New Sample Complexity Bounds for Sample Average Approximation in Heavy-Tailed Stochastic Programming

Hongcheng Liu [1] **Jindong Tong** [2]

Abstract

This paper studies sample average approximation (SAA) and its simple regularized variation in solving convex or strongly convex stochastic programming problems. Under heavy-tailed assumptions and comparable regularity conditions as in the typical SAA literature, we show — perhaps for the first time — that the sample complexity can be completely free from any complexity measure (e.g., logarithm of the covering number) of the feasible region. As a result, our new bounds can be more advantageous than the state-of-the-art in terms of the dependence on the problem dimensionality.

1. Introduction.

This paper is focused on a convex or strongly convex stochastic programming (SP) problem of the following form:

$$\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}) := \mathbb{E}[f(\mathbf{x},\xi)], \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a non-empty convex feasible region with integer d being the number of decision variables (a.k.a., dimensionality), ξ is a random vector of problem parameters whose probability distribution $\mathbb{P}$ is supported on $\Theta \subseteq \mathbb{R}^m$, and the cost function $f : \mathcal{X} \times \Theta \to \mathbb{R}$ is deterministic and measurable. Throughout this paper, we assume that $f(\,\cdot\,,\xi)$ is everywhere differentiable for almost every $\xi \in \Theta$, the expectation $\mathbb{E}\left[f(\mathbf{x},\xi)\right] = \int_\Theta f(\mathbf{x},\xi)\,\mathrm{d}\mathbb{P}(\xi)$ is well defined for every $\mathbf{x} \in \mathcal{X}$, and F admits a finite minimizer $\mathbf{x}^*$ on $\mathcal{X}$ with a finite optimal cost. Furthermore, we also assume the presence of some structure of a composite objective function; that is, there exist two deterministic and everywhere differentiable functions, denoted by $F_1 : \mathcal{X} \to \mathbb{R}$ and $F_2 : \mathcal{X} \to \mathbb{R}$ such that

$$F(\mathbf{x}) = F_1(\mathbf{x}) + F_2(\mathbf{x}), \tag{2}$$

where F_1 and F_2 satisfy regularities in Assumption 1.1 below:

Assumption 1.1. Given $q \geq 1$, let $\varrho = q/(q-1)$. For some scalars $\mathcal{L} \geq 0$, $\mathcal{M} \geq 0$, and any pair of vectors $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^2$,

$$\|\nabla F_1(\mathbf{x}_1) - \nabla F_1(\mathbf{x}_2)\|_\varrho \leq \mathcal{L} \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|_q, \tag{3}$$

and, simultaneously,

$$\|\nabla F_2(\mathbf{x})\|_\varrho \leq \mathcal{M}, \;\; \forall \mathbf{x} \in \mathcal{X}. \tag{4}$$

Here, (3) means that the first component of the population-level objective function F_1 admits Lipschitz continuous gradient. Meanwhile, (4) essentially imposes that the second component of the population-level objective function F_2 is Lipschitz continuous. Results that apply to such a composite objective function subsume the special cases of F being smooth (with $F_2 = 0$) and F being Lipschitz (with $F_1 = 0$). Conditions closely similar to, if not more critical than, Assumption 1.1 have been considered in much SP literature (such as Ghadimi & Lan, 2012; 2013; Nemirovski et al., 2009; Rakhlin et al., 2011; Lan, 2020).

The SP problem above has been widely applied and much discussed (e.g., by Shapiro et al., 2021; Birge, 1997; Birge & Louveaux, 2011; Ruszczyński & Shapiro, 2003; Lan, 2020, to name only a few). Particularly, it has extensive connections with many machine learning problems (as per, e.g., Bartlett et al., 2006; Liu et al., 2019). Indeed, the suboptimality gap in solving (1) can often be interpreted as the excess risk, an important metric of generalizability, when the SP problem is constructed for fitting/training a statistical or machine learning model. Due to the (increasingly) frequent need to perform data-driven modeling or decision-making in the presence of extreme values or outliers in data, studying solution techniques for (1) under heavy-tailed-ness becomes growingly more important (Oliveira & Thompson, 2023).

[1]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611 USA. [2]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611 USA. Email: jindongtong@ufl.edu. Correspondence to: Hongcheng Liu <liu.h@ufl.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

This paper revisits one of the most traditional but popular solution methods for the SP called the *sample average approximation* (SAA). Following the SAA literature (Dupacová & Wets, 1988; Ruszczyński & Shapiro, 2003; Kleywegt et al., 2002; Shapiro et al., 2021; Oliveira & Thompson, 2023; King & Wets, 1991, among many others), we particularly focus on the canonical formulation of the SAA and one of its simple, regularized variations — both in heavy-tailed settings:

(i) In particular, the canonical SAA is as below:

$$\min_{\mathbf{x}\in\mathcal{X}} F_N(\mathbf{x}) := N^{-1}\sum_{j=1}^{N} f(\mathbf{x},\xi_j), \tag{5}$$

where $\boldsymbol{\xi}_{1,N} := (\xi_j : j = 1, ..., N)$ is an i.i.d. random sample of ξ. Our analysis on this formulation is centered around its effectiveness for strongly convex SP problems.

(ii) On top of (5), we also consider the SAA that incorporates a Tikhonov-like regularization (referred to as the RSAA) in the following:

$$\min_{\mathbf{x}\in\mathcal{X}} F_{\lambda_0,N}(\mathbf{x}) := F_N(\mathbf{x}) + \lambda_0 V_{q'}(\mathbf{x}), \tag{6}$$

where $\lambda_0 \geq 0$ is a tuning parameter, and $V_{q'} : \mathcal{X} \to \mathbb{R}_+$ for a user's choice of q'-norm (with $q' \in (1, 2]$), is defined as

$$V_{q'}(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^0\|_{q'}^2, \tag{7}$$

for any initial guess $\mathbf{x}^0 \in \mathbb{R}^d$ that does not have to be feasible to $\mathcal{X}$. Particularly in the case of $q' = 2$ and $\mathbf{x}^0 = \mathbf{0}$, we have $V_{q'}(\mathbf{x}) = 0.5\|\mathbf{x}\|_2^2$, which becomes the canonical Tikhonov regularization (Golub et al., 1999) commonly employed in ridge regression (Hoerl & Kennard, 1970). The same type of regularziation approach has been discussed in (R)SAA theories for (general) convex SP, among others, by Hu et al. (2020), Feldman & Vondrak (2019), and Shalev-Shwartz et al. (2010; 2009) under Lipschitz continuity and by Lei & Ying (2020) under gradient dominance. Similarly in this paper, we also study the RSAA in (general) convex SP problems.

Both (5) and (6) avoid the multi-dimensional integral involved in (1) and thus render the SP problem to be solvable as a "deterministic" nonlinear program, often improving the tractability substantially (Shapiro et al., 2021).

Hereafter, consistent with the literature (e.g., Shapiro, 1993), we refer to the random variable $\widehat{\mathbf{x}} := \widetilde{\mathbf{x}}(\boldsymbol{\xi}_{1,N})$ for a deterministic and measurable function $\widetilde{\mathbf{x}} : \Theta^N \to \mathcal{X}$ such that $\widetilde{\mathbf{x}}(\boldsymbol{\xi}_{1,N}) \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_N(\mathbf{x})$ (or $\widetilde{\mathbf{x}}(\boldsymbol{\xi}_{1,N}) \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_{\lambda_0,N}(\mathbf{x})$) as an optimal solution to (5) (or (6), resp). Sufficient conditions for the measurability of $\widetilde{\mathbf{x}}$ have been established in different scenarios (Shapiro et al., 2021; Rockafellar & Wets, 2009; Krätschmer, 2023). Particularly of our interest is the quality of solution $\widehat{\mathbf{x}}$ in terms of its sample complexity; how large the (finite) sample size N should be in order to ensure that $\widehat{\mathbf{x}}$ is within the set of ϵ-suboptimal solutions to (1) with probability at least $1 - \beta$, for a user-specified accuracy threshold $\epsilon > 0$ and a given significance level $\beta \in (0, 1)$.

While much literature has studied the effectiveness of SAA and its regularized variations (e.g., Artstein & Wets, 1995; Dupacová & Wets, 1988; King & Rockafellar, 1993; King & Wets, 1991; Pflug, 1995; 1999; 2003; Shapiro, 1989; 1993; 2003; Shapiro et al., 2021; Guigues et al., 2017; Liu et al., 2016; 2022), most existing finite-sample (non-asymptotic) results assume light-tailed-ness for the underlying randomness; that is, its pth moments are finite for all $p \geq 1$. From this body of literature, a typical non-asymptotic result is in the form below:

A typical result under light-tailed-ness (e.g., as per Shapiro et al., 2021): *Given $q \geq 1$, under the Lipschitz assumption that, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and every $\xi \in \Theta$,*

$$|f(\mathbf{x},\xi) - f(\mathbf{y},\xi)| \leq M(\xi)\cdot\|\mathbf{x}-\mathbf{y}\|_q, \tag{8}$$

where $M : \Theta \to \mathbb{R}_+$ is some deterministic and measurable function, the optimal solution to the SAA in (5), denoted by $\widehat{\mathbf{x}}$, admits the following sample complexity: For any $\epsilon > 0,\ \beta \in (0, 1)$:

$$\text{Prob}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \leq \epsilon\right] \geq 1 - \beta,$$
$$\text{if } N \geq O\left(\max\left\{\frac{\upsilon_f^2(\Gamma_\epsilon(\mathcal{X}) + \ln\frac{1}{\beta})}{\epsilon^2}, \upsilon_L \ln\frac{1}{\beta}\right\}\right), \tag{9}$$

where υ_f and υ_L are the parameters of the sub-Gaussian (or subexponential) distributions assumed for random variables $Y_{\mathbf{x},\mathbf{x}'} := [f(\mathbf{x},\xi) - F(\mathbf{x})] - [f(\mathbf{x}',\xi) - F(\mathbf{x}')]$, for any given solution pair $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2$, and $M(\xi)$, respectively, and $\Gamma_\epsilon(\mathcal{X})$ is a complexity measure of the feasible region. Here, sub-Gaussian and subexponential refer to distributions whose tails vanish no slower than Gaussian or exponential distributions, respectively.

Note that a frequent choice of the complexity measure $\Gamma_\epsilon(\mathcal{X})$ of the feasible region is the logarithm of the covering number (or cardinality of the ϵ-net) for $\mathcal{X}$. This complexity measure grows polynomially with d in general. We also note that the Lipschitz condition in (8) is comparable to, if not more stringent than, Assumption 1.1.

Beyond the light-tailed assumptions, non-asymptotic sample complexity of (R)SAA (in either (5) or (6)) is much less

visited, especially when the consideration is under regularity conditions comparable to (8) or Assumption 1.1. Among the few, a very recent state-of-the-art result shows the below:

State-of-the-art result under heavy-tailed-ness (Oliveira & Thompson, 2023): *Suppose that, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}^{*,\epsilon}$ (instead of $\mathcal{X}$ in (8)) and every $\xi \in \Theta$, it holds, for some given $q \geq 1$, that*

$$|f(\mathbf{x}, \xi) - f(\mathbf{y}, \xi)| \leq M(\xi) \cdot \|\mathbf{x} - \mathbf{y}\|_q, \tag{10}$$

then for given $p' \geq 2$, $\epsilon > 0$, and $\beta \in (0, 1)$,

$$\text{Prob}[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \leq \epsilon] \geq 1 - \beta,$$
$$\text{if } N \geq O\left(\frac{\mathbf{M}_2 \cdot \left(\left(\gamma(\mathcal{X}^{*,\epsilon})\right)^2 + (\mathcal{D}^{*,\epsilon})^2 \cdot \ln \frac{1}{\beta}\right)}{\epsilon^2} + p' \cdot \left(\frac{\widetilde{\mathbf{M}}_{p'}}{\mathbf{M}_2} + \frac{\widetilde{\upsilon}_{\mathbf{x}^*,p'}}{\upsilon_{\mathbf{x}^*}}\right) \cdot \beta^{-2/p'}\right). \tag{11}$$

where $\mathcal{X}^{*,\epsilon}$ is the set of ϵ-suboptimal solutions, $\gamma(\mathcal{X}^{*,\epsilon})$ is some complexity measure of the feasible region based on generic chaining, $\mathcal{D}^{*,\epsilon}$ denotes the diameter of $\mathcal{X}^{*,\epsilon}$, $\mathbf{M}_2$ is the second moment of $M(\xi)$ in (10), $\upsilon_{\mathbf{x}^*}$ is the variance of $f(\mathbf{x}^*, \xi)$, and $\widetilde{\mathbf{M}}_{p'}$ and $\widetilde{\upsilon}_{\mathbf{x}^*,p'}$ are the p'th central moments of $[M(\xi)]^2$ and $(f(\mathbf{x}^*, \xi) - F(\mathbf{x}^*))^2$, respectively.

Closest to our settings of consideration is (11) in its "most heavy-tailed" scenario with $p' = 2$; namely, when the second central moment of $[M(\xi)]^2$ is finite — and thus the fourth moment of $M(\xi)$ is also finite. In such a case, the benchmark sample complexity bound in (11) is reduced to

$$\text{Prob}[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \leq \epsilon] \geq 1 - \beta,$$
$$\text{if } N \geq O\left(\frac{\mathbf{M}_2 \cdot \left(\left(\gamma(\mathcal{X}^{*,\epsilon})\right)^2 + (\mathcal{D}^{*,\epsilon})^2 \cdot \ln \frac{1}{\beta}\right)}{\epsilon^2} + \left(\frac{\widetilde{\mathbf{M}}_2}{\mathbf{M}_2} + \frac{\widetilde{\upsilon}_{\mathbf{x}^*,2}}{\upsilon_{\mathbf{x}^*}}\right) \cdot \frac{1}{\beta}\right). \tag{12}$$

The value of the complexity measure of the feasible region $\gamma(\mathcal{X}^{*,\epsilon})$ above can sometimes be opaque and hard to estimate. It is known that $\gamma(\mathcal{X}^{*,\epsilon}) \leq O(\sqrt{\ln d})$ when the feasible region is a simplex. Nonetheless, the best-known upper bound is $\gamma(\mathcal{X}^{*,\epsilon}) \leq O(\sqrt{d} \cdot \mathcal{D}^{*,\epsilon})$ in general. Furthermore, it is also worth noting that $\mathcal{D}^{*,\epsilon}$ is comparable to the diameter of the feasible region $\mathcal{X}$ in general. Indeed, particularly for general convex SP problems, it is easy to construct scenarios where the largest distance between any two ϵ-suboptimal solutions can be closely similar to the largest possible distance between any two feasible solutions. One such example is for the expected objective function to be close to a constant. Likewise, in many cases, $M(\xi)$ has to be large enough such that (10) holds for all $\mathbf{x}$, $\mathbf{y} \in \mathcal{X}$, making (10) not necessarily a significantly weaker condition than (8). Thus, (10) is also comparable to, if not more critical than, Assumption 1.1.

In the results (9), (11), and (12) above — as well as in most literature on SAA — a seemingly unavoidable term in the sample complexity is $\Gamma_\epsilon(\mathcal{X})$, $\gamma(\mathcal{X}^{*,\epsilon})$, or alike — the complexity measure of the feasible region — which often leads to a high dependence on problem dimensionality d. Furthermore, their dependence on the problem quantities tend to be opaque; other than some generic and conservative upper bounds, how large the $\mathcal{X}$-specific values are for the complexity measures of $\mathcal{X}$ can be hard to assess in many cases. In response to the observations above, this paper is focused on the following research question:

Question: *Does the SAA, or its simple variations, admit complexity bounds that are completely free from any complexity measure of the feasible region (even if the underlying randomness is heavy-tailed)?*

To the question above, the literature has provided positive results in both strongly convex and (general) convex cases under more critical regularity conditions than Assumption 1.1 or (8) through the notion of uniform stability and its variations (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010; 2009; Hu et al., 2020). (See some additional discussions on this in Section 2). In contrast, this paper advances those results and presents perhaps the first set of affirmative answers under standard assumptions (Assumption 1.1) for the SP, as summarized below in Main Theorems 1 and 2:

Main Theorem 1 (Informal statement of Theorem 4.8): *Let $f(\cdot, \xi)$ be μ-strongly convex ($\mu > 0$) w.r.t. the q-norm ($q \geq 1$) for almost every $\xi \in \Theta$, and F be a composite function satisfying Assumption 1.1 also w.r.t. the q-norm. Suppose that the variance of the gradient $\nabla f(\cdot, \xi)$ is bounded as per*

$$\mathbb{E}[\|\nabla F(\mathbf{x}) - \nabla f(\mathbf{x}, \xi)\|_p^2] \leq \sigma_p^2\,, \ \ \forall \mathbf{x} \in \mathcal{X}, \tag{13}$$

for some $\sigma_p \geq 1$ and some $p \in \left[1, \frac{q}{q-1}\right]$. Consider an optimal solution $\widehat{\mathbf{x}}$ to the SAA formulation in (5). Both of the two inequalities below explicate the SAA's sample complexity: For any $\epsilon > 0$,

$$\mathbb{E}[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*)] \leq \epsilon,$$
$$\text{if } N \geq O\left(\max\left\{\frac{\mathcal{L}}{\mu}, \frac{\sigma_p^2 + \mathcal{M}^2}{\mu \cdot \epsilon}\right\}\right); \tag{14}$$

and, meanwhile, for any $\epsilon > 0$ *and* $\beta \in (0,1)$

$$\text{Prob}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \le \epsilon\right] \ge 1-\beta, \quad \text{if } N \ge O\left(\max\left\{\frac{\mathcal{L}}{\mu}, \frac{\sigma_p^2+\mathcal{M}^2}{\mu\cdot\epsilon\cdot\beta}\right\}\right). \quad (15)$$

Main Theorem 2 (Informal statement of Theorem 4.11): *Let* $f(\cdot,\xi)$ *be (general) convex for almost every* $\xi \in \Theta$*. Suppose that Assumption 1.1 holds w.r.t. the q-norm (*$q > 1$*). For some tractable choices of the hyper-parameters, such as* $q' \in (1,2]:\ q' \le q$ *in* $V_{q'}$*, if (13) holds with* $p \in \left[1, \frac{q}{q-1}\right]$*, any solution* $\widehat{\mathbf{x}}$ *to the RSAA in (6) entails the following sample complexity: For any* $\epsilon \in (0,1]$*,*

$$\mathbb{E}[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*)] \le \epsilon, \quad \text{if } N \ge O\left(\frac{V_{q'}(\mathbf{x}^*)}{q'-1}\cdot\max\left\{\frac{\mathcal{L}}{\epsilon}, \frac{\sigma_p^2+\mathcal{M}^2}{\epsilon^2}\right\}\right); \quad (16)$$

and, meanwhile, for any $\epsilon \in (0,1]$ *and* $\beta \in (0,1)$*,*

$$\text{Prob}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \le \epsilon\right] \ge 1-\beta, \quad \text{if } N \ge O\left(\frac{V_{q'}(\mathbf{x}^*)}{q'-1}\cdot\max\left\{\frac{\mathcal{L}}{\epsilon}, \frac{\sigma_p^2+\mathcal{M}^2}{\epsilon^2\beta}\right\}\right). \quad (17)$$

Here, $V_{q'}(\mathbf{x}^*)$ *is half of the squared* q'*-norm distance between an optimal solution* $\mathbf{x}^*$ *and a specified vector* $\mathbf{x}^0$ *as hyper-parameters of* $V_{q'}$ *(defined in (7)).*

Note that the assumption on the underlying randomness as in (13) is a standard condition in much SP literature (e.g., Ghadimi & Lan, 2012; 2013; Lan, 2020). Nonetheless, non-asymptotic analysis of SAA under this condition has been scarcely visited, to our knowledge. Indeed, (13) allows for the consideration of some even "heavier-tailed" scenarios than the benchmark results as in (9) and (11) (or (12)). See Remark 4.2 later for more discussions.

Our new sample complexity rates summarized above can also be more advantageous than the concurrent results. More specifically, in comparison with the benchmarks in (9) and (12), observing that $\mathbf{M}_2 \approx \max_{\mathbf{x}\in\mathcal{X}^{*,\epsilon}} \mathbb{E}[\|\nabla f(\mathbf{x},\xi)\|_p^2]$ is comparable to $\sigma_p^2 + \mathcal{M}^2$ in general, we may see that our results can be more appealing in the following aspects:

- First, both main theorems are independent of the complexity measures of the feasible region, such as $\Gamma_\epsilon(\mathcal{X})$ in (9) and $\gamma(\mathcal{X}^{*,\epsilon})$ in (12), which usually elevate the dependence of the sample complexity on d. By avoiding these complexity measures, our results are less dimension-sensitive than both (9) and (12), as well as than most existing results under comparable conditions. (See more detail in Remark 4.15 subsequently). While σ_p^2 may depend on d implicitly, as we clarify later in Remark 4.16, this quantity may only grow with $d^{2/p}$ when each dimension of $\nabla f(\cdot,\xi)$ has a fixed upper bound on the central moments to the pth order. Consequently, if $p > 2$, the dependence on d becomes better than any polynomial. When $p \ge c\ln d$, for some constant $c > 0$, the sample complexity becomes dimension free (when the other quantities such as $V_{q'}(\mathbf{x}^*)$, $\mathcal{L}$ and/or $\mathcal{M}$ are fixed).

- Second, our results make use of the potential smoothness of the objective function to obtain sharper bounds. E.g., for (general) convex SP, in the more adversarial case of $\mathcal{L} = 0$, our derived complexity grows linearly with $\mathcal{M}^2$, leading to comparable complexity rates between (12) and (17). Meanwhile, as $\mathcal{L}$ becomes more dominant than $\mathcal{M}$, the rates in our new complexity (17) improves and becomes potentially more efficient than the benchmark. A similar trend can also be observed for the strongly convex case in (15). Particularly, in the most desirable case of $\mathcal{M} = 0$ and ϵ is reasonably small, our sample complexities can be simplified into

$$\text{Prob}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \le \epsilon\right] \ge 1-\beta, \quad \text{if } N \ge \begin{cases} O\left(\frac{\sigma_p^2}{\mu\cdot\epsilon\cdot\beta}\right) & \mu\text{-strongly convex SP;} \\ O\left(\frac{V_{q'}(\mathbf{x}^*)}{q'-1}\frac{\sigma_p^2}{\epsilon^2\cdot\beta}\right) & \text{general convex SP,} \end{cases} \quad (18)$$

  which identify a region of parameters free from the impact of Lipschitz constants $\mathcal{L}$ and $\mathcal{M}$ (of ∇F_1 and F_2, respectively).

- Third, in the case of strongly convex SP, the sample efficiency presented in Main Theorem 1 shows a significant improvement in terms of dependence on ϵ as compared to results in (9) — the margin of enhancement is by an order of magnitude. While this advantage is made possible via better exploiting the μ-strong convexity, there is no need to estimate μ when constructing the SAA formulation. It is also worth noting that, in the same μ-strongly convex case, the benchmark in (12) can achieve a comparably advantageous rate with ϵ. To see this, one can show that μ-strong convexity leads to $(\mathcal{D}^{*,\epsilon})^2 \le 2\mu^{-1}\epsilon$. Plugging this, as well as the fact that $\gamma(\mathcal{X}^{*,\epsilon}) \le O(\sqrt{d}\cdot\mathcal{D}^{*,\epsilon})$, into (12) leads to a sample requirement of $N \ge O\left(\frac{\mathbf{M}_2(d+\ln(1/\beta))}{\mu\epsilon} + \left(\frac{\widetilde{\mathbf{M}}_{p'}}{\mathbf{M}_2} + \frac{\widetilde{v}_{\mathbf{x}^*,2}}{v_{\mathbf{x}^*}}\right)\cdot\frac{1}{\beta}\right)$, which, nonetheless, still exhibits a significantly higher dependence on d than our result in (15).

1.1. Organizations

The rest of this paper is organized as follows: Section 2 summarizes related works. Section 3 discusses preliminary results. Our main theorems (Theorems 4.8 and 4.11) are presented in Section 4. Finally, Section 5 concludes the paper.

1.2. Notations

Denote by $\mathbb{R}$ the collection of all real numbers, and by $\mathbb{R}_+$ that of the non-negative ones. $\mathbf{0}$ is the all-zero vector of some proper dimension. We at times use (x_i) or $(x_i : i = 1, ..., d)$ to denote a d-dimensional vector $\mathbf{x} = (x_1, \cdots, x_d)^\top$ for convenience. For a function g, denote by ∇g the gradient and by $\nabla_i g$ its ith element. For any vector $\mathbf{v} = (v_i : i = 1, ..., d) \in \mathbb{R}^d$, denote by $\|\cdot\|_p := \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$ the p-norm ($p \geq 1$). Meanwhile, we define the L^p-norm of a random vector $\boldsymbol{\zeta} = (\zeta_i) \in \mathbb{R}^d$ to be $\|\boldsymbol{\zeta}\|_{L^p} := \left(\sum_{i=1}^d \mathbb{E}_{\zeta_i}\left[|\zeta_i|^p\right]\right)^{1/p}$. Here, we use $|v|$ to denote the absolute value of v if it is a real number; otherwise, $|\mathcal{V}|$ is the cardinality of $\mathcal{V}$, when it is a set. $\mathbb{E}[\cdot]$ denotes the expectation over all the randomness in "$\cdot$". Finally, "w.r.t." and "a.s." are short-hands for "with respect to" and "almost surely", respectively.

2. Related work

There is a rich body of literature on (5) and (6). As a result, many (e.g., Artstein & Wets, 1995; Dupacová & Wets, 1988; King & Rockafellar, 1993; King & Wets, 1991; Pflug, 1995; 1999; 2003; Shapiro, 1989; 2003; Shapiro et al., 2021; Guigues et al., 2017) have provided theoretical guarantees on the efficacy of the (R)SAA. However, their results are either asymptotic or focused on light-tailed problems (such as summarized in (9)). In contrast, non-asymptotic sample complexities under heavy-tailed-ness and comparable Lipschitz conditions as in (8) are much less. Among the few available, the state-of-the-art — and E.q. (9)-comparable — bound is proven very recently by Oliveira & Thompson (2023) and summarized in (11) and (12). However, the sample complexity benchmarks in both light-tailed or heavy-tailed scenarios are mostly polynomial in the complexity measures of the feasible region, such as $\Gamma_\epsilon(\mathcal{X})$ and $\gamma(\mathcal{X}^{*,\epsilon})$ in (9) and (11) (or (12)), respectively. The presence of those complexity measures typically increases the SAA's predicted sample requirement w.r.t. the dependence on d.

Sample complexities free from the said complexity measures of the feasible region have actually been shown possible for to the (R)SAA under more critical conditions. Indeed, through the argument of uniform stability or its variations, it has been proven (e.g., by Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010; 2009; Hu et al., 2020) that an optimal solution $\widehat{\mathbf{x}}$ to (R)SAA satisfies the below:

$$\text{Prob}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \leq \epsilon\right] \geq 1 - \beta, \quad \text{if } N \geq \begin{cases} O(\frac{M}{\mu\cdot\epsilon\cdot\beta}) & \mu\text{-strongly convex SP;} \\ O(\frac{MV_{q'}(\mathbf{x}^*)}{\epsilon^2\cdot\beta}) & \text{general convex SP,} \end{cases} \tag{19}$$

where $V_{q'}$ is the same as in (7), when it holds that, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and $\xi \in \Theta$, and for some Lipschitz constant $M > 0$,

$$|f(\mathbf{x}, \xi) - f(\mathbf{y}, \xi)| \leq M \cdot \|\mathbf{x} - \mathbf{y}\|_{q'}. \tag{20}$$

Meanwhile, high probability bounds (that are logarithmic in $1/\beta$) are also obtained under the same Lipschitz condition as in (20) (e.g., by Feldman & Vondrak, 2018; 2019; Bousquet & Elisseeff, 2002; Feldman & Vondrak, 2018; Klochkov & Zhivotovskiy, 2021, when their results are applied to the analysis of (5) or (6)). Nonetheless, almost all these current stability-based analyses on (R)SAA seem indicate the necessity of the Lipschitz condition in (20), which can be overly critical for many applications of the SP. Indeed, because M is independent of ξ, this quantity can be undesirably large and even unbounded under the counterpart condition of (8) or Assumption 1.1 of our consideration. To see this, one may consider a simple stochastic quadratic program of $\min\{\mathbb{E}[(\alpha^\top \mathbf{x})^2] : \mathbf{x} \in [-1, 1]^d\}$, where $\alpha \in \mathbb{R}^d$ is some Gaussian random vector. While many applications can be subsumed by simple variations of this SP problem, it does not admit a finite "M" to satisfy (20). In contrast, in the most comparable (and actually more adversarial) settings of our results (i.e., when $\mathcal{L} = 0$), our theorems only require $F(\cdot) = \mathbb{E}[f(\cdot, \xi)]$ — the population-level objective function — to be Lipschitz continuous. This can sometimes be a non-trivially weaker condition relative to (20).

This current paper frequently refers to existing complexity bounds, e.g., by Shapiro et al. (2021); Shapiro (2003); Shapiro & Nemirovski (2005), and Oliveira & Thompson (2023) as benchmarks in order to explain the claimed advantages of our results. Yet, it is worth noting that those concurrent works apply to many important scenarios that are not covered by the results of this paper. For instance, the SAA theores by Shapiro et al. (2021) and Oliveira & Thompson (2023) can handle nonconvex problems. The findings by Oliveira & Thompson (2023) further admit stochasticity in the feasible region. Nonetheless, when applied to the settings of our consideration — the convex SP problems with deterministic constraints — the results by Shapiro et al. (2021); Shapiro (2003); Shapiro & Nemirovski (2005) and Oliveira & Thompson (2023) are known to be the best available benchmarks. We would like to also argue that the SP problems considered herein are still flexible enough to cover a very wide spectrum of applications, and that our proof

arguments, which seem to differentiate from most SAA literature, may be further extended to nonconvex problems and scenarios with uncertain constraints.

3. Preliminaries

In the SAA literature, two common ways to quantify the complexity of the feasible region are: (i) the logarithm of the covering number such as $\Gamma_\epsilon(\mathcal{X})$ in (9); and (ii) the "generic chaining" functional such as $\gamma(\mathcal{X}^{*,\epsilon})$ as in (11) (and in (12) as well). Particularly for (i), the covering number of $\mathcal{X}$ is the smallest number of closed balls that satisfy the two requirements below: (a). their centers are in $\mathcal{X}$ and their radii are equal to a prescribed (small) value; and (b). their union is a superset of $\mathcal{X}$. Provably tight overestimates of the covering number (as is used, e.g., by Shapiro et al., 2021) grow exponentially with the dimensionality d of the feasible set $\mathcal{X}$, causing $\Gamma_\epsilon(\mathcal{X})$, the logarithm of the covering number, to grow polynomially with d. One may refer to Vershynin (2018) for more comprehensive discussions.

The consideration of the "generic chaining" functional $\gamma(\cdot)$ in the sample complexity analysis of the SAA has been discussed by Oliveira & Thompson (2023). According to them, the definition of $\gamma(\cdot)$ involves a notion called the *admissible sequences* as discussed below. For a set $\mathcal{S}$, a sequence $\mathbb{A}_\mathcal{S} := \{\mathcal{A}_j\}_{j\geq 0}$ is said to be *admissible* if each $\mathcal{A}_j$ in this sequence is a partition of $\mathcal{S}$ and satisfies that

$$\begin{cases} |\mathcal{A}_j| = 1 & \text{if } j = 0; \\ |\mathcal{A}_j| \leq 2^{2^j} & \text{if } j \geq 1. \end{cases}$$

For each j, denote by $diam_{\max}(\mathcal{A}_j)$ the largest diameter of a set in partition $\mathcal{A}_j$. Then, it is defined that

$$\gamma(\mathcal{S}) := \inf_{\mathbb{A}_\mathcal{S}} \sum_{j\geq 0} 2^{\frac{j}{2}} diam_{\max}(\mathcal{A}_j),$$

where the infimum is taken over all admissible sequences. While the definition of $\gamma(\mathcal{S})$ is not in a closed form, it is known that, when $\mathcal{S}$ is a simplex, $\gamma(\mathcal{S}) \leq \sqrt{\ln d}$; or otherwise, $\gamma(\mathcal{S}) \leq \sqrt{d}\cdot\mathcal{D}$ in general, given a finite diameter $\mathcal{D}$ of $\mathcal{S}$. Interested readers are referred to Oliveira & Thompson (2023) and Talagrand (2014) for more detailed discussions.

Below, we discuss some useful properties of $V_{q'}$ for $1 < q' \leq 2$, as defined in (7). Note that this function is differentiable and $(q'-1)$-strongly convex w.r.t. the q'-norm, according to Ben-Tal et al. (2001). Therefore,

$$\begin{aligned} &V_{q'}(\mathbf{x}_1) - V_{q'}(\mathbf{x}_2) - \langle \nabla V_{q'}(\mathbf{x}_2),\, \mathbf{x}_1 - \mathbf{x}_2\rangle \\ &\qquad \geq \frac{q'-1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_{q'}^2, \ \ \forall\, \mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}. \end{aligned} \tag{21}$$

When $f(\cdot,\xi)$ is convex for almost every $\xi \in \Theta$, we have $F_N(\cdot)$ also being convex for almost every $\boldsymbol{\xi}_{1,N} \in \Theta^N$. This combined with the fact that $V_{q'}$ is $(q'-1)$-strongly convex w.r.t. the q'-norm leads to the $(q'-1)$-strong convexity of $F_{\lambda_0,N}(\cdot)$ w.r.t. the same norm for almost every $\boldsymbol{\xi}_{1,N} \in \Theta^N$. As an immediate result, we have that the following inequality holds for almost every $\boldsymbol{\xi}_{1,N} \in \Theta^N$:

$$F_{\lambda_0,N}(\mathbf{x}) - F_{\lambda_0,N}(\widehat{\mathbf{x}}) \geq \frac{q'-1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}\|_{q'}^2, \ \ \forall \mathbf{x} \in \mathcal{X},$$

where we recall that $\widehat{\mathbf{x}}$ is the minimizer of $F_{\lambda_0,N}$ on $\mathcal{X}$.

Another important property of $V_{q'}$ is its Lipschitz continuity under mild conditions. To see this, one may observe that, for any $\mathbf{x} \in \mathcal{X}$, it holds that $V_{q'}$ is differentiable. Furthermore, for $\varrho = q'/(q'-1)$, it holds that

$$\begin{aligned} &\|\nabla V_{q'}(\mathbf{x})\|_\varrho \\ &= \|\mathbf{x} - \mathbf{x}^0\|_{q'}^{2-q'} \left(\sum_{i=1}^d \left(|x_i - x_i^0|\right)^{(q'-1)\varrho}\right)^{1/\varrho} \\ &= \|\mathbf{x} - \mathbf{x}^0\|_{q'}^{2-q'} \cdot \left(\sum_{i=1}^d \left(|x_i - x_i^0|\right)^{q'}\right)^{(q'-1)/q'} \\ &= \|\mathbf{x} - \mathbf{x}^0\|_{q'}. \end{aligned} \tag{22}$$

When the q'-norm diameter of the feasible region is bounded by $\mathcal{D}_{q'}$, we can tell that $V_{q'}$ is Lipschitz continuous in the following sense:

$$|V_{q'}(\mathbf{x}) - V_{q'}(\mathbf{y})| \leq \mathcal{D}_{q'} \cdot \|\mathbf{x} - \mathbf{y}\|_{q'}, \ \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}. \tag{23}$$

4. Main Results

This section presents the formal statements of our results. Subsection 4.1 discusses our assumptions, and then Subsection 4.2 provides our theorems, whose proofs are in the appendices.

4.1. Assumptions

Our first assumption is on the underlying randomness; as in (13), the variance of $\nabla f(\cdot,\xi)$ is assumed to be bounded everywhere on $\mathcal{X}$. We formalize this condition below:

Assumption 4.1 (Bounded variance)**.** For a given $p \geq 1$, there exists a scalar $\sigma_p \geq 1$ such that

$$\mathbb{E}\left[\|\nabla f(\mathbf{x},\xi) - \nabla F(\mathbf{x})\|_p^2\right] \leq \sigma_p^2 \text{ for every } \mathbf{x} \in \mathcal{X}. \tag{24}$$

Remark 4.2. We sometimes refer to this condition as "Assumption 4.1 w.r.t. the p-norm", which is common in the SP literature, especially in the discussions of stochastic first-order methods (SFOM), a mainstream alternative solution method for SP than the SAA (as discussed by, e.g., Ghadimi & Lan, 2013; 2016; Lan, 2020). This assumption is more general than the conditions on the underlying randomness

imposed by the benchmark results mentioned in (9) and (11) (or (12)). Indeed, the benchmark result in (9) imposes light-tailed-ness. Meanwhile, (11) assumes a finite p'th central moment of $[M(\xi)]^2$ with $p' \geq 2$, which implies a bounded fourth moment of $\|\nabla f(\cdot,\xi)\|_p$. (Here, $p = q/(q-1)$ with q given as in (10)). In contrast, Assumption 4.1 concerns only the second moment of $\|\nabla f(\cdot,\xi) - \nabla F(\cdot)\|_p$ and thus applies to the case with $p' = 1$, which is an unaddressed scenario for (11) (as well as for (12)). A comparable condition in the form of a finite second moment of $M(\xi)$ has also been considered by Oliveira & Thompson (2023), yet the corresponding complexity bounds are not fully explicit w.r.t. the dependence on β.

Remark 4.3. We hypothesize that the analyses in this paper can be extended to scenarios where Assumption 4.1 is replaced by the following relatively more flexible condition: $\mathbb{E}[\|\nabla F(\mathbf{x}) - \nabla f(\mathbf{x},\xi)\|_p^2] \leq \mathcal{C} \cdot \|\mathbf{x} - \mathbf{x}^*\|_q^2 + \sigma_p^2$ for all $\mathbf{x} \in \mathcal{X}$, where $\mathcal{C} \geq 0$ is some problem quantity. Nonetheless, we will leave the verification of this hypothesis to future work.

Finally, we would also like to remark that the stipulation of $\sigma_p \geq 1$ is non-critical; it is only for the simplification of notations in our results.

We formalize our assumptions of strong convexity and (general) convexity below:

Assumption 4.4 (μ-strong convexity w.r.t. the q-norm)**.** The following inequality holds for every pair of solutions $\mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}$ and almost every $\xi \in \Theta$:

$$f(\mathbf{x}_1,\xi) - f(\mathbf{x}_2,\xi) \geq \langle \nabla f(\mathbf{x}_2,\xi),\, \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\mu}{2} \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|_q^2, \quad (25)$$

for some given $\mu > 0$ and $q \geq 1$.

Remark 4.5. We refer to the above as "Assumption 4.4 w.r.t. the q-norm" or "μ-strong convexity w.r.t. the q-norm", which is common in the SAA literature (e.g., by Milz, 2023; Shalev-Shwartz et al., 2010). Some SP literature (e.g., by Ghadimi & Lan, 2012) assumes a relatively more flexible version of strong convexity than Assumption 4.4; more specifically, the below is stipulated instead therein:

$$F(\mathbf{x}_1) - F(\mathbf{x}_2) \geq \langle \nabla F(\mathbf{x}_2),\, \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\mu}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_q^2, \;\; \forall \mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}. \quad (26)$$

This condition is considered mostly in the discussions of the SFOM as an alternative SP's solution method mentioned above. Although one may easily verify that (26) is an immediate result of Assumption 4.4, the seemingly higher stringency in Assumption 4.4 does not make the SP problem much easier. Indeed, lower complexity bounds for the SFOM (such as by Rakhlin et al., 2011; Agarwal et al., 2009) are derived based on the identification of adversarial problems that satisfy Assumption 4.4. Based on these adversarial problems, one can infer that, when shifting from the assumption of (26) to Assumption 4.4, typical SFOMs may not achieve faster sample complexity rates in general.

Our second result in this section relaxes the condition of strong convexity in Assumption 4.4 into the condition of (general) convexity below:

Assumption 4.6 (General convexity)**.** The following inequality holds for all $\mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}$ and almost every $\xi \in \Theta$:

$$f(\mathbf{x}_1,\xi) - f(\mathbf{x}_2,\xi) \geq \langle \nabla f(\mathbf{x}_2,\xi),\, \mathbf{x}_1 - \mathbf{x}_2 \rangle.$$

Remark 4.7. We would like to compare the above with a counterpart assumption that the population-level objective $F(\cdot)$ is convex, which is, again, a common condition in the literature on the SFOM (e.g., by Nemirovski et al., 2009; Ghadimi & Lan, 2012; 2013). Relative to this counterpart condition, the incremental stringency in Assumption 4.6 does not make the SP problems much easier; this is because, again, the adversarial problem instances used to prove lower performance limits for SFOM for the (general) convex SP problems (such as those constructed by Agarwal et al., 2009) often satisfy Assumption 4.6. From such analysis, one can see that changing from the assumption of F being convex to Assumption 4.6 does not allow the SFOM to achieve a better sample efficiency in general.

4.2. Sample complexity bounds

We are now ready to formalize the promised sample complexity bounds in both strongly convex and (general) convex cases below.

Theorem 4.8 (Sample complexity for strongly convex SP)**.** *Suppose that Assumptions 1.1 and 4.4 hold both w.r.t. the q-norm for a given $q \geq 1$, and that Assumption 4.1 holds w.r.t. the p-norm for some $p : 1 \leq p \leq \frac{q}{q-1}$. Then any optimal solution $\widehat{\mathbf{x}}$ to the SAA in (5) satisfies the below: For any $\epsilon > 0$,*

$$\mathbb{E}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*)\right] \leq \epsilon, \quad \text{if } N \geq C_1 \cdot \max\left\{\frac{\mathcal{L}}{\mu},\, \frac{\sigma_p^2 + \mathcal{M}^2}{\mu\epsilon}\right\}; \quad (27)$$

and, meanwhile, for any $\epsilon > 0$ and $\beta \in (0,1)$,

$$\text{Prob}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \leq \epsilon\right] \geq 1 - \beta, \quad \text{if } N \geq C_1 \cdot \max\left\{\frac{\mathcal{L}}{\mu},\, \frac{\sigma_p^2 + \mathcal{M}^2}{\mu\epsilon\beta}\right\}. \quad (28)$$

Here, $C_1 > 0$ is some universal constant.

Proof. See Section A. □

Remark 4.9. The theorem above confirms the promised sample complexity in both (14) and (15) for strongly convex SP problems.

Remark 4.10. The formulation of SAA for solving a μ-strongly convex SP problem does not require estimating the value of μ, nor is it necessary to estimate σ_p, $\mathcal{M}$, or $\mathcal{L}$. This observation may lead to convenience in solving an SP problem, especially when compared to some alternative SP solution techniques such as the SFOM (e.g., as discussed by Lan, 2020). The latter often requires a reasonably high-fidelity estimation of at least some of those quantities in order to achieve a comparable sample complexity.

Our next theorem is focused on (general) convex SP. Before its statement, we first introduce some choice of hyper-parameters for the Tikhonov-like penalty term in (6) for a given $q > 1$ and a user-specified accuracy threshold $\epsilon > 0$: We let

$$q' \in (1, 2] : q' \leq q; \quad R^* \geq \max\{1, V_{q'}(\mathbf{x}^*)\}; \quad \text{and} \quad \lambda_0 = \frac{\epsilon}{2R^*}. \tag{29}$$

Theorem 4.11 (Sample complexity for general convex SP)**.** *Let $q > 1$. Suppose that the hyper-parameters q', R^*, and λ_0 are specified as in (29). Assume that (i) Assumption 1.1 w.r.t. the q-norm, (ii) Assumption 4.1 w.r.t. the p-norm for some $p : 1 \leq p \leq \frac{q}{q-1}$, and (iii) Assumption 4.6 hold. Any optimal solution to RSAA in (6), denoted by $\widehat{\mathbf{x}}$, satisfies the following inequalities: For any $\epsilon \in (0, 1]$,*

$$\mathbb{E}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*)\right] \leq \epsilon, \quad \text{if } N \geq \frac{C_2 R^*}{q' - 1} \cdot \max\left\{\frac{\mathcal{L}}{\epsilon}, \frac{\sigma_p^2 + \mathcal{M}^2}{\epsilon^2}\right\}; \tag{30}$$

and, meanwhile, for any $\epsilon \in (0, 1]$ and $\beta \in (0, 1)$,

$$\text{Prob}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) \leq \epsilon\right] \geq 1 - \beta, \quad \text{if } N \geq \frac{C_2 R^*}{q' - 1} \cdot \max\left\{\frac{\mathcal{L}}{\epsilon}, \frac{\sigma_p^2 + \mathcal{M}^2}{\beta\epsilon^2}\right\}. \tag{31}$$

Here, $C_2 > 0$ is some universal constant.

Proof. See Section B. □

Remark 4.12. We observe that the theorem above confirms the promised sample complexity in both (16) and (17) for (general) convex SP problems, if we notice that R^* is comparable to $V_{q'}(\mathbf{x}^*)$ therein.

Remark 4.13. The stipulation of $q > 1$ (and thus not including the choice of $q = 1$) is non-critical. Indeed, in the non-trivial case with $d > 1$, following the existing discussions of SFOM in the 1-norm setting (Nemirovski et al., 2009), the case where Assumption 1.1 holds for $q = 1$ can be subsumed by the consideration of the case with $q = 1 + \frac{1}{\ln d} > 1$ by the fact that

$$\|\mathbf{v}\|_{1+\frac{1}{\ln d}} \leq \|\mathbf{v}\|_1 \leq e \cdot \|\mathbf{v}\|_{1+\frac{1}{\ln d}},$$

where e is the base of natural logarithms.

Remark 4.14. A proper selection of λ_0 for this theorem relies on an overestimate of $V_{q'}(\mathbf{x}^*)$, which is equal to half of the squared q'-norm distance between the optimal solution $\mathbf{x}^*$ and any user-specified initial guess $\mathbf{x}^0$. Assuming (straightforward variations of) the knowledge of such a distance is not uncommon in related literature (e.g., as in Loh & Wainwright, 2011; Loh, 2017; Liu et al., 2022). In practice, when little is known about the SP's problem structure, one may choose $\mathbf{x}^0$ to be any feasible solution and specify R^* to be coarsely large; e.g., one may let R^* be half of the squared q'-norm diameter of the feasible region, if it is bounded. Starting from this coarse selection, one may then perform some empirical hyper-parameter search for better values of R^* (and thus λ_0) with the aid of cross validation. Meanwhile, if some problem structure about the SP problem is known, one may incorporate such *a priori* knowledge into the construction of $V_{q'}$. For instance, if it is known that $\mathbf{x}^*$ satisfies the weak sparsity condition (or the budget/capacity constraint) that $\|\mathbf{x}^*\|_1 \leq r$ for some known r (Negahban et al., 2012; Bugg & Aswani, 2021), then, in view of Remark 4.13, we may construct the regularization term with $q' = 1 + \frac{1}{\ln d}$ and $\mathbf{x}^0 = \mathbf{0}$. Correspondingly, $R^* = 0.5 \cdot e^2 \cdot r^2$.

Remark 4.15. As mentioned in Section 1, most existing sample complexity bounds grow polynomially with the complexity measures of the feasible region, such as $\Gamma_\epsilon(\mathcal{X})$ in (9) and $\gamma(\mathcal{X}^{*,\epsilon})$ in (12). These complexity measures can significantly elevate the dependence on d in general. Resultantly, if we fix all other quantities, the benchmark sample complexity rates in (9) and (12) can be simplified, respectively, into

$$O\left(\frac{\Gamma_\epsilon(\xi) + \ln(1/\beta)}{\epsilon^2}\right) \approx O\left(\frac{d + \ln(1/\beta)}{\epsilon^2}\right) \tag{32}$$

under light-tailed-ness, and

$$O\left(\frac{[\gamma(\mathcal{X}^{*,\epsilon})]^2 + \ln(1/\beta)}{\epsilon^2} + \frac{1}{\beta}\right) \approx O\left(\frac{d + \ln(1/\beta)}{\epsilon^2} + \frac{1}{\beta}\right). \tag{33}$$

In contrast, the two theorems (Theorems 4.8 and 4.11) in this paper confirm that it is possible to achieve sample complexity bounds completely free from any complexity measure of

feasible region, leading to new sample complexity rates of

$$\begin{cases} O\left(\frac{1}{\epsilon\cdot\beta}\right) & \text{strongly convex SP;} \\ O\left(\frac{1}{\epsilon^2\cdot\beta}\right) & \text{general convex SP.} \end{cases} \tag{34}$$

This perhaps marks the first explication of a universally better sample efficiency intrinsic to the (R)SAA than its existing benchmarks in (32) and (33), particularly in terms of the dependence on d.

Remark 4.16. In some applications, the variance σ_p^2 may also depend on dimensionality d. This dependence can be further explicated under additional assumption that, for some $\phi_p \geq 0$, it holds that $\|\nabla_i f(\mathbf{x},\xi) - \nabla_i F(\mathbf{x})\|_{L^p} \leq \phi_p$ for all $\mathbf{x} \in \mathcal{X}$ and every $i = 1, ..., d$. Intuitively, this additional assumption means that the component-wise pth central moment of $\nabla f(\mathbf{x},\xi)$ is bounded by ϕ_p^p everywhere. Because for $p \geq 2$, the function $(\cdot)^{2/p}$ is concave in '$\cdot$', one may easily see that the following holds:

$$\begin{aligned} &\mathbb{E}[\|\nabla f(\mathbf{x},\xi) - \nabla F(\mathbf{x})\|_p^2] \\ =&\mathbb{E}\left[\left(\sum_{i=1}^d |\nabla_i f(\mathbf{x},\xi) - \nabla_i F(\mathbf{x})|^p\right)^{2/p}\right] \\ \leq&\left(\sum_{i=1}^d \mathbb{E}\left[|\nabla_i f(\mathbf{x},\xi) - \nabla_i F(\mathbf{x})|^p\right]\right)^{2/p} \leq d^{2/p}\cdot\phi_p^2. \end{aligned}$$

Namely, in this case, one may let $\sigma_p^2 := d^{2/p}\cdot\phi_p^2$, whose dependence on dimensionality reduces when p increases. Particularly, when it is admissible to let $p > 2$, the dependence of σ_p^2 on d becomes better than any polynomial. Meanwhile, when it is feasible to let $p \geq c\ln d$ for some constant $c > 0$, the quantity σ_p^2 becomes dimension-free.

Remark 4.17. An important component of our proofs resorts to a seemingly novel argument based on the "average-replace-one (average-RO) stability" (Shalev-Shwartz et al., 2010), which is related to the average stability (Rakhlin et al., 2005), uniform-RO stability (Shalev-Shwartz et al., 2010), and uniform stability (Bousquet & Elisseeff, 2002). While it is known that the average-RO stability can lead to error bounds for learning algorithms (Shalev-Shwartz et al., 2010), seldom is there a sample complexity bound for (R)SAA based on such a stability type in comparable settings of our consideration. In contrast, most existing (R)SAA theories are based on either the "uniform convergence" theories, such as the ϵ-net (Shapiro et al., 2021) and the generic chaining (Oliveira & Thompson, 2023), or the variations of uniform (RO-) stability theories, such as by Feldman & Vondrak (2019); Shalev-Shwartz et al. (2010; 2009), and Klochkov & Zhivotovskiy (2021). Therefore, we think that our average-RO stability-based proof approach may also be of independent interest to some readers. One may see more discussions on how the average-RO stability is incorporated in our proofs from Remark A.1 in the appendix.

5. Conclusion

This paper revisits the SAA and its simple variation that incorporates the Tikhonov-like regularization. We particularly study their sample complexity in strongly convex and general convex SP problems under the assumptions of (i) the smoothness/continuity of the objective function comparable to, if not weaker than, the typical regularity conditions in the (R)SAA literature; and (ii) the heavy-tailed assumption that only the variance of the underlying randomness is bounded. Our results show that the SAA and the said RSAA variation exhibit new sample complexity rates that are — perhaps for the first time — provably free from any complexity measure of the feasible region. This marks a substantial deviation from the benchmark rates in both light-tailed and heavy-tailed scenarios, where a polynomial growth in the complexity measures of feasible region seems to have been unavoidable. Because such feasible set complexity measures can elevate the dependence of the sample complexity on the problem dimensionality in general, our new sample complexity bounds can be less dimension-sensitive than the state-of-the-art results in many applications.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

The authors would like to thank the anonymous reviewers for their active participation in the review-rebuttal cycle as well as their insightful and constructive comments. This work is partially supported by NSF CMMI 2016571 and NSF CMMI 2213459.

References

Agarwal, A., Wainwright, M. J., Bartlett, P., and Ravikumar, P. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

Artstein, Z. and Wets, R. J. Consistency of minimizers and the slln for stochastic programs. *J. Convex Anal*, 2(1-2): 1–17, 1995.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Ben-Tal, A., Margalit, T., and Nemirovski, A. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12 (1):79–108, 2001.

Birge, J. R. State-of-the-art-survey — stochastic programming: Computation and applications. *INFORMS journal on computing*, 9(2):111–133, 1997.

Birge, J. R. and Louveaux, F. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bugg, C. and Aswani, A. Logarithmic sample bounds for sample average approximation with capacity-or budget-constraints. *Operations Research Letters*, 49(2):231–238, 2021.

Dupacová, J. and Wets, R. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *The annals of statistics*, 16(4): 1517–1549, 1988.

Feldman, V. and Vondrak, J. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279. PMLR, 2019.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

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

Golub, G. H., Hansen, P. C., and O'Leary, D. P. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999.

Guigues, V., Juditsky, A., and Nemirovski, A. Non-asymptotic confidence bounds for the optimal value of a stochastic program. *Optimization Methods and Software*, 32(5):1033–1058, 2017.

Hoerl, A. E. and Kennard, R. W. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1): 69–82, 1970.

Hu, Y., Chen, X., and He, N. Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*, 30(3):2103–2133, 2020.

King, A. J. and Rockafellar, R. T. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18(1):148–162, 1993.

King, A. J. and Wets, R. J. Epi-consistency of convex stochastic programs. *Stochastics and Stochastic Reports*, 34(1-2):83–92, 1991.

Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization*, 12 (2):479–502, 2002.

Klochkov, Y. and Zhivotovskiy, N. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34: 5065–5076, 2021.

Krätschmer, V. First order asymptotics of the sample average approximation method to solve risk averse stochastic programs. *Mathematical Programming*, pp. 1–34, 2023.

Lan, G. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Lei, Y. and Ying, Y. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2020.

Liu, H., Wang, X., Yao, T., Li, R., and Ye, Y. Sample average approximation with sparsity-inducing penalty for high-dimensional stochastic programming. *Mathematical programming*, 178(1):69–108, 2019.

Liu, H., Ye, Y., and Lee, H. Y. High-dimensional learning under approximate sparsity with applications to nonsmooth estimation and regularized neural networks. *Operations Research*, 70(6):3176–3197, 2022.

Liu, T., Tao, D., and Xu, D. Dimensionality-dependent generalization bounds for k-dimensional coding schemes. *Neural computation*, 28(10):2213–2249, 2016.

Loh, P.-L. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. 2017.

Loh, P.-L. and Wainwright, M. J. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, 24, 2011.

Milz, J. Sample average approximations of strongly convex stochastic programs in hilbert spaces. *Optimization Letters*, 17(2):471–492, 2023.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

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

Oliveira, R. I. and Thompson, P. Sample average approximation with heavier tails i: non-asymptotic bounds with weak assumptions and stochastic constraints. *Mathematical Programming*, 199(1-2):1–48, 2023.

Pflug, G. C. Asymptotic stochastic programs. *Mathematics of Operations Research*, 20(4):769–789, 1995.

Pflug, G. C. Stochastic programs and statistical data. *Annals of Operations Research*, 85(0):59–78, 1999.

Pflug, G. C. Stochastic optimization and statistical inference. *Handbooks in operations research and management science*, 10:427–482, 2003.

Rakhlin, A., Mukherjee, S., and Poggio, T. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Ruszczyński, A. and Shapiro, A. Stochastic programming models. *Handbooks in operations research and management science*, 10:1–64, 2003.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *COLT*, volume 2, pp. 5, 2009.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Shapiro, A. Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics*, 17 (2):841–858, 1989.

Shapiro, A. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.

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

Shapiro, A. and Nemirovski, A. On complexity of stochastic programming problems. *Continuous optimization: Current trends and modern applications*, pp. 111–146, 2005.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Talagrand, M. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

A. Proof of Theorem 4.8

Proof. The proof of the first result of this theorem as in (27) takes two steps.

Step 1. Observe that

$$\mathbb{E}\left[F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*)\right] = \mathbb{E}\left[F(\widehat{\mathbf{x}}) - F_N(\mathbf{x}^*)\right] \leq \mathbb{E}\left[F(\widehat{\mathbf{x}}) - F_N(\widehat{\mathbf{x}})\right]. \tag{35}$$

Therefore, it suffices to establish an upper bound on $\mathbb{E}\left[F(\widehat{\mathbf{x}}) - F_N(\widehat{\mathbf{x}})\right]$, which is the focus of Step 2 in this proof.

Step 2. With the observation from Step 1, we construct a sequence of alternative SAA formulations with $F_N^{(j)}(\mathbf{x}) := \frac{1}{N}\left(f(\mathbf{x}, \xi_j') + \sum_{\iota \neq j} f(\mathbf{x}, \xi_\iota)\right)$, where ξ_j' is an i.i.d. copy of ξ, for all $j = 1, ..., N$. Let $\boldsymbol{\xi}_{1,N}^{(j)} := (\xi_1, ..., \xi_{j-1}, \xi_j', \xi_{j+1}, ..., \xi_N)$, which is obtained by switching the jth entry of $\boldsymbol{\xi}_{1,N}$ with ξ_j'. Correspondingly, let $\widehat{\mathbf{x}}^{(j)} := \widetilde{\mathbf{x}}(\boldsymbol{\xi}_{1,N}^{(j)})$ (and thus, by definition, $\widehat{\mathbf{x}}^{(j)} \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_N^{(j)}(\mathbf{x})$). Below, we establish an overestimate of $N^{-1}\sum_{j=1}^N \mathbb{E}\left[\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2\right]$. This overestimate is to play a key role in bounding $\mathbb{E}\left[F(\widehat{\mathbf{x}}) - F_N(\widehat{\mathbf{x}})\right]$.

To that end, we first observe that, for any $j = 1, ..., N$:

$$\begin{aligned}
&F_N(\widehat{\mathbf{x}}^{(j)}) - F_N(\widehat{\mathbf{x}}) \\
=&\frac{f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - f(\widehat{\mathbf{x}}, \xi_j)}{N} + \sum_{\iota \neq j} \frac{f(\widehat{\mathbf{x}}^{(j)}, \xi_\iota) - f(\widehat{\mathbf{x}}, \xi_\iota)}{N} && (36)\\
=&\frac{f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - f(\widehat{\mathbf{x}}, \xi_j)}{N} - \frac{f(\widehat{\mathbf{x}}^{(j)}, \xi_j') - f(\widehat{\mathbf{x}}, \xi_j')}{N} + F_N^{(j)}(\widehat{\mathbf{x}}^{(j)}) - F_N^{(j)}(\widehat{\mathbf{x}}) && (37)\\
\leq&\frac{f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - f(\widehat{\mathbf{x}}, \xi_j)}{N} - \frac{f(\widehat{\mathbf{x}}^{(j)}, \xi_j') - f(\widehat{\mathbf{x}}, \xi_j')}{N}. && (38)
\end{aligned}$$

Here (36) and (37) are by the definitions of F_N and $F_N^{(j)}$, and (38) is due to the fact that $\widehat{\mathbf{x}}^{(j)}$ minimizes $F_N^{(j)}$.

By Assumption 4.4, we have $f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - f(\widehat{\mathbf{x}}, \xi_j) \leq \langle \nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\rangle$ for almost every $\xi_j \in \Theta$, as well as $f(\widehat{\mathbf{x}}, \xi_j') - f(\widehat{\mathbf{x}}^{(j)}, \xi_j') \leq \langle \nabla f(\widehat{\mathbf{x}}, \xi_j'), \widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\rangle$ for almost every $\xi_j' \in \Theta$. Combining this with (38) leads to the below:

$$\begin{aligned}
&F_N(\widehat{\mathbf{x}}^{(j)}) - F_N(\widehat{\mathbf{x}}) \\
\leq&\frac{1}{N}\cdot\left\langle \nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\right\rangle + \frac{1}{N}\cdot\left\langle \nabla f(\widehat{\mathbf{x}}, \xi_j'), \widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\right\rangle, \quad a.s.\\
=&\frac{1}{N}\cdot\left\langle \nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - \nabla F(\widehat{\mathbf{x}}^{(j)}), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\right\rangle + \frac{1}{N}\cdot\left\langle \nabla f(\widehat{\mathbf{x}}, \xi_j') - \nabla F(\widehat{\mathbf{x}}), \widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\right\rangle\\
&+\frac{1}{N}\cdot\left\langle \nabla F(\widehat{\mathbf{x}}^{(j)}) - \nabla F(\widehat{\mathbf{x}}), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\right\rangle. && (39)
\end{aligned}$$

Further invoking Young's inequality and Assumption 1.1, which leads to

$$\begin{aligned}
&\left\langle \nabla F(\widehat{\mathbf{x}}^{(j)}) - \nabla F(\widehat{\mathbf{x}}), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\right\rangle\\
=&\left\langle \nabla F_1(\widehat{\mathbf{x}}^{(j)}) - \nabla F_1(\widehat{\mathbf{x}}), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\right\rangle + \left\langle \nabla F_2(\widehat{\mathbf{x}}^{(j)}) - \nabla F_2(\widehat{\mathbf{x}}), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\right\rangle\\
\leq&\mathcal{L}\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\|_q^2 + 2\mathcal{M}\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\|_q, && (40)
\end{aligned}$$

we may continue from the above to obtain, for all $\alpha > 0$ and every $j = 1, ..., N$,

$$\begin{aligned}
F_N(\widehat{\mathbf{x}}^{(j)}) - F_N(\widehat{\mathbf{x}}) \leq&\frac{1}{2\alpha\mu N^2}\cdot\left\|\nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - \nabla F(\widehat{\mathbf{x}}^{(j)})\right\|_p^2 + \frac{1}{2\alpha\mu N^2}\cdot\left\|\nabla f(\widehat{\mathbf{x}}, \xi_j') - \nabla F(\widehat{\mathbf{x}})\right\|_p^2\\
&+\left(\frac{\mathcal{L}}{N} + \alpha\mu\right)\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2 + \frac{16\mathcal{M}^2}{\mu N^2} + \frac{\mu}{16}\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2, \quad a.s. && (41)
\end{aligned}$$

By strong convexity of F_N as in Assumption 4.4 as well as the fact that $\widehat{\mathbf{x}}$ minimizes F_N, we have that

$$F_N(\widehat{\mathbf{x}}^{(j)}) - F_N(\widehat{\mathbf{x}}) \geq \frac{\mu}{2}\cdot\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2, \quad a.s. \tag{42}$$

Combining (41) and (42), we immediately obtain the below after some re-organization and simplification for all $j = 1, ..., N$:

$$\left[\left(\frac{7}{16} - \alpha\right) \cdot \mu - \frac{\mathcal{L}}{N}\right] \cdot \|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2 \leq \frac{1}{2N^2\mu\alpha} \cdot \left\|\nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - \nabla F(\widehat{\mathbf{x}}^{(j)})\right\|_p^2 + \frac{1}{2N^2\mu\alpha} \cdot \left\|\nabla f(\widehat{\mathbf{x}}, \xi_j') - \nabla F(\widehat{\mathbf{x}})\right\|_p^2 + \frac{16\mathcal{M}^2}{\mu N^2}, \quad a.s. \tag{43}$$

Note that $\widehat{\mathbf{x}}^{(j)}$ and ξ_j are independent, so are $\widehat{\mathbf{x}}$ and ξ_j'. We therefore have $\mathbb{E}[\|\nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - \nabla F(\widehat{\mathbf{x}}^{(j)})\|_p^2] \leq \sigma_p^2$ and $\mathbb{E}[\|\nabla f(\widehat{\mathbf{x}}, \xi_j') - \nabla F(\widehat{\mathbf{x}})\|_p^2] \leq \sigma_p^2$ by Assumption 4.1. Further because we may let $\alpha = 1/4$ and it is assumed that $N \geq \frac{C_1\mathcal{L}}{\mu}$, where we may as well let $C_1 \geq 8$, we have

$$\mathbb{E}\left[\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2\right] \leq \left[\left(\frac{7}{16} - \alpha\right) \cdot \mu - \frac{\mathcal{L}}{N}\right]^{-1} \cdot \left(\frac{\sigma_p^2}{N^2\mu\alpha} + \frac{16\mathcal{M}^2}{\mu N^2}\right) \leq \frac{64\sigma_p^2}{N^2\mu^2} + \frac{256\mathcal{M}^2}{N^2\mu^2}, \ \forall j = 1, ..., N;$$

$$\Longrightarrow N^{-1}\sum_{j=1}^N \mathbb{E}\left[\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2\right] \leq \frac{64\sigma_p^2}{N^2\mu^2} + \frac{256\mathcal{M}^2}{N^2\mu^2}. \tag{44}$$

Because $f(\widehat{\mathbf{x}}, \xi_j')$ and $f(\widehat{\mathbf{x}}^{(j)}, \xi_j)$ are identically distributed — so are $f(\widehat{\mathbf{x}}, \xi_j)$ and $f(\widehat{\mathbf{x}}^{(j)}, \xi_j')$ — we then obtain that $\mathbb{E}[f(\widehat{\mathbf{x}}, \xi_j')] = \mathbb{E}[f(\widehat{\mathbf{x}}^{(j)}, \xi_j)]$ and that $\mathbb{E}[f(\widehat{\mathbf{x}}, \xi_j)] = \mathbb{E}[f(\widehat{\mathbf{x}}^{(j)}, \xi_j')]$. Therefore,

$$\begin{aligned}
&\mathbb{E}[F(\widehat{\mathbf{x}}) - F_N(\widehat{\mathbf{x}})] \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N [F(\widehat{\mathbf{x}}) - f(\widehat{\mathbf{x}}, \xi_j)]\right] = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \left[f(\widehat{\mathbf{x}}, \xi_j') - f(\widehat{\mathbf{x}}, \xi_j)\right]\right], \\
=&\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[f(\widehat{\mathbf{x}}, \xi_j') - f(\widehat{\mathbf{x}}^{(j)}, \xi_j')\right] + \frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - f(\widehat{\mathbf{x}}, \xi_j)\right] \\
\leq&\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\langle \nabla f(\widehat{\mathbf{x}}, \xi_j'), \widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\rangle\right] + \frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\langle \nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\rangle\right] && (45) \\
=&\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\langle \nabla f(\widehat{\mathbf{x}}, \xi_j') - \nabla F(\widehat{\mathbf{x}}), \widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\rangle\right] \\
&+\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\langle \nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j) - \nabla F(\widehat{\mathbf{x}}^{(j)}), \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\rangle\right] \\
&+\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\langle \nabla F(\widehat{\mathbf{x}}) - \nabla F(\widehat{\mathbf{x}}^{(j)}), \widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\rangle\right] \\
\leq&\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\frac{8}{N\mu}\left\|\nabla f(\widehat{\mathbf{x}}, \xi_j') - \nabla F(\widehat{\mathbf{x}})\right\|_p^2 + \frac{8}{N\mu}\left\|\nabla F(\widehat{\mathbf{x}}^{(j)}) - \nabla f(\widehat{\mathbf{x}}^{(j)}, \xi_j)\right\|_p^2\right. \\
&\left.+\left(\mathcal{L} + \frac{N\mu}{16}\right)\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\|_q^2 + 2\mathcal{M}\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\|_q\right] && (46) \\
\leq& \frac{8}{N\mu}\sigma_p^2 + \left(\frac{\mathcal{L}}{2} + \frac{N\mu}{16}\right)N^{-1}\sum_{j=1}^N \mathbb{E}\left[\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2\right] + \frac{8\mathcal{M}^2}{N\mu} && (47) \\
\leq& \frac{C\cdot(\sigma_p^2 + \mathcal{M}^2)}{N\mu}, && (48)
\end{aligned}$$

for some universal constant $C > 0$. Here, (45) above is based on the strong convexity of $f(\cdot, \xi)$ for almost every $\xi \in \Theta$ as per Assumption 4.4, (46) is by the combination of (40) (as a result of Assumption 1.1) and Young's inequality, (47) is by Assumption 4.1, and the last inequality in (48) is by (44) and the assumption that $N \geq \frac{C_1\mathcal{L}}{\mu}$.

Eq. (48) above combined with (35) leads to the desired result in the first part of theorem as in (27). The second part of this theorem as in (28) is then an immediate result by the Markov's inequality, when it is combined with (35) and (48). □

Remark A.1. An important component of this proof is to establish an upper bound on $N^{-1}\sum_{j=1}^{N}\mathbb{E}\left[\|\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\|_q^2\right]$ as in (44). This bound ensures that, if one data point is changed to a different i.i.d. copy of ξ in SAA, the output solution does not change much, on average, in terms of the squared distance w.r.t. the q-norm. This is the manifestation of the innate average-RO stability of SAA when it is applied to solving a strongly convex SP problem. This average-RO stability serves as the pillar to the proof of our error bound in Theorem 4.8. The concept of average-RO stability is introduced by Shalev-Shwartz et al. (2010). To our knowledge, our proof may have been the first to use the average-RO stability to analyze the non-asymptotic sample complexity of the SAA.

B. Proof of Theorem 4.11

Proof. The proof below follows that of Theorem 4.8 with some important modifications. First, the RSAA in (6) can be considered as the SAA to the following new SP problem:

$$\min_{\mathbf{x}\in\mathcal{X}} F_{\lambda_0}(\mathbf{x}) := F(\mathbf{x}) + \lambda_0 V_{q'}(\mathbf{x}).$$

We repeat (35) to show that $\mathbb{E}\left[F_{\lambda_0}(\widehat{\mathbf{x}})-F_{\lambda_0}(\mathbf{x}^*)\right]\le\mathbb{E}\left[F_{\lambda_0}(\widehat{\mathbf{x}})-F_{\lambda_0,N}(\widehat{\mathbf{x}})\right]$ with $F_{\lambda_0,N}$ as defined in (6). Then, by the definition of F_{λ_0}, where $\lambda_0=0.5\epsilon/R^*$, an immediate result is that

$$\begin{aligned}&\mathbb{E}[F(\widehat{\mathbf{x}})+\lambda_0 V_{q'}(\widehat{\mathbf{x}})-F(\mathbf{x}^*)-\lambda_0 V_{q'}(\mathbf{x}^*)]\le\mathbb{E}\left[F_{\lambda_0}(\widehat{\mathbf{x}})-F_{\lambda_0,N}(\widehat{\mathbf{x}})\right]\\ \Longrightarrow&\mathbb{E}[F(\widehat{\mathbf{x}})-F(\mathbf{x}^*)]\le\mathbb{E}\left[F_{\lambda_0}(\widehat{\mathbf{x}})-F_{\lambda_0,N}(\widehat{\mathbf{x}})\right]+\lambda_0 V_{q'}(\mathbf{x}^*)\le\mathbb{E}\left[F_{\lambda_0}(\widehat{\mathbf{x}})-F_{\lambda_0,N}(\widehat{\mathbf{x}})\right]+\frac{\epsilon}{2}.\end{aligned}\tag{49}$$

Let $f_{\lambda_0}(\mathbf{x},\xi):=f(\mathbf{x},\xi)+\lambda_0 V_{q'}(\mathbf{x})$. With any $j=1,...,N$, define that $\boldsymbol{\xi}_{1,N}^{(j)}:=(\xi_1,...,\xi_{j-1},\xi_j',\xi_{j+1},...,\xi_N)$, which is obtained by switching the jth entry of $\boldsymbol{\xi}_{1,N}$ with ξ_j', an i.i.d. copy of ξ. Denote that $\widehat{\mathbf{x}}^{(j)}:=\widetilde{\mathbf{x}}(\boldsymbol{\xi}_{1,N}^{(j)})\in\arg\min_{\mathbf{x}\in\mathcal{X}}N^{-1}\left[\sum_{\iota\neq j}f_{\lambda_0}(\mathbf{x},\xi_\iota)+f_{\lambda_0}(\mathbf{x},\xi_j')\right]$. Under Assumption 4.6 and by the fact that $V_{q'}(\mathbf{x}):=\frac{1}{2}\|\mathbf{x}-\mathbf{x}^0\|_{q'}^2$, which is $(q'-1)$-strongly convex w.r.t. the q'-norm (Ben-Tal et al., 2001), we can follow Step 2 of the proof for Theorem 4.8. In particular, (39) therein implies that

$$\begin{aligned}&F_{\lambda_0,N}(\widehat{\mathbf{x}}^{(j)})-F_{\lambda_0,N}(\widehat{\mathbf{x}})\\ \le&\frac{1}{N}\cdot\left\langle\nabla f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j)-\nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)}),\,\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\right\rangle+\frac{1}{N}\cdot\left\langle\nabla f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j')-\nabla F_{\lambda_0}(\widehat{\mathbf{x}}),\,\widehat{\mathbf{x}}-\widehat{\mathbf{x}}^{(j)}\right\rangle\\ &+\frac{1}{N}\cdot\left\langle\nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)})-\nabla F_{\lambda_0}(\widehat{\mathbf{x}}),\,\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\right\rangle,\quad a.s.\end{aligned}\tag{50}$$

Observe that

$$\begin{aligned}&\left\langle\nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)})-\nabla F_{\lambda_0}(\widehat{\mathbf{x}}),\,\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\right\rangle\\ =&\left\langle\nabla F_1(\widehat{\mathbf{x}}^{(j)})-\nabla F_1(\widehat{\mathbf{x}}),\,\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\right\rangle+\left\langle\nabla F_2(\widehat{\mathbf{x}}^{(j)})-\nabla F_2(\widehat{\mathbf{x}}),\,\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\right\rangle\\ &+\left\langle\lambda_0\nabla V_{q'}(\widehat{\mathbf{x}}^{(j)})-\lambda_0\nabla V_{q'}(\widehat{\mathbf{x}}),\,\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\right\rangle\\ \le&\mathcal{L}\left\|\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\right\|_q^2+2\mathcal{M}\|\widehat{\mathbf{x}}-\widehat{\mathbf{x}}^{(j)}\|_q+\lambda_0\cdot(\|\widehat{\mathbf{x}}^{(j)}-\mathbf{x}^0\|_{q'}+\|\widehat{\mathbf{x}}-\mathbf{x}^0\|_{q'})\cdot\|\widehat{\mathbf{x}}^{(j)}-\widehat{\mathbf{x}}\|_{q'},\end{aligned}\tag{51}$$

where (51) is due to Assumption 1.1 and a property of $V_{q'}(\cdot)=0.5\|\cdot-\mathbf{x}^0\|_{q'}^2$ as in (22); that is, $\|\nabla V_{q'}(\cdot)\|_{p'}=\|\cdot-\mathbf{x}^0\|_{q'}$ for $p'=q'/(q'-1)$. Note that $\widehat{\mathbf{x}}^{(j)}$ and ξ_j are independent, so are $\widehat{\mathbf{x}}$ and ξ_j'. Assumption 4.1 then implies that $\mathbb{E}\left[\left\|\nabla f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j)-\nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)})\right\|_p^2\right]\le\sigma_p^2$ and $\mathbb{E}\left[\left\|\nabla f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j')-\nabla F_{\lambda_0}(\widehat{\mathbf{x}})\right\|_p^2\right]\le\sigma_p^2$. Further noting that $q'\le q$, we

may then continue from (50) above to obtain, for any $\alpha > 0$:

$$
\begin{aligned}
&\mathbb{E}[F_{\lambda_0,N}(\widehat{\mathbf{x}}^{(j)}) - F_{\lambda_0,N}(\widehat{\mathbf{x}})] \\
\leq&\mathbb{E}\left[\frac{1}{2\alpha(q'-1)\lambda_0 N^2}\cdot\left\|\nabla f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j) - \nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)})\right\|_p^2 + \frac{1}{2\alpha(q'-1)\lambda_0 N^2}\cdot\left\|\nabla f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j') - \nabla F_{\lambda_0}(\widehat{\mathbf{x}})\right\|_p^2\right] \\
&+\left(\frac{\mathcal{L}}{N} + \frac{(q'-1)\lambda_0}{16} + 2\alpha\lambda_0\cdot(q'-1)\right)\mathbb{E}\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_q^2 + \frac{16\mathcal{M}^2}{\lambda_0\cdot(q'-1)N^2} \\
&+\frac{\lambda_0}{4\alpha N^2\cdot(q'-1)}\cdot\mathbb{E}\left[(\|\widehat{\mathbf{x}}^{(j)} - \mathbf{x}^0\|_{q'} + \|\widehat{\mathbf{x}} - \mathbf{x}^0\|_{q'})^2\right] \\
\leq&\frac{\sigma_p^2}{\alpha(q'-1)\lambda_0 N^2} + \left(\frac{\mathcal{L}}{N} + \frac{(q'-1)\lambda_0}{16} + 2\alpha\lambda_0\cdot(q'-1)\right)\mathbb{E}\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_{q'}^2 + \frac{16\mathcal{M}^2}{\lambda_0\cdot(q'-1)N^2} \\
&+\frac{\lambda_0}{\alpha N^2\cdot(q'-1)}\cdot\mathbb{E}[\|\widehat{\mathbf{x}} - \mathbf{x}^0\|_{q'}^2],
\end{aligned}
\tag{52}
$$

where the last inequality is due to the relationship that $\mathbb{E}[\|\widehat{\mathbf{x}}-\mathbf{x}^0\|_{q'}^2] = \mathbb{E}[\|\widehat{\mathbf{x}}^{(j)}-\mathbf{x}^0\|_{q'}^2]$ and the assumption that $1 < q' \leq q$. Let $\alpha = 1/32$ and recall the assumption that $N \geq \frac{C_2\mathcal{L}}{(q'-1)\lambda_0}$, where we may as well let $C_2 \geq 8$. We may further invoke the $[(q'-1)\lambda_0]$-strong convexity of $F_{\lambda_0,N}$ in the sense of Assumption 4.4 as well as the fact that $\widehat{\mathbf{x}}$ minimizes $F_{\lambda_0,N}$ to obtain:

$$
\mathbb{E}\left[\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_{q'}^2\right] \leq \frac{128\sigma_p^2 + 64\mathcal{M}^2}{(q'-1)^2\lambda_0^2 N^2} + \frac{128}{N^2(q'-1)^2}\mathbb{E}[\|\widehat{\mathbf{x}} - \mathbf{x}^0\|_{q'}^2]. \tag{53}
$$

We observe that $f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j')$ and $f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j)$ are identically distributed, so are the pair of $f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j)$ and $f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j')$. Therefore,

$$
\begin{aligned}
&\mathbb{E}[F_{\lambda_0}(\widehat{\mathbf{x}}) - F_{\lambda_0,N}(\widehat{\mathbf{x}})] \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N [F_{\lambda_0}(\widehat{\mathbf{x}}) - f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j)]\right] = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \left[f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j') - f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j)\right]\right] \\
=&\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j') - f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j')\right] + \frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j) - f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j)\right] \\
\leq&\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\langle\nabla f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j') - \nabla F_{\lambda_0}(\widehat{\mathbf{x}}),\, \widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\rangle\right] + \frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\langle\nabla f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j) - \nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)}),\, \widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\rangle\right] \\
&+\frac{1}{2N}\sum_{j=1}^N \mathbb{E}[\langle\nabla F_{\lambda_0}(\widehat{\mathbf{x}}) - \nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)}),\, \widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\rangle]
\end{aligned}
\tag{54}
$$

$$
\begin{aligned}
\leq&\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\frac{8}{N(q'-1)\lambda_0}\left\|\nabla f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j') - \nabla F_{\lambda_0}(\widehat{\mathbf{x}})\right\|_p^2 + \frac{8}{N(q'-1)\lambda_0}\left\|\nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)}) - \nabla f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j)\right\|_p^2\right. \\
&+2\mathcal{M}\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\|_{q'} + \left(\frac{N(q'-1)\lambda_0}{16} + \mathcal{L}\right)\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\|_{q'}^2 \\
&\left.+\lambda_0\cdot(\|\widehat{\mathbf{x}}^{(j)} - \mathbf{x}^0\|_{q'} + \|\widehat{\mathbf{x}} - \mathbf{x}^0\|_{q'})\cdot\|\widehat{\mathbf{x}}^{(j)} - \widehat{\mathbf{x}}\|_{q'}\right]
\end{aligned}
\tag{55}
$$

$$
\begin{aligned}
\leq&\frac{1}{2N}\sum_{j=1}^N \mathbb{E}\left[\frac{8}{N(q'-1)\lambda_0}\left\|\nabla f_{\lambda_0}(\widehat{\mathbf{x}},\xi_j') - \nabla F_{\lambda_0}(\widehat{\mathbf{x}})\right\|_p^2 + \frac{8}{N(q'-1)\lambda_0}\left\|\nabla F_{\lambda_0}(\widehat{\mathbf{x}}^{(j)}) - \nabla f_{\lambda_0}(\widehat{\mathbf{x}}^{(j)},\xi_j)\right\|_p^2\right. \\
&+\frac{8\mathcal{M}^2}{N(q'-1)\lambda_0} + \left(\frac{N(q'-1)\lambda_0}{4} + \mathcal{L}\right)\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\|_{q'}^2 \\
&\left.+\frac{4\lambda_0}{N(q'-1)}\cdot(\|\widehat{\mathbf{x}}^{(j)} - \mathbf{x}^0\|_{q'} + \|\widehat{\mathbf{x}} - \mathbf{x}^0\|_{q'})^2\right],
\end{aligned}
\tag{56}
$$

where (54) is due to Assumption 4.6 as well as the convexity of $V_{q'}$, (55) is due to (51), the Hölder's and Young's inequalities, and the assumption that $q' \leq q$. Recall that (i) it has been assumed that $N \geq \frac{C_2\mathcal{L}}{(q'-1)\lambda_0}$, where we may let $C_2 \geq 8$; (ii) $\widehat{\mathbf{x}}^{(j)}$ and $\widehat{\mathbf{x}}$ are identically distributed; and (iii) Assumption 4.1. We then may continue from (53) and (56) above to obtain

$$\begin{aligned}
&\mathbb{E}[F_{\lambda_0}(\widehat{\mathbf{x}}) - F_{\lambda_0,N}(\widehat{\mathbf{x}})] \\
\leq& \frac{8\sigma_p^2 + 4\mathcal{M}^2}{N(q'-1)\lambda_0} + \frac{3N(q'-1)\lambda_0}{16}\mathbb{E}[\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}^{(j)}\|_{q'}^2] + \frac{8\lambda_0}{N(q'-1)} \cdot \mathbb{E}[\|\widehat{\mathbf{x}} - \mathbf{x}^0\|_{q'}^2] \\
\leq& \frac{32\sigma_p^2 + 16\mathcal{M}^2}{(q'-1)\lambda_0 N} + \frac{32\lambda_0}{N(q'-1)}\mathbb{E}[\|\widehat{\mathbf{x}} - \mathbf{x}^0\|_{q'}^2] \qquad (57) \\
=& \frac{32\sigma_p^2 + 16\mathcal{M}^2}{(q'-1)\lambda_0 N} + \frac{64\lambda_0}{N(q'-1)}\mathbb{E}[V_{q'}(\widehat{\mathbf{x}})], \qquad (58)
\end{aligned}$$

where (57) holds as a result of (53), and (58) holds by the definition of $V_{q'}$. In view of (58) and the definition of $\widehat{\mathbf{x}}$,

$$\begin{aligned}
0 \geq& \mathbb{E}[F_{\lambda_0,N}(\widehat{\mathbf{x}}) - F_{\lambda_0,N}(\mathbf{x}^*)] = \mathbb{E}[F_N(\widehat{\mathbf{x}}) + \lambda_0 V_{q'}(\widehat{\mathbf{x}}) - F_N(\mathbf{x}^*) - \lambda_0 V_{q'}(\mathbf{x}^*)] \\
=& \mathbb{E}[F_N(\widehat{\mathbf{x}}) + \lambda_0 V_{q'}(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) - \lambda_0 V_{q'}(\mathbf{x}^*)] \\
\overset{\text{Eq. (58)}}{\geq}& \mathbb{E}[F(\widehat{\mathbf{x}}) + \lambda_0 V_{q'}(\widehat{\mathbf{x}}) - F(\mathbf{x}^*) - \lambda_0 V_{q'}(\mathbf{x}^*)] - \frac{32\sigma_p^2 + 16\mathcal{M}^2}{(q'-1)\lambda_0 N} - \frac{64\lambda_0}{N(q'-1)}\mathbb{E}[V_{q'}(\widehat{\mathbf{x}})] \\
\geq& \mathbb{E}[\lambda_0 V_{q'}(\widehat{\mathbf{x}}) - \lambda_0 V_{q'}(\mathbf{x}^*)] - \frac{32\sigma_p^2 + 16\mathcal{M}^2}{(q'-1)\lambda_0 N} - \frac{64\lambda_0}{N(q'-1)}\mathbb{E}[V_{q'}(\widehat{\mathbf{x}})].
\end{aligned}$$

Because of the assumption that $R^* \geq 1$, $\sigma_p \geq 1$ and $0 < \epsilon \leq 1$, we have $N \geq C_2 \frac{(\sigma_p^2 + \mathcal{M}^2)R^*}{(q'-1)\epsilon} \implies N \geq \frac{320}{q'-1}$ for any $C_2 \geq 320$. Resultantly, re-arranging the inequality above, we immediately have the below, in view of the fact that $\lambda_0 V_{q'}(\mathbf{x}^*) = \frac{\epsilon}{2R^*} \cdot V_{q'}(\mathbf{x}^*) \leq \frac{\epsilon}{2}$:

$$\frac{4}{5}\mathbb{E}[\lambda_0 V_{q'}(\widehat{\mathbf{x}})] \leq \mathbb{E}[\lambda_0 V_{q'}(\mathbf{x}^*)] + \frac{32\sigma_p^2 + 16\mathcal{M}^2}{(q'-1)\lambda_0 N} \leq \frac{\epsilon}{2} + \frac{32\sigma_p^2 + 16\mathcal{M}^2}{(q'-1)\lambda_0 N}.$$

This inequality, combined with (58), leads to

$$\begin{aligned}
\mathbb{E}[F_{\lambda_0}(\widehat{\mathbf{x}}) - F_{\lambda_0,N}(\widehat{\mathbf{x}})] \leq& \frac{32\sigma_p^2 + 16\mathcal{M}^2}{(q'-1)\lambda_0 N} + \frac{64}{N(q'-1)} \cdot \left(\frac{5\epsilon}{8} + \frac{40\sigma_p^2 + 20\mathcal{M}^2}{(q'-1)\lambda_0 N}\right) \\
\leq& \frac{40\sigma_p^2 + 20\mathcal{M}^2}{(q'-1)\lambda_0 N} + \frac{\epsilon}{8}. \qquad (59)
\end{aligned}$$

where the last inequality above is due to $N \geq \frac{320}{q'-1}$ again. Combining (49) and (59), after some re-organization, we then obtain (30) as claimed.

Furthermore, if we invoke Markov's inequality together with (49) and (59), we then have shown (31) as desired. □