This article was downloaded by: [65.202.210.42] On: 14 March 2024, At: 09:05
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA


## Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Improved Sample-Complexity Bounds in Stochastic Optimization

Alok Baveja, Amit Chavan, Andrei Nikiforov, Aravind Srinivasan, Pan Xu

**To cite this article:**
Alok Baveja, Amit Chavan, Andrei Nikiforov, Aravind Srinivasan, Pan Xu (2023) Improved Sample-Complexity Bounds in Stochastic Optimization. Operations Research

Published online in Articles in Advance 17 Oct 2023

. https://doi.org/10.1287/opre.2018.0340

**Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions**

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2023, INFORMS

**Please scroll down for article—it is on subsequent pages**

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
*Articles in Advance*, pp. 1–9
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

Methods

# Improved Sample-Complexity Bounds in Stochastic Optimization

**Alok Baveja,[a] Amit Chavan,[b] Andrei Nikiforov,[c] Aravind Srinivasan,[d,*] Pan Xu[e]**

[a] Department of Supply Chain Management, Rutgers Business School, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854; [b] Product and Engineering, Snowflake Inc, San Mateo, California 94402; [c] School of Business, Rutgers, The State University of New Jersey, Camden, New Jersey 08102; [d] Department of Computer Science, University of Maryland, College Park, Maryland 20742; [e] Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey 07102

*Corresponding author

**Contact:** baveja@rutgers.edu, https://orcid.org/0000-0001-7228-7731 (AB); amitc@terpmail.umd.edu (AC); andnikif@camden.rutgers.edu (AN); asriniv1@umd.edu, https://orcid.org/0000-0002-0062-3684 (AS); pxu@njit.edu (PX)

**Received:** June 14, 2018
**Revised:** September 13, 2020; June 13, 2022; January 12, 2023
**Accepted:** April 29, 2023
**Published Online in Articles in Advance:** October 17, 2023

**Area of Review**: Optimization

**https://doi.org/10.1287/opre.2018.0340**

**Copyright:** © 2023 INFORMS

**Abstract.** Real-world network-optimization problems often involve uncertain parameters during the optimization phase. Stochastic optimization is a key approach introduced in the 1950s to address such uncertainty. This paper presents improved upper bounds on the number of samples required for the sample-average approximation method in stochastic optimization. It enhances the sample complexity of existing approaches in this setting, providing faster approximation algorithms for any method that employs this framework. This work is particularly relevant for solving problems like the stochastic Steiner tree problem.

**Funding:** The research of A. Baveja is partially supported by the United States Department of Transportation (via the University Transportation Research Center Program) [Grant 49198-25-26] and the British Council [the UKIERI Research Program]. The work of A. Chavan was done while he was a graduate student at the University of Maryland. The research of A. Srinivasan is partially supported by the National Science Foundation [Awards CNS 1010789, CCF-1422569, CCF-1749864, and CCF-1918749]; Adobe, Inc. [research awards]; Amazon, Inc.; and Google, Inc. The research of P. Xu was supported in part by the National Science Foundation [Awards CNS 1010789 and CCF-1422569 (when he was a graduate student)] and is partially funded by the National Science Foundation [CRII Award IIS-1948157].

**Supplemental Material:** The online appendix is available at https://doi.org/10.1287/opre.2018.0340.

**Keywords:** stochastic optimization • sample-average approximation • sample complexity

## 1. Introduction

Stochastic optimization deals with problems where there is uncertainty in the input; it aims to optimize or well approximate the expected value of an objective function that involves random input parameters. This area dates back to the classical works of Beale (1955) and Dantzig (1955) from the 1950s; we refer readers to books and surveys, including Ruszczynski and Shapiro (2003), Birge and Louveaux (2011), Shapiro et al. (2014), and the references therein, for more recent treatments of this topic. In multistage stochastic optimization, we postulate a probability distribution over the uncertain input parameters and compute a (two-stage or multistage) solution that optimizes the expected value of the objective function; the uncertain data are revealed over the two or more stages, and later stages may adaptively use the values revealed in earlier stages. This approach has been very fruitful for a range of problems and in areas including network design, inventory control, facility location, kidney exchange, and social-network operations. We highlight the works of Garg et al. (2008) and Gupta and Kumar (2009) as relevant representative examples of this paradigm. More recently, there has been an increasing number of new applications of online stochastic optimization in e-commerce: for example, online customer selection (Elmachtoub and Levi 2016), where customers arrive sequentially over each phase, and online scheduling (Wang and Truong 2018), where jobs of different priorities arrive randomly over time.

As a concrete example, consider two-stage uncapacitated facility location. Here, we are given a set of facilities $\mathcal{F}_0$ and a *superset* $\mathcal{C}_0$ of the possible set of clients. We only have stochastic information about the actual set $\mathcal{C}_1 \subseteq \mathcal{C}_0$ of clients that will materialize. In the face of this, we aim to open a suitable subset $\mathcal{F}_1 \subseteq \mathcal{F}_0$ of facilities in order to minimize the expected sum of the facility costs in $\mathcal{F}_1$ and the total distance of clients in $\mathcal{C}_1$ to their closest facility in $\mathcal{F}_1$ (see, e.g., Shmoys and Swamy 2006, Srinivasan 2007) in the *two-stage*

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

*recourse model*. $\mathcal{F}_1$ is chosen here in two stages. In the first stage, one constructs an initial set of facilities $\mathcal{F}_2$ based on distributional information about $\mathcal{C}_1$; in the second stage, $\mathcal{C}_1$ is sampled from its distribution, and we can augment $\mathcal{F}_2$ (take a "recourse action") to output a complete solution $\mathcal{F}_1$. The challenge is that such "rapid response" in the second stage is costly, and the tradeoff is between solution cost (lower in the first stage and higher in the second stage) and accuracy (precise in the second stage and probabilistic in the first stage). Chapters 1 and 2 of Birge and Louveaux (2011) present several examples and extensions of this framework.

In the facility-location example, suppose in general that we wish to judiciously build a "small" number of facilities that minimize some function of the total construction cost and the total connection cost of the clients—with the precise details varying based on the context. We could possibly assume in the first stage that each candidate client $j \in \mathcal{C}_0$ will be present *independently* of the others in $\mathcal{C}_1$ with a known probability $p_j$ or perhaps more realistically, that there is a black box that can sample $\mathcal{C}_1$ from its distribution and output any number of such samples as we desire in the first stage. Under distributional assumptions such as these two, the goal in two-stage stochastic facility location is to carefully provision $\mathcal{F}_2$ in the first stage followed by a near-optimal choice of additional facilities when $\mathcal{C}_1$ is revealed in the second stage.

### 1.1. Two-Stage Stochastic Program

We consider the following two-stage stochastic minimization program (Kleywegt et al. 2002, Ruszczynski and Shapiro 2003, Shapiro 2003, Charikar et al. 2005):

$$\min_{x \in X} f(x), \quad f(x) \doteq c(x) + \mathbb{E}_{\omega}[q(x,\omega)]. \tag{1}$$

By default, we assume that $X$ is finite; this naturally models discrete-optimization problems where our solution sets come from a finite family. Extensions to continuous $X$ are discussed in Section 7. In this model, a first-stage decision $x \in X$ has to be made while having only probabilistic information about the future, represented by the probability distribution $\pi$ of a random variable $\omega$ that is drawn from a set $\Omega$. Then, after a particular future scenario $\omega \in \Omega$ is realized probabilistically from the distribution $\pi$, a *recourse action* $r \in R$ may be taken to ensure that the requirements of the scenario $\omega$ are satisfied. In the two-stage model, $c(x)$ denotes the cost of taking the first stage action $x$. Given a particular scenario $\omega$ and a first-stage action $x$, the cost of the second stage $q(x,\omega)$ is represented as

$$q(x,\omega) = \min_{r \in R}\{\text{cost}_{\omega}(x,r) \mid (x,r) \text{ is a feasible solution for scenario } \omega\}, \tag{2}$$

where $\text{cost}_{\omega}(x,r)$ denotes the second-stage cost in the scenario $\omega$ with a first-stage action $x$ and a second-stage action $r$ and where $R$ is the set of all possible actions in the second stage. The optimization program (2) can sometimes be solved optimally for each given pair $(x,\omega)$; as explained in Charikar et al. (2005), their approach can be extended to the case where we only have an approximate—not exact—minimizer, and this property holds for our results as well.

A natural approach to solve problems modeled by (1) is to take some number $N$ of independent samples $\omega_1, \ldots, \omega_N$ from the distribution $\pi$ and to approximate $f$ by the sample-average function

$$\hat{f}(x) = c(x) + \frac{1}{N}\sum_{i=1}^{N} q(x,\omega_i). \tag{3}$$

Throughout this paper, we view $\hat{f}(x)$ as a function of $x$ that is parameterized by the number $N$ of independent samples $\omega_1, \ldots, \omega_N$. One might wish to argue that for a suitably chosen sample size $N$, a good solution $\hat{x}$ to (3) would be a good solution to $f$; more precisely, we define $\hat{x} \in X$ to be an *$\alpha$-approximate minimizer* of the function $f$ defined in (1) if

$$\text{for all } x \in X,\ f(\hat{x}) \le \alpha f(x).$$

This is the powerful *sample-average approximation* (SAA) method (Kleywegt et al. 2002; Ruszczynski and Shapiro 2003; Shapiro 2003; Shmoys and Swamy 2004, 2006; Charikar et al. 2005), which we discuss further in Section 3.

As we will see, we significantly reduce the sample size $N$ of Charikar et al. (2005) to get the same approximation precision; see Theorems 1 and 2, with the relevant parameters defined in Section 3.

## 2. Related Work and an Example of Our Improvement

Several techniques have been proposed to address emerging computational challenges in different variants of stochastic optimization; the following are a few examples. Russo and Roy (2018) propose information-directed sampling, an approach to online optimization problems in which a decision maker must balance between exploration and exploitation while learning from partial feedback. Pu and Garcia (2018) present a flocking-based approach for distributed stochastic optimization, whereas Besbes et al. (2015) consider a nonstationary variant of sequential stochastic optimization and establish a strong connection between adversarial online convex optimization and the traditional stochastic-approximation paradigm. See, for example, Immorlica et al. (2004), Ravi and Sinha (2006), Shmoys and Swamy (2006), Gupta et al. (2007), Levi et al. (2007), Srinivasan (2007), Dean et al. (2008), Chen et al. (2009), Bansal et al. (2012), Abolhassani et al. (2015), Dieker et al. (2016), Goyal et al. (2016), Jaillet et al. (2016), Papier (2016), and Raginsky and Nedić (2016) for diverse applications of (multistage) stochastic optimization.

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

Our improved sample size $N$ leads to much faster network-optimization algorithms in terms of worst-case time, such as in some of the examples listed (Gupta et al. 2004, Srinivasan 2007, Garg et al. 2008, Gupta and Kumar 2009, Abolhassani et al. 2015); these include customer service under fluctuating prices, stochastic online combinatorial optimization (such as stochastic online Steiner tree), stochastic Steiner forest, stochastic vertex cover, and stochastic facility location.

As an example, the work of Gupta and Kumar (2009) considers the *stochastic Steiner forest* problem, a fundamental network-design problem with two-stage stochastic uncertainty about the input. We start with an empty graph over a set of vertices $V$; the allowed set of edges is given by the set of edges of a graph $G = (V, E)$. The stochastic input is some (random, yet unknown) set $D$ of pairs of vertices that need to be connected by a path in our network design. The network design proceeds in two stages. In the first stage, we can install any edge $e \in E$ in our to-be-designed graph; this will cost us a given amount $c_e \geq 0$. Then, in the second stage, the set $D$ is drawn from a given distribution $\pi$ (which is available as a black box from which we can sample); we can wait to install edges in this second stage, but each edge $e$ now costs $\lambda \cdot c_e$ for some given inflation factor $\lambda \geq 1$. The goal is thus to install some edges initially and then others later after the set of pairs of vertices $D$ is realized, so that the total expected network-design cost is minimized.

The SAA method is naturally applied for the stochastic Steiner forest problem in Gupta and Kumar (2009), and the resulting run time depends at least linearly on $N$. Specifically, we mean the following. The work of Gupta and Kumar (2009) samples sets $D_1, D_2, \ldots, D_N$ (each $D_i$ being a set of pairs of vertices) independently from the distribution $\pi$. It dispenses with $\pi$ from now on by replacing $\pi$ by the uniform distribution over the explicitly given sets $D_1, D_2, \ldots, D_N$. These explicit $D_i$'s now become the input to the problem, for which the run time of Gupta and Kumar (2009) is at least linear in the input size and specifically at least linear in $N$ (i.e., at least $cN$ for some constant $c > 0$). In fact, it is (substantially) superlinear in $N$ because this input is fed into a linear program whose number of variables is linear in $N$; solving this linear program takes the bulk of the running time of the overall algorithm of Gupta and Kumar (2009). Therefore, by significantly reducing $N$, we are able to substantially speed up this algorithm. More importantly, a similar comment applies to any other such application, such as those of Gupta et al. (2004), Srinivasan (2007), Garg et al. (2008), and Abolhassani et al. (2015). Because stochasticity is ubiquitous in discrete optimization, we anticipate similar theoretical improvements in network optimization in the future.

## 3. Main Results

We significantly improve upon the sample complexity of Charikar et al. (2005) for stochastic optimization in the black-box model via the SAA method; please recall the SAA approach of using (3) in the framework of (1) and (2). A key question regarding the implementation of this recipe is how many samples $N$ suffice as a function of, for example, the desired accuracy $1 - \epsilon$ and the confidence $1 - \delta$. That is, given $\epsilon$ and $\delta$, how large an $N$ will suffice for an optimal solution to $\hat{f}$ defined in Equation (3) to be an $(1 + \epsilon)$-approximate minimizer of the function $f$ defined in Equation (1), with probability at least $1 - \delta$?

The work of Charikar et al. (2005)—which has further useful properties discussed after Theorem 2—considered this setting with the following additional natural properties.

**Property 1** (Nonnegativity). *$c(x) \geq 0$ and $q(x, \omega) \geq 0$, for each $x \in X$ and $\omega \in \Omega$.*

**Property 2** (Empty First-Stage Action). *We assume there is an empty first-stage action $0 \in X$, with $c(0) = 0$ and $q(x, \omega) \leq q(0, \omega)$, for each $x \in X$, $\omega \in \Omega$. Thus, the empty action $0 \in X$ is the cheapest first-stage solution whose second-stage completion (with respect to any realization) is the most expensive.*

**Property 3** (Bounded Inflation Factor). *There is a given $\lambda \geq 1$ such that for all $x \in X, \omega \in \Omega$, we have $q(0, \omega) - q(x, \omega) \leq \lambda c(x)$. In other words, the gap of the second-stage cost between a specific choice $x \in X$ and the empty choice 0 is upper bounded by $\lambda c(x)$.*

Recall our assumption that the first-stage decision set $X$ is finite. In this setting, the result of Charikar et al. (2005) is as follows.

**Theorem 1** (Charikar et al. 2005). *There is a constant $K_0 > 0$ such that the following holds. Any exact minimizer $\overline{x}$ of the function $\hat{f}$ defined in Equation (3) constructed with $N = \lceil K_0 \cdot (\lambda^2/\epsilon^4) \cdot \ln|X| \cdot \ln(1/\delta) \rceil$ samples is, with probability at least $1 - \delta$, a $(1 + \epsilon)$-approximate minimizer of the function $f$ defined in Equation (1).*

Our result of improved sample complexity over Theorem 1 is as follows.

**Theorem 2.** *There is a constant $K_1 > 0$ such that the following holds. Any exact minimizer $\overline{x}$ of the function $\hat{f}$ defined in Equation (3) constructed with $N = \lceil K_1 \cdot \ln(|X|/\delta) \cdot \max[(\lambda^2/\epsilon^2), (\lambda/\epsilon^3)] \rceil$ samples is, with probability at least $1 - \delta$, a $(1 + \epsilon)$-approximate minimizer of the function $f$ defined in Equation (1).*

Observe that our sample complexity has a significantly lower dependence over $\lambda$ and $\epsilon$ compared with that in Charikar et al. (2005).

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

### 3.1. Sample Complexity in Related Models

When the three properties (Properties 1–3) are absent, the sample complexity of Kleywegt et al. (2002) includes as a parameter the largest variance of $(q(x^*, \omega) - q(x, \omega))$ over all possible $x \in X$, where $x^*$ is an optimal solution in Equation (1). From Property 3, we see that $|q(x^*, \omega) - q(x, \omega)| \le \lambda(c(x) + c(x^*))$. Thus, $\max_{x \in X} \mathrm{Var}[q(x^*, \omega) - q(x, \omega)] \le 2\lambda \cdot \max_{x \in X} c(x)$. It is not clear how to remove the dependence of the final complexity on $\max_{x \in X} c(x)$ when applying the approach of Kleywegt et al. (2002). The results of Theorems 1 and 2 enjoy the advantage that their sample complexities are *independent* of this largest-variance parameter; our sample complexity is bounded as stated in Theorem 2 as long as $|X|$ and $\lambda$ are bounded, regardless of the largest-variance parameter. A detailed comparison of our work with Kleywegt et al. (2002) can be seen in the online appendix.

**3.1.1. Further Useful Properties.** As mentioned in Charikar et al. (2005), their work has some useful additional properties. One is that although the works of Shmoys and Swamy (2004, 2006) require the underlying space $X$ to be continuous and amenable to algorithms such as the ellipsoid method, the approach of Charikar et al. (2005) works in the more general discrete case. (Under mild Lipschitz-type assumptions on the functions $c(\cdot)$ and $q(\cdot, \cdot)$, a simple discretizing approach can reduce the continuous case to the discrete case (Kleywegt et al. 2002, Charikar et al. 2005).) Thus, as mentioned in Charikar et al. (2005), the approaches of Kleywegt et al. (2002), Shapiro (2003), and Charikar et al. (2005) only rely on *statistical* properties of $f$ and $X$ and not—as in Shmoys and Swamy (2004, 2006)—on the *computational* property of the amenability of optimization over $X$. As mentioned before and as explained in Charikar et al. (2005), their approach can be extended to the case where we only have an approximate—not exact—minimizer for $\hat{f}$. All of these useful properties carry over directly to our approach as well.

The black-box model considered in Charikar et al. (2005) is quite general, and an improved sample complexity translates to more efficient implementations of the several applications of the work of Charikar et al. (2005); see, for example, Srinivasan (2007), Garg et al. (2008), Gupta and Kumar (2009), and Abolhassani et al. (2015). To the best of our knowledge, we are not aware of any strong lower bounds on the sample complexity in this model other than standard lower bounds from sampling of the form $\Omega((1/\epsilon^2)\ln(1/\delta))$ (Canetti et al. 1995).

## 4. Preliminaries and Main Techniques

In this paper, we use the following form of the Chernoff bound.

**Theorem 3** (Chernoff Bound (Tarjan 2009)). *Let $X_1, \ldots, X_n$ be $n$ independent random variables with $0 \le X_i \le 1$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathsf{E}[X]$. Then, for any $\epsilon > 0$,*

$$\Pr[X \ge (1+\epsilon)\mu] \le \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right). \tag{4}$$

We specialize the version to the lemma, which will be quite useful in our paper. It is a specialization for independent and identically distributed random variables; it also parametrizes the deviation above the mean in an additive fashion.

**Lemma 1.** *Let $X_1, \ldots, X_n$ be $n$ independent identical distributed random variables with each $0 \le X_i \le 1$ and $\mathsf{E}[X_i] = \mu_X$. Let $\overline{X} = (\sum_{i=1}^n X_i)/n$. Then, for any value $v > 0$,*

$$\Pr[\overline{X} \ge \mu_X + v] \le \exp\left(-\frac{nv^2}{v + 2\mu_X}\right). \tag{5}$$

Lemma 1 follows directly from Theorem 3, and its proof can be seen in the appendix.

We first set a threshold $M_1$ and partition the set of all scenarios $\omega_1, \omega_2, \ldots$ into two subgroups, one called *high* and the other called *low*. Then, we decompose the original and empirical second-stage costs each into two sums based on the two subgroups. The approach in detail is as follows. The framework of our analysis is similar to that of Charikar et al. (2005).

Let $Z^* = f(x^*)$, where $x^*$ is a minimizer for the function $f$ defined in (1). Similar to Charikar et al. (2005), we introduce a threshold $M_1 = \lambda Z^*/\epsilon$ to divide the set of scenarios into two classes. We call a scenario $\omega$ *high* if $q(0, \omega) \ge M_1$ and *low* otherwise. Suppose we have $N$ independent samples $\omega_1, \ldots, \omega_N$ from the distribution $\pi$. Let $\hat{f}(x) = c(x) + \hat{f}_l(x) + \hat{f}_h(x)$, where

$$\hat{f}_l(x) = \frac{1}{N} \sum_{\omega_i:\ \omega_i \text{ is low}} q(x, \omega_i), \quad \hat{f}_h(x) = \frac{1}{N} \sum_{\omega_i:\ \omega_i \text{ is high}} q(x, \omega_i). \tag{6}$$

Let $f_l(x) = \mathsf{E}[\hat{f}_l(x)]$ and $f_h(x) = \mathsf{E}[\hat{f}_h(x)]$, where the expectation is taken over the $N$ independent samples $\omega_1, \ldots, \omega_N$ from $\pi$. Therefore, $f(x) = c(x) + f_l(x) + f_h(x)$. Let $p = \Pr[\omega \text{ is high}]$. We have

$$\begin{aligned} f_l(x) &= \mathsf{E}[\hat{f}_l(x)] = \mathsf{E}[q(x, \omega) | \omega \text{ is low}] \cdot (1 - p), \\ f_h(x) &= \mathsf{E}[\hat{f}_h(x)] = \mathsf{E}[q(x, \omega) | \omega \text{ is high}] \cdot p. \end{aligned} \tag{7}$$

Recall that $x^*$ and $\overline{x}$ are exact minimizers for $f(x)$ and $\hat{f}(x)$, respectively. Let $A = \hat{f}_l(x^*) - f_l(x^*) + f_l(\overline{x}) - \hat{f}_l(\overline{x})$ and $B = \hat{f}_h(x^*) - \hat{f}_h(\overline{x}) + f_h(\overline{x}) - f_h(x^*)$. We prove that $f(\overline{x}) - f(x^*) \le A + B$ (with probability 1) and that $A + B = O(\epsilon f(x^*) + \epsilon f(\overline{x}))$ with high probability. Specifically, we show that the sampling bound $N$ required for $A$ to

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

be $O(\epsilon f(x^*) + \epsilon f(\overline{x}))$ strictly dominates that for $B$ to be $O(\epsilon f(x^*) + \epsilon f(\overline{x}))$, and we significantly improve the former dominant part compared with that in Charikar et al. (2005) by conducting a tighter analysis.

Let $A = A_1 + A_2$, where $A_1 = \hat{f}_l(x^*) - f_l(x^*)$ and $A_2 = f_l(\overline{x}) - \hat{f}_l(\overline{x})$. We state the main sampling bounds in the three claims, the proofs of which can be seen in Section 6.

**Claim 1.** $\Pr[A_1 \le \epsilon Z^*] \ge 1 - \delta$ when $N = \Omega(\ln(1/\delta) \cdot (\lambda/\epsilon^3))$.

**Claim 2.** $\Pr[A_2 \le \epsilon Z^*] \ge 1 - \delta$ when $N = \Omega(\ln(|X|/\delta) \cdot \max[(\lambda^2/\epsilon^2), (\lambda/\epsilon^3)])$.

**Claim 3.** $\Pr[B \le 2\epsilon f(\overline{x}) + \epsilon Z^*] \ge 1 - \delta$ when $N = \Omega((\lambda/\epsilon) \cdot \ln(1/\delta))$.

**Remark 1** (Remark on Claim 1). Note that to get the same probability result as in Claim 1, Charikar et al. (2005) require the sampling bound to be $\Omega(\ln(1/\delta) \cdot (\lambda^2/\epsilon^4))$. The main difference is that they used a weaker version of the Chernoff bound, as shown in lemma 1 of Charikar et al. (2005), than what we use here as shown in Lemma 1. They considered the case $v = \epsilon$ and gave an upper bound of $\exp(-n\epsilon^2)$, which nearly matches the result of Lemma 1 if we set $\mu = 1$. The point is that in our context, both $\mu$ and $v$ are very small, which helps us get a much stronger upper bound when using Lemma 1.

**Remark 2** (Remark on Claim 2). In order to get the same probability result as does Claim 2, Charikar et al. (2005) require the sampling bound to be $\Omega(\ln(1/\delta) \cdot \ln|X| \cdot (\lambda^2/\epsilon^4))$. To prove Claim 2, we need to upper bound the lower-tail probability of the sum of $N$ independent and identically distributed (i.i.d.) random variables, where each has a mean at most $\lambda Z^*$ and takes a value between $[0, M_1]$ with $M_1 = \lambda Z^*/\epsilon$. Our improvement is because of a careful examination of the mass distribution for each random variable: After introducing a new threshold $M_2 = 2\lambda c(x^*)$, we find that the total mass distributed over the range $[M_2, M_1]$ is at most $2Z^*$, which is a negligible fraction of the total mean of $\lambda Z^*$ as assumed in the worst case. This helps us obtain a much-improved worst-case scenario when upper bounding the lower-tail probability.

## 5. Proof of Theorem 2

In this section, we show how Claims 1–3 yield Theorem 2.

### 5.1. A Remark on $\epsilon$ and $\delta$

As in the proof of theorem 1 in Charikar et al. (2005), we will prove Theorem 2 by working with constant multiples $C_1\epsilon$ and $C_2\delta$ of $\epsilon$ and $\delta$, respectively, for convenience rather than $\epsilon$ and $\delta$ themselves. That is, for some absolute positive constants $C_1$ and $C_2$, we aim to show that any exact minimizer $\overline{x}$ of the function $\hat{f}$ defined in (3) constructed with

$$\Omega\left(\ln\frac{|X|}{\delta} \cdot \max\left[\frac{\lambda^2}{\epsilon^2}, \frac{\lambda}{\epsilon^3}\right]\right)$$

samples is, with probability at least $1 - C_1\delta$, a $(1 + C_2\epsilon)$-approximate minimizer of the function $f$ defined in (1). This would of course imply Theorem 2 by choosing the constant $K_1$ appropriately.

**Proof of Theorem 2.** Recall that $\overline{x}$ is an exact minimizer of the function $\hat{f}$ defined in (3):

$$\begin{aligned}
f(\overline{x}) &= c(\overline{x}) + f_l(\overline{x}) + f_h(\overline{x}) \\
&= \underbrace{c(\overline{x}) + \hat{f}_l(\overline{x}) + \hat{f}_h(\overline{x})}_{\hat{f}(\overline{x})} + f_l(\overline{x}) + f_h(\overline{x}) - \hat{f}_l(\overline{x}) - \hat{f}_h(\overline{x}) \\
&= \hat{f}(\overline{x}) + f_l(\overline{x}) + f_h(\overline{x}) - \hat{f}_l(\overline{x}) - \hat{f}_h(\overline{x}) \\
&\le \hat{f}(x^*) + f_l(\overline{x}) + f_h(\overline{x}) - \hat{f}_l(\overline{x}) - \hat{f}_h(\overline{x}) \\
&\quad \text{(because of the optimality of } \overline{x} \text{ for } \hat{f}(x)) \\
&= c(x^*) + \hat{f}_l(x^*) + \hat{f}_h(x^*) + f_l(\overline{x}) + f_h(\overline{x}) - \hat{f}_l(\overline{x}) - \hat{f}_h(\overline{x}) \\
&= \underbrace{c(x^*) + f_l(x^*) + f_h(x^*)}_{f(x^*)} + \underbrace{\hat{f}_l(x^*) - f_l(x^*) + f_l(\overline{x}) - \hat{f}_l(\overline{x})}_{A} \\
&\quad + \underbrace{\hat{f}_h(x^*) - f_h(x^*) + f_h(\overline{x}) - \hat{f}_h(\overline{x})}_{B} \\
&= f(x^*) + A + B.
\end{aligned}$$

Recall that $A = A_1 + A_2$:

$$\begin{aligned}
&\Pr[f(\overline{x}) - f(x^*) \ge 3\epsilon Z^* + 2\epsilon f(\overline{x})] \\
&\le \Pr[A + B \ge 3\epsilon Z^* + 2\epsilon f(\overline{x})] \\
&\le \Pr[A_1 \ge \epsilon Z^*] + \Pr[A_2 \ge \epsilon Z^*] + \Pr[B \ge 2\epsilon f(\overline{x}) + \epsilon Z^*] \\
&\le 3\delta \quad \text{(because of Claims 1, 2, and 3),}
\end{aligned}$$

which implies that $\Pr[f(\overline{x}) \le f(x^*) \cdot (1 + 3\epsilon)/(1 - 2\epsilon)] \ge 1 - 3\delta$. Thus, we get our result. □

## 6. Proofs of Claims 1, 2, and 3

### 6.1. Auxiliary Lemmas

Before presenting the proofs of the three claims, we first show three auxiliary lemmas. Recall that $Z^* = f(x^*)$ is the optimal value of $f(x)$, the first threshold $M_1 = \lambda Z^*/\epsilon$, and $p = \Pr[q(0, \omega) \ge M_1]$ denotes the probability of a sample $\omega$ being high in the distribution $\pi$. Let $M_2 = 2\lambda c(x^*)$ and $\pi(\omega)$ be the probability associated with the outcome $\omega \in \Omega$ in $\pi$. Suppose that we have $N$ independent samples $\omega_1, \ldots, \omega_N$ from the distribution $\pi$.

**Lemma 2.**

1. $p \le (\epsilon/\lambda)$.
2. $f(0) \le \lambda Z^*$.
3. *For each* $x \in X$, $\sum_{\omega \in \Omega : q(x,\omega) > M_2} q(x, \omega) \cdot \pi(\omega) \le 2(Z^* - c(x^*))$.

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

**Proof of Lemma 2.**

1. By Property 3, $q(0,\omega) \le q(x^*,\omega) + \lambda c(x^*)$ for every $\omega$. Thus,

$$p = \Pr[q(0,\omega) \ge M_1] \le \Pr[q(x^*,\omega) + \lambda c(x^*) \ge M_1] = \Pr[q(x^*,\omega) \ge M_1 - \lambda c(x^*)].$$

Recall that $Z^* = f(x^*) = c(x^*) + \mathsf{E}[q(x^*,\omega)]$, which suggests that $\mathsf{E}[q(x^*,\omega)] = Z^* - c(x^*)$. Note that $q(x^*,\omega) \ge 0$ for all $\omega$ by Property 1. By Markov's inequality,

$$p \le \frac{Z^* - c(x^*)}{M_1 - \lambda c(x^*)} = \frac{Z^* - c(x^*)}{\lambda Z^*/\epsilon - \lambda c(x^*)} = \frac{\epsilon}{\lambda} \cdot \frac{Z^* - c(x^*)}{Z^* - \epsilon c(x^*)} \le \frac{\epsilon}{\lambda}.$$

2. Note that

$$\begin{aligned} f(0) &= c(0) + \mathsf{E}[q(0,\omega)] = \mathsf{E}[q(0,\omega)] \\ &\le \mathsf{E}[q(x^*,\omega)] + \lambda c(x^*) \le \lambda(\mathsf{E}[q(x^*,\omega)] + c(x^*)) \\ &= \lambda Z^*, \end{aligned}$$

where the second equality follows from $c(0) = 0$ by Property 2, the first inequality is because of Property 3, and the second inequality is valid because $\lambda \ge 1$.

3. Recall that $M_2 = 2\lambda c(x^*)$. For each $x \in X$, let $\Omega_x = \{\omega : q(x,\omega) > M_2\}$. Consider a given $\omega \in \Omega_x$ with $q(x,\omega) > M_2$. From Property 2, we have that $q(0,\omega) \ge q(x,\omega) > M_2$, which implies that $\omega \in \Omega_0$. Thus, we have $\Omega_x \subseteq \Omega_0$ for all $x$. Observe that

$$\begin{aligned} Z^* - c(x^*) &= \mathsf{E}[q(x^*,\omega)] && \text{(by definition of } f(x^*)) \\ &\ge \sum_{\omega \in \Omega_0 q} (x^*,\omega)\pi(\omega) \\ &\ge \sum_{\omega \in \Omega_0} (q(0,\omega) - \lambda c(x^*))\pi(\omega) && \text{(Property 3)} \\ &\ge \frac{1}{2}\sum_{\omega \in \Omega_0} q(0,\omega)\pi(\omega) && \text{(by definition of } \Omega_0) \\ &\ge \frac{1}{2}\sum_{\omega \in \Omega_x} q(x,\omega)\pi(\omega) && \text{(Property 2 and } \Omega_x \subseteq \Omega_0). \end{aligned}$$

We thus have that $\sum_{\omega \in \Omega_x} q(x,\omega)\pi(\omega) \le 2(Z^* - c(x^*))$, which yields our result. □

**Lemma 3.** *Suppose we are given a convex function $g(x)$ over $[a,b]$ with $0 \le a < b$ and a value $\mu \in [a,b]$. Consider the maximization of $\mathsf{E}[g(X)]$ subject to* (i) $a \le X \le b$ *and* (ii) $\mathsf{E}[X] = \mu$; *the maximum here can be realized when $X$ follows a two-pointed distribution, where $X$ takes values $a$ and $b$ with respective probabilities $(b-\mu)/(b-a)$ and $(\mu-a)/(b-a)$.*

The proof of Lemma 3 follows directly from the convexity property of $g(x)$; see the details in the appendix. We apply Lemma 3 to prove the lemma, which is useful in Claim 2.

**Lemma 4.** *Let $X$ be a random variable satisfying the following three conditions*: (1) $0 \le a \le X \le b$, (2) $\mathsf{E}[X] = \mu$, *and* (3) $\mathsf{E}[X|X \in (c,b]] \cdot \Pr[X \in (c,b]] \le \mu'$ *for some $c \in [a,b]$. Then, for any convex function $g(x)$ over $[a,b]$, there exists a random variable $Y$ such that* (i) $\mathsf{E}[g(X)] \le \mathsf{E}[g(Y)]$; (ii) $Y$ *also satisfies conditions* (1), (2), *and* (3); *and* (iii) $Y$*'s support is a subset of, or is equal to, the set* $\{a, b, c\}$.

**Proof of Lemma 4.** Consider a given $X$ satisfying the three conditions. Let $X_1 \doteq (X|X \in [a,c])$ and $X_2 \doteq (X|X \in (c,b])$, respectively, denote the conditional random variables of $X$ when restricted to the ranges $[a,c]$ and $(c,b]$, respectively. Let $p_1 = \Pr[X \in [a,c]]$, $p_2 = \Pr[X \in (c,b]]$, $\mu_1 = \mathsf{E}[X_1]$, and $\mu_2 = \mathsf{E}[X_2]$. Thus, $p_1 + p_2 = 1$, $p_1\mu_1 + p_2\mu_2 = \mu$, and $p_2\mu_2 \le \mu'$. By Lemma 3, we see that $\mathsf{E}[g(X_1)] \le \mathsf{E}[g(Y_1)]$ for some $Y_1$, where $\mathsf{E}[Y_1] = \mathsf{E}[X_1] = \mu_1$ and where $Y_1$ has nonzero probabilities at $a$ and $c$ only, say $p_a$ and $p_c$, respectively.

Consider the maximization of $\mathsf{E}[g(X_2)]$ subject to $c < X_2 \le b$ and $\mathsf{E}[X_2] = \mu_2$. We see that the maximum value will not decrease if we relax the range of $X_2$ to $[c,b]$. From Lemma 3, we see that $\mathsf{E}[g(X_2)] \le \mathsf{E}[g(Y_2)]$ for some $Y_2$, where $\mathsf{E}[Y_2] = \mu_2$, and $Y_2$ has nonzero probabilities at $c$ and $b$ only, say $p_c'$ and $p_b$, respectively. Construct a random variable $Y$ that is distributed on $\{a, b, c\}$ as follows:

$$\Pr[Y = a] = p_1 p_a,\ \Pr[Y = c] = p_1 p_c + p_2 p_c',\ \Pr[Y = b] = p_2 p_b.$$

Recall that $\mathsf{E}[Y_2] = \mu_2 = p_c' \cdot c + p_b \cdot b$, which implies that $p_b \cdot b \le \mu_2$ because $c \ge a \ge 0$. Thus, $\mathsf{E}[Y|Y \in (c,b]] \cdot \Pr[Y \in (c,b]] = (p_2 p_b)b \le p_2\mu_2 \le \mu'$ (i.e., $Y$ meets condition (3)). We can verify that $Y$ satisfies conditions (1) and (2) as well. Also,

$$\begin{aligned} \mathsf{E}[g(X)] &= p_1\mathsf{E}[g(X_1)] + p_2\mathsf{E}[g(X_2)] \\ &\le p_1\mathsf{E}[g(Y_1)] + p_2\mathsf{E}[g(Y_2)] = \mathsf{E}[g(Y)]. \end{aligned}$$

This completes the proof. □

## 6.2. Proof of Claim 1

**Proof of Claim 1.** Recall that by definition, $A_1 = \hat{f}_l(x^*) - f_l(x^*)$, $f_l(x^*) = \mathsf{E}[\hat{f}_l(x^*)]$, and $M_1 = \lambda Z^*/\epsilon$. By definition of $\hat{f}_l(x)$ in (6), we can rewrite $\hat{f}_l(x^*)$ as $\hat{f}_l(x^*) = \left(\sum_{i=1}^N Q_i\right)/N$, with $Q_i = q(x^*,\omega_i)$ if $q(0,\omega_i) < M_1$ (i.e., if $\omega_i$ is low) and $Q_i = 0$ otherwise. Thus, we can view $\hat{f}_l(x^*)/M_1$ as a sum of $N$ i.i.d. random variables where (i) each summand takes values in $[0,1]$ because $q(x^*,\omega_i)/M_1 \le q(0,\omega_i)/M_1 < 1$ when $\omega_i$ is low and where (ii) $\hat{f}_l(x^*)/M_1$ has mean $\mathsf{E}[\hat{f}_l(x^*)/M_1] = f_l(x^*)/M_1 \le Z^*/M_1 = \epsilon/\lambda$. By Lemma 1, we have

$$\begin{aligned} \Pr[A_1 \ge \epsilon Z^*] &= \Pr[\hat{f}_l(x^*) - f_l(x^*) \ge \epsilon Z^*] \\ &= \Pr\left[\frac{\hat{f}_l(x^*)}{M_1} - \frac{f_l(x^*)}{M_1} \ge \frac{\epsilon^2}{\lambda}\right] \\ &\le \exp\left(-\frac{N\epsilon^4/\lambda^2}{2\epsilon/\lambda + \epsilon^2/\lambda}\right) \le \exp\left(-\frac{N}{3}\frac{\epsilon^3}{\lambda}\right). \end{aligned}$$

By setting the last expression equal to $\delta$, we get our claim that when $N = \Omega\left(\ln\frac{1}{\delta} \cdot \frac{\lambda}{\epsilon^3}\right)$, $\Pr[A_1 \le \epsilon Z^*] \ge 1 - \delta$. □

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

### 6.3. Proof of Claim 2

**Proof of Claim 2.** Assume for any given $x \in X$, when $N = \Omega(\ln(1/\delta) \cdot \max[(\lambda^2/\epsilon^2), (\lambda/\epsilon^3)])$,

$$\Pr[f_l(x) - \hat{f}_l(x) > \epsilon Z^*] \le \delta. \tag{8}$$

We first show how Inequality (8) implies Claim 2. Replacing $\delta$ with $\delta/|X|$, we have that when $N = \Omega(\ln |X|/\delta \cdot \max[(\lambda^2/\epsilon^2), (\lambda/\epsilon^3)])$, $\Pr[f_l(x) - \hat{f}_l(x) > \epsilon Z^*] \le \delta/|X|$ for any $x \in X$. Thus, it follows that

$$\begin{aligned}\Pr[f_l(\bar{x}) - \hat{f}_l(\bar{x}) > \epsilon Z^*] &\le \Pr[\exists x \in X : f_l(x) - \hat{f}_l(x) > \epsilon Z^*] \\ &\le \sum_{x \in X} \Pr[f_l(x) - \hat{f}_l(x) > \epsilon Z^*] \\ &\le \delta \text{ (because of the union bound).}\end{aligned}$$

Therefore, we have $\Pr[f_l(\bar{x}) - \hat{f}_l(\bar{x}) \le \epsilon Z^*] = \Pr[A_2 \le \epsilon Z^*] \ge 1 - \delta$, and we get Claim 2.

Now, we prove Inequality (8). Consider a fixed first-stage action $x \in X$. We can view $\hat{f}_l(x)$ as the arithmetic mean of $N$ i.i.d. random variables $Q_1, \ldots, Q_N$, where

$$Q_i = \begin{cases} q(x, \omega_i) & \text{if } \omega_i \text{ is low,} \\ 0 & \text{otherwise.} \end{cases}$$

For notational convenience, let $\mathsf{E}[\hat{f}_l(x)] = f_l(x) = \mu$, $\epsilon' = \epsilon Z^*/\mu$, and $Q = \sum_{i=1}^N Q_i = N\hat{f}_l(x)$. Let $t > 0$ be a parameter to be specified later:

$$\begin{aligned}&\Pr[f_l(x) - \hat{f}_l(x) > \epsilon Z^*] = \Pr[\hat{f}_l(x) - f_l(x) < -\epsilon Z^*] \\ &= \Pr[\hat{f}_l(x) < \mu(1-\epsilon')] = \Pr[Q < (1-\epsilon')N\mu] \\ &= \Pr[e^{-tQ} > e^{-(1-\epsilon')\cdot N\cdot\mu\cdot t}] \le \frac{\mathsf{E}[e^{-tQ}]}{e^{-(1-\epsilon')\cdot N\cdot\mu\cdot t}} \\ &\quad \text{(because of Markov's inequality)} \\ &= \frac{\prod_{i=1}^N \mathsf{E}[e^{-Q_i t}]}{e^{-(1-\epsilon')N\cdot\mu\cdot t}} \\ &\quad \text{(because of the independence of the } N \text{ samples).}\end{aligned}$$

We now find a good upper bound for the value $\mathsf{E}[e^{-Q_i t}]$. Consider a given $x$. Observe that when $\omega$ is low, $q(x,\omega) \le q(0,\omega) \le M_1$. Thus, $0 \le Q_i \le M_1$. Notice that

$$\begin{aligned}\mathsf{E}[Q_i] &= \mu \\ &= \sum_{\omega: q(x,\omega) \le M_2, q(0,\omega) \le M_1} q(x,\omega)\pi(\omega) \\ &\quad + \sum_{\omega: M_2 < q(x,\omega), q(0,\omega) \le M_1} q(x,\omega)\pi(\omega) \doteq \mu_2 + \mu_1,\end{aligned}$$

where $\mu_2 = \sum_{\omega: q(x,\omega) \le M_2, q(0,\omega) \le M_1} q(x,\omega)\pi(\omega)$ and $\mu_1 = \sum_{\omega: M_2 < q(x,\omega), q(0,\omega) \le M_1} q(x,\omega)\pi(\omega)$. According to the third result of Lemma 2, we have $\mu_1 \le 2Z^*$. From Property 2, $q(x,\omega) \le q(0,\omega)$. Therefore, $\mu = \mathsf{E}[Q_i] = f_l(x) \le f_l(0) \le \lambda Z^*$ because of the second result of Lemma 2. We can verify that the function $e^{-t\cdot h}$ is convex over $h$ for any given $t > 0$. By Lemma 4, we see that $\mathsf{E}[e^{-tQ_i}] \le \mathsf{E}[e^{-tQ^*}]$, where (1) $Q^*$ takes values only on three possible boundary points $\{0, M_2, M_1\}$, (2) $\mathsf{E}[Q^*] = \mu \le \lambda Z^*$, and (3) $\Pr[Q^* = M_1] \cdot M_1 \le 2Z^*$. Slightly abusing the notation, let $\mu_1 = \Pr[Q^* = M_1] \cdot M_1$ and $\mu_2 = \Pr[Q^* = M_2] \cdot M_2$—with $\mu_1 + \mu_2 = \mu \le \lambda Z^*$ and $\mu_1 \le 2Z^*$. Therefore,

$$\mathsf{E}[e^{-tQ_i}] \le \mathsf{E}[e^{-tQ^*}] \le \frac{\mu_1}{M_1}(e^{-tM_1} - 1) + \frac{\mu_2}{M_2}(e^{-tM_2} - 1) + 1.$$

Hence, we arrive at

$$\begin{aligned}&\Pr[\hat{f}_l(x) - f_l(x) \le -\epsilon Z^*] \\ &\le \frac{\left(\frac{\mu_1}{M_1}(e^{-tM_1} - 1) + \frac{\mu_2}{M_2}(e^{-tM_2} - 1) + 1\right)^N}{e^{-(1-\epsilon')tN\mu}} \\ &\le \exp\left(\frac{N\mu_1}{M_1}(e^{-tM_1} - 1) + \frac{N\mu_2}{M_2}(e^{-tM_2} - 1) + (1-\epsilon')tN\mu\right).\end{aligned}$$

In the following, we apply the second-order approximation of $e^{-h} = 1 - h + (h^2/2) + o(h^2)$ that holds for $h$ close to zero:

$$\begin{aligned}&\frac{N\mu_1}{M_1}(e^{-tM_1} - 1) + \frac{N\mu_2}{M_2}(e^{-tM_2} - 1) + (1-\epsilon')tN\mu \\ &= \frac{N\mu_1}{M_1}\left[-tM_1 + \frac{1}{2}(tM_1)^2 + o((tM_1)^2)\right] \\ &\quad + \frac{N\mu_2}{M_2}\left[-tM_2 - \frac{1}{2}(tM_2)^2 + o((tM_2)^2)\right] + tN\mu - \epsilon' tN\mu \\ &= N\left[-t\mu_1 - t\mu_2 + t\mu + \frac{\mu_1}{M_1}\frac{1}{2}(tM_1)^2 + \frac{\mu_2}{M_2}\frac{1}{2}(tM_2)^2\right. \\ &\quad \left. - t\epsilon'\mu + o\left(\frac{\mu_1}{M_1}(tM_1)^2 + \frac{\mu_2}{M_2}(tM_2)^2\right)\right] \\ &= N\left[\frac{\mu_1}{M_1}\frac{1}{2}(tM_1)^2 + \frac{\mu_2}{M_2}\frac{1}{2}(tM_2)^2 - t\epsilon Z^*\right. \\ &\quad \left. + o\left(\frac{\mu_1}{M_1}(tM_1)^2 + \frac{\mu_2}{M_2}(tM_2)^2\right)\right].\end{aligned} \tag{9}$$

The last equality is because of two facts: (1) $\mu_1 + \mu_2 = \mu$ and (2) $\epsilon'\mu = \epsilon Z^*$.

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

Let $t = \epsilon/(4\lambda Z^*)\min(\epsilon, (1/\lambda))$ and $\Lambda = (\mu_1/M_1)(1/2)(tM_1)^2 + (\mu_2/M_2)(1/2)(tM_2)^2$. Notice that

$$\begin{aligned}\Lambda &= \frac{1}{2}t^2\mu_1 M_1 + \frac{1}{2}t^2\mu_2 M_2 \\ &\le \frac{1}{2}t^2 \cdot 2Z^* \cdot \frac{\lambda Z^*}{\epsilon} + \frac{1}{2}t^2 \cdot \lambda Z^* \cdot 2\lambda Z^* \\ &\quad \text{(using facts } \mu_1 \le 2Z^*, \mu_2 \le \mu \le \lambda Z^*, M_2 \\ &= 2\lambda c(x^*) \le 2\lambda Z^*) \\ &= t\epsilon Z^* \left( t \cdot \frac{\lambda Z^*}{\epsilon^2} + t \cdot \frac{\lambda^2 Z^*}{\epsilon} \right) \\ &\le \frac{t\epsilon Z^*}{2} \quad \text{(by the definition of } t).\end{aligned}$$

Substituting this result back to Equation (9), we obtain

$$\begin{aligned}&\frac{N\mu_1}{M_1}(e^{-tM_1} - 1) + \frac{N\mu_2}{M_2}(e^{-tM_2} - 1) + (1 - \epsilon')tN\mu \\ &\quad = N(\Lambda - t\epsilon Z^* + o(\Lambda)) \le N\left( -\frac{1}{2}t\epsilon Z^* + o(t\epsilon Z^*) \right).\end{aligned}$$

Thus, we have shown that for any given $x \in X$, when $N = \Omega(\ln(1/\delta) \cdot \max[(\lambda^2/\epsilon^2), (\lambda/\epsilon^3)])$, $\Pr[\hat{f}_l(x) - f_l(x) < -\epsilon Z^*] \le \delta$. This completes the proof. □

### 6.4. Proof of Claim 3

**Proof of Claim 3.** Recall that $B = \hat{f}_h(x^*) - \hat{f}_h(\overline{x}) + f_h(\overline{x}) - f_h(x^*)$, that a sample $\omega$ is called high if $q(0,\omega) \ge M_1 = \lambda Z^*/\epsilon$, and that $p = \Pr[\omega \text{ is high}]$. Consider a given (random) set of $N$ independent i.i.d. samples, say $\mathcal{S} = \{\omega_1, \ldots, \omega_N\}$, and let $\mathcal{S}_h$ be the set of high scenarios in $\mathcal{S}$. By definition, $\mathsf{E}[|\mathcal{S}_h|/N] = p$, where $|\mathcal{S}_h|$ denotes the cardinality of $\mathcal{S}_h$:

$$\begin{aligned}B &= \hat{f}_h(x^*) - \hat{f}_h(\overline{x}) + f_h(\overline{x}) - f_h(x^*) \\ &= \frac{1}{N}\sum_{\omega \in \mathcal{S}_h} (q(x^*,\omega) - q(\overline{x},\omega)) \\ &\quad + \mathsf{E}[q(\overline{x},\omega) - q(x^*,\omega) | \omega \text{ is high}] \cdot p \quad \text{(see (6) and (7))} \\ &\le \frac{1}{N}\sum_{\omega \in \mathcal{S}_h} (q(0,\omega) - q(\overline{x},\omega)) \\ &\quad + \mathsf{E}[q(0,\omega) - q(x^*,\omega) | \omega \text{ is high}] \cdot p \quad \text{(Property 2)} \\ &\le \frac{|\mathcal{S}_h|}{N} \cdot \lambda c(\overline{x}) + \lambda c(x^*) \cdot p \quad \text{(Property 3)} \\ &\le \frac{|\mathcal{S}_h|}{N} \cdot \lambda c(\overline{x}) + \epsilon Z^* \quad \text{(because of the first result} \\ &\quad \text{of Lemma 2).}\end{aligned}$$

Notice that $|\mathcal{S}_h|/N$ can be viewed as the mean of $N$ i.i.d. Bernoulli random variables—each with mean $p \le \epsilon/\lambda$—by the first result of Lemma 2. Applying Lemma 1 with $v = \mu = \epsilon/\lambda$, we see that when $N = \Omega(\ln(1/\delta) \cdot (\lambda/\epsilon))$, $\Pr[|\mathcal{S}_h|/N - p \le \epsilon/\lambda] \ge 1 - \delta$. In this case, we have $|\mathcal{S}_h|/N \le p + \epsilon/\lambda \le 2\epsilon/\lambda$ and $B \le 2\epsilon f(\overline{x}) + \epsilon Z^*$. □

## 7. Conclusion

We have shown improved sample-complexity bounds in two-stage stochastic optimization via a careful analysis of the worst-case scenarios based on "high" and "low" scenarios. Optimal sample-complexity bounds would be very useful to know for sample-average approximation. A bold conjecture in this direction would be that $O((\lambda/\epsilon^2)\log(1/\delta))$ samples would suffice. Is such a conjecture true? It also seems reasonable to conjecture that $\Omega((\lambda/\epsilon^2)\log(1/\delta))$ would be a lower bound on the sample complexity.

Furthermore, discretization appears to increase the sample complexity for continuous domains (Charikar et al. 2005). Specifically, section 5 of Charikar et al. (2005) considers the following continuous setting for $\boldsymbol{X}$, where $\boldsymbol{X} \subseteq \Re_+^n$ (the nonnegative orthant of $n$-dimensional Euclidean space) instead of being a discrete set. Suppose the first-stage costs are linear; for some given nonnegative vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$, we take $c(x) = \boldsymbol{v}^T\boldsymbol{x}$. We assume further that the function $q(\cdot,\cdot)$ is $(\lambda, v)$-*Lipschitz*:

$$\begin{aligned}&|q(\boldsymbol{x},\omega) - q(\boldsymbol{x}',\omega)| \le \lambda \cdot \sum_{i=1}^{n} v_i \cdot |x_i - x_i'|, \\ &\forall \omega,\ \boldsymbol{x} = (x_1, \ldots, x_n),\ \ \boldsymbol{x}' = (x_1', \ldots, x_n'),\end{aligned}$$

This is easily seen to imply Property 3. The work of Charikar et al. (2005) reduces such a continuous scenario to the discrete setting via a meshing argument. Can the sample complexity obtained here be improved upon?

### Acknowledgments

The authors thank R. Ravi, the journal referees, and the referees of the 18th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems for their valuable comments regarding the details as well as the context of this work. A preliminary version of this work appeared as part of a paper in the *Proceedings of the 18th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX 2015)*.

### Appendix. Remaining Proofs

**Proof of Lemma 1.** Let $X = \sum_{i=1}^n X_i$. Notice that

$$\begin{aligned}&\Pr[\overline{X} \ge \mu_X + v] = \Pr\left[ X \ge n\mu_X \left( 1 + \frac{v}{\mu_X} \right) \right] \\ &\le \exp\left( -n\mu_X \frac{(v/\mu_X)^2}{2 + v/\mu_X} \right) \quad \text{(using Theorem 3)} \\ &= \exp\left( \frac{-nv^2}{2\mu_X + v} \right). \quad \square\end{aligned}$$

**Proof of Lemma 3.** Because $g(x)$ is convex, we have that $g(x) - g(a) \le (g(b) - g(a))/(b - a)(x - a)$ for any $x \in [a,b]$. Therefore, we get

$$\mathsf{E}[g(X)] \le \frac{g(b) - g(a)}{b - a}(\mu - a) + g(a) = g(a)\frac{b - \mu}{b - a} + g(b)\frac{\mu - a}{b - a}. \quad \square$$

Downloaded from informs.org by [65.202.210.42] on 14 March 2024, at 09:05 . For personal use only, all rights reserved.

## References

Abolhassani M, Esfandiari H, Hajiaghayi M, Mahini H, Malec D, Srinivasan A (2015) Selling tomorrow's bargains today. Weiss G, Yolum P, eds. *Proc. 2015 Internat. Conf. Autonomous Agents Multiagent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC), 337–345.

Bansal N, Gupta A, Li J, Mestre J, Nagarajan V, Rudra A (2012) When LP is the cure for your matching woes: Improved bounds for stochastic matchings. *Algorithmica* 63(4):733–762.

Beale EM (1955) On minimizing a convex function subject to linear inequalities. *J. Roy. Statist. Soc. B* 17(2):173–184.

Besbes O, Gur Y, Zeevi A (2015) Non-stationary stochastic optimization. *Oper. Res.* 63(5):1227–1244.

Birge JR, Louveaux F (2011) *Introduction to Stochastic Programming*, 2nd ed. (Springer-Verlag, New York).

Canetti R, Even G, Goldreich O (1995) Lower bounds for sampling algorithms for estimating the average. *Inform. Processing Lett.* 53(1):17–25.

Charikar M, Chekuri C, Pál M (2005) Sampling bounds for stochastic optimization. Dinur I, Jansen K, Naor J, Rolim J, eds. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques* (Springer, Berlin), 257–269.

Chen N, Immorlica N, Karlin AR, Mahdian M, Rudra A (2009) Approximating matches made in heaven. Albers S, Marchetti-Spaccamela A, Matias Y, Nikoletseas S, Thomas W, eds. *Automata, Languages and Programming* (Springer, Berlin), 266–278.

Dantzig GB (1955) Linear programming under uncertainty. *Management Sci.* 1(3–4):197–206.

Dean BC, Goemans MX, Vondrák J (2008) Approximating the stochastic knapsack problem: The benefit of adaptivity. *Math. Oper. Res.* 33(4):945–964.

Dieker A, Ghosh S, Squillante MS (2016) Optimal resource capacity management for stochastic networks. *Oper. Res.* 65(1):221–241.

Elmachtoub AN, Levi R (2016) Supply chain management with online customer selection. *Oper. Res.* 64(2):458–473.

Garg N, Gupta A, Leonardi S, Sankowski P (2008) Stochastic analyses for online combinatorial optimization problems. Teng S-H, ed. *Proc. Nineteenth Annual ACM-SIAM Sympos. Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia), 942–951.

Goyal V, Levi R, Segev D (2016) Near-optimal algorithms for the assortment planning problem under dynamic substitution and stochastic demand. *Oper. Res.* 64(1):219–235.

Gupta A, Kumar A (2009) A constant-factor approximation for stochastic Steiner forest. Mitzenmacher M, ed. *Proc. Forty-First Annual ACM Sympos. Theory Comput.* (ACM, New York), 659–668.

Gupta A, Ravi R, Sinha A (2007) LP rounding approximation algorithms for stochastic network design. *Math. Oper. Res.* 32(2):345–364.

Gupta A, Pál M, Ravi R, Sinha A (2004) Boosted sampling: Approximation algorithms for stochastic optimization. Babai L, ed. *Proc. Thirty-Sixth Annual ACM Sympos. Theory Comput.* (ACM, New York), 417–426.

Immorlica N, Karger D, Minkoff M, Mirrokni VS (2004) On the costs and benefits of procrastination: Approximation algorithms for stochastic combinatorial optimization problems. Munro I, ed. *Proc. Fifteenth Annual ACM-SIAM Sympos. Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia), 691–700.

Jaillet P, Qi J, Sim M (2016) Routing optimization under uncertainty. *Oper. Res.* 64(1):186–200.

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

Levi R, Pál M, Roundy RO, Shmoys DB (2007) Approximation algorithms for stochastic inventory control models. *Math. Oper. Res.* 32(2):284–302.

Papier F (2016) Supply allocation under sequential advance demand information. *Oper. Res.* 64(2):341–361.

Pu S, Garcia A (2018) A flocking-based approach for distributed stochastic optimization. *Oper. Res.* 66(1):267–281.

Raginsky M, Nedić A (2016) Online discrete optimization in social networks in the presence of Knightian uncertainty. *Oper. Res.* 64(3):662–679.

Ravi R, Sinha A (2006) Hedging uncertainty: Approximation algorithms for stochastic optimization problems. *Math. Programming* 108(1):97–114.

Russo D, Roy BV (2018) Learning to optimize via information-directed sampling. *Oper. Res.* 66(1):230–252.

Ruszczynski AP, Shapiro A (2003) *Stochastic Programming*, vol. 10 (Elsevier, Amsterdam).

Shapiro A (2003) Monte Carlo sampling methods. *Handbook Oper. Res. Management Sci.* 10:353–425.

Shapiro A, Dentcheva D, Ruszczynski A (2014) *Lectures on Stochastic Programming—Modeling and Theory*, vol. 16, 2nd ed. (SIAM, Philadelphia).

Shmoys DB, Swamy C (2004) The sample average approximation method for 2-stage stochastic optimization. Accessed September 30, 2023, https://www.math.uwaterloo.ca/~cswamy/papers/SAAproof.pdf.

Shmoys DB, Swamy C (2006) An approximation scheme for stochastic linear programming and its application to stochastic integer programs. *J. ACM* 53(6):978–1012.

Srinivasan A (2007) Approximation algorithms for stochastic and risk-averse optimization. Gabow H, ed. *Proc. Eighteenth Annual ACM-SIAM Sympos. Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia), 1305–1313.

Tarjan R (2009) Advanced algorithm design. Lecture notes, Princeton University. Accessed September 30, 2023, https://www.cs.princeton.edu/courses/archive/fall09/cos521/Handouts/probabilityandcomputing.pdf.

Wang X, Truong VA (2018) Multi-priority online scheduling with cancellations. *Oper. Res.* 66(1):104–122.

**Alok Baveja** is a professor of supply chain management at the Rutgers Business School. He is currently the director of strategy and innovation at the Department of Supply Chain Management. His expertise is in using innovative modeling and technologies for managing operations in the public and private sectors.

**Amit Chavan** is a software engineer at Snowflake Inc. working on database systems. His recent work focuses on workload optimization for cloud-based databases, transaction processing in data lake architectures, and data lineage and governance in the cloud.

**Andrei Nikiforov** is an assistant professor of professional practice at the Rutgers Business School. He explores financial market responses to information flow, with an emphasis on the effects of earnings seasons on stock and bond prices. Additionally, his research encompasses the field of green energy finance. He particularly investigates how the stochastic supply of energy can be matched with the asynchronous, stochastic demand from households and firms.

**Aravind Srinivasan** is a Distinguished University Professor and a professor with the Department of Computer Science and the Institute for Advanced Computer Studies at the University of Maryland, College Park. His main research interests are in algorithms, probabilistic methods, data science, network science, and machine learning: theory and applications in areas including health, e-commerce, cloud computing, internet advertising, and fairness.

**Pan Xu** is an assistant professor in the Department of Computer Science at the New Jersey Institute of Technology. His research interests broadly span the intersection of algorithms, operations research, and artificial intelligence. His recent research focuses on various stochastic optimization problems in e-commerce, internet advertising, crowdsourcing markets, data mining, databases, and revenue management.