Received: 06 June 2017 | Accepted: 14 June 2018 | Published on: 15 January 2019
DOI: 10.1002/rsa.20839

RESEARCH ARTICLE

WILEY

An optimal (ϵ, δ)-randomized approximation scheme for the mean of random variables with bounded relative variance

Mark Huber

Claremont McKenna College, Upland, CA

Correspondence
Mark Huber, Claremont McKenna College, 850 Columbia AV, Upland, CA 91786.
Email: autotomic@gmail.com

Funding information
This research was supported by the National Science Foundation United States NSF grant; DMS-1418495,

Abstract

Randomized approximation algorithms for many #P-complete problems (such as the partition function of a Gibbs distribution, the volume of a convex body, the permanent of a $\{0, 1\}$-matrix, and many others) reduce to creating random variables $X_1, X_2, \dots$ with finite mean μ and standard deviation σ such that μ is the solution for the problem input, and the relative standard deviation $|\sigma/\mu| \leq c$ for known c. Under these circumstances, it is known that the number of samples from the $\{X_i\}$ needed to form an (ϵ, δ)-approximation $\hat{\mu}$ that satisfies $\mathbb{P}(|\hat{\mu} - \mu| > \epsilon\mu) \leq \delta$ is at least $(2 - o(1))\epsilon^{-2}c^2 \ln(1/[\sqrt{2\pi}\delta])$. We present here an easy to implement (ϵ, δ)-approximation $\hat{\mu}$ that uses $(2 + o(1))c^2\epsilon^{-2} \ln(4/\delta)$ samples. This achieves the same optimal running time as other estimators, but without the need for extra conditions such as bounds on third or fourth moments.

KEYWORDS
exponential convergence, Monte Carlo, robust

1 | INTRODUCTION

Suppose $X_1, X_2, \dots \sim X$ are iid real valued random variables of arbitrary sign with mean $\mathbb{E}[X] = \mu$ and variance $\mathbb{V}(X) = \sigma^2$. The relative standard deviation is $\sigma/|\mu|$ and the relative variance is σ^2/μ^2. Say the relative standard deviation is bounded by c if

$$\left|\frac{\sigma}{\mu}\right| \leq c. \tag{1}$$

Suppose μ and σ are unknown, but c is known. Then the goal is to use as few X_i as possible to find an estimate $\hat{\mu}$ for μ that is an (ϵ, δ)-randomized approximation scheme (or (ϵ, δ)-ras for short), which means that

$$\mathbb{P}(|\hat{\mu} - \mu| > \epsilon\mu) \leq \delta.$$

Consider an (ϵ, δ)-ras that requires

$$(S + o(1))c^2\epsilon^{-2} \ln(4/\delta)$$

samples for some constant S. Here the little-o notation refers to $\epsilon \to 0$. Then call S the *scale factor* of the algorithm.

In this work we present a simple algorithm that is both easy to implement and which achieves the optimal scale factor $S = 2$ without any additional assumptions about the random variables such as bounded higher moments.

This basic problem arises often in randomized algorithms. For instance, problems for approximating the partition function of the Ising model [16], the permanent of a $\{0, 1\}$-matrix [19], the volume of a convex body [9,20], the number of solutions to a DNF logical expression [18], the number of linear extensions of a poset [13,22], and many more all have this problem as a subproblem. Any improvement in the ability to deal with this basic problem directly translates into better approximation algorithms for all of these problems.

This problem has a long history, stretching back to Nemirovsky and Yudin [25] who used the median-of-means estimator in the context of stochastic optimization. Jerrum, Valiant, and Vazirani [18] developed a similar estimator for the purposes of creating randomized approximation schemes for #P complete problems. By 1999 [1], this method was in wide use for online algorithms. Hsu and Sabato [11,12] analyzed the basic median-of-means estimator and proved that it had a scale factor of 121.5 for small enough ϵ.

Catoni [4] greatly advanced the area by presenting an approximation that used an M-estimator. This was not an (ϵ, δ)-ras, rather it gave a confidence bound based on the samples and specific values of the parameters used in the estimate.

Devroye and coworkers [8] showed that if the kurtosis of the random variables is bounded above, then the optimal scale factor $S = 2$ could be attained with a simpler estimator. Unfortunately, in order to run their algorithm, the user needed this upper bound on the kurtosis. Bounding the kurtosis can be much more challenging mathematically than bounding the variance. Minsker and Strawn [23] returned to the original median-of-means estimator. When the random variables have bounded third moment, the Berry-Esseen Theorem can be used to show quick convergence to normality, and they showed that this gave the simple median-of-means algorithm a scale factor of 4.5. As with the Devroye and coworkers method, this requires that the bound on the third moment be given explicitly before the algorithm can be used.

The approach here takes the Catoni M-estimator in a new direction. There is no unique approach to turning the Catoni M-estimator into an (ϵ, δ)-ras. One approach is to use a two-step process that works as follows. Before running the M-estimator, first generate a weaker estimate $\hat{\mu}_1$ that is an $(\sqrt{\epsilon}, \delta/2)$ randomized approximation to μ. Then use this estimate to set the parameters of the Catoni M-estimator to give an output that is provably an (ϵ, δ)-ras. Another approach is to modify the M-estimator of [4] using later ideas of Catoni in [5].

While these different approaches work, they require (like all M-estimators) the ability to find the root of a nonlinear equation. Standard binary search and Newton's method approaches can be used to complete this rootfinding step, however analysis of the number of steps needed to get a close approximation would need to be accomplished before the running time of the method is fully known.

Previous algorithms either had too large a scale factor, required rootfinding, or required knowledge of higher moments. The new method presented here solves all these issues simultaneously.

- It achieves the optimal scale factor $S = 2$.
- No rootfinding step is required. Instead, first a function is randomly chosen by some initial samples, and then the final estimator is a sample average this random function applied to new data.
- No bound on higher moments is necessary. In fact, even if the second moment is the highest moment that exists for the random variables, the new method is still an (ϵ, δ)-ras.

Our main result concerning this new method is as follows.

Theorem 1 *Let $X_1, X_2, \ldots \sim X$ with $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$ satisfying $\sigma^2/\mu^2 \leq c^2$. For $\epsilon < c^2$, there exists an (ϵ, δ)-ras that uses n samples where*

$$n = \left\lceil \frac{2c^2\epsilon^{-2}\ln(4/\delta)}{1 - \epsilon/c^2} \right\rceil + \left\lceil \frac{8c^2\epsilon^{-1}}{(1 - \sqrt{\epsilon})^2} \right\rceil \cdot \left[2 \left\lceil \ln\left(\frac{7}{48\sqrt{\pi}\delta} \right) / \ln\left(\frac{16}{7} \right) \right\rceil + 1 \right].$$

This constant in the leading order term is the best possible.

Theorem 2 *Given ϵ and δ positive, let $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}$ be an (ϵ, δ)-ras for all distributions X with $\mathbb{E}[X] = \mu$ and $\mathbb{V}(X) = \sigma^2$ satisfying $\sigma^2/\mu^2 = c^2$. Then*

$$n \geq 2\epsilon^{-2}c^2 \left[\ln\left(\frac{1}{\sqrt{2\pi}\delta} \right) - \ln\left(\frac{2\ln(1/[\sqrt{2\pi}\delta]) + 1}{\sqrt{2\ln(1/[\sqrt{2\pi}\delta])}} \right) \right].$$

The remainder of the paper is organized as follows. The next section describes the two-step algorithm and proves correctness for each step. Section 3 then shows the lower bound on the number of samples needed. Section 4 shows how Catoni's M-estimator can be modified to be turned into an (ϵ, δ)-ras with similar performance. Finally, Section 5 considers several of the applications mentioned in the introduction in more detail.

2 | THE ALGORITHM

Define the Ψ function as follows.

$$\Psi(u) = \ln(1 + u + u^2/2)\mathbb{1}(u \geq 0) - \ln(1 - u + u^2/2)\mathbb{1}(u \leq 0). \tag{2}$$

This function was used in [4] as part of the M-estimator.

For $u \in [-1, 1]$, the value of $\Psi(u)$ is approximately u. For u greater than 1 in magnitude, the value of $\Psi(u)$ is much less than u. For a constant $\alpha > 0$, $\alpha^{-1}\Psi(\alpha u)$ is a scaled version of Ψ that is close to u for $u \in [-1/\alpha, 1/\alpha]$.

Suppose that $\hat{\mu}_1$ is an initial estimate for μ. Then

$$X_i = \hat{\mu}_1 + (X_i - \hat{\mu}_1)$$

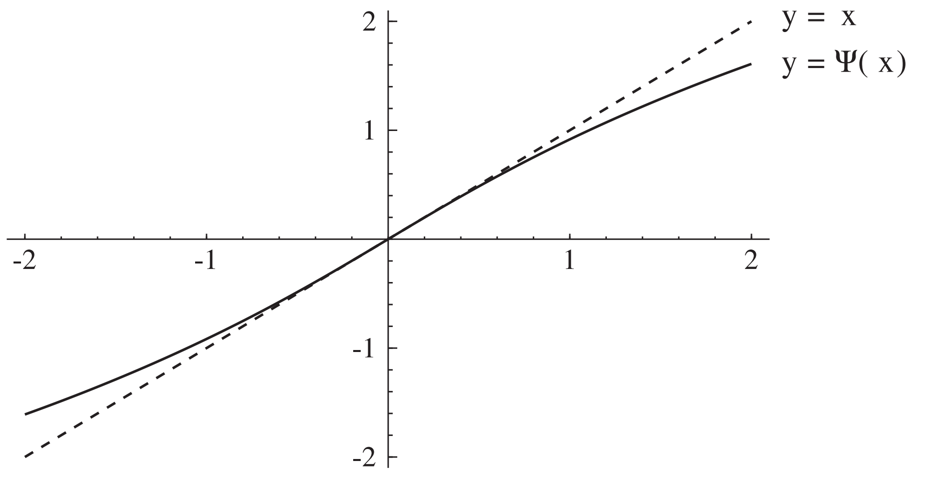

FIGURE 1 The functions $\Psi(x)$ and x over $[-2, 2]$

has mean μ but is also susceptible to outliers in the X_i distribution. By replacing this with

$$W_i = \hat{\mu}_1 + \alpha^{-1}\Psi(\alpha \cdot (X_i - \hat{\mu}_1)),$$

the value of W_i will be close to X_i when $|X_i - \hat{\mu}_1| \le \alpha^{-1}$, but always has a light-tailed distribution because of the logarithm function.

The algorithm proceeds as follows. The first step uses a median-of-means approach to find $\hat{\mu}_1$ that is a $(\sqrt{\epsilon}, \delta/2)$ approximation of μ. Given that the first step did not fail, the next step then uses the sample average of the W_i variables to create the final estimate $\hat{\mu}$ that is an $(\epsilon, \delta/2)$ approximation. The chance that either step fails is at most δ.

1. The first step is to construct a median-of-means estimator for μ [18]. Let $k = \lceil 8c^2\epsilon^{-1}/(1-\sqrt{\epsilon})^2 \rceil$, and $m = 2\lceil \ln(7/[48\sqrt{\pi}\delta)/\ln(16/7) \rceil + 1$. Let S have the distribution of the sample average of k independent draws from X. Draw $S_1, \dots, S_m \sim S$ independently, and let $\hat{\mu}_1 = \text{median}(\{S_i\})/(1-\epsilon)$.
2. Let $n = \lceil 2c^2\epsilon^{-2}\ln(2/\delta)/(1-\epsilon) \rceil$, and $\alpha = \epsilon/[c^2\hat{\mu}_1]$. Draw $X_1, \dots, X_n$ independently. For all i, let

$$W_i = \hat{\mu}_1 + \alpha^{-1}\Psi(\alpha \cdot (X_i - \hat{\mu}_1)).$$

Set $\hat{\mu} = (W_1 + \cdots + W_n)/n$.

2.1 | The first step of the algorithm

The first step of the algorithm is the powering method of Jerrum, Valiant, and Vazirani [18] applied to the sample averages. This technique was also used in [1], and later referred to as the median-of-means method [11]. These authors did not attempt to optimize the constants in their arguments, and so we repeat the proof here so we can see exactly how the choice of constant enters into the failure bound.

Suppose that we have random variables whose relative standard deviation is at most $\nu\epsilon$. What is the chance that the median of $m = 2r+1$ draws from the random variable falls into $[\mu(1-\epsilon), \mu(1+\epsilon)]$?

To answer this question, first consider the probability that a beta distributed random variable with both parameters equal to an integer r falls into a subinterval of $[0, 1]$.

Lemma 1 *Let $B \sim \mathsf{Beta}(r+1, r+1)$ denote a random variable with density $f_B(x) = [(2r+1)!/(r!r!)]x^r(1-x)^r \mathbb{1}(x \in [0,1])$. For any $0 \le a \le 1/2 \le b \le 1$ with $1-(b-a) \le 1/2$,*

$$\mathbb{P}(B \notin [a,b]) \le 2\frac{4^r}{\sqrt{\pi r}} \cdot \frac{[(1-(b-a))(b-a)]^{r+1}}{2(b-a)-1}.$$

Proof Density f_B is symmetric about its unique local maximum at $1/2$, so

$$\int_{x\in[0,a]\cup[b,1]} f_B(r)\,\mathrm{d}r \le \int_{x\in[0,a]\cup[a,a+1-b]} f_B(r) = \int_{x\in[0,1-(b-a)]} f_B(r)\,\mathrm{d}r.$$

Note $x^r(1-x)^r = (x-x^2)^r$. Let $t = 1-(b-a)$. Then $t \le 1/2$, $[x-x^2]' = 1-2x > 0$ and $[x-x^2]'' = 2x \ge 0$, so the function lies below its tangent line at t. That is,

$$(\forall x \in [0,t])(x-x^2) \le t(1-t) + (x-t)(1-2t).$$

Then

$$\int_{x\in[0,t]} [t(1-t)+(x-t)(1-2t)]^{r+1}\,\mathrm{d}r \le \frac{[t(1-t)]^r}{(r+1)(1-2t)}.$$

Using Stirling's formula to give $(2r+1)(2r)!/(r!r!) \le (2r+1)4^r/\sqrt{\pi r}$ completes the proof. ■

Lemma 2 *Let $A_1, \dots, A_{2r+1}$ be iid with mean μ and variance at most $v^2\epsilon^2\mu^2$ where $v^2 \le 1/2$. Then*

$$\mathbb{P}(\mathrm{med}(\{A_i\}) \notin [\mu(1-\epsilon), \mu(1+\epsilon)]) < \frac{(v^2)(1-v^2)}{\sqrt{\pi r}(1-2v^2)} \exp(r\ln(4(v^2)(1-v^2))).$$

Proof By Chebyshev's inequality,

$$\mathbb{P}(A_i \notin [\mu(1-\epsilon), \mu(1+\epsilon)]) \le \frac{v^2\epsilon^2\mu^2}{(\epsilon\mu)^2} = v^2.$$

Let $x_1 = \mathbb{P}(A_i < a)$ and $x_2 = \mathbb{P}(A_i \le b)$. Construct a uniform random variable over $[0,1]$ as follows. If $A_i < a$, then let $U_i \sim \mathsf{Unif}([0,x_1))$. If $A_i \in [a,b]$, let $U_i \sim \mathsf{Unif}([x_1,x_2])$. Finally, if $A_i > b$, then let $U_i \sim \mathsf{Unif}((x_2,1])$. Note that

$$\mathbb{P}(\mathrm{med}(\{U_i\}) \in [x_1,x_2]) = \mathbb{P}(\mathrm{med}(\{A_i\}) \in [a,b]).$$

The median of $2r+1$ iid uniform $[0,1]$ random variables is well known to have a beta distribution: $\mathrm{med}(\{U_i\}) \sim \mathsf{Beta}(r+1,r+1)$. So the previous lemma can be used to state

$$\mathbb{P}(\mathrm{med}(\{U_i\}) \notin [x_1,x_2]) \le \frac{(v^2)(1-v^2)[4(v^2)(1-v^2)]^r}{\sqrt{\pi r}(1-2v^2)}.$$ ■

So the failure probability is going down exponentially at rate $\ln(4(\nu^2)(1-\nu^2))$. Now for an integer k, consider $X_1, \ldots, X_k \sim X$ iid, and

$$S = (X_1 + \cdots + X_k)/k.$$

Then $\mathbb{E}[S] = \mathbb{E}[X] = \mu$ and $\mathbb{V}[S] = k\mathbb{V}[X]/k^2 = \sigma^2/k$.

In particular, for $\sigma^2/\mu^2 \leq c^2$, and $k = \lceil \epsilon^{-2}\nu^{-2}c^2 \rceil$, $\mathbb{V}[S] \leq (\nu\epsilon\mu)^2$. To take the median of $2r+1$ draws of the sample average of k draws from the $\{X_i\}$ takes $\Theta(kr) = \Theta(-1/(\nu^2 \ln(4\nu^2(1-\nu^2))))$ samples. Choose $\nu^2 = 1/8$ to roughly minimize the overall running time.

Lemma 3 (Median-of-means) *Suppose $X_1, X_2, \ldots$ are as in (1). For $k = \lceil 8c^2\epsilon^{-2}/(1-\epsilon)^2 \rceil$ let S be distributed as $(X_1 + \cdots + X_k)/k$. Let $m = 2\lceil \ln(7/[48\sqrt{\pi}\delta])/\ln(16/7) \rceil + 1$. Let $S_1, \ldots, S_m \sim S$ then*

$$\mathbb{P}(|\operatorname{med}(\{S_i\}) - \mu| > \epsilon\mu) \leq \delta.$$

Proof Follows from setting $\gamma^2 = 1/8$ in the previous lemma. ∎

We want the output of the first step to have a particular relationship between the absolute and relative error. Let $\xi = [\mu/\hat{\mu}_1] - 1$ and $\gamma = \mu - \hat{\mu}_1$.

Lemma 4 *We have $-\epsilon'\mu/(1+\xi) \leq \gamma \leq \epsilon'\mu/(1+\xi)$ if and only if $\mu/(1-\epsilon') \geq \hat{\mu}_1 \geq \mu/(1+\epsilon')$.*

Proof The statement $-\epsilon'\mu/(1+\xi) \leq \gamma \leq \epsilon'\mu/(1+\xi)$ is equivalent to

$$-\epsilon'\mu/(1+\mu/\hat{\mu}_1 - 1) \leq \mu - \hat{\mu}_1 \leq \epsilon'\mu/(1+\mu/\hat{\mu}_1 - 1).$$

Simplifying and solving these inequalities for $\hat{\mu}_1$ gives the result. ∎

The next lemma shows how to scale an unbiased estimate in an interval centered around μ to get an estimate that meets the asymmetric requirements of the previous lemma.

Lemma 5 *Suppose $\mathbb{E}[\hat{\mu}_2] = \mu$. Let $\hat{\mu}_3 = \hat{\mu}_2/(1-\epsilon'^2)$. Then*

$$\frac{\mu}{1+\epsilon'} \leq \hat{\mu}_3 \leq \frac{\mu}{1-\epsilon'} \Leftrightarrow \mu(1-\epsilon') \leq \hat{\mu}_2 \leq \mu(1+\epsilon').$$

The proof follows from simplifying the appropriate inequalities.

Now the first step of the algorithm can be shown to have the target property. Use the median-of-means approach to get a mean $\hat{\mu}_2$ with relative error $\sqrt{\epsilon}$, then let $\hat{\mu}_1 = \hat{\mu}_2/(1-(\sqrt{\epsilon})^2)$ in order to get something that satisfies the γ and ξ relationship.

Lemma 6 *The output of Step 1 of the algorithm gives $\hat{\mu}_1$ satisfying*

$$|\gamma| \leq \sqrt{\epsilon}\mu/(1+\xi) = \sqrt{\epsilon}\hat{\mu}_1,$$

where $\gamma = \mu - \hat{\mu}_1$ and $\xi = (\mu/\hat{\mu}_1) - 1$.

2.2 | The second step of the algorithm

To analyze this step, it helps to have two new functions that upper and lower bound Ψ.

$$\Psi_U(u) = \ln(1 + x + x^2/2),\ \Psi_L(u) = -\ln(1 - x + x^2/2). \tag{3}$$

Lemma 7 *For all $u \in \mathbb{R}$,*

$$\Psi_L(u) \leq \Psi(u) \leq \Psi_U(u).$$

Proof First consider $\Psi_L(u) \leq \Psi(u)$. These are equal when $x \leq 0$, so suppose $x \geq 0$. Exponentiating gives

$$\begin{aligned}\Psi_L(u) \leq \Psi(u) &\Leftrightarrow [1 - u + u^2/2]^{-1} \leq 1 + u + u^2/2,\\ &\Leftrightarrow 1 \leq 1 + u^4/4,\end{aligned}$$

therefore the inequality holds. The other inequality is shown similarly. ■

Now set

$$\begin{aligned}W_{L,i} &= \hat{\mu}_1 + \alpha^{-1}\Psi_L(\alpha \cdot (X_i - \hat{\mu}_1)),\\ W_{U,i} &= \hat{\mu}_1 + \alpha^{-1}\Psi_U(\alpha \cdot (X_i - \hat{\mu}_1)).\end{aligned}$$

By the previous lemma, $W_{L,i} \leq W_i \leq W_{U,i}$ for all i.

Lemma 8 *Suppose that $\xi \leq \sqrt{\epsilon}$ and $\gamma^2 \leq \epsilon\mu^2$. Denote $\bar{W}_U = (W_{U,1} + \cdots + W_{U,n})/n$. Then*

$$\mathbb{P}(\bar{W} > \mu(1 + \epsilon)) \leq \exp[-(n\epsilon^2/(2c^2)) \cdot (1 - \epsilon/c^2)].$$

Proof Take a Chernoff bound [7] style approach. Since $\alpha > 0$ and exp is a strictly increasing function,

$$\begin{aligned}\mathbb{P}(\bar{W}_U > \mu(1 + \epsilon) &= \mathbb{P}(W_{U,1} + \cdots + W_{U,n} > n\mu(1 + \epsilon))\\ &= \mathbb{P}(\exp(\alpha(W_{U,1} + \cdots + W_{U,n})) > \exp(\alpha n\mu(1 + \epsilon)))\\ &\leq \mathbb{E}[\exp(\alpha(W_{U,1} + \cdots + W_{U,n}]/\exp(\alpha n\mu(1 + \epsilon))\\ &= \left[\frac{\mathbb{E}[\exp(\alpha W_{U,1})]}{\exp(\alpha\mu(1 + \epsilon)}\right]^n.\end{aligned}$$

First consider the expression inside the mean in the numerator:

$$\begin{aligned}\exp(\alpha W_{U,1}) &= \exp(\alpha\hat{\mu}_1 + \ln(1 + \alpha(X - \hat{\mu}_1) + (\alpha^2/2)(X - \hat{\mu}_1)^2)\\ &= \exp(\alpha\hat{\mu}_1)[1 + \alpha(X - \hat{\mu}_1) + (\alpha^2/2)(X - \hat{\mu}_1)^2]\\ &= \exp(\alpha\hat{\mu}_1)[1 + \alpha(X - \mu + \gamma) + (\alpha^2/2)(X - \mu + \gamma)^2]\\ &= \exp(\alpha\hat{\mu}_1)[1 + \alpha(X - \mu) + \alpha\gamma + (\alpha^2/2)((X - \mu)^2 + 2(X - \mu) + \gamma^2)].\end{aligned}$$

Since $\mathbb{E}[X - \mu] = 0$ and $\mathbb{E}[(X - \mu)^2] = \sigma^2$, we have

$$\mathbb{P}(\bar{W}_U > \mu(1 + \epsilon) \leq \left[\frac{\exp(\alpha\hat{\mu}_1)[1 + \alpha\gamma + \alpha^2\gamma^2/2 + \alpha^2\sigma^2/2]}{\exp(\alpha\mu(1 + \epsilon))}\right]^n.$$

Note

$$\frac{\exp(\alpha\hat{\mu}_1)}{\exp(\alpha\mu(1+\epsilon))} = \exp(-\alpha(\mu - \hat{\mu}_1) - \alpha\epsilon\mu) = \exp(-\alpha\gamma - \alpha\epsilon\mu).$$

Next use $1 + x \leq \exp(x)$ to state

$$\mathbb{P}(\bar{W}_U > \mu(1+\epsilon) \leq \exp(-\alpha\gamma - \alpha\epsilon\mu + \alpha\gamma + \alpha^2\gamma^2/2 + \alpha^2\sigma^2/2)^n.$$

Since $\sigma^2 \leq c^2\mu^2$, $\alpha = (1+\xi)\epsilon/[c^2\mu]$, and $\gamma^2 \leq \epsilon\mu^2/(1+\xi)^2$:

$$\begin{aligned}\mathbb{P}(\bar{W}_U > \mu(1+\epsilon) &\leq \exp(-\epsilon^2/c^2 + \epsilon^2 \cdot \epsilon/[2c^4] + \epsilon^2/[2c^2])^n \\ &= \exp(-(n\epsilon^2(1+\xi)/[2c^2])(1 - \epsilon/c^2)).\end{aligned}$$

■

Lemma 9 *Suppose that* $\xi \leq \sqrt{\epsilon}$ *and* $\gamma^2 \leq \epsilon\mu^2$. *Denote* $\bar{W}_L = (W_{L,1} + \cdots + W_{L,n})/n$. *Then*

$$\mathbb{P}(\bar{W}_L < \mu(1-\epsilon)) \leq \exp(-[n\epsilon^2/(2c^2)] \cdot (1-\epsilon)).$$

Proof The proof is similar to the previous lemma: first multiply by $-\alpha$ and exponentiate to get

$$\begin{aligned}\mathbb{P}(\bar{W}_L > \mu(1-\epsilon) &= \left[\mathbb{E}[\exp(-\alpha W_{U,1})]\exp(\alpha\mu(1-\epsilon)\right]^n \\ &= [\exp(\alpha\gamma - \alpha\epsilon\mu)(1 - \alpha\gamma + \alpha^2\gamma^2/2 + \alpha^2\sigma^2/2)]^n \\ &\leq \exp(-\alpha\epsilon\mu + \alpha^2\gamma^2/2 + \alpha^2\sigma^2/2)^n,\end{aligned}$$

and the rest of the proof is the same as the previous lemma. ■

Putting these results together gives the following.

Lemma 10 *For* $n \geq 2c^2\epsilon^{-2}\ln(2/\delta)(1-\epsilon)^{-1}$ *and* $|\hat{\mu}_1 - \mu| \leq \epsilon\hat{\mu}_1$,

$$\mathbb{P}(|\bar{W} - \mu| > \epsilon\mu) \leq \delta.$$

Proof From the previous lemmas, each of the two steps of the algorithm has at most an $\exp(-\ln(2/\delta)) = \delta/2$ chance of failure. The union bound then bounds the total chance of failure by δ. ■

Theorem 1 immediately follows from summing the number of samples needed for each of the two steps, and using large enough n to make sure that each step fails with probability at most $\delta/2$.

3 | LOWER BOUND ON THE NUMBER OF SAMPLES

Begin with a rephrasing of Proposition 6.1 from [4].

Lemma 11 *Let* $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}$ *be any estimator of the mean of n iid random variables. Let* $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ *and* $\bar{Y} = (Y_1 + \cdots + Y_n)/n$. *Then either* $\mathbb{P}(\hat{\mu}(Y_1, \ldots, Y_n) \geq \mu(1+\epsilon)) \geq \mathbb{P}(\bar{Y} \geq \mu(1+\epsilon))$ *or* $\mathbb{P}(\hat{\mu} \leq \mu(1-\epsilon)) \geq \mathbb{P}(\bar{Y} \leq \mu(1-\epsilon))$ *for* $\bar{Y} = (Y_1 + \cdots Y_n)/n$.

Note that for $Y_i \sim \mathsf{N}(\mu, c^2\mu^2)$, then $\bar{Y} \sim \mathsf{N}(\mu, c^2\mu^2/n)$. Let $Z \sim \mathsf{N}(0,1)$. Then from the scaling properties of normal random variables,

$$\mathbb{P}(\bar{Y} \in [\mu(1-\epsilon), \mu(1+\epsilon)]) = \mathbb{P}(Z \in [-\epsilon\sqrt{n}/c, \epsilon\sqrt{n}/c]).$$

Since $\mathbb{P}(Z \le -a) = \mathbb{P}(Z \ge a)$ for all a, we need only bound one tail of the normal.

Lemma 12 *Let $Z \sim N(0,1)$ and a_δ satisfy $\mathbb{P}(Z \ge a_\delta) = \delta$ where $\delta \le 1/\sqrt{2\pi}$. Then*

$$a_\delta^2 \ge 2\ln\left(\frac{1}{\sqrt{2\pi}\delta}\right) + 2\ln\left(\frac{\sqrt{2\ln(1/[\sqrt{2\pi}\delta])}}{2\ln(1/[\sqrt{2\pi}\delta]) + 1}\right).$$

Proof Gordon [10] showed that for $a \ge 0$,

$$\mathbb{P}(Z \ge a) \ge \frac{a}{a^2+1}\frac{1}{\sqrt{2\pi}}\exp(-a^2/2).$$

Without the $a/(a^2+1)$ factor, the right hand side equals $\delta/2$ when $a_1 = \sqrt{2\ln(1/(\sqrt{2\pi}\delta))}$. Since $a/(a^2+1) \le 1$ we have $a_\delta \le a_1$. Also, $a/(a^2+1)$ is a decreasing function, so

$$\frac{a_1}{a_1^2+1}\frac{1}{\sqrt{2\pi}}\exp(-a_\delta^2/2) \le \delta/2.$$

Solving gives

$$a_\delta^2 \ge 2\ln\left(\frac{a_1}{a_1^2+1}\frac{1}{\sqrt{2\pi}\delta}\right)$$

as desired. ∎

Putting $a_\delta = \epsilon\sqrt{n}/c$ then gives Theorem 2.

4 | AN *M*-ESTIMATOR THAT IS AN (ϵ, δ)-RAS

Because of the rootfinding step, it is difficult to directly assess the running time of the M-estimator approach. However, in practice rootfinding is very fast, and for X_i of constant sign it is possible to show that using a modification of the estimator from [4] based on ideas appearing in [5], an (ϵ, δ)-ras can be created. (As long as the running time is measured solely in terms of the number n of $\{X_i\}$ draws that are needed.)

The concepts of this section were given to the author by an anonymous referee of an earlier draft of this paper along with permission to include these ideas.

If n samples are drawn, let

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \lambda^{-1}\Psi(\lambda(X_i/x - 1)).$$

Since Ψ is an increasing function, $f(x)$ is decreasing in x if the X_i are all nonnegative. (If any of the $\{X_i\}$ are nonzero, then $f(x)$ is strictly decreasing.) Therefore assume in this section that the $\{X_i\}$ are all nonnegative. (Of course the results of this section can be applied when the X_i are all nonpositive as well.)

The function $f(x)$ is also continuous, and so if $f((1-\epsilon)\mu) > 0$ and $f((1+\epsilon)\mu) < 0$, then any roots of f must lie in $[(1-\epsilon)\mu, (1+\epsilon)\mu]$.

Lemma 13 *Let $\epsilon' = \epsilon/(1+\epsilon)$. Suppose*

$$n = \lceil 2c^2\epsilon^{-2}\ln(2/\delta)(1+\epsilon)^2\rceil,$$

where $\epsilon \leq (4+c^{-2})^{-1}$. For $\lambda = \epsilon'/c^2$, with probability at least $1-\delta$, $f((1-\epsilon)\mu) \geq (\epsilon')^2/2$ and $f((1+\epsilon)\mu) \leq -(\epsilon')^2/2$.

Hence any value of $\hat{\mu}$ such that $f(\hat{\mu}) \in [-(\epsilon')^2/2, (\epsilon')^2/2]$ will be an (ϵ,δ)-ras for μ.

Before proving this, it will be helpful to have another form of Chernoff's bound.

Lemma 14 *For a random variable X where $\mathbb{E}[\exp(X)]$ is finite,*

$$\mathbb{P}(X \geq \ln(\mathbb{E}[\exp(X)]) + s) \leq \exp(-s).$$

Similarly, if $\mathbb{E}[\exp(-X)]$ is finite,

$$\mathbb{P}(X \leq -\ln(\mathbb{E}[\exp(-X)]) - s) \leq \exp(-s).$$

Proof Since the exponential function is increasing

$$\begin{aligned}\mathbb{P}(X \geq \ln(\mathbb{E}[\exp(X)]) + s) &= \mathbb{P}(\exp(X) \geq \mathbb{E}[\exp(X)]\exp(s)) \\ &\leq \mathbb{E}[\exp(X)]/[\mathbb{E}[\exp(X)]\exp(s)] = \exp(-s).\end{aligned}$$

The other inequality is shown in a similar fashion. ■

To use this for $f(x)$ we need the following.

Lemma 15 *Let $w = \mu/x - 1$. Then*

$$\begin{aligned}\mathbb{E}[\exp(\Psi(\lambda(X_i/x-1)))] &\leq 1 + \lambda w + (1/2)\lambda^2(c^2(1+w)^2 + w^2) \\ \mathbb{E}[\exp(-\Psi(\lambda(X_i/x-1)))] &\leq 1 - \lambda w + (1/2)\lambda^2(c^2(1+w)^2 + w^2).\end{aligned}$$

Proof First

$$\exp(\Psi(\lambda(X_i/x-1))) \leq 1 + \lambda(X_i/x-1) + (1/2)\lambda^2(X_i/x-1)^2$$

from Lemma 7. Since the mean of the square of a random variable is the sum of the variance and the square of the mean,

$$\mathbb{E}[\exp(\Psi(\lambda(X_i/x-1)))] \leq 1 + \mathbb{E}[\lambda(X_i/x-1)] + \mathbb{E}[(1/2)\lambda^2(X_i/x-1)^2]$$

$$
\begin{aligned}
&= 1 + \lambda w + (1/2)\lambda^2(\mathbb{V}(X_i/x - 1) + \mathbb{E}(X_i/x - 1)^2) \\
&\leq 1 + \lambda w + (1/2)\lambda^2(c^2\mu^2/x^2 + w^2) \\
&= 1 + \lambda w + (1/2)\lambda^2(c^2(1 + w)^2 + w^2).
\end{aligned}
$$

■

Proof of Lemma 13 Let $\mu_- = (1 - \epsilon)\mu$ and $\mu_+ = (1 + \epsilon)\mu$. Then for any a,

$$\mathbb{P}(f(\mu_+) \geq a/[n\lambda]) = \mathbb{P}(n\lambda f(\mu_+) \geq a).$$

Since

$$\mathbb{E}[\exp(n\lambda f(\mu_+))] = \mathbb{E}[\exp(\Psi(\lambda(X_1/\mu_+ - 1)))]^n,$$

we have

$$\mathbb{P}(n\lambda f(\mu_+) \geq n \ln(\mathbb{E}[\exp(\Psi(\lambda(X_1/\mu_+ - 1)))]) + s) \leq \exp(-s).$$

Setting $s = \ln(2/\delta)$ makes this bound $\delta/2$.

Similarly,

$$\mathbb{P}(n\lambda f(\mu_-) \leq -n \ln(\mathbb{E}[\exp(-\Psi(\lambda(X_1/\mu_- - 1)))]) - s) \leq \exp(-s).$$

Set $w_+ = \mu/\mu_+ - 1$ and $w_- = \mu/\mu_- - 1$. Using the fact that $\ln(1 + a) \leq a$ and $-\ln(1 - a) \geq a$ for all $|a| < 1$ together with the previous lemma gives

$$
\begin{aligned}
\mathbb{P}(n\lambda f(\mu_+) \geq n\lambda w_+ + (1/2)n\lambda^2[c^2(1 + w_+)^2 + w_+^2] + \ln(2/\delta)) \leq \delta/2 \\
\mathbb{P}(-n\lambda f(\mu_-) \leq n\lambda w_- - (1/2)n\lambda^2[c^2(1 + w_-)^2 + w_-^2] - \ln(2/\delta)) \leq \delta/2.
\end{aligned}
$$

Therefore both

$$
\begin{aligned}
f(\mu_+) \leq w_+ + (1/2)\lambda[c^2(1 + w_+)^2 + w_+^2] + \ln(2/\delta)/[n\lambda] \\
f(\mu_-) \geq w_- - (1/2)\lambda[c^2(1 + w_-)^2 + w_-^2] - \ln(2/\delta)/[n\lambda]
\end{aligned}
$$

hold with probability at least $1 - \delta$.

At this point, note that $\ln(2/\delta)/[n\lambda] \leq \epsilon'/2$. Also $w_+ = \mu/[(1 + \epsilon)\mu] - 1 = -\epsilon'$ and $w_- = \mu/[(1 - \epsilon)\mu] - 1 = \epsilon'/(1 - 2\epsilon')$. Using these together with $\lambda = \epsilon'/c^2$ allows us to say that with probability at least $1 - \delta$, both

$$
\begin{aligned}
f(\mu_+) &\leq -\epsilon' + \frac{\epsilon'}{2c^2}[c^2(1 - \epsilon')^2 + (\epsilon')^2] + \frac{\epsilon'}{2} = -(\epsilon')^2\left(1 - \frac{\epsilon'}{2}(1 + c^{-2})\right) \\
f(\mu_-) &\geq \frac{\epsilon'}{1 - 2\epsilon'} - \frac{\epsilon'}{2c^2}\left[c^2\left(\frac{1 - \epsilon'}{1 - 2\epsilon'}\right)^2 - \left(\frac{\epsilon'}{1 - 2\epsilon'}\right)^2\right] - \frac{\epsilon'}{2} = \frac{(\epsilon')^2}{(1 - 2\epsilon')^2}\left(1 - \frac{\epsilon'}{2}(5 + c^{-2})\right)
\end{aligned}
$$

hold. If $\epsilon' \leq (5 + c^{-2})^{-1}$ (or equivalently $\epsilon \leq (4 + c^{-2})^{-1}$) then the right hand side factors are at least $1/2$, and so

$$\mathbb{P}(f(\mu_+) \leq -(\epsilon')^2/2 \text{ and } f(\mu_-) \geq (\epsilon')^2/2) \geq 1 - \delta.$$

■

4.1 | Avoiding the natural logarithm

The definition of the Ψ function in Equation 2 emphasizes that the distant outliers are being assigned logarithmic weight, but is not the only function that satisfies Lemma 7.

In [6], the following was shown as Lemma 3.1.

Lemma 16 *Let*

$$\Psi_2(u) = [-2\sqrt{2}/3]\mathbb{1}(u < -\sqrt{2}) + [t - t^3/6]\mathbb{1}(u \in [-\sqrt{2}, \sqrt{2}]) + [2\sqrt{2}/3]\mathbb{1}(u > \sqrt{2}). \quad (4)$$

Then

$$\Psi_L(u) \leq \Psi_2(u) \leq \Psi_U(u),$$

for Ψ_L *and* Ψ_U *defined in (3).*

Therefore Ψ_2 can be used instead of Ψ in the approach of Section 2 or in the M-estimate approach of Section 4. This avoids the need to compute the natural logarithm when computing the estimate.

5 | APPLICATIONS

This work on an (ϵ, δ)-ras with optimal scale factor directly applies to a variety of algorithms, immediately improving their runtime. Several of these arise out of work of Jerrum, Valiant, and Vazirani [18], who showed that for a large class of *self-reducible* problems, the ability to sample from a density in polynomial time lead to an (ϵ, δ)-ras for the normalizing constant of the unnormalized density. Since finding that normalizing constant is often a #P-complete problem, this has been used in many settings. Each of these leads to a problem such as that considered here where a random variable has mean μ equal to the target with bounded relative standard deviation. This method was expanded to more examples later by Jerrum and Sinclair [17].

The idea is as follows. Suppose that the goal is to find $\#A_0$ which is the size of a set (either number of elements for a finite set or the Lebesgue measure for $A_0 \subset \mathbb{R}^n$). Suppose that we can find a sequence of decreasing sets $A_0 \supseteq A_1 \supseteq A_2 \supseteq \cdots \supseteq A_k$ where $\#A_k$ is known. If each of the sets A_i represents an instance of the original problem (perhaps with a different input), the problem is self-reducible. If there is an efficient method for generating samples uniformly from the A_i, then for each $i \in \{0, 1 \ldots, k-1\}$, let $X_{i,1}, \ldots, X_{i,m} \sim \textsf{Unif}(A_i)$, and let $R_i = m^{-1} \sum_j \mathbb{1}(X_{i,j} \in A_{i+1})$ be the percentage of values that fall into A_{i+1}. Then

$$\frac{\#A_k}{\#A_0} = \mathbb{E}[R_0]\mathbb{E}[R_1] \cdots \mathbb{E}[R_{k-1}],$$

so let $\hat{r} = R_0 \cdots R_{k-1}$ be the unbiased *product estimator* for $\#A_k/\#A_0$.

Then

$$\frac{\mathbb{V}(\hat{r})}{\mathbb{E}[\hat{r}]^2} = \frac{\mathbb{E}[\hat{r}^2]}{\mathbb{E}[\hat{r}]^2} - 1 = \left[\prod_{i=1}^{k} \frac{\mathbb{E}[R_i^2]}{\mathbb{E}[R_i]^2}\right] - 1 = \left[\prod_{i=1}^{k} 1 + \frac{\mathbb{V}[R_i]}{\mathbb{E}[R_i]^2}\right] - 1.$$

Let $r_i = \#A_i/\#A_{i+1}$. Then $X_{i,1}$ has a Bernoulli distribution with mean r_i and variance $r_i(1 - r_i)$. As the sample average of m iid draws from $X_{i,1}$, R_i has mean r_i and variance $r_i(1 - r_i)/m$. Then

$$\frac{\mathbb{V}(\hat{r})}{\mathbb{E}[\hat{r}]^2} \leq \left[\prod_{i=1}^{k} 1 + \frac{1 - r_i}{mr_i}\right] - 1,$$

so if $r_i \geq 1/M$ for all i, using $1 + x \leq \exp(x)$ gives

$$\frac{\mathbb{V}(\hat{r})}{\mathbb{E}[\hat{r}]^2} \leq \exp\left(\frac{k(M-1)}{m}\right) - 1.$$

There are k different R_i each requiring m samples, therefore km are needed to generate one value of $\hat{r}$. From the above the variance is $\exp(k(M-1)/m) - 1 \approx k(M-1)/m$ for large m. Hence for large m (such as $k(M-1)\epsilon^{-2}$) using the algorithm presented here the total number of samples needed for an (ϵ, δ)-ras is (to leading order) $2k(M-1)\epsilon^{-2}\ln(4/\delta)$, with the 2 being the optimal value of the constant.

5.1 | Linear extensions of a poset

For a direct application of this process, consider the problem of counting the number of linear extensions of a partially ordered set (poset). A *poset* on n objects $\{1, \dots, n\}$ is an ordering $\preceq$ with three properties. Let $i, j, k \in \{1, \dots, n\}$. First, $i \preceq i$. Second, if $i \preceq j$ and $j \preceq i$, then $i = j$. Third, if $i \preceq j$ and $j \preceq k$, then $i \preceq k$. A *linear extension* of the poset is a permutation τ such that $\tau(i) \preceq \tau(j) \Rightarrow i \leq j$.

Brightwell and Winkler [2] showed that counting the number of linear extensions of an arbitrary poset is a #P-complete problem. Finding the number of linear extensions has applications in nonparametric statistics [24].

A sequence of results [3, 13, 21, 22] culminated in an $O(n^3 \ln(n))$ method for generating samples uniformly from the set of linear extensions. To convert this method into a method for approximately counting the number of linear extensions, use self-reducibility.

Let n_ℓ be any element of $\{1, \dots, n\}$ which is not preceded by another element in the set. Then an easy Markov chain argument gives that the probability that a uniformly chosen linear extension has $\tau(n_\ell) = n$ is at least $1/n$. Fixing $\tau(n_\ell) = n$ in the permutation leaves a linear extension problem of size $n - 1$. So the methods of this section can be applied with $k = n$ and $M = n$. Hence (to first order) $2n^2\epsilon^{-2}\ln(4/\delta)$ samples are needed to give an (ϵ, δ)-ras to the number of linear extensions.

5.2 | Permanent of a {0, 1}-matrix

Let S_n be the set of permutations on $\{1, \dots, n\}$. Then the permanent of a matrix A with entries a_{ij} is

$$\sum_{\tau \in S_n} \prod_{i=1}^{n} a_{i,\tau(i)}.$$

Calculating the permanent exactly was shown by Valiant [27] to be a #P-complete problem.

Note that if $a_{ij} \in \{0, 1\}$, then the only permutations τ that contribute to the sum have $a_{i,\tau(i)} = 1$ for all i. So the permanent is the normalizing constant of the distribution over S_n with unnormalized density $f(\tau) = \prod_{i=1}^{n} a_{i,\tau(i)}$.

Jerrum, Sinclair, and Vigoda [19] developed a polynomial time algorithm for approximately sampling from the density $f(\tau)$. As with the previous problem of linear extensions, for such a problem on permutations there exists a value i such that $\mathbb{P}(\tau(n) = i) \geq 1/n$. This can then be used with the basic self-reducibility process to get an (ϵ, δ)-ras for the permanent. Without going into details (as the method of [19] for approximation was more complex than the basic approach) the result is the same as for linear extensions: use of the methods of this paper immediately reduces the constant in the leading term down to the optimal value.

5.3 | Gibbs distributions

These distributions arise in statistical physics and other applications.

Definition 1 $\{\pi_\beta\}_{\beta\in\mathbb{R}}$ is a *Gibbs distribution with parameter* β over finite state space Ω if there exists a *Hamiltonian function* $H(x) : \Omega \to \mathbb{R}$ such that for $X \sim \pi_\beta$,

$$\mathbb{P}(X = x) = \exp(-\beta H(x))/Z(\beta),$$

where $Z(\beta) = \sum_{x\in\Omega} \exp(-\beta H(x))$ is called the *partition function* of the distribution.

A famous example of a Gibbs distribution is the Ising model [15], where the state space consists of labelings of the nodes of a graph $G = (V, E)$ by either 0 or 1, and $H(x) = \sum_{\{v,w\}\in E} -(x(v) - x(w))^2$. In [16] finding the partition function of the ferromagnetic Ising model (where $\beta > 0$) was shown to be a #P-complete problem for general graphs, but that same work showed how to generate (approximately) samples from the distribution in time polynomial in the size of the graph.

Typically it is easy to find $Z(0)$ for these problems. For the Ising model, $Z(0) = 2^{\#V}$. In [26] it was shown how to build an estimate for $Z_\beta/Z(0)$ using samples from π where the ratio σ^2/μ^2 was bounded. In [14], it was shown how to build two random variables W and V such that $\mathbb{E}[W]/\mathbb{E}[V] = Z(\beta)/Z(0)$ and each had relative variance bounded above by $2e$.

Let $\epsilon' = [-1 + \sqrt{1 + \epsilon^2}]/\epsilon \le \epsilon/2 - \epsilon^3(1.5 - \sqrt{2})$ for $\epsilon \in [0, 1]$. If

$$|\hat{\mu}_W - \mathbb{E}[W]| \le \epsilon\mathbb{E}[W] \text{ and } |\hat{\mu}_V - \mathbb{E}[V]| \le \epsilon\mathbb{E}[V],$$

then

$$\frac{\mathbb{E}[W]}{\mathbb{E}[V]}\left(\frac{1-\epsilon'}{1+\epsilon'}\right) \le \frac{\hat{\mu}_W}{\hat{\mu}_V} \le \frac{\mathbb{E}[W]}{\mathbb{E}[V]}\left(\frac{1+\epsilon'}{1-\epsilon'}\right).$$

Then it is straightforward to show that $\hat{\mu} = [\hat{\mu}_W/\hat{\mu}_V]\sqrt{1+\epsilon^2}$ satisfies

$$\hat{\mu} \in [(\mathbb{E}[W]/\mathbb{E}[V])(1-\epsilon), (\mathbb{E}[W]/\mathbb{E}[V])(1+\epsilon)],$$

thereby giving an (ϵ, δ)-ras that (to leading order) requires $2(4\epsilon^{-2})(2e)^2 \ln(4/\delta)$ samples to estimate the partition function value.

ACKNOWLEDGMENTS

This work was completed with the assistance of NSF grant DMS-1418495. The author also gratefully acknowledges the help of the anonymous referee who provided the outline of Section 4 and gave permission for inclusion of these ideas in the final paper.

REFERENCES

1. N. Alon, Y. Matias, and M. Szegedy, *The space complexity of approximating the frequency moments*, J. Comput. Syst. Sci. Int. **58** (1999), 137-147.
2. G. Brightwell and P. Winkler, *Counting linear extensions*, Order. **8** (1991), no. 3, 225-242.
3. R. Bubley and M. Dyer, *Faster random generation of linear extensions*, Discrete Math. **201** (1999), 81-88.

4. O. Catoni, *Challenging the empirical mean and empirical variance: A deviation study*, Ann. Inst. Henri Poincaré Probab. Stat. **48** (2012), 1148-1185.
5. O. Catoni, PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. arXiv: 1603.05229, 2016.
6. O. Catoni and I. Giulini, Dimension free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression with a random design. arXiv: 1712.02747, 2017.
7. H. Chernoff, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Ann. Math. Stat. **23** (1952), 493-509.
8. L. Devroye, M. Lerasle, G. Lugosi, and R.I. Oliveira, *Sub-Gaussian mean estimators*, Ann. Statist. **44** (2016), 2695-2725.
9. M. Dyer, A. Frieze, and R. Kannan, *A random polynomial-time algorithm for approximating the volume of convex bodies*, J. Assoc. Comput. Mach. **38** (1991), no. 1, 1-17.
10. R.D. Gordon, *Values of Mills' ratio of area to bounding ordinate of the normal probability integral for large values of the argument*, Ann. Math. Stat. **12** (1941), 364-366.
11. D. Hsu and S. Sabato, *Heavy-tailed regression with a generalized median-of-means*, Proceedings of the 31st International Conference on Machine Learning, PMLR. **32** (2014), no. 2, 37-45.
12. D. Hsu and S. Sabato, *Loss minimization and parameter estimation with heavy tails*, J. Mach. Learn. Res. **17** (2016), 1-40.
13. M. Huber, *Fast perfect sampling from linear extensions*, Discrete Math. **306** (2006), 420-428.
14. M. Huber, *Approximation algorithms for the normalized constant of Gibbs distributions*, Ann. Appl. Probab. **51** (2015), no. 1, 92-105.
15. E. Ising, *Beitrag zur theorie des ferromagnetismus*, Z. Phys. **31** (1925), 253-258.
16. M. Jerrum and A. Sinclair, *Polynomial-time approximation algorithms for the Ising model*, SIAM J. Comput. **22** (1993), 1087-1116.
17. M. Jerrum and A. Sinclair, *The Markov Chain Monte Carlo method: An approach to approximate counting and integration*. Approximation Algorithms for NP-Hard Problems, PWS (chapter 12, pp. 482-520, 1996.
18. M. Jerrum, L. Valiant, and V. Vazirani, *Random generation of combinatorial structures from a uniform distribution*, Theoret. Comput. Sci. **43** (1986), 169-188.
19. M.R. Jerrum, A. Sinclair, and E. Vigoda, *A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries*, J. ACM. **51** (2004), no. 4, 671-697.
20. R. Kannan, L. Lovasz, and M. Simonovits, *Random walks and an O*(n(5)) volume algorithm for convex bodies*, Random Structures Algorithms. **11** (1997), no. 1, 1-50.
21. A. Karzanov and L. Khachiyan, *On the conductance of order Markov chains*, Order. **8** (1991), no. 1, 7-15.
22. P. Matthews, *Generating a random linear extension of a partial order*, Ann. Probab. **19** (1991), no. 3, 1367-1392.
23. S. Minsker and N. Strawn, Distributed statistical estimation and rates of convergence in normal approximation. arXiv: 1704.02658, 2017.
24. J. Morton, L. Pachter, A. Shiu, B. Sturmfels, and O. Wienand, *Convex rank tests and semigraphoids*, SIAM J. Discrete Math. **23** (2009), no. 2, 1117-1134.
25. A.S. Nemirovsky and D.B. Yudin, *Problem complexity and method efficiency in optimization*, Wiley, New York, 1983.
26. D. Štefankovič, S. Vempala, and E. Vigoda, *Adaptive simulated annealing: A near-optimal connection between sampling and counting*, J. ACM. **56** (2009), no. 3, 1-36.
27. L.G. Valiant, *The complexity of computing the permanent*, Theoret. Comput. Sci. **8** (1979), 189-201.

How to cite this article: Huber M. An optimal (ϵ, δ)-randomized approximation scheme for the mean of random variables with bounded relative variance. *Random Struct Alg.* 2019;55:356–370. https://doi.org/10.1002/rsa.20839