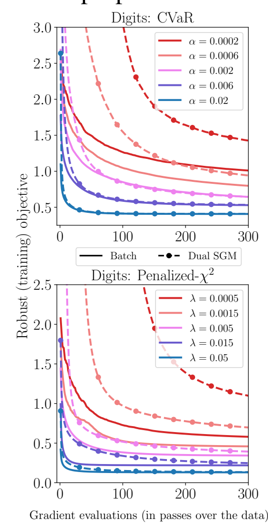
We thank the reviewers for their valuable time and insightful feedback. We begin by addressing two important points raised by multiple reviewers, and address the remaining comments of each reviewer in turn.

Intuition for the SGM bias Reviewers 1 and 5 ask for clarification and intuition for why mini-batch gradient estimates are biased for the true objective gradients but unbiased for the surrogate objective gradients. In a nutshell, the bias exists because expectation and minimization do not commute, and the true and surrogate objectives differ exactly by their order: see lines 681–682 in the supplementary material. On the other hand, the mini-batch loss estimate is unbiased for the surrogate loss *by definition*, and since ∇ and $\mathbb{E}$ do commute, we have that the mini-batch gradient is unbiased for the surrogate gradient. To see why the objective *gradient* is also similarly biased, consider the constructions in Proposition 4 (which lower bounds the bias) except with $\ell(x; s) = x \cdot s$. Then, the objective gradient at $x > 0$ is proportional to the loss must therefore be biased. We will discuss this in detail in the revised paper.



Comparison to other methods Reviewers 1, 2 and 5 ask for empirical comparison of the proposed mini-batch method with Dual SGM as well as the primal dual approaches of [15] and [43]. We agree that such comparisons are important, and we will add them to the revised manuscript. We include here preliminary results comparing our method with dual SGM for CVaR and χ^2 penalty on the digits experiment. Our theory predicts that the advantage of mini-batch over dual SGM increases as the uncertainty set grows (i.e., as α and λ decrease). Consequently, we vary the uncertainty set size (re-tuning the learning rates in a grid each time) and see results consistent with our prediction. For the revised paper we will also perform ImageNet experiments and comparison with [15] and [43]. We note that, consistently with the bound in Table 1, Namkoong and Duchi [43, Figure 1] observe that stochastic primal-dual performs on par or worse than the full-batch method; it should therefore be considerably slower than mini-batch SGM.

Reviewer 1 Thank you for the detailed comments and particularly the helpful questions and suggestions. We are glad you found our problem interesting and potentially impactful. Below we address the additional comments and questions given in point 8 of the review; we will make sure to include all clarifications in the revised paper as well. **(2)** By error floor, we refer to the suboptimality of the solution mini-batch SGM with batch size n finds when it has converged. That is, the error floor is $\mathcal{L}(\bar{x}; P_0) - \inf_{x\in\mathcal{X}} \mathcal{L}(x; P_0)$,where $\bar{x} = \arg\min_{x\in\mathcal{X}} \overline{\mathcal{L}}(x; n)$. **(3)** In this context $\mathbb{S} = \{1, ..., N\}$ and we refer to evaluating the loss over the entire dataset (see also Appendix A.2). **(4,8)** Thank you for pointing out these typos. **(6)** The runtimes presented in Table 1 correspond to the multilevel Monte Carlo guarantees in Section 4, and as such they require no assumptions on ν, β and α. **(9)** Using $\mathcal{L}_{\text{CVaR}}$ instead of $\mathcal{L}_{\text{kl-CVaR}}$ saved us tuning a parameter (the regularization strength), and still performed well compared to the full-batch method. A well-tuned smoothing parameter might obtain better results, though preliminary experiments did not show a major difference.

Reviewer 2 Thank you for the kind review and important questions. For intuition why it is possible to solve DRO problems with complexity independent of the training set size N is that the objective $\mathcal{L}(\cdot; P_0)$ is a *statistic* which one can estimate and optimize using a sufficiently large sample from P_0 ([19] proves this rigorously). This holds true even when $N = \infty$ (so P_0 has infinite support), and we therefore expect to have guarantees independent of N. The key challenge in obtaining our N-independent rates is that the standard analysis of SGM does not apply, because of the bias described above. We propose two ways to circumvent this issue. First, in Section 3 we characterize the surrogate objective for which an unbiased estimate is easy to write down. There, the key points are to bound the bias (Proposition 1) and the variance (Proposition 2). Second, in Section 4 we use multilevel Monte Carlo to formulate a sophisticated unbiased estimator of the objective (more precisely, one with arbitrarily low bias). There, the key point is bounding the second moment (Proposition 3). We will further highlight these points in the revision.

Reviewer 3 Thank you for the thoughtful suggestions; we are glad you found our paper interesting and well-written. Replacing AGD with SGD would unfortunately not allow us to obtain the guarantees in Theorem 1. To see this, note that the $O(1/T)$ in the SGD rate is proportional to the objective smoothness, which for us is $\Theta(\epsilon^{-1})$. Therefore, to make the error ϵ we would have to take T of the order ϵ^{-2}, harming our convergence guarantee. (Note also that the common $O(T^{-1/2})$ in both rates is proportional to the variance, which we make small by choosing a large batch size and appealing to Proposition 2). Regarding the suggestion to subsume the bounded loss assumption, note that if we only assume that $\mathcal{X}$ is bounded and $\ell(x; s)$ is Lipschitz in x, it does not give us bounded loss (consider $\ell(x; s) = x + s$ when $\mathbb{S} = \mathbb{R}$). We can, however, assume that $\ell(x_0; s) \in [-GR, GR]$ for all $s \in \mathbb{S}$ and some $x_0 \in \mathcal{X}$, which combined with boundedness and Lipschitz assumptions would imply a bound on the loss; we will comment on this in the revision.

Reviewer 5 Thank you for the helpful review and for highlighting important content and presentation issues. We discuss test-time robustness in our experiments in Appendix F.6.1, and note that prior work report significance robustness gains from the DRO objectives that we study [19, 31, 51, 68]. We hope that by providing efficient methods for DRO at scale, our paper will enable new demonstrations of the benefits of DRO.