Distributionally Robust Optimization via Ball Oracle Acceleration

Anonymous Author(s)
Affiliation
Address
email

Abstract

We develop and analyze algorithms for distributionally robust optimization (DRO) of convex losses. In particular, we consider group-structured and bounded f-divergence uncertainty sets. Our approach relies on an accelerated method that queries a ball optimization oracle, i.e., a subroutine that minimizes the objective within a small ball around the query point. Our main contribution is efficient implementations of this oracle for DRO objectives. For DRO with N non-smooth loss functions, the resulting algorithms find an ϵ-accurate solution with $\widetilde{O}\left(N\epsilon^{-2/3} + \epsilon^{-2}\right)$ first-order oracle queries to individual loss functions. Compared to existing algorithms for this problem, we improve complexity by a factor of up to $\epsilon^{-4/3}$.

1 Introduction

The increasing use of machine learning models in high-stakes applications highlights the importance of reliable performance across changing domains and populations [7, 39, 29]. An emerging body of research addresses this challenge by replacing Empirical Risk Minimization (ERM) with Distributionally Robust Optimization (DRO) [4, 43, 42, 16, 33], with applications in natural language processing [40, 54, 29], reinforcement learning [12, 47] and algorithmic fairness [21, 49]. While ERM minimizes the average training loss, DRO minimizes the worst-case expected loss over all probability distributions in an *uncertainty set* $\mathcal{U}$, that is, it minimizes

$$\mathcal{L}_{\text{DRO}}(x) \coloneqq \sup_{Q \in \mathcal{U}} \mathbb{E}_{S \sim Q}[\ell_S(x)], \tag{1}$$

where $\ell_S(x)$ is the loss a model $x \in \mathcal{X}$ incurs on a sample S. This work develops new algorithms for DRO, focusing on formulations where $\mathcal{U}$ contains distributions supported on N training points, where N is potentially large. We consider two well-studied DRO variants: (1) Group DRO [50, 25, 42], and (2) f-divergence DRO [4, 16].

Group DRO Machine learning models may rely on spurious correlations (that hold for most training examples but are wrongly linked to the target) and therefore suffer high loss on minority groups where these correlations do not hold [24, 21, 7]. To obtain high performance across all groups, Group DRO minimizes the worst-case loss over groups. Given a set $\mathcal{U} = \{w_1, \ldots, w_M\}$ of M distributions over $[N]$, the Group DRO objective is[1]

$$\mathcal{L}_{\text{g-DRO}}(x) \coloneqq \max_{i \in [M]} \mathbb{E}_{j \sim w_i} \ell_j(x) = \max_{i \in [M]} \sum_{j=1}^{N} w_{ij} \ell_j(x). \tag{2}$$

[1]Typically, each w_i is uniform over a subset ("group") of the N training points. However, most approaches (and ours included) extend to the setting of arbitrary w_i's, which was previously considered in [50].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Smoothness	Method	Group DRO (2)	f-divergence DRO (3)
None ($L = \infty$)	Subgradient method [38]	$N\epsilon^{-2}$	$N\epsilon^{-2}$
	Stoch. primal-dual [36] *	$M\epsilon^{-2}$	$N\epsilon^{-2}$
	MLMC stoch. gradient [32]	-	$\rho\epsilon^{-3}$ or $\alpha^{-1}\epsilon^{-2}$ †
	Ours	$N\epsilon^{-2/3} + \epsilon^{-2}$	$N\epsilon^{-2/3} + \epsilon^{-2}$
Weak ($L \approx 1/\epsilon$)	AGD on softmax [37]	$N\epsilon^{-1}$	$N\epsilon^{-1}$
	Ours	$N\epsilon^{-2/3} + N^{3/4}\epsilon^{-1}$	$N\epsilon^{-2/3} + \sqrt{N}\epsilon^{-1}$

Table 1. Number of $\nabla\ell_i$ and ℓ_i evaluations to obtain $\mathbb{E}[\mathcal{L}_{\text{DRO}}(x)] - \min_{x_\star\in\mathcal{X}} \mathcal{L}_{\text{DRO}}(x_\star) \le \epsilon$, where N is the number of training points and (for Group DRO) M is the number of groups. The stated rates omit constant and polylogarithmic factors. * Requires an additional uniform bound on losses (see Appendix C.1). † These rates hold only for specific f-divergences: CVaR at level α or χ^2-divergence with size ρ, respectively.

If we define the loss of group i as $\mathcal{L}_i(x) \coloneqq \sum_{j=1}^N w_{ij}\ell_j(x)$ then objective (2) is equivalent to $\max_{q\in\Delta^M} \sum_{i\in[M]} q_i\mathcal{L}_i(x)$ with $\Delta^M \coloneqq \{q \in \mathbb{R}^M_{\ge 0} \mid \vec{1}^T q = 1\}$. Note that, unlike ERM, Group DRO requires additional supervision in the form of subgroup identities encoded by $\{w_i\}$.

DRO with f-divergence Another approach to DRO, which requires only as much supervision as ERM, takes $\mathcal{U}$ to be an f-divergence ball around the empirical (training) distribution. For every convex function $f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ such that $f(1) = 0$, the f-divergence between distributions q and p over $[N]$ is $D_f(q, p) \coloneqq \sum_{i\in[N]} p_i f(q_i/p_i)$. The f-divergence DRO problem corresponds to the uncertainty set $\mathcal{U} = \{q \in \Delta^N : D_f(q, \frac{1}{N}\mathbf{1})) \le 1\}$, i.e.,

$$\mathcal{L}_{f\text{-div}}(x) \coloneqq \max_{q\in\Delta^N : \frac{1}{N}\sum_{i\in[N]} f(Nq_i)\le 1} \sum_{i\in[N]} q_i\ell_i(x). \tag{3}$$

This formulation generalizes several well-studied instances of DRO, with the two most notable examples being conditional value at risk (CVaR) and χ^2 uncertainty sets. CVaR at level α corresponds to $f(x) = \mathbb{1}_{\{x<\frac{1}{\alpha}\}}$, and has many applications in finance such as portfolio optimization and credit risk evaluation [41, 30] as well as in machine learning [40, 32, 14, 53, 12, 47]. The χ^2 uncertainty set with size $\rho > 0$ corresponds to $f(x) \coloneqq \frac{1}{2\rho}(x-1)^2$ and the resulting DRO problem is closely linked to variance regularization [15] and has been extensively studied in statistics and machine learning [35, 21, 15, 32, 54].

Complexity notion In this paper, we design improved-complexity methods for solving the problems (2) and (3) under the assumption that the loss ℓ_i is convex and Lipschitz for all i. We measure complexity by the (expected) requiered number of $\ell_i(x)$ and $\nabla\ell_i(x)$ evaluations to obtain ϵ-suboptimal solution, i.e., return x such that $\mathcal{L}_{\text{DRO}}(x) - \min_{x_\star\in\mathcal{X}} \mathcal{L}_{\text{DRO}}(x_\star) \le \epsilon$ with constant probability. Table 1 summarizes our complexity bounds and compares them to prior art. Throughout the introduction we assume (for simplicity) a unit domain size and that each loss is 1-Lipschitz.

Prior art Let us review existsing methods that solve Group DRO and f-divergence DRO problems (for extended discussion on other related work see Appendix B). For a dataset with N training points, the subgradient method [38] finds an ϵ approximate solution in ϵ^{-2} iterations. Computing a single subgradient costs N functions evaluations (since we need to find the maximizing q). Therefore, the complexity of this method is $O\big(N\epsilon^{-2}\big)$.

DRO can also be viewed as a game between a minimizing x-player and a maximizing q-player, which makes it amenable to primal-dual methods [36, 35, 42]. If we further assume that the losses are bounded then, for $q \in \Delta^m$, stochastic mirror descent with local norms obtains a regret bound of $O\big(\sqrt{m\log(m)/T}\big)$ (see Appendix C.1). As a consequence, for Group DRO (where $m = M$) the complexity is $\widetilde{O}\big(M\epsilon^{-2}\big)$, and for f-divergence DRO ($m = N$) the complexity is $\widetilde{O}\big(N\epsilon^{-2}\big)$.

Levy et al. [32] studied χ^2-divergence and CVaR DRO problems, and proposed using standard gradient methods with a gradient estimator based on multilevel Monte Carlo (MLMC) [5]. For

χ^2-divergence with ball of size ρ they proved a complexity bound of $\widetilde{O}\big(\rho\epsilon^{-3}\big)$, and for CVaR at level α they established complexity $\widetilde{O}\big(\alpha^{-1}\epsilon^{-2}\big)$. However, for large uncertainty sets (when ρ or α^{-1} approach N) their method does not improve over the subgradient method.

Stronger complexity bounds are available under the weak smoothness assumption that each ℓ_i has $O\big(\epsilon^{-1}\big)$-Lipschitz gradient. In particular, We can apply Nesterov's accelerated gradient descent method [37] on an entropy-regularized version of our objective to solve the problem with complexity $\widetilde{O}\big(N\epsilon^{-1}\big)$; see Appendix C.2 for more details.

Our contribution We propose algorithms that solve the problems (2) and (3) with complexity $\widetilde{O}\big(N\epsilon^{-2/3}+\epsilon^{-2}\big)$. Compared to previous works, we obtain better convergence rates for DRO with general f-divergence when $N \gg 1$ and for Group DRO when $M \gg N\epsilon^{4/3}$. When the losses have $O(\epsilon^{-1})$-Lipschitz gradient, we solve f-divergence DRO with complexity $\widetilde{O}\big(N\epsilon^{-2/3}+\sqrt{N}\epsilon^{-1}\big)$, and, under an even weaker mean-square smoothness assumption ($\mathbb{E}_{j\sim w_i}\|\nabla\ell_j(x)-\nabla\ell_j(y)\|^2 \leq O(\epsilon^{-2})\|x-y\|^2$ for all x, y and i), we solve Group DRO with complexity $\widetilde{O}\big(N\epsilon^{-2/3}+N^{3/4}\epsilon^{-1}\big)$.

Our complexity bounds are independent of the structure of f and $\{w_i\}$, allowing us to consider arbitrarily f-divergence balls and support a large number of (potentially overlapping) groups. Our rates are optimal up to logarithmic factors for the special case of minimizing $\max_{i\in[N]}\ell_i(x)$, which corresponds to Group DRO with N distinct groups and f-divergence DRO with $f=0$ [51, 55, 11].

Our approach Our algorithms are based on a technique for accelerating optimization with a *ball optimization oracle*, introduced by Carmon et al. [9] and further developed in [11, 3]. Given a function F and a query point x, the ball optimization oracle returns an approximate minimizer of F inside a ball around x with radius r; the works [9, 11, 3] show how to minimize F using $\widetilde{O}(r^{-2/3})$ oracle calls. Our development consists of efficiently implementing ball oracles with radius $r=\widetilde{O}(\epsilon)$ for the DRO problems (2) and (3), leveraging the small ball constraint to apply stochastic gradient estimators that would have exponential variance and/or cost without it.

Carmon et al. [11] previously executed this strategy for minimizing the maximum loss, i.e., $\max_{q\in\Delta^N}\sum_i q_i\ell_i(x)$, which is a special case of both Group DRO and f-divergence DRO. However, the ball-oracle implementations of [11] do not directly apply to the DRO problems that we consider; our oracle implementations differ significantly and intimately rely on the Group DRO and f-divergence problem structure. We now briefly review the main differences between our approach and [11], highlighting our key technical innovations along the way.

Since the Group DRO objective is $\max_{q\in\Delta^M}\sum_i q_i\mathcal{L}_i(x)$ for $\mathcal{L}_i(x)=\sum_{j\in[N]}w_{ij}\ell_j(x)$, one may naively apply the technique of [11] with $\mathcal{L}_i$ replacing ℓ_i. However, every step of such a method would involve computing quantities of the form $e^{\mathcal{L}_i(x)/\epsilon'}$ (for some $\epsilon'=\widetilde{\Theta}(\epsilon)$), which can be up to N times more expensive than computing $e^{\ell_j(x)/\epsilon'}$ for a single j. To avoid such expensive computation we use MLMC [5] to obtain an unbiased estimate of $e^{\mathcal{L}_i(x)/\epsilon}$ with complexity $O(1)$ and appropriately bounded variance. In the weakly-smooth case we also adapt our estimator to facilitate variance reduction [27, 2].

For f-divergence we consider the well-known dual form [4, 43]

$$\max_{q\in\Delta^N}\sum_{i\in[N]}\{q_i\ell_i(x)-\psi(q_i)\}=\min_{y\in\mathbb{R}}\left\{\Upsilon(x,y)\coloneqq\sum_{i\in[N]}\psi^*(\ell_i(x)-y)+y\right\}$$

where $\psi^*(v)=\max_{t\geq 0}\{vt-\psi(t)\}$ is the Fenchel dual of ψ. Stochastic gradient methods applied directly on the dual formulation are notoriously unstable (see e.g., [35]). This is due to the fact that the Fenchel dual ψ^* can be very badly behaved even for standard f-divergences. We solve this problem by (a) considering a small ball, (b) entropy-regularizing ψ. Our techniques rely on two technical observations: (i) $\log(\psi_\epsilon^{*\prime}(\cdot))$ is $1/\epsilon'$-Lipschitz for all convex ψ, and (ii) for 1-Lipschitz losses, $y^\star(x)=\operatorname{argmin}_{y\in\mathbb{R}}\Upsilon(x,y)$ satisfies $|y^\star(x)-y^\star(x')|\leq\|x-x'\|$ for all x, x'. To the best of our knowledge, these observations are new and potentially of independent interest.

Limitations Our work indicates that the ball optimization approach offers significant complexity gains for DRO, however, we note that turning the algorithms we propose into practical DRO methods

faces several challenges. A main challenge is the costly bisection procedure common to all Monteiro-Svaiter-type acceleration schemes [34, 19, 9, 45]. Moreover, similarly to [11], our ball radius r_ϵ and smoothing level ϵ' depend on the desired accuracy ϵ; a more adaptive setting for these parameters is likely important. For further discussion we refer to Appendix A.

2 Preliminaries

Notation We write $\|\cdot\|$ for the Euclidean norm. We denote by $\mathbb{B}_r(x_0)$ the Euclidean ball of radius r around x_0. We let $\Delta^n := \{q \in \mathbb{R}^n_{\geq 0} \mid \mathbf{1}^T q = 1\}$ denote the probability simplex in $\mathbb{R}^n$. For the sequence $z_m, \ldots, z_n$ we use the shorthand z_m^n. Using F as a generic placeholder (typically for a loss function ℓ_i), we make frequent use of the following assumption.

Assumption 1. *The function $F : \mathcal{X} \to \mathbb{R}$ is convex and and G-Lipschitz, i.e., for all $x, y \in \mathcal{X}$ we have $|F(x) - F(y)| \leq G\|x - y\|$. In addition, the set $\mathcal{X}$ has Euclidean diameter at most R.*

Throughout, N denotes the number of losses and, in Section 3, M denotes the number of groups. We use ϵ for our target accuracy and $r_\epsilon := \epsilon'/G$ for the ball radius, where $\epsilon' = \epsilon/(2\log M)$ for Group-DRO (Section 3) and $\epsilon' = \epsilon/(2\log N)$ for f-divergence DRO (Section 4).

Complexity model We measure an algorithm's complexity by its *expected* number of ℓ_i and $\nabla\ell_i$ evaluations; bounds on expected evaluation number can be readily converted to more standard probability 1 bounds [see 3, Appendix A.3]. Moreover if $\mathcal{X} \subset \mathbb{R}^d$, $d = \Omega(\log N)$,[2] and the time to evaluate ℓ_i and $\nabla\ell_i$ is $O(d)$, the expected runtime of all the algorithms we consider is at most d times the evaluation complexity.

2.1 Ball oracle acceleration

We now briefly summarize the complexity bounds given by the framework of [9, 11, 3] for accelerated minimization using queries to (inexact) ball optimization oracles, defined as follows.

Definition 1. *An algorithm is a Ball Regularized Optimization Oracle of radius r (r-BROO) for function $F : \mathcal{X} \to \mathbb{R}$ if for query point $\bar{x} \in \mathcal{X}$, regularization parameter $\lambda > 0$ and desired accuracy $\delta > 0$ it returns $\mathcal{O}_{\lambda,\delta}(\bar{x}) \in \mathcal{X}$ satisfying*

$$\mathbb{E}\left[F(\mathcal{O}_{\lambda,\delta}(\bar{x})) + \frac{\lambda}{2}\|\mathcal{O}_{\lambda,\delta}(\bar{x}) - \bar{x}\|^2\right] \leq \min_{x \in \mathbb{B}_r(\bar{x}) \cap \mathcal{X}} \left\{F(x) + \frac{\lambda}{2}\|x - \bar{x}\|^2\right\} + \frac{\lambda}{2}\delta^2. \tag{4}$$

Proposition 1. *Let F satisfy Assumption 1, let $\mathcal{C}_F$ be the complexity of evaluating F exactly, and let $\mathcal{C}_\lambda(\delta)$ bound the complexity of an r-BROO query with δ, λ. Assume that $\mathcal{C}_\lambda(\delta)$ is non-increasing in λ and at most polynomial in $1/\delta$. For any $\epsilon > 0$, Algorithm 1 returns x such that $F(x) - \min_{x_\star \in \mathcal{X}} F(x_\star) \leq \epsilon$ with probability at least $\frac{1}{2}$. For $m_\epsilon = O\big(\log \frac{GR^2}{\epsilon r}\big)$ and $\lambda_{\mathrm{m}} = O\big(\frac{m_\epsilon^2 \epsilon}{r^{4/3} R^{2/3}}\big)$, the complexity of the algorithm is*

$$O\left(\left(\frac{R}{r}\right)^{2/3}\left[\left(\sum_{j=0}^{m_\epsilon} \frac{1}{2^j}\mathcal{C}_{\lambda_{\mathrm{m}}}\left(\frac{r}{2^{j/2} m_\epsilon^2}\right)\right) m_\epsilon + (\mathcal{C}_{\lambda_{\mathrm{m}}}(r) + \mathcal{C}_F) m_\epsilon^3,\right]\right). \tag{5}$$

Informally, the proposition shows that $\widetilde{O}((R/r)^{2/3})$ BROO calls with $\lambda = \widetilde{\Omega}(\epsilon/(r^{3/4}R^{2/3}))$ and accuracy $\delta = \widetilde{O}(r)$ suffice to find an ϵ-accurate solution. As we show in the sequel, for $\mathcal{C}_\lambda(\delta) = \widetilde{O}\big(N + (\frac{G}{\lambda\delta})^2\big)$ the resulting complexity bound is $\widetilde{O}\big(N(\frac{GR}{\epsilon})^{2/3} + (\frac{GR}{\epsilon})^2\big)$. The summation over j in bound (5) stems from the use of MLMC to de-bias the BROO output (i.e., make it exact in expectation): compared to the original proposal of Asi et al. [3], our version of the procedure in Appendix D slightly alters this MLMC scheme by de-biasing one accurate BROO call instead of averaging many inaccurate de-biased calls, improving our bounds by logarithmic factors.

[2] The assumption $d = \Omega(\log N)$ is only necessary for our results on f-divergence DRO (Section 4), where the runtime of computing $\operatorname{argmin}_{y \in \mathbb{R}} \Upsilon(x, y)$ is $O(Nd + N\log N)$ due to the need to sort the losses.

3 Group DRO

In this section we develop our BROO implementations for the Group DRO objective (2). In Section 3.1 we describe an "exponentiated group-softmax" function that approximates $\mathcal{L}_{\text{g-DRO}}$ with additive error at most $\epsilon/2$. We then apply stochastic gradient methods on this function to obtain BROO implementations that yield improved rates for Group DRO via Proposition 1: we first consider the non-smooth case in Section 3.2 and then the weakly-smooth case in Section 3.3.

3.1 Exponentiated group-softmax

Given a cheap and unbiased stochastic gradient estimator of $\nabla\mathcal{L}_{\text{g-DRO}}$, we could use a variant of SGD and minimize $\mathcal{L}_{\text{g-DRO}}$ to ϵ-suboptimal solution using $O(\epsilon^{-2})$ steps. However, obtaining an unbiased estimator is challenging due to the maximum operator in $\mathcal{L}_{\text{g-DRO}}$. We first describe a straightforward extension of the exponentiated softmax of Carmon et al. [11] that approximates $\mathcal{L}_{\text{g-DRO}}$ and has the form of a weighted finite sum which is more amenable for stochastic optimization. For target accuracy ϵ, regularization parameter $\lambda \geq 0$, center point $\bar{x} \in \mathcal{X}$ and $\epsilon' = \epsilon/(2\log M) > 0$, the (regularized) group-softmax function is

$$\mathcal{L}_{\text{smax},\epsilon,\lambda}(x) \coloneqq \epsilon' \log\left(\sum_{i\in[M]} e^{\frac{\mathcal{L}_i(x)}{\epsilon'}}\right) + \frac{\lambda}{2}\|x-\bar{x}\|^2 \;\text{ where }\; \mathcal{L}_i(x) = \sum_{j\in[N]} w_{ij}\ell_j(x). \tag{6}$$

We will implement a BROO for $\mathcal{L}_{\text{smax},\epsilon} \coloneqq \mathcal{L}_{\text{smax},\epsilon,0}$, which is a uniform approximation of $\mathcal{L}_{\text{g-DRO}}$: $|\mathcal{L}_{\text{g-DRO}}(x) - \mathcal{L}_{\text{smax},\epsilon}(x)| \leq \epsilon/2$ for all $x \in \mathcal{X}$; see Appendix E.1 for details.

The (regularized) exponentiated group-softmax is

$$\Gamma_{\epsilon,\lambda}(x) \coloneqq \sum_{i\in[M]} \bar{p}_i\gamma_i(x) \;\text{ where }\; \gamma_i(x) = \epsilon' e^{\frac{\mathcal{L}_i(x)-\mathcal{L}_i(\bar{x})+\frac{\lambda}{2}\|x-\bar{x}\|^2}{\epsilon'}} \;\text{ and }\; \bar{p}_i = \frac{e^{\frac{\mathcal{L}_i(\bar{x})}{\epsilon'}}}{\sum_{i\in[M]} e^{\frac{\mathcal{L}_i(\bar{x})}{\epsilon'}}}. \tag{7}$$

In the following lemma we (easily) extend Carmon et al. [11, Lemma 1] to exponentiated group-softmax, showing that $\Gamma_{\epsilon,\lambda}$ is well-behaved inside a ball of (appropriately small) radius r around $\bar{x}$ and facilitates minimizing $\mathcal{L}_{\text{smax},\epsilon,\lambda}$ in that ball; see Appendix E.1 for the proof.

Lemma 1. *Let each ℓ_i satisfy Assumption 1, and consider the restriction of $\mathcal{L}_{\text{smax},\epsilon,\lambda}$ (6) and $\Gamma_{\epsilon,\lambda}$ (7) to $\mathbb{B}_r(\bar{x})$. Then the functions have the same minimizer $x_\star \in \mathbb{B}_r(\bar{x})$ and, if $\lambda \leq O(G/r)$ and $r \leq O(\epsilon'/G)$, then (a) $\Gamma_{\epsilon,\lambda}$ is $\Omega(\lambda)$-strongly convex, (b) each γ_i is $O(G)$-Lipschitz and (c) for every $x \in \mathbb{B}_r(\bar{x})$ we have $\mathcal{L}_{\text{smax},\epsilon,\lambda}(x) - \mathcal{L}_{\text{smax},\epsilon,\lambda}(x_\star) \leq O(\Gamma_{\epsilon,\lambda}(x) - \Gamma_{\epsilon,\lambda}(x_\star))$.*

3.2 BROO implementation for Group DRO non-smooth losses

To motivate our BROO implementation, let us review how [11] use the exponentiated softmax in the special case of size-1 groups, i.e., $\mathcal{L}_i = \ell_i$, and explain the difficulty that their approach faces when the group structure is introduced. The BROO implementation in [11] is based on SGD variant with the stochastic gradient estimator $\hat{g}(x) = e^{(\mathcal{L}_i(x)-\mathcal{L}_i(\bar{x}))/\epsilon'}\nabla\mathcal{L}_i(x)$ where $i \sim \bar{p}_i$. However, for Group DRO where $\mathcal{L}_i = \sum_{j\in[N]} w_{ij}\ell_j$, the estimator $\hat{g}(x)$ can be up to N times more expensive to compute. Approximating $\hat{g}(x)$ by drawing $j, j' \sim w_i$ and taking $e^{(\ell_j(x)-\ell_j(\bar{x}))/\epsilon'}\nabla\ell_{j'}(x)$ will result in a biased estimator since $\mathbb{E}_{j\sim w_i} e^{(\ell_j(x)-\ell_j(\bar{x}))/\epsilon'} \neq e^{(\mathcal{L}_i(x)-\mathcal{L}_i(\bar{x}))/\epsilon'}$. To address this challenge we propose a new gradient estimator based on the multilevel Monte Carlo (MLMC) method [5].

The MLMC unbiased estimator for $\gamma_i(x) = \epsilon' e^{(\mathcal{L}_i(x)-\mathcal{L}_i(\bar{x}))/\epsilon'}$, which we denote by $\widehat{\mathcal{M}}[\gamma_i(x)]$, is defined as follows:

$$\text{Draw } J \sim \text{Geom}\left(1-\tfrac{1}{\sqrt{8}}\right), S_1,\ldots,S_n \overset{\text{iid}}{\sim} w_i \text{ and let } \widehat{\mathcal{M}}[\gamma_i(x)] \coloneqq \widehat{\gamma}(x;S_1) + \frac{\widehat{\mathcal{D}}_{2^J}}{p_J},$$

where $p_j \coloneqq \mathbb{P}(J=j) = \left(1/\sqrt{8}\right)^j\left(1-\frac{1}{\sqrt{8}}\right)$ and, for $n \in 2\mathbb{N}$, we define

$$\widehat{\mathcal{D}}_n \coloneqq \widehat{\gamma}(x;S_1^n) - \frac{\widehat{\gamma}\left(x;S_1^{n/2}\right) + \widehat{\gamma}\left(x;S_{n/2+1}^n\right)}{2} \text{ and } \widehat{\gamma}(x;S_1^n) \coloneqq \epsilon' e^{\frac{1}{n}\sum_{j=1}^n \frac{\ell_{S_j}(x)-\ell_{S_j}(\bar{x})+\frac{\lambda}{2}\|x-\bar{x}\|^2}{\epsilon'}}.$$

With the MLMC estimator for γ_i in hand, we estimate the gradient of $\Gamma_{\epsilon,\lambda}$ as follows:

$$\text{Draw } i \sim p(\bar{x})\ ,\ j \sim w_i \text{ and set } \hat{g}(x) = \frac{1}{\epsilon'}\widehat{\mathcal{M}}[\gamma_i(x)](\nabla \ell_j(x) + \lambda(x - \bar{x})). \tag{8}$$

In the following lemma we summarize the important properties of the MLMC and gradient estimators; see Appendix E.2 for the proof.

Lemma 2. *Let each ℓ_i satisfy Assumption 1, and let $r \le \frac{\epsilon'}{G}$, $\lambda \le \frac{G}{r}$ and $x \in \mathbb{B}_r(\bar{x})$. Then $\widehat{\mathcal{M}}[\gamma_i(x)]$ and $\hat{g}(x)$ are unbiased for $\gamma_i(x)$ and $\nabla\Gamma_{\epsilon,\lambda}(x)$, respectively, and have bounded second moments: $\mathbb{E}\big[\widehat{\mathcal{M}}[\gamma_i(x)]\big]^2 \le O\Big(\frac{G^4\|x-\bar{x}\|^4}{\epsilon'^2} + \epsilon'^2\Big)$ and $\mathbb{E}\|\hat{g}(x)\|^2 \le O\big(G^2\big)$. In addition, the complexity of computing $\widehat{\mathcal{M}}[\gamma_i(x)]$ and $\hat{g}(x)$ is $O(1)$.*

Due to Lemma 2 and since $\Gamma_{\epsilon,\lambda}$ is $\Omega(\lambda)$-strongly convex, we can use the Epoch-SGD algorithm of Hazan and Kale [22] with our gradient estimator (8). This algorithm has rate of convergence $O\big(G^2/(\lambda T)\big)$ and our gradient estimator requires additional N function evaluations for precomputing the sampling probabilities $\{\bar{p}_i\}$. We thus arrive at the following complexity bound.

Theorem 1. *Let each ℓ_j satisfy Assumption 1, let $\epsilon, \delta, \lambda > 0$ and let $r_\epsilon = \epsilon/(2G\log M)$. For any query point $\bar{x} \in \mathbb{R}^d$, regularization strength $\lambda \le O(G/r_\epsilon)$ and accuracy δ, EpochSGD [22, Algorithm 1]) with the gradient estimator* (8) *outputs a valid r_ϵ-BROO response and has complexity $\mathcal{C}_\lambda(\delta) = O\big(N + \frac{G^2}{\lambda^2\delta^2}\big)$. Consequently, the complexity of finding an ϵ-suboptimal minimizer of $\mathcal{L}_{\text{g-DRO}}$* (2) *with probability at least $\frac{1}{2}$ is*

$$O\left(N\left(\frac{GR}{\epsilon}\right)^{2/3}\log^{11/3} H + \left(\frac{GR}{\epsilon}\right)^2 \log^2 H\right) \ \textit{where}\ \ H \coloneqq M\frac{GR}{\epsilon}.$$

We provide the proof for Theorem 1 in Appendix E.3; the final complexity bound follows from straightforward calculations which we now briefly outline. According to Proposition 1, finding an $\frac{\epsilon}{2}$-suboptimal solution for $\mathcal{L}_{\text{smax},\epsilon}$ (and consequently an ϵ-suboptimal solution for $\mathcal{L}_{\text{g-DRO}}$) involves $\widetilde{O}\big((R/r_\epsilon)^{2/3}\big)$ BROO calls with accuracy $\delta = \widetilde{\Omega}\big(r_\epsilon 2^{-J/2}\big)$ and regularization strength $\lambda \ge \lambda_{\text{m}}$, where $J = \min\{\text{Geom}(\frac{1}{2}), m\}$. We may therefore bound the complexity of each such call by

$$\sum_{j=0}^{m} 2^{-j}\mathcal{C}_{\lambda_{\text{m}}}\Big(r_\epsilon 2^{-j/2}\Big) = \sum_{j=0}^{m} 2^{-j}\widetilde{O}\left(N + \frac{2^j G^2}{\lambda_{\text{m}}^2 r_\epsilon^2}\right) \overset{(\star)}{=} \widetilde{O}\left(N + \left(\frac{GR}{\epsilon}\right)^2\left(\frac{r_\epsilon}{R}\right)^{2/3}\right),$$

where $(\star)$ follows from substituting $\lambda_{\text{m}} = \widetilde{\Omega}\big(\epsilon r_\epsilon^{-4/3}R^{-2/3}\big)$ and $m = \widetilde{O}(1)$. Multiplying this bound by $\widetilde{O}\big((R/r_\epsilon)^{2/3}\big)$ yields (up to polylogarithmic factors) the conclusion of Theorem 1.

3.3 Accelerated variance reduction for mean-square smooth losses

In this section we provide an algorithm with an improved rate of convergence under the following mean-square smoothness assumption.

Assumption 2. *For all $x, x' \in \mathbb{B}_r(\bar{x})$ and $i \in [M]$, $\mathbb{E}_{j\sim w_i}\|\nabla\ell_j(x) - \nabla\ell_j(x')\|^2 \le L^2\|x - x'\|^2$.*

Note that assuming L-Lipschitz gradient for each ℓ_i implies Assumption 2, but not the other way around. To take advantage of Assumption 2, we first rewrite the function $\Gamma_{\epsilon,\lambda}(x)$ in a way that is more amenable to variance reduction:

$$\Gamma_{\epsilon,\lambda}(x) \coloneqq \sum_{i\in[M]} c_{x',\bar{x}} p_i(x')\gamma_i(x,x'), \ \text{ where } \ \gamma_i(x,x') \coloneqq \epsilon' e^{\frac{\mathcal{L}_i(x) - \mathcal{L}_i(x') + \frac{\lambda}{2}\|x-\bar{x}\|^2}{\epsilon'}},$$

$$c_{x',\bar{x}} = \left(\frac{\sum_{j\in[M]} e^{\frac{\mathcal{L}_j(x')}{\epsilon'}}}{\sum_{j\in[M]} e^{\frac{\mathcal{L}_j(\bar{x})}{\epsilon'}}}\right) \ \text{ and } \ p_i(x') \coloneqq \frac{e^{\frac{\mathcal{L}_i(x')}{\epsilon'}}}{\sum_{j\in[M]} e^{\frac{\mathcal{L}_j(x')}{\epsilon'}}}.$$

(Note that $\gamma_i(x,\bar{x}) = \gamma_i(x)$). Given a reference point x', to compute a reduced-variance estimator of $\nabla\Gamma_{\epsilon,\lambda}(x)$, we draw $i \sim p_i(x')$ and $j \sim w_i$, and set:

$$\hat{g}_{x'}(x) \coloneqq \nabla\Gamma_{\epsilon,\lambda}(x') + \frac{c_{x',\bar{x}}}{\epsilon'}\Big[\widehat{\mathcal{M}}[\gamma_i(x,x')]\nabla\ell_j^\lambda(x) - \gamma_i(x',x')\nabla\ell_j^\lambda(x')\Big] \tag{9}$$

where $\nabla \ell_j^\lambda(x) \coloneqq \nabla \ell_j(x) + \lambda(x - \bar{x})$ and $\widehat{\mathcal{M}}[\gamma_i(x, x')]$ is an MLMC estimator for $\gamma_i(x, x')$ defined analogously to $\widehat{\mathcal{M}}[\gamma_i(x)]$ (see details in Appendix E.4). The estimator (9) is not precisely standard SVRG [27] since we use $\widehat{\mathcal{M}}[\gamma_i(x, x')]$ as an estimator for $\gamma_i(x, x')$. Simple calculations show that $\mathbb{E}\hat{g}_{x'}(x) = \nabla \Gamma_{\epsilon,\lambda}(x)$ and the following lemma shows that $\hat{g}$ satisfies a type of variance bound conducive to variance-reduction schemes; see Appendix E.4 for the proof.

Lemma 3. *Let each ℓ_j satisfy Assumptions 1 and 2. For any $\lambda \le \frac{G}{r}$, $r = \frac{\epsilon'}{G}$ and $x, x' \in \mathbb{B}_r(\bar{x})$, the variance of $\hat{g}_{x'}(x)$ is bounded by* $\mathrm{Var}(\hat{g}_{x'}(x)) \le O\Big(\big(L + \lambda + \frac{G^2}{\epsilon'}\big)^2 \|x - x'\|^2\Big)$.

Accelerated variance reduction methods for convex functions typically require a stronger variance bound of the form $\mathrm{Var}(\hat{g}_{x'}(x)) \le 2L(F(x') - F(x) - \langle \nabla F(x), x' - x\rangle)$ for every x [cf. 1, Lemma 2.4]. The guarantee of Lemma 3 is weaker, but still allows for certain accelerated rates via, e.g., the Katyusha X algorithm [2]. With it, we obtain the following guarantee.

Theorem 2. *Let each ℓ_j satisfy Assumptions 1 and 2. Let $\epsilon > 0$, $\epsilon' = \epsilon/(2\log M)$ and $r_\epsilon = \epsilon'/G$. For any query point $\bar{x} \in \mathbb{R}^d$, regularization strength $\lambda \le O(G/r_\epsilon)$ and accuracy δ, KatyushaX^s [2, Algorithm 2] with the gradient estimator* (9) *outputs a valid r_ϵ-BROO response and has complexity $\mathcal{C}_\lambda(\delta) = O\Big(\Big(N + \frac{N^{3/4}(G+\sqrt{\epsilon' L})}{\sqrt{\lambda\epsilon'}}\Big)\log\big(\frac{G r_\epsilon}{\lambda\delta^2}\big)\Big)$. Consequently, the complexity of finding an ϵ-suboptimal minimizer of $\mathcal{L}_{\text{g-DRO}}$* (2) *with probability at least $\frac{1}{2}$ is*

$$O\left(N\left(\frac{GR}{\epsilon}\right)^{2/3}\log^{14/3} H + N^{3/4}\left(\frac{GR}{\epsilon} + \sqrt{\frac{LR^2}{\epsilon}}\right)\log^{7/2} H\right) \quad \textit{where} \quad H \coloneqq M\frac{GR}{\epsilon}.$$

We provide the proof of Theorem 2 in Appendix E.5. For the special case of Group DRO with a single group satisfying Assumption 2 with $L = \Theta\big(G^2/\epsilon\big)$, i.e. minimizing the average loss, we have the lower bound $\widetilde{\Omega}\big(N + N^{3/4}\frac{GR}{\epsilon}\big)$ [55] and for the case of N distinct groups, i.e. minimizing the maximal loss, we have the lower bound $\widetilde{\Omega}\big(N\epsilon^{-2/3}\big)$ [11]. This implies that in the weakly mean-square smooth setting the term scaling as $N^{3/4}\epsilon^{-1}$ and the term scaling as $N\epsilon^{-2/3}$ are unimprovable.

4 DRO with f-divergence

In this section we develop our BROO implementation for the f-divergence objective (3). In Section 4.1 we reduce the original DRO problem to a regularized form using Lagrange multipliers. Next, in Section 4.2 we show that adding negative entropy regularization to the objective produces the stability properties necessary for efficient ball optimization. In Section 4.3 we describe a BROO implementation for the non-smooth case using a variant of Epoch-SGD [22], and in Section 4.4 we implement the BROO under a weak-smoothness assumption by carefully restarting an accelerated variance reduction method [1].

4.1 The dual problem

We first note that, by Lagrange duality, the objective (3) is equivalent to

$$\mathcal{L}_{f\text{-div}}(x) \coloneqq \max_{q \in \Delta^N : \sum_{i\in[N]} \frac{f(Nq_i)}{N} \le 1} \sum_{i\in[N]} q_i \ell_i(x) = \min_{\nu \ge 0}\left\{\nu + \max_{q\in\Delta^N}\sum_{i\in[N]}\left(q_i\ell_i(x) - \frac{\nu}{N} f(Nq_i)\right)\right\}.$$

Writing $\psi(x) \coloneqq \frac{\nu}{N} f(Nx)$, we therefore consider objectives of the form

$$\mathcal{L}_\psi(x) \coloneqq \max_{q\in\Delta^N}\sum_{i\in[N]}(q_i\ell_i(x) - \psi(q_i)) = \min_{y\in\mathbb{R}}\left\{\Upsilon(x, y) \coloneqq \sum_{i\in[N]}\psi^*(\ell_i(x) - Gy) + Gy\right\} \quad (10)$$

where for the last equality we again used Lagrange duality, with $\psi^*(v) \coloneqq \max_{t\in\mathrm{dom}(\psi)}\{vt - \psi(t)\}$ the Fenchel dual of ψ (for more details see Appendix F.1). We show that under weak assumptions (introducing logarithmic dependence on bounds on f and the losses) we can solve the constrained problem (3) to accuracy ϵ by computing a polylogarithmic number of $O(\epsilon)$-accurate minimizers of (10); see Appendix F.2 for details. Thus, for the remainder of this section we focus on minimizing $\mathcal{L}_\psi$ for arbitrary convex $\psi : \mathbb{R}_+ \to \mathbb{R}$.

4.2 Stabilizing the gradient estimator

While minimizing (10) can be viewed as ERM (over x and y), straightforward application of SGD does not solve it efficiently. To see this, consider the standard gradient estimator formed by sampling $i \sim \mathsf{Unif}([N])$ and taking $\hat{g}^{\mathrm{x}} = N\psi^{*\prime}(\ell_i(x) - Gy)\nabla\ell_i(x)$ and $\hat{g}^{\mathrm{y}} = G\big(1 - N\psi^{*\prime}(\ell_i(x) - Gy)\big)$. For general ψ, this estimator will have unbounded second moments, and therefore SGD using them would lack a convergence guarantee. As an extreme example, consider $\psi = 0$ (corresponding to minimizing the maximum loss) whose conjugate function $\psi^*(v)$ is 0 for $v \le 0$ and ∞ for $v > 0$, leading to meaningless stochastic gradients.

We obtain bounded gradient estimates in two steps. First, we find a better distribution for i using a reference point $\bar{x} \in \mathcal{X}$ with corresponding $\bar{y} = \operatorname{argmin}_{y\in\mathbb{R}} \Upsilon(\bar{x}, y)$. Namely, we note that the optimality condition for $\bar{y}$ implies that $\psi^{*\prime}(\ell_i(\bar{x}) - \bar{y})$ is a pmf over $[N]$. Therefore, we may sample $i \sim \psi^{*\prime}(\ell_i(\bar{x}) - \bar{y})$ and estimate the gradient of Υ at (x, y) using $\hat{g}^{\mathrm{x}} = \rho_i(x, y)\nabla\ell_i(x)$ and $\hat{g}^{\mathrm{y}} = G\big(1 - \rho_i(x, y)\big)$, where $\rho_i(x, y) = \frac{\psi^{*\prime}(\ell_i(x)-y)}{\psi^{*\prime}(\ell_i(\bar{x})-\bar{y})}$. However, for general ψ (and $\psi = 0$ in particular), the ratio $\rho_i(x, y)$ can be unbounded even when x, y are arbitrarily close to $\bar{x}, \bar{y}$.

Our second step ensures that $\rho_i(x, y)$ is bounded around $\bar{x}, \bar{y}$ by adding a small negative entropy term to ψ, defining

$$\psi_\epsilon(q) \coloneqq \psi(q) + \epsilon' q \log q \;\text{ where }\; \epsilon' \coloneqq \frac{\epsilon}{2\log N}, \tag{11}$$

and

$$\mathcal{L}_{\psi,\epsilon}(x) = \min_{y\in\mathbb{R}} \Upsilon_\epsilon(x, y) \quad \text{with} \quad \Upsilon_\epsilon(x, y) \coloneqq \sum_{i\in[N]} \psi^*_\epsilon(\ell_i(x) - Gy) + Gy. \tag{12}$$

Due to our choice of ϵ', we have $|\mathcal{L}_\psi(x) - \mathcal{L}_{\psi,\epsilon}(x)| \le \epsilon/2$ for all $x \in \mathbb{R}^d$, and consequently an $\epsilon/2$-accurate minimizer of $\mathcal{L}_{\psi,\epsilon}$ is also an ϵ-accurate for $\mathcal{L}_\psi$ (see Lemma 18 in Appendix F.3). When $\psi = 0$ we have $\psi^*_\epsilon(v) = e^{(v-1)/\epsilon'}$ and therefore the corresponding $\rho_i(x, y) = e^{(\ell_i(x)-\ell_i(\bar{x})-G(y-\bar{y}))/\epsilon'}$.The following lemma, which might be of independent interest, shows that the same conclusion holds for any convex ψ.

Lemma 4. *For any convex $\psi : \mathbb{R}_+ \to \mathbb{R}$ and ψ_ϵ defined in* (10)*, $\log\big(\psi^{*\prime}_\epsilon(\cdot)\big)$ is $\frac{1}{\epsilon'}$-Lipschitz.*

See proof in Appendix F.4. Thus, $\psi^{*\prime}_\epsilon(v)/\psi^{*\prime}_\epsilon(\bar{v}) = e^{\log\psi^{*\prime}_\epsilon(v)-\log\psi^{*\prime}_\epsilon(\bar{v})} \le e^{(v-\bar{v})/\epsilon'}$ and $\rho_i(x, y) \le e^{(\ell_i(x)-\ell_i(\bar{x})-G(y-\bar{y}))/\epsilon'}$ continues to hold. Therefore, if $|y - \bar{y}| \le \epsilon'/G = r_\epsilon$ and $x \in \mathbb{B}_{r_\epsilon}(\bar{x})$ (so that $|\ell_i(x) - \ell_i(\bar{x})| \le \epsilon'$ if ℓ_i satisfies Assumption 1), we have the bound $\rho_i(x, y) \le e^2$.

It remains to show that we may indeed restrict y to be within distance r_ϵ from $\bar{y}$. To this end, we make the following observation which plays a key part in our analysis and might also be of independent interest (see proof in Appendix F.4).

Lemma 5. *For $y^\star(x) = \operatorname{argmin}_{y\in\mathbb{R}} \Upsilon_\epsilon(x, y)$, we have $|y^\star(x) - y^\star(x')| \le \frac{1}{G}\|\ell(x) - \ell(x')\|_\infty$ for all $x, x' \in \mathcal{X}$. Moreover, if each ℓ_i is G-Lipschitz, we have $|y^\star(x) - y^\star(x')| \le \|x - x'\|$.*

Lemma 5 implies that $x^\star, y^\star = \operatorname{argmin}_{x\in\mathbb{B}_{r_\epsilon}(\bar{x}),y\in\mathbb{R}} \Upsilon_\epsilon(x, y)$ satisfy $|y^\star - \bar{y}| \le \|x^\star - \bar{x}\| \le r_\epsilon$. Therefore, when minimizing Υ_ϵ (or any regularized version of it) inside the ball $\mathbb{B}_{r_\epsilon}(\bar{x})$, we may restrict y to $[\bar{y} - r_\epsilon, \bar{y} + r_\epsilon]$ without loss of generality. We also note that Lemma 5 holds for all values of ϵ and is therefore valid even without entropy regularization.

4.3 BROO implementation for f-divergence DRO with non-smooth losses

By the discussion above, to implement a BROO for $\mathcal{L}_{\psi,\epsilon}(x)$ (with radius $r_\epsilon = \epsilon'/G$, regularization λ, and query $\bar{x} \in \mathcal{X}$) it suffices to minimize $\Upsilon_{\epsilon,\lambda}(x, y) \coloneqq \Upsilon_\epsilon(x, y) + \frac{\lambda}{2}\|x - \bar{x}\|^2$ over $x \in \mathbb{B}_{r_\epsilon}(\bar{x})$ and $y \in [\bar{y} - r_\epsilon, \bar{y} + r_\epsilon]$, where $\bar{y} = \operatorname{argmin}_{y\in\mathbb{R}} \Upsilon_\epsilon(\bar{x}, y)$. To that end we estimate the gradient of $\Upsilon_{\epsilon,\lambda}(x, y)$ as follows. Letting $\bar{p}_i = \psi^{*\prime}_\epsilon(\ell_i(\bar{x}) - G\bar{y})$ (making $\bar{p}$ a pmf by optimality of $\bar{y}$), we sample $i \sim \bar{p}$ and set

$$\hat{g}^{\mathrm{x}}(x, y) = \frac{\psi^{*\prime}_\epsilon(\ell_i(x) - Gy)}{\bar{p}_i}\nabla\ell_i(x, y) \;\text{ and }\; \hat{g}^{\mathrm{y}}(x, y) = G\left(1 - \frac{\psi^{*\prime}_\epsilon(\ell_i(x) - Gy)}{\bar{p}_i}\right). \tag{13}$$

Lemma 4 implies the following bounds on our gradient estimator; see proof in Appendix F.5.

Lemma 6. *Let each ℓ_i be G-Lipschitz, let $\bar{x} \in \mathcal{X}$ and $\bar{y} = \operatorname{argmin}_{y\in\mathbb{R}} \Upsilon_{\epsilon,\lambda}(\bar{x}, y)$. Let $r_\epsilon = \frac{\epsilon'}{G}$, then for all $x \in \mathbb{B}_{r_\epsilon}(\bar{x})$ and $y \in [\bar{y} - r_\epsilon, \bar{y} + r_\epsilon]$, the gradient estimators $\hat{g}^{\mathrm{x}}$ and $\hat{g}^{\mathrm{y}}$ satisfy the following:*

1. $\mathbb{E}_{i\sim\bar{p}_i}[\hat{g}^{\mathrm{x}}(x, y)] = \nabla_x \Upsilon_\epsilon(x, y)$ *and* $\mathbb{E}_{i\sim\bar{p}_i}[\hat{g}^{\mathrm{y}}(x, y)] = \nabla_y \Upsilon_\epsilon(x, y)$.
2. $\mathbb{E}_{i\sim\bar{p}_i}\|\hat{g}^{\mathrm{x}}(x, y)\|^2 \le e^4 G^2$ *and* $\mathbb{E}_{i\sim\bar{p}_i}|\hat{g}^{\mathrm{y}}(x, y)|^2 \le e^4 G^2$.

To implement the BROO using our gradient estimator we develop a variant of the Epoch-SGD algorithm of Hazan and Kale [22] (Algorithm 3 in Appendix F.5). Similarly to Epoch-SGD, we apply standard SGD on $\Upsilon_{\epsilon,\lambda}$ (with gradient estimator (13)) in "epochs" whose length doubles in every repetition. Our algorithm differs slightly in how each epoch is initialized. Standard Epoch-SGD initializes with the average of the previous epoch's iterates, and strong convexity shows that the suboptimality and distance to the optimum shrink by a constant factor after every epoch. However, since $\Upsilon_{\epsilon,\lambda}$ is strongly convex only in x and not in y, we cannot directly use this scheme. Instead, we set the initial y variable to be $\operatorname{argmin}_y \Upsilon_{\epsilon,\lambda}(x', y)$, where x' is the initial x variable still defined as the previous epoch's average; this initialization has complexity N, but we only preform it a logarithmic number of times. Using our initialization scheme and Lemma 5, we recover the original Epoch-SGD contraction argument, yielding the following complexity bound (see proof in Appendix F.5).

Theorem 3. *Let each ℓ_i satisfy Assumption 1. Let $\epsilon, \lambda, \delta > 0$, and $r_\epsilon = \epsilon/(2G\log N)$. For any query point $\bar{x} \in \mathbb{R}^d$, regularization strength $\lambda \le O(G/r_\epsilon)$ and accuracy $\delta < r_\epsilon/2$, Algorithm 3 outputs a valid r_ϵ-BROO response for $\mathcal{L}_{\psi,\epsilon}$ and has complexity $\mathcal{C}_\lambda(\delta) = O\big(\frac{G^2}{\lambda^2\delta^2} + N\log\big(\frac{r_\epsilon}{\delta}\big)\big)$. Consequently, the complexity of finding an ϵ-suboptimal minimizer of $\mathcal{L}_\psi$ (10) with probability at least $\frac{1}{2}$ is*

$$O\left(N\left(\frac{GR}{\epsilon}\right)^{2/3}\log^{11/3} H + \left(\frac{GR}{\epsilon}\right)^2 \log^2 H\right) \text{ where } H \coloneqq N\frac{GR}{\epsilon}.$$

4.4 Accelerated variance reduction for smooth losses

In this section, we take advantage of the following smoothness assumption.

Assumption 3. *For every $i \in [N]$ the loss ℓ_i is L-smooth, i.e., has L-Lipschitz gradient.*

Let us rewrite $\Upsilon_{\epsilon,\lambda}$ in a form that is more amenable to variance reduction techniques:

$$\Upsilon_{\epsilon,\lambda}(x, y) = \sum_{i\in[N]} \bar{p}_i \upsilon_i(x, y) \text{ where } \upsilon_i(x, y) \coloneqq \frac{\psi_\epsilon^*(\ell_i(x) - Gy)}{\bar{p}_i} + Gy + \frac{\lambda}{2}\|x - \bar{x}\|^2$$

and, as before $\bar{p}_i = \psi_\epsilon^*(\ell_i(\bar{x}) - G\bar{y})$ for some ball center $\bar{x} \in \mathcal{X}$ and $\bar{y} = \operatorname{argmin}_{y\in\mathbb{R}} \Upsilon_\epsilon(\bar{x}, y)$. In the following lemma, we bound the smoothness of the functions υ_i, deferring the proof to Appendix F.6.

Lemma 7. *For any $i \in [N]$, let ℓ_i be G-Lipschitz and L -smooth, let $r_\epsilon = \frac{\epsilon'}{G}$ and $\lambda = O\big(\frac{G}{r_\epsilon}\big)$. The restriction of υ_i to $x \in \mathbb{B}_{r_\epsilon}(\bar{x})$ and $y \in [\bar{y} - r_\epsilon, \bar{y} + r_\epsilon]$ is $O(G)$-Lipschitz and $O\big(L + \frac{G^2}{\epsilon'}\big)$-smooth.*

Since $\Upsilon_{\epsilon,\lambda}$ is a finite sum of smooth functions, we can obtain reduced-variance gradient estimates by the standard SVRG technique [27]. For any reference point x', y', the estimator is

$$\hat{g}_{x',y'}(x, y) = \nabla\Upsilon_\epsilon(x', y') + \nabla\upsilon_i(x, y) - \nabla\upsilon_i(x', y'), \tag{14}$$

where ∇ is with respect to the vector $[x, y]$. Similar to non-smooth case, obtaining an efficient BROO implementation is complicated by the fact that $\Upsilon_{\epsilon,\lambda}$ is strongly-convex in x but not in y. Our solution is also similar: we propose a restart scheme and minimize over y exactly between restarts (Algorithm 4 in Appendix F.6), that gives the following complexity bound (see proof in Appendix F.6).

Theorem 4. *Let each ℓ_i satisfy Assumptions 1 and 3, let $\epsilon, \lambda, \delta > 0$, and $r_\epsilon = \frac{\epsilon}{2G\log N}$. For any query point $\bar{x} \in \mathbb{R}^d$, regularization strength $\lambda \le O(\frac{G}{r_\epsilon})$ and accuracy δ, Algorithm 4 outputs a valid r_ϵ-BROO response for $\mathcal{L}_{\psi,\epsilon}$ and has complexity $\mathcal{C}_\lambda(\delta) = O\Big(\Big(N + \frac{\sqrt{N}(G+\sqrt{\epsilon' L})}{\sqrt{\lambda\epsilon'}}\Big)\log\frac{Gr_\epsilon}{\lambda\delta^2}\Big)$. Consequently, the complexity of finding an ϵ-suboptimal minimizer of $\mathcal{L}_\psi$ (10) with probability at least $\frac{1}{2}$ is*

$$O\left(N\left(\frac{GR}{\epsilon}\right)^{2/3}\log^{14/3} H + \sqrt{N}\left(\frac{GR}{\epsilon} + \sqrt{\frac{LR^2}{\epsilon}}\right)\log^{5/2} H\right) \text{ where } H \coloneqq N\frac{GR}{\epsilon}.$$

References

[1] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2017.

[2] Z. Allen-Zhu. Katyusha X: Practical momentum method for stochastic sum-of-nonconvex optimization. In *International Conference on Machine Learning (ICML)*, 2018.

[3] H. Asi, Y. Carmon, A. Jambulapati, Y. Jin, and A. Sidford. Stochastic bias-reduced gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

[5] 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)*, 2015.

[6] J. H. Blanchet, P. W. Glynn, and Y. Pei. Unbiased multilevel Monte Carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv:1904.09929*, 2019.

[7] 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.

[8] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International Conference on Machine Learning (ICML)*, 2017.

[9] Y. Carmon, A. Jambulapati, Q. Jiang, Y. Jin, Y. T. Lee, A. Sidford, and K. Tian. Acceleration with a ball optimization oracle. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] Y. Carmon, Y. Jin, A. Sidford, and K. Tian. Coordinate methods for matrix games. In *Foundations of Computer Science (FOCS)*, 2020.

[11] Y. Carmon, A. Jambulapati, Y. Jin, and A. Sidford. Thinking inside the ball: Near-optimal minimization of the maximal loss. In *Conference on Learning Theory (COLT)*, 2021.

[12] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: a CVaR optimization approach. In *Advances in neural information processing systems (NeurIPS)*, 2015.

[13] M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford. Geometric median in nearly linear time. In *ACM symposium on Theory of Computing*, 2016.

[14] S. Curi, K. Y. Levy, S. Jegelka, and A. Krause. Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

[16] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[17] 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):115–166, 2018.

[18] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv:1604.02199*, 2016.

[19] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, C. A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck, Q. Jiang, Y. T. Lee, Y. Li, and A. Sidford. Near optimal methods for minimizing convex functions with Lipschitz p-th derivatives. In *Conference on Learning Theory (COLT)*, 2019.

[20] M. B. Giles. Multilevel monte carlo path simulation. *Operations research*, 56(3):607–617, 2008.

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

[22] E. Hazan and S. Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(71):2489–2512, 2014.

[23] S. Heinrich. Multilevel Monte Carlo methods. In *International Conference on Large-Scale Scientific Computing*, 2001.

[24] D. Hovy and A. Søgaard. Tagging performance correlates with author age. In *Association for Computational Linguistics (ACL)*, pages 483–488, 2015.

[25] W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, 2018.

[26] J. Jin, B. Zhang, H. Wang, and L. Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[27] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems (NeurIPS)*, 2013.

[28] C. Kent, J. Blanchet, and P. Glynn. Frank-wolfe methods in probability space. *arXiv:2105.05352*, 2021.

[29] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

[30] P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[31] G. Lan, Z. Li, and Y. Zhou. A unified variance-reduced accelerated gradient method for convex optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[32] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[33] F. Lin, X. Fang, and Z. Gao. Distributionally robust optimization: a review on theory and applications. *Numerical Algebra, Control & Optimization*, 12(1):159, 2022.

[34] R. D. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

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

[36] 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.

[37] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming, Series A*, 103(1):127–152, 2005.

[38] Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[39] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *ACM conference on health, inference, and learning*, 2020.

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

[41] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2: 21–41, 2000.

[42] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

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

[44] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018.

[45] C. Song, Y. Jiang, and Y. Ma. Unified acceleration of high-order algorithms under Hölder continuity and uniform convexity. *SIAM journal of optimization*, 2021.

[46] C. Song, C. Y. Lin, S. J. Wright, and J. Diakonikolas. Coordinate linear variance reduction for generalized linear programming. *arXiv:2111.01842*, 2021.

[47] N. A. Urpí, S. Curi, and A. Krause. Risk-averse offline reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2021.

[48] K. A. Wang, N. S. Chatterji, S. Haque, and T. Hashimoto. Is importance weighting incompatible with interpolating classifiers? In *International Conference on Learning Representations*, 2022.

[49] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. R. Gupta, and M. I. Jordan. Robust optimization for fairness with noisy protected groups. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[50] J. Wen, C.-N. Yu, and R. Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning (ICML)*, pages 631–639, 2014.

[51] B. E. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in neural information processing systems (NeurIPS)*, 2016.

[52] R. Zhai, C. Dan, Z. Kolter, and P. Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning (ICML)*, 2021.

[53] R. Zhai, C. Dan, A. Suggala, J. Z. Kolter, and P. Ravikumar. Boosted CVaR classification. In *Advances in neural information processing systems (NeurIPS)*, 2021.

[54] C. Zhou, D. Levy, X. Li, M. Ghazvininejad, and G. Neubig. Distributionally robust multilingual machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[55] D. Zhou and Q. Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning (ICML)*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Theorem 1, Theorem 2, Theorem 3 and Theorem 4.
   (b) Did you describe the limitations of your work? [Yes] See "Limitations" paragraph in Section 1.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumption 1 Assumption 2 and Assumption 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]