{nm})]$ into three terms:

$$\begin{aligned}\mathbb{E}\mathcal{E}_1 =& \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n F(\hat{x}_{nm}) - \frac{1}{n}\sum_{k=1}^n f_{\xi_k}\left(\mathbb{E}_{\eta|\xi_k} g_\eta(\hat{x}_{nm}^{(k)},\xi_k)\right)\right] \\ &+ \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n f_{\xi_k}\left(\mathbb{E}_{\eta|\xi_k} g_\eta(\hat{x}_{nm}^{(k)},\xi_k)\right) - \frac{1}{n}\sum_{k=1}^n f_{\xi_k}\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_{kj}}(\hat{x}_{nm}^{(k)},\xi_k)\right)\right] \\ &+ \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n f_{\xi_k}\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_{kj}}(\hat{x}_{nm}^{(k)},\xi_k)\right) - \hat{F}_{nm}(\hat{x}_{nm})\right].\end{aligned} \tag{4.13}$$

Note that $\mathbb{E}[F(\hat{x}_{nm})] = \mathbb{E}[F(\hat{x}_{nm}^{(k)})]$ since ξ_k and ξ_k' are i.i.d., which implies that $\hat{x}_{nm}$ and $\hat{x}_{nm}^{(k)}$ follow an identical distribution. Since $\hat{x}_{nm}^{(k)}$ is independent of ξ_k, $\mathbb{E}[F(\hat{x}_{nm}^{(k)})] = \mathbb{E}[f_{\xi_k}(\mathbb{E}_{\eta|\xi_k} g(\hat{x}_{nm}^{(k)},\xi_k))]$ for any k. Then the first term in (4.13) is 0. As $\hat{x}_{nm}^{(k)}$ is independent of $\{\eta_{kj}\}_{j=1}^m$, the second term in (4.13) could be bounded by Lemma 3.1, and it holds that

$$\mathbb{E}\left[f_{\xi_k}\left(\mathbb{E}_{\eta|\xi_k} g_\eta(\hat{x}_{nm}^{(k)},\xi_k)\right) - f_{\xi_k}\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_{kj}}(\hat{x}_{nm}^{(k)},\xi_k)\right)\right] \leq \Delta(m). \tag{4.14}$$

Next, we upper bound the third term in (4.13). By definition, we have

$$\begin{aligned}\hat{F}_{nm}(\hat{x}_{nm}^{(k)}) - \hat{F}_{nm}(\hat{x}_{nm}) =& \hat{F}_{nm}^{(k)}(\hat{x}_{nm}^{(k)}) - \hat{F}_{nm}^{(k)}(\hat{x}_{nm}) \\ &+ \frac{1}{n} f_{\xi_k}\bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta_{kj}}(\hat{x}_{nm}^{(k)}, \xi_k)\bigg) - \frac{1}{n} f_{\xi_k}\bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta_{kj}}(\hat{x}_{nm}, \xi_k)\bigg) \\ &+ \frac{1}{n} f_{\xi'_k}\bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta'_{kj}}(\hat{x}_{nm}, \xi'_k)\bigg) - \frac{1}{n} f_{\xi'_k}\bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta'_{kj}}(\hat{x}_{nm}^{(k)}, \xi'_k)\bigg).\end{aligned} \tag{4.15}$$

By Lipschitz continuity of f_ξ and g_η and that $\hat{F}_{nm}^{(k)}(\hat{x}_{nm}^{(k)}) - \hat{F}_{nm}^{(k)}(\hat{x}_{nm}) \leq 0$, it holds that

$$\hat{F}_{nm}(\hat{x}_{nm}^{(k)}) - \hat{F}_{nm}(\hat{x}_{nm}) \leq \frac{2}{n} L_f L_g ||\hat{x}_{nm}^{(k)} - \hat{x}_{nm}||_2. \tag{4.16}$$

Since $\hat{x}_{nm}$ is the unique minimizer of $\hat{F}_{nm}(x)$, and $\hat{F}_{nm}(x)$ satisfies the QG condition with parameter μ, we have

$$\hat{F}_{nm}(\hat{x}_{nm}^{(k)}) - \hat{F}_{nm}(\hat{x}_{nm}) \geq \mu ||\hat{x}_{nm}^{(k)} - \hat{x}_{nm}||_2^{1+\delta}. \tag{4.17}$$

Combining with (4.16), we obtain

$$||\hat{x}_{nm}^{(k)} - \hat{x}_{nm}||_2 \leq \bigg(\frac{2L_f L_g}{\mu n}\bigg)^{1/\delta}. \tag{4.18}$$

By Lipschitz continuity of $f_\xi(\cdot)$ and $g_\eta(\cdot, \xi)$ and definition of $\hat{F}_{nm}(\hat{x}_{nm})$, we obtain

$$\mathbb{E}\bigg[\frac{1}{n}\sum_{k=1}^{n} f_{\xi_k}\bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta_{kj}}(\hat{x}_{nm}^{(k)}, \xi_k)\bigg) - \hat{F}_{nm}(\hat{x}_{nm})\bigg] \leq L_f L_g \bigg(\frac{2L_f L_g}{\mu n}\bigg)^{1/\delta}. \tag{4.19}$$

Combining (4.13), (4.19), and (4.14), we obtain

$$\mathbb{E}\mathcal{E}_1 \leq L_f L_g \bigg(\frac{2L_f L_g}{\mu n}\bigg)^{1/\delta} + \Delta(m). \tag{4.20}$$

Second, by optimality of $\hat{x}_{nm}$ of $\hat{F}_{nm}$, we have

$$\mathbb{E}\mathcal{E}_2 = \mathbb{E}[\hat{F}_{nm}(\hat{x}_{nm}) - \hat{F}_{nm}(x^*)] \leq 0. \tag{4.21}$$

Next, we bound $\mathbb{E}\mathcal{E}_3$. Define $\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} f_{\xi_i}\big(\mathbb{E}_{\eta|\xi_i}[g_\eta(x, \xi_i)]\big)$. Notice that x^* is independent of $\{\eta_{ij}\}_{j=1}^{m}$ for any $i = \{1, \ldots, n\}$ and $\mathbb{E}[\hat{F}_n(x^*) - F(x^*)] = 0$. By Lemma 3.1, it holds that

$$\mathbb{E}\mathcal{E}_3 = \mathbb{E}[\hat{F}_{nm}(x^*) - \hat{F}_n(x)] + \mathbb{E}[\hat{F}_n(x) - F(x)] \leq \Delta(m). \tag{4.22}$$

Combining (4.20), (4.21), and (4.22) with the Markov inequality, we obtain the desired result. ∎

The sample complexity of SAA under the Hölderian error bound condition follows directly.

Corollary 4.3. *Under Assumptions* 4.1 *and* 4.2*, with probability at least* $1-\alpha$*, the solution to the SAA problem is* ϵ*-optimal to the original CSO problem if the sample sizes* n *and* m *satisfy that*

$$n \geq \frac{(2L_f L_g)^{\delta+1}}{\mu(\alpha\epsilon)^{\delta}}, \quad m \geq \begin{cases} \frac{16L_f^2\sigma_g^2}{\alpha^2\epsilon^2}, & \textit{under Assumption } 3.1, \\ \frac{2S\sigma_g^2}{\alpha\epsilon}, & f_\xi(\cdot) \textit{ is also Lipschitz smooth.} \end{cases}$$

Hence, the total sample complexity of SAA for achieving an ϵ*-optimal solution is at most* $T = mn + n = \mathcal{O}(1/\epsilon^{\delta+2})$*; when* $f_\xi(\cdot)$ *is Lipschitz smooth, the total sample complexity reduces to* $T = mn + n = \mathcal{O}(1/\epsilon^{\delta+1})$.

In particular, when the empirical function is strongly convex or satisfies the QG condition, i.e., Assumption 4.1 with $\delta = 1$, this leads to the total sample complexity of $\mathcal{O}(1/\epsilon^3)$ for the Lipschitz continuous case and $\mathcal{O}(1/\epsilon^2)$ for the Lipschitz smooth case, respectively. From the above corollary, the error bound condition only affects the sample complexity of the outer samples, and the sample size decreases as δ decreases. As δ gets closer to zero, the sample complexity will essentially be dominated by the inner sample size.

A key difference between the results in Theorems 4.1 and 4.2 lies in the dependence on the problem dimension d and confidence level α. While the sample complexity under the Hölderian error bound condition is dimension-free, the dependence on the confidence level $1-\alpha$ grows from $\mathcal{O}(\log(1/\alpha))$ to $\mathcal{O}(1/\alpha^\delta)$. This is similar to classical results on SO for strongly convex objectives [36]. Theorem 4.2 could also be used to derive a dimension-free sample complexity of l_2 regularized SAA for a general convex CSO problem. See Appendix E for more details.

5. Sample complexity of SAA for CSO with independent random variables.

In this section, we consider the special case of CSO when the random variables ξ and η are independent. The objective then simplifies to

$$\min_{x\in\mathcal{X}} \quad F(x) := \mathbb{E}_\xi[f_\xi(\mathbb{E}_\eta[g_\eta(x,\xi)])]. \tag{5.1}$$

This is similar to yet slightly more general than (1.8), the compositional objective considered in [43, 42]. Note that the inner cost function we consider here is dependent on both ξ and η and thus cannot be written as a composition of two deterministic functions.

The sample complexity of SAA under the conditional sampling setting achieved in section 4 applies to this setting since it can be viewed as a special case of the former. However, since the inner expectation is no longer a conditional expectation, we now consider an alternative modified SAA, using the independent sampling scheme, in which we use the same set of samples to estimate the inner expectation. The procedure of the independent sampling scheme for solving (5.1) works as follows: first generate n i.i.d. samples $\{\xi_i\}_{i=1}^n$ from the distribution of ξ and m i.i.d. samples $\{\eta_j\}_{j=1}^m$ from the distribution of η, and then solve the following approximation problem:

$$\min_{x\in\mathcal{X}} \quad \hat{F}_{nm}(x) := \frac{1}{n}\sum_{i=1}^n f_{\xi_i}\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(x,\xi_i)\right). \tag{5.2}$$

As a result, the total sample complexity becomes $T = m + n$. Recently, in [9], a central limit theorem result for the SAA (5.2) with $m = n$ was established. The authors show that for Lipschitz smooth functions $f_\xi(\cdot)$ and $g_\eta(\cdot,\xi)=g_\eta(\cdot)$, the SAA estimator converges in distribution as follows:

$$\sqrt{m}\left(\min_{x\in\mathcal{X}}\hat{F}_{mm}(x)-\min_{x\in\mathcal{X}}F(x)\right)\to Z(W),$$

where $W(\cdot)=(W_1(\cdot),W_2(\cdot))$ is a zero-mean Brownian process with certain covariance functions, and $Z(\cdot)$ is a function that depends on the first-order information. This result only yields an asymptotic convergence rate of order $\mathcal{O}(1/\sqrt{m})$ for the SAA with $m=n$. Below, we will provide a finite-sample analysis for SAA and establish refined sample complexity results based on concentration inequality techniques.

In the SAA problem (5.2), the component functions $f_{\xi_i}\big(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(x,\xi_i)\big)$ share the same random vectors $\{\eta_j\}_{j=1}^m$ and are dependent. This is distinct from the SAA (1.5) considered in the previous section. Because of this key difference, the previous analysis will no longer apply to this modified SAA. We will resort to a different analysis for deriving the sample complexity. Similarly, we consider two structural assumptions, when the empirical objective is only known to be Lipschitz continuous and when the empirical objective also satisfies the error bound condition.

5.1. Sample complexity for Lipschitz continuous problems.

We first consider the case when the objective is Lipschitz continuous. We make the same basic assumptions on the Lipschitz continuity of $f_\xi(\cdot)$ and $g_\eta(\cdot,\xi)$ and boundedness of variances as described in Assumption 3.1. Our main result is summarized below.

THEOREM 5.1. *Under the independent sampling scheme and Assumption* 3.1*, for any $\delta>0$, there exists an $\epsilon_1>0$ such that for any $\epsilon\in(0,\epsilon_1)$, it holds that*

(5.3)
$$\begin{aligned}&\mathbb{P}\left(\sup_{x\in\mathcal{X}}|\hat{F}_{nm}(x)-F(x)|>\epsilon\right)\\ \leq&\mathcal{O}(1)\left(\frac{4L_fL_gD_{\mathcal{X}}}{\epsilon}\right)^d\left(\exp\left(-\frac{n\epsilon^2}{16(\delta+2)\sigma_f^2}\right)+nk\exp\left(-\frac{m\epsilon^2}{16(\delta+2)L_f^2\sigma_g^2}\right)\right).\end{aligned}$$

Here, d is the dimension of the decision set, and k is the dimension of the range of function g.

Proof. First, we pick a υ-net $\{x_l\}_{l=1}^Q$ on the decision set $\mathcal{X}$, such that $L_fL_g\upsilon=\epsilon/4$. Using an argument similar to that in the proof of Theorem 4.1, we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{x\in\mathcal{X}}|\hat{F}_{nm}(x)-F(x)|>\epsilon\right)&\leq\sum_{l=1}^Q\mathbb{P}\left(|\hat{F}_{nm}(x_l)-F(x_l)|>\frac{\epsilon}{2}\right)\\ \leq\sum_{l=1}^Q\mathbb{P}\left(|\hat{F}_{nm}(x_l)-\hat{F}_n(x_l)|>\frac{\epsilon}{4}\right)&+\sum_{l=1}^Q\mathbb{P}\left(|\hat{F}_n(x_l)-F(x_l)|>\frac{\epsilon}{4}\right).\end{aligned}\tag{5.4}$$

By Lipschitz continuity of $f_\xi(x)$ and Lemma 2.1, we have

(5.5)
$$\begin{aligned}\mathbb{P}\left(|\hat{F}_{nm}(x_l)-\hat{F}_n(x_l)|\geq\frac{\epsilon}{4}\right)\leq&\sum_{i=1}^n\mathbb{P}\left(||\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(x_l,\xi_i)-\mathbb{E}_\eta g_\eta(x_l,\xi_i)||_2\geq\frac{\epsilon}{4L_f}\right)\\ \leq&2nk\exp\left(-\frac{m\epsilon^2}{16(\delta+2)L_f^2\sigma_g^2}\right).\end{aligned}$$

By Lemma 2.1, we obtain

$$\mathbb{P}\biggl(|\hat{F}_n(x_l) - F(x_l)| \geq \frac{\epsilon}{4}\biggr) \leq 2\exp\biggl(-\frac{n\epsilon^2}{16(\delta+2)\sigma_f^2}\biggr). \tag{5.6}$$

Combining this with the fact that $Q \leq \mathcal{O}(1)(\frac{4L_gL_fD_{\mathcal{X}}}{\epsilon})^d$, we obtain the desired result.□

Invoking the relation in (4.5), the above theorem implies the following.

Corollary 5.1. *Under Assumption* 3.1, *with probability at least* $1-\alpha$, *the solution to the modified SAA problem* (5.2) *is* ϵ*-optimal to the original problem* (5.1) *if the sample sizes* n *and* m *satisfy*

$$n \geq \frac{\mathcal{O}(1)\sigma_f^2}{\epsilon^2}\biggl[d\log\biggl(\frac{8L_fL_gD_{\mathcal{X}}}{\epsilon}\biggr) + \log\biggl(\frac{1}{\alpha}\biggr)\biggr],$$
$$m \geq \frac{\mathcal{O}(1)L_f^2\sigma_g^2}{\epsilon^2}\biggl[d\log\biggl(\frac{8L_fL_gD_{\mathcal{X}}}{\epsilon}\biggr) + \log\biggl(\frac{1}{\alpha}\biggr) + \log(nk)\biggr].$$

Ignoring the log factors, under Assumption 3.1, *the total sample complexity of the modified SAA for achieving an* ϵ*-optimal solution is* $T = m + n = \mathcal{O}(d/\epsilon^2)$.

Note that this sample complexity is significantly smaller than that for the general CSO. The $\mathcal{O}(d/\epsilon^2)$ sample complexity also matches the lower bounds on sample complexity of SAA for classical SO with Lipschitz continuous objectives [25]; therefore, this result is unimprovable without further assumptions.

5.2. Sample complexity under error bound conditions. We now consider the case when the empirical objective satisfies Assumptions 4.1 and 4.2; i.e., the empirical objective $\hat{F}_{nm}(x)$ satisfies the error bound condition and has a unique minimizer for any integers n, m. Our main result is summarized as follows.

Theorem 5.2. *Under Assumptions* 3.1, 4.1, *and* 4.2, *for any* $\epsilon > 0$ *and* $\upsilon > 0$, *we have*

$$\begin{aligned} &\mathbb{P}(F(\hat{x}_{nm}) - F(x^*) \geq \epsilon) \\ &\leq \frac{1}{\epsilon}\biggl(L_fL_g\biggl(\frac{2L_fL_g}{\mu n}\biggr)^{1/\delta} + \mathcal{O}(1)\frac{L_fM_g\sqrt{d\log(D_{\mathcal{X}}/\upsilon)}}{\sqrt{m}} + \frac{L_f\sigma_g}{\sqrt{m}} + 2\upsilon L_fL_g\biggr). \end{aligned} \tag{5.7}$$

The solution to the modified SAA problem (5.2) *is* ϵ*-optimal to the problem* (5.1) *with probability at least* $1-\alpha$, *if* $\upsilon = \frac{\epsilon\alpha}{12L_fL_g}$, *and the sample sizes* n *and* m *satisfy that*

$$n \geq \frac{(2L_fL_g)^{\delta+1}}{\mu(\alpha\epsilon)^\delta},\ m \geq \max\biggl\{\biggl(\frac{12L_f\sigma_g}{\alpha\epsilon}\biggr)^2, \mathcal{O}(1)\biggl(\frac{6L_fM_g}{\alpha\epsilon}\biggr)^2 d\log\biggl(\frac{12D_{\mathcal{X}}L_fL_g}{\alpha\epsilon}\biggr)\biggr\}. \tag{5.8}$$

Similar to Theorem 4.2, the outer sample size is independent of dimension and decreases as δ decreases. As δ gets closer to zero, the sample complexity will essentially be dominated by the inner sample size. In particular, when the empirical function satisfies the QG condition or is strongly convex, i.e., Assumption 4.1 holds with $\delta = 1$, the outer sample size is reduced from $\mathcal{O}(d/\epsilon^2)$ in the Lipschitz continuous case to $\mathcal{O}(1/\epsilon)$. Yet, the total sample complexity remains $\mathcal{O}(d/\epsilon^2)$.

For a CSO problem with independent random vectors (5.1), both SAA approaches, through conditional sampling, or independent sampling, can be applied to solve the problem. Comparing Theorem 4.2 and Theorem 5.2, when smoothness and the quadratic growth condition are satisfied, the sample complexities of these two SAA approaches achieve the same order $\mathcal{O}(1/\epsilon^2)$, except for an extra $O(d)$ factor for the independent sampling. Interestingly, for a given small dimension d and the same sample budget T, the independent sampling might outperform the conditional sampling scheme since the constant factor in the sample complexity of conditional sampling is much larger. The numerical experiment on our testing cases in the next section further supports the finding.

In contrast to the sample complexity established in section 4 for the conditional sampling setting, a notable difference here is that the Lipschitz smoothness condition does not necessarily help reduce the sample complexity. This result aligns with the central limit theorem established in [9]. One of the reasons arises from the interdependence among the component functions in the modified SAA objective, leading to extra variance. Because of that, the analysis requires sophisticated arguments to handle the dependence and is much more involved. We defer the proof to Appendix B.

Remark 5.1. Although the overall $\mathcal{O}(1/\epsilon^2)$ sample complexity cannot be further improved in general, it is worth pointing out that, for some interesting specific instances, the modified SAA could achieve lower sample complexity than what is described from theory. We illustrate this with the following example.

Example 3. For $\gamma > 0$, consider the following problem:

$$\min_{x\in\mathcal{X}} F(x) := H(\mathbb{E}_\eta[x+\eta],\gamma) + (\mathbb{E}_\eta[x+\eta])^2,$$

where $\eta \sim \mathcal{N}(0,\sigma_\eta^2)$ and $H(\cdot,\gamma)$ is the Huber function, i.e.,

$$H(x,\gamma) = \begin{cases} |x| - \dfrac{1}{2}\gamma & \text{for } |x| > \gamma, \\ \dfrac{1}{2\gamma}x^2 & \text{for } |x| \le \gamma. \end{cases} \tag{5.9}$$

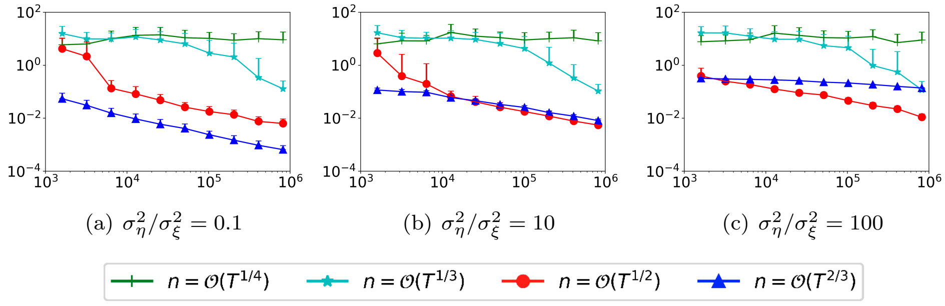
Note that here $f_\xi(x) := f(x) = H(x,\gamma) + x^2$ is deterministic, and $g_\eta(x,\xi) = x + \eta$. When $\gamma > 0$, $f(x)$ is $1/\gamma$-Lipschitz smooth. When $\gamma \to 0$, $f(x) \to |x| + x^2$, which is no longer differentiable. In this example, $x^* = \operatorname{argmin}_{x\in\mathcal{X}} F(x) = -\mathbb{E}\eta$; $F^* = \min_{x\in\mathcal{X}} F(x) = 0$. The empirical objective becomes $\hat{F}_m(x) = H(x+\bar{\eta},\gamma) + (x+\bar{\eta})^2$, where $\bar{\eta} = \frac{1}{m}\sum_{j=1}^m \eta_j$. Thus, $\hat{x}_m = \operatorname{argmin}_{x\in\mathcal{X}} \hat{F}_m(x) = -\bar{\eta}$. We show that the error of SAA satisfies

$$0 \le \mathbb{E}F(\hat{x}_m) - F(x^*) - \left(\frac{\sigma_\eta^2}{2\gamma m}\operatorname{erf}\left(\sqrt{\frac{\gamma^2 m}{2\sigma_\eta^2}}\right) + \frac{\sigma_\eta^2}{m}\right) \le \sqrt{\frac{\sigma_\eta^2}{2\pi m}}\exp\left(-\frac{m\gamma^2}{2\sigma_\eta^2}\right), \tag{5.10}$$

where $\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}}\int_0^x \exp(-x^2)dx$. As a result, when $\gamma \to 0$,

$$\lim_{\gamma\to 0} \mathbb{E}F(\hat{x}_m) - F(x^*) = \sqrt{\frac{\sigma_\eta^2}{2\pi m}} + \frac{\sigma_\eta^2}{m}. \tag{5.11}$$

For completeness, we provide detailed derivation in Appendix C. This example shows that the SAA error improves from $\mathcal{O}(1/\sqrt{m})$ to $\mathcal{O}(1/m)$ as the objective transitions from nonsmooth to smooth. When $\gamma \to 0$, the function becomes non-Lipschitz differentiable and the $O(1/\sqrt{m})$ bound for this setting is indeed tight. It remains an interesting open problem to identify sufficient conditions for achieving theoretically better sample complexity under the independent sampling scheme.



(a) $\sigma_\eta^2/\sigma_\xi^2 = 0.1$

(b) $\sigma_\eta^2/\sigma_\xi^2 = 10$

(c) $\sigma_\eta^2/\sigma_\xi^2 = 100$

FIG. 6.1. *Logistic regression; conditional sampling; dimension $d = 10$.*

6. Numerical experiments.

In this section, we conduct numerical experiments based on two applications, logistic regression and robust regression, to demonstrate the performance of SAA for solving CSO problems. For a fixed sample budget T, we adopt different sample allocation strategies for (m, n) and compute the corresponding accuracy of the SAA estimators. We repeat 30 runs for each sample allocation and report the average performance. The SAA problems are solved by CVXPY 1.0.9 [10].

6.1. Robust logistic regression.

We consider the robust logistic regression problem in Example 2. The problem is formulated in (4.7) and its SAA counterpart is of the form (4.8) with domain $\mathcal{X} = \{x | x \in \mathbb{R}^d, \|x\|_2 \leq 100\}$.

Note that from Example 2, f is Lipschitz smooth, and $\hat{F}_{nm}(x)$ satisfies QG condition on any compact convex set and with high probability has a unique minimizer for large n. Theorem 4.2 implies that the theoretical optimal sample allocation strategy is $n = \mathcal{O}(1/\sqrt{T})$ and $m = \mathcal{O}(1/\sqrt{T})$.

In the experiment, we set $d = 10$ and the samples of $\xi = (a, b)$ and η are generated as follows: $a_i \sim \mathcal{N}(0, \sigma_\xi^2 I_d)$, $b_i = \{\pm 1\}$ according to the sign of $a_i^T x^*$, and $\eta_{ij} \sim \mathcal{N}(a_i, \sigma_\eta^2 I_d)$. We set $\sigma_\xi^2 = 1$ and consider three cases for σ_η: $\sigma_\eta^2 = \{0.1, 10, 100\}$, corresponding to low, medium, and high variances from inner randomness. For a given sample budget T ranging from 10^3 to 10^6, four different sample allocation strategies are considered, i.e., $n = [T^{1/4}]$, $n = [T^{1/3}]$, $n = [T^{1/2}]$, and $n = [T^{2/3}]$. We then compute the average estimation error $F(\hat{x}_{nm}) - F^*$ over 30 runs and its standard deviation. The results are summarized in Figure 6.1, where the x-axis denotes the sample budget T, and the y-axis shows the estimation error. Each curve represents a sampling scheme, showing the average error and upper confidence bound.

The trend from Figure 6.1(a)–(c) shows that when the inner variance is relatively large, setting $n = \mathcal{O}(T^{1/2})$ consistently outperforms the other sampling strategies, which matches our analysis. The error bar suggests that a larger number of outer samples results in a smaller deviation of the estimation accuracy.

6.2. Robust regression.

We now examine the robust regression problem, where the objective is no longer Lipschitz differentiable. The problem is as follows:

$$\min_{x \in \mathcal{X}} F(x) = \mathbb{E}_{\xi=(a,b)} |\mathbb{E}_{\eta|\xi} \eta^\top x - b|, \tag{6.1}$$

where $a \in \mathbb{R}^d$ is a random feature vector, $b \in \mathbb{R}$ is the label, $\eta = a + \mathcal{N}(0, \sigma_\eta^2 I_d)$ is a perturbed noisy observation of the input feature vector a, and the domain is $\mathcal{X} = \{x | x \in \mathbb{R}^d, \|x\|_2 \leq 100\}$. For comparison purposes, we also consider the smoothed version of this problem based on the Huber function:

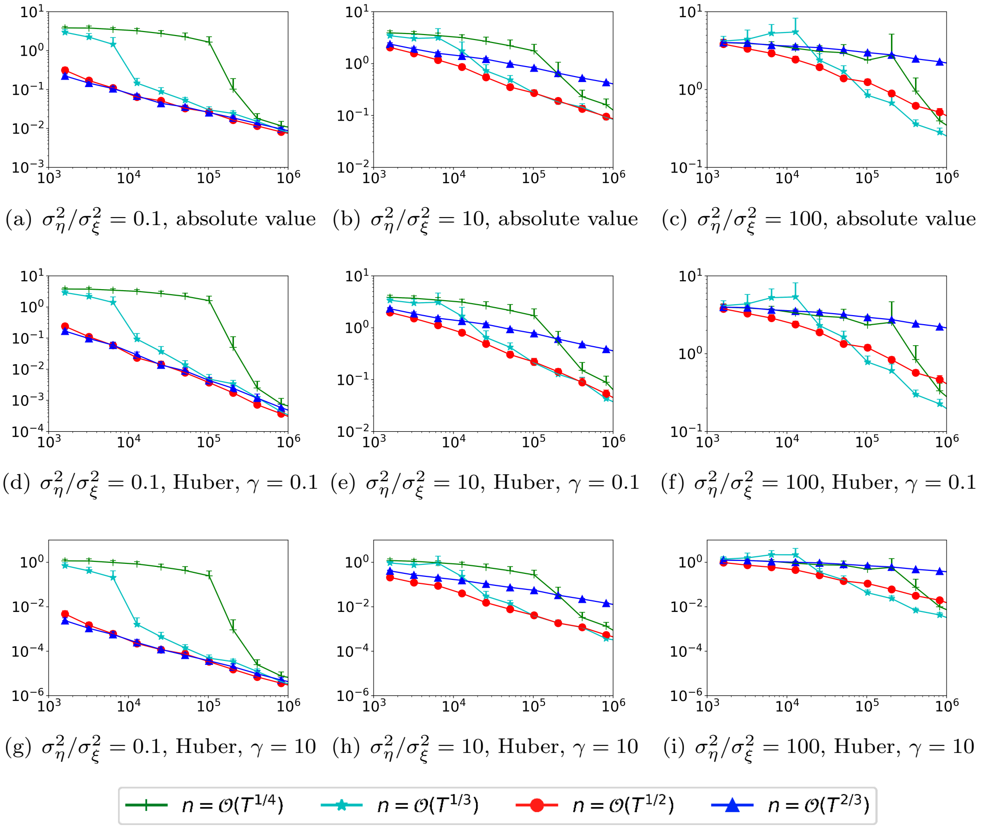
$$\min_{x\in\mathcal{X}} F^{\gamma}(x) = \mathbb{E}_{\xi=(a,b)} H\left(\mathbb{E}_{\eta\,|\,\xi}\eta^{\top}x - b, \gamma\right),$$

where $\gamma > 0$ is the smoothness parameter. The empirical functions for these two objectives are given by

$$\hat{F}_{nm}(x) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{m}\sum_{i=1}^{m}\eta_{ij}^{\top}x - b_i\right|, \quad \hat{F}^{\gamma}_{nm}(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{1}{m}\sum_{i=1}^{m}\eta_{ij}^{\top}x - b_i, \gamma\right).$$

Theorems 4.1 and 4.2 indicate that Lipschitz smoothness of outer function $f_\xi(x)$ helps reduce the inner sample size required to achieve the same level of accuracy. For a given budget T, the theoretical optimal sample allocation strategy for these two problems is $n = \mathcal{O}(T^{1/2})$ and $n = \mathcal{O}(T^{2/3})$, respectively.

In our experiment, we set $d = 20$. Samples of $\xi = (a, b)$ and η are generated as follows: $a_i \sim \mathcal{N}(0, \sigma_\xi^2 I_d)$, $b_i = a_i^{\top}x^*$, $\eta_{ij} \sim \mathcal{N}(a_i, \sigma_\eta^2 I_d)$. As in the previous experiment, we measure the average error and upper confidence bound for both problems with sample budget T ranging from 10^3 to 10^6 under four different sample allocation strategies over 30 runs. We also consider two sets of smoothness parameters, $\gamma \in \{0.1, 10\}$. The results are summarized in Figure 6.2.

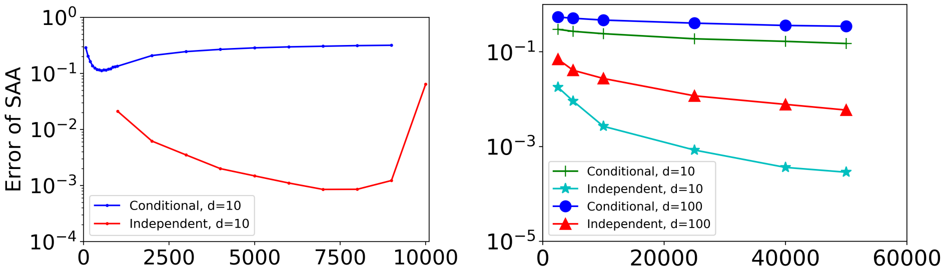


(a) $\sigma_\eta^2/\sigma_\xi^2 = 0.1$, absolute value (b) $\sigma_\eta^2/\sigma_\xi^2 = 10$, absolute value (c) $\sigma_\eta^2/\sigma_\xi^2 = 100$, absolute value

(d) $\sigma_\eta^2/\sigma_\xi^2 = 0.1$, Huber, $\gamma = 0.1$ (e) $\sigma_\eta^2/\sigma_\xi^2 = 10$, Huber, $\gamma = 0.1$ (f) $\sigma_\eta^2/\sigma_\xi^2 = 100$, Huber, $\gamma = 0.1$

(g) $\sigma_\eta^2/\sigma_\xi^2 = 0.1$, Huber, $\gamma = 10$ (h) $\sigma_\eta^2/\sigma_\xi^2 = 10$, Huber, $\gamma = 10$ (i) $\sigma_\eta^2/\sigma_\xi^2 = 100$, Huber, $\gamma = 10$

FIG. 6.2. *Error of SAA for absolute value loss and Huber loss; dimension $d = 20$.*



(a) $d = 10$, $T = 10^4$, x-axis: varying batch n (b) Various d, x-axis: varying total samples T

FIG. 6.3. *Comparison of conditional sampling and independent sampling schemes.*

Figure 6.2 (a)–(c) shows that setting $n = \mathcal{O}(\sqrt{T})$ indeed yields almost the best accuracy for absolute value loss minimization, which again matches our analysis. The overall performance of SAA for the original and that of the smoothed problems behaves quite similarly in this case, yet solving the smoothed problem yields much better accuracy under the same budget. This also supports our theoretical findings that the sample complexity is lower for smooth problems.

6.3. Comparison of conditional sampling and independent sampling. In this experiment, we consider a modified logistic regression example that falls into the special case with independent inner and outer randomness:

$$\min_{x\in\mathcal{X}} F(x) = \mathbb{E}_{\xi=(a,b)} \log(1+\exp(-b(\mathbb{E}_\eta \eta + a)^\top x)),$$

where $a \sim \mathcal{N}(0, \sigma_\xi^2 I_d) \in \mathbb{R}^d$ is a random feature vector, $b \in \{\pm 1\}$, and $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I_d)$ is the noise. The empirical function of the two sampling schemes $\hat{F}_{nm}(x)$ is of the form

$$\hat{F}_{nm}(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-b_i\left(\frac{1}{m}\sum_{j=1}^{m}\eta_{ij}+a_i\right)^\top x\right)\right).$$

When employing the independent sampling scheme, we generate $\{\eta_{1j}\}_{j=1}^m$ and let $\eta_{ij} = \eta_{1j}$ for all $i > 1$.

For both sampling schemes, the optimal allocation for n is on the order of $\mathcal{O}(\sqrt{T})$, and m is set to $m = T/n$ or $m = T - n$. In the experiment, $d = \{10, 100\}$, $\sigma_\xi^2 = 1$, and $\sigma_\eta^2 = 10$, and the samples are generated accordingly. For any given sample budget T, we compare the performance of the two sampling schemes under different choices of outer sample n varying from 0 to 10000.

Figure 6.3(a) illustrates the comparison when $d = 10$ and $T = 10000$. The bell shape in Figure 6.3(a) reflects a clear bias-variance tradeoff for different n and m.

In Figure 6.3(b), we report the best performance (by choosing the best n) of these two sampling schemes with $d \in \{10, 100\}$ and T ranging from 1000 to 50000. Figure 6.3(b) shows that the independent sampling scheme always achieves a smaller error for the logistic regression problem. The gap between the two schemes decreases as the dimension increases, which also matches our analysis.

7. Conclusion. In this paper, we introduce the class of conditional stochastic optimization problems and provide sample complexity analysis of sample average approximation under different structural assumptions. Our results show that the overall sample complexity can be significantly reduced under the Lipschitz smoothness condition, which is very different from the theory of classical SO and multistage stochastic programming. By exploiting error bound conditions, the sample complexity could be further reduced. To the best of our knowledge, these are the first nonasymptotic sample complexity results established in the context of conditional stochastic optimization. For future work, we will investigate stochastic approximation algorithms for solving this family of problems and establish their sample complexities.

Appendix A. Proof of propositions.

A.1. Proof of Lemma 2.1.

Proof. The proof of the one-dimensional random variable case was given in [20] using the Chernoff bound. Based on that, we consider the case when X is a zero-mean random vector in $\mathbb{R}^k$. Denote $X_i = (X_i^1, X_i^2, \dots, X_i^k)^\top$ for $i = 1, \dots, n$, $\sigma_j^2 = \mathbb{V}(X^j)$, $z_j = \frac{\sum_{j=1}^k \sigma_j^2}{\sigma_j^2}$, and $I_j(\cdot)$ the rate function of the jth coordinate of the random vector X. We have

$$\begin{aligned}\mathbb{P}(||\bar{X}||_2 \geq \epsilon) &= \mathbb{P}\bigg(\sum_{j=1}^k (\bar{X}^j - \mathbb{E}X^j)^2 \geq \epsilon^2\bigg) \leq \sum_{j=1}^k \mathbb{P}\bigg((\bar{X}^j)^2 \geq \frac{\epsilon^2}{z_j}\bigg)\\ &= \sum_{j=1}^k \mathbb{P}\bigg(|\bar{X}^j| \geq \frac{\epsilon}{\sqrt{z_j}}\bigg) \leq \sum_{j=1}^k \exp\bigg(-n\min\bigg\{I_j\bigg(\frac{\epsilon}{\sqrt{z_j}}\bigg); I_j\bigg(-\frac{\epsilon}{\sqrt{z_j}}\bigg)\bigg\}\bigg).\end{aligned} \tag{A.1}$$

By Lemma 2.1 and by definition, we get

$$\mathbb{P}(||\bar{X}||_2 \geq \epsilon) \leq 2\sum_{j=1}^k \exp\bigg(-\frac{n\epsilon^2}{(\delta+2)z_j\sigma_j^2}\bigg) = 2k\exp\bigg(-\frac{n\epsilon^2}{(\delta+2)\sum_{j=1}^k \sigma_j^2}\bigg).$$

Using the fact that $\sum_{j=1}^k \sigma_j^2 \leq \mathbb{E}||X||_2^2$, we obtain the desired result. □

Appendix B. Proof of Theorem 5.2.

Convergence analysis. We follow a decomposition similar to the one we followed in proving Theorem 4.2 and use the same notations, like $\hat{F}_{nm}^{(k)}(x)$ and $\hat{x}_{nm}^{(k)}$, the perturbed empirical function and its minimizer, except that we replace all the η_{kj} with η_j for $k = 1, \dots, n$ and replace the conditional expectation $\mathbb{E}_{\eta\,|\,\xi}$ with $\mathbb{E}_\eta$. Unfortunately, one will immediately notice that Lemma 3.1 is no longer applicable for bounding the second term in (4.13):

$$\mathbb{E}\bigg[\frac{1}{n}\sum_{k=1}^n f_{\xi_k}\bigg(\mathbb{E}_\eta g_\eta(\hat{x}_{nm}^{(k)}, \xi_k)\bigg) - \frac{1}{n}\sum_{k=1}^n f_{\xi_k}\bigg(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(\hat{x}_{nm}^{(k)}, \xi_k)\bigg)\bigg].$$

Because the minimizer $\hat{x}_{nm}^{(k)}$ depends on $\{\eta_j\}_{j=1}^m$, Lemma 3.1 is not applicable. Below we provide the detailed proof of Theorem 5.2.

Proof. Define $\mathcal{E}_1 := F(\hat{x}_{nm}) - \hat{F}_{nm}(\hat{x}_{nm})$ and

$$\hat{F}_{nm}^{(k)}(x) := \frac{1}{n}\sum_{i\neq k}^n f_{\xi_i}\bigg(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(x, \xi_i)\bigg) + \frac{1}{n} f_{\xi_k'}\bigg(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(x, \xi_k')\bigg),$$

the empirical function, by replacing the outer sample ξ_k with an i.i.d. sample ξ_k'. Denote $\hat{x}_{nm}^{(k)} = \operatorname{argmin}_{x\in\mathcal{X}} \hat{F}_{nm}^{(k)}(x)$. Then, $\mathbb{E}\mathcal{E}_1$ could be written as

$$\begin{aligned}\mathbb{E}\mathcal{E}_1 =&\mathbb{E}\bigg[F(\hat{x}_{nm}) - \frac{1}{n}\sum_{k=1}^{n} f_{\xi_k}\bigg(\mathbb{E}_\eta g_\eta(\hat{x}_{nm}^{(k)},\xi_k)\bigg)\bigg]\\ &+\mathbb{E}\bigg[\frac{1}{n}\sum_{k=1}^{n} f_{\xi_k}\bigg(\mathbb{E}_\eta g_\eta(\hat{x}_{nm}^{(k)},\xi_k)\bigg) - \frac{1}{n}\sum_{k=1}^{n} f_{\xi_k}\bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta_j}(\hat{x}_{nm}^{(k)},\xi_k)\bigg)\bigg]\\ &+\mathbb{E}\bigg[\frac{1}{n}\sum_{k=1}^{n} f_{\xi_k}\bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta_j}(\hat{x}_{nm}^{(k)},\xi_k)\bigg) - \hat{F}_{nm}(\hat{x}_{nm})\bigg].\end{aligned} \tag{B.1}$$

Since ξ_k and ξ_k' are i.i.d., $\hat{x}_{nm}$ and $\hat{x}_{nm}^{(k)}$ follow identical distribution. Then $\mathbb{E}F(\hat{x}_{nm}) = \mathbb{E}F(\hat{x}_{nm}^{(k)})$. As $\hat{x}_{nm}^{(k)}$ is independent of ξ_k, by definition of $F(x)$, we know $\mathbb{E}F(\hat{x}_{nm}^{(k)}) = \mathbb{E}f_{\xi_k}(\mathbb{E}_\eta g_\eta(\hat{x}_{nm}^{(k)},\xi_k))$ for any $k=1,\dots,n$. As a result, the first term is 0.

To analyze the second term, denote

$$H_k(x) := f_{\xi_k}\bigg(\mathbb{E}_\eta g_\eta(x,\xi_k)\bigg) - f_{\xi_k}\bigg(\frac{1}{m}\sum_{j=1}^{m} g_{\eta_j}(x,\xi_k)\bigg).$$

We pick a υ-net $\{x_l\}_{l=1}^{Q}$ for the decision set $\mathcal{X}$, such that for any $x\in\mathcal{X}$, there exist $l_0\in\{1,\dots,Q\}$, $\|x-x_{l_0}\|\le \upsilon$. Then it holds for any $s>0$ that

$$\begin{aligned}&\exp\left(s\mathbb{E}H_k(\hat{x}_{nm}^{(k)})\right) \le \exp\left(s\mathbb{E}\max_{l=1,\dots,Q} H_k(x_l) + 2s\upsilon L_f L_g\right)\\ \le&\mathbb{E}\exp\left(s\max_{l=1,\dots,Q} H_k(x_l) + 2s\upsilon L_f L_g\right) = \mathbb{E}\max_{l=1,\dots,Q}\exp\left(sH_k(x_l) + 2s\upsilon L_f L_g\right)\\ \le&\mathbb{E}\sum_{l=1}^{Q}\exp\left(sH_k(x_l) + 2s\upsilon L_f L_g\right) = \sum_{l=1}^{Q}\mathbb{E}\exp\left(sH_k(x_l) + 2s\upsilon L_f L_g\right).\end{aligned} \tag{B.2}$$

The first inequality holds as $\hat{x}_{nm}^{(k)}$ is independent of ξ_k, and $f_\xi(\cdot)$ and $g_\eta(\cdot,\xi)$ are Lipschitz continuous, which implies

$$H_k(\hat{x}_{nm}^{(k)}) \le \sup_{x\in\mathcal{X}} H_k(x) \le \max_{l=1,\dots,Q} H_k(x_l) + 2\upsilon L_f L_g.$$

The second inequality holds by Jensen's inequality. Next we show that $H_k(x_l) - \mathbb{E}H_k(x_l)$ is a sub-Gaussian random variable for any given ξ_k. Since $H_k(x_l)$ is a function of $\{\eta_j\}_{j=1}^{m}$, denote $H_k(x_l) := \tilde{H}(\eta_1,\dots,\eta_m)$. Then for any $p\in[m]$, and given $\eta_1,\dots,\eta_{p-1},\eta_{p+1},\dots,\eta_m$, we have

$$\begin{aligned}&\sup_{\eta_p'}\tilde{H}(\eta_1,\dots,\eta_{p-1},\eta_p',\eta_{p+1},\dots,\eta_m) - \inf_{\eta_p''}\tilde{H}(\eta_1,\dots,\eta_{p-1},\eta_p'',\eta_{p+1},\dots,\eta_m)\\ =&\sup_{\eta_p',\eta_p''}\mathbb{E}_{\xi_k} f_{\xi_k}\bigg(\frac{1}{m}\sum_{j\ne p}^{m} g_{\eta_j}(x,\xi_k) + \frac{1}{m}g_{\eta_p''}(x,\xi_k)\bigg) - f_{\xi_k}\bigg(\frac{1}{m}\sum_{j\ne p}^{m} g_{\eta_j}(x,\xi_k) + \frac{1}{m}g_{\eta_p'}(x,\xi_k)\bigg)\\ \le&\sup_{\eta_p',\eta_p''}\mathbb{E}_{\xi_k}\frac{L_f}{m}\bigg|g_{\eta_p''}(x,\xi_k) - g_{\eta_p'}(x,\xi_k)\bigg|\\ \le&\frac{2M_gL_f}{m},\end{aligned}$$

where M_g is the upper bound of $g_\eta(\cdot, \xi)$ on $\mathcal{X}$. This implies that $H_k(x_l) = \tilde{H}(\eta_1, \dots, \eta_m)$ has bounded difference $\frac{2M_g L_f}{m}$. By McDiarmid's inequality [26], for any $r > 0$,

$$\mathbb{P}(H_k(x_l) - \mathbb{E}H_k(x_l) \geq r) \leq 2\exp\left(- \frac{r^2 m}{2M_g^2 L_g^2} \right).$$

This implies that $H_k(x_l) - \mathbb{E}H_k(x_l)$ is a sub-Gaussian random variable with zero mean and variance proxy $2M_g^2 L_f^2/m$ for any given ξ_k. By definition it yields

$$\mathbb{E}\exp\left(s\left[H_k(x_l) - \mathbb{E}H_k(x_l)\right]\right) \leq \exp\left(\frac{2M_g^2 L_f^2 s^2}{m}\right).$$

Since x_l is independent of random vectors $\{\eta_j\}_{j=1}^m$, by Lemma 3.1, we know $\mathbb{E}H_k(x_l) \leq \frac{L_f \sigma_g}{\sqrt{m}}$. This further implies

$$\mathbb{E}\exp(sH_k(x_l)) \leq \exp\left(\frac{2M_g^2 L_f^2 s^2}{m} + \frac{sL_f\sigma_g}{\sqrt{m}}\right).$$

With (B.2), we have

$$\exp\left(s\mathbb{E}H_k(\hat{x}_{nm}^{(k)})\right) \leq Q\exp\left(\frac{2M_g^2 L_f^2 s^2}{m} + \frac{sL_f\sigma_g}{\sqrt{m}} + 2s\upsilon L_f L_g\right).$$

Taking the logarithm, dividing s on each side, and minimizing over s yields

$$\mathbb{E}H_k(\hat{x}_{nm}^{(k)}) \leq 2\sqrt{\frac{2\log(Q)L_f^2 M_g^2}{m}} + \frac{L_f\sigma_g}{\sqrt{m}} + 2\upsilon L_f L_g.$$

Since $Q \leq \mathcal{O}(1)(D_{\mathcal{X}}/\upsilon)^d$, we have

$$\mathbb{E}H_k(\hat{x}_{nm}^{(k)}) \leq \mathcal{O}(1)\frac{L_f M_g}{\sqrt{m}}\sqrt{d\log\left(\frac{D_{\mathcal{X}}}{\upsilon}\right)} + \frac{L_f\sigma_g}{\sqrt{m}} + 2\upsilon L_f L_g. \tag{B.3}$$

For the third term in (B.1), by following the similar steps from (4.15) to (4.18), we obtain

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n f_{\xi_k}\left(\frac{1}{m}\sum_{j=1}^m g_{\eta_j}(\hat{x}_{nm}^{(k)}, \xi_k)\right) - \hat{F}_{nm}(\hat{x}_{nm})\right] \leq L_f L_g\left(\frac{2L_f L_g}{\mu n}\right)^{1/\delta}. \tag{B.4}$$

Combining with (B.1), (B.3), and (B.4), we have
(B.5)

$$\mathbb{E}\mathcal{E}_1 \leq L_f L_g\left(\frac{2L_f L_g}{\mu n}\right)^{1/\delta} + \mathcal{O}(1)\frac{L_f M_g}{\sqrt{m}}\sqrt{d\log\left(\frac{12D_{\mathcal{X}}L_f L_g}{\alpha\epsilon}\right)} + \frac{L_f\sigma_g}{\sqrt{m}} + 2\upsilon L_f L_g.$$

Similarly to the steps from (4.21) and (4.22), by optimality of $\hat{x}_{nm}$ of $\hat{F}_{nm}$ and Lemma 3.1, we obtain

$$\mathbb{E}[\hat{F}_{nm}(\hat{x}_{nm}) - \hat{F}_{nm}(x^*)] \leq 0; \quad \mathbb{E}[\hat{F}_{nm}(x^*) - F(x^*)] \leq \frac{L_f\sigma_g}{\sqrt{m}}. \tag{B.6}$$

Finally, combining (B.5) and (B.6) with the Markov inequality, we obtain (5.7).

Let

$$L_f L_g\left(\frac{2L_f L_g}{\mu n\epsilon^\delta}\right)^{1/\delta} \le \frac{\alpha}{2};\quad \mathcal{O}(1)\frac{L_f M_g}{\sqrt{m\epsilon^2}}\sqrt{d\log\left(\frac{D_{\mathcal{X}}}{\upsilon}\right)} \le \frac{\alpha}{6};\quad \frac{2L_f\sigma_g}{\sqrt{m\epsilon^2}} \le \frac{\alpha}{6}.$$

We obtain the desired sample complexity (5.8). □

Appendix C. Example of Huber loss minimization. To show (5.10), denote $Y = \mathbb{E}\eta - \bar{\eta}$; then $Y \sim \mathcal{N}(0, \frac{\sigma_\eta^2}{m})$. Then the error of SAA is

$$\begin{aligned}\mathbb{E}F(\hat{x}_m) - F(x^*) &= \mathbb{E}H(\mathbb{E}\eta - \bar{\eta}, \gamma) + \mathbb{E}(\bar{\eta} - \mathbb{E}\eta)^2 \\ &= \int_0^\gamma \frac{1}{\gamma}y^2 p(y)dy + 2\int_\gamma^{+\infty}\left(y - \frac{1}{2}\gamma\right)p(y)dy + \mathbb{E}Y^2,\end{aligned} \tag{C.1}$$

where $p(y) = \frac{\sqrt{m}}{\sqrt{2\pi\sigma_\eta^2}}\exp\left(-\frac{my^2}{2\sigma_\eta^2}\right)$ is the PDF of Y, and $\mathbb{E}Y^2 = \frac{\sigma_\eta^2}{m}$. Denote $\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}}\int_0^x \exp(-x^2)dx$, $y_1 := y\sqrt{\frac{m}{2\sigma_\eta^2}}$. The first term in (C.1) is

$$\begin{aligned}\int_0^\gamma \frac{1}{\gamma}y^2 p(y)dy =& \frac{2\sigma_\eta^2}{m\gamma\sqrt{\pi}}\int_0^{\gamma\sqrt{\frac{m}{2\sigma_\eta^2}}} y_1^2\exp(-y_1^2)dy_1 \\ =& \frac{\sigma_\eta^2}{2\gamma m}\operatorname{erf}\left(\sqrt{\frac{\gamma^2 m}{2\sigma_\eta^2}}\right) - \sqrt{\frac{\sigma_\eta^2}{2\pi m}}\exp\left(-\frac{\gamma^2 m}{2\sigma_\eta^2}\right).\end{aligned}$$

We use the fact that

$$\int_0^z x^2\exp(-x^2)dx = \frac{1}{4}\sqrt{\pi}\operatorname{erf}(z) - \frac{1}{2}\exp(-z^2)z.$$

The second term in (C.1) is bounded by

$$\sqrt{\frac{\sigma_\eta^2}{2\pi m}}\exp\left(-\frac{m\gamma^2}{2\sigma_\eta^2}\right) = \int_\gamma^{+\infty} yp(y)dy \le 2\int_\gamma^{+\infty}\left(y - \frac{1}{2}\gamma\right)p(y)dy \le 2\int_\gamma^{+\infty} yp(y)dy.$$

Combining them together, we have (5.10). For a given $\gamma > 0$, $\operatorname{erf}\left(\sqrt{\frac{\gamma^2 m}{2\sigma_\eta^2}}\right) \to 1$ as $m \to \infty$. By (5.10), we have

$$\mathbb{E}F(\hat{x}_{nm}) - F(x^*) = \mathcal{O}\left(\frac{1}{m}\right).$$

When $\gamma \to 0$, (C.1) becomes

$$\begin{aligned}\lim_{\gamma\to 0}\mathbb{E}F(\hat{x}_m) - F(x^*) &= \lim_{\gamma\to 0}\int_0^\gamma \frac{1}{\gamma}y^2 p(y)dy + 2\int_\gamma^{+\infty}\left(y - \frac{1}{2}\gamma\right)p(y)dy + \frac{\sigma_\eta^2}{m} \\ &= \sqrt{\frac{\sigma_\eta^2}{2\pi m}} + \frac{\sigma_\eta^2}{m} = \mathcal{O}\left(\frac{1}{\sqrt{m}}\right).\end{aligned}$$

Appendix D. Empirical objectives satisfying quadratic growth condition.

Strongly convex function composed with linear function. The empirical objective function is $\hat{F}_{nm}(x) = \frac{1}{n}\sum_{i=1}^{n} f_{\xi_i}(A_i x)$, where $f_\xi(\cdot)$ is μ-strongly convex, $A_i x := \frac{1}{m}\sum_{j=1}^{m} g_{\eta_{ij}}(x, \xi_i)$, and the average of linear inner function $g_{\eta_{ij}}(x, \xi_i) := A_{\eta_{ij}} x$. To show that $\hat{F}_{nm}(x)$ satisfies the QG condition, denote $u_i = A_i y$, $v_i = A_i x$. Since $f_\xi(\cdot)$ is strongly convex,

$$f_{\xi_i}(u_i) - f_{\xi_i}(v_i) - \nabla f_{\xi_i}(v_i)^\top (u_i - v_i) \geq \frac{\mu}{2}||u_i - v_i||_2^2.$$

Taking the average over n such inequalities, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} f_{\xi_i}(u_i) - f_{\xi_i}(v_i) - \nabla f_{\xi_i}(v_i)^\top (u_i - v_i) \geq \frac{1}{n}\sum_{i=1}^{n}\frac{\mu}{2}||u_i - v_i||_2^2.$$

Replacing u_i and v_i with $A_i y$ and $A_i x$, we have

$$\frac{1}{n}\sum_{i=1}^{n} f_{\xi_i}(A_i y) - f_{\xi_i}(A_i x) - \nabla f_{\xi_i}(A_i x)^\top A_i (y - x) \geq \frac{1}{n}\sum_{i=1}^{n}\frac{\mu}{2}(y-x)^\top A_i^\top A_i (y - x).$$

Since $\nabla \hat{F}_{nm}(x)^\top = \frac{1}{n}\sum_{i=1}^{n}(A_i^\top \nabla f_{\xi_i}(A_i x))^\top = \frac{1}{n}\sum_{i=1}^{n}\nabla f_{\xi_i}(A_i x)^\top A_i$, we get

$$\hat{F}_{nm}(y) - \hat{F}_{nm}(x) - \nabla \hat{F}_{nm}(x)^\top (y - x) \geq \frac{1}{n}\sum_{i=1}^{n}\frac{\mu}{2}||A_i(y-x)||_2^2 \geq \frac{\mu}{2}||\frac{1}{n}\sum_{i=1}^{n} A_i(y-x)||_2^2.$$

Let z be a point in $\mathcal{X}^*$; we have

$$\begin{aligned}\hat{F}_{nm}(x) - \hat{F}_{nm}(z) &\geq \frac{\mu}{2}||\frac{1}{n}\sum_{i=1}^{n} A_i(x-z)||_2^2 \geq \frac{\mu\theta(\frac{1}{n}\sum_{i=1}^{n} A_i)}{2}||x - z||_2^2 \\ &\geq \min_{z\in\mathcal{X}^*}\frac{\mu\theta(\frac{1}{n}\sum_{i=1}^{n} A_i)}{2}||x - z||_2^2.\end{aligned} \tag{D.1}$$

Here $\theta(A)$ is the smallest nonzero singular of A. Thus $\hat{F}_{nm}(x)$ satisfies the quadratic growth (QG) condition for any n and m. A special case is when $n = m = 1$, i.e., a strongly convex objective composed with a linear function satisfies the QG condition.

Some strictly convex functions composed with linear function on a compact set. Consider Example 2, the logistic regression problem with the objective

$$F(x) = \mathbb{E}_{\xi=(a,b)}\log(1 + \exp(-b\mathbb{E}_{\eta|\xi}[\eta]^T x)),$$

where $a \in \mathbb{R}^d$ is a random feature vector and $b \in \{1, -1\}$ is the label, and $\eta = a + \mathcal{N}(0, \sigma^2 I_d)$ is a perturbed noisy observation of the input feature vector a. Its empirical objective function $\hat{F}_{nm}(x)$ is given by

$$\hat{F}_{nm}(x) = \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + \exp\left(-b_i\frac{1}{m}\sum_{j=1}^{m}\eta_{ij}^\top x\right)\right),$$

where $\mathbb{E}\eta_{ij} = a_i$. Here $f_{\xi_i}(u) = \log\big(1 + \exp(b_i u)\big)$. $\hat{F}_{nm}(x) = 1/n\sum_{i=1}^{n} f(u_i)$, where $f(u) = \log\big(1 + \exp(u)\big)$ is strictly convex, and $u_i = \frac{1}{m}\sum_{j=1}^{m}\eta_{ij}^\top x$ is bounded for any

$x \in \mathcal{X}$ and realization η_{ij}. It is easy to verify that on any compact set, $f(u)$ is strongly convex. The strong convexity parameter is related to the compact set. With (D.1), $\hat{F}_{nm}(x)$ satisfies the QG condition.

Note that the result is not necessarily true for all strictly convex functions. For instance, $||x||_2^4$ is strictly convex, but $||Ax||_2^4$ does not satisfy the QG condition on any compact set containing $x = 0$.

Appendix E. Other results on regularized SAA. Theorem 4.2 discusses the sample complexity of SAA for strongly convex and QG condition cases. We show that the result obtained in Theorem 4.2 can be used to obtain dimension-free sample complexity for general convex objective by adding l_2-regularization.

LEMMA E.1 ([36]). *Consider a stochastic convex optimization problem,*

$$\min_{x\in\mathcal{X}} G(x),$$

where $G(x)$ is the expectation over some convex random function. Suppose that the decision set $\mathcal{X} \in \mathbb{R}^d$ has bounded diameter $D_{\mathcal{X}}$. Denote $G_\mu(x) := G(x) + \frac{\mu}{2}||x||_2^2$, where $\mu > 0$ is a strongly convex parameter. Denote $\hat{G}(x)$ as the SAA counterpart of $G(x)$, $x^* \in \operatorname{argmin}_{x\in\mathcal{X}} G(x)$, $\hat{x} \in \operatorname{argmin}_{x\in\mathcal{X}} \hat{G}(x)$, $x_\mu^* = \operatorname{argmin}_{x\in\mathcal{X}} G_\mu(x)$, and $\hat{x}_u$ the minimizer of SAA of the regularized objective, namely $\hat{x}_u = \operatorname{argmin}_{x\in\mathcal{X}} \hat{G}_\mu(x) := \hat{G}(x) + \frac{\mu}{2}||x||_2^2$. If $\mathbb{E}[G_\mu(\hat{x}_\mu) - G_\mu(x_\mu^*)] \le \beta(\mu)$, then

$$\mathbb{E}[G(\hat{x}_\mu) - G(x^*)] \le \beta(\mu) + \frac{\mu}{2}D_{\mathcal{X}}^2.$$

Remark E.1. This theorem shows that the minimum point $\hat{x}_\mu$ to a l_2-regularized empirical function $\hat{G}_\mu$ could be a good solution to the original convex function $G(x)$ as long as one selects μ properly. Note that $\hat{x}_\mu$ might not be a minimum point of the empirical function $\hat{G}(x)$. In the CSO case, according to Theorem 4.2, if $F(x)$ is convex, the expected error of the SAA method for $\min_{x\in\mathcal{X}} F(x) + \frac{\mu}{2}||x||_2^2$ is bounded by $\beta(\mu) = \frac{4L_f^2L_g^2}{\mu n} + 2\Delta(m)$. Then, $\mathbb{E}F(\hat{x}_{nm}) - F(x^*) \le \frac{4L_f^2L_g^2}{\mu n} + \frac{\mu}{2}D_{\mathcal{X}}^2 + 2\Delta(m)$. Minimizing over μ, and by the Markov inequality, we obtain

$$\mathbb{P}(F(\hat{x}_{nm}) - F(x^*) \ge \epsilon) \le \frac{2\sqrt{2}L_fL_gD_{\mathcal{X}}}{\sqrt{n\epsilon^2}} + \frac{2\Delta(m)}{\epsilon}.$$

We notice that the outer sample size, $n = \mathcal{O}(1/\epsilon^2)$, is dimension-free, while in Theorem 4.1, $n = \mathcal{O}(d/\epsilon^2)$ depends linearly on dimension; the inner sample size m is not affected. For high-dimensional problems, adding regularization is sometimes more favorable as it lowers the sample complexity by d and also helps boost the convergence when solving the SAA.

Acknowledgments. The authors would like to acknowledge Alexander Shapiro and Lin Xiao for fruitful discussions and the reviewers for their helpful comments. The authors would also like to thank the associate editor and reviewers for their valuable feedback and suggestions.

REFERENCES

[1] D. P. BERTSEKAS, *Dynamic Programming and Optimal Control, Vol.* I, 4th ed., Athena Scientific, Nashua, NH, 2000, http://www.athenasc.com/dpbook.html.

[2] D. Bertsimas, V. Gupta, and N. Kallus, *Robust sample average approximation*, Math. Program., 171 (2017), pp. 217–282, https://doi.org/10.1007/s10107-017-1174-z.

[3] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, *Enhancing robustness of machine learning systems via data transformations*, in Proceedings of the 2018 52nd Annual Conference on Information Sciences and Systems (CISS), IEEE, Washington, DC, 2018, pp. 1–5, https://doi.org/10.1109/ciss.2018.8362326.

[4] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Math. Program., 165 (2017), pp. 471–507, https://doi.org/10.1007/s10107-016-1091-6.

[5] L. Bottou, F. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311, https://doi.org/10.1137/16m1080173.

[6] Z. Charles and D. Papailiopoulos, *Stability and generalization of learning algorithms that converge to global optima*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, 2018, pp. 745–754, http://proceedings.mlr.press/v80/charles18a.html (accessed 2019-07-16).

[7] B. Dai, N. He, Y. Pan, B. Boots, and L. Song, *Learning from conditional distributions via dual embeddings*, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 54, PMLR, 2017, pp. 1458–1467, http://proceedings.mlr.press/v54/dai17a.html (accessed 2019-07-05).

[8] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song, *SBEED: Convergent reinforcement learning with nonlinear function approximation*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, 2018, pp. 1125–1134, http://proceedings.mlr.press/v80/dai18c.html (accessed 2019-07-15).

[9] D. Dentcheva, S. Penev, and A. Ruszczyński, *Statistical estimation of composite risk functionals and risk optimization problems*, Ann. Inst. Stat. Math., 69 (2016), pp. 737–760, https://doi.org/10.1007/s10463-016-0559-8.

[10] S. Diamond and S. Boyd, *CVXPY: A Python-embedded modeling language for convex optimization*, J. Mach. Learn. Res., 17 (2016), pp. 1–5, https://web.stanford.edu/∼boyd/papers/pdf/cvxpy_paper.pdf (accessed 2019-07-16).

[11] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Math. Oper. Res., 43 (2018), pp. 919–948, https://doi.org/10.1287/moor.2017.0889.

[12] Y. M. Ermoliev and V. I. Norkin, *Sample average approximation method for compound stochastic optimization problems*, SIAM J. Optim., 23 (2013), pp. 2231–2263, https://doi.org/10.1137/120863277.

[13] S. Ghadimi, A. Ruszczyński, and M. Wang, *A Single Time-Scale Stochastic Approximation Method for Nested Stochastic Optimization*, preprint, https://arxiv.org/abs/1812.01094, 2018.

[14] P. Gong and J. Ye, *Linear Convergence of Variance-Reduced Projected Stochastic Gradient without Strong Convexity*, preprint, https://arxiv.org/abs/1406.1102, 2014.

[15] L. J. Hong and S. Juneja, *Estimating the mean of a non-linear function of conditional expectation*, in Proceedings of the 2009 Winter Simulation Conference (WSC), IEEE, Washington, DC, 2009, https://doi.org/10.1109/wsc.2009.5429428.

[16] L. J. Hong, S. Juneja, and G. Liu, *Kernel smoothing for nested estimation with application to portfolio risk measurement*, Oper. Res., 65 (2017), pp. 657–673, https://doi.org/10.1287/opre.2017.1591.

[17] Z. Huo, B. Gu, J. Liu, and H. Huang, *Accelerated method for stochastic composition optimization with nonsmooth regularization*, in 32nd AAAI Conference on Artificial Intelligence, 2018, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/download/17203/16700 (accessed 2019-07-16).

[18] M. Jaskowski and S. Jaroszewicz, *Uplift modeling for clinical trial data*, in ICML Workshop on Clinical Data Analysis, 2012, http://people.cs.pitt.edu/∼milos/icml_clinicaldata_2012/Papers/Oral_Jaroszewitz_ICML_Clinical_2012.pdf (accessed 2019-07-15).

[19] H. Karimi, J. Nutini, and M. Schmidt, *Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition*, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, New York, 2016, pp. 795–811, https://doi.org/10.1007/978-3-319-46128-1_50.

[20] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello, *The sample average approximation method for stochastic discrete optimization*, SIAM J. Optim., 12 (2002), pp. 479–502, https://doi.org/10.1137/s1052623499363220.

[21] E. Kubińska, *Approximation of Carathéodory functions and multifunctions*, Real Anal. Exchange, 30 (2005), pp. 351–359, https://doi.org/10.14321/realanalexch.30.1.0351.

[22] X. Lian, M. Wang, and J. Liu, *Finite-sum composition optimization via variance reduced gradient descent*, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 54, 2017, pp. 1159–1167, http://proceedings.mlr.press/v54/lian17a.html (accessed 2019-07-16).

[23] H. Liu, X. Wang, T. Yao, R. Li, and Y. Ye, *Sample average approximation with sparsity-inducing penalty for high-dimensional stochastic programming*, Math. Program., 178 (2018), pp. 69–108, https://doi.org/10.1007/s10107-018-1278-0.

[24] J. Liu and S. J. Wright, *Asynchronous stochastic coordinate descent: Parallelism and convergence properties*, SIAM J. Optim., 25 (2015), pp. 351–376, https://doi.org/10.1137/140961134.

[25] P. Massart and É. Nédélec, *Risk bounds for statistical learning*, Ann. Statist., 34 (2006), pp. 2326–2366, https://doi.org/10.1214/009053606000000786.

[26] C. McDiarmid, *On the method of bounded differences*, in Surveys in Combinatorics 1989, Cambridge University Press, Cambridge, UK, 1989, pp. 148–188, https://doi.org/10.1017/cbo9781107359949.008.

[27] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj, *Dual* IV: *A Single Stage Instrumental Variable Regression*, preprint, https://arxiv.org/abs/1910.12358v1, 2019.

[28] P. Niyogi, F. Girosi, and T. Poggio, *Incorporating prior information in machine learning by creating virtual examples*, Proc. IEEE, 86 (1998), pp. 2196–2209, https://doi.org/10.1109/5.726787.

[29] V. Norkin, *Convergence of the empirical mean method in statistics and stochastic programming*, Cybernet. Syst. Anal., 28 (1992), pp. 253–264, https://doi.org/10.1007/BF01126212.

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

[31] M. V. F. Pereira and L. M. V. G. Pinto, *Multi-stage stochastic optimization applied to energy planning*, Math. Program., 52 (1991), pp. 359–375, https://doi.org/10.1007/bf01582895.

[32] B. Polyak, *Minimization of composite regression functions*, Cybernetics, 14 (1978), pp. 642–644, https://doi.org/10.1007/BF01069853.

[33] R. T. Rockafellar and R. J.-B. Wets, *Scenarios and policy aggregation in optimization under uncertainty*, Math. Oper. Res., 16 (1991), pp. 119–147, https://doi.org/10.1287/moor.16.1.119.

[34] A. Ruszczyński, *Decomposition methods in stochastic programming*, Math. Program., 79 (1997), pp. 333–353, https://doi.org/10.1007/bf02614323.

[35] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge, UK, 2014, https://doi.org/10.1017/cbo9781107298019.

[36] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Learnability, stability and uniform convergence*, J. Mach. Learn. Res., 11 (2010), pp. 2635–2670, https://doi.org/10.1007/978-3-642-34106-9_3.

[37] A. Shapiro, *On complexity of multistage stochastic programs*, Oper. Res. Lett., 34 (2006), pp. 1–8, https://doi.org/10.1016/j.orl.2005.02.003.

[38] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, MOS–SIAM Ser. Optim. 9, SIAM, Philadelphia, 2009, https://doi.org/10.1137/1.9780898718751.

[39] A. Shapiro and A. Nemirovski, *On complexity of stochastic programming problems*, in Continuous Optimization, Springer-Verlag, New York, 2005, pp. 111–146, https://doi.org/10.1007/0-387-26771-9_4.

[40] S. Shen, L. Xu, J. Liu, J. Guo, and Q. Ling, *Asynchronous Stochastic Composition Optimization with Variance Reduction*, preprint, https://arxiv.org/abs/1811.06396, 2018.

[41] R. S. Sutton, H. R. Maei, and C. Szepesvári, *A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation*, in Advances in Neural Information Processing Systems 21, Curran Associates, 2009, pp. 1609–1616, http://papers.nips.cc/paper/3626-a-convergent-on-temporal-difference-algorithm-for-off-policy-learning-with-linear -function-approximation.pdf.

[42] M. Wang, E. X. Fang, and H. Liu, *Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions*, Math. Program., 161 (2017), pp. 419–449, https://doi.org/10.1007/s10107-016-1017-3.

[43] M. Wang, J. Liu, and E. X. Fang, *Accelerating stochastic composition optimization*, J. Mach. Learn. Res., 18 (2017), pp. 1–23, http://jmlr.org/papers/v18/16-504.html.

[44] Y. Xu, Q. Lin, and T. Yang, *Accelerate Stochastic Subgradient Method by Leveraging Local Error Bound*, preprint, https://arxiv.org/abs/1607.01027v1, 2016.

[45] I. Yamane, F. Yger, J. Atif, and M. Sugiyama, *Uplift modeling from separate labels*, in Advances in Neural Information Processing Systems 31, Curran Associates, 2018, pp. 9927–9937, http://papers.nips.cc/paper/8198-uplift-modeling-from-separate-labels.pdf.

[46] Y. M. Yermol'yev, *A general stochastic programming problem*, J. Cybernet., 1 (1971), pp. 106–112, https://doi.org/10.1080/01969727108542906.