Large-Scale Methods for Distributionally Robust Optimization

Daniel Levy[*] **Yair Carmon**[*], **John C. Duchi and Aaron Sidford**
Stanford University
{danilevy,jduchi,sidford}@stanford.edu,, ycarmon@cs.tau.ac.il

Abstract

We propose and analyze algorithms for distributionally robust optimization of convex losses with conditional value at risk (CVaR) and χ^2 divergence uncertainty sets. We prove that our algorithms require a number of gradient evaluations independent of training set size and number of parameters, making them suitable for large-scale applications. For χ^2 uncertainty sets these are the first such guarantees in the literature, and for CVaR our guarantees scale linearly in the uncertainty level rather than quadratically as in previous work. We also provide lower bounds proving the worst-case optimality of our algorithms for CVaR and a penalized version of the χ^2 problem. Our primary technical contributions are novel bounds on the bias of batch robust risk estimation and the variance of a multilevel Monte Carlo gradient estimator due to Blanchet and Glynn [8]. Experiments on MNIST and ImageNet confirm the theoretical scaling of our algorithms, which are 9–36 times more efficient than full-batch methods.

1 Introduction

The growing role of machine learning in high-stakes decision-making raises the need to train reliable models that perform robustly across subpopulations and environments [11, 29, 70, 58, 36, 53, 39]. Distributionally robust optimization (DRO) [3, 66] shows promise as a way to address this challenge, with recent interest in both the machine learning community [68, 74, 22, 69, 34, 55] and in operations research [20, 3, 5, 27]. Yet while DRO has had substantial impact in operations research, a lack of scalable optimization methods has hindered its adoption in common machine learning practice.

In contrast to empirical risk minimization (ERM), which minimizes an expected loss $\mathbb{E}_{S\sim P_0}\,\ell(x;S)$ over $x \in \mathcal{X} \subset \mathbb{R}^d$ with respect to a training distribution P_0, DRO minimizes the expected loss with respect to the worst distribution in an uncertainty set $\mathcal{U}(P_0)$, that is, its goal is to solve

$$\operatorname*{minimize}_{x\in\mathcal{X}}\ \mathcal{L}(x;P_0) \coloneqq \sup_{Q\in\mathcal{U}(P_0)} \mathbb{E}_{S\sim Q}\,\ell(x;S). \tag{1}$$

The literature considers several uncertainty sets [3, 5, 7, 27], and we focus on two particular choices: (a) the set of distributions with bounded likelihood ratio to P_0, so that $\mathcal{L}$ becomes the conditional value at risk (CVaR) [59, 67], and (b) the set of distributions with bounded χ^2 divergence to P_0 [3, 16]. Some of our results extend to more general ϕ-divergence (or Rényi divergence) balls [72]. Minimizers of these objectives enjoy favorable statistical properties [22, 34], but finding them is more challenging than standard ERM. More specifically, stochastic gradient methods solve ERM with a number of $\nabla\ell$ computations independent of both N, the support size of P_0 (i.e., number of data points), and d, the dimension of x (i.e., number of parameters). These guarantees do not directly apply to DRO because the supremum over Q in (1) makes cheap sampling-based gradient estimates biased. As a consequence, existing techniques for minimizing the χ^2 objective [2, 20, 3, 5, 47, 22] have $\nabla\ell$

[*] Equal contribution. Code is available on GitHub at https://github.com/daniellevy/fast-dro/.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

	CVaR at level α	χ^2 constraint ρ	χ^2 penalty λ
Objective $\mathcal{L}(x; P_0) =$	$\sup_{\|q\|_\infty \le \frac{1}{\alpha N}} q^\top \ell(x)$	$\sup_{\mathrm{D}_{\chi^2}(q) \le \rho} q^\top \ell(x)$	$\sup_{q \in \Delta^N} q^\top \ell(x) - \lambda \mathrm{D}_{\chi^2}(q)$
Subgradient method	$N\epsilon^{-2}$	$N\epsilon^{-2}$	$N\epsilon^{-2}$
Dual SGM [Appendix A.3]	$\alpha^{-2}\epsilon^{-2}$	-	$\lambda^{-2}\epsilon^{-2}$
Subsampling [22]	-	$\rho^2 d\epsilon^{-4}$	-
Stoch. primal-dual [17, 47]	$N\epsilon^{-2}$	$N\rho\epsilon^{-2}$	-
Ours	$\alpha^{-1}\epsilon^{-2}$ (Thm. 2)	$\rho\epsilon^{-3}$ (Thm. 4)	$\lambda^{-1}\epsilon^{-2}$ (Thm. 2)
Lower Bound	$\alpha^{-1}\epsilon^{-2}$ (Thm. 3)	$\rho\epsilon^{-2}$ [22]	$\lambda^{-1}\epsilon^{-2}$ (Thm. 3)

Table 1. Number of $\nabla\ell$ evaluations to obtain $\mathbb{E}[\mathcal{L}(x; P_0)] - \inf_{x' \in \mathcal{X}} \mathcal{L}(x'; P_0) \le \epsilon$ when P_0 is uniform on N training points. For simplicity we omit the Lipschitz constant of ℓ, the size of the domain $\mathcal{X}$, and logarithmic factors. We define $\ell_i(x) \coloneqq \ell(x; S_i)$ and $\mathrm{D}_{\chi^2}(q) \coloneqq \frac{N}{2}\|q - \frac{1}{N}\mathbf{1}\|_2^2$. The suprema are over q in the simplex.

evaluation complexity scaling linearly (or worse) in either N or d, which is prohibitive in large-scale applications.

In this paper, we consider the setting in which ℓ is a Lipschitz convex loss, a prototype case for stochastic optimization and machine learning [75, 49], and we propose methods for solving the problem (1) with $\nabla\ell$ complexity independent of sample size N and dimension d, and with optimal (linear) dependence on the uncertainty set size. In Table 1 we summarize their complexities and compare them to previous work. Each entry of the table shows the number of (sub)gradient evaluations to obtain a point with optimality gap ϵ; for reference, recall that for ERM the stochastic subgradient method requires order ϵ^{-2} evaluations, independent of d and N. We discuss related work further in Section 1.1 after outlining our approach.

We begin our development in Section 3 by considering the surrogate objective$\overline{\mathcal{L}}(x; n) = \mathbb{E}\,\mathcal{L}(x; \widehat{P}_n)$ corresponding to the average empirical robust objective over random batches of size n sampled from P_0. In contrast to (1), it is straightforward to obtain unbiased gradient estimates for $\overline{\mathcal{L}}$—using the mini-batch estimator $\nabla\mathcal{L}(x; \widehat{P}_n)$—and to optimize it efficiently with stochastic gradient methods. To obtain guarantees for the true objective $\mathcal{L}$, we establish uniform bounds on the error $|\mathcal{L}(x; P_0) - \overline{\mathcal{L}}(x; n)|$. For CVaR we prove a bound scaling as $1/\sqrt{n}$ and extend it to other uncertainty sets, including χ^2 balls, via the Kusuoka representation [42]. Notably, for the penalty version of the χ^2 objective (Table 1 right column) we prove a stronger bound scaling as $1/n$. This analysis implies that, for large enough batch size n, an $\epsilon/2$-minimizer of $\overline{\mathcal{L}}$ is also an ϵ-minimizer of $\mathcal{L}$. Furthermore, for CVaR and χ^2 penalty we show that the variance of the gradient estimator decreases as $1/n$, and we use Nesterov acceleration to decrease the required number of gradient steps.

To obtain stronger guarantees, in Section 4 we present a theoretically more efficient multi-level Monte Carlo (MLMC) [31, 32] gradient estimator which is a slight modification of the general technique of Blanchet and Glynn [8]. The resulting estimator is unbiased for $\nabla\overline{\mathcal{L}}(x; n)$ but requires only a *logarithmic number of samples* in n. For CVaR and χ^2 penalty we control the second moment of the gradient estimator, resulting in complexity bounds scaling with ϵ^{-2}. We further prove that these rates are worst-case optimal up to logarithmic factors.

Unfortunately, direct application of the MLMC estimator for the χ^2 uncertainty set (Table 1 center column) demonstrably fails for certain inputs. Instead, in Appendix E we optimize its Lagrange dual—the χ^2 penalty—with respect to x and Lagrange multiplier λ. Using a doubling scheme on the λ domain, we obtain a complexity guarantee scaling as ϵ^{-3}.

Section 5 presents experiments where we use DRO to train linear models for digit classification (on a mixture between MNIST [44] and typed digits [19]), and ImageNet [60]. To the best of our knowledge, the latter is the largest DRO problem solved to date. In both experiments DRO provides generalization improvements over ERM, and we show that our stochastic gradient estimators require far less $\nabla\ell$ computations—between 9× and 40×— than full-batch methods. Our experiments also reveal two facts that our theory only hints at. First, using the mini-batch gradient estimator the error floor due to the difference between $\overline{\mathcal{L}}(x; n)$ and $\mathcal{L}(x; P_0)$ becomes negligible even for batch sizes as small as 10. Second, while the MLMC estimator avoids these error floors altogether, its increased variance makes it practically inferior to the mini-batch estimator with properly tuned batch size and learning rate. Our code implements our gradient estimators in PyTorch [56] and combines them seamlessly with the framework's optimizers; we show an example code snippet in Appendix F.3.

1.1 Related work

Distributionally robust optimization grows from the robust optimization literature in operations research [3, 2, 4, 5], and the fundamental uncertainty about the data distribution at test time makes its application to machine learning natural. Experiments in the papers [47, 28, 22, 34, 17, 40] show promising results for CVaR and χ^2-constrained DRO, while other works highlight the importance of incorporating additional constraints into the uncertainty set definition [38, 24, 55, 61]. Below, we review the prior art on solving these DRO problems at scale.

Full-batch subgradient method. When P_0 has support of size N it is possible to compute a subgradient of the objective $\mathcal{L}(x; P_0)$ by evaluating $\ell(x; s_i)$ and $\nabla\ell(x; s_i)$ for $i = 1, \ldots, N$, computing the $q \in \Delta^N$ attaining the supremum (1), whence $g = \sum_{i=1}^{N} q_i \nabla\ell(x; s_i)$ is a subgradient of $\mathcal{L}$ at x. As the Lipschitz constant of $\mathcal{L}$ is at most that of ℓ, we may use these subgradients in the subgradient method [51] and find an ϵ approximate solution in order ϵ^{-2} steps. This requires order $N\epsilon^{-2}$ evaluations of $\nabla\ell$, regardless of the uncertainty set.

CVaR. Robust objectives of the form (1) often admit tractable expression in terms of joint minimization over x and the Lagrange multipliers associated with the constrained maximization over Q [e.g., 59, 66]. For CVaR, this dual formulation is an ERM problem in x and $\eta \in \mathbb{R}$, which we can solve in time independent of N using stochastic gradient methods. We refer to this as "dual SGM," providing the associated complexity bounds in Appendix A.3. Fan et al. [28] apply dual SGM for learning linear classifiers, and Curi et al. [17] compare it to their proposed stochastic primal-dual method based on determinantal point processes. While the latter performs better in practice, its worst-case guarantees scale roughly as $N\epsilon^{-2}$, similarly to the full-batch method. Kawaguchi and Lu [40] propose to only use gradients from the highest k losses in every batch, which is essentially identical to our mini-batch estimator for CVaR; they do not, however, relate their algorithm to CVaR optimization. We contribute to this line of work by obtaining tight characterizations of the mini-batch and MLMC gradient estimators, resulting in optimal complexity bounds scaling as $\alpha^{-1}\epsilon^{-2}$.

DRO with χ^2 divergence. Similar dual formulations exist for both the constrained and penalized χ^2 objectives, and dual SGM provides similar guarantees to CVaR for the penalized χ^2 objective. For the constrained-χ^2 problem, the additional Lagrange multiplier associated with the constraint induces a "perspective transform" [3, 22], making the method unstable. Indeed, Namkoong and Duchi [47] report that it fails to converge in practice and instead propose a stochastic primal-dual method with convergence rate $(1 + \rho N)\epsilon^{-2}$. Their guarantee is optimal in the weak regularization regime where $\rho \lesssim 1/N$, but is worse than the full-batch method in the setting where $\rho \gtrsim 1$. Hashimoto et al. [34] propose a different scheme alternating between ERM on x and line search over a Lagrange multiplier, but do not provide complexity bounds. Duchi and Namkoong [22] prove that for a sample of size $N' \approx \rho^2 d\epsilon^{-2}$ the empirical objective converges to $\mathcal{L}(x; P_0)$ uniformly in $x \in \mathcal{X}$; substituting N' into the full-batch complexity bound implies a rate of $\rho^2 d\epsilon^{-4}$. This guarantee is independent of N, but features an undesirable dependence on d. Ghosh et al. [30] use the mini-batch gradient estimator and gradually increase the batch size to N as optimization progresses; they do not provide convergence rate bounds. We establish concrete rates for fixed batch sizes independent of N.

MLMC gradient estimators. Multi-level Monte Carlo techniques [31, 32] facilitate the estimation of expectations of the form $\mathbb{E}\,\mathsf{F}(S_1, \ldots, S_n)$, where the S_i are i.i.d. In this work we leverage a variant of a particular MLMC estimator proposed by Blanchet and Glynn [8]. Prior work [6] uses the estimator of [8] in a DRO formulation of semi-supervised learning with Wasserstein uncertainty sets and $\mathsf{F}(\cdot)$ a ratio of expectations, as opposed to a supremum of expectations in our setting.

2 Preliminaries

We collect notation, establish a few assumptions, and provide the most important definitions for the remainder of the paper in this section.

Notation. We denote the optimization variable by $x \in \mathbb{R}^d$, and use s (or S when it is random) for a data sample in $\mathbb{S}$. We use z_l^m as shorthand for the sequence $z_l, \ldots, z_m$. For fixed x we denote the cdf of $\ell(x, S)$ by $F(t) := \mathbb{P}(\ell(x, S) \le t)$ and its inverse by $F^{-1}(u) := \inf\{t : F(t) > u\}$, leaving the dependence on x and P_0 implicit. We use $\|\cdot\|$ to denote Euclidean norm, but remark that many of our results carry over to general norms. We let Δ^m denote the simplex in m dimensions. We write $1_{\{A\}}$ for the indicator of event A, i.e., 1 if A holds and 0 otherwise, and write $\mathbb{I}_{\mathcal{C}}$ for the infinite indicator of the set $\mathcal{C}$, $\mathbb{I}_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$ and $\mathbb{I}_{\mathcal{C}}(x) = \infty$ otherwise. The Euclidean projection to a set $\mathcal{C}$ is $\Pi_{\mathcal{C}}$. We use ∇ to denote gradient with respect to x, or, for non-differentiable convex functions, an arbitrary subgradient. We denote the positive part of $t \in \mathbb{R}$ by $(t)_+ := \max\{t, 0\}$. Finally, $f \lesssim g$ means that there exists $C \in \mathbb{R}_+$, independent of any problem parameters, such that $f \le Cg$ holds; we also write $f \asymp g$ if $f \lesssim g \lesssim f$.

Assumptions. Throughout, we assume that the domain $\mathcal{X}$ is closed convex and satisfies $\|x - y\| \le R$ for all $x, y \in \mathcal{X}$. Moreover, we assume the loss function $\ell : \mathcal{X} \times \mathbb{S} \to [0, B]$ is convex and G-Lipschitz in x, i.e., $0 \le \ell(x, s) \le B$ and $|\ell(x; s) - \ell(y; s)| \le G\|x - y\|$ for $x, y \in \mathcal{X}$ and $s \in \mathbb{S}$.[2] In some cases, we entertain two additional assumptions:

Assumption A1. *The gradient $\nabla\ell(x, s)$ is H-Lipschitz in x.*

Assumption A2. *The inverse cdf F^{-1} of $\ell(x; S)$ is G_{icdf}-Lipschitz for each $x \in \mathcal{X}$.*

Most of our bounds do not require Assumptions A1 and A2. Moreover, in Appendix B.2 we argue that these assumptions are frequently not restrictive.

The distributionally robust objective. We consider a slight generalization of ϕ-divergence distributionally robust optimization (DRO). For a convex $\phi : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ satisfying $\phi(1) = 0$, the ϕ-divergence between distributions P and Q absolutely continuous w.r.t. P by $\mathrm{D}_\phi(Q, P) := \int \phi(\frac{\mathrm{d}Q}{\mathrm{d}P}(s))\mathrm{d}P(s)$. Then, for convex ϕ, ψ with $\phi(1) = \psi(1) = 0$, a constraint radius $\rho \ge 0$, and penalty $\lambda \ge 0$ *the general form of the objectives we consider* is

$$\mathcal{L}(x; P) := \sup_{Q : \mathrm{D}_\phi(Q,P) \le \rho} \Big\{ \mathbb{E}_Q[\ell(x; S)] - \lambda \mathrm{D}_\psi(Q, P) \Big\}. \tag{2}$$

As previewed, we consider the following objectives for general P_0 (nonuniform with infinite support):

- χ^2 **constraint.** $\mathcal{L}_{\chi^2}$ corresponds to $\phi(t) = \chi^2(t) := \frac{1}{2}(t-1)^2$ and $\psi = 0$.
- χ^2 **penalty.** $\mathcal{L}_{\chi^2\text{-pen}}$ corresponds to $\phi = 0$ and $\psi(t) = \chi^2(t) = \frac{1}{2}(t-1)^2$.
- **Conditional value at risk $\alpha \in (0, 1]$ (CVaR).** $\mathcal{L}_{\text{CVaR}}$ corresponds to $\phi = 0$ and $\psi = \mathbb{I}_{[0,1/\alpha)}$.

Additionally, define the following smoothed version of the CVaR objective, which we use in Section 3.

- **KL-regularized CVaR.** $\mathcal{L}_{\text{kl-CVaR}}$ corresponds to $\phi = 0$ and and $\psi(t) = \mathbb{I}_{[0,1/\alpha]}(t) + t\log t - t + 1$.

In Appendix A we present additional standard formulations and useful properties of these objectives.

With mild abuse of notation, for a sample $s_1^n \in \mathbb{S}^n$, we let

$$\mathcal{L}(x; s_1^n) := \mathcal{L}(x; \widehat{P}[s_1^n]) = \sup_{q \in \Delta^n : \sum_{i \le n} \frac{1}{n}\phi(nq_i) \le \rho} \left\{ \sum_{i=1}^{n} \big(q_i \ell(x; s_i) - \tfrac{1}{n}\psi(nq_i)\big) \right\} \tag{3}$$

denote the loss with respect to the empirical distribution on s_1^n. Averaging the robust objective over random batches of size n, we define the surrogate objective

$$\overline{\mathcal{L}}(x; n) := \mathbb{E}_{S_1^n \sim P_0^n}\, \mathcal{L}(x; S_1^n). \tag{4}$$

Complexity metrics. We measure complexity of our methods by the number of computations of $\nabla\ell(x; s)$ they require to reach a solution with accuracy ϵ. We can bound (up to a constant factor) the runtime of every method we consider by our complexity measure multiplied by $d + \mathsf{T}_{\text{eval}}$, where T_{eval} denotes the time to evaluate $\ell(x; s)$ and $\nabla\ell(x; s)$ at a single point x and sample s, and is typically $O(d)$. (In the problems we study, solving the problem (4) given $\ell(x; S_1^n)$ takes $O(n \log n)$ time; see Appendix A.2).

[2] Our results hold also when B denotes $\sup_{x \in \mathcal{X}, s, s' \in \mathbb{S}}\{\ell(x; s) - \ell(x; s')\}$. The Lipschitz loss and bounded domain assumptions imply $B \le B_0 + GR$ if $\inf_{x \in \mathcal{X}} \ell(x; s) - \inf_{x' \in \mathcal{X}} \ell(x'; s') \le B_0$ for all $s, s' \in \mathbb{S}$, which typically holds with $B_0 \approx 0$ in regression and classification problems.

3 Mini-batch gradient estimators

In this section, we develop and analyze stochastic subgradient methods using the subgradients of the mini-batch loss (3). That is, we estimate $\nabla\mathcal{L}(x;P_0)$ by sampling a mini-batch $S_1,\ldots,S_n \overset{\text{iid}}{\sim} P_0$ and computing

$$\nabla\mathcal{L}(x;S_1^n) = \sum_{i=1}^n q_i^\star \nabla\ell(x;S_i),$$

where $q^\star \in \Delta^n$ attains the supremum in Eq. (3). By definition (4) of the surrogate objective $\overline{\mathcal{L}}$, we have that $\mathbb{E}\,\nabla\mathcal{L}(x;S_1^n) = \nabla\overline{\mathcal{L}}(x;n)$. Therefore, we expect stochastic subgradient methods using $\nabla\mathcal{L}(x;S_1^n)$ to minimize $\overline{\mathcal{L}}$. However, in general, $\overline{\mathcal{L}}(x;n) \neq \mathcal{L}(x;P_0)$ and $\mathbb{E}\,\nabla\mathcal{L}(x;S_1^n) \neq \nabla\mathcal{L}(x;P_0)$.

To show that the mini-batch gradient estimator is nevertheless effective for minimizing $\mathcal{L}$, we proceed in three steps. First, in Section 3.1 we prove uniform bounds on the bias $\mathcal{L}-\overline{\mathcal{L}}$ that tend to zero with n. Second, in Section 3.2 we complement them with $1/n$ variance bounds on $\nabla\mathcal{L}(x;S_1^n)$. Finally, Section 3.3 puts the pieces together: we apply the SGM guarantees to bound the complexity of minimizing $\overline{\mathcal{L}}$ to accuracy $\epsilon/2$, using Nesterov acceleration to exploit our variance bounds, and choose the mini-batch size n large enough to guarantee (via our bias bounds) that the resulting solution is also an ϵ minimizer of the original objective $\mathcal{L}$.

3.1 Bias analysis

Proposition 1 (Bias of the batch estimator)**.** *For all $x \in \mathcal{X}$ and $n \in \mathbb{N}$ we have*

$$0 \le \mathcal{L}(x;P_0) - \overline{\mathcal{L}}(x;n) \lesssim \begin{cases} B\min\{1,(\alpha n)^{-1/2}\} & \text{for } \mathcal{L}=\mathcal{L}_{\text{CVaR}} & (5)\\ B\sqrt{(1+\rho)(\log n)/n} & \text{for } \mathcal{L}=\mathcal{L}_{\chi^2} & (6)\\ B^2(\lambda n)^{-1} & \text{for } \mathcal{L}=\mathcal{L}_{\chi^2\text{-pen}} & (7)\\ G_{\text{icdf}}\, n^{-1} & \text{for any loss (2)}, & (8)\end{cases}$$

where the bound (8) holds under Assumption A1.

We present the proof in Appendix B.1.1 and make a few remarks before proceeding to discuss the main proof ideas. First, the bounds (5), (6) and (7) are all tight up to constant or logarithmic factors when $\ell(x,S)$ has a Bernoulli distribution, and so are unimprovable without further assumptions (see Proposition 4 in Appendix B.1.2). One such assumption is that $\ell(x;S)$ has G_{icdf}-Lipschitz inverse-cdf, and it allows us to obtain a general $1/n$ bias bound (8) independent of the uncertainty set size. As we discuss in Appendix B.2.2, this assumption has natural relaxations for uniform distributions with finite supports and, for CVaR at level α, we only need the inverse cdf $F^{-1}(\beta)$ to be Lipschitz around $\beta=\alpha$, a common assumption in the risk estimation literature [71].

Proof sketch. To show that $\mathcal{L}(x;P_0) \ge \overline{\mathcal{L}}(x;n)$ for every loss of the form (2), we use Lagrange duality to write

$$\mathcal{L}(x;P_0) = \inf_{\eta,\nu}\mathbb{E}_{S_1^n\sim P_0^n}\frac{1}{n}\sum_{i=1}^n \Upsilon(x;\eta,\nu;S_i) \ \text{ and } \ \overline{\mathcal{L}}(x;n) = \mathbb{E}_{S_1^n\sim P_0^n}\inf_{\eta,\nu}\frac{1}{n}\sum_{i=1}^n \Upsilon(x;\eta,\nu;S_i),$$

for some $\Upsilon : \mathcal{X}\times\mathbb{R}\times\mathbb{R}_+\times\mathbb{S}\to\mathbb{R}$. This exposes the fundamental source of the mini-batch estimator bias: when infimum and expectation do not commute (as is the case in general), exchanging them strictly decreases the result.

Our upper bound analysis begins with CVaR, where $\mathcal{L}_{\text{CVaR}} = \frac{1}{\alpha}\int 1_{\{\beta\ge 1-\alpha\}}F^{-1}(\beta)\mathrm{d}\beta$ and $\overline{\mathcal{L}}_{\text{CVaR}} = \frac{1}{\alpha}\int \mathcal{I}_\alpha(\beta)F^{-1}(\beta)\mathrm{d}\beta$, with F^{-1} the inverse cdf of $\ell(x,S)$ and $\mathcal{I}_\alpha$ a "soft step function" that we write in closed form as a sum of Beta densities. To obtain the bound (5) we express $\int(1_{\{\beta\ge 1-\alpha\}} - \mathcal{I}_\alpha(\beta))_+\mathrm{d}\beta$ as a sum of binomial tail probabilities and apply Chernoff bounds. For CVaR only, the improved bound (8) follows from arguing that replacing $F^{-1}(\beta)$ with $G_{\text{icdf}}\cdot\beta$ overestimates the bias, and showing that $\int(1_{\{\beta\ge 1-\alpha\}} - \mathcal{I}_\alpha(\beta))\beta\mathrm{d}\beta \le (n+1)^{-1}$ for any α.

To transfer the CVaR bounds to other objectives we express the objective (2) as a weighted CVaR average over different α values, essentially using the Kusuoka representation of coherent risk measures [42]. Given any bias bound $\mathrm{bb}(\alpha)$ for CVaR at level α, this expression implies the bound

$\mathcal{L} - \overline{\mathcal{L}} \le \sup_{w \in \mathcal{W}(\mathcal{L})} \int \mathrm{bb}(\alpha) \mathrm{d}w(\alpha)$, where $\mathcal{W}(\mathcal{L})$ is a set of probability measures. Substituting $\mathrm{bb}(\alpha) = 1/\sqrt{n\alpha}$ and using the Cauchy-Schwartz inequality gives the bound (6), while substituting $\mathrm{bb}(\alpha) = G_{\text{icdf}}/n$ shows this bound in fact holds for any $\mathcal{L}$, as we claim in (8).

Showing the bound (7) requires a fairly different argument. Our proof uses the dual representation of $\mathcal{L}_{\chi^2\text{-pen}}$ as a minimum of an expected risk over a Lagrange multiplier η imposing the constraint that q in (3) sums to 1 (or that Q in (2) integrates to 1). Using convexity with respect to η we relate the value of the risk at η_n (the minimizer for sample S_1^n) to $\eta^\star$ (the population minimizer), which on expectation are $\overline{\mathcal{L}}_{\chi^2\text{-pen}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$, respectively. We then apply Cauchy-Schwartz and bound the variance of η_n with the Efron-Stein inequality [26] to obtain a $1/n$ bias bound. □

3.2 Variance analysis

With the bias bounds in Proposition 4 established, we analyze the variance of the stochastic gradient estimators $\nabla \mathcal{L}(x; S_1^n)$. More specifically, we prove that the variance of the mini-batch gradient estimator decreases as $1/n$ for penalty-type robust objectives (with $\phi = 0$) for which the maximizing Q has bounded χ^2 divergence from P_0, which we call "χ^2-bounded objectives" (see Appendix A.4). Noting that $\mathcal{L}_{\text{kl-CVaR}}$ (with $\mathcal{L}_{\text{CVaR}}$ as a special case) and $\mathcal{L}_{\chi^2\text{-pen}}$ are χ^2-bounded yields the following.

Proposition 2 (Variance of the batch estimator)**.** *For all $n \in \mathbb{N}$, $x \in \mathcal{X}$, and $S_1^n \sim P_0^n$,*

$$\mathrm{Var}\Big[\nabla \mathcal{L}_{\text{kl-CVaR}}(x; S_1^n)\Big] \lesssim \frac{G^2}{\alpha n} \quad \textit{and} \quad \mathrm{Var}\Big[\nabla \mathcal{L}_{\chi^2\text{-pen}}(x; S_1^n)\Big] \lesssim \frac{G^2(1 + B/\lambda)}{n}.$$

(Note that the variance bound on $\mathcal{L}_{\text{kl-CVaR}}$ is independent of λ and therefore holds also for $\mathcal{L}_{\text{CVaR}}$ where $\lambda = 0$). We prove Proposition 2 in Appendix B.3 and provide a proof sketch below.[3] Unfortunately, the bounds do not extend to the χ^2 constrained formulation: in Appendix B.3 (Proposition 5) we prove that for any n there exist ℓ, P_0, and x such that $\mathrm{Var}[\nabla \mathcal{L}_{\chi^2}(x; P_0)] \gtrsim \rho$. Whether Proposition 2 holds when adding a χ^2 penalty to the χ^2 constraint remains an open question.

Proof sketch. The Efron-Stein inequality [26] is $\mathrm{Var}[\nabla \mathcal{L}(x; S_1^n)] \le \frac{n}{2} \mathbb{E} \|\nabla \mathcal{L}(x; S_1^n) - \nabla \mathcal{L}(x; \tilde{S}_1^n)\|^2$, where S_1^n and $\tilde{S}_1^n$ are identical except in a random entry $I \in [n]$ for which $\tilde{S}_I$ is an i.i.d. copy of S_I. We bound $\|\nabla \mathcal{L}(x; S_1^n) - \nabla \mathcal{L}(x; \tilde{S}_1^n)\| \le G q_I + G\|q - \tilde{q}\|_1$ with the triangle inequality, where q and $\tilde{q}$ attain the maximum in (3) for S and $\tilde{S}$, respectively. The crux of our proof is the equality $\|q - \tilde{q}\|_1 = 2|q_I - \tilde{q}_I|$, which holds since increasing one coordinate of $\ell(x; S_1), \ldots, \ell(x; S_n)$ must decrease all other coordinates in q. Noting that $\mathbb{E}\,(q_I - \tilde{q}_I)^2 \le 4\,\mathbb{E}(q_I - 1/n)^2 = \frac{8}{n^2} \mathbb{E}\, \mathrm{D}_{\chi^2}(q, \frac{1}{n}\mathbf{1})$, the results follow by observing that $\mathrm{D}_{\chi^2}(q, \frac{1}{n}\mathbf{1})$ is bounded by $1/\alpha$ and B/λ for $\mathcal{L}_{\text{kl-CVaR}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$, respectively. □

3.3 Complexity guarantees

With the bias and variance guarantees established, we now provide bounds on the complexity of minimizing $\mathcal{L}(x; P_0)$ to arbitrary accuracy ϵ using standard gradient methods with the gradient estimator $\tilde{g}(x) = \nabla \mathcal{L}(x; S_1^n)$. (Recall from Section 2 that we measure complexity by the number of individual first order evaluations $(\ell(x; s), \nabla \ell(x; s))$.) Writing $\Pi_{\mathcal{X}}$ for the Euclidean projection onto $\mathcal{X}$, the stochastic gradient method (SGM) with fixed step-size η and $x_0 \in \mathcal{X}$ iterates

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta \tilde{g}(x_t)), \text{ and } \bar{x}_t = \frac{1}{t} \sum_{\tau \le t} x_\tau. \tag{9}$$

We also consider Nesterov's accelerated gradient method [50, 43]. For $x_0 = y_0 = z_0 \in \mathcal{X}$, a fixed step-size $\eta > 0$ and a sequence $\{\theta_t\}$, we iterate

$$z_{t+1} = \Pi_{\mathcal{X}}(z_t - \tfrac{\eta}{\theta_t} \tilde{g}(x_t)),\; y_{t+1} = \theta_t z_{t+1} + (1 - \theta_t) y_t, \text{ and } x_{t+1} = \theta_{t+1} z_{t+1} + (1 - \theta_t) y_{t+1}. \tag{10}$$

In Appendix B.4, we state the rates of convergence of the iterations (9) and (10) following the analysis in [43], with a small variation where the stochastic gradient estimates are unbiased for a uniform approximation of the true objective with additive error δ. Since our gradient estimator has norm bounded by G, SGM allows us to find an ϵ-minimizer of $\overline{\mathcal{L}}$ in $T \asymp (GR)^2/\epsilon^2$ steps. Therefore,

[3] In the appendix we provide bounds on the variance of $\mathcal{L}(x; S_1^n)$ in addition to $\nabla \mathcal{L}(x; S_1^n)$.

choosing n large enough in accordance to Proposition 1 guarantees that we find an ϵ-minimizer of $\mathcal{L}$. The accelerated scheme (10) admits convergence guarantees that scale with the gradient estimator variance instead of its second moment, allowing us to leverage Proposition 2 to reduce T to the order of $1/\epsilon$. The accelerated guarantees require the loss $\mathcal{L}$ to have order $1/\epsilon$-Lipschitz gradients—fortunately, this holds for $\mathcal{L}_{\chi^2\text{-pen}}$ and $\mathcal{L}_{\text{kl-CVaR}}$.

Claim 1. *Let Assumption A1 hold. For all P, $\nabla\mathcal{L}_{\text{kl-CVaR}}(x;P)$ and $\nabla\mathcal{L}_{\chi^2\text{-pen}}(x;P)$ are $(\frac{G^2}{\lambda}+H)$-Lipschitz in x, and $0 \le \mathcal{L}_{\text{CVaR}}(x;P) - \mathcal{L}_{\text{kl-CVaR}}(x;P) \le \lambda\log(1/\alpha)$ for all x.*

See proof in Appendix A.1.6. Thus, to minimize $\mathcal{L}_{\text{CVaR}}$ we instead minimize $\mathcal{L}_{\text{kl-CVaR}}$ and choose $\lambda \asymp \epsilon/\log(1/\alpha)$ to satisfy the smoothness requirement while incurring order ϵ approximation error. For $\mathcal{L}_{\chi^2\text{-pen}}$ with $\lambda \ge \epsilon$ we get sufficient smoothness for free.[4]

As computing every gradient estimator requires n evaluations of $\nabla\ell$, the total gradient complexity is nT, and we have the following suite of guarantees (see Appendix B.5 for proof).

Theorem 1. *Let Assumptions A1 and A2 hold, possibly trivially (with $H=\infty$ or $G_{\text{icdf}}=\infty$). Let $\epsilon \in (0,B)$ and write $\nu = \frac{H}{G^2}\epsilon$. With suitable choices of the batch size n and iteration count T, the gradient methods (9) and (10) find $\bar{x}$ satisfying $\mathbb{E}\,\mathcal{L}(\bar{x},P_0) - \inf_{x'\in\mathcal{X}}\mathcal{L}(x';P_0) \le \epsilon$ with complexity nT admitting the following bounds.*

- *For $\mathcal{L}=\mathcal{L}_{\text{CVaR}}$, we have $nT \lesssim \frac{(GR)^2}{\alpha\epsilon^2}\left(1+\min\left\{\frac{\alpha G_{\text{icdf}}\sqrt{\log\frac{1}{\alpha}+\nu}}{GR}, \frac{B^2\sqrt{\log\frac{1}{\alpha}+\nu}}{GR\epsilon}, \frac{B^2}{\epsilon^2}\right\}\right)$.*
- *For $\mathcal{L}=\mathcal{L}_{\chi^2\text{-pen}}$ with $\lambda \le B$, we have $nT \lesssim \frac{(GR)^2B}{\lambda\epsilon^2}\left(1+\min\left\{\frac{B}{GR}\sqrt{\frac{\epsilon(1+\nu)}{\lambda}}, \frac{B}{\epsilon}\right\}\right)$.*
- *For $\mathcal{L}=\mathcal{L}_{\chi^2}$, we have $nT \lesssim \frac{(1+\rho)(GR)^2B^2}{\epsilon^4}\log\frac{(1+\rho)B^2}{\epsilon^2}$.*
- *For any loss of the from (2), we have $nT \lesssim \frac{(GR)^2G_{\text{icdf}}}{\epsilon^3}$.*

The smoothness parameter H only appears in rates resulting from Nesterov acceleration. Even there, H appears in lower-order terms in ϵ since $\nu = \frac{H}{G^2}\epsilon$. We also note that the final $G_{\text{icdf}}\epsilon^{-3}$ rate holds even when the uncertainty set is the entire simplex; therefore, when $G_{\text{icdf}} < \infty$ it is possible to approximately minimize the maximum loss [64] in sublinear time. Theorem 1 achieves the claimed rates of convergence in Table 1 in certain settings. In particular, it recovers the rates for $\mathcal{L}_{\text{CVaR}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$ (the first and last column of the table) when $\nu \lesssim 1$, $\lambda \gtrsim (B/(GR))^2\epsilon$, and $\alpha \lesssim GR/G_{\text{icdf}}$. In the next section, we show how to attain the claimed optimal rates for $\mathcal{L}_{\text{CVaR}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$ without conditions, returning to address the rates for the constrained χ^2 objective $\mathcal{L}_{\chi^2}$ in Appendix E.

4 Multi-level Monte Carlo (MLMC) gradient estimators

In the previous section, we optimized the mini-batch surrogate $\overline{\mathcal{L}}(x;n)$ to the risk $\mathcal{L}(x;P_0)$, using Proposition 1 to guarantee the surrogate's fidelity for sufficiently large n. The increasing (linear) complexity of computing the estimator $\nabla\mathcal{L}(x;S_1^n)$ as n grows limits the (theoretical) efficiency of the method. To that end, in this section we revisit a multi-level Monte Carlo (MLMC) gradient estimator of Blanchet and Glynn [8] to form an unbiased approximation to $\nabla\overline{\mathcal{L}}(x;n)$ whose sample complexity is logarithmic in n. We provide new bounds on the variance of this MLMC estimator, leading immediately to improved (and, as we shall see, optimal) efficiency estimates for stochastic gradient methods using it.

To define the estimator, let $J \sim \min\{\mathsf{Geo}(1/2), j_{\max}\}$ be a truncated geometric random variable supported on $\{1,\ldots,j_{\max}\}$, and let $q(j) = \mathbb{P}(J=j) = 2^{-j+1_{\{j=j_{\max}\}}}$. For a realization of J we draw a sample of size $2^J n_0$ and compute the multi-level Monte-Carlo estimator as follows:

$$\widehat{\mathcal{M}}[\nabla\mathcal{L}] := \nabla\mathcal{L}(x;S_1^{n_0}) + \frac{1}{q(J)}\widehat{\mathcal{D}}_{2^J n_0}, \text{ where } \widehat{\mathcal{D}}_k := \nabla\mathcal{L}(x;S_1^k) - \frac{\nabla\mathcal{L}(x;S_1^{k/2}) + \nabla\mathcal{L}(x;S_{k/2+1}^k)}{2}. \tag{11}$$

[4] We can also handle the case $\lambda < \epsilon$ by adding a KL-divergence term to ψ for $\mathcal{L}_{\chi^2\text{-pen}}$.

Our estimator differs from the proposal [8] in two aspects: the distribution of J and the option to set $n_0 > 1$. As we further discuss in Appendix C.3, the former difference is crucial for our setting, while the latter is pratically and theoretically helpful yet not crucial. The following properties of the MLMC estimator are key to our analysis (see Appendix C.1 for proofs).

Claim 2. *The estimator $\widehat{\mathcal{M}}[\nabla \mathcal{L}]$ with parameters $n = 2^{j_{\max}} n_0$ satisfies*

$\mathbb{E}\,\widehat{\mathcal{M}}[\nabla L] = \mathbb{E}\,\nabla \mathcal{L}(x; S_1^n) = \nabla \overline{\mathcal{L}}(x; n)$, *requiring expected sample size* $\mathbb{E}\,2^J n_0 = n_0(1+\log_2(n/n_0))$.

Proposition 3 (Second moment of MLMC gradient estimator)**.** *For all $x \in \mathcal{X}$, the multi-level Monte Carlo estimator with parameters n and n_0 satisfies*

$$\mathbb{E}\left\|\widehat{\mathcal{M}}[\nabla \mathcal{L}_{\text{CVaR}}]\right\|^2 \lesssim \left(1 + \frac{\log \frac{n}{n_0}}{\alpha n_0}\right) G^2 \;\; \textit{and} \;\; \mathbb{E}\left\|\widehat{\mathcal{M}}[\nabla \mathcal{L}_{\chi^2\text{-pen}}]\right\|^2 \lesssim \left(1 + \frac{B \log \frac{n}{n_0}}{\lambda n_0}\right) G^2.$$

Claim 2 follows from a simple calculation, while the core of Proposition 3 is a sign-consistency argument for simplifying a 1-norm, similar to the proof of Proposition 2.

Further paralleling Proposition 2, we obtain similar bounds on the MLMC estimates of $\mathcal{L}_{\text{CVaR}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$ (in addition to their gradients), and demonstrate that similar bounds fail to hold for $\nabla \mathcal{L}_{\chi^2}$ (Proposition 7 in Appendix C.1). Therefore, directly using the MLMC estimator on $\nabla \mathcal{L}_{\chi^2}$ cannot provide guarantees for minimizing $\mathcal{L}_{\chi^2}$; instead, in Appendix E we develop a doubling scheme that minimizes the dual objective $\mathcal{L}_{\chi^2\text{-pen}}(x; P_0) + \lambda \rho$ jointly over x and λ. This scheme relies on MLMC estimators for both the gradient $\nabla \mathcal{L}_{\chi^2\text{-pen}}$ and the derivative of $\mathcal{L}_{\chi^2\text{-pen}}$ with respect to λ.

Proposition 3 guarantees that the second moment of our gradient estimators remain bounded by a quantity that depends logarithmically on n. For these estimators, Proposition 6 thus directly provides complexity guarantees to minimize $\mathcal{L}_{\text{CVaR}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$. We also provide a high probability bound on the total complexity of the algorithm using a one-sided Bernstein concentration bound. We state the guarantee below and present a short proof in Appendix C.2.

Theorem 2 (MLMC complexity guarantees)**.** *For $\epsilon \in (0, B)$, set $n \asymp \frac{B^2}{\alpha \epsilon^2}$, $1 \lesssim n_0 \lesssim \frac{\log n}{\alpha}$ and $T \asymp \frac{(GR)^2}{n_0 \alpha \epsilon^2} \log^2 n$. The stochastic gradient iterates* (9) *with $\tilde{g}(x) = \widehat{\mathcal{M}}[\nabla \mathcal{L}_{\text{CVaR}}(x; \cdot)]$ satisfy $\mathbb{E}[\mathcal{L}_{\text{CVaR}}(\bar{x}_T; P_0)] - \inf_{x \in \mathcal{X}} \mathcal{L}_{\text{CVaR}}(x; P_0) \le \epsilon$ with complexity at most*

$$n_0 \log_2\left(\frac{n}{n_0}\right) T + 5\sqrt{(n \log n)^2 + n_0 n T \log n} \lesssim \frac{(GR + B)^2}{\alpha \epsilon^2} \log^2 \frac{B^2}{\alpha \epsilon^2} \;\; \textit{w.p} \;\; \ge 1 - \frac{1}{n}.$$

The same conclusion holds when replacing $\mathcal{L}_{\text{CVaR}}$ with $\mathcal{L}_{\chi^2\text{-pen}}$ and α^{-1} with $1 + B/\lambda$.

Lower bounds. We match the guarantees of Theorem 2 with lower bounds that hold in a standard *stochastic oracle* model [48, 43, 10], where algorithms interact with a problem instance by iteratively querying $x_t \in \mathcal{X}$ (for $t \in \mathbb{N}$) and observing $\ell(x_t; S)$ and $\nabla \ell(x_t; S)$ with $S \sim P_0$ (independent of x_t). All algorithms we consider fit into this model, with each gradient evaluation corresponding to an oracle query. Therefore, to demonstrate that our MLMC guarantees are unimprovable in the worst case (ignoring logarithmic factors), we formulate a lower bound on the number of queries any oracle-based algorithm requires.

Theorem 3 (Minimax lower bounds)**.** *Let $G, R, \alpha, \lambda > 0$, $\epsilon \in (0, GR/64)$, and sample space $\mathbb{S} = [-1, 1]$. There exists a numerical constant $c > 0$ such that the following holds.*

- *For each $d \ge 1$, domain $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| \le R\}$, and any algorithm, there exists a distribution P_0 on $\mathbb{S}$ and convex G-Lipschitz loss $\ell : \mathcal{X} \times \mathbb{S} \to [0, GR]$ such that*

$$T \le c \frac{(GR)^2}{\alpha \epsilon^2} \;\; \textit{implies} \;\; \mathbb{E}[\mathcal{L}_{\text{CVaR}}(x_T; P_0)] - \inf_{x' \in \mathcal{X}} \mathcal{L}_{\text{CVaR}}(x'; P_0) > \epsilon.$$

- *There exists $d_\epsilon \lesssim (GR)^2 \epsilon^{-2} \log \frac{GR}{\epsilon}$ such that for $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| \le R\}$, the same conclusion holds when replacing $\mathcal{L}_{\text{CVaR}}$ with $\mathcal{L}_{\chi^2\text{-pen}}$ and α with $\lambda/(GR)$.*

We present the proof in Appendix D. Our proof for the penalized χ^2 lower bound leverages a classical high-dimensional hard instance construction for oracle-based optimization, while our proof for CVaR is information-theoretic. Consequently, the CVaR lower bound is stronger: it holds for $d = 1$ and extends to a global model where at every round the oracle provides the entire function $\ell(\cdot; S)$ rather than $\ell(x; S)$ and $\nabla \ell(x; S)$ at the query point x.

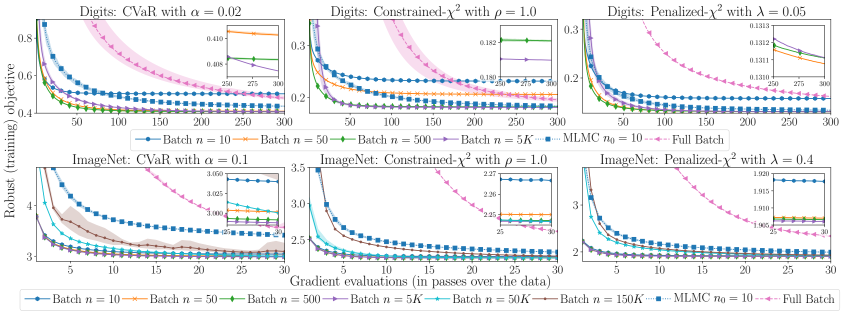

Figure 1. Convergence of DRO objective in our digits and ImageNet classification experiments. Shaded areas indicate range of variability across 5 repetitions (minimum to maximum), and the zoomed-in regions highlight the (often very low) “bias floor” of small batch sizes.

5 Experiments

We test our theoretical predictions with experiments on two datasets. Our main focus is measuring how the total work in solving the DRO problems depends on different gradient estimators. In particular, we quantify the tradeoffs in choosing the mini-batch size n in the estimator $\nabla\mathcal{L}(x; S_1^n)$ of Section 3 and the effect of using the MLMC technique of Section 4. To ensure that we operate in practically meaningful settings, our experiments involve heterogeneous data, and we tune the DRO objective to improve the generalization performance of ERM on the hardest subpopulation. We provide a full account of experiments in Appendix F and summarize them below.

Our digit recognition experiment reproduces [22, Section 3.2], where the training data includes the 60K MNIST training images mixed with 600 images of typed digits from [19], while our ImageNet experiment uses the ILSVRC-2012 1000-way classification task. In each experiment we use DRO to learn linear classifiers on top of pretrained neural network features (i.e., training the head of the network), taking ℓ to be the logarithmic loss with squared-norm regularization; see Appendix F.1. Each experiments compares different gradient estimators for minimizing the $\mathcal{L}_{\text{CVaR}}$, $\mathcal{L}_{\chi^2}$ and $\mathcal{L}_{\chi^2\text{-pen}}$ objectives. Appendix F.2 details our hyper-parameter settings and their tuning procedures.

Figure 1 plots the training objective as optimization progresses. In Appendix F.4 we provide expanded figures that also report the robust generalization performance. We find that the benefits of DRO manifest mainly when the metric of interest is continuous (e.g., log loss) as opposed to the 0-1 loss.

Our analysis in Section 3.1 bounds the suboptimality of solutions resulting from using a mini-batch estimators with batch size n, showing it must vanish as n increases. Figure 1 shows that smaller batch sizes indeed converge to suboptimal solutions, and that their suboptimality becomes negligible very quickly: essentially every batch size larger than 10 provides fairly small bias (with the exception of $\mathcal{L}_{\chi^2}$ in the digits experiment). The effect of bias is particularly weak for $\mathcal{L}_{\chi^2\text{-pen}}$, consistent with its superior theoretical guarantees. We note, however, that the suboptimality we see in practice is far smaller than the worst-case bounds in Proposition 1; we investigate this in detail in Appendix F.5.

While the MLMC estimator does not suffer from a bias floor (by design), it is also much slower to converge. This may appear confusing, since the MLMC convergence guarantees are optimal (for $\mathcal{L}_{\text{CVaR}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$) while the mini-batch estimator achieves the optimal rate only under certain assumptions. Recall, however, that these assumptions are smoothness of the loss (which holds in our experiments) and—for CVaR—sufficiently rapid decay of the bias floor, which we verify empirically.

For batch sizes in the range 50–5K, the traces in Figure 1 look remarkably similar. This is consistent with our theoretical analysis for $\mathcal{L}_{\text{CVaR}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$, which shows that the variance decreases linearly with the batch size and we may therefore (with Nesterov acceleration) increase the step size proportionally and expect the total work to remain constant. As theory predicts, this learning rate increase is only possible up to a certain batch size (roughly 5K in our experiments), after which larger batches become less efficient. Indeed, to reach within 2% of the optimal value, the full-batch method requires 27–36× more work than batch sizes 50–5K for ImageNet, and 9–16× more work for the digits experiment (see Table 4 and 5 for a precise breakdown of these numbers).

We also repeat our experiments with the dual SGM and prima-dual methods mentioned in Table 1 and compare them with them our proposed method; see Appendix F.6 for details.

Broader Impact

The robustness of machine learning (ML) models, or lack thereof, has far-reaching present and future societal consequences: in autonomous vehicles [39, 18], medical diagnosis [53], facial recognition [11], credit scoring [29], and recidivism prediction [12, 1], failure of ML to perform robustly across sub-population or under distribution shift can have disastrous real-life consequences, particularly for members of underserved and/or under-represented groups.

Distributionally robust optimization (DRO) is emerging as a methodology for imposing the constraint that models perform uniformly well across subgroups, and several works conduct experiments demonstrating its benefit in promoting fairness [34, 24, 74] and robustness [68, 61] in ML. However, the computational experiments in these works are relatively small in scale, and there exist serious computational impediments to scaling up DRO. Consequently, the potential benefits of several DRO formulations remain unexplored.

The main contribution of our work is in strengthening the theoretical and algorithmic foundations of two fundamental DRO formulations. In particular, for χ^2-divergence uncertainty sets we give the first proof that stochastic gradient methods can scale to large data similarly to they way they scale for standard empirical risk minimization. We believe that our algorithms will serve a basis for future experimentation with CVaR and χ^2 divergence DRO, and we hope that the resulting findings would lead to more robust and fair machine learning algorithms with positive societal impact. Towards that end, we will release an implementation of our DRO gradient estimators that integrates seamlessly into the PyTorch optimization framework and is therefore suitable for application in a wide range of ML tasks.

In addition, we believe that our work is a step towards a suite of algorithms capable of solving a broader class of DRO problems at scale, including e.g., uncertainty set with explicit group structure as proposed in [38, 61]. We believe that such algorithm suite will empower machine learning researchers and engineers to create more reliable and ethical systems.

However, greater applicability and simplicity always comes with the risk of irresponsible and superficial use. In particular, we are concerned with the possibility that DRO might become a marketing scheme to sell off ML systems as robust without proper verification. Therefore, the development of robust training procedures must go hand-in-hand with the development of rigorous and independent evaluation methodologies for auditing of algorithms [36, 54, 41, 14, 45].

Acknowledgments

The authors would like to thank Hongseok Namkoong for discussions and insights, as well as Nimit Sohoni for comments on an earlier draft.

Sources of funding

DL, YC and JCD were supported by the NSF under CAREER Award CCF-1553086 and HDR 1934578 (the Stanford Data Science Collaboratory) and Office of Naval Research YIP Award N00014-19-2288. YC was supported by the Stanford Graduate Fellowship. AS is supported by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research gift, and a Sloan Research Fellowship.

Competing interests

The authors declare no competing interests.

References

[1] S. Barocas and A. D. Selbst. Big data's disparate impact. *104 California Law Review*, 3: 671–732, 2016.

[2] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[3] A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357, 2013.

[4] D. Bertsimas, D. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[5] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming, Series A*, 167(2):235–292, 2018.

[6] J. Blanchet and Y. Kang. *Semi-supervised Learning Based on Distributionally Robust Optimization*, chapter 1, pages 1–33. John Wiley & Sons, Ltd, 2020. ISBN 9781119721871.

[7] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[8] J. H. Blanchet and P. W. Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE, 2015.

[9] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[10] G. Braun, C. Guzmán, and S. Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7), 2017.

[11] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[12] A. Chouldechova. A study of bias in recidivism prediciton instruments. *Big Data*, pages 153–163, 2017.

[13] K. Clarkson, E. Hazan, and D. Woodruff. Sublinear optimization for machine learning. *Journal of the Association for Computing Machinery*, 59(5), 2012.

[14] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023*, 2018.

[15] N. Cressie and T. R. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, pages 440–464, 1984.

[16] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary*, 2:299–318, 1967.

[17] S. Curi, K. Levy, S. Jegelka, A. Krause, et al. Adaptive sampling for stochastic risk-averse learning. *arXiv:1910.12511 [cs.LG]*, 2019.

[18] D. Dai and L. Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018.

[19] T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images. In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, February 2009.

[20] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[21] J. C. Duchi. Introductory lectures on stochastic convex optimization. In *The Mathematics of Data*, IAS/Park City Mathematics Series. American Mathematical Society, 2018.

[22] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, to appear, 2020.

[23] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[24] J. C. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses against mixture covariate shifts. 2019.

[25] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge University Press, 2019.

[26] B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3): 586–596, 1981.

[27] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming, Series A*, 171(1–2):115–166, 2018.

[28] Y. Fan, S. Lyu, Y. Ying, and B. Hu. Learning with average top-k loss. In *Advances in Neural Information Processing Systems 30*, pages 497–505, 2017.

[29] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. Predictably unequal? the effects of machine learning on credit markets. *Social Science Research Network: 3072038*, 2018.

[30] S. Ghosh, M. Squillante, and E. Wollega. Efficient stochastic gradient descent for distributionally robust learning. *arXiv:1805.08728 [stats.ML]*, 2018.

[31] M. B. Giles. Multilevel Monte Carlo path simulation. *Operations research*, 56(3):607–617, 2008.

[32] M. B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.

[33] C. Guzmán and A. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015.

[34] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[35] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[36] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the Seventh International Conference on Learning Representations*, 2019.

[37] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1993.

[38] W. Hu, G. Niu, I. Sato, and M. Sugiayma. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[39] N. Kalra and S. M. Paddock. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, 94:182–193, 2016.

[40] K. Kawaguchi and H. Lu. Ordered SGD: A new stochastic optimization framework for empirical risk minimization. In *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics*, 2020.

[41] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv:1711.05144 [cs.LG]*, 2018.

[42] S. Kusuoka. On law invariant coherent risk measures. In *Advances in Mathematical Economics*, pages 83–95. Springer, 2001.

[43] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming, Series A*, 133(1–2):365–397, 2012.

[44] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. In *International Conference on Artificial Neural Networks*, pages 53–60, 1995.

[45] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Conference on Fairness, Accountability and Transparency*, pages 220–229, 2019.

[46] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[47] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, 2016.

[48] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

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

[50] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[51] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[52] Y. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming, Series A*, 103:127–152, 2005.

[53] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020.

[54] M. O'Kelly, A. Sinha, H. Namkoong, J. Duchi, and R. Tedrake. Scalable end-to-end autonomous vehicle testing via rare-event simulation. In *Advances in Neural Information Processing Systems 30*, 2018.

[55] Y. Oren, S. Sagawa, T. Hashimoto, and P. Liang. Distributionally robust language modeling. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[56] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *Neural Information Processing Systems (NIPS) Workshop on Automatic Differentiation*, 2017.

[57] J. Pitman. *Probability*. Springer-Verlag, 1993.

[58] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[59] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2: 21–42, 2000.

[60] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[61] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the Eighth International Conference on Learning Representations*, 2020.

[62] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[63] S. Shalev-Shwartz and Y. Singer. Convex repeated games and fenchel duality. In *Advances in Neural Information Processing Systems 19*, 2006.

[64] S. Shalev-Shwartz and Y. Wexler. Minimizing the maximal loss: How and why? In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[65] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 71–79, 2013.

[66] A. Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27 (4):2258–2275, 2017.

[67] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society, 2009.

[68] A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the Sixth International Conference on Learning Representations*, 2018.

[69] M. Staib and S. Jegelka. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems 32*, pages 9134–9144, 2019.

[70] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011.

[71] A. A. Trindade, S. Uryasev, A. Shapiro, and G. Zrazhevsky. Financial prediction with constrained tail risk. *Journal of Banking & Finance*, 31(11):3524–3538, 2007.

[72] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[73] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[74] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. I. Jordan. Robust optimization for fairness with noisy protected groups. *arXiv:2002.09343 [cs.LG]*, 2020.

[75] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

Appendix

A Extended preliminaries

In this section we collect several basic results which we use in subsequent derivations in the paper: Section A.1 gives several additional characterization of the robust objective $\mathcal{L}$, Section A.2 briefly discusses the computation of $\mathcal{L}$ and its costs, Section A.3 gives a short derivation of the complexity guarantees for "dual SGM" in Table 1, and Section A.4 introduces the notion of losses contained in a χ^2 divergence ball. Finally, Section A.5 lists a few standard probabilistic bounds.

A.1 Characterization of the robust objective

Here we give several equivalent characterizations of the robust objective

$$\mathcal{L}(x; P) \coloneqq \sup_{Q \ll P : \mathrm{D}_\phi(Q,P) \le \rho} \Big\{ \mathbb{E}_{S \sim Q}[\ell(x; S)] - \lambda \mathrm{D}_\psi(Q, P) \Big\}. \tag{12}$$

where ψ, ϕ are closed convex functions from $\mathbb{R}_+$ to $\mathbb{R}$ satisfying $\psi(1) = \phi(1) = 0$,

$$\mathrm{D}_\phi(Q, P) \coloneqq \int \phi\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}P, \text{ and } \mathrm{D}_\psi(Q, P) \coloneqq \int \psi\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}P.$$

For $\widehat{P}[s_1^n]$ uniform on $s_1, s_2, \ldots, s_n$ (which we abbreviate s_1^n), we write

$$\mathcal{L}(x; s_1^n) \coloneqq \mathcal{L}(x; \widehat{P}[s_1^n]) = \sup_{q \in \Delta^n : \sum_{i \le n} \frac{1}{n} \phi(nq_i) \le \rho} \left\{ \sum_{i \le n} \big(q_i \ell(x; s_i) - \tfrac{1}{n} \psi(nq_i) \big) \right\}. \tag{13}$$

A.1.1 Inverse-cdf formulation

Instead of expressing the objective in terms of distribution over $\mathbb{S}$, we can characterize the robust loss in terms of the inverse cdf of the distribution (over $\mathbb{R}$) of $\ell(x; S)$. Let F^{-1} denotes the inverse cdf of $\ell(x; S)$ under P. Note that $\ell(x; S)$ with $S \sim P$ is equal in distribution to $F^{-1}(U)$ with $U \sim \mathsf{Unif}([0, 1])$. Therefore,

$$\begin{aligned} \mathcal{L}(x; P) &\coloneqq \sup_{Q' : \mathrm{D}_\phi(Q', \mathsf{Unif}([0,1])) \le \rho} \Big\{ \mathbb{E}_{U \sim Q'}[F^{-1}(U)] - \lambda \mathrm{D}_\psi(Q', \mathsf{Unif}([0, 1])) \Big\} \\ &= \sup_{r \in \mathcal{R}} \int_0^1 \Big[r(u) F^{-1}(u) - \lambda \psi(r(u)) \Big] \mathrm{d}u, \end{aligned} \tag{14}$$

where the last equality follows from writing $r(u) = \frac{\mathrm{d}Q'}{\mathrm{d}\mathsf{Unif}([0,1])}(u)$, and the set $\mathcal{R}$ is

$$\mathcal{R} \coloneqq \left\{ r : [0, 1] \to \mathbb{R}_+ \;\middle|\; \int_0^1 r(u) \mathrm{d}u = 1 \text{ and } \int_0^1 \phi(r(u)) \mathrm{d}u \le \rho \right\}. \tag{15}$$

A.1.2 Dual formulation

We can convert the maximization over r in Eq. (14) (or Q in (12)) with minimization over Lagrange multipliers for the constraint that r sums to 1 and the ϕ-divergence constraint, yielding

$$\mathcal{L}(x; P) = \inf_{\eta \in \mathbb{R}, \nu \ge 0} \Upsilon(x, \eta, \nu; P), \text{ where}$$

$$\Upsilon(x, \eta, \nu; P) \coloneqq \int_0^1 \sup_{r \in \mathbb{R}_+} \Big[r F^{-1}(u) - \eta(r - 1) - \nu(\phi(r) - \rho) - \lambda \psi(r) \Big] \mathrm{d}u, \tag{16}$$

where the strong duality follows Shapiro [66, Sec. 3.2]. Writing $(g)^*[v] \coloneqq \sup_{t \in \mathrm{dom}(g)} \{vt - g(t)\}$ for the conjugate function of g, we may express Υ as

$$\Upsilon(x, \eta, \nu; P) = \int_0^1 (\nu\phi + \lambda\psi)^*[F^{-1}(u) - \eta] \mathrm{d}u + \eta + \nu\rho = \mathbb{E}(\nu\phi + \lambda\psi)^*[\ell(x; S) - \eta] + \eta + \nu\rho, \tag{17}$$

where the expectation is over $S \sim P$, i.e. the distribution from which we observe samples. On a finite sample s_1^n we have

$$\Upsilon(x, \eta, \nu; s_1^n) \coloneqq \Upsilon(x, \eta, \nu; \widehat{P}[s_1^n]) = \frac{1}{n} \sum_{i \le n} (\nu\phi + \lambda\psi)^*[\ell(x; s_i) - \eta] + \eta + \nu\rho.$$

For pure-constraint objectives (with $\psi = 0$), Υ simplifies to

$$\psi = 0 \implies \Upsilon(x, \eta, \nu; P) = \nu \, \mathbb{E}_{S \sim P} \, \phi^* \left[\frac{\ell(x; S) - \eta}{\nu} \right] + \eta + \nu\rho. \tag{18}$$

For pure-penalty objective (with $\phi = 0$) the Lagrange multiplier ν is unnecessary and we have

$$\phi = 0 \implies \Upsilon(x, \eta; P) = \lambda \, \mathbb{E}_{S \sim P} \, \psi^* \left[\frac{\ell(x; S) - \eta}{\lambda} \right] + \eta. \tag{19}$$

Note that Υ is an expectation (i.e., an empirical risk) which means that to minimize $\mathcal{L}(x; P)$ we can, in principle, apply ERM jointly on x, η and ν, as we further discuss in Appendix A.3.

Finally, we note that any $Q^\star$ attaining the supremum in (12) is of the form

$$\frac{\mathrm{d}Q^\star}{\mathrm{d}P}(s) = (\nu^\star + \lambda\psi)^{*\prime}[\ell(x; s) - \eta^\star].$$

where $\eta^\star$ and $\nu^\star$ are optimal Lagrange multipliers in (16) and $(\nu^\star + \lambda\psi)^{*\prime}$ is a subderivative of $(\nu^\star + \lambda\psi)^*$. For $\phi = 0$ this specializes to

$$\frac{\mathrm{d}Q^\star}{\mathrm{d}P}(s) = \psi^{*\prime} \left[\frac{\ell(x; s) - \eta^\star}{\lambda} \right].$$

For a finite sample, we have

$$q_i^\star = \frac{1}{n} \psi^{*\prime} \left[\frac{\ell(x; s_i) - \eta^\star}{\lambda} \right]. \tag{20}$$

A.1.3 Expressions for CVaR

Recall that CVaR at level α corresponds to $\phi = 0$ and $\psi = \mathbb{I}_{[0, 1/\alpha)}$. The dual expression of CVaR simplifies to [67, Example 6.16]

$$\mathcal{L}_{\text{CVaR}}(x; P) = \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha} \, \mathbb{E}_{S \sim P} (\ell(x; S) - \eta)_+ + \eta \right\}.$$

It also has a simple closed-form expression in terms of the inverse cdf of $\ell(x; S)$ [67, Theorem 6.2]:

$$\mathcal{L}_{\text{CVaR}}(x; P) = \frac{1}{\alpha} \int_{1-\alpha}^{1} F^{-1}(u) \mathrm{d}u. \tag{21}$$

We note that this last expression is a direct consequence of (14), since $\mathcal{R}$ is the set of measures never exceeding $\frac{1}{\alpha}$. On a finite sample s_1^n this gives the closed-form expression

$$\mathcal{L}_{\text{CVaR}}(x; s_1^n) = \frac{1}{\alpha n} \sum_{i=1}^{\lfloor \alpha n \rfloor} \ell(x; s_{(i)}) + \left(1 - \frac{\lfloor \alpha n \rfloor}{\alpha n} \right) \ell(x; s_{(\lfloor \alpha n \rfloor + 1)}), \tag{22}$$

where $s_{(1)}, \ldots, s_{(n)}$ are a permutation of s_1^n satisfying $\ell(x; s_{(1)}) \ge \ell(x; s_{(2)}) \ge \cdots \ge \ell(x; s_{(n)})$. For $\alpha \le 1/n$ we simply have $\mathcal{L}_{\text{CVaR}}(x; s_1^n) = \max_{i \le n} \ell(x; s_i)$.

The KL-divergence penalized CVaR at level α corresponds to $\psi(t) = \mathbb{I}_{[0, 1/\alpha]}(t) + t \log t - t + 1$, for which

$$\psi^*[v] = \begin{cases} e^v - 1 & v < \log \frac{1}{\alpha} \\ \frac{1}{\alpha} - 1 + \frac{1}{\alpha}(v - \log \frac{1}{\alpha}) & \text{otherwise,} \end{cases}$$

and the dual expression for $\mathcal{L}_{\text{kl-CVaR}}$ is given by (19). In the special case $\alpha \le 1/n$ the CVaR constraint becomes inactive, and we can minimize over η in closed form to obtain the standard "soft max" objective $\mathcal{L}_{\text{kl-CVaR}}(x; s_1^n) = \lambda \log\big(\frac{1}{n} \sum_{i \le n} \exp(\ell(x; s_i)/\lambda) \big)$.

A.1.4 Expressions for $\mathcal{L}_{\chi^2\text{-pen}}$ and $\mathcal{L}_{\chi^2}$

The penalized version of the χ^2 objective corresponds to $\phi(t) = 0$ and $\psi(t) = \frac{1}{2}(t-1)^2$. Note that $\mathrm{D}_\phi(Q, P)$ is invariant under $\psi(t) \mapsto \psi(t) + c \cdot (t-1)$ for any $c \in \mathbb{R}$ because $\int (\frac{\mathrm{d}Q}{\mathrm{d}P} - 1)\mathrm{d}P = 0$. We find it more convenient to work with $\psi(t) = \frac{1}{2}(t-1)^2 + (t-1) = \frac{1}{2}(t^2-1)$, for which the conjugate is simply $\psi^*[v] = \frac{1}{2}((v)_+^2 + 1)$. The dual form (19) gives

$$\mathcal{L}_{\chi^2\text{-pen}}(x; P) = \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{2\lambda} \mathbb{E}_{S \sim P} \left(\ell(x; S) - \eta\right)_+^2 + \frac{\lambda}{2} + \eta \right\}. \tag{23}$$

The infimum is attained at the $\eta^\star$ solving $\mathbb{E}(\ell(x; S) - \eta^\star)_+ = \lambda$. In other words,

$$\eta^\star = \mathbb{E}[\ell(x; S) \mid \ell(x; S) \ge \eta^\star] - \frac{\lambda}{\mathbb{P}(\ell(x; S) \ge \eta^\star)} = \mathcal{L}_{\text{CVaR}}^{F(\eta^\star)}(x; P) - \frac{\lambda}{1 - F(\eta^\star)},$$

where $F(t) = \mathbb{P}(\ell(x; S) \le t)$ is the cdf of $\ell(x; S)$. Letting $\mathfrak{G}(\eta^\star)$ denote the event that $\ell(x; S) \ge \eta^\star$, substituting back to the expression for $\mathcal{L}_{\chi^2\text{-pen}}$ gives

$$\mathcal{L}_{\chi^2\text{-pen}}(x; P) = \mathbb{E}[\ell(x; S) \mid \mathfrak{G}(\eta^\star)] + \frac{1}{2\lambda}\mathrm{Var}[\ell(x; S) \mid \mathfrak{G}(\eta^\star)] + \frac{\lambda}{2}\left(\frac{1}{\mathbb{P}(\mathfrak{G}(\eta^\star))} - 1\right)^2.$$

In words, $\mathcal{L}_{\chi^2\text{-pen}}$ is a sum of a CVaR (at level $F(\eta^\star)$), a conditional variance regularization term and an outage probability regularization term. This expression simplifies considerably when λ is sufficiently large. Specifically, we have,

$$\begin{aligned} \lambda \ge B &\implies \lambda \ge \mathbb{E}\,\ell(x; S) - F^{-1}(0) \\ &\implies \eta^\star = \mathbb{E}\,\ell(x; S) - \lambda \text{ and } \mathbb{P}(\mathfrak{G}(\eta^\star)) = 1 \\ &\implies \mathcal{L}_{\chi^2\text{-pen}}(x; P) = \mathbb{E}\,\ell(x; S) + \frac{1}{2\lambda}\mathrm{Var}[\ell(x; S)]. \end{aligned} \tag{24}$$

That is, for sufficiently large λ the objective $\mathcal{L}_{\chi^2\text{-pen}}$ is simply the empirical risk with variance regularization (see also [22]).

For a finite sample we have

$$\mathcal{L}_{\chi^2\text{-pen}}(x; s_1^n) = \frac{1}{2\lambda n} \sum_{i \le n} (\ell(x; s_i) - \eta_n^\star)_+^2 + \frac{\lambda}{2} + \eta_n^\star.$$

Where $\eta_n^\star$ is the solution to $\sum_{i \le n} (\ell(x; s_i) - \eta_n^\star)_+ = n\lambda$, or equivalently

$$\eta_n^\star = \frac{1}{i^\star} \sum_{i \le i^\star} \ell(x; s_{(i)}) - \frac{\lambda n}{i^\star} \text{ for the unique } i^\star \text{ such that } \ell(x; s_{(i^\star+1)}) \le \eta_n^\star \le \ell(x; s_{(i^\star)}), \tag{25}$$

where $\{\ell(x; s_{(i)})\}$ are the sorted $\{\ell(x; s_i)\}$ and $\ell(x; s_{(n+1)}) \coloneqq -\infty$.

An expression for $\mathcal{L}_{\chi^2}$ follows via (23)

$$\mathcal{L}_{\chi^2}(x; P) = \inf_{\lambda \ge 0} \left\{ \mathcal{L}_{\chi^2\text{-pen}}(x; P) + \lambda\rho \right\} = \inf_{\eta \in \mathbb{R}} \left\{ \sqrt{1 + 2\rho}\sqrt{\mathbb{E}_{S \sim P} \left(\ell(x; S) - \eta\right)_+^2} + \eta \right\}, \tag{26}$$

and the maximizing Q is

$$\frac{\mathrm{d}Q^\star}{\mathrm{d}P}(s) = \frac{(\ell(x; s) - \eta^\star)_+}{\mathbb{E}_{S \sim P}(\ell(x; S) - \eta^\star)_+}. \tag{27}$$

A.1.5 Expression for $\nabla\mathcal{L}$

Let $Q^\star$ by a distribution attaining the supremum in (12) and recall that $\nabla\ell(x; s)$ denotes an element in the sub-differential of $\ell(x; s)$ w.r.t. x. Then the following vector is a subgradient of $\mathcal{L}$ [37, Corollary 4.4.4],

$$\nabla\mathcal{L}(x; P) = \mathbb{E}_{S \sim Q^\star} \nabla\ell(x; S).$$

Similarly, for a sample of size n and a maximizing $q^\star$, we have

$$\nabla\mathcal{L}(x; s_1^n) = \sum_{i \le n} q_i^\star \nabla\ell(x; s_i). \tag{28}$$

A.1.6 Smoothness of $\mathcal{L}_{\chi^2\text{-pen}}$ and $\mathcal{L}_{\text{kl-CVaR}}$

The smoothness of $\mathcal{L}$ (i.e., Lipschitz continuity of its gradient) plays a role in our mini-batch gradient estimator complexity guarantees. When the penalty term ψ is strongly convex, the maximizing $Q^\star$ (or $q^\star$) is unique, and if $\nabla\ell$ is H-Lipschitz then $\mathcal{L}$ is differentiable [37, Corollary 4.4.5]. In particular, writing $Q^\star_x$ for the maximizing Q at point x, we have

$$\begin{aligned}\|\nabla\mathcal{L}(x;P)-\nabla\mathcal{L}(y;P)\| &= \left\|\int\left\{\nabla\ell(x;s)\mathrm{d}Q^\star_x(y)-\nabla\ell(y;s)\mathrm{d}Q^\star_y(s)\right\}\right\| \\ &\le \int\|\nabla\ell(x;s)-\nabla\ell(y;s)\|\mathrm{d}Q^\star_x(s)+\int\|\nabla\ell(y;s)\|\left|\frac{\mathrm{d}Q^\star_x}{\mathrm{d}P}(s)-\frac{\mathrm{d}Q^\star_y}{\mathrm{d}P}(s)\right|\mathrm{d}P(s) \\ &\le H\|x-y\|+G\left\|Q^\star_x-Q^\star_y\right\|_1. \qquad (29)\end{aligned}$$

Therefore, if $Q^\star_x$ is Lipschitz w.r.t. x in the 1-norm then $\nabla\mathcal{L}$ is Lipschitz as well. This is indeed the case $\mathcal{L}_{\chi^2\text{-pen}}$ and $\mathcal{L}_{\text{kl-CVaR}}$.

Claim 1. *Let Assumption A1 hold. For all P, $\nabla\mathcal{L}_{\text{kl-CVaR}}(x;P)$ and $\nabla\mathcal{L}_{\chi^2\text{-pen}}(x;P)$ are $(\frac{G^2}{\lambda}+H)$-Lipschitz in x, and $0\le\mathcal{L}_{\text{CVaR}}(x;P)-\mathcal{L}_{\text{kl-CVaR}}(x;P)\le\lambda\log(1/\alpha)$ for all x.*

Proof. Since entropy is 1-strongly-convex w.r.t. the 1-norm, for $\mathcal{L}_{\text{kl-CVaR}}$ we have that the penalty $\lambda\psi$ is λ-strongly-convex w.r.t. the 1-norm and therefore [63, Lemma 2]

$$\left\|Q^\star_x-Q^\star_y\right\|_1\le\frac{1}{\lambda}\|\ell(x;\cdot)-\ell(y;\cdot)\|_\infty\le\frac{G}{\lambda}\|x-y\|,$$

which by (29) implies that $\nabla\mathcal{L}_{\text{kl-CVaR}}$ is $(H+G^2/\lambda)$-Lipschitz as required. For $\mathcal{L}_{\chi^2\text{-pen}}$, we find it easier to argue for a finite sample s_1^n. By (13) we have $q^\star_x=\arg\max_{q\in\Delta^n}\left\{q\top\ell(x)-\frac{1}{2}\lambda n\|q\|_2^2\right\}$, where $\ell_i(x)=\ell(x;s_i)$. Therefore, by λn -strong-convexity w.r.t. the 2-norm, we have

$$\|q^\star_x-q^\star_y\|_1\le\sqrt{n}\|q^\star_x-q^\star_y\|_2\le\frac{1}{\lambda\sqrt{n}}\|\ell(x)-\ell(y)\|_2\le\frac{G}{\lambda}\|x-y\|,$$

establishing that $\nabla\mathcal{L}_{\chi^2\text{-pen}}$ is also $(H+G^2/\lambda)$-Lipschitz.

Finally, we note that $\mathcal{L}_{\text{kl-CVaR}}(x;P)\le\mathcal{L}_{\text{CVaR}}$ because $D_\psi(Q,P)\ge0$ for all Q. Conversely since any feasible Q satisfies $\mathrm{d}Q/\mathrm{d}P\le1/\alpha$ we have $D_\psi(Q,P)=\int\mathrm{d}Q\log\frac{\mathrm{d}Q}{\mathrm{d}P}\le\log\frac{1}{\alpha}$ and therefore $\mathcal{L}_{\text{kl-CVaR}}(x;P)\ge\mathcal{L}_{\text{CVaR}}(x;P)-\lambda\log\frac{1}{\alpha}$. □

A.2 Computational cost

To compute $\mathcal{L}(x;s_1^n)$ and its (sub)gradient from $\{\ell(x;s_i)\}_{i\le n}$ and $\{\nabla\ell(x;s_i)\}_{i\le n}$ we compute $q^\star$ that maximizes (13) and substitute it back in (28). The substitution requires $O(nd)$ work, so it remains to account for the work in computing $q^\star$.

For CVaR, this clearly amounts to sorting $\{\ell(x;s_i)\}_{i\le n}$ and therefore takes $O(n\log n)$ time. Similarly, for $\mathcal{L}_{\chi^2\text{-pen}}$ we may find sort the losses and find $i^\star$ in (25), and hence $\eta^\star_n$ and $q^\star$, in $O(n)$ time. Alternatively, for any objective with $\phi=0$ (including $\mathcal{L}_{\chi^2\text{-pen}}$ and $\mathcal{L}_{\text{kl-CVaR}}$ we can bisect directly on η, either to minimize the expression (19) or to satisfy the the simplex constraint $\sum_{i\le n}q^\star_i=\frac{1}{n}\sum_{i\le n}(\psi^\star)'[(\ell(x;s_i)-\eta^\star)/\lambda]=1$.

For $\mathcal{L}_{\chi^2}$ we may find $q^\star$ by performing similar bisection over η via the expression (26), again either minimizing it or solving for the condition $\frac{1}{n}\sum_{i\le n}(\ell(x;s_i)-\eta^\star)_+^2=(1+2\rho)\big(\frac{1}{n}\sum_{i\le n}(\ell(x;s_i)-\eta^\star)_+\big)^2$. Finding an ε accurate solution via bisection requires roughly $n\log\frac{B}{\varepsilon}$ time.

Since we are interested in large-scale application, we assume that $d\gg\log(nB/\varepsilon)$ and therefore the time to compute the objective and its gradient is $O(nd)$.

For simplicity and stability, our code implements the computation of $q^\star$ using bisection over η for each of $\mathcal{L}_{\chi^2\text{-pen}},\mathcal{L}_{\chi^2}$ and $\mathcal{L}_{\text{kl-CVaR}}$.

A.3 Stochastic gradient method on the dual objective

Here we discuss the convergence guarantees for a simple stochastic gradient method using the dual expression (16) for $\mathcal{L}(x; P_0)$ in order to minimize it over x. While several works consider such methods (see Section 1.1), we could not find direct reference for their runtime guarantees, and we therefore briefly derive it below.

Focusing on objectives with $\phi = 0$ (as in (19)), and writing γ_x and γ_η for step sizes, we write the iterations on x and the Lagrange multiplier η as

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \gamma_x \nabla \Upsilon(x_t, \eta_t; P_0)) = \Pi_{\mathcal{X}}\left(x_t - \gamma_x \psi^{*\prime}\left[\frac{\ell(x; S_i) - \eta_t}{\lambda}\right] \nabla \ell(x; S_i)\right), \text{ and}$$
$$\eta_{t+1} = \Pi_{[\underline{\eta}, \overline{\eta}]}\left(\eta_t - \gamma_\eta \frac{\partial}{\partial \eta} \Upsilon(x_t, \eta_t; P_0)\right) = \Pi_{[\underline{\eta}, \overline{\eta}]}\left(\eta_t + \gamma_\eta \psi^{*\prime}\left[\frac{\ell(x; S_i) - \eta_t}{\lambda}\right] - \gamma_\eta\right), \quad (30)$$

Where $S_1, S_2, \ldots$ are drawn iid from P_0.

For CVaR, we have $(\psi^*)'[v] = \frac{1}{\alpha} 1_{\{v \geq 0\}}$ and we may restrict η to the range $[\underline{\eta}, \overline{\eta}] = [0, B]$, as the optimal η is the value at risk level α and therefore in the range of ℓ. For $\mathcal{L}_{\chi^2\text{-pen}}$ we have $(\psi^*)'[v] = (v)_+$ and we may take $[\underline{\eta}, \overline{\eta}] = [-\lambda, B]$ due to the condition $\mathbb{E}(\ell(x; S) - \eta^\star)_+ = \lambda$. In these settings, the method (30) has the following guarantee

Claim 3. *Let $\epsilon \in (0, B)$. For CVaR and a suitable choice of γ_x, γ_η the average iterate $\bar{x}_T = \frac{1}{T} \sum_{t \leq T} x_t$ satisfies*

$$\mathbb{E}\, \mathcal{L}_{\text{CVaR}}(\bar{x}_T; P_0) - \min_{x' \in \mathcal{X}} \mathcal{L}_{\text{CVaR}}(x'; P_0) \leq \epsilon \text{ for } T \asymp \frac{(GR)^2 + B^2}{\alpha^2 \epsilon^2}.$$

Similarly, for χ^2 penalty we have

$$\mathbb{E}\, \mathcal{L}_{\chi^2\text{-pen}}(\bar{x}_T; P_0) - \min_{x' \in \mathcal{X}} \mathcal{L}_{\chi^2\text{-pen}}(x'; P_0) \leq \epsilon \text{ for } T \asymp \frac{(GR)^2 + B^2}{\epsilon^2}\left(1 + \frac{B^2}{\lambda^2}\right).$$

Proof. By Proposition 6, the expected sub-optimality of $\bar{x}_T$ is $\lesssim (\Gamma_x R + \Gamma_\eta(\overline{\eta} - \underline{\eta})/\sqrt{T}$, where Γ_x^2 (respectively Γ_η) is an upper bound on the second moment of $\nabla \Upsilon(x, \eta; S)$ (respectively $\frac{\partial}{\partial \eta} \Upsilon(x, \eta; S)$). For $\mathcal{L}_{\text{CVaR}}$ we have $\Gamma_x \leq G/\alpha$, $\Gamma_\eta \leq 1/\alpha$ and $\overline{\eta} - \underline{\eta} = B$. For $\mathcal{L}_{\chi^2\text{-pen}}$ we have $\Gamma_x \leq G(1 + B/\lambda)$, $\Gamma_\eta = 1 + B/\lambda$ and $\overline{\eta} - \underline{\eta} = B + \lambda$. The result follows from substituting $T \asymp \left(\Gamma_x^2 R^2 + \Gamma_\eta^2 (\overline{\eta} - \underline{\eta})^2\right) \epsilon^{-2}$. □

A.4 Uncertainty sets contained in χ^2 divergence balls

A number of our results hold for general subclass of the objective (12) with the following property.

Definition 1 (χ^2-bounded objective)**.** *An objective $\mathcal{L}(x; P_0)$ is C-χ^2-bounded if for all x and all $Q^\star$ attaining the supremum in (12) we have $\mathrm{D}_{\chi^2}(Q^\star, P_0) \leq C$.*

The three objectives we focus on are χ^2-bounded.

Claim 4. *The objectives $\mathcal{L}_{\text{kl-CVaR}}$, $\mathcal{L}_{\chi^2}$ and $\mathcal{L}_{\chi^2\text{-pen}}$ are χ^2-bounded with constants $C = \frac{1}{\alpha} - 1$, $C = \rho$ and $C = B/\lambda$, respectively.*

Proof. That $\mathcal{L}_{\chi^2}$ is ρ-χ^2-bounded is obvious from definition. For $\mathcal{L}_{\chi^2\text{-pen}}$ we have

$$\begin{aligned} \mathcal{L}_{\chi^2\text{-pen}}(x; P_0) &= \mathbb{E}_{S \sim Q^\star} \ell(x; S) - \lambda \mathrm{D}_{\chi^2}(Q^\star, P_0) \\ &\geq \mathbb{E}_{S \sim P_0} \ell(x; S) - \lambda \mathrm{D}_{\chi^2}(P_0; P_0) = \mathbb{E}_{S \sim P_0} \ell(x; S) \end{aligned}$$

and consequently

$$\mathrm{D}_{\chi^2}(Q^\star, P_0) \leq \frac{E_{S \sim Q^\star} \ell(x; S) - \mathbb{E}_{S \sim P_0} \ell(x; S)}{\lambda} \leq \frac{B}{\lambda}.$$

Finally, for $\mathcal{L}_{\text{kl-CVaR}}$ every feasible Q satisfies $\mathrm{d}Q/\mathrm{d}P_0 \leq 1/\alpha$ and therefore

$$\mathrm{D}_{\chi^2}(Q, P_0) = \int \left(\frac{\mathrm{d}Q}{\mathrm{d}P_0}(s)\right)^2 \mathrm{d}P_0(s) - 1 \leq \frac{1}{\alpha} \left(\frac{\mathrm{d}Q}{\mathrm{d}P_0}(s)\right) \mathrm{d}P_0(s) - 1 = \frac{1}{\alpha} - 1.$$

□

A.5 General results

We conclude this section of the appendix by stating three general results that aid our analysis. First, we give a lemma stating that a binomial random variable with parameters n and α has a constant probability of being at least $\sqrt{\alpha(1-\alpha)n}$ below its mean.

Lemma 1. *Let $n \in \mathbb{N}$ and $\alpha \in (0,1)$. There exists a numerical constant $C \in \mathbb{R}$ such that*

$$\mathbb{P}\big(\mathsf{Bin}(n,\alpha) \le n\alpha - \sqrt{n\alpha(1-\alpha)}\big) \ge \mathbb{P}(\mathcal{N}(0,1) \le -1) - \frac{C}{\sqrt{\alpha(1-\alpha)n}}.$$

Proof. Note that $\mathbb{P}\big(\mathsf{Bin}(n,\alpha) \le n\alpha - \sqrt{n\alpha(1-\alpha)}\big) = \mathbb{P}(Y\sqrt{n} \le -1)$ where $Y = \frac{1}{n}\frac{\mathsf{Bin}(n,\alpha)-n\alpha}{\sqrt{\alpha(1-\alpha)}}$ is the mean of n independent random variable with zero mean, unit variance, and absolute third moment $\rho = \frac{\alpha^2+(1-\alpha)^2}{\sqrt{\alpha(1-\alpha)}} \le \frac{1}{\sqrt{\alpha(1-\alpha)}}$. The Berry-Esseen theorem [25, Theorem 3.4.17] states that for such Y we have $|\mathbb{P}(Y\sqrt{n} \le t) - \mathbb{P}(\mathcal{N}(0,1) \le t)| \le C\rho/\sqrt{n}$, for all $t \in \mathbb{R}$; substituting $t = 1$ and $\rho \le \frac{1}{\sqrt{\alpha(1-\alpha)}}$ concludes the proof. □

Second, we state the Efron-Stein inequality in vector form, which follows from applying the standard scalar bound element-wise.

Lemma 2 (Efron-Stein inequality [9, Theorem 3.1])**.** *Let X_1^{n+1} be i.i.d random variables and $f : \mathcal{X}^n \to \mathbb{R}^m$. Let I be uniform on $\{1,\ldots,n\}$ and let $\tilde{X}_1^N$ be such that $\tilde{X}_i = X_i$ for $i \ne I$ and $\tilde{X}_I = X_{n+1}$. Then*

$$\mathrm{Var}[f(X_1^n)] \le \frac{n}{2}\,\mathbb{E}\,\|f(X_1^n) - f(\tilde{X}_1^n)\|^2. \tag{31}$$

Third, we give a general lemma on the variance of sampling without replacement, which we specialize to the simplex for later use.

Lemma 3. *Let $p \in \Delta^k$ and let $\mathcal{I}$ be a random subset of $[k]$ of size $k/2$. Then*

$$\mathbb{E}\left(\sum_{i\in\mathcal{I}} p_i - \frac{1}{2}\right)^2 \le \frac{1}{2}\left\|p - \tfrac{1}{k}\mathbf{1}\right\|^2 = \frac{1}{2k}\mathrm{D}_{\chi^2}(p, \tfrac{1}{k}\mathbf{1}).$$

Proof. Let us denote $q = p - \frac{1}{k}\mathbf{1}$. We have

$$\begin{aligned}
\mathbb{E}\left(\sum_{i\le k} p_i 1_{\{i\in\mathcal{I}\}} - \frac{1}{2}\right)^2 &= \mathbb{E}\left(\sum_{i\le k} q_i 1_{\{i\in\mathcal{I}\}}\right)^2 \\
&\overset{(i)}{=} \frac{1}{2}\sum_{i\le k} q_i^2 + \sum_{i\ne j} q_i q_j\,\mathbb{E}\,1_{\{i\in\mathcal{I} \text{ and } j\in\mathcal{I}\}} \\
&\overset{(ii)}{=} \frac{1}{2}\|q\|^2 + \frac{k-2}{4(k-1)}\sum_{i\le k}\sum_{j\ne i} q_i q_j = \frac{1}{2}\|q\|^2 + \frac{k-2}{4(k-1)}\sum_{i\le k} q_i(1-q_i) \\
&\overset{(iii)}{=} \left(\frac{1}{2} - \frac{k-2}{4(k-1)}\right)\|q\|^2 \le \frac{1}{2}\|q\|^2,
\end{aligned}$$

where (i) stems from $\mathbb{P}(i \in \mathcal{I}) = \frac{1}{2}$, (ii) from $\mathbb{P}(i \in I \text{ and } j \in I) = \frac{1}{2}\frac{k/2-1}{k-1}$ and (iii) from $\sum_{i\le k} q_i = 0$. Noting that $\mathrm{D}_{\chi^2}(p, \frac{1}{k}\mathbf{1}) = k\|q\|^2$ concludes the proof. □

B Proofs from Section 3

This section completes the proof and discussion of the results in Section 3. First, in Section B.1, we prove the bias bounds in Proposition 1 and argue their tightness in the worst case. Section B.2 provides additional discussion of the smoothness and Lipschitz inverse-cdf assumptions sometimes used in this section. Then, in Section B.3 we bound the variance of the mini-batch estimators for

χ^2-bounded penalty objectives and their gradient, obtaining Proposition 2 as a corollary. We also argue that similar bounds do not hold for the χ^2 constraint objective. In Section B.4 we review the standard convergence guarantees for stochastic gradients iterations with and without Nesterov acceleration, and in Section B.5 we combine all these ingredients to prove Theorem 1.

B.1 Bias of batch estimator

B.1.1 Proof of Proposition 1

Proposition 1 (Bias of the batch estimator)**.** *For all $x \in \mathcal{X}$ and $n \in \mathbb{N}$ we have*

$$0 \le \mathcal{L}(x; P_0) - \overline{\mathcal{L}}(x; n) \lesssim \begin{cases} B \min\{1, (\alpha n)^{-1/2}\} & \textit{for } \mathcal{L} = \mathcal{L}_{\text{CVaR}} & (5) \\ B\sqrt{(1+\rho)(\log n)/n} & \textit{for } \mathcal{L} = \mathcal{L}_{\chi^2} & (6) \\ B^2(\lambda n)^{-1} & \textit{for } \mathcal{L} = \mathcal{L}_{\chi^2\text{-pen}} & (7) \\ G_{\text{icdf}}\, n^{-1} & \textit{for any loss } (2), & (8) \end{cases}$$

where the bound (8) *holds under Assumption A1.*

Proof. We first show that the bound $\mathcal{L} \ge \overline{\mathcal{L}}$ holds for any loss of the form (12) and then proceed to show each of the bounds (5)–(8). We remark here that the bound (6) actually holds for any ρ-χ^2-bounded objective (Definition 1).

Proof of $\mathcal{L}(x; P_0) \ge \overline{\mathcal{L}}(x; n)$. The dual expression (17) gives

$$\begin{aligned} \mathcal{L}(x; P_0) &= \inf_{\eta \in \mathbb{R}, \nu \ge 0} \mathbb{E}_{S \sim P_0} \{(\nu\phi + \lambda\psi)^*[\ell(x; S) - \eta] + \eta + \nu\rho\} \\ &= \inf_{\eta \in \mathbb{R}, \nu \ge 0} \mathbb{E}_{S_1^n \sim P_0^n} \left\{ \frac{1}{n} \sum_{i \le n} (\nu\phi + \lambda\psi)^*[\ell(x; S_i) - \eta] + \eta + \nu\rho \right\} \\ &\ge \mathbb{E}_{S_1^n \sim P_0^n} \inf_{\eta \in \mathbb{R}, \nu \ge 0} \left\{ \frac{1}{n} \sum_{i \le n} (\nu\phi + \lambda\psi)^*[\ell(x; S_i) - \eta] + \eta + \nu\rho \right\} = \mathbb{E}\, \mathcal{L}(x; S_1^n) = \overline{\mathcal{L}}(x; n), \end{aligned}$$

where the inequality follows from exchanging the expectation and the infimum.

Proof of the CVaR bias bound (5)**.** By Eq. (22) we have

$$\overline{\mathcal{L}}_{\text{CVaR}}(x; n) = \mathbb{E}\, \mathcal{L}_{\text{CVaR}}(x; S_1^n) = \frac{1}{\alpha n} \sum_{i=1}^{\lfloor \alpha n \rfloor} \mathbb{E}\, \ell(x; S_{(i)}) + \left(1 - \frac{\lfloor \alpha n \rfloor}{\alpha n}\right) \mathbb{E}\, \ell(x; S_{(\lfloor \alpha n \rfloor + 1)}),$$

where $\ell(x; S_{(i)})$ is the ith order statistic of $\ell(x; S_1^n)$ (in decreasing order). Recalling that F denotes the cdf of $\ell(x; S)$, we may write $\ell(x; S) = F^{-1}(U)$ with U uniform on $[0, 1]$. Therefore, $\ell(x; S_{(i)}) = F^{-1}(U_{(i)})$ where $U_{(i)} \sim \mathsf{Beta}(n - i + 1, i)$ is the ith order statistic of n iid $\mathsf{Unif}([0, 1])$ random variables [57, Sec. 4.6]. Taking expectation, we have

$$\mathbb{E}_{S_1^n \sim P_0^n}\, \ell(x; S_{(i)}) = \int_0^1 F_Z^{-1}(u) f_{\mathsf{Beta}(n-i+1,i)}(u) \mathrm{d}u,$$

where $f_{\mathsf{Beta}(a,b)}$ is the density function of the Beta random variable of parameters a, b. Substituting back, we have

$$\begin{aligned} &\overline{\mathcal{L}}_{\text{CVaR}}(x; n) = \frac{1}{\alpha} \int_0^1 \mathcal{I}_\alpha(u) F^{-1}(u) \mathrm{d}u, \quad \text{where} \\ &\qquad \mathcal{I}_\alpha(u) = \frac{1}{n} \sum_{i=1}^{\lfloor \alpha n \rfloor} f_{\mathsf{Beta}(n-i+1,i)}(u) + \left(\alpha - \frac{\lfloor \alpha n \rfloor}{n}\right) f_{\mathsf{Beta}(n - \lfloor \alpha n \rfloor, \lfloor \alpha n \rfloor + 1)}(u). \end{aligned} \tag{32}$$

Using

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} f_{\mathsf{Beta}(n-i+1,i)}(u) &= \frac{1}{n} \sum_{i=1}^{n} \frac{n!}{(n-i)!(i-1)!} u^{n-i} (1-u)^{i-1} \\ &= \sum_{i=0}^{n-1} \binom{n-1}{i-1} (1-u)^i u^{n-1-i} = 1, \end{aligned} \tag{33}$$

we have that

$$1 - \mathcal{I}_\alpha(u) \le \frac{1}{n} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n} f_{\mathsf{Beta}(n-i+1,i)}(u).$$

Recalling Eq. (21) for $\mathcal{L}_{\text{CVaR}}(x; P_0)$, and recalling that $F^{-1}(u) \in [0, B]$ for all u by assumption, we bound the bias as

$$\begin{aligned}
\mathcal{L}_{\text{CVaR}}(x; P_0) - \overline{\mathcal{L}}_{\text{CVaR}}(x; n) &= \frac{1}{\alpha} \int_0^1 \big[1_{\{u \ge 1-\alpha\}} - \mathcal{I}_\alpha(u)\big] F^{-1}(u) \mathrm{d}u \\
&\le \frac{B}{\alpha} \int_{1-\alpha}^1 \big[1_{\{u \ge 1-\alpha\}} - \mathcal{I}_\alpha(u)\big] \mathrm{d}u \\
&\le \frac{B}{\alpha} \int_{1-\alpha}^1 \left(\frac{1}{n} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n} f_{\mathsf{Beta}(n-i+1,i)}(u) \right) \mathrm{d}u \\
&= \frac{B}{\alpha n} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n} \mathbb{P}(\mathsf{Beta}(n-i+1, i) \ge 1-\alpha).
\end{aligned} \tag{34}$$

To conclude, it suffices to bound the tail probability of the Beta random variables. We have [see, e.g., 57, Ex. 5 in Sec. 4.6]

$$\begin{aligned}
\mathbb{P}(\mathsf{Beta}(n-i+1, i) \ge 1-\alpha) &= 1 - \mathbb{P}(\mathsf{Beta}(n-i+1, i) \le 1-\alpha) \\
&= \mathbb{P}(\mathsf{Bin}(n; 1-\alpha) \le n-i) = \mathbb{P}(\mathsf{Bin}(n; \alpha) \ge i),
\end{aligned}$$

and the multiplicative Chernoff bound [46, Theorem 4.3] gives

$$\mathbb{P}(\mathsf{Bin}(n; \alpha) \ge i) \le \exp\left(-\frac{i - n\alpha}{3} \min\left\{ \frac{i - n\alpha}{n\alpha}, 1 \right\} \right).$$

Therefore, for $\alpha n \ge 9$,

$$\begin{aligned}
\sum_{i=\lfloor \alpha n \rfloor + 1}^{n} \mathbb{P}(\mathsf{Bin}(n; \alpha) \ge i) &\le \sum_{i=\lfloor \alpha n \rfloor + 1}^{2\lfloor \alpha n \rfloor} \exp\left(-\frac{(i - n\alpha)^2}{3n\alpha} \right) + \sum_{i=2\lfloor \alpha n \rfloor + 1}^{\infty} \exp\left(-\frac{i - n\alpha}{3} \right) \\
&\le 1 + \int_0^\infty \exp\left(-\frac{u^2}{3n\alpha} \right) \mathrm{d}u + \exp\left(-\frac{2\lfloor \alpha n \rfloor + 1 - \alpha n}{3} \right) \sum_{i=0}^{\infty} e^{-i/3} \\
&\le 1 + \frac{\sqrt{3\pi\alpha n}}{2} + \frac{e^{-3}}{1 - e^{-1/3}} \le 3\sqrt{\alpha n}.
\end{aligned}$$

Substituting into (34) and using $\mathcal{L}_{\text{CVaR}}(x; P_0) \le B$ when $\alpha n \le 9$ gives the final bound

$$\mathcal{L}_{\text{CVaR}}(x; P_0) - \overline{\mathcal{L}}_{\text{CVaR}}(x; n) \le B \min\left\{ \frac{3}{\sqrt{\alpha n}}, 1 \right\}. \tag{35}$$

Proof of the bound (6). We start with the expression (14) specialized for the $\mathcal{L}_{\chi^2}$,

$$\mathcal{L}_{\chi^2}(x; P_0) = \sup_{r \in \mathcal{R}} \int_0^1 r(\beta) F^{-1}(1-\beta) \mathrm{d}\beta,$$

where

$$\mathcal{R} = \left\{ r : [0,1] \to \mathbb{R}_+ \;\middle|\; \|r\|_1 = 1,\ \|r\|_2^2 \le 1 + 2\rho,\ \text{and } r \text{ is non-increasing} \right\};$$

The restriction of $\mathcal{R}$ to non-increasing functions is "free" since F^{-1} is non-decreasing. Our strategy is to relate F^{-1} to CVaR and then apply the corresponding bias bounds (5)—this type of transformation is closely related to the Kusuoka representation of coherent risk measures [42]. Specifically, note that

$$\mathcal{L}^{\alpha}_{\text{CVaR}} = \frac{1}{\alpha} \int_0^\alpha F_Z^{-1}(1-\beta) \mathrm{d}\beta \implies F_Z^{-1}(1-\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}(\alpha \mathcal{L}^{\alpha}_{\text{CVaR}}).$$

Therefore, for any $r \in \mathcal{R}$ integration by parts gives

$$\int_0^1 r(\beta)F_Z^{-1}(1-\beta)\mathrm{d}\beta = \int_0^1 r(\alpha)\frac{\mathrm{d}}{\mathrm{d}\alpha}(\alpha\mathcal{L}^{\alpha}_{\text{CVaR}})\mathrm{d}\alpha = r(1)\mathcal{L}^{1}_{\text{CVaR}} - \int_0^1 r'(\alpha)\alpha\mathcal{L}^{\alpha}_{\text{CVaR}}d\alpha.$$

The CVaR bias bound (35) tells us that $\mathcal{L}^{\alpha}_{\text{CVaR}} \le \overline{\mathcal{L}}^{\alpha}_{\text{CVaR}} + \mathrm{bb}(\alpha)$ where $\mathrm{bb}(\alpha) = 3B\min\Big\{\sqrt{\frac{1}{\alpha n}}, 1\Big\}$. Moreover, we may write $\overline{\mathcal{L}}^{\alpha}_{\text{CVaR}} = \frac{1}{\alpha}\int_0^{\alpha}\mathbb{E}\,\widehat{F}^{-1}(1-\beta)\mathrm{d}\beta$, where $\widehat{F}$ denotes the empirical cdf of the losses $\ell(x;S_1),\ldots,\ell(x;S_n)$. Noting that $r'(\alpha) \le 0$ for all α, we may write

$$\begin{aligned}-\int_0^1 r'(\alpha)\alpha\mathcal{L}^{\alpha}_{\text{CVaR}}d\alpha &\le -\int_0^1 r'(\alpha)\alpha\overline{\mathcal{L}}^{\alpha}_{\text{CVaR}}d\alpha - \int_0^1 r'(\alpha)\alpha\cdot\mathrm{bb}(\alpha)d\alpha \\ &= \mathbb{E}\int_0^1 r(\beta)\widehat{F}^{-1}(1-\beta)\mathrm{d}\beta - r(1)\overline{\mathcal{L}}^{1}_{\text{CVaR}} + \int_0^1 [r(\alpha)-r(1)](\alpha\cdot\mathrm{bb}(\alpha))'\mathrm{d}\alpha,\end{aligned}$$

where in the final equality we used again integration by parts along with $\mathbb{E}\,\widehat{F}^{-1}(1-\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}(\alpha\overline{\mathcal{L}}_{\text{CVaR}})$.

Substituting back and using $\mathcal{L}^{1}_{\text{CVaR}} = \overline{\mathcal{L}}^{1}_{\text{CVaR}} = \mathbb{E}\,\ell(x;S)$, we obtain

$$\int_0^1 r(\beta)F_Z^{-1}(1-\beta)\mathrm{d}\beta - \mathbb{E}\int_0^1 r(\beta)\widehat{F}^{-1}(1-\beta)\mathrm{d}\beta \le \sup_{r\in\mathcal{R}}\int_0^1 [r(\alpha)-r(1)](\alpha\cdot\mathrm{bb}(\alpha))'\mathrm{d}\alpha =: E$$

Taking a supremum over $r \in \mathcal{R}$, we conclude that

$$\begin{aligned}\mathcal{L}_{\chi^2}(x;P_0) = \sup_{r\in\mathcal{R}}\int_0^1 r(\beta)F_Z^{-1}(1-\beta)\mathrm{d}\beta &\le \sup_{r\in\mathcal{R}}\mathbb{E}\int_0^1 r(\beta)\widehat{F}^{-1}(1-\beta)\mathrm{d}\beta + E \\ &\le \mathbb{E}\sup_{r\in\mathcal{R}}\int_0^1 r(\beta)\widehat{F}^{-1}(1-\beta)\mathrm{d}\beta + E = \overline{\mathcal{L}}_{\chi^2}(x;n) + E. \qquad (36)\end{aligned}$$

It remains to bound the quantity E, which we do via the the Cauchy-Schwarz inequality and the definition of $\mathcal{R}$, which gives

$$E = \int_0^1 [r(\alpha)-r(1)](\alpha\cdot\mathrm{bb}(\alpha))'\mathrm{d}\alpha \le \|r\|_2\|(\alpha\cdot\mathrm{bb}(\alpha)'\|_2 \le \sqrt{1+2\rho}\cdot\|(\alpha\cdot\mathrm{bb}(\alpha))'\|_2$$

for all $r \in \mathcal{R}$. We calculate $(\alpha\cdot\mathrm{bb}(\alpha))' = \mathrm{bb}(0)1_{\{\alpha\le 1/n\}} + \frac{1}{2}\mathrm{bb}(\alpha)1_{\{\alpha>1/n\}}$, so that

$$\|(\alpha\cdot\mathrm{bb}(\alpha)'\|_2^2 = \frac{\mathrm{bb}^2(0)}{n}\left(1+\int_{1/n}^1\frac{\mathrm{d}\beta}{4\beta}\right) \le (3B)^2\cdot\frac{4+\log n}{4n}.$$

for all $r \in \mathcal{R}$, giving the required bound.

Remark 1. The bound (36) hold for any loss (12) and not just $\mathcal{L}_{\chi^2}$. Moreover, the final bound using Cauchy-Schwarz is equally valid for any ρ-χ^2-bounded uncertainty set. In particular, consider the Cressie-Read uncertainty sets [15] corresponding to k-norm the constraint $\|r\|_k^2 \le 1+2\rho$. For $k > 2$ they satisfy $\|r\|_2^2 \le 1+2\rho$ and our bias bounds holds (using Hölder's inequality instead of Cauchy-Schwarz removes the logarithmic factor). For $k \in (1,2)$ Hölder's inequality gives bounds decaying as $n^{-(k-1)/k}$.

Proof of the bound (8) Starting with CVaR, we return to the expression (32) for the bias and note that

$$\big[1_{\{u\ge 1-\alpha\}} - \mathcal{I}_\alpha(u)\big]F^{-1}(u) \le \big[1_{\{u\ge 1-\alpha\}} - \mathcal{I}_\alpha(u)\big]\big\{F^{-1}(1-\alpha) + G_{\text{icdf}}\cdot(u-(1-\alpha))\big\}$$

holds for all u, because when $u < 1-\alpha$ we have that $1_{\{u\ge 1-\alpha\}} - \mathcal{I}_\alpha(u) \le 0$ and so we increase the LHS by replacing F^{-1} with an under-estimate, while for $u \ge 1-\alpha$ we have $1-\mathcal{I}_\alpha(u) \ge 0$ due to (33) and we increase the LHS be replacing it with an F^{-1} with an over-estimate. Substituting into (32) and calculating gives

$$\begin{aligned}
&\mathcal{L}_{\text{CVaR}}(x;P_0) - \overline{\mathcal{L}}_{\text{CVaR}}(x;n) \\
&\quad\le \frac{1}{\alpha}\int_0^1 \left(1_{\{u\ge 1-\alpha\}} - \mathcal{I}_\alpha(u)\right)\left[F^{-1}(1-\alpha) + G_{\text{icdf}}\cdot(u-[1-\alpha])\right]\mathrm{d}u \\
&\quad\overset{(i)}{=} \frac{G_{\text{icdf}}}{\alpha}\int_0^1 \left(1_{\{u\ge 1-\alpha\}} - \mathcal{I}_\alpha(u)\right)u\mathrm{d}u \\
&\quad\overset{(ii)}{=} G_{\text{icdf}}\left[\frac{1}{2\alpha}\left(1-(1-\alpha)^2\right) - \frac{1}{\alpha n}\sum_{i=n-\lfloor\alpha n\rfloor+1}^{n}\frac{i}{n+1} - \left(1-\frac{\lfloor\alpha n\rfloor}{\alpha n}\right)\frac{n-\lfloor\alpha n\rfloor}{n+1}\right] \\
&\quad= G_{\text{icdf}}\left[1-\frac{\alpha}{2} - \frac{1}{\alpha n(n+1)}\frac{\lfloor\alpha n\rfloor}{2}\left(2n-\lfloor\alpha n\rfloor+1\right) - \left(1-\frac{\lfloor\alpha n\rfloor}{\alpha n}\right)\frac{n-\lfloor\alpha n\rfloor}{n+1}\right] \\
&\quad= G_{\text{icdf}}\left[1-\frac{\alpha}{2} - \frac{1}{2}\frac{\lfloor\alpha n\rfloor}{\alpha n(n+1)}\left(\lfloor\alpha n\rfloor+1\right) - \frac{n-\lfloor\alpha n\rfloor}{n+1}\right] \\
&\quad= G_{\text{icdf}}\left[\frac{1}{n+1} + \frac{\lfloor\alpha n\rfloor}{n+1}\left[1-\frac{\lfloor\alpha n\rfloor}{2\alpha n}\right] - \frac{\alpha}{2} - \frac{1}{2}\frac{\lfloor\alpha n\rfloor}{\alpha n(n+1)}\right] \\
&\quad\le G_{\text{icdf}}\left[\frac{1}{n+1} + \frac{\alpha n}{2(n+1)} - \frac{\alpha}{2}\right] \le \frac{G_{\text{icdf}}}{n+1}. \qquad (37)
\end{aligned}$$

Above, (i) uses the fact that $\frac{1}{\alpha}\mathcal{I}_\alpha$ is a convex combination of densities to deduce that

$$\int_0^1 \left(1_{\{u\ge 1-\alpha\}} - \mathcal{I}_\alpha(u)\right)\left[F^{-1}(1-\alpha) - G_{\text{icdf}}\cdot(1-\alpha)\right]\mathrm{d}u = 0,$$

and (ii) uses the definition (32) of $\mathcal{I}_\alpha$ along with the fact that $\mathbb{E}\,\mathsf{Beta}(a,b) = \frac{a}{a+b}$.

This bound extends to any $\mathcal{L}$ of the form (12) via (36), since we have $\mathrm{bb}(\alpha) = G_{\text{icdf}}/(n+1)$ independent of α and consequently $(\alpha\cdot\mathrm{bb}(\alpha))' = G_{\text{icdf}}/(n+1)$, giving

$$E = \sup_{r\in\mathcal{R}}\int_0^1 [r(\alpha)-r(1)](\alpha\cdot\mathrm{bb}(\alpha))'\mathrm{d}\alpha = \frac{G_{\text{icdf}}}{n+1}\cdot\sup_{r\in\mathcal{R}}\int_0^1[r(\alpha)-r(1)]\mathrm{d}\alpha \le \frac{G_{\text{icdf}}}{n+1},$$

since $\int r(\alpha)\mathrm{d}\alpha = 1$ for all $r\in\mathcal{R}$ regardless of ϕ and ψ.

Penalized-χ^2 We use the shorthand $Z = \ell(x,S)$ and for a sample S_1^n we let $Z_i = \ell(x,S_i)$. By Eq. (23),

$$\mathcal{L}_{\chi^2\text{-pen}}(x;P_0) = \Upsilon(\eta^\star;P_0) = \mathbb{E}\,\frac{(Z-\eta^\star)_+^2}{2\lambda} + \eta^* + \frac{\lambda}{2},$$

where η^* is the unique solution to $\mathbb{E}(Z-\eta^\star)_+ = \lambda$. (We omit the dependence of Υ on x as x is constant throughout). Similarly, we have that

$$\mathcal{L}_{\chi^2\text{-pen}}(x;S_1^n) = \Upsilon(\eta_n;S_1^n),\ \text{ where }\ \Upsilon(\eta_n;S_1^n) := \sum_{i=1}^n\frac{(Z_i-\eta)_+^2}{2\lambda n} + \eta + \frac{\lambda}{2}$$

and η_n is the unique solution to $\frac{1}{n}\sum_{i=1}^n(Z_i-\eta)_+ = \lambda$. Convexity of Υ w.r.t. η gives us

$$\Upsilon(\eta_n;S_1^n) \ge \Upsilon(\eta^\star;S_1^n) + \Upsilon'(\eta^\star;S_1^n)(\eta_n-\eta^\star).$$

Taking expectation, we observe that $\mathbb{E}\,\Upsilon(\eta_n;S_1^n) = \overline{\mathcal{L}}_{\chi^2\text{-pen}}(x;n)$ and $\mathbb{E}\,\Upsilon(\eta^\star;S_1^n) = \Upsilon(\eta^\star;P_0) = \mathcal{L}_{\chi^2\text{-pen}}(x;P_0)$. Therefore, by the Cauchy-Schwarz inequality

$$\begin{aligned}
\overline{\mathcal{L}}_{\chi^2\text{-pen}}(x;n) - \mathcal{L}_{\chi^2\text{-pen}}(x;P_0) &= \mathbb{E}\,\Upsilon'(\eta^\star;S_1^n)(\eta_n-\eta^\star) \\
&\overset{(\star)}{=} \mathbb{E}\,\Upsilon'(\eta^\star;S_1^n)(\eta_n-\mathbb{E}\,\eta_n) \ge -\sqrt{\mathrm{Var}\Upsilon'(\eta^\star;S_1^n)}\sqrt{\mathrm{Var}\eta_n}, \qquad (38)
\end{aligned}$$

where $(\star)$ uses that $\mathbb{E}\,\Upsilon'(\eta^\star; S_1^n) = \mathbb{E}\,\Upsilon'(\eta^\star; P_0) = 0$ by the definition of $\eta^\star$, and therefore we may replace $\eta_n - \eta^\star$ with $\eta_n - \mathbb{E}\,\eta_n$. We now proceed to bound each variance separately. First, we have

$$\begin{aligned}
\mathrm{Var}\Upsilon'(\eta^\star; S_1^n) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(Z_i-\eta^*)_+}{\lambda}-1\right)\right]^2 = \frac{1}{n}\,\mathbb{E}\left(\frac{(Z-\eta^*)_+}{\lambda}-1\right)^2 \\
&= \frac{1}{n}\left[\mathbb{E}\,\frac{(Z-\eta^\star)_+^2}{\lambda^2}-1\right] = \frac{1}{n}\left[\frac{1}{\lambda}(2\mathcal{L}_{\chi^2\text{-pen}}(x;P_0)-2\eta^*-\lambda)-1\right] \le \frac{2B}{\lambda n}, \quad (39)
\end{aligned}$$

where in the final transition we used $\mathcal{L}_{\chi^2\text{-pen}}(x;P_0) \le B$ and $\eta^\star \ge -\lambda$ due to $\mathbb{E}(Z-\eta^\star)_+ = \lambda$ and $Z \ge 0$.

To handle the second variance we use the Efron-Stein inequality (Lemma 2). Let I be uniformly distributed on $[n]$, and define

$$\tilde{Z}_1^n = (Z_1, \ldots, Z_{I-1}, Z_I', Z_{I+1}, \ldots, Z_n),$$

where Z' is an i.i.d. copy of Z. Let $\tilde{\eta}_n$ be the solution to $\frac{1}{n}\sum_{i=1}^n(\tilde{Z}_i-\eta)_+ = \lambda$. Then,

$$\mathrm{Var}\eta_n \le \frac{n}{2}\,\mathbb{E}(\eta_n-\tilde{\eta}_n)^2 \qquad (40)$$

Define the random set

$$\mathcal{A} := \{i \mid Z_i - \eta_n > 0\}.$$

Recalling that $\sum_{i=1}^n(\tilde{Z}_i-\tilde{\eta}_n)_+ = \sum_{i=1}^n(Z_i-\eta_n)_+ = \lambda n$, we have

$$\begin{aligned}
0 &= \sum_{i\in[n]}\Big\{(Z_i-\eta_n)_+ - (\tilde{Z}_i-\tilde{\eta}_n)_+\Big\} \le \sum_{i\in\mathcal{A}}\Big\{(Z_i-\eta_n) - (\tilde{Z}_i-\tilde{\eta}_n)_+\Big\} \\
&\le \sum_{i\in\mathcal{A}}\Big\{(Z_i-\eta_n) - (\tilde{Z}_i-\tilde{\eta}_n)\Big\} = |\mathcal{A}|(\tilde{\eta}_n-\eta_n) + (Z_I - Z_I')1_{\{I\in\mathcal{A}\}},
\end{aligned}$$

and therefore $\eta_n - \tilde{\eta}_n \le \frac{B1_{\{I\in\mathcal{A}\}}}{|\mathcal{A}|}$. Similarly defining $\tilde{\mathcal{A}} := \{i \mid \tilde{Z}_i - \tilde{\eta}_n > 0\}$ and applying the same argument with $\tilde{\eta}_n$ and η_n swapped allows us to conclude that

$$(\eta_n-\tilde{\eta}_n)^2 \le B^2\max\left\{\frac{1_{\{I\in\mathcal{A}\}}}{|\mathcal{A}|^2}, \frac{1_{\{I\in\tilde{\mathcal{A}}\}}}{|\tilde{\mathcal{A}}|^2}\right\} \le B^2\left(\frac{1_{\{I\in\mathcal{A}\}}}{|\mathcal{A}|^2} + \frac{1_{\{I\in\tilde{\mathcal{A}}\}}}{|\tilde{\mathcal{A}}|^2}\right).$$

Taking expectation, we obtain

$$\mathbb{E}(\eta_n-\tilde{\eta}_n)^2 \le 2B^2\,\mathbb{E}\left[\frac{1_{\{I\in\mathcal{A}\}}}{|\mathcal{A}|^2}\right] = \frac{2B^2}{n}\,\mathbb{E}\left[\frac{1}{|\mathcal{A}|}\right],$$

where the final transition follows from $\mathbb{E}[\,1_{\{I\in\mathcal{A}\}} \mid |\mathcal{A}|] = |\mathcal{A}|/n$ (since I is uniform on $[n]$). Assume for the moment that $\lambda \le B$. Then we must have $\eta_n \ge 0$ and moreover $|\mathcal{A}| \ge n\min\{1, \lambda/B\}$ with probability 1. Substituting back into (40), we get the variance bound

$$\mathrm{Var}\eta_n \le \frac{B^3}{\lambda n}. \qquad (41)$$

Combining (41), (39) and (38) gives the result for $\lambda \le B$.

In the edge case that $\lambda \ge B$, Eq. (24) gives us that

$$\begin{aligned}
\overline{\mathcal{L}}_{\chi^2\text{-pen}}(x;n) &= \mathbb{E}\,\frac{1}{n}\sum_{i\le n}Z_i + \frac{1}{2\lambda}\,\mathbb{E}\,\mathrm{Var}[Z_1^n] = \mathbb{E}\,Z + \frac{n-1}{2n\lambda}\mathrm{Var}[Z] \\
&= \mathcal{L}(x;P_0) - \frac{1}{2\lambda}\mathrm{Var}[Z] \ge \mathcal{L}(x;P_0) - \frac{B^2}{2\lambda n}.
\end{aligned}$$

We note that in this case may easily form an unbiased estimator of $\mathcal{L}_{\chi^2\text{-pen}}$ by using the standard unbiased variance estimator. □

B.1.2 Worst-case tightness of bias bounds

Proposition 4. *For $p \in [0,1]$, let $P_0 = \mathsf{Bernoulli}(p_0)$ and $\ell(x;s) = B \cdot s$. The following results hold.*

- *Set $p_0 = \alpha$, then*
$$\mathcal{L}_{\mathrm{CVaR}}(x;P_0) - \overline{\mathcal{L}}_{\mathrm{CVaR}}(x;n) \gtrsim \frac{B\sqrt{1-\alpha}}{\sqrt{\alpha n}}.$$

- *Set $p_0 = (1+2\rho)^{-1}$, then*
$$\mathcal{L}_{\chi^2}(x;P_0) - \overline{\mathcal{L}}_{\chi^2}(x;n) \gtrsim B\sqrt{\frac{\rho}{n}}.$$

- *Set $p_0 = \lambda/B \le 1/2$, then*
$$\mathcal{L}_{\chi^2\text{-pen}}(x;P_0) - \overline{\mathcal{L}}_{\chi^2\text{-pen}}(x;n) \gtrsim \frac{B^2}{\lambda n}.$$

Proof. As before, we treat each case separately.

CVaR. First, note that $\mathcal{L}_{\mathrm{CVaR}}(x;P_0) = B$ since for Q such that $Q(1) = 1$ we have $\frac{\mathrm{d}Q}{\mathrm{d}P_0}(s) = \frac{1}{\alpha}1_{\{s=1\}}$ and therefore $Q \in \mathcal{U}^{\alpha}_{\mathrm{CVaR}}(P_0)$. Second, for a sample $S_1^n \in \{0,1\}^n$ we have
$$\mathcal{L}_{\mathrm{CVaR}}(x;S_1^n) = B\max\left\{1, \frac{1}{\alpha n}\sum_{i\in[n]} S_i\right\}.$$

Therefore
$$\begin{aligned}\mathcal{L}_{\mathrm{CVaR}}(x;P_0) - \overline{\mathcal{L}}_{\mathrm{CVaR}}(x;n) &= \frac{B}{\alpha n}\,\mathbb{E}\left(n\alpha - \sum_{i\in[n]} S_i\right)_+ \\ &\ge \frac{B\sqrt{1-\alpha}}{\sqrt{\alpha n}}\mathbb{P}\big(\mathsf{Bin}(n,\alpha) \le n\alpha - \sqrt{n\alpha(1-\alpha)}\big) \gtrsim \frac{B\sqrt{1-\alpha}}{\sqrt{\alpha n}},\end{aligned}$$
where the final bound follows from the Berry-Esseen theorem (see Lemma 1).

Constrained-χ^2. The χ^2 divergence between two Bernoulli random variables is
$$\mathrm{D}_{\chi^2}(\mathsf{Bernoulli}(q), \mathsf{Bernoulli}(p)) = \frac{1}{2}p\left(\frac{q}{p}-1\right)^2 + \frac{1}{2}(1-p)\left(\frac{1-q}{1-p}-1\right)^2 = \frac{(q-p)^2}{2p(1-p)}.$$
Therefore, for any $p \in (0,1)$, the element in $\mathcal{U}^{\rho}_{\chi^2}(\mathsf{Bernoulli}(p))$ that maximizes $Q(1)$ is $Q = \mathsf{Bernoulli}(q)$ with $q = \min\left\{1, p + \sqrt{2\rho}\sqrt{p(1-p)}\right\}$. Set $p_0 = \frac{1}{1+2\rho}$ and note that the function $f(p) = p + \sqrt{2\rho}\sqrt{p(1-p)} = p + \sqrt{(1-p_0)/p_0}\sqrt{p(1-p)}$ satisfies $f(p_0) = 1$ and $f'(p) \ge \frac{1}{2p_0}$ for all $p \le p_0$. Therefore, we have
$$\mathcal{L}_{\chi^2}(x;\mathsf{Bernoulli}(p)) \le B\left[1 - \left(\frac{p_0 - p}{2p_0}\right)_+\right]$$
for all $p \in (0,1)$, with equality at $p = p_0$. In particular, setting $P_0 = \mathsf{Bernoulli}(p_0)$ implies $\mathcal{L}_{\chi^2}(x;P_0) = B$ and for a sample $S_1^n \sim P_0^n$ with $\hat{p} = \frac{1}{n}\sum_{i\in[n]} S_i$ we have $\mathcal{L}_{\chi^2}(x;S_1^n) \le B\left(1 - \frac{p_0 - \hat{p}}{2p_0}\right)$. Therefore
$$\mathcal{L}_{\chi^2}(x;P_0) - \overline{\mathcal{L}}_{\chi^2}(x;n) \ge \frac{B}{2p_0}\,\mathbb{E}(p_0 - \hat{p})_+ = \frac{B}{2p_0 n}\cdot\mathbb{E}\left(np_0 - \sum_{i\in[n]} S_i\right)_+ \overset{(\star)}{\gtrsim} \frac{B\sqrt{1-p_0}}{\sqrt{p_0 n}} = B\sqrt{\frac{2\rho}{n}},$$
where $(\star)$ follows from the CVaR case and for the final equality we substitute the definition of p_0.

Penalized-χ^2. For any $p \in (0,1)$ we have

$$\begin{aligned}\mathcal{L}(x; \mathsf{Bernoulli}(p)) &= \sup_{q\in[0,1]} \left\{qB - \lambda \mathrm{D}_{\chi^2}(\mathsf{Bernoulli}(q), \mathsf{Bernoulli}(p))\right\} \\ &= \sup_{q\in(0,1)} \left\{ qB - \frac{\lambda(q-p)^2}{2p(1-p)}\right\} = \begin{cases} pB\left(1+\frac{(1-p)B}{2\lambda}\right) & p \le \lambda/B \\ B - \frac{\lambda(1-p)}{2p} & \text{otherwise.}\end{cases}\end{aligned}$$

Simplifying, we have,

$$\mathcal{L}(x; \mathsf{Bernoulli}(p)) \le \frac{B+\lambda}{2} + \frac{B^2}{2\lambda} \cdot \left[\left(p - \frac{\lambda}{B}\right) - \frac{B}{\lambda} \frac{\left(p-\frac{\lambda}{B}\right)_+^2}{1 + \frac{B}{\lambda}\left(p - \frac{\lambda}{B}\right)_+}\right],$$

with equality at $p = \lambda/B$. Taking $p_0 = \lambda/B$ and and for a sample $S_1^n \sim P_0^n$ letting $\hat{p} = \frac{1}{n}\sum_{i\in[n]} S_i$, we have

$$\mathcal{L}_{\chi^2\text{-pen}}(x; P_0) - \overline{\mathcal{L}}_{\chi^2\text{-pen}}(x; n) \ge -\frac{B}{2p_0}\,\mathbb{E}(\hat{p} - p_0) + \frac{B}{2p_0^2}\,\mathbb{E}\,\frac{(\hat{p}-p_0)_+^2}{1 + \frac{1}{p_0}(\hat{p}-p_0)_+}.$$

Since $\mathbb{E}\,\hat{p} = p_0$, we may lower bound this as

$$\mathcal{L}_{\chi^2\text{-pen}}(x; P_0) - \overline{\mathcal{L}}_{\chi^2\text{-pen}}(x; n) \ge \frac{B(1-p_0)}{4np_0} \mathbb{P}(\sqrt{n^{-1}p_0(1-p_0)} \le \hat{p} - p_0 \le p_0).$$

We have

$$\begin{aligned}&\mathbb{P}(\sqrt{n^{-1}p_0(1-p_0)} \le \hat{p} - p_0 \le p_0) \\ &\quad \ge \mathbb{P}(\mathsf{Bin}(n,p_0) \ge np_0 + \sqrt{np_0(1-p_0)}) - \mathbb{P}(\mathsf{Bin}(n,p_0) \ge 2np_0) \gtrsim 1\end{aligned}$$

by Berry-Esseen and Chernoff, and the result follows by substituting $p_0 = \lambda/B$. □

B.2 Discussion of additional assumptions

B.2.1 Smoothness of ℓ

The guarantees for the accelerated gradient iterations (10), detailed in Appendix B.4, require the objective function be smooth, i.e., have Lipschitz gradient. However, the degree of smoothness need not be high: as Nesterov [52] and subsequent work [43, 23] observed, even if $\nabla\mathcal{L}$ is order G^2/ϵ Lipschitz, acceleration allows finding an ϵ accurate solution in roughly GR/ϵ steps (a quadratic improvement over the SGM rate), as long as the gradient variance is itself of order ϵ; the accelerated rates in Theorem 1 stem from this fact.

By Claim 1, for $\mathcal{L}$ to have roughly G^2/ϵ Lipschitz gradient, the loss gradients $\nabla\ell$ have to be $H = G^2/\epsilon$ Lipschitz. This is in fact a weak assumption, because every G-Lipschitz loss ℓ has a *smoothed* version $\tilde{\ell}$ that satisfies $|\tilde{\ell}(x;s) - \ell(x;s)| \lesssim \epsilon$ for all x, s and that $\nabla\tilde{\ell}(x;s)$ is G^2/ϵ Lipschitz. For example, we may replace the hinge loss $\ell(x;s) = (1 - x^\top s)_+$ with $\tilde{\ell}(x;s) = \epsilon\log(1 + \exp([1 - x^\top s]/\epsilon))$. More generally, the smoothing [33]

$$\tilde{\ell}(x;s) = \inf_{y\in\mathcal{X}} \left\{ \ell(y;s) + \frac{G^2}{2\epsilon}\|y - x\|^2 \right\} \tag{42}$$

works for any G-Lipschitz ℓ.

In practice, we are often at liberty to replace the original loss ℓ with its smoothed version $\tilde{\ell}$ and minimize the resulting objective $\tilde{\mathcal{L}}$ which is guaranteed to be sufficiently smooth and approximates $\mathcal{L}$ to accuracy ϵ. Indeed, in the "statistical learning" model where we observe the entire $\ell(\cdot;S)$ per sample of $S \sim P_0$, we can apply the smoothing (42) to enforce the smoothness requirement without loss of generality. Therefore, our smoothness assumption can fail to hold only in situation where ℓ is non-smooth and ℓ and $\nabla\ell$ are strict black-boxes, so we cannot compute (42) without multiple black-box queries.

B.2.2 Lipschitz inverse-cdf

The inverse-cdf of $\ell(x; S)$ is Lipschitz if and only if the distribution of $\ell(x; S)$ has positive density in the interval $[\min_{s \in \mathbb{S}} \ell(x; s), \max_{s \in \mathbb{S}} \ell(x; s)]$. This is a rather strong assumption that fails whenever $\mathbb{S}$ is discrete or $\ell(x; S)$ is distributed as two separate bulks. However, the conclusions of our analysis under the Lipschitz inverse-cdf assumption hold under two natural relaxations.

Near-Lipschitz inverse-cdf and discrete loss distributions. Note that if F^{-1} satisfies $|F^{-1}(u) - \tilde{F}^{-1}(u)| \le \delta$ for all $u \in [0, 1]$ and a G_{icdf}-Lipschitz $\tilde{F}^{-1}(u)$, then we can repeat the proof of the bound (8) to show that $\mathcal{L}(x; P_0) - \overline{\mathcal{L}}(x; n) \le \delta + \frac{G_{\text{icdf}}}{n+1}$ for all objectives of the form (12). Moreover, suppose that P_0 is uniform on N elements s_1^N such that $\ell(x; s_i)$ is increasing in i, and suppose that it holds that

$$\ell(x; s_{i+1}) - \ell(x; s_i) \le \frac{G_{\text{icdf}}}{N}. \tag{43}$$

That is, the increments in the loss are not too far from uniform. Then, the piecewise linear function $\tilde{F}^{-1}$ connecting the steps in F^{-1} is G_{icdf}-Lipschitz and satisfies $|F^{-1}(u) - \tilde{F}^{-1}(u)| \le G_{\text{icdf}}/N$. Therefore, the assumption (43) implies that for any mini-batch size $n < N$, we have $\mathcal{L}(x; P_0) - \overline{\mathcal{L}}(x; n) \le 2G_{\text{icdf}}/(n + 1)$. We note also that the assumption $n < N$ is essentially without loss of generality, since for $n = N$ we can simply use a full-batch method with no bias at all.

CVaR bias bounds with locally Lipschitz inverse-cdf. The proof of the bound (37) also works if F^{-1} is Lipschitz in a small neighborhood of the CVaR cutoff $1 - \alpha$, because for values of u that are roughly $\sqrt{n^{-1}\alpha}$ far from $1 - \alpha$ we may bound $|1_{\{u \ge 1-\alpha\}} - \mathcal{I}_\alpha(u)|$ via tail bounds, as in the proof of the bound (5). Therefore, we expect the bias of $\mathcal{L}_{\text{CVaR}}(x; P_0) - \overline{\mathcal{L}}(x; n)$ to vanish with rate $1/n$ whenever the distribution of $\ell(x; S)$ has a density at the $1 - \alpha$ quantile loss value. Prior work shows that, from an asymptotic perspective, the converse is also true: when $\ell(x; S)$ does not have a density at the $1 - \alpha$ quantile, the bias vanishes with asymptotic rate $n^{-1/2}$ [cf. 71, Theorem 2].

B.3 Proofs of variance bounds

We give a more general statement of the variance bound using the notion of C-χ^2-bounded objectives (Definition 1); Proposition 2 follows immediately from Claim 4.

Proposition 2'. *Let $\mathcal{L}$ be an objective of the form (12). If $\mathcal{L}$ is C-χ^2-bounded, we have that for all $n \in \mathbb{N}$ and $x \in \mathcal{X}$*

$$\operatorname{Var}[\mathcal{L}(x; S_1^n)] \le \frac{2(1 + C)}{n} B^2.$$

If in addition $\phi = 0$ and ψ is strictly convex, we have

$$\operatorname{Var}[\nabla \mathcal{L}(x; S_1^n)] \le \frac{8(1 + C)}{n} G^2.$$

Proof. We first show the bound on the objective variance. By the the Efron-Stein inequality (see Lemma 2), we have

$$\operatorname{Var}[\mathcal{L}(x; S_1^n)] \le \frac{n}{2} \mathbb{E} \left(\mathcal{L}(x; S_1^n) - \mathcal{L}(x; \tilde{S}_1^n)\right)^2, \tag{44}$$

where S and $\tilde{S}$ are identical except in a random entry I for which $\tilde{S}_I$ is an iid copy of S_I. Let q and $\tilde{q}$ denote the maximizers of (13) for samples S_1^n and $\tilde{S}_1^n$ respectively. In addition, let $Z_i = \ell(x; S_i)$ and $\tilde{Z}_i = \ell(x; \tilde{S}_i)$. Clearly, $\mathcal{L}(x; S_1^n)$ is convex in Z and satisfies $\frac{\partial}{\partial Z} \mathcal{L}(x; S_1^n) = q$. Therefore,

$$\mathcal{L}(x; S_1^n) - \mathcal{L}(x; \tilde{S}_1^n) \le \left\langle \frac{\partial}{\partial Z} \mathcal{L}(x; S_1^n), Z - \tilde{Z} \right\rangle = q_I(\ell(x; S_I) - \ell(x; \tilde{S}_I)).$$

Applying the argument again with S and $\tilde{S}$ swapped, we find that

$$|\mathcal{L}(x; S_1^n) - \mathcal{L}(x; \tilde{S}_1^n)| \le \max\{q_I, \tilde{q}_I\} |\ell(x; S_I) - \ell(x; \tilde{S}_I)| \le B\sqrt{q_I^2 + \tilde{q}_I^2}.$$

Therefore, using the fact the q_I and $\tilde{q}_I$ are identically distributed, we have

$$\mathbb{E}(\mathcal{L}(x; S_1^n) - \mathcal{L}(x; \tilde{S}_1^n))^2 \le 2B^2 \, \mathbb{E} \, q_I^2 = \frac{2G^2}{n} \mathbb{E} \|q\|_2^2 = \frac{4B^2}{n^2}(C + 1),$$

where the final bound is due to $\|q\|_2^2 = \frac{1}{n}(2\mathrm{D}_{\chi^2}(q, \frac{1}{n}\mathbf{1}) + 1)$ and the C-χ^2-bounded property of $\mathcal{L}$. Substituting back into (44) gives the claimed objective variance bound.

Next, to show the bound on the gradient variance we invoke Efron-Stein elementwise to obtain

$$\mathrm{Var}[\nabla \mathcal{L}(x; S_1^n)] \le \frac{n}{2}\,\mathbb{E}\,\|\nabla \mathcal{L}(x; S_1^n) - \nabla \mathcal{L}(x; \tilde{S}_1^n)\|^2.$$

By the expression (28) for $\nabla \mathcal{L}$ we have

$$\begin{aligned}\|\nabla \mathcal{L}(x; S_1^n) - \nabla \mathcal{L}(x; \tilde{S}_1^n)\| &= \left\| \sum_{i \ne I} (q_i - \tilde{q}_i) \nabla \ell(x; S_i) + q_I \nabla \ell(x; S_I) - \tilde{q}_I \nabla \ell(x; \tilde{S}_I) \right\| \\ &\le G \left(\sum_{i \ne I} |q_i - \tilde{q}_i| + q_I + \tilde{q}_I \right),\end{aligned}$$

where the bound follows from the triangle inequality and the fact that ℓ is G-Lipschitz.

Now, observe that $q_i = \frac{1}{n}\psi^{*\prime}[(\ell(x; S_i) - \eta)/\lambda]$ for some $\eta \in \mathbb{R}$ by Eq. (20). Similarly, $q_i = \frac{1}{n}\psi^{*\prime}[(\ell(x; S_i) - \tilde{\eta})/\lambda]$ for some $\tilde{\eta}$. Since ψ is strictly convex we have that $\psi^{*\prime}$ is continuous and monotonic non-decreasing. Consequently, either $q_i \ge \tilde{q}_i$ for all $i \ne I$ (if $\eta \le \tilde{\eta}$), or $q_i \le \tilde{q}_i$ for all $i \ne I$ (if $\eta \ge \tilde{\eta}$). In either case, we have

$$\sum_{i \ne I} |q_i - \tilde{q}_i| = \left| \sum_{i \ne I} (q_i - \tilde{q}_i) \right| = |q_I - \tilde{q}_I|,$$

where the final equality used the fact that $\sum_{i \le n} q_i = \sum_{i \le n} \tilde{q}_i = 1$. Substituting back, we find that

$$\|\nabla \mathcal{L}(x; S_1^n) - \nabla \mathcal{L}(x; \tilde{S}_1^n)\| \le 2G \max\{q_I, \tilde{q}_I\} \le 2G\sqrt{q_I^2 + \tilde{q}_I^2}.$$

The remainder of the proof is identical to that of the objective variance bound, except with $2G$ replacing B. □

Proposition 2' implies that the variance of the χ^2 constraint objective $\mathcal{L}_{\chi^2}$ is at most $2(1+\rho)B^2/n$. However, our gradient variance bound requires $\phi = 0$ and therefore does not apply to $\nabla \mathcal{L}_{\chi^2}$. The following proposition shows that the requirement $\phi = 0$ is necessary, since no upper bound of the from $O(1)(1+\rho)G^2/n$ holds for $\mathrm{Var}[\nabla \mathcal{L}_{\chi^2}]$.

Proposition 5 (Variance of the mini-batch gradient estimator for $\mathcal{L}_{\chi^2}$)**.** *For any $n > 4$ and $\rho \ge 0$, there exists a distribution P_0 over $\mathbb{S} = \{0, 1, 2\}$ and a G-Lipschitz loss $\ell : [-1, 1] \times \mathbb{S} \to [0, 1]$ such that*

$$\mathrm{Var}\big[\nabla \mathcal{L}_{\chi^2}(0; S_1^n)\big] \gtrsim \frac{\rho^2}{(1+\rho)^2} G^2.$$

Proof. We construct P_0 as follows,

$$\mathbb{P}(S = 2) = p_2 = 1 - 2^{1/n} \approx \frac{\log 2}{n} \;\text{ and }\; \mathbb{P}(S = 1) = p_1 = \frac{1}{1 + 2\rho}.$$

(Note that we may assume without loss of generality that $\rho \gtrsim 1/n$, so that $P(S = 0) = 1 - p_1 - p_2 > 0$, since for $\rho = 0$ we already have a standard $1/n$ lower bound on the variance). We set the loss values to be

$$\ell(0; 0) = 0 \;\;,\;\; \ell(0; 1) = \frac{1}{30n} \;\text{ and }\; \ell(0; 2) = 1,$$

and the loss gradients as

$$\nabla \ell(0; 2) = \nabla \ell(0; 0) = -G \;\text{ and }\; \ell(0; 1) = G.$$

The source of high variance in this construction is that, for a sample S_1^n, the maximizing $q^\star$ behaves very differently when $S_i = 2$ for some i and when $S_i \ne 2$ for all i. In the former case, we show that $q^\star$ puts significant mass on samples with $S_i \ne 1$, so $\nabla \mathcal{L}(0; S_1^n) < G(1 - c)$ for some $c \gtrsim \rho/(1+\rho)$. In the latter case, we show that with constant probability $q^\star$ places mass only on samples with $S_i = 1$, and so $\nabla \mathcal{L}(0; S_1^n) = G$. Since either scenario occurs with constant probability, the variance bound follows.

To provide a detailed proof, let $\mathrm{C}_k(S_1^n) = \sum_{i \le n} 1_{\{S_i = k\}}$ be the number of samples with value k, for $k \in \{0, 1, 2\}$, and consider the events

$$\mathfrak{E}_a(S_1^n) = \{\mathrm{C}_2(S_1^n) = 0 \text{ and } \mathrm{C}_1(S_1^n) \ge np_1\}$$

and

$$\mathfrak{E}_b(S_1^n) = \{\mathrm{C}_2(S_1^n) = 1 \text{ and } \mathrm{C}_1(S_1^n) < np_1\}.$$

Note that we chose p_2 such that $\mathbb{P}(\mathrm{C}_2(S_1^n) = 0) = (1 - p_2)^n = \frac{1}{2}$ and that $\mathbb{P}(\mathrm{C}_1(S_1^n) \ge np_1) \gtrsim 1$ since np_1 is roughly the median of $\mathrm{C}_1(S_1^n)$. Similarly, $\mathbb{P}(\mathrm{C}_2(S_1^n) = 1) = np_0(1 - p_0)^{n-1} \approx \frac{\log 2}{2}$ and $\mathbb{P}(\mathrm{C}_1(S_1^n) < np_1) \gtrsim 1$. Therefore,

$$\mathbb{P}(\mathfrak{E}_a(S_1^n)) \gtrsim 1 \;\text{ and }\; \mathbb{P}(\mathfrak{E}_b(S_1^n)) \gtrsim 1. \tag{45}$$

We bound $\nabla \mathcal{L}_{\chi^2}$ conditional on each event in turn.

Under event $\mathfrak{E}_a(S_1^n)$, the empirical loss distribution is Bernoulli with parameter $\mathrm{C}_1(S_1^n)/n \ge p_1 = 1/(1 + 2\rho)$ and consequently $q^\star$ places mass only on samples with value 1 (see further discussion in the proof of Proposition 4). Therefore, we have

$$\mathbb{E}[\nabla \mathcal{L}_{\chi^2}(0; S_1^n) \mid \mathfrak{E}_a(S_1^n)] = G. \tag{46}$$

To bound the gradient under event $\mathfrak{E}_b(S_1^n)$, assume that without loss of generality that $S_1 = 2$ is the unique sample with that value. We consider separately the cases $q_1^\star > 2/3$ and $q_1^\star \le 2/3$. In the former, we clearly have $\nabla \mathcal{L}_{\chi^2}(0; S_1^n) \le -q_1^\star G + (1 - q_1^\star) G < -G/3$. In the latter case, we recall Eq. (27) showing that $q^\star$ is of the form

$$q_i^\star = \frac{(\ell(x; S_i) - \eta^\star)_+}{\sum_{j \le n} (\ell(x; S_j) - \eta^\star)_+}$$

for some $\eta^\star \in \mathbb{R}$. The fact that $q_1^\star \le 2/3$ and that there are at most n samples with value $\ell(0; 1) = 1/(30n)$ gives the following bound on $\eta^\star$

$$\frac{2}{3} \ge q_1^\star = \frac{\ell(0; S_1) - \eta^\star}{\sum_{j \le n} (\ell(x; S_j) - \eta^\star)_+} \ge \frac{1 - \eta^\star}{31/30 - n\eta^\star} \implies \eta^\star \le -\frac{1}{3n}.$$

Suppose $S_j = 0$ and $S_i = 1$, then

$$r = \frac{q_j^\star}{q_i^\star} = \frac{\ell(0; 0) - \eta^\star}{\ell(0; 1) - \eta^\star} = 1 - \frac{\ell(0; 1)}{\ell(0; 1) - \eta^\star} \ge \frac{7}{8}.$$

Assuming that $S_1 = 2$, we have that the total weight under $q^\star$ of samples with gradient $-G$ is

$$q_1^\star + (1 - q_1^\star) \frac{r\mathrm{C}_0(S_1^n)}{\mathrm{C}_1(S_1^n) + r\mathrm{C}_0(S_1^n)} \ge \frac{7}{8}(1 - p_1) = \frac{7\rho}{4(1 + 2\rho)},$$

which implies $\nabla \mathcal{L}_{\chi^2}(0; S_1^n) \le -\frac{7\rho}{4(1+2\rho)} G + (1 - \frac{7\rho}{4(1+2\rho)}) G \le G(1 - \frac{7\rho}{2(1+2\rho)})$. We conclude that

$$\mathbb{E}[\nabla \mathcal{L}_{\chi^2}(0; S_1^n) \mid \mathfrak{E}_b(S_1^n)] \le G(1 - c) \;\text{ for }\; c = \min\left\{\frac{4}{3}, \frac{7\rho}{2(1 + 2\rho)}\right\} \gtrsim \frac{\rho}{1 + \rho}. \tag{47}$$

Let $\tilde{S}_1^n$ be an independent copy of S_1^n. We combine our conclusions (45), (46) and (47) to form a variance bound as follows,

$$\begin{aligned}
\mathrm{Var}[\nabla \mathcal{L}_{\chi^2}(0; S_1^n)] &= \frac{1}{2} \mathbb{E}\Big(\nabla \mathcal{L}_{\chi^2}(0; S_1^n) - \nabla \mathcal{L}_{\chi^2}(0; \tilde{S}_1^n)\Big)^2 \\
&\ge \frac{1}{2} \mathbb{E}\left[\Big(\nabla \mathcal{L}_{\chi^2}(0; S_1^n) - \nabla \mathcal{L}_{\chi^2}(0; \tilde{S}_1^n)\Big)^2 \,\Big|\, \mathfrak{E}_a(S_1^n), \mathfrak{E}_b(\tilde{S}_1^n)\right] \mathbb{P}(\mathfrak{E}_a(S_1^n), \mathfrak{E}_b(\tilde{S}_1^n)) \\
&\ge \frac{1}{2} c^2 G^2 \cdot \mathbb{P}(\mathfrak{E}_a(S_1^n)) \cdot \mathbb{P}(\mathfrak{E}_b(\tilde{S}_1^n)) \gtrsim \frac{\rho^2}{(1 + \rho)^2} G^2.
\end{aligned}$$

□

B.4 Convergence rates of stochastic gradient methods

We state below the classical convergence rates for standard and accelerated stochastic gradient methods, under a somewhat non-standard assumption that the stochastic gradient estimates are unbiased for a uniform approximation of the objective function with additive error δ.

Proposition 6 (Convergence of stochastic gradient methods [43, Corollary 1])**.** *Let $F : \mathcal{X} \to \mathbb{R}$ and $\overline{F} : \mathcal{X} \to \mathbb{R}$ satisfy $0 \le F(x) - \overline{F}(x) \le \delta$ for all $\mathcal{X} \in \mathbb{R}$. Assume that $\overline{F}$ is convex and that a stochastic gradient estimator $\tilde{g}$ satisfies $\mathbb{E}\,\tilde{g}(x) \in \partial\overline{F}(x)$ and $\mathbb{E}\,\|\tilde{g}(x)\|^2 \le \Gamma^2$ for all $x \in \mathcal{X}$. For $T \in \mathbb{N}$, the iterate $\bar{x}_T$ in the sequence (9) with $\eta \asymp \frac{R}{T^{1/2}\Gamma}$ satisfies*

$$\mathbb{E}\,F(\bar{x}_T) - \inf_{x'} F(x') \lesssim \delta + \frac{\Gamma R}{\sqrt{T}}. \tag{48}$$

If in addition $\nabla\overline{F}$ is Λ-Lipschitz and $\mathrm{Var}[\tilde{g}(x)] \le \sigma^2$ for all $x \in \mathcal{X}$, the iterate y_T in the sequence (10) with $\eta \asymp \min\{\frac{1}{\Lambda}, \frac{R}{T^{3/2}\sigma}\}$ and $\theta_t = \frac{2}{t+1}$ satisfies

$$\mathbb{E}\,F(y_T) - \inf_{x'} F(x') \lesssim \delta + \frac{\Lambda R^2}{T^2} + \frac{\sigma R}{\sqrt{T}}. \tag{49}$$

Proof. [43] gives us the rates (48) and (49) but for $\overline{F}$ rather than F. That is, it guarantees that SGM finds $\bar{x}_T$ such that

$$\mathbb{E}\,\overline{F}(\bar{x}_T) - \inf_{x'} \overline{F}(x') \lesssim \frac{\Gamma R}{\sqrt{T}}.$$

To remove the bars, we use $0 \le F(x) - \overline{F}(x) \le \delta$ to write

$$-\inf_{x'} F(x') \ge -\inf_{x'} \overline{F}(x') \;\text{ and }\; F(\bar{x}_T) \ge \overline{F}(\bar{x}_T) + \delta.$$

□

Remark 2. In the unconstrained case $\mathcal{X} = \mathbb{R}^d$, the recursion (10) reduces to the more familiar form

$$v_{t+1} = \omega_t v_t - \eta\tilde{g}(x_t), \;\; x_{t+1} = x_t + \omega_{t+1}v_{t+1} - \eta\tilde{g}(x_t), \tag{50}$$

where $\omega_t = (1 - \theta_{t-1})\frac{\theta_t}{\theta_{t-1}}$ is a time-varying "momentum" parameter; the sequences y_t, z_t are related to v_t via $v_t = \frac{\theta_t}{\omega_t}(z_t - y_t)$ and $y_{t+1} = x_t - \eta\tilde{g}(x_t)$.

B.5 Proofs of complexity bounds

Theorem 1. *Let Assumptions A1 and A2 hold, possibly trivially (with $H = \infty$ or $G_{\mathrm{icdf}} = \infty$). Let $\epsilon \in (0, B)$ and write $\nu = \frac{H}{G^2}\epsilon$. With suitable choices of the batch size n and iteration count T, the gradient methods (9) and (10) find $\bar{x}$ satisfying $\mathbb{E}\,\mathcal{L}(\bar{x}, P_0) - \inf_{x' \in \mathcal{X}} \mathcal{L}(x'; P_0) \le \epsilon$ with complexity nT admitting the following bounds.*

- *For $\mathcal{L} = \mathcal{L}_{\mathrm{CVaR}}$, we have $nT \lesssim \frac{(GR)^2}{\alpha\epsilon^2}\left(1 + \min\left\{\frac{\alpha G_{\mathrm{icdf}}\sqrt{\log\frac{1}{\alpha} + \nu}}{GR}, \frac{B^2\sqrt{\log\frac{1}{\alpha} + \nu}}{GR\epsilon}, \frac{B^2}{\epsilon^2}\right\}\right)$.*
- *For $\mathcal{L} = \mathcal{L}_{\chi^2\text{-pen}}$ with $\lambda \le B$, we have $nT \lesssim \frac{(GR)^2 B}{\lambda\epsilon^2}\left(1 + \min\left\{\frac{B}{GR}\sqrt{\frac{\epsilon(1+\nu)}{\lambda}}, \frac{B}{\epsilon}\right\}\right)$.*
- *For $\mathcal{L} = \mathcal{L}_{\chi^2}$, we have $nT \lesssim \frac{(1+\rho)(GR)^2B^2}{\epsilon^4}\log\frac{(1+\rho)B^2}{\epsilon^2}$.*
- *For any loss of the from (2), we have $nT \lesssim \frac{(GR)^2 G_{\mathrm{icdf}}}{\epsilon^3}$.*

Proof. To prove each bound in the theorem we choose n large enough via one of the bounds in Proposition 1 and then choose T to guarantee ϵ-accurate solution via Proposition 6. For a (potentially random) point $\bar{x} \in \mathcal{X}$ and robust risk $\mathcal{L}$, we define the shorthand

$$\mathsf{err}(x; \mathcal{L}) := \mathbb{E}\,\mathcal{L}(\bar{x}; P_0) - \inf_{x \in \mathcal{X}} \mathcal{L}(x; P_0).$$

We summarize our choices of n and T for different robust objectives, under different assumptions in Table B.5. In the statement of the theorem, we sometimes upper bound $a \vee b := \max\{a, b\}$ by $a + b$ for readability, and state the tighter rates here.

Loss	∇ est.	n	T	complexity $= nT$
$\mathcal{L}_{\text{CVaR}}$	$\nabla\mathcal{L}_{\text{CVaR}}$	$\frac{B^2}{\alpha\epsilon^2}$	$\frac{(GR)^2}{\epsilon^2}$	$\frac{(GR)^2B^2}{\alpha\epsilon^4}$
$\mathcal{L}_{\chi^2}$	$\nabla\mathcal{L}_{\chi^2}$	$\frac{(1+\rho)B^2}{\epsilon^2}\log\frac{(1+\rho)B^2}{\epsilon^2}$	$\frac{(GR)^2}{\epsilon^2}$	$\frac{(1+\rho)(GR)^2B^2}{\epsilon^4}\log\frac{(1+\rho)B^2}{\epsilon^2}$
$\mathcal{L}_{\chi^2\text{-pen}}$	$\nabla\mathcal{L}_{\chi^2\text{-pen}}$	$\frac{B^2}{\lambda\epsilon}$	$\frac{(GR)^2}{\epsilon^2}$	$\frac{(GR)^2B^2}{\lambda\epsilon^3}$
any $\mathcal{L}$ in (2)	$\nabla\mathcal{L}$	$\frac{G_{\text{icdf}}}{\epsilon}$	$\frac{(GR)^2}{\epsilon^2}$	$\frac{(GR)^2G_{\text{icdf}}}{\epsilon^3}$
$\mathcal{L}_{\text{CVaR}}$	$\nabla\mathcal{L}_{\text{kl-CVaR}}$	$\frac{B^2}{\alpha\epsilon^2}$	$\frac{GR\sqrt{\log\frac{1}{\alpha}+\nu}}{\epsilon}\vee\frac{(GR)^2}{B^2}$	$\frac{(GR)^2}{\alpha\epsilon^2}\left(1\vee\frac{B^2}{GR\epsilon}\sqrt{\log\frac{1}{\alpha}+\nu}\right)$
$\mathcal{L}_{\text{CVaR}}$	$\nabla\mathcal{L}_{\text{kl-CVaR}}$	$\frac{G_{\text{icdf}}}{\epsilon}$	$\frac{GR\sqrt{\log\frac{1}{\alpha}+\nu}}{\epsilon}\vee\frac{(GR)^2}{\alpha G_{\text{icdf}}\epsilon}$	$\frac{(GR)^2}{\epsilon^2}\left(\frac{1}{\alpha}\vee\frac{G_{\text{icdf}}}{GR}\sqrt{\log\frac{1}{\alpha}+\nu}\right)$
$\mathcal{L}_{\chi^2\text{-pen}}$	$\nabla\mathcal{L}_{\chi^2\text{-pen}}$	$\frac{B^2}{\lambda\epsilon}$	$\frac{GR\sqrt{1+\nu}}{\epsilon}\vee\frac{(GR)^2}{B\epsilon}$	$\frac{GRB}{\lambda\epsilon^2}\left(B\sqrt{1+\nu}\vee GR\right)$

Table 2. Parameter settings for Theorem 1. For $\mathcal{L}_{\text{kl-CVaR}}$ we take $\lambda \asymp \frac{\epsilon}{\log(\frac{1}{\alpha})}$. For $\mathcal{L}_{\chi^2\text{-pen}}$ assume $\lambda \ge \epsilon$. $\nu := H\epsilon/G^2$. We use the shorthand $a \vee b$ for $\max\{a, b\}$.

CVaR. We distinguish between the different possible assumptions on the loss ℓ and distribution P_0 as they yield different rates.

(a) Non-smooth ℓ: let $\bar{x}_T$ be the iterates of (9), the sub-optimality guarantee of (48) and the bias bound of Proposition 1 yield

$$\mathsf{err}(\bar{x}_T; \mathcal{L}_{\text{CVaR}}) \lesssim \frac{B}{\sqrt{\alpha n}} + \frac{GR}{\sqrt{T}}.$$

In that case, setting $n \asymp \frac{B^2}{\alpha\epsilon^2}$ guarantees that the bias is smaller than ϵ and setting $T \asymp \frac{(GR)^2}{\epsilon^2}$ yields that $\mathsf{err}(\bar{x}_T; \mathcal{L}_{\text{CVaR}}) \lesssim \epsilon$.

(b) Smooth ℓ: if ℓ is H-smooth, we consider the $\mathcal{L}_{\text{kl-CVaR}}$ objective with $\lambda = \frac{\epsilon}{\log\frac{1}{\alpha}}$. This guarantees that, for all $x \in \mathcal{X}$

$$\mathcal{L}_{\text{kl-CVaR}}(x; P_0) \le \mathcal{L}_{\text{CVaR}}(x; P_0) \le \mathcal{L}_{\text{kl-CVaR}}(x; P_0) + \epsilon.$$

$\overline{\mathcal{L}}_{\text{kl-CVaR}}$ being $(\frac{G^2\log(1/\alpha)}{\epsilon} + H)$-smooth, the final iterate of the sequence (10) achieves

$$\mathsf{err}(y_T; \mathcal{L}_{\text{CVaR}}) \lesssim \epsilon + \frac{B}{\sqrt{\alpha n}} + \frac{(GR)^2(\log\frac{1}{\alpha} + \nu)}{\epsilon T^2} + \frac{GR}{\sqrt{\alpha n T}}.$$

To make sure that the second and third terms are smaller than ϵ, we set $T = \frac{(GR)^2}{\alpha n\epsilon^2} \vee \frac{GR}{\epsilon}\sqrt{\log\frac{1}{\alpha} + \nu}$. To guarantee small bias, we set $n \asymp \frac{B^2}{\alpha\epsilon^2}$; the resulting complexity is

$$nT \asymp \frac{(GR)^2}{\alpha\epsilon^2}\max\left\{1, \frac{B^2}{GR\epsilon}\sqrt{\log\frac{1}{\alpha} + \nu}\right\}.$$

(c) Smooth ℓ and inverse cdf Lipschitz: in this case, the regret guarantees of the iterates of (49) is

$$\mathsf{err}(y_T; \mathcal{L}_{\text{CVaR}}) \lesssim \epsilon + \frac{G_{\text{icdf}}}{n} + \frac{(GR)^2(\log\frac{1}{\alpha} + \nu)}{\epsilon T^2} + \frac{GR}{\sqrt{\alpha n T}}.$$

We once again set $T = \frac{(GR)^2}{\alpha n\epsilon^2} \vee \frac{GR}{\epsilon}\sqrt{\log\frac{1}{\alpha} + \nu}$, and choosing $n \asymp \frac{G_{\text{icdf}}}{\epsilon}$ yields the result.

Penalized-χ^2. We distinguish between whether or not ℓ is smooth.

(a) Non-smooth ℓ: for the sequence of iterates of (9), we have

$$\mathsf{err}(\bar{x}_T; \mathcal{L}_{\chi^2\text{-pen}}) \lesssim \frac{B^2}{\lambda n} + \frac{GR}{\sqrt{T}},$$

and setting $n \asymp \frac{B^2}{\lambda \epsilon}$ and $T \asymp \frac{(GR)^2}{\epsilon^2}$ yields the fist rate.

(b) Smooth ℓ: We now turn to acceleration, we have

$$\mathsf{err}(y_T; \mathcal{L}_{\chi^2\text{-pen}}) \lesssim \frac{B^2}{\lambda n} + \frac{R^2\left(\frac{G^2}{\lambda} + H\right)}{T^2} + GR\sqrt{\frac{1 + \frac{B}{\lambda}}{nT}}.$$

First, noting that $\lambda \ge \epsilon$ guarantees that $R^2(\frac{G^2}{\lambda} + H) \le \frac{(GR)^2(1+\nu)}{\epsilon}$. Furthermore, we simplify the variance term since $B/\lambda \ge 1$. We thus set $T \asymp \frac{GR}{\epsilon}\sqrt{1+\nu} \vee \frac{(GR)^2 B}{\lambda n \epsilon^2}$ and choose $n \asymp \frac{B^2}{\lambda \epsilon^2}$. This yields the final result

$$nT \lesssim \frac{GRB}{\lambda \epsilon^2}\big(B\sqrt{1+\nu} \vee (GR)\big).$$

Constrained-χ^2. This case is straightforward—without any bound on the variance in the worst-case, we turn to the basic SGM guarantee (48); we have

$$\mathsf{err}(\bar{x}_T; \mathcal{L}_{\chi^2}) \lesssim B\sqrt{1+2\rho}\sqrt{\frac{\log n}{n}} + \frac{GR}{\sqrt{T}}.$$

We set $T \asymp \frac{(GR)^2}{\epsilon^2}$ and $n \asymp \frac{(1+2\rho)B^2}{\epsilon^2}\log((1+2\rho)B^2\epsilon^{-2})$. We then have

$$B\sqrt{1+2\rho}\sqrt{\frac{\log n}{n}} = \epsilon\sqrt{1 + \frac{\log\log(\frac{(1+2\rho)B^2}{\epsilon^2})}{\log(\frac{(1+2\rho)B^2}{\epsilon^2})}} \le \sqrt{2}\epsilon,$$

and this concludes the proof.

Lipschitz inverse-cdf. The sequence of iterates (9) yield error

$$\mathsf{err}(\bar{x}_T; \mathcal{L}) \le \frac{G_{\text{icdf}}}{n} + \frac{GR}{\sqrt{T}},$$

and setting $n \asymp \frac{G_{\text{icdf}}}{\epsilon}, T \asymp \frac{(GR)^2}{\epsilon^2}$ concludes the proof of the theorem. □

C Proofs of Section 4

We now provide additional discussion of the multilevel Monte Carlo estimator for general functions F, whose form we restate here for convenience

$$\widehat{\mathcal{M}}[\mathsf{F}] := \mathsf{F}(x; S_1^{n_0}) + \frac{1}{q(J)}\widehat{\mathcal{D}}_{2^J n_0}, \text{ where } \widehat{\mathcal{D}}_k := \mathsf{F}(x; S_1^k) - \frac{\mathsf{F}(x; S_1^{k/2}) + \mathsf{F}(x; S_{k/2+1}^k)}{2}. \quad (51)$$

Section C.1 provides upper bounds on the moments of $\widehat{\mathcal{M}}$ for estimating $\mathcal{L}_{\text{kl-CVaR}}$, $\mathcal{L}_{\chi^2\text{-pen}}$, and their gradients, proving Claim 2 and Proposition 3. In that section we also prove that similar second moment bounds do not always hold for $\nabla\mathcal{L}_{\chi^2}$. In Section C.2 we prove the complexity guarantees in Theorem 2, and we conclude in Section C.3 with a comparison of some of our design choices to the original proposal of Blanchet and Glynn [8].

C.1 Proofs of moment bounds

Claim 2'. *For any function* F, *the estimator* $\widehat{\mathcal{M}}[\mathsf{F}]$ *with parameters* $n = 2^{j_{\max}} n_0$ *satisfies*

$$\mathbb{E}\,\widehat{\mathcal{M}}[\mathsf{F}] = \mathbb{E}\,\mathsf{F}(S_1^n), \text{ requiring expected sample size } \mathbb{E}\,2^J n_0 = n_0(1+\log_2(n/n_0)).$$

Proof. For any even k, $\mathbb{E}\,\widehat{\mathcal{D}}_k = \mathbb{E}\,\mathsf{F}(S_1^k) - \mathbb{E}\,\mathsf{F}(S_1^{k/2})$. Therefore, the expectation of $\widehat{\mathcal{M}}[\mathsf{F}]$ telescopes: $\mathbb{E}\,\widehat{\mathcal{M}}[\mathsf{F}] = \mathbb{E}[\mathsf{F}(S_1^{n_0})] + \sum_{j=1}^{j_{\max}} \mathbb{E}\,\widehat{\mathcal{D}}_{2^j n_0} = \mathbb{E}[\mathsf{F}(S_1^n)]$. The expected number of samples follows from direct calculation: $\mathbb{E}[2^J] = \sum_{j=1}^{j_{\max}} 2^j \mathbb{P}(J=j) = j_{\max}+1$. □

We have the following bound on the second moment of the estimator,

$$\mathbb{E}\left\|\widehat{\mathcal{M}}[\mathsf{F}]\right\|^2 \le 2\left\|\mathsf{F}(S_1^{n_0})\right\|^2 + \sum_{j=1}^{j_{\max}} \frac{2}{q(j)}\,\mathbb{E}\left\|\widehat{\mathcal{D}}_{2^j n_0}\right\|^2 \le 2\left\|\mathsf{F}(S_1^{n_0})\right\|^2 + \sum_{j=1}^{j_{\max}} 2^{j+1}\,\mathbb{E}\left\|\widehat{\mathcal{D}}_{2^j n_0}\right\|^2. \quad (52)$$

For χ^2-bounded (Definition 1) pure-penalty losses such as $\mathcal{L}_{\text{kl-CVaR}}$ and $\mathcal{L}_{\chi^2\text{-pen}}$, we argue that $\mathbb{E}\|\widehat{\mathcal{D}}_k\|^2 \lesssim 1/k$, so that $2^j\,\mathbb{E}\|\widehat{\mathcal{D}}_{2^j n_0}\|^2 \lesssim 1/n_0$. Substituting into the bound (52) gives the following guarantees, from which Proposition 3 follows immediately via Claim 4.

Proposition 3'. *Let* $\mathcal{L}$ *be an objective of the form* (12) *with* $\phi = 0$ *and strictly convex* ψ. *If* $\mathcal{L}$ *is* C-χ^2*-bounded, we have that for all* $x \in \mathcal{X}$, *the multi-level Monte Carlo estimator with parameters* n *and* n_0 *satisfies*

$$\mathbb{E}\left(\widehat{\mathcal{M}}[\mathcal{L}]\right)^2 \le 2B^2\left(1 + \frac{2C}{n_0}\log_2(n/n_0)\right) \text{ and } \mathbb{E}\left\|\widehat{\mathcal{M}}[\nabla\mathcal{L}]\right\|^2 \le 2G^2\left(1+\frac{2C}{n_0}\log_2(n/n_0)\right).$$

Proof. The proof follows similarly to the proof of Proposition 2', where the key step is to bound $\widehat{\mathcal{D}}_k$ for $k \in 2\mathbb{N}$. We distinguish between estimating the gradient and the loss as, for the latter, one needs to account for estimating the regularizer D_ψ.

Gradient estimator. We start with the proof of the second moment of the gradient estimator. Let $k \in 2\mathbb{N}$ and let q, q' and q'' be the maximizer of (13) for $S_1^k, S_1^{k/2}$ and $S_{k/2+1}^k$ respectively. We have

$$\begin{aligned}\|\widehat{\mathcal{D}}_k\| &= \left\|\sum_{i\le k}\Big(q_i - \tfrac{1}{2}q_i' 1_{\{i\le k/2\}} - \tfrac{1}{2}q_{i-k/2}'' 1_{\{i>k/2\}}\Big)\nabla\ell(x;S_i)\right\| \\ &\le G\sum_{i\le k/2}|q_i - \tfrac{1}{2}q_i'| + G\sum_{i>k/2}|q_i - \tfrac{1}{2}q_{i-k/2}''|.\end{aligned}$$

For $i \in \{1,\dots,k/2\}$, it holds that $q_i = \frac{1}{n}\psi^{*\prime}[(\ell(x;S_i)-\eta)/\lambda]$ and $q_i' = \frac{2}{n}\psi^{*\prime}[(\ell(x;S_i)-\eta')/\lambda]$ for $\eta,\eta' \in \mathbb{R}$. Since that ψ is strictly convex, $\psi^{*\prime}$ is increasing and $q_i - \frac{1}{2}q_i'$ is of constant sign for $i \in \{1,\dots,k/2\}$. Therefore,

$$\sum_{i\le k/2}\left|q_i - \frac{1}{2}q_i'\right| = \left|\sum_{i\le k/2}\left(q_i - \frac{1}{2}q_i'\right)\right| = \left|\sum_{i\le k/2} q_i - \frac{1}{2}\right|.$$

By symmetry, it thus holds that

$$\mathbb{E}\,\|\widehat{\mathcal{D}}_k\|^2 \le 4G^2\,\mathbb{E}\left(\sum_{i\le k/2} q_i - \frac{1}{2}\right)^2 \overset{(i)}{\le} \frac{2}{k}\mathrm{D}_{\chi^2}(q, \tfrac{1}{k}\mathbf{1}) \overset{(ii)}{\le} \frac{2CG^2}{k},$$

where (i) is due to Lemma 3 and (ii) follows from the assumption that $\mathcal{L}$ is C-χ^2-bounded. Substituting into (52), we have

$$\begin{aligned}\mathbb{E}\,\|\widehat{\mathcal{M}}[\nabla\mathcal{L}]\|^2 &\le 2G^2 + 4CG^2\sum_{j\le 1}^{j_{\max}}\frac{1}{q(j)2^j n_0} \\ &\le 2G^2 + \frac{4CG^2}{n_0}\left(j_{\max} - \frac{1}{2}\right) \\ &\le G^2\left(2 + \frac{4C}{n_0}\log_2(n/n_0)\right).\end{aligned}$$

This concludes the argument for the gradient.

Loss estimator. With the same notation, let us define $\tilde{q} \coloneqq [\frac{1}{2}q', \frac{1}{2}q''] \in \Delta^k$. We first prove that $\widehat{\mathcal{D}}_k \geq 0$. Indeed, we have

$$\begin{aligned}\mathcal{L}(x; S_1^k) = \sum_{i \leq k} q_i \ell(x; S_i) - \mathrm{D}_\psi(q, \tfrac{1}{k}\mathbf{1}) &\overset{(i)}{\geq} \sum_{i \leq k} \tilde{q}_i \ell(x; S_i) - \mathrm{D}_\psi(\tilde{q}, \tfrac{1}{k}\mathbf{1}) \\ &\overset{(ii)}{=} \frac{1}{2}\mathcal{L}(x; S_1^{k/2}) + \frac{1}{2}\mathcal{L}(x; S_{k/2+1}^k),\end{aligned}$$

where (i) is because q is the maximizer for S_1^k and (ii) because the ψ-divergence tensorizes, i.e., $\mathrm{D}_\psi(\tilde{q}, \frac{1}{k}\mathbf{1}) = \frac{1}{2}\mathrm{D}_\psi(q', \frac{2}{k}\mathbf{1}) + \frac{1}{2}\mathrm{D}_\psi(q'', \frac{2}{k}\mathbf{1})$. This guarantees that $\widehat{\mathcal{D}}_k = \mathcal{L}(x; S_1^k) - \frac{1}{2}\mathcal{L}(x; S_1^{k/2}) - \frac{1}{2}\mathcal{L}(x; S_{k/2+1}^k) \geq 0$.

Let us now upper bound $\widehat{\mathcal{D}}_k$. To that end, we define $\tilde{q}' = 2q_1^{k/2} + \delta$ where $\delta \in \mathbb{R}^{k/2}$ is a fixed-sign vector such that $\tilde{q}'$ lies in $\Delta^{k/2}$. More precisely, if $\tilde{q}'^\top \mathbf{1} > 1$, δ decreases the mass of the largest coordinate until $\tilde{q}'_{(1)} = \frac{2}{k}$ and iterates along the sorted coordinates until $\tilde{q}' \in \Delta^{k/2}$. If $\tilde{q}'^\top \mathbf{1} < 1$, δ similarly increases the smallest coordinate to $\frac{2}{k}$ until $\tilde{q}' \in \Delta^{k/2}$. Without loss of generality, we can assume that ψ attains its minimum at $t = 1$ (otherwise may replaced it by $\psi(t) - \psi'(1)(t-1)$ without changing the objective). Therefore, since $\tilde{q}'$ is closer to $\frac{2}{k}\mathbf{1}$ than $2q_1^{k/2}$, it holds that

$$\mathrm{D}_\psi(\tilde{q}', \tfrac{2}{k}\mathbf{1}) = \frac{2}{k}\sum_{i \leq k/2} \psi(\tfrac{k\tilde{q}'_i}{2}) \leq \frac{2}{k}\sum_{i \leq k/2} \psi(kq_i)$$

Finally, we know that q' is optimal for $S_1^{k/2}$ and so

$$\begin{aligned}\mathcal{L}(x; S_1^{k/2}) &\geq \sum_{i \leq k/2} \tilde{q}'_i \ell(x; S_i) - \mathrm{D}_\psi(\tilde{q}', \tfrac{2}{k}\mathbf{1}) \\ &\geq 2\sum_{i \leq k/2} q_i \ell(x; S_i) - \frac{2}{k}\sum_{i \leq k/2} \psi(kq_i) - \sum_{i \leq k/2} [-\delta_i]_+ B \\ &= 2\sum_{i \leq k/2} q_i \ell(x; S_i) - \frac{2}{k}\sum_{i \leq k/2} \psi(kq_i) - 2\left[\sum_{i \leq k/2} q_i - \frac{1}{2}\right]_+ B.\end{aligned}$$

The same argument for the indices $\{k/2+1, \ldots, k\}$ yields

$$\widehat{\mathcal{D}}_k \leq 2B\left\{\left[\sum_{i=1}^{k/2} q_i - \frac{1}{2}\right]_+ + \left[\sum_{i=k/2+1}^{k} q_i - \frac{1}{2}\right]_+\right\} = 2B\left|\sum_{i=1}^{k/2} q_i - \frac{1}{2}\right|.$$

Therefore, we have

$$\mathbb{E}(\widehat{\mathcal{D}}_k)^2 \leq 4B^2\,\mathbb{E}\left[\sum_{i \leq k/2} q_i - \frac{1}{2}\right]^2 \overset{(i)}{\leq} \frac{2CB^2}{k},$$

where (i) follows from Lemma 3 and the C-χ^2-boundedness of $\mathcal{L}$. Substituting into (52) yields the desired bound on $\widehat{\mathcal{M}}[\mathcal{L}]$. □

Having established the gradient estimator upper bounds for pure-penalty objectives, we demonstrate that similar bounds *do not* extend to the case of χ^2 constraint.

Proposition 7 (Lower bound in the case of constrained-χ^2). *For every $\rho \geq 1$, n_0 and $n \geq 4$, there exists a distribution P_0 over $\mathbb{S} = \{0, 1, 2\}$ and a G-Lipschitz loss $\ell : [-1, 1] \times \mathbb{S} \to \mathbb{R}_+$ such that the multi-level Monte Carlo gradient estimator with parameters n_0 and n satisfies*

$$\mathbb{E}\left\|\widehat{\mathcal{M}}[\nabla\mathcal{L}_{\chi^2}]\right\|^2 \gtrsim \frac{n}{n_0}G^2.$$

Proof. We reuse the construction and notation in the proof of Proposition 5 and so do not repeat it. For a sample S_1^n, we consider the event $\mathfrak{E}_a(S_1^{n/2})$ where $S_i \neq 2$ for all $i \leq n/2$ and there are at least $np_1/2$ samples with value 1. We argue in the proof of Proposition 5 (Eq. (46)) that under this event we have

$$\mathbb{E}[\nabla\mathcal{L}_{\chi^2}(0; S_1^{n/2}) \mid \mathfrak{E}_a(S_1^{n/2})] = G.$$

Moreover, we have

$$\nabla\mathcal{L}_{\chi^2}(0; S_{n/2+1}^n) \geq -G$$

with probability 1, so overall

$$\mathbb{E}\left[\frac{1}{2}\nabla\mathcal{L}_{\chi^2}(0; S_1^{n/2}) + \frac{1}{2}\nabla\mathcal{L}_{\chi^2}(0; S_{n/2+1}^n) \;\middle|\; \mathfrak{E}_a(S_1^{n/2})\right] \geq 0.$$

We also consider the event $\mathfrak{E}_b(S_1^n)$ that there is exactly one sample with value 2 and less the np_1 samples with value 1. As per the proof of Proposition 5 (Eq. (47)) we have

$$\mathbb{E}[\nabla\mathcal{L}_{\chi^2}(0; S_1^n) \mid \mathfrak{E}_b(S_1^n)] \leq -\frac{1}{6}G,$$

where we used $\rho \geq 1$.

Moreover, by the arguments in the proof of Proposition 5, we have

$$\mathbb{P}\Big(\mathfrak{E}_a(S_1^{n/2}) \cap \mathfrak{E}_b(S_1^n)\Big) \gtrsim 1.$$

Therefore, since $\widehat{\mathcal{D}}_n = \nabla\mathcal{L}_{\chi^2}(0; S_1^n) - \frac{1}{2}\nabla\mathcal{L}_{\chi^2}(0; S_1^{n/2}) - \frac{1}{2}\nabla\mathcal{L}_{\chi^2}(0; S_1^{n/2})$, we have

$$\begin{aligned}
\mathbb{E}\|\widehat{\mathcal{D}}_n\|^2 &\geq \mathbb{E}\Big[\|\widehat{\mathcal{D}}_n\|^2 \mid \mathfrak{E}_a(S_1^{n/2}) \cap \mathfrak{E}_b(S_1^n)\Big]\mathbb{P}\Big(\mathfrak{E}_a(S_1^{n/2}) \cap \mathfrak{E}_b(S_1^n)\Big) \\
&\geq \frac{G^2}{36}\mathbb{P}\Big(\mathfrak{E}_a(S_1^{n/2}) \cap \mathfrak{E}_b(S_1^n)\Big) \gtrsim G^2.
\end{aligned}$$

The proof is complete by noting that

$$\mathbb{E}\Big|\widehat{\mathcal{M}}[\nabla\mathcal{L}_{\chi^2}(0; \cdot)]\Big|^2 \geq 2^{j_{\max}-1}\,\mathbb{E}\|\widehat{\mathcal{D}}_n\|^2 = \frac{n}{2n_0}\,\mathbb{E}\|\widehat{\mathcal{D}}_n\|^2 \gtrsim \frac{n}{n_0}G^2.$$

□

Since the number T of SGM iterations must be proportional to the second moment of the gradient estimator, Proposition 7 tells us that in the worst case we might have to set $T \asymp n(GR)^2/\epsilon^2$, in which case we might as well use a mini-batch estimator with batch size n and run $(GR)^2/\epsilon^2$ SGM steps.

C.2 Proof of complexity bounds

Theorem 2 (MLMC complexity guarantees)**.** *For $\epsilon \in (0, B)$, set $n \asymp \frac{B^2}{\alpha\epsilon^2}$, $1 \lesssim n_0 \lesssim \frac{\log n}{\alpha}$ and $T \asymp \frac{(GR)^2}{n_0\alpha\epsilon^2}\log^2 n$. The stochastic gradient iterates (9) with $\tilde{g}(x) = \widehat{\mathcal{M}}[\nabla\mathcal{L}_{\text{CVaR}}(x; \cdot)]$ satisfy $\mathbb{E}[\mathcal{L}_{\text{CVaR}}(\bar{x}_T; P_0)] - \inf_{x\in\mathcal{X}}\mathcal{L}_{\text{CVaR}}(x; P_0) \leq \epsilon$ with complexity at most*

$$n_0\log_2\left(\frac{n}{n_0}\right)T + 5\sqrt{(n\log n)^2 + n_0 n T\log n} \lesssim \frac{(GR+B)^2}{\alpha\epsilon^2}\log^2\frac{B^2}{\alpha\epsilon^2} \;\; \textit{w.p} \;\; \geq 1 - \frac{1}{n}.$$

The same conclusion holds when replacing $\mathcal{L}_{\text{CVaR}}$ with $\mathcal{L}_{\chi^2\text{-pen}}$ and α^{-1} with $1 + B/\lambda$.

Proof. The convergence guarantee of Proposition 6 and the second moment bound of Proposition 3 directly give that iterates of the form (9) with the MLMC gradient estimator guarantees a regret smaller than ϵ for $n \asymp \frac{B^2}{\alpha\epsilon^2}$, $1 \lesssim n_0 \lesssim \frac{\log n}{\alpha}$ and $T \asymp \frac{(GR)^2}{n_0\alpha\epsilon^2}\log^2 n$. However, since the multilevel estimator randomizes the batch size, it remains to show that the number of samples concentrates below the claimed bound. Let $K_t = n_0 2^{J_t}$ be the batch size at time t, and note that

$$\begin{aligned}
&\mathbb{E}\,K_1 = n_0\log_2\frac{2n}{n_0}, \\
&\mathbb{E}\,K_1^2 = 3n_0 n - 2n_0^2 \leq 3n_0 n, \;\text{ and} \\
&\;\;K_1 \leq n \;\text{ with probability } 1.
\end{aligned}$$

Therefore, since K_1^T are iid, a one-sided Bernstein bound [73, Prop. 2.14] implies that

$$\mathbb{P}\left[\sum_{t\le T} K_t \ge n_0 \log_2(2n/n_0)T + \delta\right] \le \exp\left(-\frac{\delta^2}{6n_0 nT + \frac{n\delta}{3}}\right).$$

Solving in δ for the RHS to be equal to $\frac{1}{n}$ yields $\delta = \frac{n\log n}{3}(1 + \sqrt{1 + 216\frac{Tn_0}{n\log n}})$. We replace n, n_0 and T by their values and conclude the proof. □

C.3 Comparison with Blanchet and Glynn [8]

There are two differences between our MLMC estimator and the proposal of Blanchet and Glynn [8]. First, we take J to be a truncated $\mathsf{Geo}(1/2)$ random variable while they suggest $J \sim \mathsf{Geo}(2^{-3/2})$ without truncation—as we further discuss below, this modification is crucial for ensuring a useful second moment bound in our setting. The second difference is that we allow for a minimum sample size $n_0 > 1$ as opposed to $n_0 = 1$ in [8]. This modification is somewhat less important, as $n_0 = 1$ suffices for optimal gradient complexity, but choosing slightly larger n_0 is helpful in practice and can provably reduce the sequential depth of SGM by logarithmic factors.

Let us discuss in more detail the choice $p = 1/2$ in our construction of $J \sim \min\{\mathsf{Geo}(p), j_{\max}\}$. Inspection of Claim 2 shows that $p < 1/2$ implies that the expected sample cost is $\mathbb{E}\, 2^J n_0 \le \frac{n_0}{1-2p}$ independent of $n = 2^{j_{\max}} n_0$, so in principle we could compute unbiased estimates even for $\mathbb{E}\,\mathsf{F}(S_1^\infty)$, i.e., the population objective. However, any $p < 1/2$ would result in overly large second moments: substituting $q(j) \propto p^{-j}$ and $\mathbb{E}\|\widehat{\mathcal{D}}_k\|^2 \asymp 1/k$ in (52) would result in bounds scaling with $(n/n_0)^{\log_2 1/(2p)}$. Therefore, $p = 1/2$ is the only value for which both the second moment and expected number of samples are sub-polynomial in n. In contrast, Blanchet and Glynn [8] apply the MLMC estimator to more regular functionals for which $\|\widehat{\mathcal{D}}_k\|^2 \lesssim 1/k^2$, and consequently can use a smaller value for p.

D Lower bound proofs

This section proves our lower bounds, which we restate for ease of reference.

Theorem 3 (Minimax lower bounds)**.** *Let $G, R, \alpha, \lambda > 0$, $\epsilon \in (0, GR/64)$, and sample space $\mathbb{S} = [-1, 1]$. There exists a numerical constant $c > 0$ such that the following holds.*

- *For each $d \ge 1$, domain $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| \le R\}$, and any algorithm, there exists a distribution P_0 on $\mathbb{S}$ and convex G-Lipschitz loss $\ell : \mathcal{X} \times \mathbb{S} \to [0, GR]$ such that*

$$T \le c\frac{(GR)^2}{\alpha\epsilon^2} \;\; \textit{implies} \;\; \mathbb{E}[\mathcal{L}_{\text{CVaR}}(x_T; P_0)] - \inf_{x'\in\mathcal{X}} \mathcal{L}_{\text{CVaR}}(x'; P_0) > \epsilon.$$

- *There exists $d_\epsilon \lesssim (GR)^2\epsilon^{-2}\log\frac{GR}{\epsilon}$ such that for $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| \le R\}$, the same conclusion holds when replacing $\mathcal{L}_{\text{CVaR}}$ with $\mathcal{L}_{\chi^2\text{-pen}}$ and α with $\lambda/(GR)$.*

Since our proofs for CVaR and χ^2 penalty are quite different, we present them separately in Theorems 3a and 3b, respectively.

D.1 CVaR lower bound

To prove the CVaR lower bound we use the following standard Le Cam reduction from stochastic optimization to hypothesis testing.

Lemma 4. *[21, Chapter 5] Let $\mathcal{P}$ be a set of distributions and $P_{-1}, P_1 \in \mathcal{P}$ and define*

$$\mathsf{d}_{\text{opt}}(P_1, P_{-1}) := \sup\{\delta' \ge 0 \mid \textit{no } x \in \mathcal{X} \textit{ is } \delta'\textit{-optimal for both } \mathcal{L}(\cdot; P_{-1}) \textit{ and } \mathcal{L}(\cdot; P_1)\}.$$

Then for any measurable mapping $\hat{x}_n : \mathbb{S}^n \to \mathcal{X}$ we have

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{S_1^n\sim P^n}\{\mathcal{L}(\hat{x}_n(S_1^n); P)\} - \inf_{x'\in\mathcal{X}} \mathcal{L}(x'; P) \ge \frac{\mathsf{d}_{\text{opt}}(P_1, P_{-1})}{2}\left(1 - \sqrt{\frac{n}{2}\mathrm{D}_{\text{kl}}(P_{-1}, P_1)}\right),$$

Armed with Lemma 4, we state and prove the lower bound for CVaR.

Theorem 3a (CVaR lower bound)**.** *Let $G, R, \alpha > 0$, $\epsilon \in (0, GR/64)$, $\mathbb{S} = [-G, G]$, $\mathcal{X} = [-R, R]$, and $\ell(x, s) = x \cdot s$. For any (potentially randomized) mapping $\hat{x}_n : \mathbb{S}^n \to \mathcal{X}$ there exists a distribution P_0 over $\mathbb{S}$ such that,*

$$n \le \frac{(GR)^2}{2048\alpha\epsilon^2} \text{ implies } \mathbb{E}\,\mathcal{L}_{\text{CVaR}}(\hat{x}_n(S_1^n); P_0) - \inf_{x \in \mathcal{X}} \mathcal{L}_{\text{CVaR}}(x; P_0) \ge \epsilon.$$

Proof. Let us first assume that $\alpha \le \frac{1}{2}$. For $\delta \le \min\{\alpha, 1 - 2\alpha\}$, $\mu > 0$ and $v \in \{-1, 1\}$ we consider the distributions P_v such that for $S_v \sim P_v$ we have

$$S_v = G \cdot \begin{cases} \mu & \text{with probability } \alpha + \delta v \\ -1 & \text{with probability } 1 - \alpha - \delta v \end{cases}. \tag{53}$$

For $x \in [-R, R]$, we let $\ell(x; s) = x \cdot s$. Since the CVaR objective is positively homogeneous, we have

$$\mathcal{L}_{\text{CVaR}}(x; P_v) = |x| \cdot \mathcal{L}_{\text{CVaR}}(\text{sign}(x); P_v).$$

It therefore suffices to compute $\mathcal{L}_{\text{CVaR}}(\pm 1; P_{\pm 1})$. A quick calculation yields

$$\mathcal{L}_{\text{CVaR}}(1; S_1) = G\mu, \;\; \mathcal{L}_{\text{CVaR}}(-1, S_1) = G,$$
$$\mathcal{L}_{\text{CVaR}}(1; S_{-1}) = G\mu\left(1 - \frac{\delta}{\alpha}\right) - G\frac{\delta}{\alpha} \text{ and } \mathcal{L}_{\text{CVaR}}(-1; S_{-1}) = G.$$

We thus have a closed-form expression for the CVaR objective: for P_1 we have

$$\mathcal{L}_{\text{CVaR}}(x; P_1) = -Gx1_{\{x \le 0\}} + Gx\mu 1_{\{x \ge 0\}},$$

which clearly attains its minimum at $x = 0$ where it has value 0. Choosing μ such that

$$\mu = \frac{\delta}{2\alpha}\left(1 - \frac{\delta}{2\alpha}\right)^{-1}$$

gives $\mathcal{L}_{\text{CVaR}}(1; S_{-1}) = -G\mu$ and

$$\mathcal{L}_{\text{CVaR}}(x; P_{-1}) = -Gx1_{\{x \le 0\}} - Gx\mu 1_{\{x \ge 0\}},$$

which attains its minimum at $x = R$ where it has value $-GR\mu$. We therefore have that

$$\mathsf{d}_{\text{opt}}(P_1, P_{-1}) = \frac{GR\mu}{2} \ge \frac{GR\delta}{4\alpha}. \tag{54}$$

Moreover, we have $t \log t - t + 1 \le (t-1)^2$ for all $t \ge 0$, so that $\mathrm{D}_{\text{kl}}(Q, P) \le 2\mathrm{D}_{\chi^2}(Q, P)$ for all Q, P, and in particular

$$\mathrm{D}_{\text{kl}}(P_{-1}, P_1) \le 2\mathrm{D}_{\chi^2}(P_{-1}, P_1) = \frac{4\delta^2}{(1 - \alpha - \delta)(\alpha + \delta)} \le \frac{8\delta^2}{\alpha}, \tag{55}$$

where that last transition used $\delta \le \alpha$ and $\alpha \le 1/2$.

We take

$$\delta = \sqrt{\frac{\alpha}{16(n + \alpha^{-1})}},$$

where so that $\mathrm{D}_{\text{kl}}(P_{-1}, P_1) \le 1/(2n)$ and Lemma 4 combined with (54) and (55) gives

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{S_1^n \sim P^n}\{\mathcal{L}(\hat{x}_n(S_1^n); P)\} - \inf_{x' \in \mathcal{X}} \mathcal{L}(x'; P) \ge \frac{GR}{32\sqrt{\alpha n + 1}},$$

and the result follows from substituting $n \le \frac{(GR)^2}{2048\alpha\epsilon^2}$. When $\alpha \ge 1/2$ the result follows from the standard lower bound for stochastic convex optimization (e.g. [21, Thm. 5.2.10]). □

D.2 Penalized-χ^2 lower bound

Computation of $\mathcal{L}_{\chi^2\text{-pen}}(x; P_\pm)$ for the CVaR lower bound construction (53) shows that the argument does not easily transfer to the penalized-χ^2 objective because—as opposed to constrained-χ^2 and CVaR—it is not positive homogeneous in x.

Sidestepping this difficulty, we prove our lower bound using the different machinery of high-dimensional hard instances for oracle-based optimization [48]. We consider two standard oracles. First is the deterministic first-order oracle, that for a function $f : \mathbb{R}^d \to \mathbb{R}$ and a query x returns

$$\mathsf{O}_f^{\mathrm{D}}(x) \coloneqq (f(x), \nabla f(x)),$$

where we recall that $\nabla f(x)$ is an arbitrary element of $\partial f(x)$. Second is the stochastic oracle, that for a loss function $\ell : \mathcal{X} \times \mathbb{S} \to \mathbb{R}$ and distribution P_0 returns the randomized mapping

$$\mathsf{O}_{\ell,P_0}^{\mathrm{S}}(x) \coloneqq (\ell(x; S), \nabla \ell(x; S)), \text{ for } S \sim P_0.$$

We construct the hard instance for $\mathcal{L}_{\chi^2\text{-pen}}$ based on the standard hard instance for non-stochastic convex optimization, whose properties are as follows.

Proposition 8 (Braun et al. [10], Theorem V.1)**.** *Let $\epsilon, G, R > 0$. There exists $d_\epsilon \lesssim (GR)^2\epsilon^{-2} \log \frac{GR}{\epsilon}$ such that the following holds for $\mathcal{X} = \{x \in \mathbb{R}^{d_\epsilon} \mid \|x\| \le R\}$. For any (possibly randomized) algorithm there exists $f_\epsilon : \mathcal{X} \to [0, GR]$ convex and G-Lipschitz such the query x_T $\mathsf{O}_{f_\epsilon}^{\mathrm{D}}$ at iteration T satisfies*

$$T \le c\frac{(GR)^2}{\epsilon^2} \text{ implies } \mathbb{E} f_\epsilon(x_T) - \inf_{\|x'\| \le R} f_\epsilon(x') \ge \epsilon,$$

for a numerical constant $c > 0$.

In other words, any "dimension-free" algorithm needs to interact $\Omega(\epsilon^{-2})$ times with the deterministic oracle to obtain an ϵ-suboptimal point. With this result, we prove our lower bound for optimizing $\mathcal{L}_{\chi^2\text{-pen}}$.

Theorem 3b (Penalized-χ^2 lower bound)**.** *Let $G, R, \lambda > 0$ and $\epsilon \in (0, GR)$. There exists $d_\epsilon \lesssim (GR)^2\epsilon^{-2} \log \frac{GR}{\epsilon}$ such that the following holds for $\mathcal{X} = \{x \in \mathbb{R}^{d_\epsilon} \mid \|x\| \le R\}$ and $\mathbb{S} \subseteq [-1, 1]$. For every algorithm there exists a distribution P_0 over $\mathbb{S}$ and $\ell : \mathcal{X} \times \mathbb{S} \to [-GR, GR]$ convex and G-Lipschitz in x, such that the query x_T to $\mathsf{O}_{\ell,P_0}^{\mathrm{S}}$ at iteration T satisfies*

$$T \le c\frac{(GR)^3}{\lambda\epsilon^2} \text{ implies } \mathbb{E}[\mathcal{L}_{\chi^2\text{-pen}}(x_T; P_0)] - \min_{x' \in \mathcal{X}} \mathcal{L}_{\chi^2\text{-pen}}(x'; P_0) > \epsilon,$$

for $c > 0$ independent of G, R, λ and ϵ.

Proof. Consider any convex and G Lipschitz $f : \mathcal{X} \to [0, GR]$, define $\mathbb{S} \coloneqq \{0, 1\}$ and $P_0 = \mathsf{Bernoulli}(\frac{\lambda}{GR})$, and construct the following loss

$$\ell(x; S) \coloneqq \begin{cases} f(x) & \text{if } S = 1 \\ -GR & \text{if } S = 0. \end{cases}$$

(If $\lambda > GR$ the result follows from the standard $(GR)^2/\epsilon^2$ lower bound for convex optimization). Expressing the resulting objective $\mathcal{L}_{\chi^2\text{-pen}}$ with the dual form (23) gives

$$\begin{aligned} \mathcal{L}_{\chi^2\text{-pen}}(x; P_0) &= \inf_{\eta \in \mathbb{R}} \left\{ \frac{\lambda}{2} + \eta + \frac{1}{2\lambda} \left[\frac{\lambda}{GR}(f_\epsilon(x) - \eta)_+^2 + \left(1 - \frac{\lambda}{GR}\right)(-GR - \eta)_+^2 \right] \right\} \\ &= f(x) - \frac{GR - \lambda}{2}, \end{aligned}$$

since $\eta^\star = f(x) - GR \ge -GR$. We get that minimizing $\mathcal{L}_{\chi^2\text{-pen}}$ is equivalent to optimizing f.

Fix an algorithm interacting with $\mathsf{O}_{\ell,P_0}^{\mathrm{S}}$ and note that it implies a (randomized) algorithm interacting with $\mathsf{O}_f^{\mathrm{D}}$. Therefore we may take $f = f_\epsilon$, the hard function for this algorithm that Proposition 8 guarantees. Note that an algorithm interacting with $\mathsf{O}_{\ell,P_0}^{\mathrm{S}}$ receives information on f_ϵ only when

$S = 1$. Therefore, the worst-case expected optimality gap when minimizing $\mathcal{L}_{\chi^2\text{-pen}}$ with T queries to $\mathsf{O}^{\mathrm{S}}_{\ell,P_0}$ is identical to the worst-case expected optimality gap when minimizing f_ϵ with $\mathsf{Bin}(T, \frac{\lambda}{GR})$ queries. Therefore, Proposition 8 tells us that for some $c' > 0$,

$$\mathbb{E}[\mathcal{L}(x_T; P_0) - \inf_{\|x'\| \le R} \mathcal{L}(x'; P_0)] \ge \epsilon \cdot \mathbb{P}\bigg(\mathsf{Bin}\Big(T, \frac{\lambda}{GR}\Big) \le c' \cdot \frac{(GR)^2}{\epsilon^2}\bigg).$$

Substituting $T \le \frac{c}{4} \cdot \frac{(GR)^3}{\lambda\epsilon^2}$ gives that $\mathbb{P}(\mathsf{Bin}(T, \frac{\lambda}{GR}) \le c' \cdot \frac{(GR)^2}{\epsilon^2}) \ge \frac{1}{2}$ by a standard Chernoff bound. The result follows by properly adjusting the constant factors (e.g., replacing ϵ with 2ϵ). □

E A doubling scheme for minimizing $\mathcal{L}_{\chi^2}$

In this section, we obtain a stronger guarantee for minimizing the χ^2-constraint robust objective. Namely, we leverage duality relationships to approximate the constrained objective $\mathcal{L}_{\chi^2}$ via its penalized counterpart, $\mathcal{L}_{\chi^2\text{-pen}}$. We adjust notation to make the dependence of $\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}$ on λ explicit.

Our starting point is the recognition that, by duality (cf. [66, Sec. 3.2]),

$$\mathcal{L}_{\chi^2}(x; P_0) = \inf_{\lambda \ge 0} \Big\{\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x; P_0) + \lambda\rho\Big\} = \inf_{\lambda \ge 0} \sup_{Q \ll P_0} \big\{\mathbb{E}_Q\, \ell(x; S) - \lambda\left[\mathrm{D}_{\chi^2}(Q, P_0) - \rho\right]\big\}$$

for any distribution P_0. For $0 \le \underline{\lambda} \le \overline{\lambda}$, we may thus consider the approximation

$$\mathcal{L}_{\chi^2[\underline{\lambda},\overline{\lambda}]}(x; P_0) \coloneqq \min_{\lambda \in [\underline{\lambda},\overline{\lambda}]} f_\rho(x, \lambda) \;\text{ where }\; f_\rho(x, \lambda) \coloneqq \mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x; P_0) + \lambda\rho.$$

By restricting λ to an appropriate range, we can then approximate $\mathcal{L}_{\chi^2}$ by its truncated version, as the next lemma shows.

Lemma 5. *For all P_0, ρ and ϵ,*

$$\min_{x \in \mathcal{X}} \mathcal{L}_{\chi^2[\frac{\epsilon}{2\rho}, \frac{B}{\rho}]}(x; P_0) \le \min_{x' \in \mathcal{X}} \mathcal{L}_{\chi^2}(x'; P_0) + \frac{\epsilon}{2}.$$

Proof. Le $x^\star, \lambda^\star = \arg\min_{x \in \mathcal{X}, \lambda \ge 0}\{f_\rho(x, \lambda)\}$, noting that $\min_{x' \in \mathcal{X}} \mathcal{L}_{\chi^2}(x'; P_0) = f_\rho(x^\star, \lambda^\star)$. For any x, λ let $Q^\star_{x,\lambda}$ be the maximizing Q in (12) for these values of x, λ. Moreover, let $\mathrm{D}(x, \lambda) = \mathrm{D}_{\chi^2}(Q^\star_{x,\lambda}, P_0)$. By Claim 4, for all $\lambda > B/\rho$ we have that $\mathrm{D}(x, \lambda) < \rho$, and consequently $\lambda > \lambda^\star$, i.e., $\lambda^\star \le B/\rho$, and hence that upper bound has no impact on accuracy.

When in addition we have $\lambda^\star \ge \epsilon/(2\rho)$ then clearly $\min_{x \in \mathcal{X}} \mathcal{L}_{\chi^2[\frac{\epsilon}{2\rho}, \frac{B}{\rho}]}(x; P_0) = \mathcal{L}_{\chi^2[0,\infty]}(x; P_0) = \min_{x' \in \mathcal{X}} \mathcal{L}_{\chi^2}(x'; P_0)$. Otherwise, if $\lambda^\star < \epsilon/(2\rho) =: \lambda_\epsilon$ we may write

$$\begin{aligned}
\min_{x \in \mathcal{X}} \mathcal{L}_{\chi^2[\frac{\epsilon}{2\rho}, \frac{B}{\rho}]}(x; P_0) &\overset{(i)}{\le} f_\rho(x^\star, \lambda_\epsilon) \overset{(ii)}{\le} f_\rho(x^\star, \lambda^\star) + \left[\tfrac{\partial}{\partial\lambda} f_\rho(x^\star, \lambda_\epsilon)\right](\lambda_\epsilon - \lambda^\star) \\
&= \min_{x' \in \mathcal{X}} \mathcal{L}_{\chi^2}(x'; P_0) + [\rho - \mathrm{D}(x^\star, \lambda_\epsilon)](\lambda_\epsilon - \lambda^\star) \\
&\overset{(iii)}{\le} \min_{x' \in \mathcal{X}} \mathcal{L}_{\chi^2}(x'; P_0) + \lambda_\epsilon\rho = \min_{x' \in \mathcal{X}} \mathcal{L}_{\chi^2}(x'; P_0) + \frac{\epsilon}{2}.
\end{aligned}$$

Where we used (i) that $x^\star$ and λ_ϵ are feasible points in the joint minimization of $f_\rho(x, \lambda)$ over $x \in \mathcal{X}$ and $\lambda \in [\lambda_\epsilon, B/\rho]$; (ii) the convexity of f in λ; and (iii) the fact that $\mathrm{D}(x^\star, \lambda_\epsilon) \ge 0$ and $\lambda^\star \le \lambda_\epsilon$. □

Our strategy is therefore to jointly minimize $f_\rho(x, \lambda) = \mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x; P_0) + \lambda\rho$ over both $x \in \mathcal{X}$ and $\lambda \in [\underline{\lambda}, \overline{\lambda}]$ (rather than $[0, \infty]$), using the approximation guarantee in Lemma 5 to argue that the restriction of λ will have limited effect on the quality of the resulting solution. We iterate the projected stochastic gradient method with the multi-level Monte Carlo (MLMC) gradient estimator (11) via

$$x_{t+1} = \Pi_{\mathcal{X}}\Big(x_t - \gamma_x \widehat{\mathcal{M}}\big[\nabla \mathcal{L}^{\lambda_t}_{\chi^2\text{-pen}}(x_t)\big]\Big),\; \lambda_{t+1} = \Pi_{[\underline{\lambda},\overline{\lambda}]}\Big(\lambda_t - \gamma_\lambda \widehat{\mathcal{M}}\big[\tfrac{\partial}{\partial\lambda} \mathcal{L}^{\lambda_t}_{\chi^2\text{-pen}}(x_t) + \rho\big]\Big). \tag{56}$$

If we can bound the moments of the MLMC-approximated gradients $\widehat{\mathcal{M}}$, we can then leverage standard stochastic gradient analyses to prove convergence. We use the following bound.

Lemma 6. *We have*

$$\mathbb{E}\Big(\widehat{\mathcal{M}}\big[\tfrac{\partial}{\partial\lambda}\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x;\cdot)+\rho\big]\Big)^2 \lesssim \frac{B^2}{\lambda^2}\left(1+\frac{B\log\frac{n}{n_0}}{\lambda n_0}\right)+\rho^2.$$

Proof. Recall the definition (51) of the MLMC estimator of a general F and the expression (52) for its second moment. Suppose that $\mathsf{F}(\cdot)=\mathsf{F}_1(\cdot)+\mathsf{F}_2(\cdot)+c$, where c is a constant. Then

$$\mathbb{E}\|\widehat{\mathcal{D}}_k\big[\mathsf{F}\big]\|^2 = \mathbb{E}\|\widehat{\mathcal{D}}_k\big[\mathsf{F}_1+\mathsf{F}_2\big]\|^2 \le 2\,\mathbb{E}\|\widehat{\mathcal{D}}_k\big[\mathsf{F}_1\big]\|^2+2\,\mathbb{E}\|\widehat{\mathcal{D}}_k\big[\mathsf{F}_2\big]\|^2.$$

Consequently, by (52), we have

$$\mathbb{E}\|\widehat{\mathcal{M}}\big[\mathsf{F}\big]\|^2 \le 2c^2+2\,\mathbb{E}\|\widehat{\mathcal{M}}\big[\mathsf{F}_1\big]\|^2+2\,\mathbb{E}\|\widehat{\mathcal{M}}\big[\mathsf{F}_2\big]\|^2. \tag{57}$$

We apply this observation to $\widehat{\mathcal{M}}\big[\frac{\partial}{\partial\lambda}\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x;\cdot)+\rho\big]$ by noting that

$$\tfrac{\partial}{\partial\lambda}\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x;S_1^n) = -\mathrm{D}_{\chi^2}(q^\star;\tfrac{1}{n}\mathbf{1}) = \frac{1}{\lambda}\left(\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x;S_1^n)-\frac{1}{n}\sum_{i\le n}q_i^\star\ell(x;S_i)\right).$$

Proposition 3' gives us the bound $\mathbb{E}\Big(\widehat{\mathcal{M}}\big[\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x;S_1^n)\big]\Big)^2 \lesssim B^2\Big(1+\frac{B}{\lambda n_0}\log\frac{n}{n_0}\Big)$. Moreover, we have that

$$\mathbb{E}\left(\widehat{\mathcal{M}}\Big[\frac{1}{n}\sum_{i\le n}q_i^\star\ell(x;S_i)\Big]\right)^2 \lesssim B^2\left(1+\frac{B}{\lambda n_0}\log\frac{n}{n_0}\right)$$

By exactly the same argument that proves the gradient second moment bound in Proposition 3'. The result then follows by substituting into (57). □

Therefore, we may find an ϵ approximate minimizer with complexity roughly $B^3\overline{\lambda}^2/(\underline{\lambda}^3\epsilon^2)$:

Lemma 7. *Fix $\epsilon\in(0,B)$ and $\overline{\lambda}\ge\underline{\lambda}>0$. For a suitable setting of the parameters n_0,n,T,γ_x and γ_λ, the average $\bar{x}_T=\sum_{t\le T}x_t$ of the iterates (56) satisfies $\mathbb{E}\,\mathcal{L}_{\chi^2[\underline{\lambda},\overline{\lambda}]}(\bar{x}_T;P_0)\le \min_{x\in\mathcal{X}}\mathcal{L}_{\chi^2[\underline{\lambda},\overline{\lambda}]}(x;P_0)+\epsilon$, with complexity*

$$\lesssim \left(1+\frac{B}{\underline{\lambda}}\right)\frac{(GR)^2+B^2\overline{\lambda}^2/\underline{\lambda}^2+\overline{\lambda}^2\rho^2}{\epsilon^2}\log^2\left(1+\frac{B}{\underline{\lambda}\epsilon}\right)\ \textit{with probability}\ \ge 1-\frac{\epsilon^2}{B^2}.$$

Proof. We take $n\asymp\frac{B}{\underline{\lambda}\epsilon}$ to guarantee bias below $\epsilon/2$ by Proposition 1, and we take $n_0\asymp\frac{B}{\underline{\lambda}}\log n$ to guarantee that

$$\Gamma_x^2 \coloneqq \sup_{x\in\mathcal{X},\lambda\in[\underline{\lambda},\overline{\lambda}]}\mathbb{E}\left\|\widehat{\mathcal{M}}\big[\nabla\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x;\cdot)\big]\right\|^2 \lesssim G^2$$

and, by Lemma 6,

$$\Gamma_\lambda^2 \coloneqq \sup_{x\in\mathcal{X},\lambda\in[\underline{\lambda},\overline{\lambda}]}\mathbb{E}\Big(\widehat{\mathcal{M}}\big[\tfrac{\partial}{\partial\lambda}\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}(x;\cdot)+\rho\big]\Big)^2 \lesssim \frac{B^2}{\underline{\lambda}^2}+\rho^2.$$

Let $\bar{\lambda}_T=\sum_{t\le T}\lambda_t$ be the average of the λ iterates in (56). By appropriate choice of η and η' we guarantee (via Proposition 6) that

$$\mathbb{E}\,\mathcal{L}_{\chi^2[\underline{\lambda},\overline{\lambda}]}(\bar{x}_T;P_0)\le\mathbb{E}\,f_\rho(\bar{x}_T,\bar{\lambda}_T)\le\min_{x\in\mathcal{X},\lambda\in[\underline{\lambda},\overline{\lambda}]}f_\rho(x,\lambda)+\mathrm{err}_T=\min_{x\in\mathcal{X}}\mathcal{L}_{\chi^2[\underline{\lambda},\overline{\lambda}]}(x;P_0)+\mathrm{err}_T,$$

where

$$\mathrm{err}_T\lesssim\frac{\epsilon}{2}+\frac{\Gamma_xR+\Gamma_\lambda(\overline{\lambda}-\underline{\lambda})}{\sqrt{T}}.$$

Therefore, by taking

$$T\asymp\frac{\Gamma_x^2R^2+\Gamma_\lambda^2\overline{\lambda}^2}{\epsilon^2}\asymp\frac{(GR)^2+B^2\overline{\lambda}^2/\underline{\lambda}^2+\overline{\lambda}^2\rho^2}{\epsilon^2}$$

we guarantee that $\mathrm{err}_T\le\epsilon$, and the complexity bound follows from substituting n_0,n and T in the high probability upper bound $n_0\log_2\left(\frac{n}{n_0}\right)T+5\sqrt{(n\log n)^2+n_0nT\log n}$ shown in Theorem 2. □

Directly substituting $\underline{\lambda} = \frac{\epsilon}{2\rho}$ and $\overline{\lambda} = \frac{B}{\rho}$ results in a guarantee scaling as ϵ^{-5}, which is worse than the mini-batch rate of ϵ^{-4}. To improve on this, we divide $[\frac{\epsilon}{2\rho}, \frac{B}{\rho}]$ into $K = \log_2 \frac{B}{\epsilon}$ sub-intervals $[\lambda^{(i+1)}, \lambda^{(i)}]$ satisfying $\lambda^{(i+1)}/\lambda^{(i)} = 2$. We then perform the stochastic gradient method (56) on each of these intervals $[\lambda^{(i+1)}, \lambda^{(i)}]$ in turn, yielding estimates $\bar{x}^{(i)}$ that are each $\lesssim \epsilon$-suboptimal for the approximate objective $\mathcal{L}_{\chi^2[\lambda^{(i+1)},\lambda^{(i)}]}$. Using the bounded ratio $\lambda^{(i+1)}/\lambda^{(i)} = 2$, this requires complexity roughly $1/(\lambda^{(i+1)}\epsilon^2) \lesssim \rho/\epsilon^3$, giving the following theorem.

Theorem 4. *Fix $\epsilon \in (0, B)$, and for $i \in \mathbb{N}$ set $\lambda^{(i)} = \frac{B}{\rho} 2^{-i+1}$ and let $\bar{x}^{(i)}$ be an $\epsilon/2$-approximate minimizer of $\mathcal{L}_{\chi^2[\lambda^{(i+1)},\lambda^{(i)}]}$ computed via stochastic gradient iterations according to Lemma 7. Then, for $1 + K = \lceil \log_2 \frac{2B}{\epsilon} \rceil$ and some $i^\star \le K$ we have $\mathbb{E}\, \mathcal{L}_{\chi^2}(\bar{x}^{(i^\star)}; P_0) \le \min_{x\in\mathcal{X}} \mathcal{L}_{\chi^2}(x; P_0) + \epsilon$. Computing $\bar{x}^{(1)}, \ldots, \bar{x}^{(K)}$ requires a total number of $\nabla\ell$ evaluations*

$$\lesssim \frac{(GR)^2(\rho B + \epsilon \log_2 \frac{B}{\epsilon})}{\epsilon^3} \log^2\left(1 + \frac{\rho B}{\epsilon^2}\right) \text{ with probability } \ge 1 - \frac{\epsilon}{B}.$$

Proof. By Lemma 7, finding an ϵ approximate solution in the interval $[\lambda^{(i+1)}, \lambda^{(i)}]$ requires

$$\lesssim \left(1 + \frac{\rho B}{2^{K-i}\epsilon}\right) \frac{(GR)^2 + B^2}{\epsilon^2} \log^2\left(1 + \frac{\rho B}{\epsilon^2}\right)$$

gradient computations, where we have used $\lambda^{(i)}/\lambda^{(i+1)} \le 2$, $\lambda^{(i)} \le \frac{B}{\rho}$, and $\lambda^{(i+1)} \ge \frac{\epsilon}{2\rho} 2^{K-i}$. Summing over i (and applying a union bound) gives the claimed guarantee. Since the minimizer of $f_\rho(x, \lambda)$ over $x \in \mathcal{X}$ and $\lambda \in [\frac{\epsilon}{2\rho}, \frac{B}{\rho}]$ is equivalent is identical to its minimizer in one of the intervals $[\lambda^{(i+1)}, \lambda^{(i)}]$ for $i \le K$, the result follows from Lemma 5. □

The index $i^\star$ is independent of randomness in our procedure, but we do not know it in advance. Instead, we may estimate the minimized objective for each i and select the index with the lowest estimate. Let $\hat{\lambda}^{(i)}$ be the average of the λ iterations of our stochastic gradient method (56) for a particular interval $[\lambda^{(i+1)}, \lambda^{(i)}]$. Our bias and variance bounds on $\mathcal{L}_{\chi^2\text{-pen}}$ (Proposition 1 and Proposition 2' in the appendix) imply the we can estimate[5] $f_\rho(\bar{x}^{(i)}, \hat{\lambda}^{(i)})$ to accuracy $\lesssim \epsilon$ with a sample of size $\asymp B^2/(\lambda^{(i)}\epsilon^2) \asymp 2^{i-K}B^3\rho\epsilon^{-3}$. Taking $i^\star$ to be the index i minimizing this estimate, it is straightforward to argue that $\mathbb{E}\, \mathcal{L}_{\chi^2}(\bar{x}^{(i^\star)}; P_0) - \min_{x\in\mathcal{X}} \mathcal{L}_{\chi^2}(x; P_0) \lesssim \epsilon$. Therefore, the cost of selecting the best i is at most the cost of performing the optimization.

Theorem 4 provides a rigorous guarantee on the complexity of minimizing $\mathcal{L}_{\chi^2}$ with a fixed constraint ρ by optimizing the parameter λ of $\mathcal{L}^{\lambda}_{\chi^2\text{-pen}}$. In practice, we usually have no prior knowledge of ρ, so it will often make sense to directly tune λ according to validation criteria rather than a target ρ. We also note that Duchi and Namkoong [22] prove a lower bound of order $\rho\epsilon^{-2}$, which is smaller than our $\rho\epsilon^{-3}$ rate. Establishing the optimal rate for this problem remains an open question.

F Experiments

In this section we give a detailed description of our experiments. We begin with a description of the problems we study (Section F.1) followed by our hyperparameter settings (Section F.2) and brief remarks about our PyTorch implementation (Section F.3). Then, in Sections F.4 and F.5 we present and discuss our results in detail, including speed-up factors over full-batch optimization, a study of the generalization impacts of the DRO objective, and direct empirical evaluation of the bias $\mathcal{L} - \overline{\mathcal{L}}$ which we bound in Proposition 1.

F.1 Dataset description

Digits. We consider the MNIST handwritten digit recognition dataset with the standard train/test split into with $6 \cdot 10^4$ and 10^4 training and test images, respectively. There are 10 classes corresponding to the ten digits. We augment the training set with $N_{\text{typed}} = 600$ randomly chosen digits from the characters dataset [19], i.e., 1% of the hand-written digits. Our test set includes the MNIST test set as

[5] To obtain an estimate that has error $\lesssim \epsilon$ with high probability, we can use the median of a logarithmic number of iid copies of the batch estimator.

well as a class-balanced sample of 8K typed digits not included in the training data. Creating an 8K image test set requires that we disregard the original test/train split of [19], but is important in order to make estimates of per-class accuracy reliable. To featurize our data, we train a small convolutional Neural Network (two convolutional layers, two fully-connected layers with ReLU activation function) with a standard ERM objective and 10 epochs of SGM on the MNIST training set (with no typed digits). For both handwritten and typed digits, we use the activations of the last layer as the feature vector.

We perform DRO to learn a linear classifier x on our features, taking the loss ℓ to be multi-class logarithmic loss with a quadratic regularization term on x (the weight part only, not the bias), namely, for a data point $s = (z, y)$ with $z \in \mathbb{R}^d, y \in [C]$ (with C the number of classes) and regularization strength $\mu \geq 0$, we use

$$\ell([x, b]; (z, y)) \coloneqq \log\left(\sum_{c=1}^{C} \exp(\langle x_c - x_y, z\rangle + b_c - b_y)\right) + \frac{\mu}{2}\sum_{c=1}^{C} \|x_c\|_2^2,$$

where $x \in \mathbb{R}^{C\times d}, b \in \mathbb{R}^C$ and x_c denotes the c-th row of x. As the generalization metric, we report accuracy and log loss on the worst sub-group of the data—where a sub-group corresponds to a tuple (subpopulation, class), e.g., (typed, 9).

ImageNet. The ImageNet dataset comprises of $1.2 \cdot 10^6$ training images and $5 \cdot 10^4$ test images with 1000 different classes. We featurize the dataset using a pre-trained ResNet-50 [35] (trained on ImageNet itself with an ERM objective). We use those features as the input to a linear classifier, with regularized multi-class logarithmic loss as in the previous experiment. As the robust generalization metric, we report the average loss and accuracy on the 10 classes with highest test loss.

F.2 Hyperparameter tuning

We fix the budget of our algorithms to 300 epochs for Digits and 30 epochs for ImageNet, where an epoch corresponds to N computations of $\nabla\ell$, where N is the training set size. For all (mini)-batch methods we use Nesterov acceleration (50) with constant momentum $\omega = 0.9$; we did not carefully tune this parameter but did observe it performs better than no momentum. For MLMC using no momentum ($\omega = 0$) performs slightly better than momentum 0.9, so we use no momentum in this case. We also perform iterate averaging with the scheme of Shamir and Zhang [65] with parameter 3 (roughly averaging over the last third of the iterates). Our experiments with CVaR use $\nabla\mathcal{L}_{\text{CVaR}}$ rather than $\nabla\mathcal{L}_{\text{kl-CVaR}}$, in contrast to our theory; we leave empirical exploration of entropy smoothing for CVaR to future work.

Stepsizes. We tune our stepsizes with a coarse-to-fine strategy. More precisely, for each stepsize in $\{10^i\}_{-5\leq i\leq 0}$, we perform a single run of the experiment, and pick the best two stepsizes in terms of the final training value. For these two stepsizes, we evaluate $\frac{\eta}{2}, 2\eta$ and select the stepsize that gives the best value of the training loss. For this final stepsize, we repeat the experiments with 5 different seeds (affecting weight initialization and mini batch samples but not the dataset structure) and report the minimum and maximum across seeds at each iteration. We select all the stepsizes in our experiments using this strategy, except for batch size $n = 10$ in ImageNet where we extrapolated the stepsize from other batch sizes. Table 3 summarizes our step size choices—for batch sizes up to 5K we see a clear linear relationship between the batch size and optimal step size.

ℓ_2-regularization and parameters of the robust loss We choose the strength of the regularizer in the set $\{0, 10^{-5}, \ldots, 10^{-1}\}$. For each robust loss, we consider an appropriate grid of either the size of the uncertainty set (α and ρ) or the strength of the penalty (λ). We evaluate each configuration (ℓ_2 regularization and robust loss parameters) with the stepsizes from the coarse grid and pick the configuration that achieves a good trade-off in terms worst-subgroup and average-case generalization. For simplicity, we choose the same regularization strenght for all the robust losses—$\mu = 10^{-3}$ for ImageNet and $\mu = 10^{-2}$ for Digits. For ERM, we choose the two values of ℓ_2 regularization that optimize either worst subgroup loss or worst subgroup accuracy. That is, for ImageNet we tune the ℓ_2 regularization for the best result on either the worst 10 classes loss and worst 10 classes accuracy respectively, and for Digits we choose the values that optimize loss/accuracy on the hardest typed class—for both experiments, this results in $\mu \in \{10^{-4}, 10^{-3}\}$ for ERM.

F.3 PyTorch Integration

Figure 2 illustrates our integration of DRO into PyTorch. Users simply define the robust loss they wish to use (in the example $\mathcal{L}_{\chi^2}$ with $\rho = 1$) and feed the loss for the examples in the batch to the robust

		ImageNet			Digits		
		$\mathcal{L}_{\text{CVaR}}$	$\mathcal{L}_{\chi^2}$	$\mathcal{L}_{\chi^2\text{-pen}}$	$\mathcal{L}_{\text{CVaR}}$	$\mathcal{L}_{\chi^2}$	$\mathcal{L}_{\chi^2\text{-pen}}$
Algorithm		$\alpha = 0.1$	$\rho = 1$	$\lambda = 0.4$	$\alpha = 0.02$	$\rho = 1$	$\lambda = 0.05$
Batch	$n = 10$	$1 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$
	$n = 50$	$5 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$1 \cdot 10^{-4}$
	$n = 500$	$5 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$1 \cdot 10^{-3}$
	$n = 5K$	$5 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$5 \cdot 10^{-3}$	$2 \cdot 10^{-2}$	$1 \cdot 10^{-2}$
	$n = 50K$	$2 \cdot 10^{-1}$	$5 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	–	–	–
	$n = 150K$	$5 \cdot 10^{-1}$	$5 \cdot 10^{-1}$	$5 \cdot 10^{-1}$	–	–	–
MLMC	$n_0 = 10$	$1 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$
Full-batch		$5 \cdot 10^{-1}$	$5 \cdot 10^{-1}$	$5 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$1 \cdot 10^{-2}$

Table 3. Stepsizes for the experiments we present in this work. We use momentum 0.9 for all configurations except MLMC, where we do not use momentum. We select the stepsizes according to the 'coarse-to-fine' strategy we describe in this section.

```
1  from robust_losses import RobustLoss
2
3  # we define the usual variables but also our robust loss
4  model = ...
5  criterion = ...
6  robust_loss = RobustLoss(geometry='chi-square', size=1.0)
7  # [...]
8  # training loop
9      outputs = model(inputs)
10     if not robust:
11         loss = criterion(outputs, targets).backward()
12     else:
13         loss = robust_loss(
14                criterion(outputs, targets, reduction='none')
15         ).backward()
16 # rest of the training loop
```

Figure 2. An example training loop in PyTorch where one can decide to use the robust training objective at the cost of three extra lines of code (lines 1, 6 and 13).

layer. While our current implementation only supports the robust objectives we analyze—namely, CVaR, KL-regularized CVaR, constrained-χ^2 and penalized-χ^2—it is easy to extend to other choices of ϕ and ψ.

F.4 Experiment results

We complement the training curves in Figure 1 with comparisons of robust generalization metrics and training efficiency. In Figures 3 and 4 we show the training curves of Figure 1 along with two "robust" generalization metrics and two "average" performance metrics. For Digits, we consider the loss and accuracy on the worst sub-group—typically the typed digit 9—as the robust generalization metrics. For ImageNet, we look at the average loss (resp. accuracy) on the 10 labels with highest loss (resp. lowest accuracy). In each figure we also show the values achieved by ERM with two different regularization strengths chosen to optimize either loss or accuracy on the worst-subgroup. In Tables 4 and 5 we compare the number of epochs the various algorithms require to reach a training loss within 2% of the minimal value found across all runs. To achieve such convergence with the full batch method we run it for much longer: 30K epochs for Digits and 1K epochs for ImageNet.

F.5 Discussion

F.5.1 Generalization performance

We now take a closer look at the curves presented in Figures 3 and 4. We first note that, in the context of machine learning, one does not wish to reach the minimum of the training objective but rather find a model that achieves good generalization performance. From that perspective, we observe that mini-batch methods achieve their best generalization performance in a shorter time than necessary to

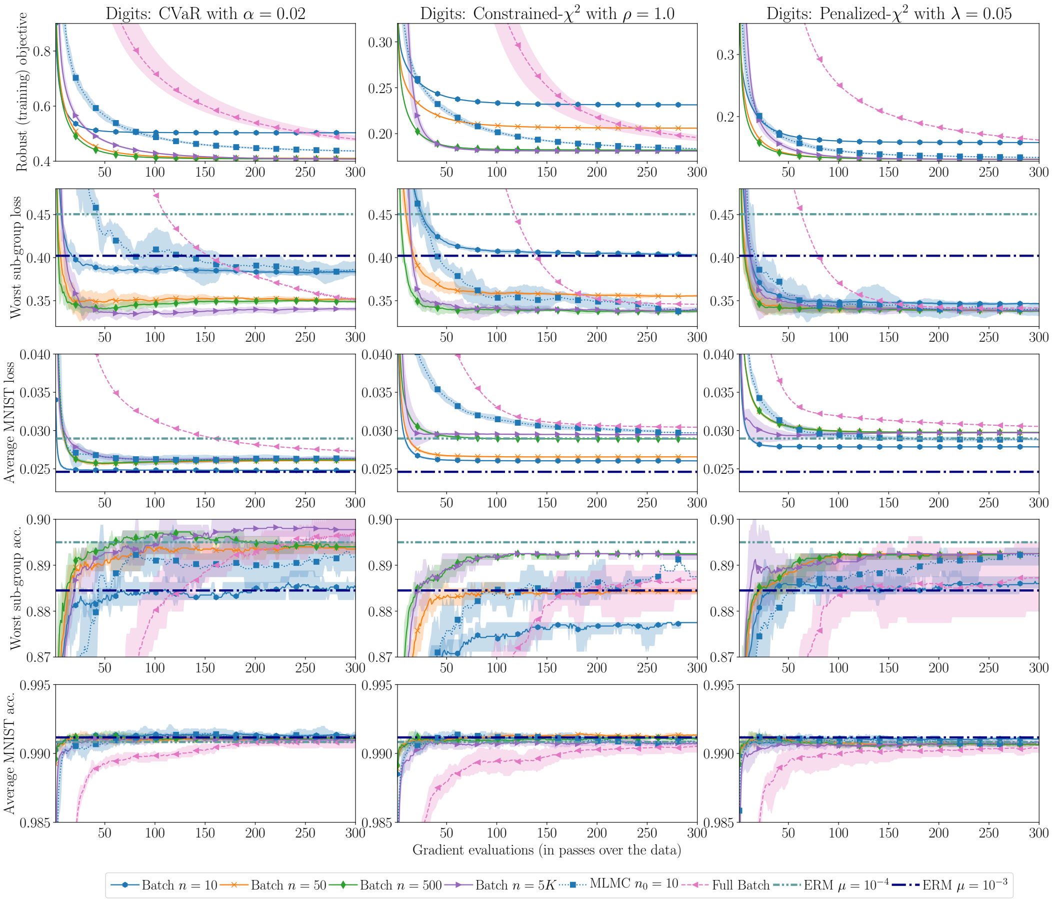

Figure 3. Detailed results from our digit recognition experiment. Shaded areas indicate range of variability across 5 repetitions (minimum to maximum), and the zoomed-in regions highlight the (often very small) "bias floor" of small batch sizes.

converge on the training objective, e.g., less than 50 epochs for CVaR on Digits when the training objective always requires more than 115 epochs.

In the case of Digits, we observe that DRO achieves a better trade-off than ERM in all settings. More precisely, DRO achieves better worst sub-group loss and accuracy than either of the ERM runs with no visible degradation in average accuracy and slightly worse average loss. We observe a similar trend in the case of ImageNet, albeit with a more visible degradation in average loss and accuracy.

We note that in the Digits experiment batch size $n = 10$ has generalization performance more similar to ERM. This is an expected by-product of the bias inherent in small batch size, as in the edge case $n = 1$, the mini-batch method degenerates to ERM.

Hu et al. [38] observe that applying DRO objectives of the form (12) directly on the 0-1 loss amounts to a simple monotonic transformation of the average accuracy, and is therefore equivalent to minimizing average accuracy. Thus, in as far as the logarithmic loss is a surrogate to the 0-1 loss (which is arguably the case in near realizable-settings), DRO might not provide improvements in robust accuracy. This is consistent with the observations in our experiments, where we see only small effects on the accuracy in the Digits experiments (which is close to realizable), and a somewhat more pronounced but still modest effect on ImageNet (which is not quite realizable, as the training accuracy is below 90%). Nevertheless, these observation do not preclude DRO from improvement the subpopulation test loss itself, as we see in our experiments: for Digits DRO provides between between 17.5% and 27% reduction in worst subgroup loss compared to ERM, and for ImageNet the reduction is a more modest 5.6% and 9%. While the common practice in machine learning is to view accuracy as the more important performance metric, logarithmic loss is also operationally

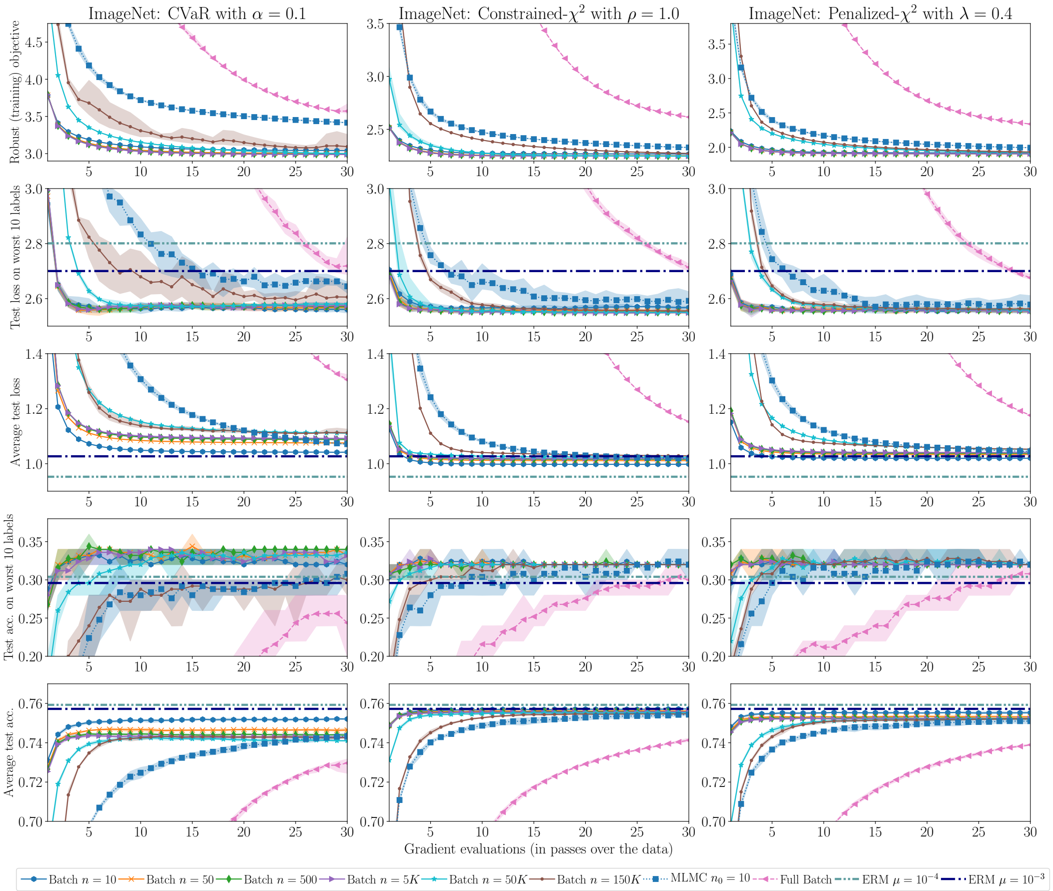

Figure 4. Detailed results from our ImageNet classification experiment. Shaded areas indicate range of variability across 5 repetitions (minimum to maximum), and the zoomed-in regions highlight the (often very small) "bias floor" of small batch sizes.

	Number of epochs to 2% of opt				Speed-up
	$n = 50$	$n = 500$	$n = 5K$	Full-batch	vs. full-batch
$\mathcal{L}_{\text{CVaR}}, \alpha = 0.02$	189 ± 3	$\mathbf{115} \pm 1$	193 ± 4	1035	9.0×
$\mathcal{L}_{\chi^2}, \rho = 1$	∞	74 ± 1	$\mathbf{60} \pm 3$	570	9.5×
$\mathcal{L}_{\chi^2\text{-pen}}, \lambda = 0.05$	107 ± 1	$\mathbf{104} \pm 1$	131 ± 5	1680	16.2×

Table 4. Empirical complexity for the Digits experiment in terms of number of epochs required to reach within 2% of the optimal training objective value, averaged across 5 seeds ± one standard deviation. (For the full-batch experiments we only ran one seed). The "speed-up" column gives the ratio between the full batch complexity and the best mini-batch complexity.

meaningful, as it measures the calibration of the model predictions. Thus, DRO is potentially helpful in situations where robust precise uncertainty estimates are important.

We remark that approaches that explicitly target the subgroups on which we measure the generalization [e.g., 61] will likely perform better than DRO. However, in contrast to these methods DRO is agnostic to the subgroup definition—except that we use a subgroup validation set in order to tune its uncertainty set size—and therefore requires less data annotation.

F.5.2 Optimization performance

As Figure 1 and Tables 4 and 5 indicate, mini-batch methods converge significantly faster than full-batch. We also see that, while theoretically optimal, MLMC methods are slower to converge.

	Number of epochs to 2% of opt							Speed-up
$n =$	10	50	500	$5K$	$50K$	$150K$	Full-batch	vs. full-batch
$\mathcal{L}_{\text{CVaR}}, \alpha = 0.1$	20	10	**9**	**9**	19	-	245	$27\times$
$\mathcal{L}_{\chi^2}, \rho = 1$	6	**5**	**5**	**5**	8 ± 1	23	160	$32\times$
$\mathcal{L}_{\chi^2\text{-pen}}, \lambda = 0.4$	7	**5**	**5**	**5**	22	26	180	$36\times$

Table 5. Empirical complexity for the ImageNet experiment in terms of number of epochs required to reach within 2% of the optimal training objective value, averaged across 5 seeds ± one standard deviation, whenever it is not zero. (For the full-batch experiments we only ran one seed). The "speed-up" column gives the ratio between the full batch complexity and the best mini-batch complexity.

		ImageNet times [minutes]			Digits times [minutes]		
Algorithm		per epoch	to 2% of opt	# epochs	per epoch	to 2% of opt	# epochs
Batch	$n = 10$	120 ± 5	850 ± 30	7	0.80 ± 0.1	∞	∞
	$n = 50$	23 ± 0.7	116 ± 4	**5**	0.23 ± 0.01	24 ± 1	107 ± 1
	$n = 500$	5.9 ± 0.2	29 ± 1	**5**	0.056 ± 0.004	5.8 ± 0.4	$\mathbf{104} \pm 1$
	$n = 5K$	3.3 ± 0.04	$\mathbf{16.5} \pm 0.2$	**5**	0.033 ± 0.004	$\mathbf{4.4} \pm 0.7$	131 ± 6
	$n = 50K$	2.2 ± 0.03	50 ± 0.9	22	–	–	–
	$n = 150K$	2.1 ± 0.03	55 ± 0.7	26	–	–	–
MLMC	$n_0 = 10$	16 ± 1	∞	∞	0.34 ± 0.02	∞	∞
Full-batch		2.1	380	180	0.022	37.0	1680

Table 6. Comparison wallclock time (in minutes) of the different algorithms, in terms of time per epoch and time to reach within 2% of the best training loss. In the last two columns, we report the number of epochs required to reach within 2% of the best training loss. We report ∞ for configurations that do not reach the sub-optimality goal for the duration of the experiment, and omit standard deviations when then they are 0.

Furthermore, the bias is empirically much smaller than what the theory predicts and setting the batch size as small as 50 guarantees negligible bias; we investigate this further below. As the theory predicts, the MLMC method (for corresponding values of n_0) effectively counteracts this bias, and is able to converge to the optimal value even when n_0 is 10.

We also note that the effect of batch size on the depth of the algorithm (number of iterations) is remarkably consistent with the theoretical prediction of the variance-based analysis in Section 3: for smaller batch sizes the number of steps is roughly inversely proportional to the batch size, and the total amount of work is constant. The best stepsize also grows linearly with the batch size (see Table 3). As batch sizes grow, the best stepsize plateaus and the number of steps required for convergence also stops decreasing with the batch size, making the total work become larger.

Runtime comparison. In Table 6 we report the gradient complexity and wallclock time to reach accuracy within 2% of the optimal value. For brevity, we show it for a single robust objective (penalized-χ^2), but we observe that similar results across robust objectives. We note that for small batch sizes the time per epoch is significantly larger than for larger batch sizes, this due in part to parallelization in evaluating ℓ and $\nabla\ell$ and in part to logging and Python interpreter overhead, which increase linearly with the number of iterations. However, these effects diminish as the batch size grows, and for batch size 5K the wallclock time to reach an accurate solution is an order of magnitude smaller than with the full-batch method. We run our experiments with 4 Intel Xeon E5-2699 CPUs and 12–32Gb of memory. Increasing the number of CPUs or using GPUs would allow for greater parallelism and improve the runtime at greater batch sizes. However, increasing the model complexity (e.g., to a deep neural network) would have the opposite effect. Using 4 CPUs for linear classification gives roughly the same range of feasible batch sizes as a ResNet-50 on large GPU arrays.

Bias analysis. Figure 1 shows that even for small batch sizes—where the guarantees of Proposition 1 are essentially vacuous—stochastic gradient steps with the mini-batch gradient estimator find solutions very close to optimal. There could be two explanations for this finding: (a) $\mathcal{L}$ and $\overline{\mathcal{L}}$ are actually much closer to each other than the theory predicts, or (b) $\mathcal{L}$ and $\overline{\mathcal{L}}$ are far apart as expected, but still their minimizers are close.

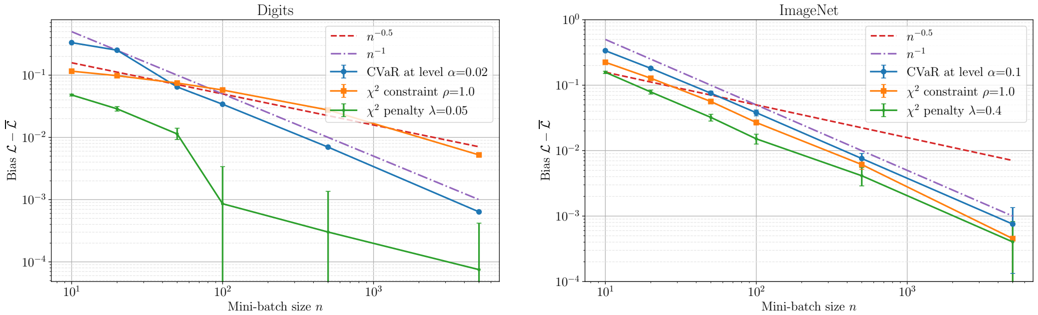

Figure 5. Evaluation of the bias $\mathcal{L}(\bar{x}_T; P_0) - \overline{\mathcal{L}}(\bar{x}_T; n)$ at the last iterate $\bar{x}_T$ of the experiments in Figure 1, for different batch sizes n. (These batch sizes n are not the same as the mini-batch size used to compute $\bar{x}_T$; we take the latter to be 10). Error bars indicate a 95% confidence interval computed using the bootstrap.

To test hypothesis (a), we examine the loss values at the last iterate $\bar{x}_T$ of our Digits and ImageNet experiments with mini-batch size 10. For each objective, we estimate $\overline{\mathcal{L}}(\bar{x}_T; n)$ for various values of n by averaging 50K evaluations of $\mathcal{L}(\bar{x}_T; S_1^n)$, and use it to compute an estimate of the bias $\mathcal{L}(\bar{x}_T; P_0) - \overline{\mathcal{L}}(\bar{x}_T; n)$.[6] In Figure 5 we plot the bias estimate against the mini-batch size n. We see that hypothesis (a) is false: for both ImageNet and Digits, the difference $\mathcal{L}(\bar{x}_T; P_0) - \overline{\mathcal{L}}(\bar{x}_T; n)$ is quite large at small n, as our upper bounds and matching lower bounds in the Bernoulli case would suggest. We also see that the bias decays as $1/n$ in all cases except for χ^2 constraint in Digits; this is again consistent with our theory as we expect the inverse-cdf assumption to be relevant in practice and particularly for CVaR where it only needs to hold around the $1-\alpha$ quantile. We conclude that despite the significant bias at small batch size n, approximate minimizers of $\overline{\mathcal{L}}(x; n)$ are also approximate minimizers of $\mathcal{L}(x; P_0)$. This is possibly due to the fact that the bias $\mathcal{L}(x; P_0) - \overline{\mathcal{L}}(x; n)$ is nearly constant as a function of x. We leave further study of this hypothesis to future work.

F.6 Comparison with alternative optimization methods

We complement the worst-case complexity comparison in Table 1 by repeating our experiments with two alternative optimization methods: dual SGM and primal-dual methods.

F.6.1 Comparison with dual SGM

Experiment description. Recall the dual SGM method we describe and analyze in Section A.3. The complexity guarantees of dual SGM depend quadratically on the size of the uncertainty set—scaling with α^{-2} for CVaR and with λ^{-2} for the penalized version of the χ^2 objective. In contrast, our theory predicts that the method we propose have an optimal linear dependence on the size of the uncertainty set. Here we empirically test this prediction on the Digits experiment. To do so, we compare the performance of our proposed mini-batch method with dual SGM for uncertainty sets of increasing size. For CVaR we consider

$$\alpha \in \{0.02, 0.006, 0.002, 0.0006, 0.0002\},$$

and for penalized χ^2 we consider

$$\lambda \in \{0.05, 0.015, 0.005, 0.0015, 0.0005\}.$$

Parameter tuning. For each uncertainty set size, we jointly tune the stepsizes γ_x and γ_η over the following grids

$$\gamma_x \in \{1 \cdot 10^{-i}, 3 \cdot 10^{-i}\}_{3 \le i \le 6},\ \gamma_\eta \in \{1 \cdot 10^{-i}\}_{2 \le i \le 5}.$$

We choose a coarser grid for γ_η as we noticed that the value of γ_η had a marginal influence on the final performance. For both the mini-batch algorithm and dual SGM, we pick the batch size $n = 500$ We follow the same averaging scheme and momentum as in our previous experiments.

[6] For CVaR it is fact possible to compute $\overline{\mathcal{L}}(\bar{x}_T; n)$ in closed form via (22); we do that for the Digits experiment. Scaling the computation to ImageNet is nontrivial, so there we use an empirical estimate instead.

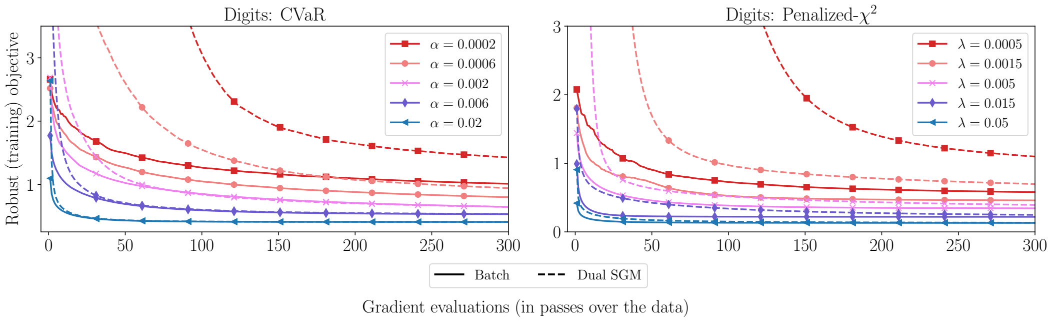

Figure 6. Comparison of batch methods to dual SGM on the digits experiments for increasing sizes of uncertainty set sizes or regularization. We observe that as the size grows, dual SGM performs increasingly worse.

Discussion of results. We plot the results of the experiment in Figure 6. As the theory predicts, when the size of the uncertainty set grows, dual SGM performs significantly worse than batch methods. Conversely, as expected, for small uncertainty sets dual SGM performs on par with the mini-batch method. We empirically observe that the performance of dual SGM depends only weakly on the choice of γ_η. As a result, dual SGM is not much more difficult to tune than the mini-batch method.

F.6.2 Comparison with primal-dual methods

Experiment description. We now turn to primal-dual methods, whose complexity guarantees scale as ϵ^{-2} but are linear in N, and are therefore expected to become less efficient as the size of the training set grows. To test this prediction, we repeat our Digits and ImageNet experiments (with $N = 60.6$K and $N = 1.2$M, respectively) using these alternative methods for the contrained-χ^2 and CVaR objectives. We then compare their performance to that of gradient methods with our mini-batch estimator.

Method description. Primal-dual methods maintain an iterate sequence $\{x_t, q_t\}_{t\in\mathbb{N}}$, where $q_t \in \mathcal{U}(P_0) \subset \Delta^N$ represent an online estimate of the distribution q attaining the maximum in (3) at $x_1, \ldots, x_t$. To compute x_{t+1}, q_{t+1}, we sample a batch of n indices J_1^n drawn independently from q_t, and (denoting $S_i = s_{J_i}$) estimate the gradient of $\sum_{i=1}^N q_i \ell(x; s_i)$ with respect to x and q as follows:

$$\tilde{g}_t^x = \frac{1}{n}\sum_{i=1}^n \nabla \ell(x_t; S_i) \text{ and } [\tilde{g}_t^q]_j = \frac{1}{n}\sum_{i=1}^n \frac{1}{q_{J_i}} \ell(x_t; S_i) 1_{\{J_i = j\}}.$$

To compute x_{t+1} from x_t and $\tilde{g}_t^x$ we apply the same stochastic gradient scheme we use in our previous experiments (Nesterov momentum 0.9).[7] We also use the averaging scheme in [65] with parameter 3 as before. To compute q_{t+1} from q_t and $\tilde{g}_t^q$ we apply a mirror descent step. For the constrained χ^2 problem the step is of the form

$$q_{t+1} = \mathop{\arg\max}_{q: \mathrm{D}_{\chi^2}(q, \frac{1}{N}\mathbf{1}) \le \rho} \left\{ \langle q, \gamma_q \tilde{g}_t^q \rangle + \frac{1}{2}\|q - q_t\|^2 \right\} = \mathop{\arg\min}_{q \in \Delta^N : \|q - \frac{1}{N}\mathbf{1}\|^2 \le 2\rho/n} \|q - (q_t + \gamma_q \tilde{g}_t^q)\|^2, \quad (58)$$

i.e., a Euclidean projection of the unconstrained gradient step on q_t to the uncertainty set. For the CVaR problem, the step is of the form

$$q_{t+1} = \mathop{\arg\max}_{q \in \Delta^N : \|q\|_\infty \le \frac{1}{\alpha N}} \{\langle q, \mathrm{clip}(\gamma_q \tilde{g}_t^q)\rangle + \mathrm{D}_{\mathrm{kl}}(q, q_t)\}, \quad (59)$$

where $\mathrm{clip}(x)$ is the Euclidean projection of x to $[-1, 1]^N$.

The χ^2 step is essentially the same as in [47], while the CvaR step is different from the proposal by Curi et al. [17]. Nevertheless, local norms regret analysis [13, 62] readily shows that with appropriate γ_x and γ_q the step (59) allows us to find ϵ-optimal solutions within $\lesssim \frac{N \log \frac{1}{\alpha} B^2 + G^2 R^2}{\epsilon^2}$

[7] The performance of the primal-dual method appears fairly insensitive to the use of momentum so we keep the parameter the same as in our previous experiments for simplicity.

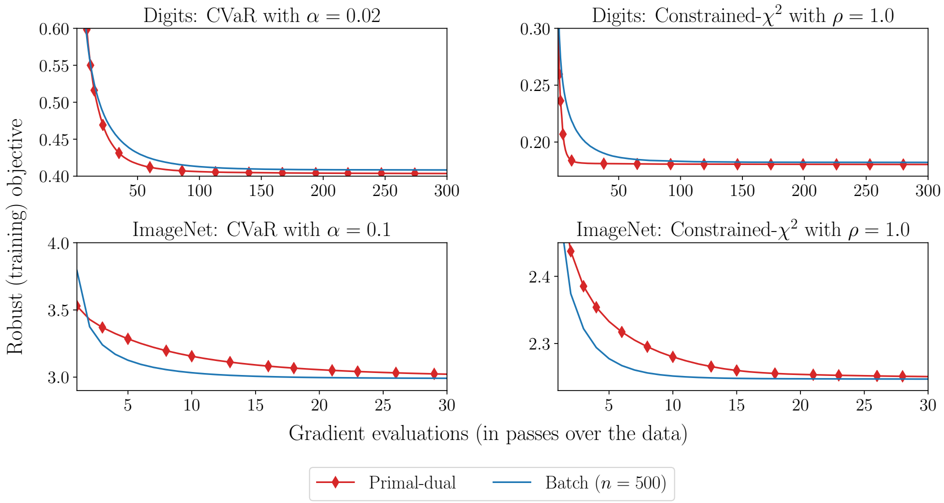

Figure 7. Comparison of batch methods to primal-dual methods. We observe that the primal-dual methods are more efficient on the Digits experiment, but the trend reverses on the large-scale ImageNet experiment.

iterations, similarly to the guarantee that Curi et al. [17] show for a computationally intractable determinantal point process scheme. They also propose a tractable approximation for this scheme, but do not prove that it converges to the solution of the CVaR problem.

Parameter tuning. For every training task we jointly tune the parameters γ_x and γ_q. We tune γ_x over the values 10^{-i}, $2 \cdot 10^{-i}$ and $5 \cdot 10^{-i}$ for $i \geq 1$ (similarly to our previous experiments) and we tune γ_q over the values 10^{-i} and $3 \cdot 10^{-i}$ for $i \geq 1$. The best-performing values of (γ_x, γ_q) are $(0.02, 0.003)$ for Digits/CVaR; $(0.02, 3 \cdot 10^{-7})$ for Digits/χ^2; $(0.05, 3 \cdot 10^{-5})$ for ImageNet/CVaR; and $(0.02, 3 \cdot 10^{-11})$ for ImageNet/χ^2. We use batch size $n = 500$ throughout.

Discussion of results. Figure 7 compares primal-dual and mini-batch primal methods with the best-performing hyperparameters, for two datasets and two objectives. For the Digits experiment, the primal-dual method perform better that the primal-only method (for χ^2 significantly so). This may appear surprising, since the primal-dual complexity guarantees are larger by an additional factor of $N = 60.6$K for this dataset. However, a closer look at the analysis of primal-dual methods shows that the term NB^2 is actually an upper bound on $\sum_{i=1}^{N}[\ell(x; s_i)]^2$ at $x = x_1, x_2, \ldots$. As the method converges, many data points are correctly classified with high confidence and therefore have very low value of $[\ell(x; s_i)]^2$. Hence, a more realistic complexity estimate would replace N by the number of incorrectly classified training points, which for Digits is quite small (less than 100). Moreover, we observe that the optimal value of γ_x for primal-dual methods is significantly larger than the corresponding step size for the primal-only method, likely because $\tilde{g}^x$ gives uniform weights to each s_i as opposed to the adversarial weight of the primal-only method. The larger step sizes enable more rapid optimization over x.

For the larger-scale ImageNet experiment, the primal-only method significantly outperforms the primal-dual method. This is consistent with the above discussion, since here the number of misclassified training examples is large (more than 100K).

As an additional illustration of the superior scalability of primal-only method, consider a thought experiment where we replicate each element in our dataset m times to form a new dataset of size mN. Clearly, this will have no impact on the primal-only method. In contrast, the norm of $\tilde{g}^q$ will grow by a factor of m, and we may expect the complexity of the method to increase by that factor as well.

Finally, we remark that tuning the primal-dual method is considerably more difficult than tuning the primal-only method. In addition to having two learning rates to search over, using an overly large value for γ_q typically causes the algorithm to converge to a suboptimal point rather than diverge. Therefore, the common procedure of decreasing the learning rate until divergence no longer occurs will fail for the primal-dual method.