Operations Research Letters 49 (2021) 231–238

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Logarithmic sample bounds for Sample Average Approximation with capacity- or budget-constraints

Caleb Bugg *, Anil Aswani

Industrial Engineering and Operations Research, University of California, Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 14 September 2020
Received in revised form 20 November 2020
Accepted 8 January 2021
Available online 15 January 2021

Keywords:
Sample average approximation
Sample bounds
Stochastic complexity

ABSTRACT

Sample Average Approximation (SAA) is used to approximately solve stochastic optimization problems. In practice, SAA requires much fewer samples than predicted by existing theoretical bounds that ensure the SAA solution is close to optimal. Here, we derive new sample-size bounds for SAA that, for certain problems, are logarithmic (existing bounds are polynomial) in problem dimension. Notably, our new bounds provide a theoretical explanation for the success of SAA for many capacity- or budget-constrained optimization problems.

1. Introduction

In this paper, we consider a generic stochastic optimization problem of the form

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

where $\xi \in \Xi$ is some random variable with known distribution $P(\cdot)$, and $\mathcal{X} \subset \mathbb{R}^p$ is the (deterministic) feasible set. That is, we consider problems where stochasticity enters into the objective function and not the constraints. Note $F(x^*) = \mathbb{E}_\xi f(x^*,\xi)$ is the corresponding optimal value, where

$$x^* \in \arg\min_{x\in\mathcal{X}} F(x)$$

is any optimal point. Such problems are often difficult to solve because the expectation cannot be analytically computed except in cases when the distribution $P(\cdot)$ or the function $f(x,\xi)$ have very specific mathematical forms.

Sample Average Approximation (SAA) is a commonly-used procedure for solving (1), and it works by approximating the stochastic optimization problem using a deterministic optimization problem that is easier to solve [5,11,12,28,30]. The idea of SAA is to first generate an i.i.d. sample $\xi_1,\ldots,\xi_n$ of the random variable ξ, and then approximate the expectation $\mathbb{E}_\xi f(x,\xi)$ using its sample average

$$\min_{x\in\mathcal{X}}\{F_n(x) := \frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i)\}. \tag{2}$$

Note the objective function value of the original stochastic optimization problem (1) with the optimal solution of the SAA problem (2) is given by $F(\hat{x}_n) = \mathbb{E}_\xi f(\hat{x}_n,\xi)$, where we have that

$$\hat{x}_n \in \arg\min_{x\in\mathcal{X}} F_n(x)$$

is any optimal point of the SAA problem (2). By definition of x^* and $\hat{x}_n$, we have that $F(x^*) \le F(\hat{x}_n)$, and clearly we expect to see that $F(\hat{x}_n) \to F(x^*)$ almost surely as $n \to \infty$ by an argument using the uniform law of large numbers.

1.1. Sample bounds

Two practical considerations necessitate that the number of samples n in the SAA problem (2) be as small as possible. First, for many applications it is computationally costly to generate any single sample ξ_i of the random variable ξ. Second, for many functional forms of $f(\cdot,\cdot)$ it is the case that larger values of n require greater computation (e.g., more function evaluations, more gradient evaluations, etc.) in order to numerically solve the SAA problem (2).

Towards this goal, a now classical analysis [12,27,28] showed that in order to ensure

$$\mathbb{P}\big(F(\hat{x}_n) - F(x^*) \le \delta\big) \ge 1-\alpha \tag{3}$$

for any $\delta \in (0,1]$ and $\alpha \in (0,1]$, the number of samples n should satisfy

$$n \gtrsim \frac{p}{\delta^2}\log\frac{1}{\delta} + \frac{1}{\delta^2}\log\frac{1}{\alpha}. \tag{4}$$

Here, we have used the notation $x \gtrsim y$ of [16] which means $x \ge cy$ for some constant $c > 0$ that is independent of p, δ, α and which may depend polynomially upon other parameters of the

* Corresponding author.
E-mail addresses: caleb_bugg@berkeley.edu (C. Bugg), aaswani@berkeley.edu (A. Aswani).

https://doi.org/10.1016/j.orl.2021.01.007
0167-6377/

optimization problem (1). This bound says the required number of samples depends polynomially on the dimension p of the decision variable x, and this bound is prohibitively restrictive when p is high-dimensional. In fact, applications of SAA for large p are common in many domains.

However, the experience of many practitioners has been that a much smaller number of samples n (as compared to the above bound) is needed in order to ensure the SAA approximation is close to the true optimal value [11,17,25] — that is, the sample bound (4) is often overly conservative. Motivated by this empirical observation, there has been work on algorithmic approaches that iteratively or adaptively choose sample sizes in order to get good solutions with SAA with a small number of samples [23,24].

1.2. Contributions and outline

We briefly outline our paper and highlight our main contributions. To make our paper self-contained, we first provide in Section 2 an overview of the stochastic process theory of Rademacher complexity [1,14]. We include this overview because Rademacher complexity theory is not a common subject within the operations community. Our contributions begin in Section 3, where we use this stochastic process theory to derive a new bound for n to ensure (3). This bound depends in a non-trivial way upon the complex, stochastic interplay between $\mathcal{X}$ and $f(\cdot,\cdot)$. In Section 4, we describe an algorithmic procedure that can be used to numerically upper-bound this stochastic quantity for some stochastic optimization problems (1). Then in Section 4 we give examples of stochastic optimization problems where our approach yields explicit symbolic bounds on the number of samples n needed. Notably, we show that single-index models with an ℓ_1 constraint yields logarithmic bounds on the number of samples needed. We conclude with Section 5, where we conduct numerical experiments with the Markowitz portfolio selection problem to demonstrate the significant improvement of our bound relative to the classical bound (4).

1.3. Comparison to other sample bounds

One set of previous results [12,27,28] are based on an intermediate bound: When $\mathcal{X}$ is a discrete set of points, then (3) holds whenever $n \gtrsim \frac{1}{\delta^2}\log\frac{\#\mathcal{X}}{\alpha}$, where $\#\mathcal{X}$ is the cardinality of the finite set $\mathcal{X}$. Such a counting approach is effective for some problems: For example, these ideas have been used to prove that nonconvex (both integer and continuous) optimization problems with an ℓ_1 constraint have a significantly reduced theoretical computational complexity [20].

However, the above counting bound obscures the complex interplay between the feasible set $\mathcal{X}$ and the mathematical structure of the function $f(\cdot,\cdot)$, which is what actually governs the behavior of the SAA solutions. More recent work [21,22] uses empirical process theory to derive sample bounds, which is better able to capture the interplay between the objective and the feasible set. This work uses a chaining argument to characterize stochastic complexity, and this approach is in fact closely related our use of Rademacher theory that also characterizes stochastic complexity.

All sample bounds make assumptions about the continuity of $f(x,\xi)$ and about the problem stochasticity. It is common to assume $f(x,\xi)$ satisfies stochastic Lipschitz [6,12,27,28] or stochastic Hölder continuity [21,22] conditions. It is also common to assume stochastic regularity, such as having sub-Gaussian distributions [6,12,27,28] or assuming that sample averages are well-behaved [21,22]. Past results find sample bounds that are linear in the dimension p [6,12,27,28], or find sample bounds for specific problems like lasso that have effective dimensions that are smaller than p (and even logarithmic in p in the case of lasso) [21,22].

In this paper, we assume deterministic Lipschitz continuity on $f(x,\xi)$ and ensure stochastic regularity by requiring boundedness. We show sample bounds that are logarithmic in dimension p when the underlying stochastic optimization problem has ℓ_1 or nuclear norm constraints. The sample bounds of [6,12,27,28] are linear in p because they consider generic feasible sets, whereas the faster-than-linear bounds of [21,22] are calculated only for specific problems because their general sample bounds require knowing a difficult-to-compute quantity. However, we stress that our logarithmic sample bounds do not arise because of stronger assumptions, but are instead due to the geometry of the ℓ_1 or nuclear norm constraints. We demonstrate this by providing in this paper an alternative proof of logarithmic bounds under the more general assumptions of [12,27,28]. Our stronger assumptions are due to the proof technique we use. For instance, our boundedness assumption can be relaxed using the results of [13] though we do not consider this generalization here because it requires more notation that hides the main ideas.

Another related line of work has explored sample bounds under assumptions of sparsity. The work in [15,16] used sparsity-based techniques to study the relationship between sample size and SAA solution quality for the special case where the optimal solution x^* is sparse. (Here we define sparsity to include both low cardinality vectors and low rank matrices.) The approach of [15,16] adds a regularization term to the SAA (2), which induces sparsity in the optimal solution $\hat{x}$ to the SAA. The authors prove that this leads the sample size n to have poly-logarithmic dependence on the dimension p. However, there is an alternative explanation for these derived bounds. Since the optimal solution is known to be sparse, the effective feasible set $\mathcal{X}'$ (which incorporates the fact that x^* is sparse) can be taken to be much smaller than the stated feasible set $\mathcal{X}$. We use this idea in Section 4 to derive similar logarithmic bounds for a similar (to the ones studied by [15,16]) class of problems; these original bounds were achieved using regularization and using a much more technically difficult argument than the one we provide here.

2. Rademacher complexities

Before deriving our results, we need to provide a brief introduction to the stochastic process theory of Rademacher complexity [1,14], which is key to understanding our results. We provide this introduction because these results and underlying methodology are not generally known within the operations and control communities, though they are generally well-known within statistics and probability theory.

Let $\epsilon_1,\ldots,\epsilon_n$ be i.i.d. Rademacher random variables, where ϵ is a Rademacher random variable if its distribution is $\mathbb{P}(\epsilon = \pm 1) = \frac{1}{2}$; and let $f(x,\xi)$ be the function from the objective of (1). We define the *Rademacher complexity* of the function set $\mathcal{F} := \{f(x,\xi) : x \in \mathcal{X}\}$ to be

$$\mathcal{R}_n[f] = \mathbb{E}_\xi\left(\sup_{x\in\mathcal{X}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x,\xi_i)\right|\right). \tag{5}$$

Note the Rademacher complexity is often defined without an absolute value when the set $\mathcal{F}$ is symmetric, and we can use an equivalent definition without an absolute value

$$\mathcal{R}_n[f] = \mathbb{E}_\xi\left(\sup_{s\in\pm 1, x\in\mathcal{X}}\frac{1}{n}\sum_{i=1}^n \epsilon_i s f(x,\xi_i)\right)$$

by defining the augmented function set $\mathcal{F}' := \{sf(x,\xi) : s \in \pm 1, x \in \mathcal{X}\}$. Because both definitions are equivalent, we use the representation (5) to maintain consistency. Without loss of

generality, we will make the following assumption for the remainder of the paper:

Assumption. We have that $-\Delta/2 \leq f(x,\xi) \leq \Delta/2$ for all $(x,\xi) \in \mathcal{X} \times \Xi$, for some finite constant $\Delta \in \mathbb{R}_+$.

This assumption is without loss of generality because if the function $f(x,\xi)$ is bounded on its domain $(x,\xi) \in \mathcal{X} \times \Xi$ then we can always define an equivalent stochastic optimization problem by defining $f'(x,\xi) = f(x,\xi) - m - \Delta/2$ for some finite constant $m \in \mathbb{R}$ such that the above is satisfied.

We would like to emphasize that the above boundedness assumption can be relaxed by using recent generalizations of McDiarmid's inequality for unbounded random variables [13], but we do not consider this generalization here because it adds considerable notational complexity while obscuring the underlying ideas. In this paper, we assume boundedness so as to most clearly focus on the key underlying ideas.

The Rademacher complexity can be used to construct inequalities that bound the probability of large deviations of certain stochastic processes from their expectations. We begin with such a result on deviation between the sample average (2) from the objective of SAA with its expectation (1) in the objective of the stochastic optimization problem. The result below is similar to existing ones on concentration of measure [1,3,14], though our result differs somewhat from the standard forms. Its proof uses what is known as a *symmetrization* argument, and we repeat this argument here to introduce it to readers who are unfamiliar with it.

Proposition 1. *If the assumption holds, then we have*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\Big|F_n(x) - F(x)\Big| > t\Big) \leq \exp\Big(-2n\Big(\frac{t - 2\mathcal{R}_n[f]}{\Delta}\Big)^2\Big). \tag{6}$$

Proof. For notational convenience, we define

$$\mathcal{E} = \sup_{x\in\mathcal{X}}\Big|\frac{1}{n}\sum_{i=1}^n f(x,\xi_i) - \mathbb{E}_\xi f(x,\xi)\Big|$$

Observe that Jensen's inequality gives

$$\mathbb{E}\Big(\mathcal{E}\Big) \leq \mathbb{E}\Big(\sup_{x\in\mathcal{X}}\Big|\frac{1}{n}\sum_{i=1}^n f(x,\xi_i) - f(x,\xi_i')\Big|\Big),$$

where ξ_i, ξ_i' are i.i.d. But $f(x,\xi_i) - f(x,\xi_i')$ has a symmetric distribution, and so its distribution is equivalent to the distribution of $\epsilon_i \cdot (f(x,\xi_i) - f(x,\xi_i'))$ since the ϵ_i have a Rademacher distribution. And so we get

$$\mathbb{E}\Big(\mathcal{E}\Big) \leq \mathbb{E}\Big(\sup_{x\in\mathcal{X}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(x,\xi_i) - f(x,\xi_i'))\Big|\Big).$$

Applying the triangle inequality yields

$$\mathbb{E}\Big(\mathcal{E}\Big) \leq 2\mathbb{E}\Big(\sup_{x\in\mathcal{X}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x,\xi_i)\Big|\Big). \tag{7}$$

But observe that the right-hand side is $2\mathcal{R}_n[f]$. Since the function $f(\cdot,\cdot)$ is bounded by assumption, we can use the standard McDiarmid's inequality [3,19] to get

$$\mathbb{P}\Big(\mathcal{E} - \mathbb{E}\Big(\mathcal{E}\Big) > u\Big) \leq \exp\Big(-2n\Big(\frac{u}{\Delta}\Big)^2\Big).$$

Combining this with (7) implies that

$$\mathbb{P}\Big(\mathcal{E} - 2\mathcal{R}_n[f] > u\Big) \leq \exp\Big(-2n\Big(\frac{u}{\Delta}\Big)^2\Big),$$

or with the substitution $u = t - 2\mathcal{R}_n[f]$ that (6) holds. □

We also prove a nonstandard result (i.e., to the best of our knowledge this result is not found in the literature on Rademacher complexity) involving functions of (1).

Corollary 1. *Let $h : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and suppose the assumption holds. Then we have*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\Big|h(F_n(x)) - h(F(x))\Big| > t\Big) \leq \exp\Big(-2n\Big(\frac{t - 2L\mathcal{R}_n[f]}{L\Delta}\Big)^2\Big). \tag{8}$$

Proof. We first note that Lipschitz continuity of $h(\cdot)$ implies $|h(F_n(x)) - h(F(x))| \leq L|F_n(x) - F(x)|$. This means by Proposition 1 we have

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}\Big|h(F_n(x)) - h(F(x))\Big| > t\Big) \leq \mathbb{P}\Big(\sup_{x\in\mathcal{X}}L\Big|F_n(x) - F(x)\Big| > t\Big) \leq \exp\Big(-2n\Big(\frac{t/L - 2\mathcal{R}_n[f]}{\Delta}\Big)^2\Big).$$

The bound (8) now follows. □

It is pragmatically useful to interpret this corollary. In essence, the result says that applying a Lipschitz function $h(\cdot)$ to (1) is the same in terms of the concentration of measure as scaling the Rademacher complexity by L to $L\mathcal{R}_n[f]$ and scaling the assumption bound by L to $L\Delta$.

3. New sample bounds

The above proposition can be used to construct new sample bounds that ensure (3) holds for a generic stochastic optimization problem (1), and our next result gives an implicit formula for such a bound when $\mathcal{R}_n[f]$ is strictly decreasing.

Proposition 2. *Suppose $R_n[f]$ is strictly decreasing in n. If the assumption holds and $n \geq N$ with*

$$\begin{aligned} N &= \min_{\gamma_1,\gamma_2}\max\{N_1, N_2\} \\ N_1 &= (\Delta/\gamma_1\delta)^2\log(2/\alpha)/2 \\ N_2 &= \min\{n : 2R_n[f] \leq \gamma_2\delta\} \end{aligned} \tag{9}$$

and $\gamma_1, \gamma_2 \in (0,1)$ such that $2\gamma_1 + \gamma_2 = 1$, then (3) holds.

Proof. First observe that

$$F(\hat{x}_n) - F(x^*) = F(\hat{x}_n) - F_n(\hat{x}_n) + F_n(\hat{x}_n) - F_n(x^*) + F_n(x^*) - F(x^*).$$

Since $\hat{x}_n$ minimizes the SAA, we have $F_n(\hat{x}_n) \leq F_n(x^*)$. Let $\gamma_1, \gamma_2 \in (0,1)$ be such that $2\gamma_1 + \gamma_2 = 1$, and note the union bound gives

$$\mathbb{P}\Big(F(\hat{x}_n) - F(x^*) \leq \delta\Big) \geq 1 + -\mathbb{P}\Big(F(\hat{x}_n) - F_n(\hat{x}_n) > (\gamma_1+\gamma_2)\delta\Big) + -\mathbb{P}\Big(F_n(x^*) - F(x^*) > \gamma_1\delta\Big). \tag{10}$$

Next, observe that Hoeffding's inequality [3,7] implies

$$\mathbb{P}\Big(F_n(x^*) - F(x^*) > t\Big) \leq \exp\Big(-2N\Big(\frac{t}{\Delta}\Big)^2\Big)$$

since $n \geq N$. Note we use Hoeffding's inequality (i.e., not a uniform convergence result) since x^* is fixed. So if $-2N\cdot(\gamma_1\delta/\Delta)^2$

$= \log(\alpha/2)$ then $\mathbb{P}(F_n(x^*) - F(x^*) > \gamma_1\delta) \leq \alpha/2$. Now if $\gamma_2\delta \geq 2\mathcal{R}_n[f]$, then by Proposition 1 we have

$$\mathbb{P}\Big(F(\hat{x}_n) - F_n(\hat{x}_n) > (\gamma_1+\gamma_2)\delta\Big) \leq \exp\Big(-2n\Big(\frac{\gamma_1\delta}{\Delta}\Big)^2\Big).$$

Since $n \geq N$, then if $-2N\cdot(\gamma_1\delta/\Delta)^2 = \log(\alpha/2)$ then $\mathbb{P}(F(\hat{x}_n) - F_n(\hat{x}_n) > (\gamma_1+\gamma_2)\delta) \leq \alpha/2$. Combining the above with (10) gives the desired result. □

The above result can be difficult to interpret because of the implicit equation that specifies the minimum sample size N for (3) to hold. So we next present a simplified result for the case where the Rademacher complexity can be bounded in the form $\mathcal{R}_n[f] \leq \frac{c(p)}{\sqrt{n}}$ for some function $c(p)$. Such a bound can be constructed for many cases [8–10,14,26].

Corollary 2. *If the assumption holds, $\mathcal{R}_n[f] = \frac{c(p)}{\sqrt{n}}$, and*

$$n \geq \frac{1}{2}\Big(\frac{4\Delta}{\delta}\Big)^2\log\Big(\frac{2}{\alpha}\Big) + \Big(\frac{4c(p)}{\delta}\Big)^2$$

then we have that (3) holds.

Proof. We compute $N_2 = \min\{n : 2R_n[f] \leq \gamma_2\delta\}$ under the additional assumption. Specifically, $2c(p)/\sqrt{N_2} = \gamma_2\delta$, or rewritten that $N_2 = (2c(p)/\gamma_2\delta)^2$. Next consider an $N' := N_1 + N_2 \geq N = \min_{\gamma_1,\gamma_2}\max\{N_1, N_2\}$. Noting that $\gamma_2 = 1 - 2\gamma_1$, we have that

$$N' = (\Delta/\gamma_1\delta)^2\log(2/\alpha)/2 + (2c(p)/(1-2\gamma_1)\delta)^2.$$

Our result follows by choosing $\gamma_1 = 1/4$. □

4. Bounds for problems

The prior two results bound the sample size n required to achieve (3), but their use requires knowing the Rademacher complexity $\mathcal{R}_n[f]$ for a particular stochastic optimization problem (1). Here, we discuss how the Rademacher complexity can be bounded for various classes of stochastic optimization problems. We describe a Monte Carlo approach, and we also give explicit bounds for specific problems.

4.1. Monte Carlo bounds

The first class of problems we consider are those where there exists a surrogate optimization problem that provides an upper bound of the form

$$\sup_{x\in\mathcal{X}}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x,\xi_i)\Big| \leq \max_{y\in\mathcal{Y}} G(y,\{\epsilon_i,\xi_i\}_{i=1}^n). \tag{11}$$

Our scenario is where the optimization problem on the right-hand side is easily-solvable. For example, when the problem on the left-hand side can be represented as a mixed-integer linear program (MILP) then the problem on the right-hand side could be its continuous relaxation. Various approximation techniques exist, and how different classes of problems for the left-hand side can be upper-bounded by a surrogate problem are beyond the scope of this present paper.

For this scenario, we define a Monte Carlo estimate. Let ξ_{ij} and ϵ_{ij} be i.i.d. samples of ξ and of the Rademacher random variable, respectively. Then a Monte Carlo estimate $\widehat{\mathcal{R}}[f]$ of the Rademacher complexity is given by

$$\widehat{\mathcal{R}}_n[f] = \frac{1}{m}\sum_{j=1}^{m}\Big(\max_{y_j\in\mathcal{Y}} G(y_j,\{\epsilon_{ij},\xi_{ij}\}_{i=1}^n)\Big)$$

where m is the number of repetitions. Our next result concerns the correctness of this Monte Carlo estimate.

Proposition 3. *Suppose $0 \leq G(y,\{\epsilon_i,\xi_i\}^n) \leq \frac{\sigma}{2}$ for all $(y,\{\epsilon_i,\xi_i\}^n) \in \mathcal{Y}\times\{\pm 1,\Xi\}^n$, for some finite constant $\sigma\in\mathbb{R}_+$. Then we have $\mathcal{R}_n[f] \leq \widehat{\mathcal{R}}_n[f] + d$ with probability at least $1 - \exp(-2m(\frac{d}{\sigma})^2)$.*

Proof. First note that Hoeffding's inequality implies

$$\mathbb{P}\Big(\widehat{\mathcal{R}}_n[f] < \mathbb{E}\Big(\widehat{\mathcal{R}}_n[f]\Big) - d\Big) \leq \exp\Big(-2m\Big(\frac{d}{\sigma}\Big)^2\Big).$$

We next define the quantity

$$\mathcal{F} = \frac{1}{m}\sum_{j=1}^{m}\Big(\sup_{x_j\in\mathcal{X}}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,j}f(x_j,\xi_{i,j})\Big|\Big).$$

But by definition $\mathcal{F} \leq \widehat{\mathcal{R}}_n[f]$ and $\mathbb{E}\mathcal{F} = \mathcal{R}_n[f] \leq \mathbb{E}\widehat{\mathcal{R}}_n[f]$, and so this means that we have

$$\mathbb{P}\Big(\widehat{\mathcal{R}}_n[f] < \mathcal{R}_n[f] - d\Big) \leq \exp\Big(-2m\Big(\frac{d}{\sigma}\Big)^2\Big).$$

The result follows by taking the complement of this. □

The above result says the Monte Carlo estimate $\widehat{\mathcal{R}}_n[f]$ upper bounds the Rademacher complexity $\mathcal{R}_n[f]$ with high probability. The quality of the estimate depends upon two factors. The first is how weak the upper bound (11) of the surrogate optimization problem is. The second is how large m is, with larger values corresponding to more accurate estimates. We reiterate that this approach is only feasible when the surrogate problem is easily-solvable, which enables the use of large values of m in computing the estimate. We will provide a specific example in the next section.

4.2. Explicit bounds

Next, we provide explicit bounds for specific classes of stochastic optimization problems.

Proposition 4. *Let $g:\mathbb{R}\to\mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{x\in\mathcal{S}}\Big\{\mathbb{E}_\xi\big(g(\xi^\mathsf{T}x)\big) \;\Big|\; \|x\|_1 \leq \lambda\Big\} \tag{12}$$

where $\mathcal{S}\subseteq\mathbb{R}^p$ and $\max_{\xi\in\Xi}\|\xi\|_\infty \leq C < +\infty$. Then the Rademacher complexity of the above problem is bounded by $\mathcal{R}_n[f] \leq \lambda LC\sqrt{2\log 2p/n}$, and we need

$$n \geq \Big(\frac{4\lambda LC}{\delta}\Big)^2\cdot\Big(2\log\Big(\frac{2}{\alpha}\Big) + 2\log 2p\Big) \tag{13}$$

samples to ensure that (3) holds.

Proof. Let $\Lambda = \{x : \|x\|_1 \leq \lambda\}$, and note that we have

$$\max_{x,y\in\Lambda}\Big|g(\xi^\mathsf{T}x) - g(\xi^\mathsf{T}y)\Big| \leq L\max_{x,y\in\Lambda}\Big|\xi^\mathsf{T}(x-y)\Big| \leq 2\lambda LC$$

where the first inequality follows by Lipschitz continuity of $g(\cdot)$, and the second inequality follows by Hölder's inequality. This means the assumption holds for $\Delta = 2\lambda LC$. Next we bound the Rademacher complexity of the problem (12), which is bounded by L times the Rademacher complexity for the stochastic optimization problem where $g(\cdot)$ is the identity function (see Lemma 26.9 in [26], which was originally Lemma 1.1 from the lecture notes [10]; a slightly less general version of this result was first shown by [14]). This second Rademacher complexity for when $g(\cdot)$ is the identity was bounded in [9], and the final result is as above. The sample bound (13) now follows from Corollary 2. □

The above single-index model with an ℓ_1 constraint is a situation where we need logarithmic in p samples for SAA, which is a substantial improvement over the standard bound (4) showed

by [12,27,28] that is polynomial in p. Continuity is needed for logarithmic bounds. For instance, Proposition 2 of [6] gives an example of a particular $f(x,\xi)$ with an ℓ_2 constraint $\mathcal{X} = \{x : \|x\|_2 \leq \lambda\}$, where $F(x)$ is Lipschitz and linear bounds are necessary for a small optimality gap. By defining $g(x)$ such that $g(0) = 0$ and $g(x) = x \cdot \|x\|_1/\|x\|_2$ otherwise, we can thus use that particular $f(\cdot,\cdot)$ to construct an example $f(g(x),\xi)$ with an ℓ_1 constraint $\mathcal{X} = \{x : \|x\|_1 \leq \lambda\}$ where $F(g(x))$ is not Lipschitz and a linear number of samples is necessary for a small gap.

Next we consider a class of problems similar to [16], and we show a similar logarithmic in p bound but without using regularization and with a much simpler technical argument.

Corollary 3. *Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{x\in\mathcal{X}} \mathbb{E}_\xi\big(g(\xi^\mathsf{T}x)\big) \tag{14}$$

where $\max_{\xi\in\Xi}\|\xi\|_\infty \leq C < +\infty$. Suppose there is an optimal solution x^* to (14) that is sparse, meaning that $s := \sum_{i=1}^p \mathbf{1}(x_i^* \neq 0)$ is small with $\|x^*\|_\infty \leq \mu < +\infty$. Then

$$n \geq \left(\frac{4\mu sLC}{\delta}\right)^2 \cdot \left(2\log\left(\frac{2}{\alpha}\right) + 2\log 2p\right)$$

samples ensures that (3) holds when $\hat{x}$ is the SAA solution to the stochastic optimization problem

$$\min_{x\in\mathcal{X}}\left\{\mathbb{E}_\xi\big(g(\xi^\mathsf{T}x)\big) \;\middle|\; \|x\|_1 \leq \mu s\right\}. \tag{15}$$

Proof. Note that $\|x^*\|_1 \leq \mu s$ by assumption. Thus x^* is an optimal solution for both (15) and (14), and both problems have the same minimum value $F(x^*)$. The result now follows by applying Proposition 4 to (15). □

Last we note that our sample bound can be improved when the problem has nonnegativity constraints.

Proposition 5. *Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{x\in\mathcal{S}}\left\{\mathbb{E}_\xi\big(g(\xi^\mathsf{T}x)\big) \;\middle|\; x \geq 0, \|x\|_1 \leq \lambda\right\}$$

where $\mathcal{S} \subseteq \mathbb{R}^p$ and $\max_{\xi\in\Xi}\|\xi\|_\infty \leq C < +\infty$. Then the Rademacher complexity of the above problem is bounded by $\mathcal{R}_n[f] \leq \lambda LC\sqrt{\log p/n}$, and we need

$$n \geq \left(\frac{4\lambda LC}{\delta}\right)^2 \cdot \left(\frac{1}{2}\log\left(\frac{2}{\alpha}\right) + \log p\right)$$

samples to ensure that (3) holds.

The proof is omitted because it is essentially identical to that of Proposition 4, the main difference being a different bound from [9] is used for the Rademacher complexity.

4.3. Explicit bounds for matrix optimization

Next, we provide bounds for specific classes of stochastic matrix optimization problems. We use $\|X\|_*$ to denote the nuclear norm of $X \in \mathbb{R}^{p\times q}$, and $\|\xi\|_2$ in this section denotes the spectral norm of the random matrix $\xi \in \Xi \subset \mathbb{R}^{p\times q}$.

Proposition 6. *Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{X\in\mathcal{S}}\left\{\mathbb{E}_\xi\big(g(\mathrm{tr}(\xi^\mathsf{T}X))\big) \;\middle|\; \|X\|_* \leq \lambda\right\}$$

where $\mathcal{S} \subseteq \mathbb{R}^{p\times q}$ and $\max_{\xi\in\Xi}\|\xi\|_2 \leq C < +\infty$. Then the Rademacher complexity of the above stochastic optimization problem is bounded by $\mathcal{R}_n[f] \leq \lambda LC\sqrt{3\log(\min\{p,q\})/n}$, and we need

$$n \geq \left(\frac{4\lambda LC}{\delta}\right)^2 \cdot \left(2\log\left(\frac{2}{\alpha}\right) + 3\log\big(\min\{p,q\}\big)\right)$$

samples to ensure that (3) holds.

Proof. Let $\Lambda = \{X : \|X\|_* \leq \lambda\}$, and note that we have

$$\begin{aligned}\max_{X,Y\in\Lambda}\Big|g(\mathrm{tr}(\xi^\mathsf{T}X)) - g(\mathrm{tr}(\xi^\mathsf{T}Y))\Big| &\leq L\max_{X,Y\in\Lambda}\Big|\mathrm{tr}(\xi^\mathsf{T}(X-Y))\Big| \\ &\leq L\max_{X,Y\in\Lambda}\|\xi\|_2\|X-Y\|_* \\ &\leq 2\lambda LC\end{aligned}$$

where the first inequality follows by Lipschitz continuity of $g(\cdot)$, and the second inequality follows by Hölder's inequality for unitarily invariant norms [2]. This means the assumption holds for $\Delta = 2\lambda LC$. Next we bound the Rademacher complexity of (12): Observe that this is bounded by L times the Rademacher complexity for when the function $g(\cdot)$ is the identity function (see Lemma 26.9 in [26] and Lemma 1.1 in [10]). The Rademacher complexity for when $g(\cdot)$ is the identity was bounded in [8], and the final result is as above. The sample bound (13) now follows from Corollary 2. □

The above single-index model with a nuclear norm constraint needs logarithmic in $\min\{p,q\}$ samples for SAA. This is a substantial improvement over the standard bound (4) showed by [12,27,28] that in this case is

$$n \gtrsim \frac{pq}{\delta^2}\log\frac{1}{\delta} + \frac{1}{\delta^2}\log\frac{1}{\alpha}.$$

which is polynomial in p and q.

Next we consider a class of problems similar to [15], and we show a logarithmic in $\min\{p,q\}$ bound but without using regularization and with a much simpler technical argument.

Corollary 4. *Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz with constant L, and consider the stochastic optimization problem*

$$\min_{X\in\mathcal{X}} \mathbb{E}_\xi\big(g(\mathrm{tr}(\xi^\mathsf{T}X))\big) \tag{16}$$

where $\mathcal{X} \subseteq \mathbb{R}^{p\times q}$ and $\max_{\xi\in\Xi}\|\xi\|_2 \leq C < +\infty$. Suppose there is an optimal solution X^* to (16) that is low rank, meaning $r := \mathrm{rank}(X^*)$ is small with $\|X^*\|_2 \leq \mu < +\infty$. Then we need

$$n \geq \left(\frac{4\mu rLC}{\delta}\right)^2 \cdot \left(2\log\left(\frac{2}{\alpha}\right) + 3\log\big(\min\{p,q\}\big)\right)$$

samples to ensure that (3) holds when $\hat{X}$ is the SAA solution to the stochastic optimization problem

$$\min_{x\in\mathcal{X}}\left\{\mathbb{E}_\xi\big(g(\mathrm{tr}(\xi^\mathsf{T}X))\big) \;\middle|\; \|X\|_* \leq \mu r\right\}. \tag{17}$$

Proof. Note that $\|X^*\|_* \leq \mu r$ by assumption. Thus X^* is an optimal solution for both (17) and (16), and both problems have the same minimum value $F(x^*)$. The result now follows by applying Proposition 6 to (17). □

4.4. Alternative proof

We conclude this section by considering the more general setting of the past bound from [28]. We give an alternative proof of our result for problems with an ℓ_1 constraint, which modifies the proof of Theorem 5.18 in [28].

Proposition 7. *Consider the stochastic optimization*

$$\min_{x\in\mathcal{S}}\left\{\mathbb{E}_\xi f(x,\xi)\;\middle|\;\|x\|_1\le\lambda\right\} \tag{18}$$

where $\mathcal{S}\subseteq\mathbb{R}^p$. Let $\Lambda=\{x\in\mathcal{S}:\|x\|_1\le\lambda\}$, and suppose two assumptions hold. First, for any $x',x\in\Lambda$ there exists constant $\sigma_{x',x}>0$ such that the moment-generating function $M_{x',x}(t)=\mathbb{E}_\xi\exp(tY_{x',x})$ of random variable $Y_{x',x}=[f(x',\xi)-F(x')]-[f(x,\xi)-F(x)]$ satisfies $M_{x',x}(t)\le\exp(\sigma_{x',x}^2t^2/2)$ for all $t\in\mathbb{R}$. Second, there exists a (measurable) function $\kappa:\Xi\to\mathbb{R}_+$ such that its moment-generating function $M_\kappa(t)$ is finite valued for all t in a neighborhood of zero and $|f(x',\xi)-f(x,\xi)|\le\kappa(\xi)\|x'-x\|$ for almost everywhere $\xi\in\Xi$ and all $x',x\in\Lambda$. Then (3) holds whenever

$$n\ge\frac{8\sigma^2}{\delta^2}\cdot\log\frac{64\lambda^2L^2p}{\delta^2}+\left(\frac{8\sigma^2}{\delta^2}+\frac{1}{\beta}\right)\cdot\log\left(\frac{2}{\alpha}\right), \tag{19}$$

where $\sigma^2=\sup_{x',x\in\Lambda}(\sigma_{x',x})^2$, $L=\mathbb{E}_\xi\kappa(\xi)$, and we have that $\beta=\sup_{t\in\mathbb{R}}\big(2Lt-\log M_\kappa(t)\big)$.

Proof. We first show there exists a set $\mathcal{V}=\{x_1,\ldots,x_k\}$ with $\log k\le 32(\lambda L/\delta)^2\log p$ so $\max_{x\in\Lambda}\min_{x'\in\mathcal{X}}\|x-x'\|\le\delta/8L$. To show this, we use the Sudakov minoration [29] that says $\sqrt{\log k}\le\mathbb{E}(\sup_{x\in\Lambda}g^{\mathsf T}x)/2(\delta/8L)$, where $g\in\mathbb{R}^p$ is a vector whose entries are i.i.d. Gaussian random variables with zero mean and unit variance. Hölder's inequality and the symmetry of Λ imply that we have $\mathbb{E}(\sup_{x\in\Lambda}g^{\mathsf T}x)\le\mathbb{E}(\sup_{x\in\Lambda}\|x\|_1\cdot\max_j|g_j|)\le\lambda\sqrt{2\log p}$ where we have used the basic bound $\mathbb{E}(\max_j|g_j|)\le\sqrt{2\log p}$ for g_j that are the jth entry of the vector g. Thus Sudakov's minoration gives that $\sqrt{\log k}\le\lambda\sqrt{2\log p}/2(\delta/8L)$. Rearranging this inequality gives the desired bound $\log k\le 32(\lambda L/\delta)^2\log p$.

Next choose any x^* that minimizes (18) and consider the modified stochastic optimization problem $\min_{x\in\mathcal{V}\cup\{x^*\}}F(x)$. Let $x^{*,v}\in\arg\min_{x\in\mathcal{V}\cup\{x^*\}}F(x)$, and define the solution $\hat{x}_n^v\in\arg\min_{x\in\mathcal{V}\cup\{x^*\}}F_n(x)$. Note $F_n(\hat{x}_n)\le F_n(\hat{x}_n^v)$ since $\mathcal{V}\cup\{x^*\}\subseteq\Lambda$. The second assumption implies that with probability one we have $|F_n(x')-F_n(x)|\le\hat{\kappa}_n\|x'-x\|$ for all $x',x\in\Lambda$, where $\hat{\kappa}_n=\frac{1}{n}\sum_{i=1}^n\kappa(\xi_i)$. By construction of $\mathcal{V}$, there exists x' with $\|x'-\hat{x}_n\|\le\delta/8L$. Thus $F_n(x')\le F_n(\hat{x}_n)+\hat{\kappa}_n\cdot(\delta/8L)\le F_n(\hat{x}_n^v)+\hat{\kappa}_n\cdot(\delta/8L)$.

We continue our analysis under the event that $\hat{\kappa}_n\le 2L$. Here, $F_n(x')\le F_n(\hat{x}_n^v)+\delta/4$. Thus Theorem 5.17 of [28] says that $\mathbb{P}\big(F(x')-F(x^{*,v})\le 3\delta/4\big)\ge 1-\alpha/2$ for $n\ge(8\sigma^2/\delta^2)\times\log(2(k+1)/\alpha)$. But $F(x^{*,v})=F(x^*)$ by construction of the modified stochastic optimization, and also $F(\hat{x}_n)\le F(x')+L\cdot(\delta/8L)$ since the second assumption implies that $|F(x')-F(x)|\le L\|x'-x\|$ for all $x',x\in\Lambda$. This means we have $F(\hat{x}_n)\le F(x^*)+3\delta/4+\delta/8$ with probability at least $1-\alpha/2$. Note the Chernoff bound implies that $\hat{\kappa}_n\le 2L$ with probability at least $1-\exp(-n\beta)=1-\alpha/2$ when $n\ge\log(2/\alpha)/\beta$. Thus (3) holds when (19) holds. □

The significance of this alternative proof under more general conditions is that it shows that the logarithmic sample bounds arise because of the properties of ℓ_1 constraint. (A similar alternative proof of logarithmic sample bounds under more general conditions can also be constructed for nuclear norm constraints.) The boundedness and Lipschitz continuity assumptions we make in previous subsections are due to the technical details of the proof technique that we use.

5. Numerical experiments

Consider a scenario where we would like to choose a portfolio that allocates investments into some combination of p risky assets and 1 risk-free asset, while considering a tradeoff between maximizing the expected return of the portfolio and the risk tolerance of the investor. The Markowitz portfolio selection model [4,18] is a simple framework to pose such a problem. Let $\xi\in\mathbb{R}^p$ be a random variable of the returns from the p risky assets, and define $\mu=\mathbb{E}_\xi\xi$ and $\Sigma=\mathbb{E}_\xi((\xi-\mu)(\xi-\mu)^{\mathsf T})$. Then one formulation of the problem involves solving a convex quadratic program

$$\min_{x\in\mathbb{R}^p}\left\{x^{\mathsf T}\Sigma x-\gamma\cdot x^{\mathsf T}(\mu-r\mathbf{1})\;\middle|\;x\ge 0,\|x\|_1\le 1\right\} \tag{20}$$

where: r is the rate of return for the risk-free asset, $\gamma>0$ trades-off between the returns and risk of the portfolio, and each entry of the vector x gives the fraction of the portfolio allocated to the p risky assets; hence $1-\sum_{i=1}^p x_i$ is the fraction of the portfolio allocated to the risk-free asset.

5.1. Sample bounds

To bound the Rademacher complexity of (20), we can use an existing calculus for Rademacher complexity [1,14].

Proposition 8. *If $\|\xi\|_\infty\le s$, then the Rademacher complexity for (20) is bounded by $\mathcal{R}_n[f]\le(4s^2+\gamma s)\sqrt{\log p/n}$. Also, the assumption is satisfied for $\Delta=4s^2+\gamma s$.*

Proof. Using the identity $\Sigma=\mathbb{E}_\xi(\xi\xi^{\mathsf T})-\mu\mu^{\mathsf T}$, we have $x^{\mathsf T}\Sigma x-\gamma\cdot(\mu-r\mathbf{1})^{\mathsf T}x=\mathbb{E}_\xi\big((\xi^{\mathsf T}x)^2-\gamma\cdot\xi^{\mathsf T}x\big)-\big(\mathbb{E}_\xi\big(\xi^{\mathsf T}x\big)\big)^2+\gamma\cdot r\mathbf{1}^{\mathsf T}x$. Thus we can rewrite (20) as

$$\begin{aligned}&\min\ \mathbb{E}_\xi\big((\xi^{\mathsf T}x)^2-\gamma\cdot\xi^{\mathsf T}x\big)-\big(\mathbb{E}_\xi\big(\xi^{\mathsf T}x\big)\big)^2+\gamma\cdot r\mathbf{1}^{\mathsf T}x\\&\text{s.t. } x\ge 0,\ \|x\|_1\le 1\end{aligned}$$

The above is useful for bounding Rademacher complexity. Deterministic terms have a Rademacher complexity of zero, and the Rademacher complexity for the sum of problems is upper-bounded by the sum of the individual Rademacher complexities [1,14]. So we conclude the proof by bounding the Rademacher complexity of three problems: First, consider the problem $\min\{\mathbb{E}_\xi((\xi^{\mathsf T}x)^2)\mid x\ge 0,\|x\|_1\le 1\}$. Since $\|\xi\|_\infty\le s$, Proposition 5 says $\mathcal{R}_n[f_1]\le 2s^2\sqrt{\log p/n}$ with $\Delta_1=2s^2$, since $g(u)=u^2$ is Lipschitz with $L=2s$ when $u\in[-s,s]$. Second, consider the optimization problem $\min\{-(\mathbb{E}_\xi(\xi^{\mathsf T}x))^2\mid x\ge 0,\|x\|_1\le 1\}$. Proposition 5 with Corollary 1 gives $R_n[f_2]\le 2s^2\sqrt{\log p/n}$ with $\Delta_2=2s^2$, since $h(u)=u^2$ is Lipschitz with $L=2s$ when $u\in[-s,s]$. Third, consider the problem $\min\{\mathbb{E}_\xi(-\gamma\cdot\xi^{\mathsf T}x)\mid x\ge 0,\|x\|_1\le 1\}$. Proposition 5 gives $R_n[f_3]\le\gamma s\sqrt{\log p/n}$ with $\Delta_3=\gamma s$. The result follows by noting that $\mathcal{R}_n[f]\le\mathcal{R}_n[f_1]+\mathcal{R}_n[f_2]+\mathcal{R}_n[f_3]$ and that $\Delta\le\Delta_1+\Delta_2+\Delta_3$. □

We can compare various sample bounds on n to ensure (3) holds. We begin by calculating the specific previous bound from [28]: Hoeffding's lemma [7] bounds variance of (20) by $\frac{\Delta^2}{4}$ for the Δ from Proposition 8. Moreover, the Lipschitz constant of the objective (without expectation) in (20) is $L\le\sqrt{p}(4s^2+\gamma s)$. Consequently, the previous bound from [28] is

$$n\ge 2\cdot\left(\frac{4s^2+\gamma s}{\delta}\right)^2\cdot\left(p\log\frac{8\sqrt{p}(4s^2+\gamma s)}{\delta}+\log\frac{2}{\alpha}\right). \tag{21}$$

In contrast, combining Proposition 8 with Corollary 2 gives our bound to be

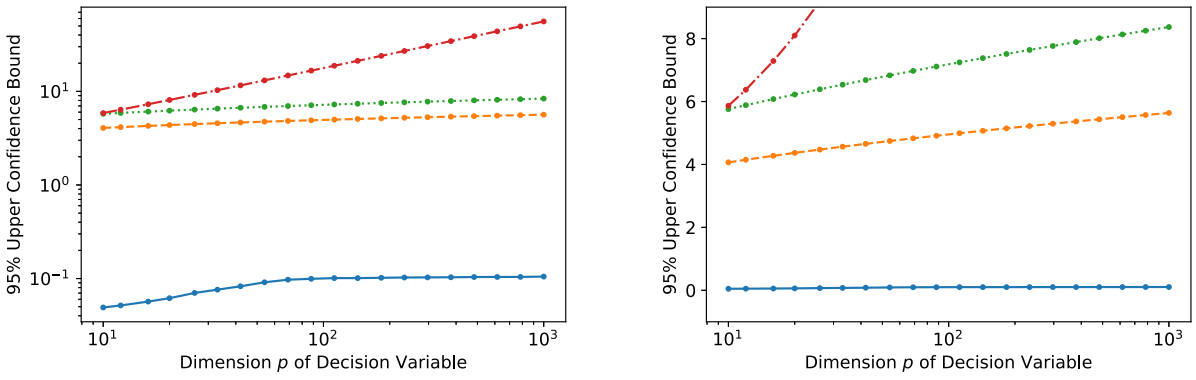
$$n\ge\left(\frac{16s^2+4\gamma s}{\delta}\right)^2\cdot\left(\frac{1}{2}\log\left(\frac{2}{\alpha}\right)+\log p\right). \tag{22}$$

The difference is our bound (22) is logarithmic in p whereas the past bound (21) is quasi-linear in p.

5.2. Results of experiment

To experimentally compare the above bounds with the actual SAA performance, we consider this next scenario: We assume returns for the assets are distributed as

$$\xi\sim\mathcal{U}^1(-\tfrac{s}{2},\tfrac{s}{2})\mathbf{1}_p+\mathcal{U}^p(-\tfrac{s}{2},\tfrac{s}{2}),$$



(a) log-log plot (b) semi-log plot

Fig. 1. Comparison of 95% upper confidence bound of SAA solution gap (solid blue) with bounds on 95% upper confidence bound gap predicted by [12,27,28] (dash-dotted red), our Proposition 2 (dashed orange), and our Corollary 2 (dotted green). The left shows results on a log–log scale, and the right shows results (excluding most of the [12,27,28] bound) on a semi-log scale. In both plots, the x-axis is the dimension p of the decision variable, and the y-axis is the 95% upper confidence bound gap.

where $\mathcal{U}^k(l,u)$ is a k-dimensional uniform distribution with the support of the distribution in each dimension of $[l,u]$, and $\mathbf{1}_p$ is a p-dimensional vector of ones. The interpretation is that the first term $\mathcal{U}^1(-\frac{s}{2},\frac{s}{2})\mathbf{1}_p$ describes a strongly correlated component of the returns, while the second term $\mathcal{U}^p(-\frac{s}{2},\frac{s}{2})$ describes an independent component of the returns. We also assume $\gamma = 1$, $s = 1$, and $r = 0$.

To compare the bounds with the actual SAA solution gaps, we used $n = 50$ and computed the 95% upper confidence bound of the SAA solution gap by solving SAA a total of 10,000 times each for different values of p. The 95% upper confidence bound is the smallest value of δ for $\alpha = 0.05$ in (3). We also compute the smallest value of δ for $\alpha = 0.05$ for the different bounds available to us. Specifically, we compare the actual upper confidence bound to

- bound (21), which is the past bound from [12,27,28]
- bound (9), which is the implicit sample bound from our Proposition 2
- bound (22), which is the simplified sample bound from our Corollary 2 for when $\lim_n \delta = 0$

The results are shown in Fig. 1. The past bound from [12,27,28] grows much faster than the actual SAA solution gap. In contrast, our bound (9) from Proposition 2 visually matches the growth rate of the actual SAA solution gap. Our bound (22), which is the simplified bound using Corollary 2, grows faster than the actual SAA solution gap; however, the bound (22) is a reasonably accurate approximation to (9) from Proposition 2. This suggests the simplified bound of Corollary 2 is useful for qualitative understanding of scaling, whereas the more accurate bound of Proposition 2 is more useful for determining necessary sample sizes for SAA.

Acknowledgments

The authors would like to thank Deepak Rajan for providing useful discussions and suggestions about this work. This material is based upon work supported by the National Science Foundation, USA under Grant CMMI-1847666.

References

[1] P.L. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results, J. Mach. Learn. Res. 3 (2002) 463–482.
[2] R. Bhatia, Matrix Analysis, Springer-Verlag, 1997.
[3] S. Boucheron, G. Lugosi, P. Massart, Concentration Inequalities: A Nonasymptotic Theory of Independence, Oxford University Press, 2013.
[4] B. Bruder, N. Gaussel, J.C. Richard, T. Roncalli, Regularization of portfolio allocation, 2013, Available at SSRN 2767358.
[5] J. Dupacová, R. Wets, Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems, Ann. Statist. 16 (1988) 1517–1549.
[6] V. Guigues, A. Juditsky, A. Nemirovski, Non-asymptotic confidence bounds for the optimal value of a stochastic program, Optim. Methods Softw. 32 (2017) 1033–1058.
[7] W. Hoeffding, Probability inequalities for sums of bounded random variables, JASA 58 (1963) 13–30.
[8] S.M. Kakade, S. Shalev-Shwartz, A. Tewari, Regularization techniques for learning with matrices, J. Mach. Learn. Res. 13 (2012) 1865–1890.
[9] S.M. Kakade, K. Sridharan, A. Tewari, On the complexity of linear prediction: Risk bounds, margin bounds, and regularization, in: NeurIPS, 2009, pp. 793–800.
[10] S. Kakade, A. Tewari, Rademacher composition and linear prediction, in: Lecture 17 notes for 'CMSC 35900 Learning Theory', 2008.
[11] S. Kim, R. Pasupathy, S.G. Henderson, A guide to sample average approximation, in: Handbook of Simulation Optimization, Springer, 2015, pp. 207–243.
[12] A.J. Kleywegt, A. Shapiro, T. Homem-de Mello, The sample average approximation method for stochastic discrete optimization, SIAM J. Optim. 12 (2002) 479–502.
[13] A. Kontorovich, Concentration in unbounded metric spaces and algorithmic stability, in: ICML, 2014, pp. 28–36.
[14] M. Ledoux, M. Talagrand, Probability in Banach Spaces: Isoperimetry and Processes, Springer-Verlag, 1991.
[15] H. Liu, C. Hernandez, H.Y. Lee, Regularized sample average approximation for high-dimensional stochastic optimization under low-rankness, 2019, arXiv preprint arXiv:1904.03453.
[16] H. Liu, X. Wang, T. Yao, R. Li, Y. Ye, Sample average approximation with sparsity-inducing penalty for high-dimensional stochastic programming, Math. Program. (2018) 1–40.
[17] J. Luedtke, S. Ahmed, A sample approximation approach for optimization with probabilistic constraints, SIAM J. Optim. 19 (2008) 674–699.
[18] H. Markowitz, Portfolio selection, J. Finance 7 (1952) 77–91.
[19] C. McDiarmid, On the method of bounded differences, Surv. Combin. 141 (1989) 148–188.
[20] Y. Mintz, A. Aswani, Polynomial-time approximation for nonconvex optimization problems with an L1-constraint, in: IEEE CDC, IEEE, 2017, pp. 682–687.
[21] R.I. Oliveira, P. Thompson, Sample average approximation with heavier tails i: non-asymptotic bounds with weak assumptions and stochastic constraints, 2017a, arXiv:1705.00822.
[22] R.I. Oliveira, P. Thompson, Sample average approximation with heavier tails ii: localization in stochastic convex optimization and persistence results for the lasso, 2017b, arXiv:1711.04734.
[23] J.O. Royset, On sample size control in sample average approximations for solving smooth stochastic programs, Comput. Optim. Appl. 55 (2013) 265–309.
[24] J.O. Royset, R. Szechtman, Optimal budget allocation for sample average approximation, Oper. Res. 61 (2013) 762–776.
[25] A. Ruszczyński, A. Shapiro, Stochastic programming, in: Handbooks in Operations Research and Management Science, Elsevier, 2003.

[26] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.
[27] A. Shapiro, Monte Carlo sampling methods, in: Handbooks in Operations Research and Management Science, Vol. 10, 2003, pp. 353–425.
[28] A. Shapiro, D. Dentcheva, A. Ruszczyński, Lectures on Stochastic Programming: Modeling and Theory, SIAM, 2009.
[29] V. Sudakov, Gaussian random processes and solid angle measures in Hilbert space, Dokl. Akad. Nauk SSSR 197 (1971) 43–45.
[30] B. Verweij, S. Ahmed, A.J. Kleywegt, G. Nemhauser, A. Shapiro, The sample average approximation method applied to stochastic routing problems: a computational study, Comput. Optim. Appl. 24 (2003) 289–333.