Efficient Algorithms for Empirical Group Distributionally Robust Optimization and Beyond

Dingzhi Yu [1 2] **Yunuo Cai** [2] **Wei Jiang** [1] **Lijun Zhang** [1 3]

Abstract

In this paper, we investigate the empirical counterpart of Group Distributionally Robust Optimization (GDRO), which aims to minimize the maximal empirical risk across m distinct groups. We formulate empirical GDRO as a *two-level* finite-sum convex-concave minimax optimization problem and develop an algorithm called ALEG to benefit from its special structure. ALEG is a double-looped stochastic primal-dual algorithm that incorporates variance reduction techniques into a modified mirror prox routine. To exploit the two-level finite-sum structure, we propose a simple group sampling strategy to construct the stochastic gradient with a smaller Lipschitz constant and then perform variance reduction for all groups. Theoretical analysis shows that ALEG achieves ε-accuracy within a computation complexity of $\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$, where $\bar{n}$ is the average number of samples among m groups. Notably, our approach outperforms the state-of-the-art method by a factor of $\sqrt{m}$. Based on ALEG, we further develop a two-stage optimization algorithm called ALEM to deal with the empirical Minimax Excess Risk Optimization (MERO) problem. The computation complexity of ALEM nearly matches that of ALEG, surpassing the rates of existing methods.

1. Introduction

Recently, a popular class of Distributionally Robust Optimization (DRO) problem named as Group DRO (GDRO), has drawn significant attention in machine learning (Oren et al., 2019; Mohri et al., 2019; Carmon & Hausler, 2022; Zhang et al., 2023; Mehta et al., 2024). GDRO optimizes the maximal risk over a group of m distributions, which can be formulated as the following stochastic minimax optimization problem:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \mathbb{E}_{\xi\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\xi)\right], \tag{1}$$

where $\{\mathcal{P}_i\}_{i\in[m]}$ is a group of distributions and $\ell(\mathbf{w};\xi)$ is the loss function, measuring the predictive error of the model $\mathbf{w}\in\mathcal{W}$ for a sample ξ. Similar to Empirical Risk Minimization (ERM) (Shalev-Shwartz & Ben-David, 2014), which replaces the risk minimization problem over a distribution by its empirical counterpart, this paper focuses on the empirical variant of problem (1):

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{R_i(\mathbf{w}) := \frac{1}{n_i}\sum_{j=1}^{n_i}\ell(\mathbf{w};\xi_{ij})\right\}, \tag{2}$$

where $\{R_i(\cdot)\}_{i\in[m]}$ are empirical risk functions for each group, n_i is the number of samples for group i, and ξ_{ij} is the j-th sample in group i. We further denote $\bar{n} := \frac{1}{m}\sum_{i=1}^{m} n_i$ as the average number of samples for m groups. Many popular machine learning tasks can be mathematically modeled by (empirical) GDRO, including federated learning (Mohri et al., 2019; Laguel et al., 2021a;b), robust language modeling (Oren et al., 2019), robust neural network training (Sagawa et al., 2020), and collaborative PAC learning (Blum et al., 2017; Nguyen & Zakynthinou, 2018; Rothblum & Yona, 2021).

To leverage the powerful first-order optimization methods, we transform (2) into an equivalent finite-sum convex-concave saddle-point problem (Nemirovski et al., 2009):

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{F(\mathbf{w},\mathbf{q}) := \sum_{i=1}^{m}\mathbf{q}_i R_i(\mathbf{w})\right\}, \tag{3}$$

where $\Delta_m = \{\mathbf{q}\geq\mathbf{0}, \mathbf{1}^T\mathbf{q} = 1 | \mathbf{q}\in\mathbb{R}^m\}$ is the $(m-1)$-dimensional simplex, and $\{R_i(\cdot)\}_{i\in[m]}$ are assumed to be convex.

The existing vast amount of work on general convex-concave optimization can be applied to (3). For instance,

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [2]School of Data Science, Fudan University, Shanghai, China [3]Pazhou Laboratory (Huangpu), Guangzhou, China. Correspondence to: Lijun Zhang <zhanglj@lamda.nju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Stochastic Mirror Descent (SMD) (Nemirovski et al., 2009), originally targeting problem (1), can be utilized for (3) with a complexity of $\mathcal{O}\left(\frac{m \ln m}{\varepsilon^2}\right)$. However, the quadratic dependency on ε prevents us from getting a high-accuracy solution. A recent work of Alacaoglu & Malitsky (2022) proposes Mirror Prox with Variance Reduction (MPVR) to solve the general finite-sum convex-concave optimization problem. Running MPVR on (3), we get an $\mathcal{O}\left(m\bar{n} + \frac{m\sqrt{m\bar{n}}\ln m}{\varepsilon}\right)$ complexity, but this rate is suboptimal for (3), as shown below.

Inspired by MPVR, this paper goes go one step further by exploiting the *two-level* structure of (3) to derive a better $\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$ complexity. We call (3) "two-level finite-sum" because $F(\mathbf{w}, \mathbf{q})$ can be decomposed into two nested finite summations. Similar to MPVR, we also incorporate variance reduction into the stochastic mirror prox algorithm, but with a novel way to construct the stochastic gradients. Specifically, we develop a simple yet effective group sampling technique, which uses m samples, one for each group, to construct the stochastic gradients, and in this way, the Lipschitz constant is reduced by a factor of $1/m$. Then, we employ variance reduction techniques within each group to attain faster convergence. Our algorithm, named as ALEG, improves the state-of-the-art computation complexity by $\sqrt{m}$. Another advantage of ALEG is its support for changeable hyperparameters, achieved by leveraging the one-index-shifted weighted average to compute (mirror) snapshot points and Lyapunov functions. This enhances ALEG's practical effectiveness, as variable learning rates have proven beneficial in real-world machine learning tasks (Liner & Miikkulainen, 2021).

Furthermore, we extend our methodology to solve the empirical Minimax Excess Risk Optimization (MERO) problem (Agarwal & Zhang, 2022) given below:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{\underline{R}_i(\mathbf{w}) := R_i(\mathbf{w}) - R_i^*\right\}, \tag{4}$$

where we define $R_i^* := \min_{\mathbf{w}\in\mathcal{W}} R_i(\mathbf{w})$ as the minimal empirical risk of group i, usually unknown. Compared with empirical GDRO, empirical MERO replaces the raw empirical risk $R_i(\mathbf{w})$ with the *excess empirical risk* $\underline{R}_i(\mathbf{w}) := R_i(\mathbf{w}) - R_i^*$ to cope with heterogeneous noise. We provide an efficient two-stage **Al**gorithm for **E**mpirical **M**ERO (ALEM) by following the two-stage schema (Zhang et al., 2024), and making use of our empirical GDRO algorithm. In the first stage, R_i^* is estimated by an approximate minimal empirical risk $\hat{R}_i^*$ through running ALEG for m groups. In the second stage, we approximate the original problem by replacing the minimal empirical risks $\{R_i^*\}_{i\in[m]}$ with the estimated values $\{\hat{R}_i^*\}_{i\in[m]}$, and optimize the new problem by ALEG. We demonstrate that the computation complexity of ALEM is $\tilde{\mathcal{O}}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$[1], which nearly matches that of the empirical GDRO problem.

2. Related Work

In this section, we review and compare some previous work that can be used to solve the empirical GDRO problem in (3) and the empirical MERO problem in (5).

2.1. Group Distributionally Robust Optimization

For the original GDRO, Nemirovski et al. (2009) propose a stochastic approximation (SA) (Robbins & Monro, 1951) approach named Stochastic Mirror Descent (SMD) and obtain a sample complexity of $\mathcal{O}\left(\frac{m \ln m}{\varepsilon^2}\right)$, which nearly matches the lower bound $\Omega\left(\frac{m}{\varepsilon^2}\right)$ (Soma et al., 2022) up to a logarithmic factor. Soma et al. (2022) cast problem (1) as a two-player zero-sum game to utilize SMD and no-regret algorithms from multi-armed bandits (MAB). Recently, Zhang et al. (2023) refine their analysis by borrowing techniques from non-oblivious MAB (Neu, 2015) to establish a sample complexity of $\mathcal{O}\left(\frac{m \ln m}{\varepsilon^2}\right)$.

Although SA approaches primarily aim at solving (1), they can also be directly applied to empirical GDRO. Running SMD or non-oblivious online learning algorithm on (3) yields a computation complexity of $\mathcal{O}\left(\frac{m \ln m}{\varepsilon^2}\right)$, which serves as the baseline of the problem. Although the complexity does not depend on the average number of samples $\bar{n}$, it suffers from a quadratic dependency on ε.

2.2. Empirical GDRO and Empirical MERO

Carmon & Hausler (2022) optimize (3) directly by leveraging exponentiated group-softmax and the Nesterov smoothing (Nesterov, 2005) technique. Then they run a Ball Regularized Optimization Oracle (Carmon et al., 2020; 2021) on Katyusha X (Allen-Zhu, 2018) (BROO-KX) to get a complexity of $\mathcal{O}\left(\frac{m\bar{n}}{\varepsilon^{2/3}} \ln^{14/3}(\frac{\ln m}{\varepsilon}) + \frac{(m\bar{n})^{3/4}}{\varepsilon} \ln^{7/2}(\frac{\ln m}{\varepsilon})\right)$. We note that their complexity is measured by the number of oracle queries, i.e., $\ell(\cdot;\xi_{ij})$ and $\nabla\ell(\cdot;\xi_{ij})$ evaluations, instead of total computations. In fact, the BROO-KX algorithm includes expensive bisections, which are computationally expensive.

As an extension of the empirical GDRO problem, Agarwal & Zhang (2022) study the empirical MERO problem. They cast (4) as a two-player zero-sum game:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{\underline{F}(\mathbf{w}, \mathbf{q}) := \sum_{i=1}^{m} \mathbf{q}_i \underline{R}_i(\mathbf{w})\right\}, \tag{5}$$

where $\underline{R}_i(\mathbf{w}) := R_i(\mathbf{w}) - R_i^*$ is the excess empirical risk. In

[1]We use the $\tilde{\mathcal{O}}$ notation to hide constant factors as well as logarithmic factors in ε.

Table 1: Comparisons of computation complexities for empirical GDRO.*

Algorithm	Computation Complexity
SMD (Nemirovski et al., 2009)	$\mathcal{O}\left(\frac{m\ln m}{\varepsilon^2}\right)$
AL-SVRE (Luo et al., 2021)	$\mathcal{O}\left(\left(m\bar{n}+\frac{m\sqrt{m\bar{n}\ln m}}{\varepsilon}+m^{5/4}\bar{n}^{3/4}\sqrt{\frac{\ln m}{\varepsilon}}\right)\ln\frac{m}{\varepsilon}\right)$
BROO-KX (Carmon & Hausler, 2022)	$\mathcal{O}\left(\frac{m\bar{n}}{\varepsilon^{2/3}}\ln^{14/3}(\frac{\ln m}{\varepsilon})+\frac{(m\bar{n})^{3/4}}{\varepsilon}\ln^{7/2}(\frac{\ln m}{\varepsilon})\right)^{**}$
MPVR (Alacaoglu & Malitsky, 2022)	$\mathcal{O}\left(m\bar{n}+\frac{m\sqrt{m\bar{n}\ln m}}{\varepsilon}\right)$
ALEG (Theorem 4.4 and Corollary 4.6)	$\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$

* The results are transformed in our formulation in terms of $m, \bar{n}$.
** BROO-KX relies on expensive bisection operations, which is not reflected in this complexity measure.

each iteration, the maximizing player calls the exponentiated gradient algorithm (Kivinen & Warmuth, 1997) to update $\mathbf{q}$ while the minimizing player calls an expensive ERM oracle to minimize $\underline{F}(\cdot, \mathbf{q}_t)$ for the current $\mathbf{q}_t$. Under the condition that the ERM oracle is perfect, their algorithm converges at a rate of $\mathcal{O}\left(\sqrt{\frac{\ln(m\bar{n})}{T}}\right)$. For simplicity, we assume the cost of the ERM oracle is proportional to the number of samples $m\bar{n}$. This leads to a computation complexity of $\mathcal{O}\left(\frac{m\bar{n}\ln(m\bar{n})}{\varepsilon^2}\right)$ in total.

2.3. Finite-Sum Convex-Concave Optimization

The general finite-sum convex-concave optimization problem is given by

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}}\left\{G(\mathbf{x},\mathbf{y}):=\sum_{i=1}^{n}G_i(\mathbf{x},\mathbf{y})\right\}, \tag{6}$$

where $G(\mathbf{x},\mathbf{y})$ is convex w.r.t. $\mathbf{x}$ and concave w.r.t. $\mathbf{y}$. Comparing with (3), we call (6) "*one-level* finite-sum" because the objective $G(\mathbf{x},\mathbf{y})$ is written as a single finite summation. Existing algorithms for this problem can also be applied to (3), but they suffer from a suboptimal complexity due to neglecting the special two-level finite-sum structure of $F(\mathbf{w},\mathbf{q})$. We review two representative works in the sequel.

Alacaoglu & Malitsky (2022) develop the Mirror Prox with Variance Reduction (MPVR) algorithm. To incorporate variance reduction, they modify the classic mirror prox by replacing a stochastic gradient at the last iterated point with a full gradient at the current snapshot. Moreover, they use "negative momentum" (Driggs et al., 2022) in the dual space to further accelerate the algorithm. MPVR can be applied to the empirical GDRO problem with its uniform sampling or importance sampling technique. However, both of them fail to capture the intrinsic two-level finite-sum structure in (3) and therefore suffer an additional $\sqrt{m}$ factor in the total complexity of $\mathcal{O}\left(m\bar{n}+\frac{m\sqrt{m\bar{n}\ln m}}{\varepsilon}\right)$.

Based on MPVR, Luo et al. (2021) use Accelerated Loopless Stochastic Variance-Reduced Extragradient (AL-SVRE) to tackle with the one-level finite-sum convex-concave problems. AL-SVRE uses MPVR as the subproblem solver and then conducts a catalyst acceleration scheme (Lin et al., 2015). Applying it to the empirical GDRO problem gives a complexity of $\mathcal{O}\left(\left(m\bar{n}+\frac{m\sqrt{m\bar{n}\ln m}}{\varepsilon}+m^{5/4}\bar{n}^{3/4}\sqrt{\frac{\ln m}{\varepsilon}}\right)\ln\frac{m}{\varepsilon}\right)$, which is also worse than our result by $\sqrt{m}$.

2.4. Complexity Comparisons

We summarize the existing results for empirical GDRO in Table 1. Under the circumstance of moderately high accuracy $\varepsilon \leq \mathcal{O}\left(\sqrt{\frac{\ln m}{\bar{n}}}\right)$, our $\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$ complexity beats the baseline $\mathcal{O}\left(\frac{m\ln m}{\varepsilon^2}\right)$ of SMD. AL-SVRE has a complexity whose dominating term $\mathcal{O}\left(\frac{m\sqrt{m\bar{n}\ln m}}{\varepsilon}\ln\frac{m}{\varepsilon}\right)$ is worse than us by a factor of $\sqrt{m}\ln\frac{m}{\varepsilon}$. BROO-KX provides an oracle complexity with a leading term of $\mathcal{O}\left(\frac{(m\bar{n})^{3/4}}{\varepsilon}\ln^{7/2}(\frac{\ln m}{\varepsilon})\right)$, which is less favorable than our approach under the common situations where $m \leq \mathcal{O}(\bar{n})$. MPVR exhibits a complexity of $\mathcal{O}\left(m\bar{n}+\frac{m\sqrt{m\bar{n}\ln m}}{\varepsilon}\right)$, which remains inferior to our approach by a factor of $\sqrt{m}$.

We also briefly discuss the empirical MERO algorithms. Agarwal & Zhang (2022) use Empirical Risk Minimization with Exponential Gradient algorithm (ERMEG) to optimize (5), thereby exhibiting a complexity of

$\mathcal{O}\left(\frac{m\bar{n}\ln(m\bar{n})}{\varepsilon^2}\right)$. Note that this result underestimates the optimization complexity of ERM oracles and thus is excessively conservative. Zhang et al. (2024) develop a two-stage stochastic approximation (TSA) approach to target the population MERO. TSA can also be utilized in the empirical scenario, yielding a computation complexity of $\mathcal{O}\left(\frac{m\ln m}{\varepsilon^2}\right)$. We observe that both methods suffer from the quadratic dependency on ε. Based on our ALEG algorithm, we further develop a two-stage **Al**gorithm for **E**mpirical **M**ERO (ALEM), which attains an $\tilde{\mathcal{O}}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$ complexity, outperforming existing algorithms.

3. Preliminaries

In this section, we present notations, definitions, assumptions, and the Bregman setup used in the paper.

3.1. Notations

Denote by $\|\cdot\|_x$ a general norm on finite dimensional Banach space $\mathcal{E}_x$ and its dual norm $\|\mathbf{x}\|_{x,*} = \sup_{\mathbf{y}\in\mathcal{E}_x}\{\langle\mathbf{x},\mathbf{y}\rangle \mid \|\mathbf{y}\|_x \le 1\}$. We use $[S] = \{1,2,\cdots,S\}$ and $[S]^0 = \{0,1,\cdots,S-1\}$ for some positive integer S. We denote $\mathcal{W}\times\Delta_m$ by $\mathcal{Z}$. In view of $\mathbf{z} = (\mathbf{w};\mathbf{q}) \in \mathcal{Z}$ as the concatenation of $\mathbf{w}$ and $\mathbf{q}$, we use $F(\mathbf{z}) = F(\mathbf{w},\mathbf{q})$ and $\nabla F(\mathbf{z}) = (\nabla_{\mathbf{w}}F(\mathbf{w},\mathbf{q}); -\nabla_{\mathbf{q}}F(\mathbf{w},\mathbf{q}))$ to denote the merged function value and the merged gradient, respectively.

3.2. Definitions and Assumptions

For mirror descent (Beck & Teboulle, 2003) type of primal-dual methods, we need to construct the distance-generating function and the corresponding Bregman divergence.

Definition 3.1. We call a continuous function $\psi_x : X \mapsto \mathbb{R}$ a distance-generating function modulus α_x w.r.t. $\|\cdot\|_x$, if (i) the set $X^o = \{\mathbf{x} \in X | \partial\psi_x(\mathbf{x}) \neq 0\}$ is convex; (ii) ψ_x is continuously differentiable and α_x-strongly convex w.r.t. $\|\cdot\|_x$, i.e., $\langle\nabla\psi_x(x_1) - \nabla\psi_x(x_2), x_1 - x_2\rangle \ge \alpha_x \|x_1 - x_2\|_x^2$, $\forall x_1, x_2 \in X^o$.

Definition 3.2. Define Bregman function $B_x : X \times X^o \mapsto \mathbb{R}_+$ associated with distance-generating function ψ_x as

$$B_x(x, x^o) = \psi_x(x) - \psi_x(x^o) - \langle\nabla\psi_x(x^o), x - x^o\rangle. \quad (7)$$

We equip $\mathcal{W}$ with a distance-generating function $\psi_w(\cdot)$ modulus α_w w.r.t. a norm $\|\cdot\|_w$ endowed on $\mathcal{E}$. Similarly, we have $\psi_q(\cdot)$ modulus α_q w.r.t. $\|\cdot\|_q$. The choice of such ψ_x and $\|\cdot\|_x$ should rely on the geometric structure of our domain. In this paper, we stick to $\psi_q(\mathbf{q}) = \sum_{i=1}^m \mathbf{q}_i \ln \mathbf{q}_i$ as the entropy function, which is 1-strongly convex w.r.t $\|\cdot\|_1$.

The following assumptions will be used in our Bregman setup analysis, which are commonly adopted in the existing literature (Nemirovski et al., 2009).

Assumption 3.3. (Boundness on the domain) The domain $\mathcal{W}$ is convex and its diameter measured by $\psi_w(\cdot)$ is bounded by a positive constant D_w, i.e.

$$\max_{\mathbf{w}\in\mathcal{W}}\psi_w(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{W}}\psi_w(\mathbf{w}) \le D_w^2. \quad (8)$$

Similarly, we assume Δ_m is bounded by D_q. Since ψ_q is specified as the entropy function, we have $D_q = \sqrt{\ln m}$.

Assumption 3.4. (Smoothness and Lipschitz continuity) For any $i \in [m]$, $j \in [n_i]$, $\ell(\cdot;\xi_{ij})$ is L-smooth and G-Lipschitz continuous.

Remark 3.5. In the context of stochastic convex optimization, smoothness is of the essence to obtain a variance-based convergence rate (Lan, 2012).

Assumption 3.6. (Convexity) For every $i \in [m]$, empirical risk function $R_i(\cdot)$ is convex.

Remark 3.7. Our convexity assumption is *weaker* than Luo et al. (2021) and Carmon & Hausler (2022), due to not requiring each component loss function $\ell(\cdot;\xi_{ij})$ to be convex.

3.3. Bregman Setups

We endow the Cartesian product space $\mathcal{E}\times\mathbb{R}^m$ and its dual space $\mathcal{E}^*\times\mathbb{R}^m$ with the following norm and dual norm (Nemirovski et al., 2009). For any $\mathbf{z} = (\mathbf{w};\mathbf{q}) \in \mathcal{E}\times\mathbb{R}^m$ and any $\mathbf{z}^* = (\mathbf{w}^*;\mathbf{q}^*) \in \mathcal{E}^*\times\mathbb{R}^m$,

$$\begin{aligned}\|\mathbf{z}\| &:= \sqrt{\frac{\alpha_w}{2D_w^2}\|\mathbf{w}\|_w^2 + \frac{\alpha_q}{2D_q^2}\|\mathbf{q}\|_q^2},\\ \|\mathbf{z}^*\|_* &:= \sqrt{\frac{2D_w^2}{\alpha_w}\|\mathbf{w}^*\|_{w,*}^2 + \frac{2D_q^2}{\alpha_q}\|\mathbf{q}^*\|_{q,*}^2}.\end{aligned} \quad (9)$$

The corresponding distance-generating function has the following form:

$$\psi(\mathbf{z}) := \frac{1}{2D_w^2}\psi_w(\mathbf{w}) + \frac{1}{2D_q^2}\psi_q(\mathbf{q}). \quad (10)$$

It's easy to verify that $\psi(\mathbf{z})$ is 1-strongly w.r.t. the norm $\|\cdot\|$ in (9). So now we can define the Bregman divergence $B : \mathcal{Z}\times\mathcal{Z}^o \mapsto \mathbb{R}_+$ used in our algorithm:

$$B(\mathbf{z},\mathbf{z}^o) := \psi(\mathbf{z}) - \psi(\mathbf{z}^o) - \langle\nabla\psi(\mathbf{z}^o), \mathbf{z}-\mathbf{z}^o\rangle. \quad (11)$$

To analyze the quality of an approximate solution, we adopt a commonly used performance measure in existing literature (Luo et al., 2021; Carmon & Hausler, 2022; Zhang et al., 2023), known as the duality gap of any given $\bar{\mathbf{z}} = (\bar{\mathbf{w}};\bar{\mathbf{q}})$ for (3):

$$\epsilon(\bar{\mathbf{z}}) := \max_{\mathbf{q}\in\Delta_m} F(\bar{\mathbf{w}},\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w},\bar{\mathbf{q}}). \quad (12)$$

We aim to find a solution $\bar{\mathbf{z}} = (\bar{\mathbf{w}};\bar{\mathbf{q}})$ that is ε-accuracy of (3), i.e., $\epsilon(\bar{\mathbf{z}}) \le \varepsilon$. It's obvious that $\epsilon(\bar{\mathbf{z}})$ is an upper bound for the optimality of $\mathbf{w}$ to (3), since

$$\begin{aligned}&\max_{i\in[m]} R_i(\bar{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}}\max_{i\in[m]} R_i(\mathbf{w})\\ &= \max_{\mathbf{q}\in\Delta_m}\sum_{i=1}^m \mathbf{q}_i R_i(\bar{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{q}\in\Delta_m}\sum_{i=1}^m \mathbf{q}_i R_i(\mathbf{w}) \\ &\leq \max_{\mathbf{q}\in\Delta_m}\sum_{i=1}^m \mathbf{q}_i R_i(\bar{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}}\sum_{i=1}^m \bar{\mathbf{q}}_i R_i(\mathbf{w}) = \epsilon(\bar{\mathbf{z}}).\end{aligned} \tag{13}$$

4. Algorithm for Empirical GDRO

Inspired by Alacaoglu & Malitsky (2022), we follow the common double-loop structure of variance reduction. The outer loop computes snapshot points in the primal space and the dual space, respectively. The inner loop runs a modified mirror prox scheme with a full gradient plus a stochastic gradient, rather than a pair of stochastic gradients in the classic mirror prox algorithm (Nemirovski, 2004; Juditsky et al., 2011). We emphasize the following two techniques that separate our algorithm from other similar work (Luo et al., 2021; Carmon & Hausler, 2022; Alacaoglu & Malitsky, 2022).

Variance Reduction Based on Group Sampling To improve complexity bounds, we propose the group sampling technique which samples uniformly from all groups per iteration and further reduces the variance of the stochastic gradient for each group. This strategy captures the two-level finite-sum structure by eliminating the randomness on the first level of the finite-sum structure, i.e., the summation $\sum_{i=1}^m \mathbf{q}_i R_i(\mathbf{w})$. Although the group sampling of ALEG queries m times more stochastic gradients than uniform sampling or importance sampling of MPVR, it produces a better stochastic gradient with a $1/m$ lower Lipschitz constant (cf. Lemma 4.3), which ultimately improves the complexity by a factor of $\sqrt{m}$. The motivation behind this technique is twofold: (i) identifying the nested finite-sum structure of the objective F, and (ii) leveraging the inherent properties of the stochastic gradient under $\|\cdot\|_\infty$. Comprehensive discussions can be found in Appendix C.

Alterable Hyperparameters To support variable algorithmic hyperparameters, we use the one-index-shifted weighted average rather than the naive ergodic average to construct the (mirror) snapshot points and the Lyapunov functions. In this way, we provide theoretical support for non-constant learning rates (Liner & Miikkulainen, 2021), which have been proven competitive in many practical scenarios.

We introduce the proposed ALEG in Algorithm 1. The detailed algorithm is elaborated in Section 4.1 and the corresponding theoretical guarantee is presented in Section 4.2.

4.1. Our Algorithm

In the outer loop of ALEG, we follow the standard variance reduction procedure (Zhang et al., 2013; Johnson & Zhang, 2013) by periodically computing the snapshot points as well as the full gradient. In the inner loop of ALEG, we maintain two sets of solutions as mirror prox algorithm (Nemirovski, 2004; Juditsky et al., 2011) and further accelerate it via snapshot points.

At the beginning of the s-th outer loop, we compute the snapshot $\mathbf{z}^s$ using the weighted average from the previous trajectory. The mirror snapshot $\nabla\psi(\bar{\mathbf{z}}^s)$ is constructed similarly by the weighted average of the current trajectory mapped in the dual space[2]. Formally, we compute them via the one-index-shifted weighted average as follows

$$\mathbf{z}^s = \left(\sum_{k=1}^{K_s}\alpha_{k-1}^{s-1}\right)^{-1}\sum_{k=1}^{K_{s-1}}\alpha_{k-1}^{s-1}\mathbf{z}_k^{s-1}, \tag{14}$$

$$\nabla\psi(\bar{\mathbf{z}}^s) = \left(\sum_{k=1}^{K_s}\alpha_{k-1}^{s-1}\right)^{-1}\sum_{k=1}^{K_{s-1}}\alpha_{k-1}^{s-1}\nabla\psi(\mathbf{z}_k^{s-1}). \tag{15}$$

Then, the full gradient $\nabla F(\mathbf{z}^s)$ is computed as

$$\nabla F(\mathbf{z}^s) = \begin{pmatrix} \sum_{i=1}^m \mathbf{q}_i^s \nabla R_i(\mathbf{w}^s) \\ -\left[R_1(\mathbf{w}^s), \cdots, R_m(\mathbf{w}^s)\right]^T\end{pmatrix}. \tag{16}$$

In the inner loop k, we first compute the prox point using the full gradient from the last epoch:

$$\begin{aligned}\mathbf{z}_{k+1/2}^s = \arg\min_{\mathbf{z}\in\mathcal{Z}}\{&\eta_k^s\left\langle\nabla F(\mathbf{z}^s),\mathbf{z}\right\rangle + \alpha_k^s B(\mathbf{z},\bar{\mathbf{z}}^s)\\ &+(1-\alpha_k^s)B(\mathbf{z},\mathbf{z}_k^s)\}.\end{aligned} \tag{17}$$

The above update is different from the traditional mirror prox algorithm (Nemirovski, 2004; Juditsky et al., 2011) because it uses the full gradient $\nabla F(\mathbf{z}^s)$ instead of the stochastic gradient at $\mathbf{z}_{k-1}^s$.

Remark 4.1. The update in (17) utilizes the "negative momentum" (Driggs et al., 2022) technique to achieve acceleration with the help of snapshot points. Similar ideas are also used in recent studies on variance reduction (Shang et al., 2017; Zhou et al., 2018). However, their momentum is performed in the primal space, which is different from MPVR and our method.

After $\mathbf{z}_{k+1/2}^s$ is calculated, we leverage the two-level structure to construct the stochastic gradient based on our group sampling technique. For each group i, we sample uniformly from the i-th group's samples $\{\xi_{ij}\}_{j=1}^{n_i}$. The m samples generated by the group sampling technique are denoted as:

$$\xi_k^s := \{\xi_{k,i}^s\}_{i=1}^m, \quad \xi_{k,i}^s \sim \mathrm{Unif}\left(\{\xi_{ij}\}_{j=1}^{n_i}\right), \forall\, i\in[m]. \tag{18}$$

[2]Actually, the value of $\nabla\psi(\bar{\mathbf{z}}^s)$ is sufficient for Algorithm 1. The inverse or conjugate of $\nabla\psi$ is not necessary to calculate and thus no additional cost is incurred.

Algorithm 1 Variance-Reduced Stochastic Mirror Prox **Al**gorithm for **E**mpirical **G**DRO (ALEG)

Input: Risk functions $\{R_i(\mathbf{w})\}_{i\in[m]}$, epoch number S, iteration numbers $\{K_s\}$, learning rates $\{\eta_k^s\}$, and weights $\{\alpha_k^s\}$.

Initialize $\mathbf{z}_0 = (\mathbf{w}_0; \mathbf{q}_0) = \arg\min_{\mathbf{z}\in\mathcal{Z}} \psi(\mathbf{z})$ as the starting point.
For each $j \in [K_{-1}]$, set $\mathbf{z}_j^{-1} = \mathbf{z}_0^0 = \mathbf{z}_0$.
for $s = 0$ **to** $S-1$ **do**
  Compute the snapshot $\mathbf{z}^s$ and the mirror snapshot $\nabla\psi(\bar{\mathbf{z}}^s)$ according to (14) and (15), respectively.
  Compute the full gradient $\nabla F(\mathbf{z}^s)$ according to (16).
  for $k = 0$ **to** $K_s - 1$ **do**
    Compute $\mathbf{z}_{k+1/2}^s$ according to (17).
    For each $i \in [m]$, sample $\xi_{k,i}^s$ uniformly from $\{\xi_{ij}\}_{j=1}^{n_i}$.
    Compute the variance-reduced stochastic gradient estimator $\mathbf{g}_k^s$ defined in (20).
    Compute $\mathbf{z}_{k+1}^s$ according to (21).
  end for
  Set $\mathbf{z}_0^{s+1} = \mathbf{z}_{K_s}^s$.
end for
Return $\mathbf{z}_S$ according to (22).

Remark 4.2. Compared to uniform sampling or importance sampling in Alacaoglu & Malitsky (2022), we incorporate the finite-sum components of the objective in (3) so that our stochastic gradients make use of a total of m samples.

Then, we construct our stochastic gradient based on the group sampling technique, as specified below:

$$\nabla F(\mathbf{z}^s; \xi_k^s) := \begin{pmatrix} \sum_{i=1}^m \mathbf{q}_i^s \nabla\ell(\mathbf{w}^s; \xi_{k,i}^s) \\ -\left[\ell(\mathbf{w}^s; \xi_{k,1}^s), \cdots, \ell(\mathbf{w}^s; \xi_{k,m}^s)\right]^T \end{pmatrix}. \tag{19}$$

The above construction equally weighs the randomness from every group, which is the key step to exploit the two-level finite-sum structure of (3). Next, we introduce the following lemma to quantify its Lipschitz continuity.

Lemma 4.3. *For any $s \in [S]^0, k \in [K_s]^0$, $\nabla F(\mathbf{z}; \xi_k^s)$ is L_z-Lipschitz continuous, where*

$$L_z := 2D_w \max\left\{\sqrt{2D_w^2 L^2 + G^2 \ln m}, G\sqrt{2\ln m}\right\}.$$

Comparing with MPVR (Alacaoglu & Malitsky, 2022, Definition 7), Lemma 4.3 shows that our group sampling technique can reduce the Lipschitz constant of the stochastic gradient by a factor of m, which is the reason for the $\sqrt{m}$ improvement of the complexity.

Inspired by variance reduction (Zhang et al., 2013; Johnson & Zhang, 2013) techniques, we construct the following variance-reduced stochastic gradient estimator using the stochastic gradient at the snapshot $\mathbf{z}^s$ and the full gradient in (16):

$$\mathbf{g}_k^s = \nabla F(\mathbf{z}_{k+1/2}^s; \xi_k^s) - \nabla F(\mathbf{z}^s; \xi_k^s) + \nabla F(\mathbf{z}^s). \tag{20}$$

With expectation taken over ξ_k^s, it is easy to verify that $\mathbf{g}_k^s$ is an unbiased estimator. The final step is to compute $\mathbf{z}_{k+1}^s$ according to the mirror prox scheme:

$$\begin{aligned} \mathbf{z}_{k+1}^s = \arg\min_{\mathbf{z}\in\mathcal{Z}} \{ &\eta_k^s \langle \mathbf{g}_k^s, \mathbf{z}\rangle + \alpha_k^s B(\mathbf{z}, \bar{\mathbf{z}}^s) \\ &+ (1-\alpha_k^s) B(\mathbf{z}, \mathbf{z}_k^s)\}. \end{aligned} \tag{21}$$

In (21), we use the variance-reduced stochastic gradient estimator $\mathbf{g}_k^s$ instead of raw stochastic gradient at $\mathbf{z}_k^s$ to achieve variance reduction. Upon completion of the double loop procedure, ALEG returns solutions in a different manner compared to MPVR, as shown below

$$\mathbf{z}_S = \left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1} \eta_k^s\right)^{-1} \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1} \eta_k^s \mathbf{z}_{k+1/2}^s. \tag{22}$$

Specifically, our utilization of alterable learning rates necessitates the computation of $\mathbf{z}_S$ through weighted averaging, as opposed to the ergodic averaging approach employed by Alacaoglu & Malitsky (2022).

4.2. Theoretical Guarantee

Now we present our theoretical result for Algorithm 1. The proofs are deferred to Appendix A.

Theorem 4.4. *Under Assumptions 3.3, 3.4 and 3.6, by setting $K_s = K, \alpha_k^s = \frac{1}{K}, \frac{1}{10L_z\sqrt{K}} \le \eta_k^s \le \frac{1}{L_z\sqrt{5K}}$, our Algorithm 1 ensures that*

$$\mathbb{E}\left[\epsilon(\mathbf{z}_S)\right] \le \mathcal{O}\left(\frac{1}{S}\sqrt{\frac{\ln m}{K}}\right). \tag{23}$$

Remark 4.5. We stick to the constant parameter setting in terms of $\{K_s\}$ and $\{\alpha_k^s\}$, while allowing the learning rates $\{\eta_k^s\}$ to remain adjustable. Variable $\{K_s\}$ and $\{\alpha_k^s\}$ complicate analysis and presentation with an additional summation term in overall complexity. For brevity, we set $\{K_s\}$ and $\{\alpha_k^s\}$ to be constant.

Corollary 4.6. *Under conditions in Theorem 4.4, by setting $K = \Theta(\bar{n})$, the computation complexity for Algorithm 1 to reach ε-accuracy of* (3) *is $\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$.*

Our complexity in Corollary 4.6 is better than the state-of-the-art rate (Alacaoglu & Malitsky, 2022) by a factor of $\sqrt{m}$. From a practical perspective, ALEG still maintains a low per-iteration complexity. The main step (17) and (21) only involves projections onto $\mathcal{W}$ and Δ_m respectively. With $\psi_w(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ and $\psi_q(\mathbf{q}) = \sum_{i=1}^m \mathbf{q}_i \ln \mathbf{q}_i$, the updates are equivalent to Stochastic Gradient Descent (SGD) and Hedge (Freund & Schapire, 1997).

5. Algorithm for Empirical MERO

In this section, we extend our methodology in Section 4 to solve the empirical MERO in (5). Inspired by the efficient stochastic approximation approach to MERO (Zhang et al., 2024), we propose a two-stage **Al**gorithm for **E**mpirical **M**ERO (ALEM) as shown in Algorithm 2. In the first stage, ALEM runs m ALEG instances to estimate the minimal empirical risk as $\{\hat{R}_i^*\}_{i\in[m]}$. In the second stage, we focus on the following approximate problem:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{ \hat{\underline{F}}(\mathbf{w},\mathbf{q}) := \sum_{i=1}^{m} \mathbf{q}_i \hat{\underline{R}}_i(\mathbf{w}) \right\}, \tag{24}$$

which can be regarded as a substitution of the excess empirical risk $\underline{R}_i(\mathbf{w})$ in (5) with the approximated excess empirical risk $\hat{\underline{R}}_i(\mathbf{w}) := R_i(\mathbf{w}) - \hat{R}_i^*$. Then ALEM treats (24) as an empirical GDRO problem and calls ALEG again to optimize it.

Similar to Zhang et al. (2024, Lemma 4.2), we present the following lemma to show that the optimization error for (5) is under control, provided that the optimization error for (24) is small and the approximation error $\hat{R}_i^* - R_i^*$ is close to zero for all groups. All the proofs are deferred to Appendix B.

Lemma 5.1. *For any $\bar{\mathbf{z}} = (\bar{\mathbf{w}};\bar{\mathbf{q}}) \in \mathcal{Z}$, define the duality gap for the empirical MERO problem (5) and the approximated empirical MERO problem (24) as*

$$\begin{aligned} \underline{\epsilon}(\bar{\mathbf{z}}) &:= \max_{\mathbf{q}\in\Delta_m} \underline{F}(\bar{\mathbf{w}},\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \underline{F}(\mathbf{w},\bar{\mathbf{q}}), \\ \hat{\underline{\epsilon}}(\bar{\mathbf{z}}) &:= \max_{\mathbf{q}\in\Delta_m} \hat{\underline{F}}(\bar{\mathbf{w}},\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \hat{\underline{F}}(\mathbf{w},\bar{\mathbf{q}}), \end{aligned} \tag{25}$$

respectively. It holds that

$$\underline{\epsilon}(\bar{\mathbf{z}}) \le \hat{\underline{\epsilon}}(\bar{\mathbf{z}}) + 2\max_{i\in[m]}\{\hat{R}_i^* - R_i^*\}. \tag{26}$$

Stage 1: Empirical Risk Minimization Noticing that when the number of groups reduces to 1, the minimax problem in (2) degenerates to the classical one-level finite-sum empirical risk minimization problem, which ALEG can still handle. Therefore, we can apply ALEG to each group to get an estimator $\hat{R}_i^*$ for R_i^*. The extra input T can be viewed as a budget to control the cost of each group. To deal with the max operator in (26), we provide the following high-probability bound.

Theorem 5.2. *Under Assumptions 3.3, 3.4 and 3.6, running Algorithm 1 as an ERM oracle by setting $S = \left\lceil \frac{T}{\sqrt{\bar{n}}} \right\rceil$, $K_s = \bar{n}, \alpha_k^s = \frac{1}{\bar{n}}, \frac{1}{10L\sqrt{\bar{n}}} \le \eta_k^s \le \frac{1}{L\sqrt{5\bar{n}}}$, for any group i, the following holds with probability at least $1-\delta$,*

$$\hat{R}_i^* - R_i^* \le \mathcal{O}\left(\frac{1}{T}\left(\sqrt{\ln\frac{mT}{\sqrt{\bar{n}}\delta}} + \frac{1}{\sqrt{\bar{n}}}\ln\frac{mT}{\sqrt{\bar{n}}\delta}\right)\right). \tag{27}$$

Algorithm 2 Two-Stage **Al**gorithm for **E**mpirical **M**ERO (ALEM)

Input: Risk functions $\{R_i(\mathbf{w})\}_{i\in[m]}$, epoch number S, iteration numbers $\{K_s\}$, learning rates $\{\eta_k^s\}$, weights $\{\alpha_k^s\}$, and budget T.

1: **for** $i = 1$ **to** m **do**
2: Empirical risk minimization: $\bar{\mathbf{w}}_i \leftarrow \text{ALEG}(R_i(\mathbf{w}), S, \{K_s\}, \{\eta_k^s\}, \{\alpha_k^s\})$.
3: Estimate the minimal empirical risk: $\hat{R}_i^* = R_i(\bar{\mathbf{w}}_i)$.
4: **end for**
5: Run our empirical GDRO solver to optimize (24): $\bar{\mathbf{z}} \leftarrow \text{ALEG}\Big(\{R_i(\mathbf{w}) - \hat{R}_i^*\}_{i\in[m]}, S, \{K_s\}, \{\eta_k^s\}, \{\alpha_k^s\}\Big)$.
6: **Return** $\bar{\mathbf{z}}$, $\{\bar{\mathbf{w}}_i\}_{i\in[m]}$, $\{\hat{R}_i^*\}_{i\in[m]}$.

Stage 2: Solving Empirical GDRO After we managed to minimize the empirical risks for m groups, we can approximate the excess empirical risk $\underline{R}_i(\mathbf{w})$ by $R_i(\mathbf{w}) - \hat{R}_i^*$. Then, we send the modified risk function $\hat{\underline{R}}_i(\cdot)$ into Algorithm 1 to get a solution for (24). We stick to the aforementioned budget T to regulate the cost. Based on Lemma 5.1 and Theorem 4.4, we have the following convergence guarantee for Algorithm 2.

Theorem 5.3. *Under the conditions in Theorem 5.2, our Algorithm 2 ensures that with probability at least $1-\delta$,*

$$\begin{aligned} \underline{\epsilon}(\bar{\mathbf{z}}) \le \mathcal{O}\Bigg(\frac{1}{T}\Bigg(&\sqrt{\ln m \ln\frac{T}{\sqrt{\bar{n}}\delta}} + \sqrt{\ln\frac{mT}{\sqrt{\bar{n}}\delta}} \\ &+ \sqrt{\frac{\ln m}{\bar{n}}}\ln\frac{T}{\sqrt{\bar{n}}\delta}\Bigg)\Bigg). \end{aligned} \tag{28}$$

Corollary 5.4. *Based on Theorems 5.2 and 5.3, the total computation complexity for Algorithm 2 to reach ε-accuracy of (5) is $\tilde{\mathcal{O}}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$.*

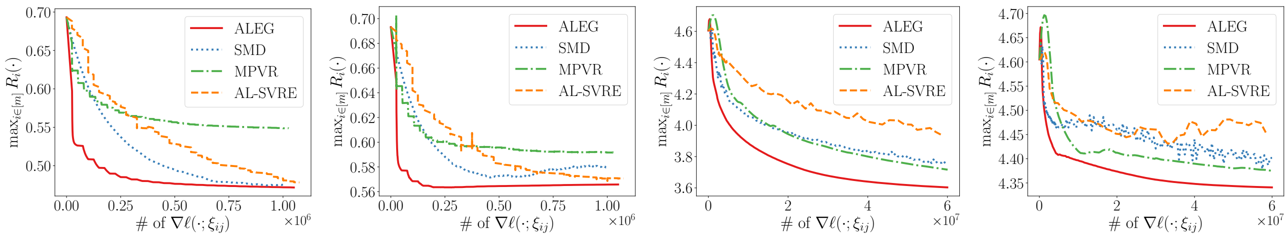
Remark 5.5. Corollary 5.4 shows that the complexity of ALEM nearly matches that of the empirical GDRO problem, which significantly improves over the $\mathcal{O}\left(\frac{m\bar{n}\ln(m\bar{n})}{\varepsilon^2}\right)$ computation complexity of ERMEG (Agarwal & Zhang, 2022).

6. Experiments

In this section, we conduct numerical experiments on empirical GDRO and empirical MERO to evaluate the performance of our algorithms.

6.1. Setup

For the empirical GDRO problem, We follow the setup in previous literature (Namkoong & Duchi, 2016; Soma et al.,



(a) Training on the synthetic data (b) Test on the synthetic data (c) Training on CIFAR-100 (d) Test on CIFAR-100

Figure 1: Comparison of the max empirical risk $\max_{i\in[m]} R_i(\cdot)$ with respect to the number of stochastic gradient evaluations # of $\nabla\ell(\cdot;\xi_{ij})$ on the synthetic dataset and the CIFAR-100 dataset.

2022; Zhang et al., 2023), using both synthetic and real-world datasets. Our goal is to find a single classifier to minimize the maximal empirical risk across all categories.

For the synthetic dataset, we set the number of groups to be 25. For each $i \in [25]$, we draw $\mathbf{w}_i^* \in \mathbb{R}^{1024}$ from the uniform distribution over the unit sphere. The data sample $\{\xi_{ij}\}_{j\in[n_i]}$ of group i is generated by $\xi_{ij} = (\mathbf{x}_{ij}, y_{ij})$, where

$$\mathbf{x}_{ij} \sim \mathcal{N}(0, I),$$
$$y_{ij} = \begin{cases} \operatorname{sign}\left(\mathbf{x}_{ij}^T\mathbf{w}_i^*\right), & \text{with probability } 0.9, \\ -\operatorname{sign}\left(\mathbf{x}_{ij}^T\mathbf{w}_i^*\right), & \text{with probability } 0.1. \end{cases}$$

We set $\ell(\cdot;\cdot)$ as the logistic loss and use different methods to train a linear model for this binary classification problem.

For the real-world dataset, we use CIFAR-100 (Krizhevsky et al., 2009), which has 100 classes containing 500 training images and 100 testing images for each class. Our goal is to determine the class for each image. We set $m = 100$ according to the number of categories and therefore the empirical risk function for group i is exactly the average loss function amongst all images of this class. We set $\ell(\cdot;\cdot)$ as the softmax loss function for this multi-class classification problem. The underlying predictive model remains to be linear, which satisfies the convex-concave setting.

For the empirical MERO problem, we aim to conduct a similar task as before. We stick to CIFAR-100 as the real-world dataset. However, to simulate the scenario where the groups of distributions differ from each other, we introduce heterogeneous noise in the synthetic dataset, and generate $\xi_{ij} = (\mathbf{x}_{ij}, y_{ij})$, where

$$\mathbf{x}_{ij} \sim \mathcal{N}(0, I),$$
$$y_{ij} = \begin{cases} \operatorname{sign}(\mathbf{x}_{ij}^T\mathbf{w}_i^*), & \text{with probability } p_i = 0.95 - \frac{i}{160}, \\ -\operatorname{sign}(\mathbf{x}_{ij}^T\mathbf{w}_i^*), & \text{with probability } 1 - p_i, \end{cases}$$

for $i = 0, 1, \cdots, 24$. The rest of the construction process follows the same steps as those used in the empirical GDRO experiments.

Different from the empirical GDRO experiments, we need to calculate the minimal empirical risk for all groups so as to evaluate our performance. We pass the data of each group to an Empirical Risk Minimization (ERM) oracle. Due to the convexity of the problem, running the oracle adequately long ensures the solution will closely approximate the true minimal empirical risk. After this process, we regard the outputs of the oracle as the true values $\{R_i^*\}_{i\in[m]}$.

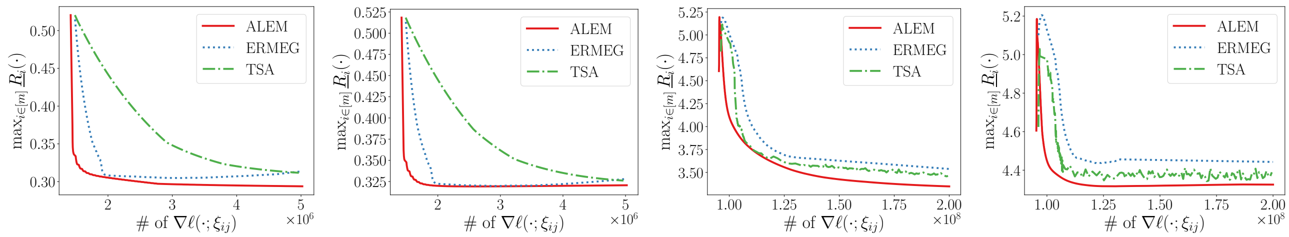
6.2. Results for Empirical GDRO

To evaluate the performance measure, we report the maximal empirical risk $\max_{i\in[m]} R_i(\cdot)$ on the training set. In order to show the generalization abilities, we also report the maximal empirical risk on the test set. To evaluate different algorithms, we use the number of stochastic gradient evaluations to reflect the computation complexity.

We compare our algorithm ALEG with SMD (Nemirovski et al., 2009), MPVR (Alacaoglu & Malitsky, 2022) and AL-SVRE (Luo et al., 2021). The results are shown in Figure 1. We emphasize that our implementation of ALEG supports changeable hyperparameters in terms of $\{K_s\}, \{\alpha_k^s\}, \{\eta_k^s\}$, which helps to boost its overall performance according to our observations. However, to fairly compare ALEG with others, we stick to the settings in Theorem 4.4, i.e., constant $\{K_s\}, \{\alpha_k^s\}$ and alterable $\{\eta_k^s\}$.

On the synthetic dataset, ALEG demonstrates notably faster convergence compared to other methods, in terms of both the training set and test set. Additionally, it achieves a lower maximal empirical risk than the alternatives. For the CIFAR-100 dataset, finding a single classifier becomes challenging because the computation complexity is proportional to the number of groups m. Under such a challenging task, ALEG still performs significantly better than others.

On the training set of CIFAR-100, ALEG also significantly outperforms other methods in terms of convergence speed and quality of the solution. On the CIFAR-100 test set, ALEG demonstrates faster convergence and greater stability than the other three algorithms, showcasing its superior



(a) Training on the synthetic data (b) Test on the synthetic data (c) Training on CIFAR-100 (d) Test on CIFAR-100

Figure 2: Comparison of the max excess empirical risk $\max_{i\in[m]} \underline{R}_i(\cdot)$ with respect to the number of stochastic gradient evaluations $\#$ of $\nabla\ell(\cdot;\xi_{ij})$ on the synthetic dataset and the CIFAR-100 dataset.

generalization capability. While SMD and MPVR behave similarly on the training set, MPVR demonstrates its robustness on the test set, indicating variance reduction has an edge over pure stochastic algorithms like SMD.

6.3. Results for Empirical MERO

We compare ALEM with ERMEG (Agarwal & Zhang, 2022, Section 6) and TSA (Zhang et al., 2024, Algorithm 2). Similarly, we report the max excess empirical risk $\max_{i\in[m]} \underline{R}_i(\cdot)$ on both the training set and the test set. The results are presented in Figure 2.

Note that all three algorithms follow the two-stage schema, which means they have to estimate m minimal empirical risks before solving the approximated empirical GDRO problem. This explains why the x-axis of the four figures in Figure 2 doesn't start at zero. The result in Figures 2(a) and 2(c) shows that our algorithm converges more rapidly and is more stable than ERMEG or TSA. The result in Figures 2(b) and 2(d) validates the strong generalization ability of ALEM.

In terms of the running time, we also observe that our algorithm performs significantly faster than TSA and ERMEG, especially for the latter. ERMEG not only needs to estimate the minimal empirical risk for all groups, but it also runs an ERM oracle every iteration, which is highly time-consuming and impractical.

7. Conclusion

We develop a variance-reduced stochastic mirror prox algorithm called ALEG to target the empirical GDRO problem. Specifically, we propose a simple yet effective group sampling strategy and a novel variable routine to reduce the complexity and enhance the algorithmic flexibility, respectively. ALEG attains an $\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$ computation complexity, which improves the state-of-the-art result by a factor of $\sqrt{m}$.

Based on ALEG, we develop a two-stage algorithm called ALEM to cope with the empirical MERO problem. In the first stage, ALEM runs ALEG to estimate the minimal empirical risk for all groups. In the second stage, ALEM utilizes ALEG to solve an approximate empirical MERO problem. We also establish an $\tilde{\mathcal{O}}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$ computation complexity, improving over the existing methods. Finally, we conduct experiments on the synthetic dataset as well as the real-world dataset to validate the effectiveness of our algorithms.

Acknowledgements

This work was partially supported by the National Science and Technology Major Project (2022ZD0114801) and NSFC (62122037, 61921006).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, A. and Zhang, T. Minimax regret optimization for robust machine learning under distribution shift. In *Proceedings of the 35th Conference on Learning Theory*, pp. 2704–2729, 2022.

Alacaoglu, A. and Malitsky, Y. Stochastic variance reduction for variational inequality methods. In *Proceedings of the 35th Conference on Learning Theory*, pp. 778–816, 2022.

Allen-Zhu, Z. Katyusha X: Simple momentum method for stochastic sum-of-nonconvex optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 179–185, 2018.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 699–707, 2016.

Allen-Zhu, Z. and Orecchia, L. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference*, pp. 3:1–3:22, 2017.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Bertsimas, D., Gupta, V., and Kallus, N. Robust sample average approximation. *Mathematical Programming*, 171: 217–282, 2014.

Blum, A., Haghtalab, N., Procaccia, A. D., and Qiao, M. Collaborative PAC learning. In *Advances in Neural Information Processing Systems 30*, 2017.

Carmon, Y. and Hausler, D. Distributionally robust optimization via ball oracle acceleration. In *Advances in Neural Information Processing Systems 35*, pp. 35866–35879, 2022.

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems 32*, 2019.

Carmon, Y., Jambulapati, A., Jiang, Q., Jin, Y., Lee, Y. T., Sidford, A., and Tian, K. Acceleration with a ball optimization oracle. In *Advances in Neural Information Processing Systems 33*, pp. 19052–19063, 2020.

Carmon, Y., Jambulapati, A., Jin, Y., and Sidford, A. Thinking inside the ball: Near-optimal minimization of the maximal loss. In *Proceedings of the 34th Conference on Learning Theory*, pp. 866–882, 2021.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chen, L., Yao, B., and Luo, L. Faster stochastic algorithms for minimax optimization under Polyak-Łojasiewicz condition. In *Advances in Neural Information Processing Systems 35*, pp. 13921–13932, 2022.

Condat, L. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1-2):575–585, 2016.

Dang, C. D. and Lan, G. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and Applications*, 60(2):277–310, 2015.

Driggs, D., Ehrhardt, M. J., and Schönlieb, C.-B. Accelerating variance-reduced stochastic gradient methods. *Mathematical Programming*, 191(2):1–45, 2022.

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

Duchi, J. C., Glynn, P. W., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Gao, R. and Kleywegt, A. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Han, Y., Xie, G., and Zhang, Z. Lower complexity bounds of finite-sum optimization problems: The results and construction. *Journal of Machine Learning Research*, 25 (2):1–86, 2024.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2029–2037, 2018.

Jiang, W., Li, G., Wang, Y., Zhang, L., and Yang, T. Multi-block-single-probe variance reduced estimator for coupled compositional optimization. In *Advances in Neural Information Processing Systems 35*, pp. 32499–32511, 2022.

Jin, J., Zhang, B., Wang, H., and Wang, L. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Advances in Neural Information Processing Systems 34*, pp. 2771–2782, 2021.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advance in Neural Information Processing Systems 26*, pp. 315–323, 2013.

Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Kakade, S. M. and Tewari, A. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21*, 2008.

Kivinen, J. and Warmuth, M. K. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Laguel, Y., Pillutla, K., Malick, J., and Harchaoui, Z. A superquantile approach to federated learning with heterogeneous devices. In *Proceedings of the 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2021a.

Laguel, Y., Pillutla, K., Malick, J., and Harchaoui, Z. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, 2021b.

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

Lee, S. and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems 34*, pp. 22588–22600, 2021.

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems 28*, 2015.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6083–6093, 2020a.

Lin, T., Jin, C., and Jordan, M. I. Near-optimal algorithms for minimax optimization. In *Proceedings of the 33rd Conference on Learning Theory*, pp. 2738–2779, 2020b.

Liner, E. and Miikkulainen, R. Improving neural network learning through dual variable learning rates. In *International Joint Conference on Neural Networks*, pp. 1–7, 2021.

Liu, M., Rafique, H., Lin, Q., and Yang, T. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *The Journal of Machine Learning Research*, 22(1):7651–7684, 2021.

Liu, Y., Shang, F., An, W., Liu, J., Liu, H., and Lin, Z. A single-loop accelerated extra-gradient difference algorithm with improved complexity bounds for constrained minimax optimization. In *Advances in Neural Information Processing Systems 36*, pp. 61699–61711, 2023.

Luo, L., Xie, G., Zhang, T., and Zhang, Z. Near optimal stochastic algorithms for finite-sum unbalanced convex-concave minimax optimization. *arXiv preprint arXiv:2106.01761*, 2021.

Mehta, R., Roulet, V., Pillutla, K., Liu, L., and Harchaoui, Z. Stochastic optimization for spectral risk measures. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 10112–10159, 2023.

Mehta, R., Roulet, V., Pillutla, K., and Harchaoui, Z. Distributionally robust optimization with bias and variance reduction. In *The 12th International Conference on Learning Representations*, 2024.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4615–4625, 2019.

Mokhtari, A., Ozdaglar, A. E., and Pattathil, S. Convergence rate of $O(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020.

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

Nemirovski, A. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

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

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

Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems 28*, pp. 3168–3176, 2015.

Nguyen, H. and Zakynthinou, L. Improved algorithms for collaborative PAC learning. In *Advances in Neural Information Processing Systems 31*, 2018.

Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4227–4237, 2019.

Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1–2):1–35, 2021a.

Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2):1–35, 2021b.

Palaniappan, B. and Bach, F. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems 29*, 2016.

Pillutla, K., Laguel, Y., Malick, J., and Harchaoui, Z. Federated learning with superquantile aggregation for heterogeneous data. *Machine Learning*, pp. 1–68, 2023.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1571–1578, 2012.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 314–323, 2016.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Rockafellar, R. T. *Convex Analysis:(PMS-28)*. Princeton University Press, 2015.

Rothblum, G. N. and Yona, G. Multi-group agnostic PAC learnability. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9107–9115, 2021.

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 29563–29648, 2023.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *The 8th International Conference on Learning Representations*, 2020.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *Proceedings of the 22nd Conference on Learning Theory*, pp. 5, 2009.

Shang, F., Liu, Y., Cheng, J., and Zhuo, J. Fast stochastic variance reduced gradient method with momentum acceleration for machine learning. *arXiv preprint arXiv:1703.07948*, 2017.

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

Smith, L. N. Cyclical learning rates for training neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 464–472, 2017.

Soma, T., Gatmiry, K., and Jegelka, S. Optimal algorithms for group distributionally robust optimization and beyond. *arXiv preprint arXiv:2212.13669*, 2022.

Xie, G., Luo, L., Lian, Y., and Zhang, Z. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10504–10513, 2020.

Yang, J., Kiyavash, N., and He, N. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems 33*, pp. 1153–1165, 2020.

Yazdandoost Hamedani, E. and Jalilzadeh, A. A stochastic variance-reduced accelerated primal-dual method for finite-sum saddle-point problems. *Computational Optimization and Applications*, 85(2):1–27, 2023.

Zhang, L., Mahdavi, M., and Jin, R. Linear convergence with condition number independent access of full gradients. In *Advance in Neural Information Processing Systems 26*, pp. 980–988, 2013.

Zhang, L., Zhao, P., Zhuang, Z.-H., Yang, T., and Zhou, Z.-H. Stochastic approximation approaches to group distributionally robust optimization. In *Advances in Neural Information Processing Systems 36*, pp. 52490–52522, 2023.

Zhang, L., Bai, H., Tu, W.-W., Yang, P., and Hu, Y. Efficient stochastic approximation of minimax excess risk optimization. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Zhao, R. Accelerated stochastic algorithms for convex-concave saddle-point problems. *Mathematics of Operations Research*, 47(2):1443–1473, 2022.

Zhou, K., Shang, F., and Cheng, J. A simple stochastic variance reduced algorithm with fast convergence rates. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5980–5989, 2018.

A. Analysis for Empirical GDRO

We present the omitted proofs for empirical GDRO in this section. Firstly, we conduct some necessary technical preparations in Appendix A.1. Then, we analyze the algorithmic variance-reduced behaviors in Appendix A.2 and give an implicit upper bound for the variances of the stochastic gradients along the trajectory. Finally, we proceed to show the convergence and the derived complexity in Appendix A.3.

A.1. Preparations

Here we provide some definitions to facilitate understanding and bring convenience.

Definition A.1. (Saddle point) Define any solution to (3) as $\mathbf{z}_* = (\mathbf{w}_*; \mathbf{q}_*)$.

Definition A.2. (Martingale difference sequence) Define $\Delta_k^s = \mathbf{g}_k^s - \nabla F(\mathbf{z}_{k+1/2}^s)$.

Definition A.3. (Periodically decaying sequence) We call $\{\alpha_{k-1}^{s-1}\}_{s\in[S],k\in[K_s]}$ a periodically decaying sequence if it satisfies: (i) $\alpha_{K_s-1}^s \le \alpha_0^{s+1}$; (ii) $\sum_{k=1}^{K_s} \alpha_{k-1}^s \le \sum_{k=1}^{K_{s-1}} \alpha_{k-1}^{s-1}$.

Definition A.4. (Lyapunov function) For Bregman divergence defined in (11), we define

$$\Psi_s(\mathbf{z}) := (1-\alpha_0^s)B(\mathbf{z}, \mathbf{z}_0^s) + \sum_{k=1}^{K_{s-1}} \alpha_{k-1}^{s-1} B(\mathbf{z}, \mathbf{z}_k^{s-1}). \tag{29}$$

Remark A.5. For Definition A.2, we know that Δ_k^s equals zero in conditional expectation, which is verified in Proposition A.7. Definition A.3 could be easily satisfied both in theoretical analysis and real-world experimental settings. Intuitively, our design of $\{\alpha_k^s\}_{s\in[S],k\in[K_s]}$ is inspired by cyclical learning rate (Smith, 2017), with the property of periodic decay.

In the following analysis, we stick to the choice $\|\cdot\|_q = \|\cdot\|_1$ not just for simplicity, but also more practical when it comes to implementation. This choice enables the mirror descent step for $\mathbf{q}$ to have a closed-form solution. Note that in this case, we have $D_q^2 = \ln m$ and $\alpha_q = 1$. Without loss of generality, we also assume $\alpha_w = 1$. There are some important facts shown in the following lemmas.

Proposition A.6. *For any* $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$*,* $\|\mathbf{z}_1 - \mathbf{z}_2\| \le 2\sqrt{2}$.

Proof. Define $\mathbf{z}_0 = \arg\min_{\mathbf{z}\in\mathcal{Z}} \psi(\mathbf{z})$. By the fact that $\max_{z\in\mathcal{Z}} B(\mathbf{z}, \mathbf{z}_0) \le \max_{z\in\mathcal{Z}} \psi(\mathbf{z}) - \min_{z\in\mathcal{Z}} \psi(\mathbf{z}) \le 1$, we have

$$\begin{aligned} \|\mathbf{z}_1 - \mathbf{z}_2\| &\le \|\mathbf{z}_1 - \mathbf{z}_0\| + \|\mathbf{z}_2 - \mathbf{z}_0\| \le \sqrt{2B(\mathbf{z}_1, \mathbf{z}_0)} + \sqrt{2B(\mathbf{z}_2, \mathbf{z}_0)} \\ &\le \sqrt{2\max_{z\in\mathcal{Z}} B(\mathbf{z}, \mathbf{z}_0)} + \sqrt{2\max_{z\in\mathcal{Z}} B(\mathbf{z}, \mathbf{z}_0)} \le 2\sqrt{2}. \end{aligned} \tag{30}$$

□

Proposition A.7. *(Unbiasedness of the merged stochastic gradient operator) The stochastic gradient* $\nabla F(\mathbf{z}^s; \xi_k^s)$ *defined in* (19) *is unbiased.*

Proof. From our group sampling strategy defined in (18), we have

$$\forall \mathbf{z} \in \mathcal{Z}, \forall\, i \in [m]: \quad \mathbb{E}\left[\ell(\mathbf{w}; \xi_{k,i}^s)\right] = R_i(\mathbf{w}), \quad \mathbb{E}\left[\nabla \ell(\mathbf{w}; \xi_{k,i}^s)\right] = \nabla R_i(\mathbf{w}). \tag{31}$$

From the linearity of expectation, it's easy to deduce that $\mathbb{E}\left[\nabla F(\mathbf{z}; \xi_k^s)\right] = \nabla F(\mathbf{z})$ for any $\mathbf{z} \in \mathcal{Z}$. Hence the unbiasedness of the stochastic gradient $\nabla F(\mathbf{z}^s; \xi_k^s)$ is proved naturally. □

Proof of Lemma 4.3. Pick two arbitrary points $\mathbf{z}^+ = (\mathbf{w}^+, \mathbf{q}^+) \in \mathcal{Z}, z = (\mathbf{w}, \mathbf{q}) \in \mathcal{Z}$. First, we bound the gradient in $\mathbf{w}$

as follows:

$$\begin{aligned}
&\left\|\nabla_{\mathbf{w}} F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{w}} F(\mathbf{z};\xi_k^s)\right\|_{w,*}^2 \\
=&\left\|\sum_{i=1}^m \mathbf{q}_i^+[\nabla \ell(\mathbf{w}^+;\xi_{k,i}^s) - \nabla \ell(\mathbf{w};\xi_{k,i}^s)] + \sum_{i=1}^m (\mathbf{q}_i^+ - \mathbf{q}_i)\nabla \ell(\mathbf{w};\xi_{k,i}^s)\right\|_{w,*}^2 \\
\leq&2\left\|\sum_{i=1}^m \mathbf{q}_i^+[\nabla \ell(\mathbf{w}^+;\xi_{k,i}^s) - \nabla \ell(\mathbf{w};\xi_{k,i}^s)]\right\|_{w,*}^2 + 2\left\|\sum_{i=1}^m (\mathbf{q}_i^+ - \mathbf{q}_i)\nabla \ell(\mathbf{w};\xi_{k,i}^s)\right\|_{w,*}^2 \\
\leq&2\sum_{i=1}^m \mathbf{q}_i^+\left\|\nabla \ell(\mathbf{w}^+;\xi_{k,i}^s) - \nabla \ell(\mathbf{w};\xi_{k,i}^s)\right\|_{w,*}^2 + 2\left(\sum_{i=1}^m |\mathbf{q}_i^+ - \mathbf{q}_i|\left\|\nabla \ell(\mathbf{w};\xi_{k,i}^s)\right\|_{w,*}\right)^2 \\
\leq&2\sum_{i=1}^m \mathbf{q}_i^+ L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 + 2\left(\sum_{i=1}^m |\mathbf{q}_i^+ - \mathbf{q}_i| G\right)^2 \\
=&2L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 + 2G^2\left\|\mathbf{q}^+ - \mathbf{q}\right\|_1^2.
\end{aligned} \tag{32}$$

The second inequality uses Assumption 3.4. Next, we bound the gradient in $\mathbf{q}$. Again from Assumption 3.4, we have

$$\forall\, i \in [m] :, \quad \ell(\mathbf{w}^+;\xi_{k,i}^s) - \ell(\mathbf{w};\xi_{k,i}^s) \leq G\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w. \tag{33}$$

Notice that $\nabla_{\mathbf{q}} F(\mathbf{z};\xi_k^s) = \left[\ell(\mathbf{w};\xi_{k,1}^s), \cdots, \ell(\mathbf{w};\xi_{k,m}^s)\right]^T$. Squaring it on both sides of (33) and taking maximum over all $i \in [m]$, we have

$$\left\|\nabla_{\mathbf{q}} F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{q}} F(\mathbf{z};\xi_k^s)\right\|_\infty^2 = \max_{i\in[m]}[\ell(\mathbf{w}^+;\xi_{k,i}^s) - \ell(\mathbf{w};\xi_{k,i}^s)]^2 \leq G^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2. \tag{34}$$

By merging the two component's gradients, we get the desired result by simple calculation:

$$\begin{aligned}
&\left\|\nabla F(\mathbf{z}^+;\xi_k^s) - \nabla F(\mathbf{z};\xi_k^s)\right\|_*^2 \\
=&2D_w^2\left\|\nabla_{\mathbf{w}} F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{w}} F(\mathbf{z};\xi_k^s)\right\|_{w,*}^2 + 2\ln m\left\|\nabla_{\mathbf{q}} F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{q}} F(\mathbf{z};\xi_k^s)\right\|_\infty^2 \\
\leq&(4D_w^2 L^2 + 2G^2\ln m)\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 + 4D_w^2 G^2\left\|\mathbf{q}^+ - \mathbf{q}\right\|_1^2 \\
\leq&\frac{1}{2D_w^2}\left[4D_w^2(2D_w^2 L^2 + G^2\ln m)\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2\right] + \frac{1}{2\ln m}\left[4D_w^2(2G^2\ln m)\left\|\mathbf{q}^+ - \mathbf{q}\right\|_1^2\right] \\
\leq&\frac{L_z^2}{2D_w^2}\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 + \frac{L_z^2}{2\ln m}\left\|\mathbf{q}^+ - \mathbf{q}\right\|_1^2 \\
=&L_z^2\left\|\mathbf{z}^+ - \mathbf{z}\right\|^2.
\end{aligned} \tag{35}$$

□

Lemma A.8. *(Optimality condition for mirror descent) For any* $\mathbf{g} \in \mathcal{E}^* \times \mathbb{R}^m$*, let* $\mathbf{z}^{t+1} = \arg\min_{\mathbf{z}\in\mathcal{Z}}\{\langle \mathbf{g}, \mathbf{z}\rangle + \alpha B(\mathbf{z}, \mathbf{z}_1) + (1-\alpha)B(\mathbf{z}, \mathbf{z}_2)\}$*. It holds that*

$$\langle \mathbf{g}, \mathbf{z}^{t+1} - \mathbf{z}\rangle \leq -B(\mathbf{z}, \mathbf{z}^{t+1}) + \alpha\left[B(\mathbf{z}, \mathbf{z}_1) - B(\mathbf{z}^{t+1}, \mathbf{z}_1)\right] + (1-\alpha)\left[B(\mathbf{z}, \mathbf{z}_2) - B(\mathbf{z}^{t+1}, \mathbf{z}_2)\right], \quad \forall\, \mathbf{z} \in \mathcal{Z}. \tag{36}$$

Proof. The proof is straightforward by applying the first-order optimality condition for the definition of $\mathbf{z}^{t+1}$. Then a direct application of three point equality of Bregman divergence yields the result. By the first order optimality of $\mathbf{z}^{t+1}$,

$$\mathbf{0} \in \mathbf{g} + \alpha\left[\nabla\psi(\mathbf{z}^{t+1}) - \nabla\psi(\mathbf{z}_1)\right] + (1-\alpha)\left[\nabla\psi(\mathbf{z}^{t+1}) - \nabla\psi(\mathbf{z}_2)\right] + N_{\mathcal{Z}}(\mathbf{z}^{t+1}) \tag{37}$$

where $N_{\mathcal{Z}}(\mathbf{z}^{t+1}) := \{\mathbf{g} \in \mathcal{E}^* \times \mathbb{R}^m | \langle \mathbf{g}, \mathbf{z} - \mathbf{z}^{t+1}\rangle \leq 0, \forall \mathbf{z} \in \mathcal{Z}\}$ is the normal cone (subdifferential of indicator function) at point $\mathbf{z}^{t+1}$ for convex set $\mathcal{Z}$. The above relation further implies

$$\langle \mathbf{g} + \alpha\left[\nabla\psi(\mathbf{z}_1) - \nabla\psi(\mathbf{z}^{t+1})\right] + (1-\alpha)\left[\nabla\psi(\mathbf{z}_2) - \nabla\psi(\mathbf{z}^{t+1})\right], \mathbf{z} - \mathbf{z}^{t+1}\rangle \leq 0,\ \forall\, \mathbf{z} \in \mathcal{Z}. \tag{38}$$

According to the generalized triangle inequality for Bregman divergence, we have

$$\langle \nabla\psi(\mathbf{z}_i) - \nabla\psi(\mathbf{z}^{t+1}), \mathbf{z} - \mathbf{z}^{t+1}\rangle = B(\mathbf{z}, \mathbf{z}^{t+1}) + B(\mathbf{z}^{t+1}, \mathbf{z}_i) - B(\mathbf{z}, \mathbf{z}_i), \quad \forall\, i \in \{1, 2\}. \tag{39}$$

Applying (39) to the LHS of (38) to derive that for any $\mathbf{z} \in \mathcal{Z}$,

$$\begin{aligned} &\langle \mathbf{g}, \mathbf{z} - \mathbf{z}^{t+1}\rangle + \alpha\left[B(\mathbf{z}, \mathbf{z}^{t+1}) + B(\mathbf{z}^{t+1}, \mathbf{z}_1) - B(\mathbf{z}, \mathbf{z}_1)\right] + (1-\alpha)\left[B(\mathbf{z}, \mathbf{z}^{t+1}) + B(\mathbf{z}^{t+1}, \mathbf{z}_2) - B(\mathbf{z}, \mathbf{z}_2)\right] \\ =&\langle \mathbf{g}, \mathbf{z} - \mathbf{z}^{t+1}\rangle + B(\mathbf{z}, \mathbf{z}^{t+1}) + \alpha\left[B(\mathbf{z}^{t+1}, \mathbf{z}_1) - B(\mathbf{z}, \mathbf{z}_1)\right] + (1-\alpha)\left[B(\mathbf{z}^{t+1}, \mathbf{z}_2) - B(\mathbf{z}, \mathbf{z}_2)\right] \le 0, \end{aligned} \tag{40}$$

which concludes our proof by a simple rearrangement. □

A.2. Variance-Reduced Routine

Lemma A.9. *If $\{\alpha_k^s\}_{s\in[S]^0, k\in[K_s]^0}$ is a periodically decaying sequence as defined in Definition A.3, for any $\mathbf{z} \in \mathcal{Z}$:*

$$\begin{aligned} &\sum_{k=0}^{K_s-1} \eta_k^s \left\langle \nabla F(\mathbf{z}_{k+1/2}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z}\right\rangle \\ \le&\Psi_s(\mathbf{z}) - \Psi_{s+1}(\mathbf{z}) + \sum_{k=0}^{K_s-1}\left[\eta_k^s \left\langle \Delta_k^s, \mathbf{z} - \mathbf{z}_{k+1/2}^s\right\rangle + \frac{(\eta_k^s L_z)^2 - \alpha_k^s}{2}\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\|^2\right]. \end{aligned} \tag{41}$$

Proof. Using Lemma A.8 on $\mathbf{z}_{k+1/2}^s$ and taking arbitrary $\mathbf{z}$ as $\mathbf{z}_{k+1}^s$, we have

$$\begin{aligned} \eta_k^s \left\langle \nabla F(\mathbf{z}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z}_{k+1}^s\right\rangle \le -B(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s) + \alpha_k^s\left[B(\mathbf{z}_{k+1}^s, \bar{\mathbf{z}}^s) - B(\mathbf{z}_{k+1/2}^s, \bar{\mathbf{z}}^s)\right] \\ +(1-\alpha_k^s)\left[B(\mathbf{z}_{k+1}^s, \mathbf{z}_k^s) - B(\mathbf{z}_{k+1/2}^s, \mathbf{z}_k^s)\right]. \end{aligned} \tag{42}$$

Using Lemma A.8 on $\mathbf{z}_{k+1}^s$, we have for any $\mathbf{z} \in \mathcal{Z}$,

$$\begin{aligned} \eta_k^s \left\langle \mathbf{g}_k^s, \mathbf{z}_{k+1}^s - \mathbf{z}\right\rangle \le -B(\mathbf{z}, \mathbf{z}_{k+1}^s) + \alpha_k^s\left[B(\mathbf{z}, \bar{\mathbf{z}}^s) - B(\mathbf{z}_{k+1}^s, \bar{\mathbf{z}}^s)\right] \\ +(1-\alpha_k^s)\left[B(\mathbf{z}, \mathbf{z}_k^s) - B(\mathbf{z}_{k+1}^s, \mathbf{z}_k^s)\right]. \end{aligned} \tag{43}$$

Adding them together:

$$\begin{aligned} &\eta_k^s \left\langle \nabla F(\mathbf{z}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z}_{k+1}^s\right\rangle + \eta_k^s \left\langle \mathbf{g}_k^s, \mathbf{z}_{k+1}^s - \mathbf{z}\right\rangle \\ \le& - B(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s) - B(\mathbf{z}, \mathbf{z}_{k+1}^s) + \alpha_k^s\left[B(\mathbf{z}, \bar{\mathbf{z}}^s) - B(\mathbf{z}_{k+1/2}^s, \bar{\mathbf{z}}^s)\right] \\ &+ (1-\alpha_k^s)\left[B(\mathbf{z}, \mathbf{z}_k^s) - B(\mathbf{z}_{k+1/2}^s, \mathbf{z}_k^s)\right]. \end{aligned} \tag{44}$$

According to (20) and Definition A.2 we have

$$\begin{aligned} &\eta_k^s \left\langle \nabla F(\mathbf{z}_{k+1/2}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z}\right\rangle \\ =&\eta_k^s \left\langle \mathbf{g}_k^s, \mathbf{z}_{k+1/2}^s - \mathbf{z}\right\rangle + \eta_k^s \left\langle \nabla F(\mathbf{z}_{k+1/2}^s) - \mathbf{g}_k^s, \mathbf{z}_{k+1/2}^s - \mathbf{z}\right\rangle \\ =&\eta_k^s \left\langle \mathbf{g}_k^s, \mathbf{z}_{k+1/2}^s - \mathbf{z}_{k+1}^s\right\rangle + \eta_k^s \left\langle \mathbf{g}_k^s, \mathbf{z}_{k+1}^s - \mathbf{z}\right\rangle - \eta_k^s \left\langle \Delta_k^s, \mathbf{z}_{k+1/2}^s - \mathbf{z}\right\rangle \\ =&\eta_k^s \left\langle \nabla F(\mathbf{z}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z}_{k+1}^s\right\rangle + \eta_k^s \left\langle \mathbf{g}_k^s, \mathbf{z}_{k+1}^s - \mathbf{z}\right\rangle \\ &+ \eta_k^s \left\langle \nabla F(\mathbf{z}_{k+1/2}^s; \xi_k^s) - \nabla F(\mathbf{z}^s; \xi_k^s), \mathbf{z}_{k+1/2}^s - \mathbf{z}_{k+1}^s\right\rangle - \eta_k^s \left\langle \Delta_k^s, \mathbf{z}_{k+1/2}^s - \mathbf{z}\right\rangle. \end{aligned} \tag{45}$$

Adding (44) and (45), we have

$$\begin{aligned} &\eta_k^s \left\langle \nabla F(\mathbf{z}_{k+1/2}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z}\right\rangle \\ \le&\eta_k^s \left\langle \nabla F(\mathbf{z}^s; \xi_k^s) - \nabla F(\mathbf{z}_{k+1/2}^s; \xi_k^s), \mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s\right\rangle \\ &+ \alpha_k^s\left[B(\mathbf{z}, \bar{\mathbf{z}}^s) - B(\mathbf{z}_{k+1/2}^s, \bar{\mathbf{z}}^s)\right] + (1-\alpha_k^s)\left[B(\mathbf{z}, \mathbf{z}_k^s) - B(\mathbf{z}_{k+1/2}^s, \mathbf{z}_k^s)\right] \\ &- B(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s) - B(\mathbf{z}, \mathbf{z}_{k+1}^s) + \eta_k^s \left\langle \Delta_k^s, \mathbf{z} - \mathbf{z}_{k+1/2}^s\right\rangle. \end{aligned} \tag{46}$$

Applying Young's inequality to the inner product and further using the smoothness of the stochastic gradient (cf. Lemma 4.3), the following estimation holds:

$$
\begin{aligned}
&\eta_k^s \left\langle \nabla F(\mathbf{z}^s;\xi_k^s) - \nabla F(\mathbf{z}^s_{k+1/2};\xi_k^s), \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2} \right\rangle \\
\leq & \frac{(\eta_k^s)^2}{2} \left\| \nabla F(\mathbf{z}^s;\xi_k^s) - \nabla F(\mathbf{z}^s_{k+1/2};\xi_k^s) \right\|^2 + \frac{1}{2} \left\| \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2} \right\|^2 \\
\leq & \frac{(\eta_k^s L_z)^2}{2} \left\| \mathbf{z}^s_{k+1/2} - \mathbf{z}^s \right\|^2 + \frac{1}{2} \left\| \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2} \right\|^2 .
\end{aligned} \tag{47}
$$

Recall the definition of $\nabla\psi(\bar{\mathbf{z}}^s)$ and use the linearity of Bregman functions to yield:

$$
B(\mathbf{z}, \bar{\mathbf{z}}^s) - B(\mathbf{z}^s_{k+1/2}, \bar{\mathbf{z}}^s) = \left(\sum_{k=1}^{K_{s-1}} \alpha_{k-1}^{s-1} \right)^{-1} \sum_{j=1}^{K_{s-1}} \alpha_{j-1}^{s-1} \left[B(\mathbf{z}, \mathbf{z}_j^{s-1}) - B(\mathbf{z}^s_{k+1/2}, \mathbf{z}_j^{s-1}) \right]. \tag{48}
$$

According to the definition of $\mathbf{z}^s$, Jensen's Inequality and the strong-convexity of $\psi(\cdot)$, we get

$$
\begin{aligned}
&\left(\sum_{k=1}^{K_{s-1}} \alpha_{k-1}^{s-1} \right)^{-1} \sum_{j=1}^{K_{s-1}} -\alpha_{j-1}^{s-1} B(\mathbf{z}^s_{k+1/2}, \mathbf{z}_j^{s-1}) \\
\leq & \left(\sum_{k=1}^{K_{s-1}} \alpha_{k-1}^{s-1} \right)^{-1} \sum_{j=1}^{K_{s-1}} -\frac{\alpha_{j-1}^{s-1}}{2} \left\| \mathbf{z}^s_{k+1/2} - \mathbf{z}_j^{s-1} \right\|^2 \leq -\frac{1}{2} \left\| \mathbf{z}^s_{k+1/2} - \mathbf{z}^s \right\|^2 ,
\end{aligned} \tag{49}
$$

and

$$
-B(\mathbf{z}^s_{k+1}, \mathbf{z}^s_{k+1/2}) \leq -\frac{1}{2} \left\| \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2} \right\|^2 . \tag{50}
$$

Combining (47) (48) (49) (50) with (46) and casting out $-B(\mathbf{z}^s_{k+1/2}, \mathbf{z}^s_k)$, we have

$$
\begin{aligned}
&\eta_k^s \left\langle \nabla F(\mathbf{z}^s_{k+1/2}), \mathbf{z}^s_{k+1/2} - \mathbf{z} \right\rangle \\
\leq & \frac{(\eta_k^s L_z)^2}{2} \left\| \mathbf{z}^s_{k+1/2} - \mathbf{z}^s \right\|^2 + \frac{1}{2} \left\| \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2} \right\|^2 - (1-\alpha_k^s) B(\mathbf{z}^s_{k+1/2}, \mathbf{z}^s_k) \\
& + (1-\alpha_k^s) B(\mathbf{z}, \mathbf{z}^s_k) - B(\mathbf{z}, \mathbf{z}^s_{k+1}) + \left(\sum_{k=1}^{K_{s-1}} \alpha_{k-1}^{s-1} \right)^{-1} \alpha_k^s \sum_{j=1}^{K_{s-1}} \alpha_{j-1}^{s-1} B(\mathbf{z}, \mathbf{z}_j^{s-1}) \\
& - \frac{1}{2} \left\| \mathbf{z}^s_{k+1/2} - \mathbf{z}^s_{k+1} \right\|^2 - \frac{\alpha_k^s}{2} \left\| \mathbf{z}^s_{k+1/2} - \mathbf{z}^s \right\|^2 + \eta_k^s \left\langle \Delta_k^s, \mathbf{z} - \mathbf{z}^s_{k+1/2} \right\rangle \\
\leq & (1-\alpha_k^s) B(\mathbf{z}, \mathbf{z}^s_k) - B(\mathbf{z}, \mathbf{z}^s_{k+1}) + \left(\sum_{k=1}^{K_{s-1}} \alpha_{k-1}^{s-1} \right)^{-1} \alpha_k^s \sum_{j=1}^{K_{s-1}} \alpha_{j-1}^{s-1} B(\mathbf{z}, \mathbf{z}_j^{s-1}) \\
& + \eta_k^s \left\langle \Delta_k^s, \mathbf{z} - \mathbf{z}^s_{k+1/2} \right\rangle + \frac{(\eta_k^s L_z)^2 - \alpha_k^s}{2} \left\| \mathbf{z}^s_{k+1/2} - \mathbf{z}^s \right\|^2 .
\end{aligned} \tag{51}
$$

Recall the definition of Lyapunov function in Definition A.4 and periodically decaying sequence in Definition A.3. Together with the fact that $\mathbf{z}_{K_s}^s = \mathbf{z}_0^{s+1}$, we have

$$\begin{aligned}
&\sum_{k=0}^{K_s-1}\left[(1-\alpha_k^s)B(\mathbf{z},\mathbf{z}_k^s) - B(\mathbf{z},\mathbf{z}_{k+1}^s) + \left(\sum_{k=1}^{K_{s-1}}\alpha_{k-1}^{s-1}\right)^{-1}\alpha_k^s\sum_{j=1}^{K_{s-1}}\alpha_{j-1}^{s-1}B(\mathbf{z},\mathbf{z}_j^{s-1})\right] \\
=&\sum_{k=0}^{K_s-1}(1-\alpha_k^s)\left[B(\mathbf{z},\mathbf{z}_k^s) - B(\mathbf{z},\mathbf{z}_{k+1}^s)\right] - \sum_{k=0}^{K_s-1}\alpha_k^s B(\mathbf{z},\mathbf{z}_{k+1}^s) + \frac{\sum_{k=0}^{K_s-1}\alpha_k^s}{\sum_{k=1}^{K_{s-1}}\alpha_{k-1}^{s-1}}\sum_{j=1}^{K_{s-1}}\alpha_{j-1}^{s-1}B(\mathbf{z},\mathbf{z}_j^{s-1}) \\
=&(1-\alpha_0^s)B(\mathbf{z},\mathbf{z}_0^s) - (1-\alpha_{K_s-1}^s)B(\mathbf{z},\mathbf{z}_0^{s+1}) - \sum_{k=1}^{K_s}\alpha_{k-1}^s B(\mathbf{z},\mathbf{z}_k^s) + \frac{\sum_{k=1}^{K_s}\alpha_{k-1}^s}{\sum_{k=1}^{K_{s-1}}\alpha_{k-1}^{s-1}}\sum_{k=1}^{K_{s-1}}\alpha_{k-1}^{s-1}B(\mathbf{z},\mathbf{z}_k^{s-1}) \\
\le&(1-\alpha_0^s)B(\mathbf{z},\mathbf{z}_0^s) + \sum_{k=1}^{K_{s-1}}\alpha_{k-1}^{s-1}B(\mathbf{z},\mathbf{z}_k^{s-1}) - (1-\alpha_0^{s+1})B(\mathbf{z},\mathbf{z}_0^{s+1}) - \sum_{k=1}^{K_s}\alpha_{k-1}^s B(\mathbf{z},\mathbf{z}_k^s) \\
=&\Psi_s(\mathbf{z}) - \Psi_{s+1}(\mathbf{z}).
\end{aligned} \tag{52}$$

we complete the proof by summing both sides of (51) by index k from 0 to $K_s - 1$ and use the above relation. ☐

Lemma A.10. *Denote the filtration generated by our algorithm by* $\mathcal{F} = \{\mathcal{F}_k^s\}_{k\in[K_s]^0, s\in[S]^0}$. *Let* $\eta_k^s = \frac{\sqrt{\alpha_k^s(1-\theta_k^s)}}{L_z}$, $\theta_k^s \in (0.8, 0.99)$. *Then the following recurrence holds:*

$$\mathbb{E}[\Psi_{s+1}(\mathbf{z}_*)\Big|\mathcal{F}_0^s,\cdots,\mathcal{F}_{K_s-1}^s] \le \Psi_s(\mathbf{z}_*) - \frac{1}{2}\sum_{k=0}^{K_s-1}\mathbb{E}\left[\alpha_k^s\theta_k^s\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\|^2\Big|\mathcal{F}_0^s,\cdots,\mathcal{F}_{K_s-1}^s\right]. \tag{53}$$

Proof. Since $\mathbf{z}_* = (\mathbf{w}_*; \mathbf{q}_*)$ is the solution to (3), then we have

$$F(\mathbf{w}_*, \mathbf{q}_{k+1/2}^s) \le F(\mathbf{w}_*, \mathbf{q}_*) \le F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_*). \tag{54}$$

Recall convexity assumption (Assumption 3.6) and the linearity of $\mathbf{q}$, we have

$$\begin{aligned}
F(\mathbf{w}_*, \mathbf{q}_{k+1/2}^s) &\ge F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_{k+1/2}^s) + \left\langle\nabla_{\mathbf{w}}F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_{k+1/2}^s), \mathbf{w}_* - \mathbf{w}_{k+1/2}^s\right\rangle, \\
F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_*) &= F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_{k+1/2}^s) + \left\langle\nabla_{\mathbf{q}}F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_{k+1/2}^s), \mathbf{q}_* - \mathbf{q}_{k+1/2}^s\right\rangle.
\end{aligned} \tag{55}$$

Therefore,

$$\begin{aligned}
&\left\langle\nabla F(\mathbf{z}_{k+1/2}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z}_*\right\rangle \\
=&\left\langle\nabla_{\mathbf{w}}F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_{k+1/2}^s), \mathbf{w}_{k+1/2}^s - \mathbf{w}_*\right\rangle - \left\langle\nabla_{\mathbf{q}}F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_{k+1/2}^s), \mathbf{q}_{k+1/2}^s - \mathbf{q}_*\right\rangle \\
\ge&F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_{k+1/2}^s) - F(\mathbf{w}_*, \mathbf{q}_{k+1/2}^s) + F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_*) - F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_{k+1/2}^s) \\
=&F(\mathbf{w}_{k+1/2}^s, \mathbf{q}_*) - F(\mathbf{w}_*, \mathbf{q}_{k+1/2}^s) \ge 0.
\end{aligned} \tag{56}$$

Then plugging the above inequality to the LHS of Lemma A.9, we have

$$0 \le \Psi_s(\mathbf{z}_*) - \Psi_{s+1}(\mathbf{z}_*) + \sum_{k=0}^{K_s-1}\left[\eta_k^s\left\langle\Delta_k^s, \mathbf{z} - \mathbf{z}_{k+1/2}^s\right\rangle + \frac{(\eta_k^s L_z)^2 - \alpha_k^s}{2}\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\|^2\right]. \tag{57}$$

For any *fixed* $\mathbf{z} \in \mathcal{Z}$, Δ_k^s is conditional independent from $\mathbf{z}_{k+1/2}^s - \mathbf{z}$. By the tower rule of expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{K_s-1}\eta_k^s\left\langle\Delta_k^s, \mathbf{z}_{k+1/2}^s - \mathbf{z}_*\right\rangle\Big|\mathcal{F}_0^s,\cdots,\mathcal{F}_{K_s-1}^s\right] \\
=&\sum_{k=0}^{K_s-1}\mathbb{E}\left[\eta_k^s\left\langle\mathbb{E}\left[\Delta_k^s|\mathcal{F}_k^s\right], \mathbf{z}_{k+1/2}^s - \mathbf{z}_*\right\rangle\Big|\mathcal{F}_{k+1}^s,\cdots,\mathcal{F}_{K_s-1}^s\right] = 0.
\end{aligned} \tag{58}$$

Notice that $\mathbb{E}\left[\Psi_s(\mathbf{z}_*)|\mathcal{F}_0^s,\cdots,\mathcal{F}_{K_s-1}^s\right] = \Psi_s(\mathbf{z}_*)$. By Lemma A.9 and a simple rearrangement we can get the result. ☐

Corollary A.11. *Under the conditions of Lemma A.10, we have*

$$\sum_{s=0}^{\infty}\sum_{k=0}^{K_s-1}\mathbb{E}\left[\alpha_k^s\theta_k^s\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2\right]\leq 2\Psi_0(\mathbf{z}_*)\leq 2\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z}) \tag{59}$$

Proof. Summing the inequality in Lemma A.10, noticing the non-negativity of $\Psi_s(\mathbf{z})$ together with the tower rule suffices to prove this corollary. □

Lemma A.12. *With η_k^s set in Lemma A.10, we have*

$$\left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\right)\epsilon(\mathbf{z}_S)\leq\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})+\max_{\mathbf{z}\in\mathcal{Z}}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\left\langle\Delta_k^s,\mathbf{z}-\mathbf{z}_{k+1/2}^s\right\rangle-\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\frac{\alpha_k^s\theta_k^s}{2}\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2. \tag{60}$$

Proof. The following result is a direct derivation from the Assumption 3.6 and Lemma A.9:

$$\begin{aligned}\left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\right)\epsilon(\mathbf{z}_S)&\leq\left[\max_{\mathbf{q}\in\Delta_m}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^sF(\mathbf{w}_{k+1/2}^s,\mathbf{q})-\min_{\mathbf{w}\in\mathcal{W}}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^sF(\mathbf{w},\mathbf{q}_{k+1/2}^s)\right]\\&\leq\max_{\mathbf{z}\in\mathcal{Z}}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\left\langle\nabla F(\mathbf{z}_{k+1/2}^s),\mathbf{z}_{k+1/2}^s-\mathbf{z}\right\rangle\\&\leq\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})+\max_{\mathbf{z}\in\mathcal{Z}}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\left\langle\eta_k^s\Delta_k^s,\mathbf{z}-\mathbf{z}_{k+1/2}^s\right\rangle-\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\frac{\alpha_k^s\theta_k^s}{2}\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2.\end{aligned} \tag{61}$$

□

A.3. Convergence and Complexity

Lemma A.13. *Under the conditions in Lemma A.10, we have*

$$\mathbb{E}\left[\epsilon(\mathbf{z}_S)\right]\leq L_z\left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\sqrt{\alpha_k^s(1-\theta_k^s)}\right)^{-1}\left[1+\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})\right]. \tag{62}$$

Proof. It's obvious to note that $\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\left\langle\Delta_k^s,\mathbf{z}-\mathbf{z}_{k+1/2}^s\right\rangle$ is a martingale difference sequence for any *fixed* $\mathbf{z}\in\mathcal{Z}$. The existence of the maximum operation on the RHS of (60) deprives the inner product $\langle\Delta_k^s,\mathbf{z}-\mathbf{z}_{k+1/2}^s\rangle$ from being a martingale difference sequence. We need to apply a classical technique called "ghost iterate" (Nemirovski et al., 2009) to switch the order of maximization and expectation, and thus eliminating the dependency for $\mathbf{z}$. Image there is an online algorithm performing stochastic mirror descent (SMD):

$$\mathbf{y}_{k+1}^s=\arg\min_{\mathbf{y}\in\mathcal{Z}}\{\langle-\eta_k^s\Delta_k^s,\mathbf{y}-\mathbf{y}_k^s\rangle+B(\mathbf{y},\mathbf{y}_k^s)\},\ \mathbf{y}_0^{s+1}=\mathbf{y}_{K_s}^s\quad\forall\, s\in[S]^0,k\in[K_s]^0. \tag{63}$$

Also, we define $\mathbf{y}_0^0=\mathbf{z}_0^0=\arg\min_{\mathbf{z}\in\mathcal{Z}}\psi(\mathbf{z})$. According to Nemirovski et al. (2009, Lemma 6.1), we have for any $\mathbf{z}\in\mathcal{Z}$:

$$\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\langle\eta_k^s\Delta_k^s,\mathbf{z}-\mathbf{y}_k^s\rangle\leq B(\mathbf{z},\mathbf{z}_0^0)+\frac{1}{2}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}(\eta_k^s)^2\left\|\Delta_k^s\right\|_*^2. \tag{64}$$

Now that we have decoupled the dependency, it's safe for us to assert that $\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\langle\eta_k^s\Delta_k^s,\mathbf{y}_k^s-\mathbf{z}_{k+1/2}^s\rangle$ is a martingale difference sequence, since $\mathbf{y}_k^s-\mathbf{z}_{k+1/2}^s$ is conditionally independent of Δ_k^s.

Firstly, we show that Δ_k^s is uniformly bounded above:

$$
\begin{aligned}
\|\Delta_k^s\|_* &= \left\|\nabla F(\mathbf{z}_{k+1/2}^s;\xi_k^s) - \nabla F(\mathbf{z}^s;\xi_k^s) + \nabla F(\mathbf{z}^s) - \nabla F(\mathbf{z}_{k+1/2}^s)\right\|_* \\
&\le \left\|\nabla F(\mathbf{z}_{k+1/2}^s;\xi_k^s) - \nabla F(\mathbf{z}^s;\xi_k^s)\right\|_* + \left\|\mathbb{E}\left[\nabla F(\mathbf{z}_{k+1/2}^s;\xi_k^s) - \nabla F(\mathbf{z}^s;\xi_k^s)\right]\right\|_* \\
&\le \left\|\nabla F(\mathbf{z}_{k+1/2}^s;\xi_k^s) - \nabla F(\mathbf{z}^s;\xi_k^s)\right\|_* + \mathbb{E}\left[\left\|\nabla F(\mathbf{z}_{k+1/2}^s;\xi_k^s) - \nabla F(\mathbf{z}^s;\xi_k^s)\right\|_*\right] \\
&\le L_z\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\| + \mathbb{E}\left[L_z\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\|\right] \\
&\le 2\sqrt{2}L_z + \mathbb{E}\left[2\sqrt{2}L_z\right] \\
&= 4\sqrt{2}L_z.
\end{aligned} \tag{65}
$$

The above inequality is ensured by the continuity introduced in Lemma 4.3.

Then we define $V_k^s = \langle \eta_k^s\Delta_k^s, \mathbf{y}_k^s - \mathbf{z}_{k+1/2}^s\rangle$. The following holds:

$$
\begin{aligned}
&\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{Z}}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\left\langle\Delta_k^s, \mathbf{z} - \mathbf{z}_{k+1/2}^s\right\rangle\right] \\
=&\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{Z}}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\left\langle\eta_k^s\Delta_k^s, \mathbf{z} - \mathbf{y}_k^s\right\rangle\right] + \mathbb{E}\left[\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\underbrace{\left\langle\eta_k^s\Delta_k^s, \mathbf{y}_k^s - \mathbf{z}_{k+1/2}^s\right\rangle}_{V_k^s}\right] \\
\overset{(64)}{\le}&\max_{\mathbf{z}\in\mathcal{Z}} B(\mathbf{z}, \mathbf{z}_0^0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[\frac{(\eta_k^s)^2}{2}\|\Delta_k^s\|_*^2\right] + \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[\mathbb{E}[V_k^s|\mathcal{F}_k^s]\right] \\
\overset{(65)}{\le}&1 + \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[2(\eta_k^s L_z)^2\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\|^2\right] \\
=&1 + \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[2\alpha_k^s(1-\theta_k^s)\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\|^2\right].
\end{aligned} \tag{66}
$$

Combining it with Lemma A.12:

$$
\begin{aligned}
&\left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\right)\mathbb{E}\left[\epsilon(\mathbf{z}_S)\right] \\
\le&\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z}) + \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{Z}}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\left\langle\Delta_k^s, \mathbf{z} - \mathbf{z}_{k+1/2}^s\right\rangle\right] - \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[\frac{\alpha_k^s\theta_k^s}{2}\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\|^2\right] \\
\le&\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z}) + 1 + \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[\alpha_k^s(2 - \frac{5}{2}\theta_k^s)\left\|\mathbf{z}_{k+1/2}^s - \mathbf{z}^s\right\|^2\right] \\
\le&1 + \max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z}),
\end{aligned} \tag{67}
$$

which concludes our proof by dividing $\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s$ to both sides of the above inequality. □

Proof of Theorem 4.4. First, we show that $\max_{\mathbf{z}\in\mathcal{Z}} \Psi_0(\mathbf{z})$ is bounded under the given conditions:

$$\begin{aligned}\max_{\mathbf{z}\in\mathcal{Z}} \Psi_0(\mathbf{z}) &= \max_{\mathbf{z}\in\mathcal{Z}} \left\{ (1-\alpha_0^0) B(\mathbf{z}, \mathbf{z}_0^0) + \sum_{k=1}^{K_{-1}} \alpha_{k-1}^{-1} B(\mathbf{z}, \mathbf{z}_k^{-1}) \right\} \\ &= \max_{\mathbf{z}\in\mathcal{Z}} \left\{ (1-\alpha_0^0) B(\mathbf{z}, \mathbf{z}_0) + \sum_{k=1}^{K_{-1}} \alpha_{k-1}^{-1} B(\mathbf{z}, \mathbf{z}_0) \right\} \\ &\leq 1 - \alpha_0^0 + \sum_{k=1}^{K_{-1}} \alpha_{k-1}^{-1} \leq 2.\end{aligned} \tag{68}$$

Under the given parameters, we have

$$\mathbb{E}\left[\epsilon(\mathbf{z}_S)\right] \leq L_z \left(\sum_{s=0}^{S-1} \sum_{k=0}^{K_s-1} \sqrt{\frac{(1-\theta_k^s)}{K}} \right)^{-1} \left[1 + \max_{\mathbf{z}\in\mathcal{Z}} \Psi_0(\mathbf{z}) \right] \overset{(68)}{\leq} \frac{30 L_z}{S\sqrt{K}}. \tag{69}$$

From Lemma 4.3 we know that $L_z = \mathcal{O}(\sqrt{\ln m})$, which concludes our proof. □

Proof of Corollary 4.6. The inner loop of Algorithm 1 consumes $\mathcal{O}(m)$ computations per iteration. For outer loop s, the full gradient $\nabla F(\mathbf{z}^s)$ is calculated, requiring $\mathcal{O}(m\bar{n})$ computations. So the algorithms consume $\mathcal{O}\left(mKS + m\bar{n}S\right)$ in total. Aiming to set the two terms at the same order, we choose $K = \Theta(\bar{n})$. As a consequence, $S = \mathcal{O}\left(\frac{L_z}{\varepsilon\sqrt{\bar{n}}}\right)$. Plugging this into $\mathcal{O}\left(mKS + m\bar{n}S\right)$ we derive an $\mathcal{O}\left(\frac{L_z m\sqrt{\bar{n}}}{\varepsilon}\right)$ computation complexity. With $L_z = \mathcal{O}(\sqrt{\ln m})$ taken into consideration, the total computation complexity to reach ε-accuracy is $\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$. □

B. Analysis for Empirical MERO

We present the omitted proofs for empirical MERO in this section.

B.1. Optimization Error

The following proof verifies that the optimization error for the approximated problem is under control, which is proved to be essential in our two-stage schema for empirical MERO.

Proof of Lemma 5.1. For any $\mathbf{z} = (\mathbf{w}; \mathbf{q}) \in \mathcal{Z}$, we have

$$\left|\underline{F}(\mathbf{z}) - \hat{\underline{F}}(\mathbf{z})\right| = \left|\sum_{i=1}^{m} \mathbf{q}_i \left[\underline{R}_i(\mathbf{w}) - \hat{\underline{R}}_i(\mathbf{w})\right]\right| = \left|\sum_{i=1}^{m} \mathbf{q}_i \left[\hat{R}_i^* - R_i^*\right]\right| \leq \max_{i\in[m]} \left\{\hat{R}_i^* - R_i^*\right\}. \tag{70}$$

For convenience, we denote $\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \underline{F}(\mathbf{w}, \bar{\mathbf{q}})$ and $\tilde{\mathbf{q}} = \arg\min_{\mathbf{q}\in\Delta_m} \underline{F}(\bar{\mathbf{w}}, \mathbf{q})$. Therefore, we can complete our proof by simple algebra.

$$\begin{aligned}
\underline{\epsilon}(\bar{\mathbf{z}}) = \underline{F}(\bar{\mathbf{w}}, \tilde{\mathbf{q}}) - \underline{F}(\tilde{\mathbf{w}}, \bar{\mathbf{q}}) &\overset{(70)}{\leq} \hat{\underline{F}}(\bar{\mathbf{w}}, \tilde{\mathbf{q}}) - \hat{\underline{F}}(\tilde{\mathbf{w}}, \bar{\mathbf{q}}) + 2\max_{i\in[m]} \left\{\hat{R}_i^* - R_i^*\right\} \\
&\leq \max_{\mathbf{q}\in\Delta_m} \hat{\underline{F}}(\bar{\mathbf{w}}, \mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \hat{\underline{F}}(\mathbf{w}, \bar{\mathbf{q}}) + 2\max_{i\in[m]} \left\{\hat{R}_i^* - R_i^*\right\} \\
&= \hat{\underline{\epsilon}}(\bar{\mathbf{z}}) + 2\max_{i\in[m]} \left\{\hat{R}_i^* - R_i^*\right\}.
\end{aligned} \tag{71}$$

□

B.2. Stage 1: Excess Empirical Risk Convergence

At the beginning of this section, we present a useful lemma to bridge the variance-reduced property of ALEG and the martingale difference sequence.

Lemma B.1. *(Bernstein's Inequality for Martingales (Cesa-Bianchi & Lugosi, 2006)) Let $\{V_t\}_{t=1}^T$ be a martingale difference sequence with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t=1}^T$ bounded above by V, i.e. $|V_t| \leq V$. If the sum of the conditional variances is bounded, i.e. $\sum_{t=1}^T \mathbb{E}[V_t^2|\mathcal{F}_t] \leq \sigma^2$, then for any $\delta \in (0, 1]$,*

$$\mathbb{P}\left(\sum_{t=1}^{T} V_t > \sigma\sqrt{2\ln\frac{1}{\delta}} + \frac{2}{3}V\ln\frac{1}{\delta}\right) \leq \delta. \tag{72}$$

Next, we present the proof of the high probability bound in Theorem 5.2.

Proof of Theorem 5.2. At the beginning, we briefly discuss how ALEG can be used as an ERM oracle. Under the circumstance of $m = 1$, we notice that Δ_m reduces to a singleton. The original empirical GDRO problem (3) can be rewritten as

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_1} \{F(\mathbf{w}, \mathbf{q}) = R_i(\mathbf{w})\} \Longleftrightarrow \min_{\mathbf{w}\in\mathcal{W}} \{R_i(\mathbf{w})\}. \tag{73}$$

As a consequence, the merged gradient w.r.t. $\mathbf{q}$ vanishes. In this case, when we talk about the smoothness of $F(\mathbf{w}, \mathbf{q})$, we are actually focusing on the smoothness of $R_i(\mathbf{w})$. We can conclude from Assumption 3.4 that $R_i(\cdot)$ is L-smooth. Moreover, the duality gap for the output of Algorithm 1 $\bar{\mathbf{z}}_i = (\bar{\mathbf{w}}_i, \bar{\mathbf{q}}_i)$ of (73) satisfies:

$$\epsilon(\bar{\mathbf{z}}_i) = R_i(\bar{\mathbf{w}}_i) - \min_{\mathbf{w}\in\mathcal{W}} R_i(\mathbf{w}) = \hat{R}_i^* - R_i^*, \quad \forall\, i \in [m], \tag{74}$$

which is exactly the excess empirical risk for group i. Recall the "ghost iterate" technique demonstrated in Appendix A.3, we define $V_k^s = \langle \eta_k^s \Delta_k^s, \mathbf{y}_k^s - \mathbf{z}_{k+1/2}^s \rangle$. Firstly, we show that V_k^s is uniformly bounded above:

$$|V_k^s| \leq \eta_k^s \|\Delta_k^s\|_* \left\|\mathbf{y}_k^s - \mathbf{z}_{k+1/2}^s\right\| \leq 2\sqrt{2}\eta_k^s \|\Delta_k^s\|_* \leq 16\eta_k^s L = 16\sqrt{\alpha_k^s(1-\theta_k^s)} \leq \frac{16}{\sqrt{5\bar{n}}}. \tag{75}$$

Secondly, we bound the sum of conditional variance of $\{V_k^s\}_{k\in[K_s]^0, s\in[S]^0}$. By (65), the definition of θ_k^s and Corollary A.11,

$$\begin{aligned}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[(V_k^s)^2|\mathcal{F}_k^s\right] &\leq \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[8(\eta_k^s)^2\left\|\Delta_k^s\right\|_*^2\middle|\mathcal{F}_k^s\right]\\ &\leq \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\mathbb{E}\left[32(\eta_k^s L)^2\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2\middle|\mathcal{F}_k^s\right]\\ &= \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}32\mathbb{E}\left[\alpha_k^s(1-\theta_k^s)\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2\middle|\mathcal{F}_k^s\right]\\ &\leq \sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}8\mathbb{E}\left[\alpha_k^s\theta_k^s\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2\middle|\mathcal{F}_k^s\right]\\ &\leq 32\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z}).\end{aligned} \tag{76}$$

Finally, we can use Lemma B.1 together with the union bound to conclude that with probability at least $1-\delta$

$$\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}V_k^s \leq 32\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})\sqrt{2\ln\frac{S}{\delta}}+\frac{32}{3\sqrt{5\bar{n}}}\ln\frac{S}{\delta}. \tag{77}$$

From Lemma A.12 and the previous result (64), along with the boundness of Bregman function and the definition of θ_k^s, we have

$$\begin{aligned}\epsilon(\bar{\mathbf{z}}_i) \leq& \left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\right)^{-1}\Bigg[\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})+\max_{\mathbf{z}\in\mathcal{Z}}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\langle\eta_k^s\Delta_k^s,\mathbf{z}-\mathbf{y}_k^s\rangle\\ &+\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\langle\eta_k^s\Delta_k^s,\mathbf{z}_{k+1/2}^s-\mathbf{y}_k^s\rangle-\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\frac{\alpha_k^s\theta_k^s}{2}\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2\Bigg]\\ \leq& \left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\right)^{-1}\Bigg[\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})+\max_{\mathbf{z}\in\mathcal{Z}}B(\mathbf{z},\mathbf{z}_0^0)+\frac{1}{2}\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}(\eta_k^s)^2\left\|\Delta_k^s\right\|_*^2\\ &+\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}V_k^s-\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\frac{\alpha_k^s\theta_k^s}{2}\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2\Bigg]\\ \leq& \left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\eta_k^s\right)^{-1}\Bigg[\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})+1+\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\frac{2\alpha_k^s(1-\theta_k^s)}{L^2}L^2\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2\\ &-\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\frac{\alpha_k^s\theta_k^s}{2}\left\|\mathbf{z}_{k+1/2}^s-\mathbf{z}^s\right\|^2+\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}V_k^s\Bigg]\\ \leq& L\left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\sqrt{\alpha_k^s(1-\theta_k^s)}\right)^{-1}\left[1+\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})+\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}V_k^s\right].\end{aligned} \tag{78}$$

Combining (77) and (78) we derive that with probability at least $1-\delta$,

$$\begin{aligned}\hat{R}_i^*-R_i^* \leq& L\left(\sum_{s=0}^{S-1}\sum_{k=0}^{K_s-1}\sqrt{\alpha_k^s(1-\theta_k^s)}\right)^{-1}\left[1+\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})+32\max_{\mathbf{z}\in\mathcal{Z}}\Psi_0(\mathbf{z})\sqrt{2\ln\frac{S}{\delta}}+\frac{32}{3\sqrt{5\bar{n}}}\ln\frac{S}{\delta}\right]\\ \leq& \frac{10L}{S\sqrt{\bar{n}}}\left(3+64\sqrt{2\ln\frac{S}{\delta}}+\frac{32}{3\sqrt{5\bar{n}}}\ln\frac{S}{\delta}\right),\end{aligned} \tag{79}$$

where the last inequality uses (68). The above relation needs to hold for every group i, which necessitates the usage of the union bound tool. We can deduce that with probability at least $1-\delta$,

$$\max_{i\in[m]}\left\{\hat{R}_i^* - R_i^*\right\} \leq \frac{10L}{S\sqrt{\bar{n}}}\left(3+64\sqrt{2\ln\frac{mS}{\delta}}+\frac{32}{3\sqrt{5\bar{n}}}\ln\frac{mS}{\delta}\right). \tag{80}$$

Recall the relation between S and T in Theorem 5.2, we derive that

$$\max_{i\in[m]}\left\{\hat{R}_i^* - R_i^*\right\} \leq \mathcal{O}\left(\frac{1}{T}\left(\sqrt{\ln\frac{mT}{\sqrt{\bar{n}}\delta}}+\frac{1}{\sqrt{\bar{n}}}\ln\frac{mT}{\sqrt{\bar{n}}\delta}\right)\right), \tag{81}$$

which is equivalent to (27). □

B.3. Stage 2: Empirical GDRO Solver Convergence

The second stage of Algorithm 2 solves (24) by Algorithm 1. The main idea is to combine the convergence results in Section 4 with Lemma 5.1.

Proof of Theorem 5.3. The proof of Theorem 5.2 actually provides a high probability bound for empirical GDRO. By substituting L with L_z in the previous derivations, the following holds with probability at least $1-\frac{\delta}{2}$,

$$\underline{\hat{\epsilon}}(\bar{\mathbf{z}}) \leq \frac{10L_z}{T}\left(3+32\sqrt{2\ln\frac{2T}{\sqrt{\bar{n}}\delta}}+\frac{64}{3\sqrt{5\bar{n}}}\ln\frac{2T}{\sqrt{\bar{n}}\delta}\right). \tag{82}$$

Since we only need to provide a theoretical guarantee for a single approximate MERO problem, the union bound appeared in (81) is unnecessary. Based on (81), the following holds with probability with probability at least $1-\frac{\delta}{2}$,

$$\max_{i\in[m]}\left\{\hat{R}_i^* - R_i^*\right\} \leq \frac{10L}{T}\left(3+64\sqrt{2\ln\frac{2mT}{\sqrt{\bar{n}}\delta}}+\frac{32}{3\sqrt{5\bar{n}}}\ln\frac{2mT}{\sqrt{\bar{n}}\delta}\right). \tag{83}$$

Together, we immediately derive that with probability at least $1-\delta$,

$$\begin{aligned}
\underline{\epsilon}(\bar{\mathbf{z}}) &\overset{(26)}{\leq} \underline{\hat{\epsilon}}(\bar{\mathbf{z}}) + 2\max_{i\in[m]}\left\{\hat{R}_i^* - R_i^*\right\}\\
&\leq \frac{10}{T}\left[3(L_z+2L)+64\sqrt{2}\left(L_z\sqrt{\ln\frac{2T}{\sqrt{\bar{n}}\delta}}+2L\sqrt{\ln\frac{2mT}{\sqrt{\bar{n}}\delta}}\right)+\frac{32}{3\sqrt{5\bar{n}}}\left(L_z\ln\frac{2T}{\sqrt{\bar{n}}\delta}+2L\ln\frac{2m}{\sqrt{\bar{n}}\delta}\right)\right]\\
&= \mathcal{O}\left(\frac{1}{T}\left(\sqrt{\ln m\ln\frac{T}{\sqrt{\bar{n}}\delta}}+\sqrt{\ln\frac{mT}{\sqrt{\bar{n}}\delta}}+\frac{1}{\sqrt{\bar{n}}}\left(\sqrt{\ln m}\ln\frac{T}{\sqrt{\bar{n}}\delta}+\ln\frac{mT}{\sqrt{\bar{n}}\delta}\right)\right)\right)\\
&\leq \mathcal{O}\left(\frac{1}{T}\left(\sqrt{\ln m\ln\frac{T}{\sqrt{\bar{n}}\delta}}+\sqrt{\ln\frac{mT}{\sqrt{\bar{n}}\delta}}+\sqrt{\frac{\ln m}{\bar{n}}}\ln\frac{T}{\sqrt{\bar{n}}\delta}\right)\right),
\end{aligned} \tag{84}$$

where the last inequality holds under the ordinary case where $m \leq \mathcal{O}(\bar{n})$. □

At the end of this section, we calculate the number of computations needed to run Algorithm 2.

Proof of Corollary 5.4. The proof of Corollary 5.4 is analogous to the proof of Corollary 4.6. According to Theorem 5.3, we need a budget of $T=\mathcal{O}\left(\frac{\sqrt{\ln m}}{\varepsilon}\right)$ to reach ε-accuracy if logarithmic factors for $\frac{T}{\sqrt{\bar{n}}}$ are overlooked. This requires $S=\mathcal{O}\left(\frac{\sqrt{\ln m}}{\sqrt{\bar{n}}\varepsilon}\right)$ correspondingly. In the first stage of Algorithm 2, the computation complexity is $\mathcal{O}\left(\sum_{i=1}^m \bar{n}S+n_iS\right)=\mathcal{O}\left(m\bar{n}S\right)=\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$. In the second stage of Algorithm 2, the computation complexity is $\mathcal{O}\left(mKS+m\bar{n}S\right)=\mathcal{O}\left(m\bar{n}S\right)=\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$. As a consequence, adding the computation complexity from two stages together gives the computation complexity of $\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$. To derive the final result, the neglected $\ln\frac{T}{\sqrt{\bar{n}}}$ needs to be taken into consideration. Therefore, the total complexity for Algorithm 2 to reach ε-accuracy is $\mathcal{O}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\ln\frac{\sqrt{\ln m}}{\sqrt{\bar{n}}\varepsilon}\right)$, which is also denoted by $\tilde{\mathcal{O}}\left(\frac{m\sqrt{\bar{n}}\ln m}{\varepsilon}\right)$. $\square$

C. Revisiting MPVR as a Black Box

In this section, we illustrate the extension of MPVR (Alacaoglu & Malitsky, 2022) to solve the empirical GDRO problem defined in (3). To run MPVR as a black box, we need to specify the problem formulation, the construction of stochastic gradients, and the Lipschitz continuity of the stochastic gradients. After these steps, we can derive a suboptimal complexity and then discuss how the proposed approach goes beyond the application of existing techniques.

C.1. Problem Formulation

Since MPVR only deals with one-level finite-sums, the objectives in the empirical GDRO problem should be rewritten as

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{ F(\mathbf{w},\mathbf{q}) := \sum_{l=1}^{m\bar{n}} \frac{\mathbf{q}_{l_i}}{n_{l_i}} \ell(\mathbf{w};\xi_l) \right\}. \tag{85}$$

where

$$\{\xi_l\}_{l=1}^{m\bar{n}} := \left\{\xi_{l_i l_j}\right\}_{l=1}^{m\bar{n}} = \{\xi_{ij}\}_{j\in[n_i],i\in[m]} \tag{86}$$

with l_i, l_j determined by

$$l_i \in [m] : \sum_{p=1}^{l_i-1} n_p < l \leq \sum_{p=1}^{l_i} n_p \quad \text{and} \quad l_j = l - \sum_{p=1}^{l_i-1} n_p. \tag{87}$$

The above relations show the obvious bijective mapping between one-level finite-sum index l and two-level finite-sum index (l_i, l_j).

From (85) we can see that the number of finite-sum components for MPVR is $m\bar{n}$. In the next part, we construct the stochastic gradient, which comes from accessing one of the members of $m\bar{n}$ components.

C.2. Construction of Stochastic Gradients

The construction of the stochastic gradients is essential for variance-reduced methods. Such construction is also closely related to the sampling strategy. Alacaoglu & Malitsky (2022) introduce uniform sampling and importance sampling as two main stochastic oracles into their approach. The former strategy treats every member of $m\bar{n}$ equally and therefore generates a uniform distribution among all indices. The latter considers the continuity of each component, assigning each component a probability proportional to its Lipschitz constants.

C.2.1. Uniform Sampling

Denote by Unif$(\cdot)$ as the uniform distribution and $\nabla F(\mathbf{z};\xi_k^s)$ as the stochastic gradient for the k-th inner loop, s-th outer loop. From the above discussions, the chosen index is generated by

$$l \sim \text{Unif}(m\bar{n}). \tag{88}$$

Consequently, the stochastic gradient can be constructed as follows

$$\nabla F(\mathbf{z};\xi_k^s) := \begin{pmatrix} \frac{m\bar{n}}{n_{l_i}} \mathbf{q}_{l_i} \nabla \ell(\mathbf{w};\xi_l) \\ -\left[0,\cdots,\underbrace{\frac{m\bar{n}}{n_{l_i}}\ell(\mathbf{w};\xi_l)}_{l_i\text{-th element}},\cdots,0\right]^T \end{pmatrix}. \tag{89}$$

The following property is necessary for MPVR. The proof is deferred to Appendix C.4.

Lemma C.1. *(First-order and second-order moment information of the stochastic gradient from uniform sampling) With expectation taken over ξ_k^s, the following properties holds:*

1. $\mathbb{E}\left[\nabla F(\mathbf{z};\xi_k^s)\right] = \nabla F(\mathbf{z})$;

2. $\mathbb{E}\left[\|\nabla F(\mathbf{z}_1;\xi_k^s) - \nabla F(\mathbf{z}_2;\xi_k^s)\|_*^2\right] \le L_{\mathbf{u}}^2 \|\mathbf{z}_1 - \mathbf{z}_2\|^2, \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$;

where the Lipschitz constant $L_{\mathbf{u}}$ is defined as

$$L_{\mathbf{u}} := 2D_w \max\left\{\sqrt{2D_w^2 L^2 m \frac{\bar{n}}{n_{\min}} + G^2 m^2 \ln m \frac{\bar{n}}{\bar{n}_h}}, G\sqrt{2m \ln m \frac{\bar{n}}{n_{\min}}}\right\}. \tag{90}$$

with n_h being the harmonic average of the number of samples and $n_{\min}$ being the minimal number of samples amongst m groups, i.e.,

$$\bar{n}_h := \frac{m}{\sum_{i=1}^m \frac{1}{n_i}}, \quad n_{\min} := \min_{i\in[m]} n_i. \tag{91}$$

C.2.2. IMPORTANCE SAMPLING

At first glance, importance sampling might be the same as uniform sampling because the Lipschitz constant for each loss function is the same according to Assumption 3.4. Whereas, from (85) we can see that the finite-sum component has a coefficient of $\frac{1}{n_i}$. Hence, the Lipschitzness is shifted by a factor of $\frac{1}{n_i}$. Therefore, we assign the probability proportional to this coefficient, i.e., we endow a larger sampling probability for a group with more samples (larger n_i).

Based on the above discussions, we present the process of importance sampling as follows:

$$l_i \sim \text{Unif}(m), \quad l_j | l_i \sim \text{Unif}(n_{l_i}). \tag{92}$$

Naturally, the following stochastic gradient is produced:

$$\nabla F(\mathbf{z};\xi_k^s) := \begin{pmatrix} m\mathbf{q}_{l_i} \nabla \ell(\mathbf{w};\xi_l) \\ -\left[0,\cdots,\underbrace{m\ell(\mathbf{w};\xi_l)}_{l_i\text{-th element}},\cdots,0\right]^T \end{pmatrix}, \tag{93}$$

Similar to Appendix C.2.1, we have the following lemma to quantify the first-order and second-order moments of the stochastic gradient in (93).

Lemma C.2. *(First-order and second-order moment information of the stochastic gradient from importance sampling) With expectation taken over ξ_k^s, the following properties holds:*

1. $\mathbb{E}\left[\nabla F(\mathbf{z};\xi_k^s)\right] = \nabla F(\mathbf{z})$;
2. $\mathbb{E}\left[\|\nabla F(\mathbf{z}_1;\xi_k^s) - \nabla F(\mathbf{z}_2;\xi_k^s)\|_*^2\right] \le L_{\mathbf{i}}^2 \|\mathbf{z}_1 - \mathbf{z}_2\|^2, \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$;

where the Lipschitz constant $L_{\mathbf{i}}$ is defined as

$$L_{\mathbf{i}} := 2D_w \max\left\{\sqrt{2D_w^2 L^2 m + G^2 m^2 \ln m}, G\sqrt{2m \ln m}\right\}. \tag{94}$$

Remark C.3. Our Lipschitzness in Lemma 4.3 is different from the ones in Lemmas C.1 and C.2. The definition of the Lipschitz continuity in Alacaoglu & Malitsky (2022) is reflected by the second-order moment of the stochastic gradient. While ours is reflected by the square dual norm of the stochastic gradient, which makes our formulation stronger than MPVR. The high probability bound we provide justifies this formulation since the continuity w.r.t. the second-order moment does not hold with probability 1.

C.3. Loose Complexity Bound

With the preparations in previous sections, we can formally apply MPVR to solve empirical GDRO as shown in Algorithm 3. For completeness, we will present the following theoretical guarantee for Algorithm 3. The proofs are deferred to Appendix C.4.

Algorithm 3 MPVR for Empirical GDRO

Input: Risk function $\{R_i(\mathbf{w})\}_{i\in[m]}$, epoch number S, iteration number K, learning rate η, weight α.

Initialize starting point $\mathbf{z}_0 = (\mathbf{w}_0; \mathbf{q}_0) = \arg\min_{\mathbf{z}\in\mathcal{Z}} \psi(\mathbf{z})$.
For each $j \in [K]$, set $\mathbf{z}_j^{-1} = \mathbf{z}_0^0 = \mathbf{z}_0$.
for $s = 0$ **to** $S-1$ **do**
  for $k = 0$ **to** $K-1$ **do**
    $\mathbf{z}_{k+1/2}^s = \arg\min_{\mathbf{z}\in\mathcal{Z}}\{\langle \eta\nabla F(\mathbf{z}^s), \mathbf{z}\rangle + \alpha B(\mathbf{z}, \bar{\mathbf{z}}^s) + (1-\alpha)B(\mathbf{z}, \mathbf{z}_k^s)\}$.
    Option I: Uniform Sampling.
      Sample according to (88) and compute stochastic gradient according to (89).
    Option II: Importance Sampling.
      Sample according to (92) and compute stochastic gradient according to (93).
    Compute stochastic gradient estimator $\mathbf{g}_k^s$ defined in (20).
    $\mathbf{z}_{k+1}^s = \arg\min_{\mathbf{z}\in\mathcal{Z}}\{\langle \eta\mathbf{g}_k^s, \mathbf{z}\rangle + \alpha B(\mathbf{z}, \bar{\mathbf{z}}^s) + (1-\alpha)B(\mathbf{z}, \mathbf{z}_k^s)\}$.
  end for
  Compute full gradient $\nabla F(\mathbf{z}^s)$ according to (16).
  Compute snapshot point: $\mathbf{z}^s = \frac{1}{K}\sum_{k=0}^{K-1}\mathbf{z}_k^s$.
  Compute mirror snapshot point: $\nabla\psi(\bar{\mathbf{z}}^s) = \frac{1}{K}\sum_{k=0}^{K-1}\nabla\psi(\mathbf{z}_k^s)$.
  $\mathbf{z}_0^{s+1} = \mathbf{z}_K^s$.
end for
Return $\mathbf{z}_S = \frac{1}{SK}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\mathbf{z}_{k+1/2}^s$.

Theorem C.4. *Under Assumptions 3.3, 3.4 and 3.6, choose $0 \le \alpha < 1$, $0 < \gamma < 1$, and $\eta = \frac{\gamma\sqrt{1-\alpha}}{L_{\mathbf{c}}}$, with $L_{\mathbf{c}} = L_{\mathbf{u}}$ for uniform sampling and $L_{\mathbf{c}} = L_{\mathbf{i}}$ for importance sampling, Algorithm 3 ensures that*

$$\mathbb{E}\left[\epsilon(\mathbf{z}_S)\right] \le \mathcal{O}\left(\frac{L_{\mathbf{c}}}{S\sqrt{K}}\right). \tag{95}$$

Corollary C.5. *Under conditions in Theorem C.4, by setting $K = \Theta(m\bar{n})$ and employing either uniform sampling or importance sampling, the computation complexity for Algorithm 3 to reach ε-accuracy of (3) is $\mathcal{O}\left(m\bar{n} + \frac{m\sqrt{m\bar{n}}\ln m}{\varepsilon}\right)$.*

Corollary C.5 tells us that when utilizing MPVR to solve empirical GDRO, the complexity is $\sqrt{m}$ worse than ALEG. Such discrepancy stems from the dependence on the Lipschitz constant $L_{\mathbf{c}}$. Both uniform sampling and importance sampling will produce a Lipschitz constant worse than us by a factor of m, as shown in both (90) and (94). One can easily discover that if the Lipschitz constant for MPVR were only $\sqrt{m}$ worse than us, then the total complexity would be the same. Unfortunately, this could not happen because the sampling pattern ignores the nested finite-sum structure of the original problem. Technically, this additional factor is inherently due to the property of the infinity norm (dual norm of ℓ_1 norm), which scales up the factors from m coordinates when added together. Mathematical details are explicitly explained in Lemma C.6 and Remark C.7.

C.4. Omitted Proofs

We present the omitted proofs for the lemmas and theorems in Appendix C.

Proof of Lemma C.1. First, we study the first-order moment information of the stochastic gradient in (89). Considering the stochastic gradient w.r.t. $\mathbf{w}$, it follows that

$$\mathbb{E}[\nabla_{\mathbf{w}}F(\mathbf{z};\xi_k^s)] = \mathbb{E}\left[\frac{m\bar{n}}{n_{l_i}}\mathbf{q}_{l_i}\nabla\ell(\mathbf{w};\xi_l)\right] = \sum_{l=1}^{m\bar{n}}\frac{1}{m\bar{n}}\cdot\frac{m\bar{n}}{n_{l_i}}\mathbf{q}_{l_i}\nabla\ell(\mathbf{w};\xi_l) = \nabla_{\mathbf{w}}F(\mathbf{z}). \tag{96}$$

Then we consider the stochastic gradient w.r.t. $\mathbf{q}$. Denote by $\mathbf{e}_i \in \mathbb{R}^m$ as the one-hot vector that has an i-th component 1 with the rest component being 0. It follows that

$$\mathbb{E}[\nabla_{\mathbf{q}} F(\mathbf{z};\xi_k^s)] = \mathbb{E}\left[\frac{m\bar{n}}{n_{l_i}}\ell(\mathbf{w};\xi_l)\cdot\mathbf{e}_{l_i}\right] = \sum_{l=1}^{m\bar{n}}\frac{1}{m\bar{n}}\cdot\frac{m\bar{n}}{n_{l_i}}\ell(\mathbf{w};\xi_l)\cdot\mathbf{e}_{l_i} = \nabla_{\mathbf{q}}F(\mathbf{z}). \tag{97}$$

Secondly, we analyze the continuity of the second-order moment of the stochastic gradient in (89). According to Lemma C.6, we specify the following upper bound for the stochastic gradient w.r.t. $\mathbf{w}$ with $C_i = \frac{m\bar{n}}{n_{l_i}}$:

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_{\mathbf{w}}F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{w}}F(\mathbf{z};\xi_k^s)\right\|_{w,*}^2\right] \le& 2L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2\sum_{l_i=1}^m\frac{m\bar{n}}{n_{l_i}}\mathbf{q}_{l_i}^+ + 2G^2\sum_{l_i=1}^m\frac{m\bar{n}}{n_{l_i}}\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)^2\\ \le& 2m\bar{n}L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2\sum_{l_i=1}^m\frac{\mathbf{q}_{l_i}^+}{n_{\min}} + 2m\bar{n}G^2\sum_{l_i=1}^m\frac{\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)^2}{n_{\min}}\\ \le& 2m\frac{\bar{n}}{n_{\min}}L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 + 2m\frac{\bar{n}}{n_{\min}}G^2\left\|\mathbf{q}^+ - \mathbf{q}\right\|_1^2.\end{aligned} \tag{98}$$

Correspondingly, we consider the stochastic gradient w.r.t. $\mathbf{q}$ based on Lemma C.6,

$$\mathbb{E}\left[\left\|\nabla_{\mathbf{q}}F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{q}}F(\mathbf{z};\xi_k^s)\right\|_\infty^2\right] \le \sum_{l_i=1}^m\frac{m\bar{n}}{n_{l_i}}\cdot G^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 = m^2\frac{\bar{n}}{\bar{n}_h}G^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2. \tag{99}$$

We finish the proof by following the similar merging process in (35) (cf. the proof of Lemma 4.3). □

The following derivation process is analogous to the proof of Lemma C.1.

Proof of Lemma C.2. First, we study the first-order moment information of the stochastic gradient in (93). Considering the stochastic gradient w.r.t. $\mathbf{w}$, it follows that

$$\mathbb{E}[\nabla_{\mathbf{w}}F(\mathbf{z};\xi_k^s)] = \mathbb{E}\left[\mathbb{E}\left[m\mathbf{q}_{l_i}\nabla\ell(\mathbf{w};\xi_l)\middle|l_i\right]\right] = \mathbb{E}\left[m\mathbf{q}_{l_i}\nabla R_{l_i}(\mathbf{w})\right] = \sum_{l_i=1}^m\frac{1}{m}\cdot m\mathbf{q}_{l_i}\nabla R_{l_i}(\mathbf{w}) = \nabla_{\mathbf{w}}F(\mathbf{z}). \tag{100}$$

Then we consider the stochastic gradient w.r.t. $\mathbf{q}$.

$$\mathbb{E}[\nabla_{\mathbf{q}}F(\mathbf{z};\xi_k^s)] = \mathbb{E}\left[\mathbb{E}\left[m\ell(\mathbf{w};\xi_l)\cdot\mathbf{e}_{l_i}\middle|l_i\right]\right] = \mathbb{E}\left[mR_{l_i}(\mathbf{w})\cdot\mathbf{e}_{l_i}\right] = \nabla_{\mathbf{q}}F(\mathbf{z}). \tag{101}$$

Secondly, we analyze the continuity according to Lemma C.6. With $C_i = m$, it holds that:

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_{\mathbf{w}}F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{w}}F(\mathbf{z};\xi_k^s)\right\|_{w,*}^2\right] \le& 2L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 + 2G^2\sum_{l_i=1}^m m\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)^2\\ \le& 2mL^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2\sum_{l_i=1}^m\frac{\mathbf{q}_{l_i}^+}{n_{\min}} + 2mG^2\left\|\mathbf{q}^+ - \mathbf{q}\right\|_1^2.\end{aligned} \tag{102}$$

$$\mathbb{E}\left[\left\|\nabla_{\mathbf{q}}F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{q}}F(\mathbf{z};\xi_k^s)\right\|_\infty^2\right] \le \sum_{l_i=1}^m m\cdot G^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 = m^2G^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2. \tag{103}$$

The proof is finished by a simple merging process as in (35). □

Proof of Theorem C.4. The theoretical guarantee is a direct application of Alacaoglu & Malitsky (2022, Theorem 8). We only need to substitute the Lipschitz constant by $L_{\mathbf{c}}$. □

Proof of Corollary C.5. First, we investigate the order of $L_{\mathbf{c}}$ according to the sampling strategy.

For uniform sampling, we have a *data-dependent* $L_{\mathbf{u}}$ in (90). The value of $\frac{\bar{n}}{n_{\min}}$ ranges from 1 to $\bar{n}$, indicating that the order could be worse as $\Theta(m)$. Due to the fact that $\frac{\bar{n}}{\bar{n}_h} \geq 1$, we can derive that $L_{\mathbf{u}} = \mathcal{O}\left(m\sqrt{\ln m}\right)$. For importance sampling, we can directly assert that $L_{\mathbf{i}} = \mathcal{O}\left(m\sqrt{\ln m}\right)$ in view of (94). Therefore, we conclude that $L_{\mathbf{c}} = \mathcal{O}\left(m\sqrt{\ln m}\right)$.

Before calculating the computation complexity, we need to verify the tightness of the Lipschitzness. A linear loss item for $\ell(\cdot;\xi)$ realizes the lower bound. From Lemma C.6 and Remark C.7 we can further confirm that the order of $L_{\mathbf{c}}$ could not be improved.

The inner loop and outer loop of Algorithm 3 require $\Theta(KS)$ and $\Theta(m\bar{n}S)$ computations in total, respectively. To set the two costs in the same order, we choose $K = \Theta(m\bar{n})$. Consequently, we have $S = \mathcal{O}\left(\frac{L_{\mathbf{c}}}{\varepsilon\sqrt{m\bar{n}}}\right)$. Therefore, we derive a complexity of $\mathcal{O}\left(\frac{L_{\mathbf{c}}\sqrt{m\bar{n}}}{\varepsilon}\right)$. Because the initial cost for the full gradient is $m\bar{n}$ and $L_{\mathbf{c}} = \mathcal{O}\left(m\sqrt{\ln m}\right)$, we get a total complexity of $\mathcal{O}\left(m\bar{n} + \frac{m\sqrt{m\bar{n}\ln m}}{\varepsilon}\right)$. □

Finally, we present a technical lemma used in the previous analysis, in which we can see why ALEG goes beyond the pure extension of MPVR.

Lemma C.6. *Let $C_i = \frac{m\bar{n}}{n_{l_i}}$ for uniform sampling and $C_i = m$ for importance sampling, then for any $\mathbf{z}^+, \mathbf{z} \in \mathcal{Z}$, we have the following estimate*

$$\mathbb{E}\left[\left\|\nabla_{\mathbf{w}} F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{w}} F(\mathbf{z};\xi_k^s)\right\|_{w,*}^2\right] \leq 2L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 \sum_{l_i=1}^m C_i \mathbf{q}_{l_i}^+ + 2G^2 \sum_{l_i=1}^m C_i\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)^2 \tag{104}$$

for the stochastic gradient w.r.t. $\mathbf{w}$*, and*

$$\mathbb{E}\left[\left\|\nabla_{\mathbf{q}} F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{q}} F(\mathbf{z};\xi_k^s)\right\|_\infty^2\right] \leq \sum_{l_i=1}^m C_i \cdot G^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 \tag{105}$$

for the stochastic gradient w.r.t. $\mathbf{q}$.

Proof. Using the tower rule, it's straightforward to verify that

$$\mathbb{E}\left[C_i^2\right] = \sum_{l_i=1}^m \frac{C_i^2}{C_i} \sum_{j=1}^{n_{l_i}} \frac{1}{n_{l_i}} = \sum_{l_i=1}^m C_i \tag{106}$$

holds for both sampling techniques. For the stochastic gradient w.r.t. $\mathbf{w}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_{\mathbf{w}} F(\mathbf{z}^+;\xi_k^s) - \nabla_{\mathbf{w}} F(\mathbf{z};\xi_k^s)\right\|_{w,*}^2\right]\\
=&\mathbb{E}\left[C_i^2\left\|\mathbf{q}_{l_i}^+[\nabla\ell(\mathbf{w}^+;\xi_l) - \nabla\ell(\mathbf{w};\xi_l)] + \left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)\nabla\ell(\mathbf{w};\xi_l)\right\|_{w,*}^2\right]\\
\leq&2\mathbb{E}\left[C_i^2\left\|\mathbf{q}_{l_i}^+\left[\nabla\ell(\mathbf{w}^+;\xi_l) - \nabla\ell(\mathbf{w};\xi_l)\right]\right\|_{w,*}^2\right] + 2\mathbb{E}\left[C_i^2\left\|\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)\nabla\ell(\mathbf{w};\xi_l)\right\|_{w,*}^2\right]\\
\leq&2\sum_{l_i=1}^m C_i \sum_{j=1}^{n_{l_i}} \frac{1}{n_{l_i}}\left\|\mathbf{q}_{l_i}^+\left[\nabla\ell(\mathbf{w}^+;\xi_l) - \nabla\ell(\mathbf{w};\xi_l)\right]\right\|_{w,*}^2 + 2\mathbb{E}\left[C_i^2\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)^2\left\|\nabla\ell(\mathbf{w};\xi_l)\right\|_{w,*}^2\right]\\
\leq&2\sum_{l_i=1}^m \sum_{j=1}^{n_{l_i}} \frac{C_i}{n_{l_i}}\left\|\left[\nabla\ell(\mathbf{w}^+;\xi_l) - \nabla\ell(\mathbf{w};\xi_l)\right]\right\|_{w,*}^2 + 2\sum_{l_i=1}^m C_i\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)^2 \sum_{j=1}^{n_{l_i}} \frac{1}{n_{l_i}}\left\|\nabla\ell(\mathbf{w};\xi_l)\right\|_{w,*}^2\\
\leq&2\sum_{l_i=1}^m \sum_{j=1}^{n_{l_i}} \frac{C_i\mathbf{q}_{l_i}^+}{n_{l_i}} L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 + 2\sum_{l_i=1}^m C_i\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)^2 \sum_{j=1}^{n_{l_i}} \frac{1}{n_{l_i}} G^2\\
\leq&2L^2\left\|\mathbf{w}^+ - \mathbf{w}\right\|_w^2 \sum_{l_i=1}^m C_i\mathbf{q}_{l_i}^+ + 2G^2 \sum_{l_i=1}^m C_i\left(\mathbf{q}_{l_i}^+ - \mathbf{q}_{l_i}\right)^2.
\end{aligned} \tag{107}$$

As for the stochastic gradient w.r.t. $\mathbf{q}$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_{\mathbf{q}}F(\mathbf{z}^+;\xi_k^s)-\nabla_{\mathbf{q}}F(\mathbf{z};\xi_k^s)\right\|_\infty^2\right] \leq & \mathbb{E}\left[\left\|C_i\left[\ell(\mathbf{w}^+;\xi_l)-\ell(\mathbf{w};\xi_l)\right]\cdot\mathbf{e}_{l_i}\right\|_\infty^2\right] \\ \leq & \mathbb{E}\left[C_i^2\left[\ell(\mathbf{w}^+;\xi_l)-\ell(\mathbf{w};\xi_l)\right]^2\right] \\ \leq & \mathbb{E}\left[C_i^2\right]\cdot G^2\left\|\mathbf{w}^+-\mathbf{w}\right\|_w^2.\end{aligned} \tag{108}$$

Plug the result for $\mathbb{E}\left[C_i^2\right]$ into (108) yields the desired estimate. ☐

Remark C.7. The subtle issue is closely related to the property of the infinity norm as shown in the second inequality of (108), where the aggregation of the infinity norm for m one-hot vectors scales up by m times. However, our *group sampling*-based stochastic gradient construction can solve this issue by switching the order of the summation and infinity norm (check (33) to see our estimate), and thereby produce a better stochastic gradient with lower Lipschitz constant.