Efficient Stochastic Approximation of Minimax Excess Risk Optimization

Lijun Zhang [1 2] **Haomin Bai** [1] **Wei-Wei Tu** [3] **Ping Yang** [4] **Yao Hu** [4]

Abstract

While traditional distributionally robust optimization (DRO) aims to minimize the maximal risk over a set of distributions, Agarwal & Zhang (2022) recently proposed a variant that replaces risk with *excess risk*. Compared to DRO, the new formulation—minimax excess risk optimization (MERO) has the advantage of suppressing the effect of heterogeneous noise in different distributions. However, the choice of excess risk leads to a very challenging minimax optimization problem, and currently there exists only an inefficient algorithm for empirical MERO. In this paper, we develop efficient stochastic approximation approaches which directly target MERO. Specifically, we leverage techniques from stochastic convex optimization to estimate the minimal risk of every distribution, and solve MERO as a stochastic convex-concave optimization (SCCO) problem with biased gradients. The presence of bias makes existing theoretical guarantees of SCCO inapplicable, and fortunately, we demonstrate that the bias, caused by the estimation error of the minimal risk, is under-control. Thus, MERO can still be optimized with a nearly optimal convergence rate. Moreover, we investigate a practical scenario where the quantity of samples drawn from each distribution may differ, and propose a stochastic approach that delivers *distribution-dependent* convergence rates.

1. Introduction

With the widespread application of machine learning, it is common to encounter situations where the test distribution differs from the training distribution (Sugiyama et al., 2007; Zhang et al., 2013a; Gama et al., 2014). When faced with distribution shifts, learning models trained by conventional techniques (e.g., empirical risk minimization) tend to suffer significant performance degradation (Koh et al., 2021). Distributionally robust optimization (DRO) seeks a model with minimax risk over a set of potential distributions, offering a principled approach for generalization across distributions (Duchi & Namkoong, 2021). In general, DRO can be formulated as the following minimax optimization problem

$$\min_{\mathbf{w}\in\mathcal{W}} \sup_{\mathcal{P}\in\mathcal{S}} \left\{ \mathrm{E}_{\mathbf{z}\sim\mathcal{P}}\big[\ell(\mathbf{w};\mathbf{z})\big] \right\} \tag{1}$$

where $\mathcal{S}$ is a set of distributions, $\mathbf{z} \in \mathcal{Z}$ denotes a random sample drawn from $\mathcal{P}$, $\mathcal{W}$ is a hypothesis class, and $\ell(\cdot;\cdot)$ denotes a loss function that measures the performance. When $\mathcal{S}$ contains a finite number of distributions, (1) is referred to as Group DRO (GDRO) (Sagawa et al., 2020). One motivating example involves deploying a common classifier across multiple hospitals, each serving populations with varied demographics, to predict disease occurrences (Blum et al., 2017).

Although the minimax formulation enhances the model's robustness to distribution shifts, it carries the potential of being sensitive to heterogeneous noise, especially in GDRO where the candidate distributions might exhibit substantial differences. The rationale is evident: a single distribution with high levels of noise could easily dominate the maximum in (1), causing other distributions to be essentially overlooked. To overcome this limitation, Agarwal & Zhang (2022) recently proposed a variant of DRO, namely minimax regret optimization (MRO), that replaces the raw risk $\mathrm{E}_{\mathbf{z}\sim\mathcal{P}}[\ell(\mathbf{w};\mathbf{z})]$ in (1) with *excess risk* (or regret). In particular, they consider the setting of GDRO, and seek to minimize the worst-case excess risk over m distributions $\mathcal{P}_1,\ldots,\mathcal{P}_m$:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ R_i(\mathbf{w}) - R_i^* \right\}. \tag{2}$$

where $R_i(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\big[\ell(\mathbf{w};\mathbf{z})\big]$ and $R_i^* = \min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\big[\ell(\mathbf{w};\mathbf{z})\big]$. To avoid confusion with the term "regret" commonly used in online learning (Cesa-Bianchi & Lugosi, 2006), in this paper, we refer to MRO as minimax excess risk optimization (MERO).

MERO can be understood as subtracting the intrinsic difficulty of each distribution (i.e., the minimal risk R_i^*) from

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [2]Pazhou Laboratory (Huangpu), Guangzhou, China [3]Artificial Productivity Inc., Beijing, China [4]Xiaohongshu Inc., Beijing, China. Correspondence to: Lijun Zhang <zhanglj@lamda.nju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

the risk, making the remaining quantities more comparable. However, since the value of R_i^* is typically unknown, the optimization problem in (2) poses a significant challenge. In the previous work, Agarwal & Zhang (2022) choose to optimize the empirical counterpart of MERO, but the resulting algorithm requires solving an empirical risk minimization problem in each iteration, which is computationally prohibitive when the number of samples is large. Along this line of research, this paper aims to develop efficient stochastic approximation approaches that optimize (2) *directly*. It is worth mentioning that by doing so, we also eliminate the need to analyze the discrepancy between the empirical MERO and the population MERO. To exploit techniques from stochastic approximation (Nemirovski et al., 2009), we cast (2) as a stochastic convex-concave saddle-point problem:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{ \phi(\mathbf{w},\mathbf{q}) = \sum_{i=1}^m q_i \left[R_i(\mathbf{w}) - R_i^* \right] \right\} \tag{3}$$

where $\Delta_m = \{\mathbf{q} \in \mathbb{R}^m : \mathbf{q} \geq 0, \sum_{i=1}^m q_i = 1\}$ is the $(m-1)$-dimensional simplex.

If the value of R_i^* is known, we can readily employ stochastic mirror descent (SMD) to optimize (3), mirroring the application of SMD to GDRO (Zhang et al., 2023, Algorithm 1). An intuitive approach thus emerges, which first estimates the value of R_i^*, denoted by $\widehat{R}_i$, replaces the function $\phi(\mathbf{w},\mathbf{q})$ in (3) with $\sum_{i=1}^m q_i[R_i(\mathbf{w}) - \widehat{R}_i]$, and subsequently applies SMD to the revised problem. As elaborated later, this multi-stage approach is indeed feasible, but it needs to fix the total number of iterations (or the total number of random samples), and cannot return a solution upon demand. To possess the *anytime* ability, i.e., the capability to deliver a solution at any round, we propose a stochastic approximation approach that alternates between estimating R_i^* and optimizing (3) in an iterative manner. Specifically, we execute m instances of SMD, each designed to minimize the risk of an individual distribution. In this way, at each round t, we have a model $\bar{\mathbf{w}}_t^i$ that attains a low risk for each distribution $\mathcal{P}_i$. Leveraging these $\bar{\mathbf{w}}_t^i$s, we are able to construct biased stochastic gradients for (3), and proceed to optimize the problem by SMD. Our theoretical analysis demonstrates that the bias is under-control, and the optimization error reduces at an $\widetilde{O}(\sqrt{(\log m)/t})$ rate,[1] which is optimal up to a polylogarithmic factor.

Furthermore, we investigate MERO under the imbalanced setting where the number of samples drawn from each distribution is different. To this end, we develop a stochastic approximation approach for a weighted formulation of MERO, which yields a convergence rate for each distribution depending on the budget of that particular distribution. Under appropriate conditions, we demonstrate that for smooth risk functions, the excess risk of the i-th distribution decreases at an $O((\log m)/\sqrt{n_i})$ rate, where n_i is the sample budget. Finally, we conduct experiments to validate the efficiency and effectiveness of our algorithms.

[1]We use the $\widetilde{O}$ notation to hide constant factors as well as polylogarithmic factors in t.

2. Related Work

There is an extensive body of literature on DRO, with various problem formulations and research focuses (Rahimian & Mehrotra, 2022). Below, we offer a concise overview of the literature related to GDRO and MERO. Further works are detailed in Appendix A.

When the number of distributions in $\mathcal{S}$ is finite, (1) becomes GDRO (Sagawa et al., 2020), which can be formulated as the following stochastic convex-concave problem:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{ \sum_{i=1}^m q_i R_i(\mathbf{w}) \right\}. \tag{4}$$

By drawing 1 random sample in each round, Sagawa et al. (2020) have applied SMD to (4), and established an $O(m\sqrt{(\log m)/T})$ convergence rate. As a result, their algorithm achieves an $O(m^2(\log m)/\epsilon^2)$ sample complexity for finding an ϵ-optimal solution, which is suboptimal according to the $\Omega(m/\epsilon^2)$ lower bound (Soma et al., 2022, Theorem 5). In the literature, there exist 3 different ways to obtain a (nearly) optimal $O(m(\log m)/\epsilon^2)$ sample complexity.

1. In each round, we draw m random samples, one from each distribution, and then update $\mathbf{w}$ and $\mathbf{q}$ in (4) by SMD. In this way, the optimization error reduces at an $O(\sqrt{(\log m)/T})$ rate (Nemirovski et al., 2009, §3.2), implying an $O(m(\log m)/\epsilon^2)$ sample complexity.
2. In each round, we only draw 1 sample from one specific distribution, update $\mathbf{w}$ by SMD, and update $\mathbf{q}$ by an online algorithm for non-oblivious multi-armed bandits (Lattimore & Szepesvári, 2020). It can be shown that the optimization error reduces at an $O(\sqrt{m(\log m)/T})$ rate (Soma et al., 2022; Zhang et al., 2023), keeping the same sample complexity.
3. The third approach is similar to the second one, but it updates $\mathbf{q}$ by integrating SMD with gradient clipping (Carmon & Hausler, 2022, §A.1).

Recently, Zhang et al. (2023) have considered a practical scenario where the number of samples that can be drawn from each distribution is different. To this end, they introduce a weighted formulation of GDRO:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{ \sum_{i=1}^m q_i p_i R_i(\mathbf{w}) \right\} \tag{5}$$

where $p_i > 0$ is a weight assigned to distribution $\mathcal{P}_i$. Then, they develop two stochastic algorithms to solve (5), and

establish distribution-wise convergence rates. In this way, the rate is not dominated by the distribution with the smallest budget. Inspired by Zhang et al. (2023), we will investigate MERO with varying sample sizes across distributions in Section 4. The technique of introducing scale factors has been previously presented in the study of MERO under heterogeneous distributions (Agarwal & Zhang, 2022, §5), where all distributions have the same number of samples, but with different complexities.

To optimize MERO, Agarwal & Zhang (2022) consider the empirical version:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ \frac{1}{n}\sum_{j=1}^{n} \ell(\mathbf{w};\mathbf{z}^{(i,j)}) - \min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{j=1}^{n} \ell(\mathbf{w};\mathbf{z}^{(i,j)}) \right\}$$

where $\{\mathbf{z}^{(i,j)} : j = 1,\ldots,n\}$ are random samples drawn from distribution $\mathcal{P}_i$, and analyze the generalization performance using classical tools from learning theory. They have developed an iterative method for solving the above problem, which needs to address an empirical risk minimization problem in each iteration, rendering the process inefficient. It's worth noting that the principle of minimizing the worst-case excess risk has surfaced in various other fields (Eldar et al., 2004; Alaiz-Rodríguez et al., 2007; Jiang et al., 2013; Abusorrah et al., 2019).

3. Stochastic Approximation of MERO

We first describe preliminaries of stochastic approximation, including the setup and assumptions, then develop a multi-stage approach, and finally propose an anytime approach. Due to space limitations, we defer all proofs to appendices.

3.1. Preliminaries

We first present the standard setup of mirror descent (Nemirovski et al., 2009). We endow the domain $\mathcal{W}$ with a distance-generating function $\nu_w(\cdot)$, which is 1-strongly convex w.r.t. a specific norm $\|\cdot\|_w$. We define the Bregman distance corresponding to $\nu_w(\cdot)$ as

$$B_w(\mathbf{u},\mathbf{v}) = \nu_w(\mathbf{u}) - \left[\nu_w(\mathbf{v}) + \langle \nabla\nu_w(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle\right].$$

For the simplex Δ_m, we select the entropy function $\nu_q(\mathbf{q}) = \sum_{i=1}^m q_i \ln q_i$, which demonstrates 1-strong convexity w.r.t. the vector ℓ_1-norm $\|\cdot\|_1$, as the distance-generating function. In a similar manner, $B_q(\cdot,\cdot)$ represents the Bregman distance associated with $\nu_q(\cdot)$, which is the Kullback–Leibler divergence between distributions.

Next, we introduce standard assumptions.

Assumption 3.1. All the risk functions $R_1(\cdot),\cdots,R_m(\cdot)$ and the domain $\mathcal{W}$ are convex.

Assumption 3.2. The domain $\mathcal{W}$ is bounded in the sense that

$$\max_{\mathbf{w}\in\mathcal{W}} B_w(\mathbf{w},\mathbf{o}_w) \le D^2. \tag{6}$$

where $\mathbf{o}_w = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w})$.

For the simplex Δ_m, we have $\max_{\mathbf{q}\in\Delta_m} B_q(\mathbf{q},\mathbf{o}_q) \le \ln m$ where $\mathbf{o}_q = \frac{1}{m}\mathbf{1}_m \in \mathbb{R}^m$ and $\mathbf{1}_m$ is the m-dimensional vector of all ones (Beck & Teboulle, 2003, Proposition 5.1).

We assume that the gradient is bounded, and the loss belongs to $[0,1]$.

Assumption 3.3. For all $i\in[m]$, we have

$$\|\nabla\ell(\mathbf{w};\mathbf{z})\|_{w,*} \le G,\ \forall \mathbf{w}\in\mathcal{W},\ \mathbf{z}\sim\mathcal{P}_i \tag{7}$$

where $\|\cdot\|_{w,*}$ is the dual norm of $\|\cdot\|_w$.

Assumption 3.4. For all $i\in[m]$, we have

$$0 \le \ell(\mathbf{w};\mathbf{z}) \le 1,\ \forall \mathbf{w}\in\mathcal{W},\ \mathbf{z}\sim\mathcal{P}_i. \tag{8}$$

Given an solution $(\bar{\mathbf{w}},\bar{\mathbf{q}})$ to (3), the optimization error is defined as

$$\epsilon_\phi(\bar{\mathbf{w}},\bar{\mathbf{q}}) = \max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}},\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \phi(\mathbf{w},\bar{\mathbf{q}}). \tag{9}$$

3.2. A Multi-Stage Stochastic Approximation Approach for MERO

As mentioned in the introduction, we can design a multi-stage stochastic approach for MERO. Notably, analogous methodologies have found their application in the empirical research conducted on language modeling (Oren et al., 2019; Xie et al., 2023).

Stage 1: Minimizing the risk For each distribution $\mathcal{P}_i$, we run an instance of SMD to minimize the risk $R_i(\cdot)$, and obtain an approximate solution $\bar{\mathbf{w}}^{(i)}$. We execute each SMD for T iterations, and thus consume mT samples. From the theoretical guarantee of SMD (Nemirovski et al., 2009), with probability at least $1-\delta$, we have $R_i(\bar{\mathbf{w}}^{(i)}) - R_i^* = O(\sqrt{\log(1/\delta)/T})$, for each $i\in[m]$. By the union bound, with high probability, we have $\max_{i\in[m]}[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*] = O(\sqrt{(\log m)/T})$.

Stage 2: Estimating the minimal risk To estimate the value of $R_i(\bar{\mathbf{w}}^{(i)})$, we draw T samples $\mathbf{z}_1^{(i)},\ldots,\mathbf{z}_T^{(i)}$ from each distribution $\mathcal{P}_i$, and calculate $\widehat{R}_i(\bar{\mathbf{w}}^{(i)}) = \frac{1}{T}\sum_{j=1}^T \ell(\bar{\mathbf{w}}^{(i)};\mathbf{z}_j^{(i)})$. From standard concentration inequalities (Lugosi, 2009) and the union bound, with high probability, we have $\max_{i\in[m]} |\widehat{R}_i(\bar{\mathbf{w}}^{(i)}) - R_i(\bar{\mathbf{w}}^{(i)})| = O(\sqrt{(\log m)/T})$. In this step, we also use mT samples.

Stage 3: Applying SMD to an approximate problem From the above two steps, with high probability, we have

$$\max_{i\in[m]} |\widehat{R}_i(\bar{\mathbf{w}}^{(i)}) - R_i^*| = O(\sqrt{(\log m)/T}). \tag{10}$$

Then, we formulate the following problem

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \hat{\phi}(\mathbf{w},\mathbf{q}) = \sum_{i=1}^m q_i\big[R_i(\mathbf{w}) - \widehat{R}_i(\bar{\mathbf{w}}^{(i)})\big] \tag{11}$$

which serves as an approximation to (3). Since $\widehat{R}_i(\bar{\mathbf{w}}^{(i)})$ is a constant, we can directly apply SMD to (11). After T iterations, we obtain solutions $\bar{\mathbf{w}}$ and $\bar{\mathbf{q}}$ such that, with high probability $\max_{\mathbf{q}\in\Delta_m} \hat{\phi}(\bar{\mathbf{w}},\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \hat{\phi}(\mathbf{w},\bar{\mathbf{q}}) = O(\sqrt{(\log m)/T})$ (Zhang et al., 2023). From (10), it is easy to prove that $\max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}},\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \phi(\mathbf{w},\bar{\mathbf{q}}) = O(\sqrt{(\log m)/T})$. This step also requires mT samples.

In summary, the above approach reduces the optimization error to $O(\sqrt{(\log m)/T})$, at the cost of $3mT$ samples. In other words, it attains an $O(m(\log m)/\epsilon^2)$ sample complexity, which is nearly optimal according to the lower bound of GDRO (Soma et al., 2022). We would like to highlight that the 2nd stage is not essential, and is included to facilitate understanding. In fact, we can omit the 2nd stage, and define $\hat{\phi}(\mathbf{w},\mathbf{q}) = \sum_{i=1}^m q_i[R_i(\mathbf{w}) - R_i(\bar{\mathbf{w}}^{(i)})]$ in (11). The resulting optimization problem can still be solved by SMD, and the sample complexity remains in the same order. An illustrative example of this two-stage approach can be found in Section 4.

Remark 3.5. While the multi-stage approach achieves a nearly optimal sample complexity, it suffers two *limitations*: (i) the total number of iterations must be predetermined; (ii) a solution is available only when the algorithm enters the final stage. To circumvent these drawbacks, we put forth a stochastic approximation approach that interleaves the aforementioned three stages together, being able to return a solution at any time.

3.3. An Anytime Stochastic Approximation Approach for MERO

We maintain m instances of SMD to minimize all the risk functions $R_1(\cdot),\cdots,R_m(\cdot)$, and meanwhile utilize their solutions to optimize (3) according to SMD.

For the purpose of minimizing $R_i(\cdot)$, we denote by $\mathbf{w}_t^{(i)}$ the solution in the t-th iteration. We first draw 1 sample $\mathbf{z}_t^{(i)}$ from each distribution $\mathcal{P}_i$, and calculate the stochastic gradient $\nabla\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)})$ which is an unbiased estimator of $\nabla R_i(\mathbf{w}_t^{(i)})$. According to SMD (Nemirovski et al., 2009), we update $\mathbf{w}_t^{(i)}$ by

$$\begin{aligned}\mathbf{w}_{t+1}^{(i)} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \Big\{&\eta_t^{(i)}\langle\nabla\ell(\mathbf{w}_t^{(i)};\mathbf{z}_t^{(i)}),\mathbf{w}-\mathbf{w}_t^{(i)}\rangle \\ &+B_w(\mathbf{w},\mathbf{w}_t^{(i)})\Big\},\ \forall i\in[m]\end{aligned} \tag{12}$$

where $\eta_t^{(i)} > 0$ is the step size. Due to technical reasons, we will use the weighted average of iterates

$$\bar{\mathbf{w}}_t^{(i)} = \sum_{j=1}^t \frac{\eta_j^{(i)}\mathbf{w}_j^{(i)}}{\sum_{k=1}^t \eta_k^{(i)}} = \frac{(\sum_{j=1}^{t-1}\eta_j^{(i)})\bar{\mathbf{w}}_{t-1}^{(i)} + \eta_t^{(i)}\mathbf{w}_t^{(i)}}{\sum_{k=1}^t \eta_k^{(i)}} \tag{13}$$

as an approximate solution to $\min_{\mathbf{w}\in\mathcal{W}} R_i(\mathbf{w})$. While selecting the last iterate $\mathbf{w}_t^{(i)}$ is also an option (Shamir & Zhang, 2013; Harvey et al., 2019; Jain et al., 2019), this choice leads to a more complex analysis. Therefore, we prefer to employ $\bar{\mathbf{w}}_t^{(i)}$.

We proceed to minimize (3) by SMD. Let $\mathbf{w}_t$ and $\mathbf{q}_t$ be the solutions in the t-th round. Based on the random samples $\mathbf{z}_t^{(1)},\ldots,\mathbf{z}_t^{(m)}$, we define the stochastic gradient of $\phi(\cdot,\cdot)$ at $(\mathbf{w}_t,\mathbf{q}_t)$ w.r.t. $\mathbf{w}$ as

$$\mathbf{g}_w(\mathbf{w}_t,\mathbf{q}_t) = \sum_{i=1}^m q_{t,i}\nabla\ell(\mathbf{w}_t;\mathbf{z}_t^{(i)}), \tag{14}$$

which is an unbiased estimator of the true gradient $\nabla_{\mathbf{w}}\phi(\mathbf{w}_t,\mathbf{q}_t) = \sum_{i=1}^m q_{t,i}\nabla R_i(\mathbf{w}_t)$. The challenge lies in the construction of the stochastic gradient w.r.t. $\mathbf{q}$. To this end, we define

$$\begin{aligned}\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t) = \Big[&\ell(\mathbf{w}_t;\mathbf{z}_t^{(1)}) - \ell(\bar{\mathbf{w}}_t^{(1)};\mathbf{z}_t^{(1)}),\ldots, \\ &\ell(\mathbf{w}_t;\mathbf{z}_t^{(m)}) - \ell(\bar{\mathbf{w}}_t^{(m)};\mathbf{z}_t^{(m)})\Big]^\top\end{aligned} \tag{15}$$

which is a *biased* estimator of the true gradient $\nabla_{\mathbf{q}}\phi(\mathbf{w}_t,\mathbf{q}_t) = [R_1(\mathbf{w}_t) - R_1^*,\ldots,R_m(\mathbf{w}_t) - R_m^*]^\top$, since

$$\begin{aligned}&\mathrm{E}_{t-1}\left[\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t)\right] \\ =&\Big[R_1(\mathbf{w}_t) - R_1(\bar{\mathbf{w}}_t^{(1)}),\ldots,R_m(\mathbf{w}_t) - R_m(\bar{\mathbf{w}}_t^{(m)})\Big]^\top \\ \neq&\nabla_{\mathbf{q}}\phi(\mathbf{w}_t,\mathbf{q}_t)\end{aligned}$$

where $\mathrm{E}_{t-1}[\cdot]$ represents the expectation conditioned on the randomness until round $t-1$. Thanks to the SMD update in (12), we know that $R_i(\bar{\mathbf{w}}_t^{(i)})$ is close to R_i^*, for all $i\in[m]$. As a result, the bias in $\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t)$, determined by $R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i^*$, is effectively managed, making it possible to maintain a (nearly) optimal convergence rate.

Equipped with the stochastic gradients in (14) and (15), we

Algorithm 1 An Anytime Stochastic Approximation Approach for MERO

Initialize $\mathbf{w}_1 = \mathbf{w}_1^{(1)} = \cdots = \mathbf{w}_1^{(m)} = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w})$, and $\mathbf{q}_1 = \frac{1}{m}\mathbf{1}_m \in \mathbb{R}^m$
for $t = 1$ to T **do**
  For each $i \in [m]$, draw a sample $\mathbf{z}_t^{(i)}$ from distribution $\mathcal{P}_i$
  For each $i \in [m]$, calculate $\nabla\ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})$ and update $\mathbf{w}_t^{(i)}$ according to (12)
  For each $i \in [m]$, calculate the weighted average $\bar{\mathbf{w}}_t^{(i)}$ in (13)
  Construct the stochastic gradients in (14) and (15)
  Update $\mathbf{w}_t$ and $\mathbf{q}_t$ according to (16) and (17), respectively
  Calculate the the weighted averages $\bar{\mathbf{w}}_t$ and $\bar{\mathbf{q}}_t$ in (18)
end for

update $\mathbf{w}_t$ and $\mathbf{q}_t$ by SMD:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \big\{\eta_t^w \langle \mathbf{g}_w(\mathbf{w}_t, \mathbf{q}_t), \mathbf{w} - \mathbf{w}_t\rangle + B_w(\mathbf{w}, \mathbf{w}_t)\big\}, \tag{16}$$

$$\mathbf{q}_{t+1} = \operatorname*{argmin}_{\mathbf{q}\in\Delta_m} \big\{\eta_t^q \langle -\mathbf{g}_q(\mathbf{w}_t, \mathbf{q}_t), \mathbf{q} - \mathbf{q}_t\rangle + B_q(\mathbf{q}, \mathbf{q}_t)\big\} \tag{17}$$

where $\eta_t^w > 0$ and $\eta_t^q > 0$ are step sizes. We will maintain the weighted averages of iterates:

$$\bar{\mathbf{w}}_t = \sum_{j=1}^{t} \frac{\eta_j^w \mathbf{w}_j}{\sum_{k=1}^{t} \eta_k^w}, \text{ and } \bar{\mathbf{q}}_t = \sum_{j=1}^{t} \frac{\eta_j^q \mathbf{q}_j}{\sum_{k=1}^{t} \eta_k^q} \tag{18}$$

which can be returned as solutions if necessary. The completed procedure is given in Algorithm 1.

Next, we discuss the theoretical guarantee of Algorithm 1. To this end, we first present the optimization error of $\bar{\mathbf{w}}_t^{(i)}$ in (13) for each risk function (Nemirovski et al., 2009, §2.3).

Theorem 3.6. *Under Assumptions 3.1, 3.2, and 3.3, by setting $\eta_t^{(i)} = \frac{D}{G\sqrt{t}}$ in Algorithm 1, with probability at least $1-\delta$, we have*

$$R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i^* \le \frac{DG\big[3 + \ln t + 16\sqrt{(1+\ln t)\ln(2mt^2/\delta)}\big]}{4(\sqrt{t+1}-1)} \tag{19}$$

for all $i \in [m]$, and $t \in \mathbb{Z}_+$.

Finally, we examine the optimization error of $\bar{\mathbf{w}}_t$ and $\bar{\mathbf{q}}_t$ for (3). Because of the biased stochastic gradient in (15), we cannot apply existing guarantees of SMD for stochastic convex-concave optimization (Nemirovski et al., 2009, §3.1). Therefore, we provide a novel analysis of SMD with biased gradients, and utilize Theorem 3.6 to demonstrate that the bias does not significantly affect the convergence behavior.

Theorem 3.7. *Under Assumptions 3.1, 3.2, 3.3, and 3.4, and setting appropriate parameters $\eta_t^{(i)}$, η_t^w, and η_t^q in Algorithm 1, with high probability, we have*

$$\epsilon_\phi(\bar{\mathbf{w}}_t, \bar{\mathbf{q}}_t) = O\left(\frac{\log^2 t + \log^{1/2} m \log^{3/2} t}{\sqrt{t}}\right)$$

for all $t \in \mathbb{Z}_+$.

Remark 3.8. The above theorem demonstrates that Algorithm 1 converges at a (nearly) optimal rate of $O([\log^2 t + \log^{1/2} m \log^{3/2} t]/\sqrt{t})$. Given a fixed number of iterations T, this rate is slightly slower than the $O(\sqrt{(\log m)/T})$ rate achieved by the multi-stage approach in Section 3.2. Nonetheless, a considerable advantage of Algorithm 1 is its anytime characteristic, indicating it is capable of returning a solution at any round.

Remark 3.9. Compared to stochastic algorithms for GDRO (Nemirovski et al., 2009; Zhang et al., 2023), our method requires the maintenance and updating of m additional models, resulting in a higher computational complexity. When m is very large and the computational cost becomes prohibitive, we may opt for a simplified version of MERO (Xie et al., 2023):

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \left\{\sum_{i=1}^{m} q_i \left[R_i(\mathbf{w}) - R_i(\mathbf{w}_r)\right]\right\} \tag{20}$$

where $\mathbf{w}_r$ is a reference model shared by all distributions. In (20), we use the risk of $\mathbf{w}_r$ on each distribution to replace the minimal risk of that distribution, and thus only need to estimate one additional model, namely $\mathbf{w}_r$. A possible way for selecting $\mathbf{w}_r$ is to choose the model that minimizes the average risk across all distributions. To efficiently solve (20), we could develop a multi-stage stochastic method, akin to the approach in Section 3.2, or an anytime stochastic method, similar to Algorithm 1.

4. Stochastic Approximation of Weighted MERO

In practice, the costs associated with gathering samples can differ among distributions, and thus we may allocate diverse budgets for different distributions.

4.1. Preliminaries

Let n_i represent the sample budget of distribution $\mathcal{P}_i$. For the sake of simplicity, we assume an order of $n_1 \ge n_2 \ge \cdots \ge n_m$. To satisfy the budget, we can simply run the multi-stage approach in Section 3.2 with $T = n_m/3$,

or Algorithm 1 for n_m iterations. However, the optimization error decreases only at an $O(\sqrt{(\log m)/n_m})$ or $\widetilde{O}(\sqrt{(\log m)/n_m})$ rate, which is determined by the smallest budget n_m. In other words, for a distribution $\mathcal{P}_i$ with $n_i > n_m$, the extra $n_i - n_m$ samples are wasted.

Analogous to the weighted GDRO in (5), we also formulate a weighted version of MERO:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \varphi(\mathbf{w},\mathbf{q}) = \sum_{i=1}^{m} q_i p_i \left[R_i(\mathbf{w}) - R_i^*\right] \tag{21}$$

where the value of the weight p_i will be determined later. As demonstrated by Agarwal & Zhang (2022), the heterogeneous scaling in (21) allows us to establish distribution-specific bounds for the excess risk. In our paper, the goal is to make the excess risk of distribution $\mathcal{P}_i$ reducing at an $O(\sqrt{(\log m)/n_i})$ rate. Again, the optimization problem in (21) is more challenging than the counterpart in (5), due to the existence of R_i^*s. We notice that because the budgets are fixed and known, there is no need to pursue the anytime ability. In the following, we will develop a two-stage stochastic approach for weighted MERO.

4.2. A Two-Stage Stochastic Approximation Approach for Weighted MERO

Our approach consists of two stages: minimizing each risk and minimizing an approximate problem. The complete procedure is shown in Algorithm 2, where Steps 1-8 belong to the 1st stage, and others correspond to the 2nd stage.

Stage 1: Minimizing the risk Similar to the first stage in Section 3.2, we will deploy an instance of SMD to minimize each individual risk $R_i(\cdot)$. The difference is that the number of iterations is set to be $n_i/2$ for distribution $\mathcal{P}_i$. Consequently, a larger budget yields a smaller error. Because the total number of iterations is fixed, we will use a fixed step size for each SMD. Specifically, the update rule for the i-th distribution at the t-th round is given by

$$\begin{aligned}\mathbf{w}_{t+1}^{(i)} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \Big\{ &\eta^{(i)} \langle \nabla \ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)}), \mathbf{w} - \mathbf{w}_t^{(i)} \rangle \\ &+ B_w(\mathbf{w}, \mathbf{w}_t^{(i)}) \Big\}\end{aligned} \tag{22}$$

where $\mathbf{w}_t^{(i)}$ is the current solution, $\mathbf{z}_t^{(i)}$ is a random sample drawn from $\mathcal{P}_i$, and $\eta^{(i)} > 0$ is the step size. After $n_i/2$ iterations, we will use the average of iterates $\bar{\mathbf{w}}^{(i)} = \frac{1}{n_i/2}\sum_{t=1}^{n_i/2} \mathbf{w}_t^{(i)}$ as an approximate solution to $\min_{\mathbf{w}\in\mathcal{W}} R_i(\mathbf{w})$.

Similar to Theorem 3.6, we have the following guarantee for the excess risk of each $\bar{\mathbf{w}}^{(i)}$.

Theorem 4.1. *Under Assumptions 3.1, 3.2, and 3.3, by setting $\eta^{(i)} = \frac{2D}{G\sqrt{n_i}}$ in Algorithm 2, with probability at least $1-\delta$, we have*

$$R_i(\bar{\mathbf{w}}^{(i)}) - R_i^* \leq \frac{2DG}{\sqrt{n_i}} \left(1 + 4\sqrt{2\ln\frac{m}{\delta}}\right), \tag{23}$$

for all $i \in [m]$.

Stage 2: Minimizing an approximate problem Based on the solutions from the 1st stage, we construct the problem

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \widehat{\varphi}(\mathbf{w},\mathbf{q}) = \sum_{i=1}^{m} q_i p_i \left[R_i(\mathbf{w}) - R_i(\bar{\mathbf{w}}^{(i)})\right] \tag{24}$$

to approximate (21). The following lemma shows that, the optimization error of any $(\bar{\mathbf{w}}, \bar{\mathbf{q}})$ for (21) is close to that for (24), provided $R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*$ is small, for all $i \in [m]$. As a result, we can focus on optimizing (24), which is more manageable, as the stochastic gradient of $\widehat{\varphi}(\mathbf{w},\mathbf{q})$ can be easily constructed.

Lemma 4.2. *For any $(\bar{\mathbf{w}}, \bar{\mathbf{q}})$, we have*

$$\epsilon_{\varphi}(\bar{\mathbf{w}}, \bar{\mathbf{q}}) \leq \epsilon_{\widehat{\varphi}}(\bar{\mathbf{w}}, \bar{\mathbf{q}}) + 2\max_{i\in[m]} \left\{ p_i \left[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*\right] \right\}.$$

From Theorem 4.1, we know that a larger n_i leads to a smaller $R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*$. Therefore, when the budget n_i is large, we can set a large p_i without influencing $\max_{i\in[m]} \left\{p_i[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*]\right\}$, which is crucial for achieving faster rates in distributions with larger budgets.

Inspired by Zhang et al. (2023, Algorithm 4), we solve (24) by stochastic mirror-prox algorithm (SMPA) (Juditsky et al., 2011). One notable merit of SMPA is that its optimization error depends on the variance of the gradient. For a distribution $\mathcal{P}_i$ with a larger n_i, we can leverage mini-batches (Roux et al., 2008; Zhang et al., 2013c) to estimate $R_i(\mathbf{w}) - R_i(\bar{\mathbf{w}}^{(i)})$ in (24) more accurately, (i.e., with a smaller variance), which again makes it possible to use a larger p_i.

Within the framework of SMPA, we keep two sets of solutions: $(\mathbf{w}_t, \mathbf{q}_t)$ and $(\mathbf{w}_t', \mathbf{q}_t')$. In the t-th iteration, we first draw n_i/n_m samples from every distribution $\mathcal{P}_i$, denoted by $\mathbf{z}_t^{(i,1)}, \ldots, \mathbf{z}_t^{(i,n_i/n_m)}$. Then, we use them to construct stochastic gradients of $\widehat{\varphi}(\mathbf{w},\mathbf{q})$ at $(\mathbf{w}_t', \mathbf{q}_t')$:

$$\begin{aligned}&\mathbf{g}_w(\mathbf{w}_t', \mathbf{q}_t') \\ &= \sum_{i=1}^{m} q_{t,i}' p_i \left(\frac{n_m}{n_i} \sum_{j=1}^{n_i/n_m} \nabla \ell(\mathbf{w}_t'; \mathbf{z}_t^{(i,j)}) \right), \\ &\mathbf{g}_q(\mathbf{w}_t', \mathbf{q}_t') \\ &= \Bigg[p_1 \frac{n_m}{n_1} \sum_{j=1}^{n_1/n_m} \left[\ell(\mathbf{w}_t'; \mathbf{z}_t^{(1,j)}) - \ell(\bar{\mathbf{w}}^{(1)}; \mathbf{z}_t^{(1,j)})\right], \\ &\qquad \ldots, p_m \left[\ell(\mathbf{w}_t'; \mathbf{z}_t^{(m)}) - \ell(\bar{\mathbf{w}}^{(m)}; \mathbf{z}_t^{(m)})\right] \Bigg]^\top.\end{aligned} \tag{25}$$

It is easy to verify that $\mathrm{E}_{t-1}[\mathbf{g}_w(\mathbf{w}'_t, \mathbf{q}'_t)] = \nabla_{\mathbf{w}}\widehat{\varphi}(\mathbf{w}'_t, \mathbf{q}'_t)$ and $\mathrm{E}_{t-1}[\mathbf{g}_q(\mathbf{w}'_t, \mathbf{q}'_t)] = \nabla_{\mathbf{q}}\widehat{\varphi}(\mathbf{w}'_t, \mathbf{q}'_t)$. Based on (25), we use SMD to update $(\mathbf{w}'_t, \mathbf{q}'_t)$, and obtain $(\mathbf{w}_{t+1}, \mathbf{q}_{t+1})$:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}}\big\{\eta_w\langle \mathbf{g}_w(\mathbf{w}'_t, \mathbf{q}'_t), \mathbf{w} - \mathbf{w}'_t\rangle + B_w(\mathbf{w}, \mathbf{w}'_t)\big\}, \tag{26}$$

$$\mathbf{q}_{t+1} = \operatorname*{argmin}_{\mathbf{q}\in\Delta_m}\big\{\eta_q\langle -\mathbf{g}_q(\mathbf{w}'_t, \mathbf{q}'_t), \mathbf{q} - \mathbf{q}'_t\rangle + B_q(\mathbf{q}, \mathbf{q}'_t)\big\} \tag{27}$$

where $\eta_w > 0$ and $\eta_q > 0$ are step sizes. Next, we draw another n_i/n_m samples from every distribution $\mathcal{P}_i$ to construct stochastic gradients at $(\mathbf{w}_{t+1}, \mathbf{q}_{t+1})$:

$$\begin{aligned}
&\mathbf{g}_w(\mathbf{w}_{t+1}, \mathbf{q}_{t+1}) \\
&= \sum_{i=1}^{m} q_{t+1,i} p_i \left(\frac{n_m}{n_i}\sum_{j=1}^{n_i/n_m} \nabla \ell(\mathbf{w}_{t+1}; \hat{\mathbf{z}}_t^{(i,j)})\right), \\
&\mathbf{g}_q(\mathbf{w}_{t+1}, \mathbf{q}_{t+1}) \\
&= \Bigg[p_1 \frac{n_m}{n_1}\sum_{j=1}^{n_1/n_m}\big[\ell(\mathbf{w}_{t+1}; \hat{\mathbf{z}}_t^{(1,j)}) - \ell(\bar{\mathbf{w}}^{(1)}; \hat{\mathbf{z}}_t^{(1,j)})\big], \\
&\qquad \ldots, p_m\big[\ell(\mathbf{w}_{t+1}; \hat{\mathbf{z}}_t^{(m)}) - \ell(\bar{\mathbf{w}}^{(m)}; \hat{\mathbf{z}}_t^{(m)})\big]\Bigg]^\top.
\end{aligned} \tag{28}$$

where $\hat{\mathbf{z}}_t^{(i,1)}, \ldots, \hat{\mathbf{z}}_t^{(i,n_i/n_m)}$ are random samples from distribution $\mathcal{P}_i$. Then, we use them to update $(\mathbf{w}'_t, \mathbf{q}'_t)$, and obtain $(\mathbf{w}'_{t+1}, \mathbf{q}'_{t+1})$:

$$\mathbf{w}'_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}}\big\{\eta_w\langle \mathbf{g}_w(\mathbf{w}_{t+1}, \mathbf{q}_{t+1}), \mathbf{w} - \mathbf{w}'_t\rangle + B_w(\mathbf{w}, \mathbf{w}'_t)\big\}, \tag{29}$$

$$\mathbf{q}'_{t+1} = \operatorname*{argmin}_{\mathbf{q}\in\Delta_m}\big\{\eta_q\langle -\mathbf{g}_q(\mathbf{w}_{t+1}, \mathbf{q}_{t+1}), \mathbf{q} - \mathbf{q}'_t\rangle + B_q(\mathbf{q}, \mathbf{q}'_t)\big\} \tag{30}$$

Recall that after the first stage, we have $n_i/2$ samples left for each distribution $\mathcal{P}_i$. So, we repeat the above process for $n_m/4$ iterations to meet the budget constraints. Finally, we return $\bar{\mathbf{w}} = \frac{4}{n_m}\sum_{t=2}^{1+n_m/4}\mathbf{w}_t$ and $\bar{\mathbf{q}} = \frac{4}{n_m}\sum_{t=2}^{1+n_m/4}\mathbf{q}_t$ as solutions.

To analyze the performance of Algorithm 2, we introduce two additional assumptions.

Assumption 4.3. All the risk functions are L-smooth, i.e.,

$$\|\nabla R_i(\mathbf{w}) - \nabla R_i(\mathbf{w}')\|_{w,*} \le L\|\mathbf{w} - \mathbf{w}'\|_w \tag{31}$$

for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, and $i \in [m]$.

The assumption of smoothness is essential for achieving a convergence rate that depends on the variance (Lan, 2012).

Assumption 4.4. The dual norm $\|\cdot\|_{w,*}$ is κ-regular for some small constant $\kappa \ge 1$.

Algorithm 2 A Two-Stage Stochastic Approximation Approach for Weighted MERO

Input: Step sizes: $\eta^{(1)}, \ldots, \eta^{(m)}$, η_w and η_q

Initialize $\mathbf{w}_1^{(1)} = \cdots = \mathbf{w}_1^{(m)} = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w})$
for $i = 1$ to m **do**
  for $t = 1$ to $n_i/2$ **do**
    Draw a sample $\mathbf{z}_t^{(i)}$ from distribution $\mathcal{P}_i$
    Calculate $\nabla\ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)})$ and update $\mathbf{w}_t^{(i)}$ according to (22)
  end for
  Calculate the average of iterates $\bar{\mathbf{w}}^{(i)} = \frac{1}{n_i/2}\sum_{t=1}^{n_i/2}\mathbf{w}_{t+1}^{(i)}$
end for
Initialize $\mathbf{w}'_1 = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \nu_w(\mathbf{w})$, and $\mathbf{q}'_1 = \frac{1}{m}\mathbf{1}_m \in \mathbb{R}^m$
for $t = 1$ to $n_m/4$ **do**
  For each $i \in [m]$, draw n_i/n_m samples $\{\mathbf{z}_t^{(i,j)} : j = 1, \ldots, n_i/n_m\}$ from distribution $\mathcal{P}_i$
  Construct the stochastic gradients defined in (25)
  Calculate $\mathbf{w}_{t+1}$ and $\mathbf{q}_{t+1}$ according to (26) and (27), respectively
  For each $i \in [m]$, draw n_i/n_m samples $\{\hat{\mathbf{z}}_t^{(i,j)} : j = 1, \ldots, n_i/n_m\}$ from distribution $\mathcal{P}_i$
  Construct the stochastic gradients defined in (28)
  Calculate $\mathbf{w}'_{t+1}$ and $\mathbf{q}'_{t+1}$ according to (29) and (30), respectively
end for
return $\bar{\mathbf{w}} = \frac{4}{n_m}\sum_{t=2}^{1+n_m/4}\mathbf{w}_t$ and $\bar{\mathbf{q}} = \frac{4}{n_m}\sum_{t=2}^{1+n_m/4}\mathbf{q}_t$

The condition of regularity plays a role when examining the impact of mini-batches on stochastic gradients. For a comprehensive definition, please consult the work of Juditsky & Nemirovski (2008).

Following Zhang et al. (2023), we set the weight p_i in (21) as

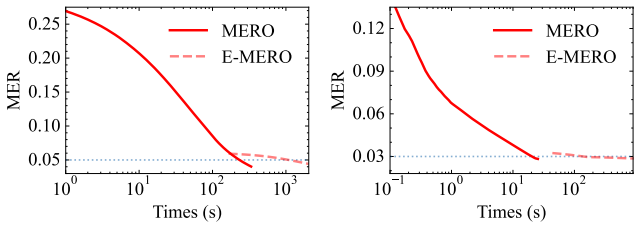
$$p_i = \frac{1/\sqrt{n_m} + 1}{1/\sqrt{n_m} + \sqrt{n_m/n_i}}. \tag{32}$$

Then, we have the following theorem regarding the excess risk of $\bar{\mathbf{w}}$ on every distribution.

Theorem 4.5. *Under Assumptions 3.1, 3.2, 3.3, 3.4, 4.3, and 4.4, and setting appropriate parameters in Algorithm 2, with high probability, we have*

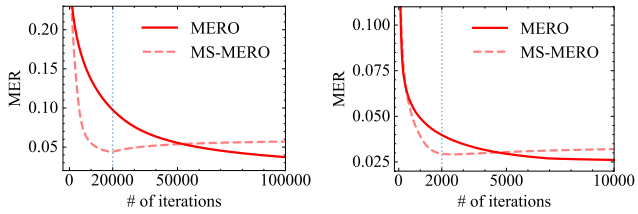
$$R_i(\bar{\mathbf{w}}) - R_i^* = \frac{1}{p_i}p_\varphi^* + O\left(\left(\frac{1}{n_m} + \frac{1}{\sqrt{n_i}}\right)\ln m\right).$$

Remark 4.6. We observe that for a distribution $\mathcal{P}_i$ with $n_i \le n_m^2$, the excess risk diminishes at a rate of $O((\log m)/\sqrt{n_i})$,



(a) The synthetic dataset (b) The Adult dataset

Figure 1. The maximal excess risk (MER) versus the running time.



(a) The synthetic dataset (b) The Adult dataset

Figure 2. The maximal excess risk (MER) versus the number of iterations.

a significant improvement over the $\widetilde{O}(\sqrt{(\log m)/n_m})$ rate outlined in Section 4.1, until it approaches p_φ^*/p_i. For any $\mathcal{P}_i$ with a very large budget, i.e., $n_i > n_m^2$, it attains an $O((\log m)/n_m)$ rate, which almost matches the convergence rate of deterministic convex-concave saddle-point optimization (Nemirovski, 2004).

Although the exact value of p_φ^* is generally unknown, we can expect it to be relatively small when there exists a single model that performs well on all distributions. In particular, if all the distributions are aligned, we can prove that $p_\varphi^* = 0$ (Agarwal & Zhang, 2022, Corollary 9), leading to the following corollary.

Corollary 4.7. *Suppose there exists a model $\mathbf{w}_* \in \mathcal{W}$ such that $R_i(\mathbf{w}_*) = R_i^*$ for all $i \in [m]$. Under the condition of Theorem 4.5, with high probability, we have*

$$R_i(\bar{\mathbf{w}}) - R_i^* = O\left(\left(\frac{1}{n_m} + \frac{1}{\sqrt{n_i}}\right)\ln m\right).$$

Remark 4.8. When all distributions are aligned, the above corollary offers upper bounds for the *standard* excess risk, making it more interpretable than the theoretical guarantee provided by Zhang et al. (2023, Theorem 4).

5. Experiments

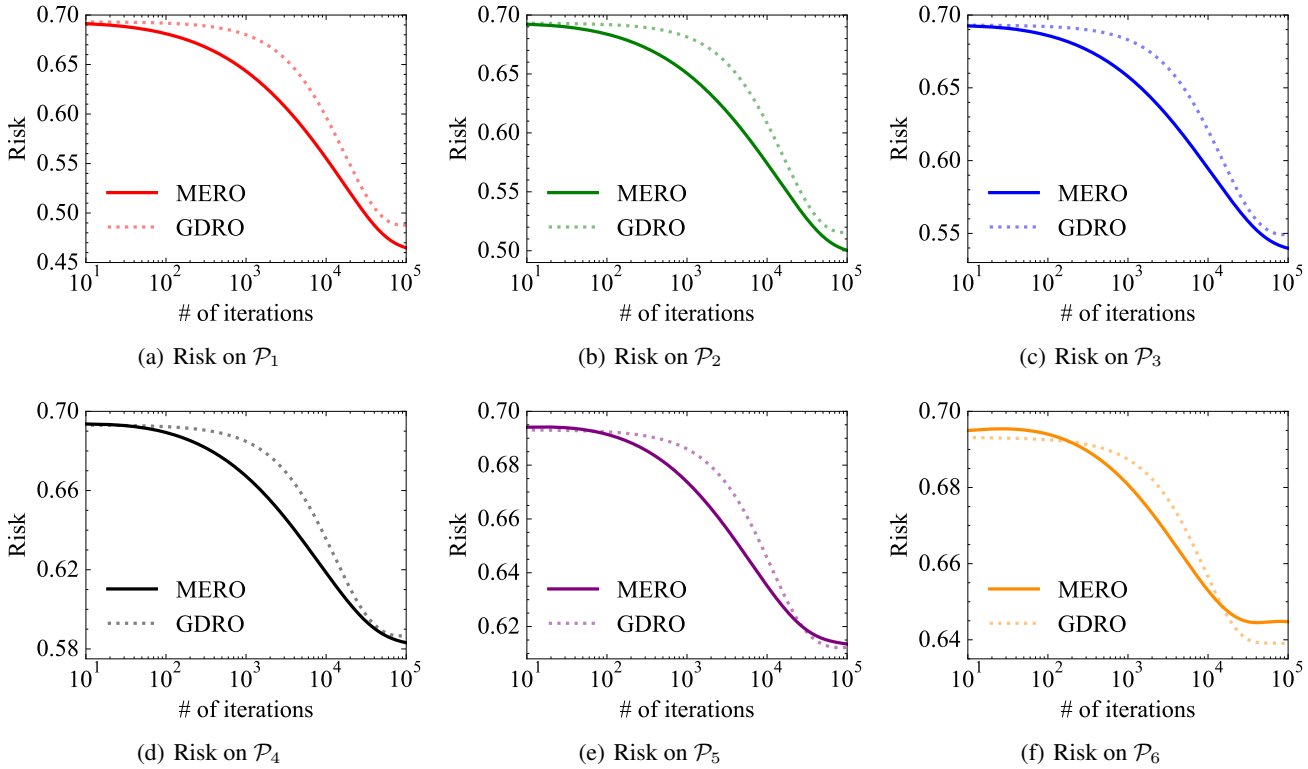
In the experiments, we investigate both the balanced and imbalanced scenarios, employing synthetic and real-world datasets. Due to limited space, we only present a subset of the experimental outcomes, with the comprehensive set of results accessible in Appendix C.

First, we demonstrate the efficiency of our anytime stochastic approximation approach in Algorithm 1, referred to as MERO, through a comparative analysis with the optimization procedure of Agarwal & Zhang (2022) for empirical MERO, denominated as E-MERO. Recall that Algorithm 1 runs for T rounds, making use of a cumulative total of mT samples. For a fair comparison, we set the number of samples from each distribution in E-MERO to be T. We assign the value of T as 10^5 for the synthetic dataset and 10^4 for the Adult dataset. In Fig. 1, we depict the relationship between the maximal excess risk (MER), i.e., $\max_{i\in[m]}[R_i(\mathbf{w}) - R_i^*]$, and the running time. The lack of the E-MERO curve at the beginning is attributed to its initialization phase, a period focused on minimizing the empirical risk for each distribution. It's evident that MERO achieves convergence significantly quicker than E-MERO for both datasets, highlighting the computational efficiency of stochastic approximation. To be more clear, we present in Table 1 the time required for MERO and E-MERO to attain a specified MER target—0.05 for the synthetic dataset and 0.03 for the Adult dataset. The data reveals that MERO outpaces E-MERO, being 5.7 times faster and 6.9 times faster on the synthetic and Adult datasets, respectively. Additionally, it's important to note that MERO is more memory-efficient, as it eliminates the need to store training data.

Table 1. Running times of MERO and E-MERO.

DATASET	ALGORITHM	MER VALUE	TIMES (S)
SYNTHETIC	MERO	0.05	255.0
	E-MERO	0.05	1470.5
ADULT	MERO	0.03	20.4
	E-MERO	0.03	141.6

Second, we illustrate the benefit of the anytime capability of Algorithm 1 by comparing with the multi-stage approach detailed in Section 3.2, denoted by MS-MERO. In the case of MS-MERO, we assign a preset value of T as 2×10^4 for the synthetic dataset and 2×10^3 for the Adult dataset. However, we continue the execution of the 3rd stage even when the count of iterations goes beyond the set value of T. In Fig. 2, the graph displays the progression of MER in relation to the number of iterations. We exclude the initial two stages of MS-MERO from the illustration, because it only produces a model during the 3rd stage. In the beginning, MS-MERO rapidly reduces the MER, but as the actual number of executed rounds exceeds the predetermined limit, its parameter settings are not optimal. Consequently, we observe a stagnation or even a slight increase in the MER. In contrast, MERO demonstrates a consistent decrease in MER and ultimately outperforms MS-MERO. Such results underscore the importance of anytime algorithms, especially when the total number of iterations is unknown.



(a) Risk on $\mathcal{P}_1$ (b) Risk on $\mathcal{P}_2$ (c) Risk on $\mathcal{P}_3$

(d) Risk on $\mathcal{P}_4$ (e) Risk on $\mathcal{P}_5$ (f) Risk on $\mathcal{P}_6$

Figure 3. Individual risk versus the number of iterations on the synthetic dataset.

Third, we contrast MERO with GDRO to analyze their unique characteristics. Notice that MERO and GDRO adopt different objectives, and therefore neither holds absolute superiority (Agarwal & Zhang, 2022, Proposition 1 and Example 1). While the MER of MERO is always better than that of GDRO, it does not necessarily mean that the model yielded by MERO is universally preferable across all distributions. To illustrate this point, we compare our Algorithm 1 with the stochastic approximation algorithm of Zhang et al. (2023, Algorithm 1) designed for GDRO, which is also referred to as GDRO for convenience. We assess the risks associated with each distribution for both MERO and GDRO, and in the initial comparison, we set the x-axis as the number of iterations to ensure that both algorithms consume the same number of samples. Experimental results on the synthetic dataset are presented in Fig. 3. As can be seen, GDRO exhibits strong performance with distributions $\mathcal{P}_5$ and $\mathcal{P}_6$, while MERO demonstrates superior results with the remaining 4 distributions. This pattern is as expected, since GDRO targets the raw risk, and the last two distributions are characterized by the high level of noise, hence the large risk. Consequently, GDRO tends to concentrate its efforts on these distributions, achieving lower risks for them. By contrast, MERO effectively mitigates the impact of noise, achieving a more balanced performance across various distributions. Experimental results on the Adult dataset, showcased in Fig. 4 of Appendix C, lead us to similar conclusions: GDRO exhibits slightly better performance on distributions ($\mathcal{P}_1$, $\mathcal{P}_3$ and $\mathcal{P}_5$) with large risk.

6. Conclusion and Future Work

This paper aims to develop efficient stochastic approximation approaches for MERO. First, we design a multi-stage stochastic algorithm, which attains a (nearly) optimal convergence rate of $O(\sqrt{(\log m)/T})$ for a fixed number of iterations T. Then, we propose an anytime stochastic method, which reduces the error at an $\widetilde{O}(\sqrt{(\log m)/t})$ rate at every iteration t. Next, we delve into the setting where different distributions possess varying sample budgets, and develop a two-stage stochastic procedure that is endowed with distribution-dependent convergence rates. Finally, we substantiate the efficiency and effectiveness of our methods through experimental validation.

A future direction involves applying stochastic approximation to empirical MERO to reduce the computational cost. By leveraging the finite-sum structure (Zhang et al., 2013b; Johnson & Zhang, 2013; Reddi et al., 2016), we have recently made progress in this direction (Yu et al., 2024).

Acknowledgements

This work was partially supported by the National Science and Technology Major Project (2022ZD0114801) and NSFC (U23A20382, 62122037).

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

Abusorrah, A., Alabdulwahab, A., Li, Z., and Shahidehpour, M. Minimax-regret robust defensive strategy against false data injection attacks. *IEEE Transactions on Smart Grid*, 10(2):2068–2079, 2019.

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

Alaiz-Rodríguez, R., Guerrero-Curieses, A., and Cid-Sueiro, J. Minimax regret classifier for imprecise class distributions. *Journal of Machine Learning Research*, 8(4): 103–130, 2007.

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

Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

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

Ben-Tal, A., Hazan, E., Koren, T., and Mannor, S. Oracle-based robust optimization via online learning. *Operations Research*, 63(3):628–638, 2015.

Bertsimas, D., Gupta, V., and Kallus, N. Robust sample average approximation. *Mathematical Programming*, 171: 217–282, 2018.

Blum, A., Haghtalab, N., Procaccia, A. D., and Qiao, M. Collaborative PAC learning. In *Advances in Neural Information Processing Systems 30*, pp. 2389–2398, 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.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

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

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.

Eldar, Y. C., , and Merhav, N. A competitive minimax approach to robust estimation of random parameters. *IEEE Transactions on Signal Processing*, 52(7):1931–1946, 2004.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Karbasi, A. Learning distributionally robust models at scale via composite optimization. In *International Conference on Learning Representations*, 2022.

Harvey, N. J. A., Liaw, C., Plan, Y., and Randhawa, S. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the 32nd Conference on Learning Theory*, pp. 1579–1613, 2019.

Jain, P., Nagaraj, D., and Netrapalli, P. Making the last iterate of sgd information theoretically optimal. In *Proceedings of the 32nd Conference on Learning Theory*, pp. 1752–1755, 2019.

Jiang, R., Wang, J., Zhang, M., and Guan, Y. Two-stage minimax regret robust unit commitment. *IEEE Transactions on Power Systems*, 28(3):2271–2282, 2013.

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 *Advances 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.

Juditsky, A. B. and Nemirovski, A. S. Large deviations of vector-valued martingales in 2-smooth normed spaces. *ArXiv e-prints*, arXiv:0809.0813, 2008.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5637–5664, 2021.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics*, pp. 130–166, 2019.

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

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems 33*, pp. 8847–8860, 2020.

Lugosi, G. Concentration-of-measure inequalities. Technical report, Department of Economics, Pompeu Fabra University, 2009.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, pp. 2216–2224, 2016.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems 30*, pp. 2971–2980, 2017.

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.

Nguyen, V. A., Shafieezadeh Abadeh, S., Yue, M.-C., Kuhn, D., and Wiesemann, W. Optimistic distributionally robust optimization for nonparametric likelihood approximation. In *Advances in Neural Information Processing Systems 32*, pp. 15872–15882, 2019.

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.

Qi, Q., Guo, Z., Xu, Y., Jin, R., and Yang, T. An online method for a class of distributionally robust optimization with non-convex objectives. In *Advances in Neural Information Processing Systems 34*, pp. 10067–10080, 2021.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Weakly-convex-concave min-max optimization: Provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

Rahimian, H. and Mehrotra, S. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3(4), 2022.

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

Roux, N. L., Manzagol, P.-A., and Bengio, Y. Topmoumoute online natural gradient algorithm. In *Advances in Neural Information Processing Systems 20*, pp. 849–856, 2008.

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 *International Conference on Learning Representations*, 2020.

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

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

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

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

Song, C., Lin, C. Y., Wright, S. J., and Diakonikolas, J. Coordinate linear variance reduction for generalized linear programming. In *Advances in Neural Information Processing Systems 35*, pp. 22049–22063, 2022.

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

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Wiesemann, W., Kuhn, D., and Sim, M. Distributionally robust convex optimization. *Operations Research*, 62(6): 1358–1376, 2014.

Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q. V., Ma, T., and Yu, A. W. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems 36*, pp. 69798–69818, 2023.

Yu, D., Cai, Y., Jiang, W., and Zhang, L. Efficient algorithms for empirical group distributional robust optimization and beyond. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *Proceedings of the30th International Conference on Machine Learning*, pp. 819–827, 2013a.

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, 2013b.

Zhang, L., Yang, T., Jin, R., and He, X. $O(\log T)$ projections for stochastic optimization of smooth and strongly convex functions. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1121–1129, 2013c.

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.

A. More Related Work

For the DRO problem in (1), the uncertain set $\mathcal{S}$ is typically chosen to be the neighborhood surrounding a target distribution $\mathcal{P}_0$, and constructed by using certain distance functions between distributions, such as the f-divergence (Ben-Tal et al., 2013), the Wasserstein distance (Esfahani & Kuhn, 2018; Kuhn et al., 2019), and the maximum mean discrepancy (Staib & Jegelka, 2019). Other ways for defining $\mathcal{S}$ include moment constraints (Delage & Ye, 2010; Wiesemann et al., 2014) and hypothesis testing of goodness-of-fit (Bertsimas et al., 2018). The nature of the loss function can exhibit variability: it could assume a convex form (Ben-Tal et al., 2015; Shapiro, 2017), a non-convex structure (Jin et al., 2021; Qi et al., 2021), or potentially incorporate a regularizer (Sinha et al., 2018). Research efforts may be directed towards diverse objectives, including the development of the optimization algorithms (Namkoong & Duchi, 2016; Levy et al., 2020; Rafique et al., 2022; Haddadpour et al., 2022; Song et al., 2022), the exploration of finite sample and asymptotic properties of the empirical solution (Namkoong & Duchi, 2017; Duchi & Namkoong, 2021), the determination of confidence intervals for the risk (Duchi et al., 2021), or the approximation of nonparametric likelihood (Nguyen et al., 2019).

B. Analysis

In this section, we present proofs of main theorems.

B.1. Proof of Theorem 3.6

Besides the high probability bound in (19), we will also establish an expectation bound:

$$\mathrm{E}\left[R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i^*\right] \leq \frac{DG(3+\ln t)}{4(\sqrt{t+1}-1)}, \quad \forall i \in [m],\ t \in \mathbb{Z}_+. \tag{33}$$

The analysis closely adheres to the content in Section 2.3 of Nemirovski et al. (2009), and for the sake of completeness, we present the proof here.

Let $\mathbf{w}_*^{(i)} \in \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} R_i(\mathbf{w})$ be the optimal solution that minimizes $R_i(\cdot)$. From the property of mirror descent, e.g., Lemma 2.1 of Nemirovski et al. (2009), we have

$$\begin{aligned}
&\eta_j^{(i)}\langle \nabla \ell(\mathbf{w}_j^{(i)};\mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle \\
\leq& B_w(\mathbf{w}_*^{(i)}, \mathbf{w}_j^{(i)}) - B_w(\mathbf{w}_*^{(i)}, \mathbf{w}_{j+1}^{(i)}) + \frac{(\eta_j^{(i)})^2}{2}\|\nabla \ell(\mathbf{w}_j^{(i)};\mathbf{z}_j^{(i)})\|_{w,*}^2 \\
\overset{(7)}{\leq}& B_w(\mathbf{w}_*^{(i)}, \mathbf{w}_j^{(i)}) - B_w(\mathbf{w}_*^{(i)}, \mathbf{w}_{j+1}^{(i)}) + \frac{(\eta_j^{(i)})^2 G^2}{2}.
\end{aligned} \tag{34}$$

Thus, we have

$$\begin{aligned}
&\eta_j^{(i)}\langle \nabla R_i(\mathbf{w}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle \\
=&\eta_j^{(i)}\langle \nabla \ell(\mathbf{w}_j^{(i)};\mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle + \eta_j^{(i)}\langle \nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)};\mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle \\
\overset{(34)}{\leq}& B_w(\mathbf{w}_*^{(i)}, \mathbf{w}_j^{(i)}) - B_w(\mathbf{w}_*^{(i)}, \mathbf{w}_{j+1}^{(i)}) + \frac{(\eta_j^{(i)})^2 G^2}{2} + \eta_j^{(i)}\langle \nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)};\mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle.
\end{aligned}$$

Summing the above inequality over $j = 1, \ldots, t$, we have

$$\begin{aligned}
&\sum_{j=1}^t \eta_j^{(i)}\langle \nabla R_i(\mathbf{w}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle \\
\leq& B_w(\mathbf{w}_*^{(i)}, \mathbf{w}_1^{(i)}) + \frac{G^2}{2}\sum_{j=1}^t (\eta_j^{(i)})^2 + \sum_{j=1}^t \eta_j^{(i)}\langle \nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)};\mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle \\
\overset{(6)}{\leq}& D^2 + \frac{G^2}{2}\sum_{j=1}^t (\eta_j^{(i)})^2 + \sum_{j=1}^t \eta_j^{(i)}\langle \nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)};\mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle.
\end{aligned} \tag{35}$$

From the convexity of the risk function, we have

$$\begin{aligned}
&R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)}) = R_i\left(\sum_{j=1}^t \frac{\eta_j^{(i)}\mathbf{w}_j^{(i)}}{\sum_{k=1}^t \eta_k^{(i)}}\right) - R_i(\mathbf{w}_*^{(i)}) \\
\leq &\left(\sum_{j=1}^t \frac{\eta_j^{(i)}}{\sum_{k=1}^t \eta_k^{(i)}} R_i(\mathbf{w}_j^{(i)})\right) - R_i(\mathbf{w}_*^{(i)}) = \sum_{j=1}^t \frac{1}{\sum_{k=1}^t \eta_k^{(i)}} \eta_j^{(i)} \left(R_i(\mathbf{w}_j^{(i)}) - R_i(\mathbf{w}_*^{(i)})\right) \\
\leq &\sum_{j=1}^t \frac{1}{\sum_{k=1}^t \eta_k^{(i)}} \eta_j^{(i)} \langle \nabla R_i(\mathbf{w}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)} \rangle \\
\overset{(35)}{\leq} &\frac{D^2 + \frac{G^2}{2}\sum_{j=1}^t (\eta_j^{(i)})^2 + \sum_{j=1}^t \eta_j^{(i)} \langle \nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)} \rangle}{\sum_{j=1}^t \eta_j^{(i)}}.
\end{aligned} \tag{36}$$

Define

$$\delta_j^{(i)} = \eta_j^{(i)} \langle \nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)} \rangle. \tag{37}$$

Recall that $\mathbf{w}_j^{(i)}$ and $\mathbf{w}_*^{(i)}$ do not depend on $\mathbf{z}_j^{(i)}$, and thus

$$\mathrm{E}_{j-1}[\delta_j^{(i)}] = \eta_j^{(i)} \left\langle \mathrm{E}_{j-1}\left[\nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)})\right], \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)} \right\rangle = 0.$$

So, $\delta_1^{(i)}, \ldots, \delta_t^{(i)}$ is a martingale difference sequence.

B.1.1. The Expectation Bound

Taking expectation over (36), we have

$$\mathrm{E}\left[R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)})\right] \leq \frac{2D^2 + G^2 \sum_{j=1}^t (\eta_j^{(i)})^2}{2\sum_{j=1}^t \eta_j^{(i)}}. \tag{38}$$

By setting $\eta_j^{(i)} = \frac{D}{G\sqrt{j}}$, we have

$$\mathrm{E}\left[R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)})\right] \leq \frac{2DG + DG\sum_{j=1}^t \frac{1}{j}}{2\sum_{j=1}^t \frac{1}{\sqrt{j}}}. \tag{39}$$

Notice that

$$\begin{aligned}
\sum_{j=1}^t \frac{1}{j} &\leq 1 + \int_1^t \frac{1}{x} dx = 1 + \ln x|_1^t = 1 + \ln t, \\
\sum_{j=1}^t \frac{1}{\sqrt{j}} &\geq \int_1^{t+1} \frac{1}{\sqrt{x}} dx = 2\sqrt{x}|_1^{t+1} = 2(\sqrt{t+1} - 1).
\end{aligned} \tag{40}$$

From (39) and (40), we have

$$\mathrm{E}\left[R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)})\right] \leq \frac{DG(3 + \ln t)}{4(\sqrt{t+1} - 1)}$$

which proves (33).

B.1.2. The High Probability Bound

To establish the high probability bound, we make use of the Hoeffding-Azuma inequality for martingales (Cesa-Bianchi & Lugosi, 2006).

Lemma B.1. *Let $V_1, V_2, \ldots$ be a martingale difference sequence with respect to some sequence $X_1, X_2, \ldots$ such that $V_i \in [A_i, A_i + c_i]$ for some random variable A_i, measurable with respect to $X_1, \ldots, X_{i-1}$ and a positive constant c_i. If $S_n = \sum_{i=1}^n V_i$, then for any $t > 0$,*

$$\Pr[S_n > t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

To apply the above lemma, we need to show that $|\delta_j^{(i)}|$ is bounded. We have

$$\begin{aligned}
&\left\|\nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)})\right\|_{w,*} \leq \left\|\nabla R_i(\mathbf{w}_j^{(i)})\right\|_{w,*} + \left\|\nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)})\right\|_{w,*} \\
\leq &\mathrm{E}_{t-1}\left[\left\|\nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)})\right\|_{w,*}\right] + \left\|\nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)})\right\|_{w,*} \overset{(7)}{\leq} 2G,
\end{aligned} \tag{41}$$

$$\begin{aligned}
&\left\|\mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\right\|_w \leq \left\|\mathbf{w}_j^{(i)} - \mathbf{w}_1^{(i)}\right\|_w + \left\|\mathbf{w}_1^{(i)} - \mathbf{w}_*^{(i)}\right\|_w \\
\leq &\left(\sqrt{2B_w(\mathbf{w}_j^{(i)}, \mathbf{w}_1^{(i)})} + \sqrt{2B_w(\mathbf{w}_*^{(i)}, \mathbf{w}_1^{(i)})}\right) \overset{(6)}{\leq} 2\sqrt{2}D.
\end{aligned} \tag{42}$$

As a result,

$$\begin{aligned}
|\delta_j^{(i)}| =&\eta_j^{(i)} \left|\langle \nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)}), \mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\rangle\right| \\
\leq&\eta_j^{(i)} \left\|\nabla R_i(\mathbf{w}_j^{(i)}) - \nabla \ell(\mathbf{w}_j^{(i)}; \mathbf{z}_j^{(i)})\right\|_{w,*} \left\|\mathbf{w}_j^{(i)} - \mathbf{w}_*^{(i)}\right\|_w \overset{(41),(42)}{\leq} 4\sqrt{2}\eta_j^{(i)} DG.
\end{aligned} \tag{43}$$

From Lemma B.1, with probability at least $1 - \delta/[2mt^2]$, we have

$$\sum_{j=1}^t \delta_j^{(i)} \leq 8DG\sqrt{\sum_{j=1}^t (\eta_j^{(i)})^2 \left(\ln \frac{2mt^2}{\delta}\right)}. \tag{44}$$

Substituting (44) into (36), with probability at least $1 - \delta/[2mt^2]$, we have

$$\begin{aligned}
R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)}) \leq &\frac{D^2 + \frac{G^2}{2}\sum_{j=1}^t (\eta_j^{(i)})^2 + 8DG\sqrt{\sum_{j=1}^t (\eta_j^{(i)})^2 \left(\ln \frac{2mt^2}{\delta}\right)}}{\sum_{j=1}^t \eta_j^{(i)}} \\
=&\frac{2DG + DG\sum_{j=1}^t \frac{1}{j} + 16DG\sqrt{\sum_{j=1}^t \frac{1}{j}\left(\ln \frac{2mt^2}{\delta}\right)}}{2\sum_{j=1}^t \frac{1}{\sqrt{j}}} \\
\overset{(40)}{\leq} &\frac{DG\left[3 + \ln t + 16\sqrt{(1 + \ln t)\ln(2mt^2/\delta)}\right]}{4(\sqrt{t+1} - 1)}.
\end{aligned}$$

We complete the proof by taking the union bound over all $i \in [m]$ and $t \in \mathbb{Z}_+$, and using the well-known fact

$$\sum_{t=1}^\infty \frac{1}{t^2} = \frac{\pi^2}{6} \leq 2.$$

B.2. Proof of Theorem 3.7

We first present the complete form of Theorem 3.7.

Theorem B.2. *Under Assumptions 3.1, 3.2, 3.3, and 3.4, by setting*

$$\eta_t^{(i)} = \frac{D}{G\sqrt{t}},\ \eta_t^w = \frac{2D^2}{\sqrt{(2D^2G^2 + 2\ln m)t}},\ \textit{and}\ \eta_t^q = \frac{2\ln m}{\sqrt{(2D^2G^2 + 2\ln m)t}} \tag{45}$$

in Algorithm 1, we have

$$\begin{aligned}
&\mathrm{E}\left[\epsilon_\phi(\bar{\mathbf{w}}_t, \bar{\mathbf{q}}_t) = \max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}}_t, \mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \phi(\mathbf{w}, \bar{\mathbf{q}}_t)\right] \\
\leq &\frac{(5+3\ln t)\sqrt{2D^2G^2+2\ln m} + 2DG\left[3+\ln t + 16(1+\sqrt{\ln m})\sqrt{2(1+\ln t)}\right](1+\ln t)}{2(\sqrt{t+1}-1)} \\
= &O\left(\frac{\log^2 t + \log^{1/2} m \log^{3/2} t}{\sqrt{t}}\right)
\end{aligned} \tag{46}$$

for all $t \in \mathbb{Z}_+$*. Furthermore, with probability at least* $1-2\delta$*,*

$$\begin{aligned}
&\epsilon_\phi(\bar{\mathbf{w}}_t, \bar{\mathbf{q}}_t) = \max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}}_t, \mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \phi(\mathbf{w}, \bar{\mathbf{q}}_t) \\
\leq &\frac{1}{2(\sqrt{t+1}-1)}\left[\sqrt{2D^2G^2+2\ln m}\left(5+3\ln t + 8\sqrt{(1+\ln t)\ln\frac{2t^2}{\delta}}\right)\right. \\
&\left. +2DG\left(3+\ln t + 16\sqrt{(1+\ln t)\ln\frac{2mt^2}{\delta}}\right)(1+\ln t)\right] = O\left(\frac{\log^2 t + \log^{1/2} m \log^{3/2} t}{\sqrt{t}}\right)
\end{aligned} \tag{47}$$

for all $t \in \mathbb{Z}_+$*.*

Following the analysis of Nemirovski et al. (2009, §3.1), we will first combine the two update rules in (16) and (17) into a single one.

B.2.1. Merging the Two Update Rules in (16) and (17)

Let $\mathcal{E}$ be the space in which $\mathcal{W}$ resides. We equip the Cartesian product $\mathcal{E} \times \mathbb{R}^m$ with the following norm and dual norm:

$$\left\|(\mathbf{w}, \mathbf{q})\right\| = \sqrt{\frac{1}{2D^2}\|\mathbf{w}\|_w^2 + \frac{1}{2\ln m}\|\mathbf{q}\|_1^2}, \text{ and } \left\|(\mathbf{u}, \mathbf{v})\right\|_* = \sqrt{2D^2\|\mathbf{u}\|_{w,*}^2 + 2\|\mathbf{v}\|_\infty^2 \ln m}. \tag{48}$$

We use the notation $\mathbf{x} = (\mathbf{w}, \mathbf{q})$, and equip the set $\mathcal{W} \times \Delta_m$ with the distance-generating function

$$\nu(\mathbf{x}) = \nu(\mathbf{w}, \mathbf{q}) = \frac{1}{2D^2}\nu_w(\mathbf{w}) + \frac{1}{2\ln m}\nu_q(\mathbf{q}). \tag{49}$$

It is easy to verify that $\nu(\mathbf{x})$ is 1-strongly convex w.r.t. the norm $\|\cdot\|$ in (48). Let $B(\cdot,\cdot)$ be the Bregman distance associated with $\nu(\cdot)$:

$$\begin{aligned}
B(\mathbf{x}, \mathbf{x}') =& \nu(\mathbf{x}) - \left[\nu(\mathbf{x}') + \langle \nabla\nu(\mathbf{x}'), \mathbf{x} - \mathbf{x}'\rangle\right] \\
=& \frac{1}{2D^2}\left(\nu_w(\mathbf{w}) - \left[\nu_w(\mathbf{w}') + \langle \nabla\nu_w(\mathbf{w}'), \mathbf{w} - \mathbf{w}'\rangle\right]\right) \\
&+ \frac{1}{2\ln m}\left(\nu_q(\mathbf{q}) - \left[\nu_q(\mathbf{q}') + \langle \nabla\nu_q(\mathbf{q}'), \mathbf{q} - \mathbf{q}'\rangle\right]\right) \\
=& \frac{1}{2D^2}B_w(\mathbf{w}, \mathbf{w}') + \frac{1}{2\ln m}B_q(\mathbf{q}, \mathbf{q}')
\end{aligned} \tag{50}$$

where $\mathbf{x}' = (\mathbf{w}', \mathbf{q}')$. Recall the definitions of $\mathbf{o}_w$ and $\mathbf{o}_q$ in Section 3.1, and we have

$$(\mathbf{o}_w, \mathbf{o}_q) = \operatorname*{argmin}_{(\mathbf{w},\mathbf{q})\in\mathcal{W}\times\Delta_m} \nu(\mathbf{w}, \mathbf{q}).$$

Then, we can show that the domain $\mathcal{W} \times \Delta_m$ is bounded since

$$\max_{(\mathbf{w},\mathbf{q})\in\mathcal{W}\times\Delta_m} B([\mathbf{w}, \mathbf{q}], [\mathbf{o}_w, \mathbf{o}_q]) = \frac{1}{2D^2}\max_{\mathbf{w}\in\mathcal{W}} B_w(\mathbf{w}, \mathbf{o}_w) + \frac{1}{2\ln m}\max_{\mathbf{q}\in\Delta_m} B_q(\mathbf{q}, \mathbf{o}_q) \overset{(6)}{\leq} 1. \tag{51}$$

With the above configurations, (16) and (17) are equivalent to

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \left\{\eta_t\big\langle[\mathbf{g}_w(\mathbf{w}_t,\mathbf{q}_t), -\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t)], \mathbf{x}-\mathbf{x}_t\big\rangle + B(\mathbf{x},\mathbf{x}_t)\right\} \tag{52}$$

where $\eta_t > 0$ is the step size that satisfies

$$\eta_t^w = 2\eta_t D^2, \text{ and } \eta_t^q = 2\eta_t \ln m. \tag{53}$$

And in the beginning, we set $\mathbf{x}_1 = \operatorname{argmin}_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \nu(\mathbf{x}) = (\mathbf{w}_1, \mathbf{q}_1) = (\mathbf{o}_w, \mathbf{o}_q)$.

B.2.2. Analysis of SMD with Biased Stochastic Gradients

To simplify the notation, we define

$$\begin{aligned} F(\mathbf{w}_t,\mathbf{q}_t) &= [\nabla_{\mathbf{w}}\phi(\mathbf{w}_t,\mathbf{q}_t), -\nabla_{\mathbf{q}}\phi(\mathbf{w}_t,\mathbf{q}_t)] \\ &= \left[\sum_{i=1}^m q_{t,i}\nabla R_i(\mathbf{w}_t), -\big[R_1(\mathbf{w}_t)-R_1^*,\ldots,R_m(\mathbf{w}_t)-R_m^*\big]^\top\right] \end{aligned} \tag{54}$$

which contains the true gradient of $\phi(\cdot,\cdot)$ at $(\mathbf{w}_t,\mathbf{q}_t)$, and

$$\begin{aligned} &\mathbf{g}(\mathbf{w}_t,\mathbf{q}_t) = [\mathbf{g}_w(\mathbf{w}_t,\mathbf{q}_t); -\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t)] \\ \overset{(14),(15)}{=}& \left[\sum_{i=1}^m q_{t,i}\nabla\ell(\mathbf{w}_t;\mathbf{z}_t^{(i)}), -\big[\ell(\mathbf{w}_t;\mathbf{z}_t^{(1)})-\ell(\bar{\mathbf{w}}_t^{(1)};\mathbf{z}_t^{(1)}),\ldots,\ell(\mathbf{w}_t;\mathbf{z}_t^{(m)})-\ell(\bar{\mathbf{w}}_t^{(m)};\mathbf{z}_t^{(m)})\big]^\top\right] \end{aligned} \tag{55}$$

which contains the stochastic gradient used in (52). The norm of the stochastic gradient is well-bounded:

$$\begin{aligned} \|\mathbf{g}_w(\mathbf{w}_t,\mathbf{q}_t)\|_{w,*} &= \left\|\sum_{i=1}^m q_{t,i}\nabla\ell(\mathbf{w}_t;\mathbf{z}_t^{(i)})\right\|_{w,*} \le \sum_{i=1}^m q_{t,i}\left\|\nabla\ell(\mathbf{w}_t;\mathbf{z}_t^{(i)})\right\|_{w,*} \overset{(7)}{\le} \sum_{i=1}^m q_{t,i}G = G, \\ \|\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t)\|_\infty &= \left\|\big[\ell(\mathbf{w}_t;\mathbf{z}_t^{(1)})-\ell(\bar{\mathbf{w}}_t^{(1)};\mathbf{z}_t^{(1)}),\ldots,\ell(\mathbf{w}_t;\mathbf{z}_t^{(m)})-\ell(\bar{\mathbf{w}}_t^{(m)};\mathbf{z}_t^{(m)})\big]^\top\right\|_\infty \overset{(8)}{\le} 1 \end{aligned}$$

and thus

$$\big\|\mathbf{g}(\mathbf{w}_t,\mathbf{q}_t)\big\|_* = \sqrt{2D^2\|\mathbf{g}_w(\mathbf{w}_t,\mathbf{q}_t)\|_{w,*}^2 + 2\|\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t)\|_\infty^2 \ln m} \le \underbrace{\sqrt{2D^2G^2+2\ln m}}_{:=M}. \tag{56}$$

The bias of $\mathbf{g}(\mathbf{w}_t,\mathbf{q}_t)$ is characterized by

$$F(\mathbf{w}_t,\mathbf{q}_t) - \mathrm{E}_{t-1}\left[\mathbf{g}(\mathbf{w}_t,\mathbf{q}_t)\right] = \left[0, -\big[R_1(\bar{\mathbf{w}}_t^{(1)})-R_1^*,\ldots,R_m(\bar{\mathbf{w}}_t^{(m)})-R_m^*\big]^\top\right]. \tag{57}$$

From the convexity-concavity of $\phi(\cdot,\cdot)$, we have (Nemirovski et al., 2009, (3.9))

$$\begin{aligned} &\max_{\mathbf{q}\in\Delta_m}\phi(\bar{\mathbf{w}}_t,\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}}\phi(\mathbf{w},\bar{\mathbf{q}}_t) \\ \overset{(18)}{=}& \max_{\mathbf{q}\in\Delta_m}\phi\left(\sum_{j=1}^t \frac{\eta_j^w\mathbf{w}_j}{\sum_{k=1}^t\eta_k^w}, \mathbf{q}\right) - \min_{\mathbf{w}\in\mathcal{W}}\phi\left(\mathbf{w}, \sum_{j=1}^t\frac{\eta_j^q\mathbf{q}_j}{\sum_{k=1}^t\eta_k^q}\right) \end{aligned}$$

$$
\begin{aligned}
&\overset{(53)}{=} \max_{\mathbf{q}\in\Delta_m} \phi\left(\sum_{j=1}^t \frac{\eta_j \mathbf{w}_j}{\sum_{k=1}^t \eta_k}, \mathbf{q}\right) - \min_{\mathbf{w}\in\mathcal{W}} \phi\left(\mathbf{w}, \sum_{j=1}^t \frac{\eta_j \mathbf{q}_j}{\sum_{k=1}^t \eta_k}\right) \\
&\leq \left(\sum_{j=1}^t \eta_j\right)^{-1} \left[\max_{\mathbf{q}\in\Delta_m} \sum_{j=1}^t \eta_j \phi(\mathbf{w}_j, \mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{j=1}^t \eta_j \phi(\mathbf{w}, \mathbf{q}_j)\right] \\
&\leq \left(\sum_{j=1}^t \eta_j\right)^{-1} \max_{(\mathbf{w},\mathbf{q})\in\mathcal{W}\times\Delta_m} \sum_{j=1}^t \eta_j \Big[\big\langle \nabla_{\mathbf{w}} \phi(\mathbf{w}_j, \mathbf{q}_j), \mathbf{w}_j - \mathbf{w}\big\rangle - \big\langle \nabla_{\mathbf{q}} \phi(\mathbf{w}_j, \mathbf{q}_j), \mathbf{q}_j - \mathbf{q}\big\rangle\Big] \\
&\overset{(54)}{=} \left(\sum_{j=1}^t \eta_j\right)^{-1} \max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \sum_{j=1}^t \eta_j \big\langle F(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\big\rangle.
\end{aligned}
\tag{58}
$$

As a result, we can decompose the optimization error as follows:

$$
\begin{aligned}
&\max_{\mathbf{q}\in\Delta_m} \phi(\bar{\mathbf{w}}_t, \mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \phi(\mathbf{w}, \bar{\mathbf{q}}_t) \\
&\leq \left(\sum_{j=1}^t \eta_j\right)^{-1} \underbrace{\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \sum_{j=1}^t \eta_j \big\langle \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\big\rangle}_{:=E_1} \\
&\quad + \left(\sum_{j=1}^t \eta_j\right)^{-1} \underbrace{\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \sum_{j=1}^t \eta_j \big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\big\rangle}_{:=E_2} \\
&\quad + \left(\sum_{j=1}^t \eta_j\right)^{-1} \underbrace{\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \sum_{j=1}^t \eta_j \big\langle F(\mathbf{w}_j, \mathbf{q}_j) - \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right], \mathbf{x}_j - \mathbf{x}\big\rangle}_{:=E_3}.
\end{aligned}
\tag{59}
$$

We proceed to bound the three terms in (59). To bound the first term E_1, we follow the derivation of (34), and have

$$
\begin{aligned}
\eta_j \big\langle \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\big\rangle &\leq B(\mathbf{x}, \mathbf{x}_j) - B(\mathbf{x}, \mathbf{x}_{j+1}) + \frac{\eta_j^2}{2} \|\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\|_*^2 \\
&\overset{(56)}{\leq} B(\mathbf{x}, \mathbf{x}_j) - B(\mathbf{x}, \mathbf{x}_{j+1}) + \frac{M^2}{2} \eta_j^2.
\end{aligned}
\tag{60}
$$

Summing the above inequality over $j = 1, \ldots, t$, we have

$$
\sum_{j=1}^t \eta_j \big\langle \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\big\rangle \leq B(\mathbf{x}, \mathbf{x}_1) + \frac{M^2}{2} \sum_{j=1}^t \eta_j^2 \overset{(51)}{\leq} 1 + \frac{M^2}{2} \sum_{j=1}^t \eta_j^2. \tag{61}
$$

Next, we consider the second term E_2. Because of the maximization operation, the variable $\mathbf{x}$ depends on the randomness of the algorithm, and thus we cannot treat $\eta_j \langle \mathrm{E}_{j-1}[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\rangle$, $j = 1, \ldots, t$ as a martingale difference sequence. To address this challenge, we make use of the "ghost iterate" technique of Nemirovski et al. (2009), and develop the following lemma.

Lemma B.3. *Under the condition of Theorem 3.7, we have*

$$
\mathrm{E}\left[\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \sum_{j=1}^t \eta_j \big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\big\rangle\right] \leq 1 + 2M^2 \sum_{j=1}^t \eta_j^2, \quad \forall t \in \mathbb{Z}_+. \tag{62}
$$

Furthermore, with probability at least $1-\delta$*,*

$$\begin{aligned}&\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m}\sum_{j=1}^t\eta_j\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j,\mathbf{q}_j)\right]-\mathbf{g}(\mathbf{w}_j,\mathbf{q}_j),\mathbf{x}_j-\mathbf{x}\rangle\\ \leq&1+2M^2\sum_{j=1}^t\eta_j^2+8M\sqrt{\sum_{j=1}^t\eta_j^2\left(\ln\frac{2t^2}{\delta}\right)},\quad\forall t\in\mathbb{Z}_+.\end{aligned}\tag{63}$$

To bound the last term E_3, we make use of Theorem 3.6, and prove the following lemma.

Lemma B.4. *Under the condition of Theorem 3.7, we have*

$$\begin{aligned}&\mathrm{E}\left[\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m}\sum_{j=1}^t\eta_j\langle F(\mathbf{w}_j,\mathbf{q}_j)-\mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j,\mathbf{q}_j)\right],\mathbf{x}_j-\mathbf{x}\rangle\right]\\ \leq&\sum_{j=1}^t\eta_j\frac{DG\big[3+\ln j+16(1+\sqrt{\ln m})\sqrt{2(1+\ln j)}\big]}{2(\sqrt{j+1}-1)},\quad\forall t\in\mathbb{Z}_+.\end{aligned}\tag{64}$$

Furthermore, with probability at least $1-\delta$*,*

$$\begin{aligned}&\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m}\sum_{j=1}^t\eta_j\langle F(\mathbf{w}_j,\mathbf{q}_j)-\mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j,\mathbf{q}_j)\right],\mathbf{x}_j-\mathbf{x}\rangle\\ \leq&\sum_{j=1}^t\eta_j\frac{DG\big[3+\ln j+16\sqrt{(1+\ln j)\ln(2mj^2/\delta)}\big]}{2(\sqrt{j+1}-1)},\quad\forall t\in\mathbb{Z}_+.\end{aligned}\tag{65}$$

Combining (59), (61), (62) and (64), we have

$$\begin{aligned}&\mathrm{E}\left[\max_{\mathbf{q}\in\Delta_m}\phi(\bar{\mathbf{w}}_t,\mathbf{q})-\min_{\mathbf{w}\in\mathcal{W}}\phi(\mathbf{w},\bar{\mathbf{q}}_t)\right]\\ \leq&\left(\sum_{j=1}^t\eta_j\right)^{-1}\left(2+\frac{5M^2}{2}\sum_{j=1}^t\eta_j^2+\sum_{j=1}^t\eta_j\frac{DG\big[3+\ln j+16(1+\sqrt{\ln m})\sqrt{2(1+\ln j)}\big]}{2(\sqrt{j+1}-1)}\right).\end{aligned}$$

By setting

$$\eta_j=\frac{1}{M\sqrt{j}},\tag{66}$$

we have

$$\begin{aligned}&\mathrm{E}\left[\max_{\mathbf{q}\in\Delta_m}\phi(\bar{\mathbf{w}}_t,\mathbf{q})-\min_{\mathbf{w}\in\mathcal{W}}\phi(\mathbf{w},\bar{\mathbf{q}}_t)\right]\\ \leq&\left(\sum_{j=1}^t\frac{1}{\sqrt{j}}\right)^{-1}\left(2M+\frac{5M}{2}\sum_{j=1}^t\frac{1}{j}+\sum_{j=1}^t\frac{1}{\sqrt{j}}\frac{DG\big[3+\ln j+16(1+\sqrt{\ln m})\sqrt{2(1+\ln j)}\big]}{2(\sqrt{j+1}-1)}\right).\end{aligned}$$

It is easy to verify that

$$2(\sqrt{j+1}-1)\geq\frac{\sqrt{j}}{2},\ \forall j\in\mathbb{Z}_+.\tag{67}$$

So, we have

$$\begin{aligned}
&\mathrm{E}\left[\max_{\mathbf{q}\in\Delta_m}\phi(\bar{\mathbf{w}}_t,\mathbf{q})-\min_{\mathbf{w}\in\mathcal{W}}\phi(\mathbf{w},\bar{\mathbf{q}}_t)\right]\\
\leq&\left(\sum_{j=1}^t\frac{1}{\sqrt{j}}\right)^{-1}\left(2M+\frac{5M}{2}\sum_{j=1}^t\frac{1}{j}+2DG\left[3+\ln t+16(1+\sqrt{\ln m})\sqrt{2(1+\ln t)}\right]\sum_{j=1}^t\frac{1}{j}\right)\\
\overset{(40)}{\leq}&\frac{1}{2(\sqrt{t+1}-1)}\left(M(5+3\ln t)+2DG\left[3+\ln t+16(1+\sqrt{\ln m})\sqrt{2(1+\ln t)}\right](1+\ln t)\right)
\end{aligned}$$

which proves (46). The setting of step sizes η_t^w and η_t^q in (45) is derived by combining (53) and (66).

From (59) , (61), (63) and (65), with probability at least $1-2\delta$, we have

$$\begin{aligned}
&\max_{\mathbf{q}\in\Delta_m}\phi(\bar{\mathbf{w}}_t,\mathbf{q})-\min_{\mathbf{w}\in\mathcal{W}}\phi(\mathbf{w},\bar{\mathbf{q}}_t)\\
\leq&\left(\sum_{j=1}^t\eta_j\right)^{-1}\left(2+\frac{5M^2}{2}\sum_{j=1}^t\eta_j^2+8M\sqrt{\sum_{j=1}^t\eta_j^2\left(\ln\frac{2t^2}{\delta}\right)}\right.\\
&\left.+\sum_{j=1}^t\eta_j\frac{DG\left[3+\ln j+16\sqrt{(1+\ln j)\ln(2mj^2/\delta)}\right]}{2(\sqrt{j+1}-1)}\right)\\
\overset{(66),(67)}{\leq}&\left(\sum_{j=1}^t\frac{1}{\sqrt{j}}\right)^{-1}\left(2M+\frac{5M}{2}\sum_{j=1}^t\frac{1}{j}+8M\sqrt{\left(\ln\frac{2t^2}{\delta}\right)\sum_{j=1}^t\frac{1}{j}}\right.\\
&\left.+2DG\left(3+\ln t+16\sqrt{(1+\ln t)\ln\frac{2mt^2}{\delta}}\right)\sum_{j=1}^t\frac{1}{j}\right)\\
\overset{(40)}{\leq}&\frac{1}{2(\sqrt{t+1}-1)}\left[M\left(5+3\ln t+8\sqrt{(1+\ln t)\ln\frac{2t^2}{\delta}}\right)\right.\\
&\left.+2DG\left(3+\ln t+16\sqrt{(1+\ln t)\ln\frac{2mt^2}{\delta}}\right)(1+\ln t)\right]
\end{aligned}$$

which proves (47).

B.3. Proof of Theorem 4.1

Besides the high probability bound in (23), we will also establish an expectation bound:

$$\mathrm{E}\left[R_i(\bar{\mathbf{w}}^{(i)})-R_i^*\right]\leq\frac{2DG}{\sqrt{n_i}},\quad\forall i\in[m].\tag{68}$$

We follow the analysis of Theorem 3.6, and use a fixed step size to simplify the results.

Let $t=\frac{n_i}{2}$. From (38), we have

$$\mathrm{E}\left[R_i(\bar{\mathbf{w}}^{(i)})-R_i(\mathbf{w}_*^{(i)})\right]\leq\frac{2D^2+G^2\sum_{j=1}^t(\eta^{(i)})^2}{2\sum_{j=1}^t\eta^{(i)}}=\frac{D^2}{t\eta^{(i)}}+\frac{G^2\eta^{(i)}}{2}=DG\sqrt{\frac{2}{t}}=\frac{2DG}{\sqrt{n_i}}$$

where we set $\eta^{(i)}=\frac{D}{G}\sqrt{\frac{2}{t}}=\frac{2D}{G\sqrt{n_i}}$, which proves (68).

Repeating the proof in Section B.1.2, with probability at least $1-\delta/m$, we have

$$\begin{aligned}R_i(\bar{\mathbf{w}}^{(i)}) - R_i(\mathbf{w}_*^{(i)}) \leq & \frac{D^2 + \frac{G^2}{2}\sum_{j=1}^t (\eta^{(i)})^2 + 8DG\sqrt{\sum_{j=1}^t (\eta^{(i)})^2 \ln\frac{m}{\delta}}}{\sum_{j=1}^t \eta^{(i)}} \\ = & \frac{D^2}{t\eta^{(i)}} + \frac{G^2\eta^{(i)}}{2} + \frac{8DG\sqrt{\ln\frac{m}{\delta}}}{\sqrt{t}} \\ = & DG\sqrt{\frac{2}{t}}\left(1+4\sqrt{2\ln\frac{m}{\delta}}\right) = \frac{2DG}{\sqrt{n_i}}\left(1+4\sqrt{2\ln\frac{m}{\delta}}\right)\end{aligned}$$

for any $i \in [m]$. We obtain (23) by taking the union bound over all $i \in [m]$.

B.4. Proof of Theorem 4.5

Theorem 4.5 is a condensed version of the following theorem and corollary.

Theorem B.5. *Define*

$$\begin{aligned}&p_{\max} = \max_{i\in[m]} p_i, \quad \omega_{\max} = \max_{i\in[m]} \frac{p_i^2 n_m}{n_i}, \quad r_{\max} = \max_{i\in[m]} \frac{p_i}{\sqrt{n_i}} \\ &\widetilde{L} = 2\sqrt{2}p_{\max}(D^2L + D^2G\sqrt{\ln m}), \text{ and } \sigma^2 = 2c\omega_{\max}(\kappa D^2G^2 + \ln^2 m)\end{aligned} \tag{69}$$

where $c > 0$ is an absolute constant. Under Assumptions 3.1, 3.2, 3.3, 3.4, 4.3, and 4.4, and setting

$$\eta^{(i)} = \frac{2D}{G\sqrt{n_i}},\ \eta_w = 2D^2\min\left(\frac{1}{\sqrt{3}\widetilde{L}}, 2\sqrt{\frac{2}{7\sigma^2 n_m}}\right),\ \text{and } \eta_q = 2\min\left(\frac{1}{\sqrt{3}\widetilde{L}}, 2\sqrt{\frac{2}{7\sigma^2 n_m}}\right)\ln m$$

in Algorithm 2, with probability at least $1-2\delta$, we have

$$\begin{aligned}&R_i(\bar{\mathbf{w}}) - R_i^* \leq \frac{1}{p_i}p_\varphi^* \\ &+ \frac{1}{p_i}\left[\frac{14\widetilde{L}}{n_m} + \sqrt{\frac{\sigma^2}{n_m}}\left(\frac{28}{\sqrt{3}} + 7\sqrt{6\log\frac{2}{\delta}} + \frac{28\sqrt{2}}{n_m}\log\frac{2}{\delta}\right) + 4DG\left(1+4\sqrt{2\ln\frac{m}{\delta}}\right)r_{\max}\right]\end{aligned} \tag{70}$$

where p_φ^* is the optimal value of (21).

Corollary B.6. *Under the condition of Theorem B.5 and (32), with high probability, we have*

$$R_i(\bar{\mathbf{w}}) - R_i^* = \frac{1}{p_i}p_\varphi^* + O\left(\left(\frac{1}{n_m} + \frac{1}{\sqrt{n_i}}\right)\ln m\right).$$

B.5. Proof of Theorem B.5

Following the analysis of Zhang et al. (2023, Theorem 4), we bound the optimization error of $(\bar{\mathbf{w}}, \bar{\mathbf{q}})$ for (24) below.

Theorem B.7. *Under the condition of Theorem B.5, with probability at least $1-\delta$, we have*

$$\epsilon_{\widehat{\varphi}}(\bar{\mathbf{w}}, \bar{\mathbf{q}}) \leq \frac{14\widetilde{L}}{n_m} + \sqrt{\frac{\sigma^2}{n_m}}\left(\frac{28}{\sqrt{3}} + 7\sqrt{6\log\frac{2}{\delta}} + \frac{28\sqrt{2}}{n_m}\log\frac{2}{\delta}\right).$$

Then, we make use of Lemma 4.2 to bound the optimization error of $(\bar{\mathbf{w}}, \bar{\mathbf{q}})$ for problem (21). From Theorem 4.1 and Lemma 4.2, with probability at least $1-\delta$, we have

$$\begin{aligned}\epsilon_\varphi(\bar{\mathbf{w}}, \bar{\mathbf{q}}) \leq & \epsilon_{\widehat{\varphi}}(\bar{\mathbf{w}}, \bar{\mathbf{q}}) + 2\max_{i\in[m]}\left\{p_i\left[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*\right]\right\} \\ \leq & \epsilon_{\widehat{\varphi}}(\bar{\mathbf{w}}, \bar{\mathbf{q}}) + 2\max_{i\in[m]}\left\{p_i\frac{2DG}{\sqrt{n_i}}\left(1+4\sqrt{2\ln\frac{m}{\delta}}\right)\right\} \\ = & \epsilon_{\widehat{\varphi}}(\bar{\mathbf{w}}, \bar{\mathbf{q}}) + 4DG\left(1+4\sqrt{2\ln\frac{m}{\delta}}\right)\max_{i\in[m]}\frac{p_i}{\sqrt{n_i}}.\end{aligned}$$

Combining with Theorem B.7, with probability at least $1-2\delta$, we have

$$\begin{aligned}&\epsilon_\varphi(\bar{\mathbf{w}},\bar{\mathbf{q}})\\ \leq&\frac{14\widetilde{L}}{n_m}+\sqrt{\frac{\sigma^2}{n_m}}\left(\frac{28}{\sqrt{3}}+7\sqrt{6\log\frac{2}{\delta}}+\frac{28\sqrt{2}}{n_m}\log\frac{2}{\delta}\right)+4DG\left(1+4\sqrt{2\ln\frac{m}{\delta}}\right)\max_{i\in[m]}\frac{p_i}{\sqrt{n_i}}.\end{aligned}\tag{71}$$

Next, we bound the excess risk of $\bar{\mathbf{w}}$ on every distribution. To this end, we have

$$\begin{aligned}&\max_{i\in[m]}\left\{p_i\left[R_i(\bar{\mathbf{w}})-R_i^*\right]\right\}-\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{q}\in\Delta_m}\varphi(\mathbf{w},\mathbf{q})\\ =&\max_{\mathbf{q}\in\Delta_m}\left\{\sum_{i=1}^m q_ip_i\left[R_i(\bar{\mathbf{w}})-R_i^*\right]\right\}-\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{q}\in\Delta_m}\varphi(\mathbf{w},\mathbf{q})\\ \leq&\max_{\mathbf{q}\in\Delta_m}\left\{\sum_{i=1}^m q_ip_i\left[R_i(\bar{\mathbf{w}})-R_i^*\right]\right\}-\min_{\mathbf{w}\in\mathcal{W}}\left\{\sum_{i=1}^m \bar{q}_ip_i\left[R_i(\mathbf{w})-R_i^*\right]\right\}\\ =&\max_{\mathbf{q}\in\Delta_m}\varphi(\bar{\mathbf{w}},\mathbf{q})-\min_{\mathbf{w}\in\mathcal{W}}\varphi(\mathbf{w},\bar{\mathbf{q}})=\epsilon_\varphi(\bar{\mathbf{w}},\bar{\mathbf{q}}).\end{aligned}$$

Thus, for every distribution $\mathcal{P}_i$, the excess risk can be bounded in the following way:

$$\begin{aligned}&R_i(\bar{\mathbf{w}})-R_i^*\\ \leq&\frac{1}{p_i}\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{q}\in\Delta_m}\varphi(\mathbf{w},\mathbf{q})+\frac{1}{p_i}\epsilon_\varphi(\bar{\mathbf{w}},\bar{\mathbf{q}})\\ \overset{(71)}{\leq}&\frac{1}{p_i}\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{q}\in\Delta_m}\varphi(\mathbf{w},\mathbf{q})\\ &+\frac{1}{p_i}\left[\frac{14\widetilde{L}}{n_m}+\sqrt{\frac{\sigma^2}{n_m}}\left(\frac{28}{\sqrt{3}}+7\sqrt{6\log\frac{2}{\delta}}+\frac{28\sqrt{2}}{n_m}\log\frac{2}{\delta}\right)+4DG\left(1+4\sqrt{2\ln\frac{m}{\delta}}\right)\max_{i\in[m]}\frac{p_i}{\sqrt{n_i}}\right].\end{aligned}$$

which proves (70).

B.6. Proof of Corollary B.6

From (69) and (70) of Zhang et al. (2023), we have

$$\frac{1}{p_i}\frac{\widetilde{L}}{n_m}=O\left(\left(\frac{1}{n_m}+\frac{1}{\sqrt{n_i}}\right)\sqrt{\ln m}\right)\text{ and }\frac{1}{p_i}\sqrt{\frac{\sigma^2}{n_m}}=O\left(\left(\frac{1}{n_m}+\frac{1}{\sqrt{n_i}}\right)\sqrt{\kappa+\ln^2 m}\right).\tag{72}$$

Furthermore,

$$\begin{aligned}\frac{r_{\max}}{p_i}=&\frac{1}{p_i}\max_{i\in[m]}\frac{p_i}{\sqrt{n_i}}=\frac{1}{p_i}\max_{i\in[m]}\left(\frac{1/\sqrt{n_m}+1}{1/\sqrt{n_m}+\sqrt{n_m/n_i}}\frac{1}{\sqrt{n_i}}\right)\\ \leq&\frac{1}{p_i}\max_{i\in[m]}\left(\left(\frac{1}{\sqrt{n_m}}+1\right)\sqrt{\frac{n_i}{n_m}}\frac{1}{\sqrt{n_i}}\right)=\frac{1}{p_i}\left(\frac{1}{\sqrt{n_m}}+1\right)\frac{1}{\sqrt{n_m}}\\ =&\frac{1/\sqrt{n_m}+\sqrt{n_m/n_i}}{1/\sqrt{n_m}+1}\left(\frac{1}{\sqrt{n_m}}+1\right)\frac{1}{\sqrt{n_m}}=\frac{1}{n_m}+\frac{1}{\sqrt{n_i}}.\end{aligned}\tag{73}$$

Combining (70), (72), and (73), we have

$$R_i(\bar{\mathbf{w}})-R_i^*=\frac{1}{p_i}\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{q}\in\Delta_m}\varphi(\mathbf{w},\mathbf{q})+O\left(\left(\frac{1}{n_m}+\frac{1}{\sqrt{n_i}}\right)\sqrt{\kappa+\ln^2 m}\right).$$

B.7. Proof of Corollary 4.7

Under the assumption of this corollary, we have

$$\begin{aligned} 0 \leq & p_\varphi^* = \min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \varphi(\mathbf{w},\mathbf{q}) \\ & \leq \max_{\mathbf{q}\in\Delta_m} \varphi(\mathbf{w}_*,\mathbf{q}) = \max_{\mathbf{q}\in\Delta_m} \left\{ \sum_{i=1}^m q_i p_i \big[R_i(\mathbf{w}_*) - R_i^*\big] \right\} = 0. \end{aligned}$$

B.8. Proof of Theorem B.7

Our analysis is similar to that of Theorem 4 of Zhang et al. (2023). For brevity, we will only highlight the differences.

We first introduce the monotone operator $F(\mathbf{w},\mathbf{q})$ associated with (24):

$$\begin{aligned} F(\mathbf{w},\mathbf{q}) =& [\nabla_{\mathbf{w}}\widehat{\varphi}(\mathbf{w},\mathbf{q}); -\nabla_{\mathbf{q}}\widehat{\varphi}(\mathbf{w},\mathbf{q})] \\ =& \left[\sum_{i=1}^m q_i p_i \nabla R_i(\mathbf{w}); -\Big[p_1[R_1(\mathbf{w}) - R_1(\bar{\mathbf{w}}^{(1)})], \ldots, p_m\big[R_m(\mathbf{w}) - R_m(\bar{\mathbf{w}}^{(m)})\big]\Big]^\top \right]. \end{aligned}$$

It is slightly different from the monotone operator defined by Zhang et al. (2023, (65)), attributed to the inclusion of $R_1(\bar{\mathbf{w}}^{(1)}), \ldots, R_m(\bar{\mathbf{w}}^{(m)})$. However, these additional terms do not alter the continuity of $F(\mathbf{w},\mathbf{q})$. In particular, Lemma 3 of Zhang et al. (2023) remains applicable, leading to the following lemma.

Lemma B.8. *For the monotone operator $F(\mathbf{w},\mathbf{q})$, we have*

$$\|F(\mathbf{w},\mathbf{q}) - F(\mathbf{w}',\mathbf{q}')\|_* \leq \widetilde{L}\big\|(\mathbf{w}-\mathbf{w}',\mathbf{q}-\mathbf{q}')\big\|$$

where $\widetilde{L}$ is defined in (69).

Then, we investigate the stochastic oracle in Algorithm 2:

$$\mathbf{g}(\mathbf{w},\mathbf{q}) = [\mathbf{g}_w(\mathbf{w},\mathbf{q}); -\mathbf{g}_q(\mathbf{w},\mathbf{q})]$$

where

$$\begin{aligned} \mathbf{g}_w(\mathbf{w},\mathbf{q}) &= \sum_{i=1}^m q_i p_i \left(\frac{n_m}{n_i} \sum_{j=1}^{n_i/n_m} \nabla \ell(\mathbf{w};\mathbf{z}^{(i,j)}) \right), \\ \mathbf{g}_q(\mathbf{w},\mathbf{q}) &= \left[p_1 \frac{n_m}{n_1} \sum_{j=1}^{n_1/n_m} \big[\ell(\mathbf{w};\mathbf{z}^{(1,j)}) - \ell(\bar{\mathbf{w}}^{(1)};\mathbf{z}^{(1,j)})\big], \ldots, p_m\big[\ell(\mathbf{w};\mathbf{z}^{(m)}) - \ell(\bar{\mathbf{w}}^{(m)};\mathbf{z}^{(m)})\big] \right]^\top \end{aligned}$$

and $\mathbf{z}^{(i,j)}$ is the j-th sample drawn from distribution $\mathcal{P}_i$. Again, $\mathbf{g}(\mathbf{w},\mathbf{q})$ is different from that of Zhang et al. (2023, (67)), because of the additional terms. However, it is easy to verify that the variance only changes by a constant factor, and Lemma 4 of Zhang et al. (2023) still holds with a different constant.

Lemma B.9. *For the stochastic oracle $\mathbf{g}(\mathbf{w},\mathbf{q})$, we have*

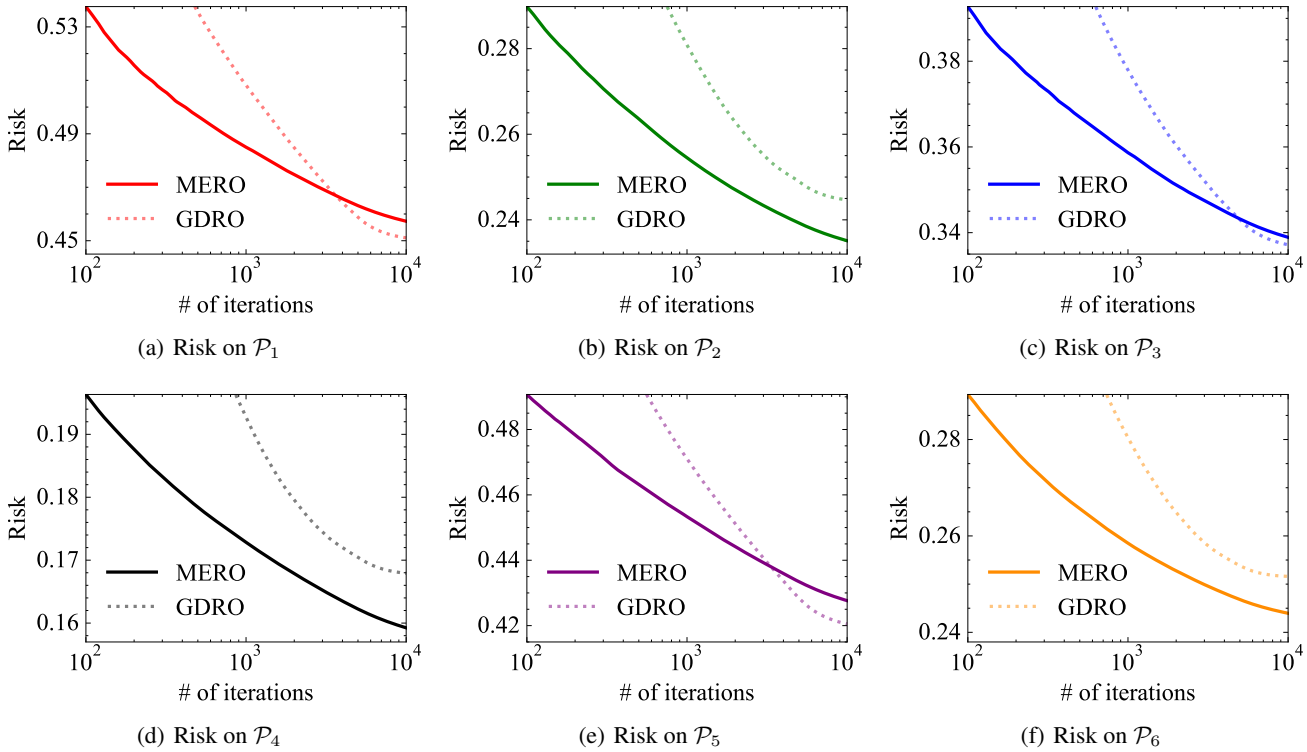
$$\mathrm{E}\left[\exp\left(\frac{\|F(\mathbf{w},\mathbf{q}) - \mathbf{g}(\mathbf{w},\mathbf{q})\|_*^2}{\sigma^2}\right)\right] \leq 2$$

where σ^2 is defined in (69).

The final difference lies in the number of iterations for SMPA, which is $n_m/4$ in our Algorithm 2, and $n_m/2$ in the work of Zhang et al. (2023).

From Corollary 1 of Juditsky et al. (2011), by setting

$$\eta = \min\left(\frac{1}{\sqrt{3}\widetilde{L}}, 2\sqrt{\frac{2}{7\sigma^2 n_m}}\right)$$



(a) Risk on $\mathcal{P}_1$ (b) Risk on $\mathcal{P}_2$ (c) Risk on $\mathcal{P}_3$

(d) Risk on $\mathcal{P}_4$ (e) Risk on $\mathcal{P}_5$ (f) Risk on $\mathcal{P}_6$

Figure 4. Individual risk versus the number of iterations on the Adult dataset.

we have

$$\Pr\left[\epsilon_\varphi(\bar{\mathbf{w}}, \bar{\mathbf{q}}) \geq \frac{14\widetilde{L}}{n_m} + 28\sqrt{\frac{\sigma^2}{3n_m}} + 7\Lambda\sqrt{\frac{2\sigma^2}{n_m}}\right] \leq \exp\left(-\frac{\Lambda^2}{3}\right) + \exp\left(-\frac{\Lambda n_m}{4}\right)$$

for all $\Lambda > 0$. Choosing Λ such that $\exp(-\Lambda^2/3) \leq \delta/2$ and $\exp(-\Lambda n_m/4) \leq \delta/2$, we have with probability at least $1-\delta$

$$\epsilon_\varphi(\bar{\mathbf{w}}, \bar{\mathbf{q}}) \leq \frac{14\widetilde{L}}{n_m} + 28\sqrt{\frac{\sigma^2}{3n_m}} + 7\left(\sqrt{3\log\frac{2}{\delta}} + \frac{4}{n_m}\log\frac{2}{\delta}\right)\sqrt{\frac{2\sigma^2}{n_m}}.$$

C. Full Experiments

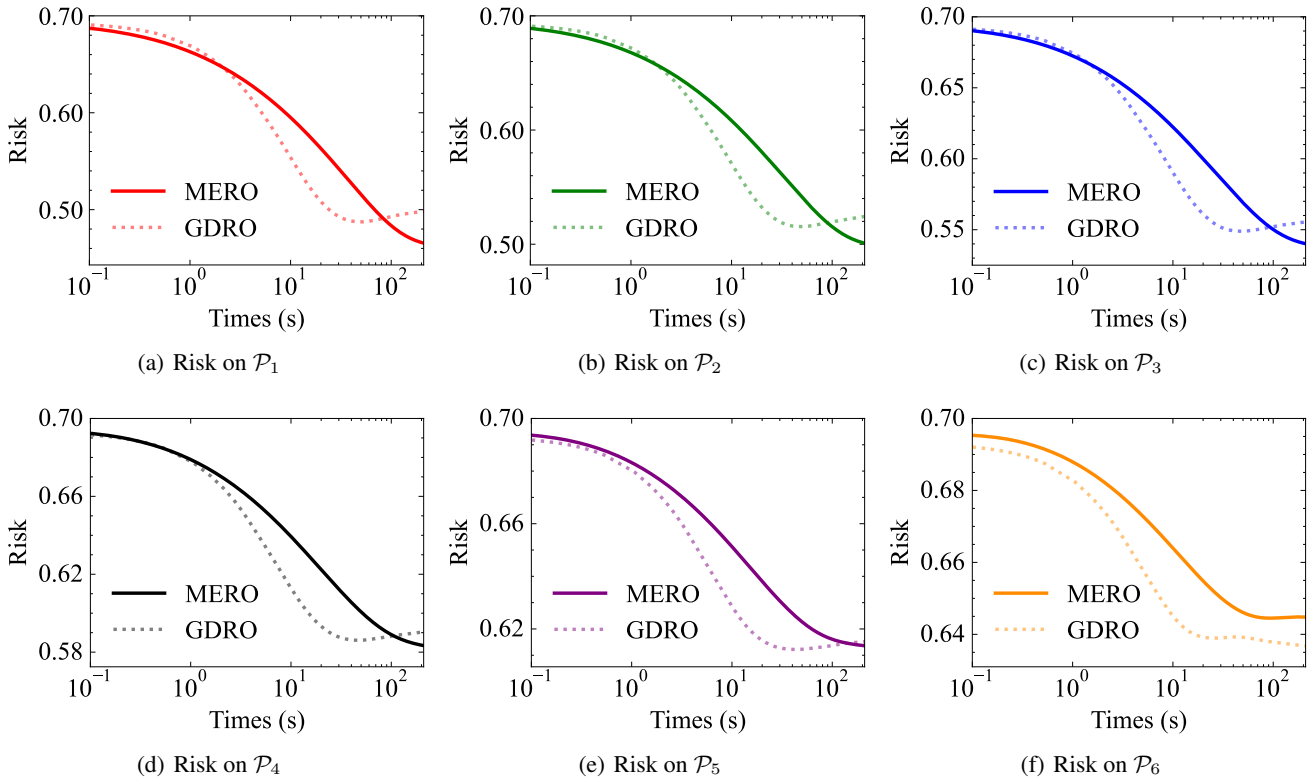
In this section, we conduct empirical studies to evaluate our proposed algorithms.

C.1. Datasets and Experimental Settings

Following previous work (Namkoong & Duchi, 2016; Soma et al., 2022), we employ both synthetic and real-world datasets.

For the synthetic dataset, we construct $m = 6$ distributions, each of which is associated with a true classifier $\mathbf{w}_i^* \in \mathbb{R}^{1000}$. The selection process is as follows: we initially choose an arbitrary $\mathbf{w}_0^*$ on the unit sphere. Subsequently, we randomly pick m points on a sphere with radius d, centered at $\mathbf{w}_0^*$. These points are then projected onto the unit sphere to form the set $\{\mathbf{w}_i^*\}_{i\in[m]}$. We set $d = 0.2$ to keep the classifiers $\{\mathbf{w}_i^*\}_{i\in[m]}$ close, thereby emphasizing the optimization challenges due to the varying noise across the distributions. For each distribution $i \in [m]$, a sample $(\mathbf{x}, y)$ is generated where $\mathbf{x}$ follows a standard normal distribution $\mathcal{N}(0, I)$, and $y = \text{sign}(\mathbf{x}^\top \mathbf{w}_i^*)$ or its inverse, each with respective probabilities $p_i = 1 - 0.05 \times i$ and $1 - p_i$.

We additionally utilize the Adult dataset (Becker & Kohavi, 1996), which encompasses a variety of attributes, including age, gender, race, and educational background, for a total of 48842 individuals. The samples are classified into 6 distinct groups,



(a) Risk on $\mathcal{P}_1$ (b) Risk on $\mathcal{P}_2$ (c) Risk on $\mathcal{P}_3$

(d) Risk on $\mathcal{P}_4$ (e) Risk on $\mathcal{P}_5$ (f) Risk on $\mathcal{P}_6$

Figure 5. Individual risk versus the running time on the synthetic dataset.

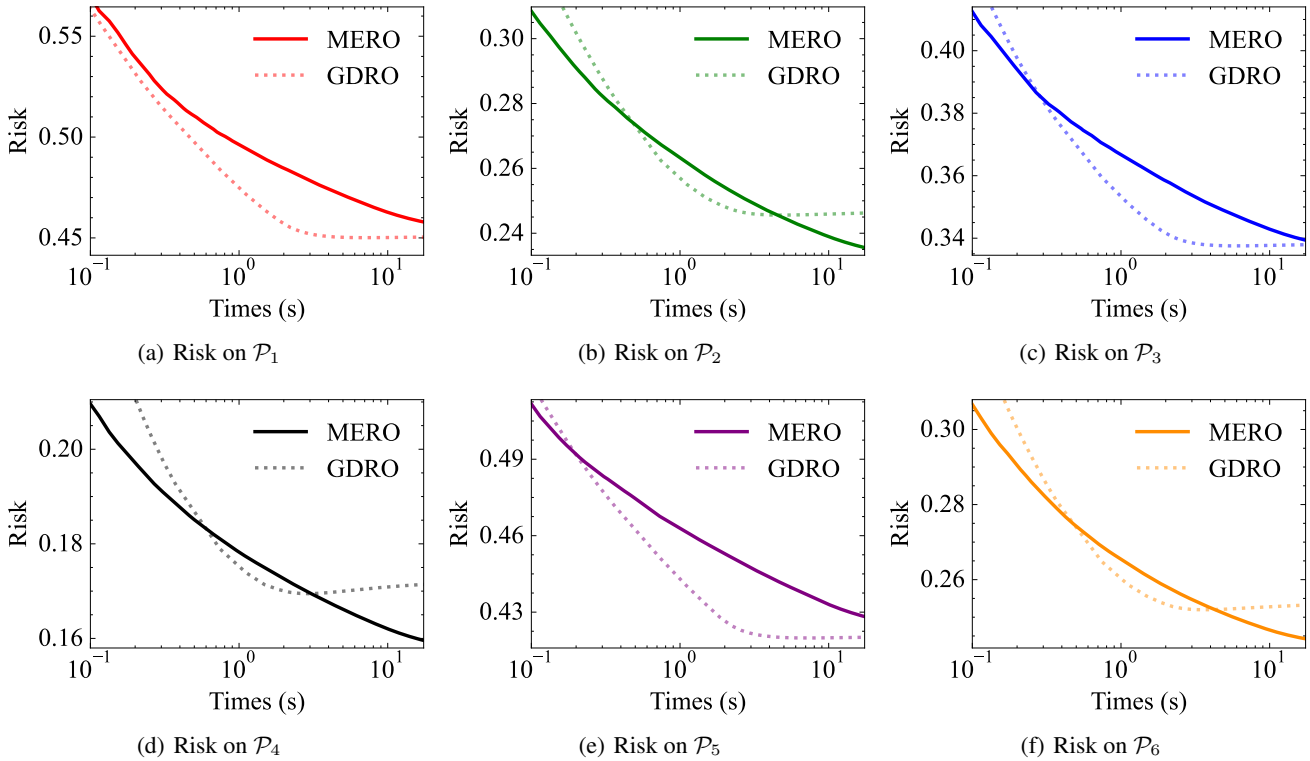
based on a combination of race {black, white, others} and gender {female, male}. By employing one-hot encoding for 12 attributes, 103-dimensional feature vectors are generated to classify whether the income surpasses \$5000.

We designate the logistic loss as our loss function, and utilize various methods to train a linear model. In assessing their performance, it's essential to determine the model's risk for each distribution. To approximate the risk value, we will draw a specific number of samples and use the average risk calculated from these samples as an estimation. The minimal risk R_i^* for each distribution is estimated in a similar way: initially, a model is trained to minimize the empirical risk using a large set of samples; subsequently, the risk is calculated based on a freshly drawn sample set.

C.2. Experiments on Balanced Data

In our experiments with the synthetic dataset, samples will be dynamically generated in real time, adhering to the generation protocol in Section C.1. Regarding the Adult dataset, we designate the distribution $\mathcal{P}_i$ to represent the uniform distribution over the data within the i-th group, and thus the sample generalization process simplifies to randomly selecting samples from each group with replacement.

Besides the results in Section 5, we also compare the running time of MERO and GDRO. Note that the stochastic algorithm for GDRO is more efficient than our Algorithm 1, as it does not require estimating the minimal risk for each distribution. To examine the difference, we plot the risk relative to the running time in Fig. 5 and Fig. 6. We observe that the risk decreases more rapidly for GDRO compared to MERO, primarily because GDRO processes a greater number of samples in the same amount of time. Nevertheless, the ultimate performance of GDRO and MERO is consistent with previous findings presented in Fig. 3 and Fig. 4, with GDRO generally performing better on harder distributions. Additionally, an upward trend in the final curve of GDRO is noted. This is because Algorithm 1 of Zhang et al. (2023), like MS-MERO, is not anytime. As a result, its fixed step size becomes suboptimal as time progresses.



(a) Risk on $\mathcal{P}_1$ (b) Risk on $\mathcal{P}_2$ (c) Risk on $\mathcal{P}_3$

(d) Risk on $\mathcal{P}_4$ (e) Risk on $\mathcal{P}_5$ (f) Risk on $\mathcal{P}_6$

Figure 6. Individual risk versus the running time on the Adult dataset.

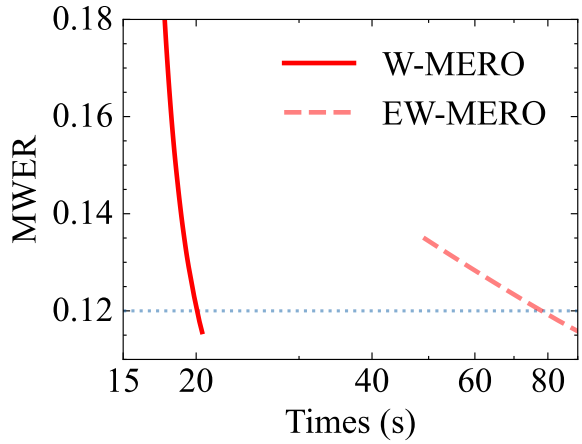


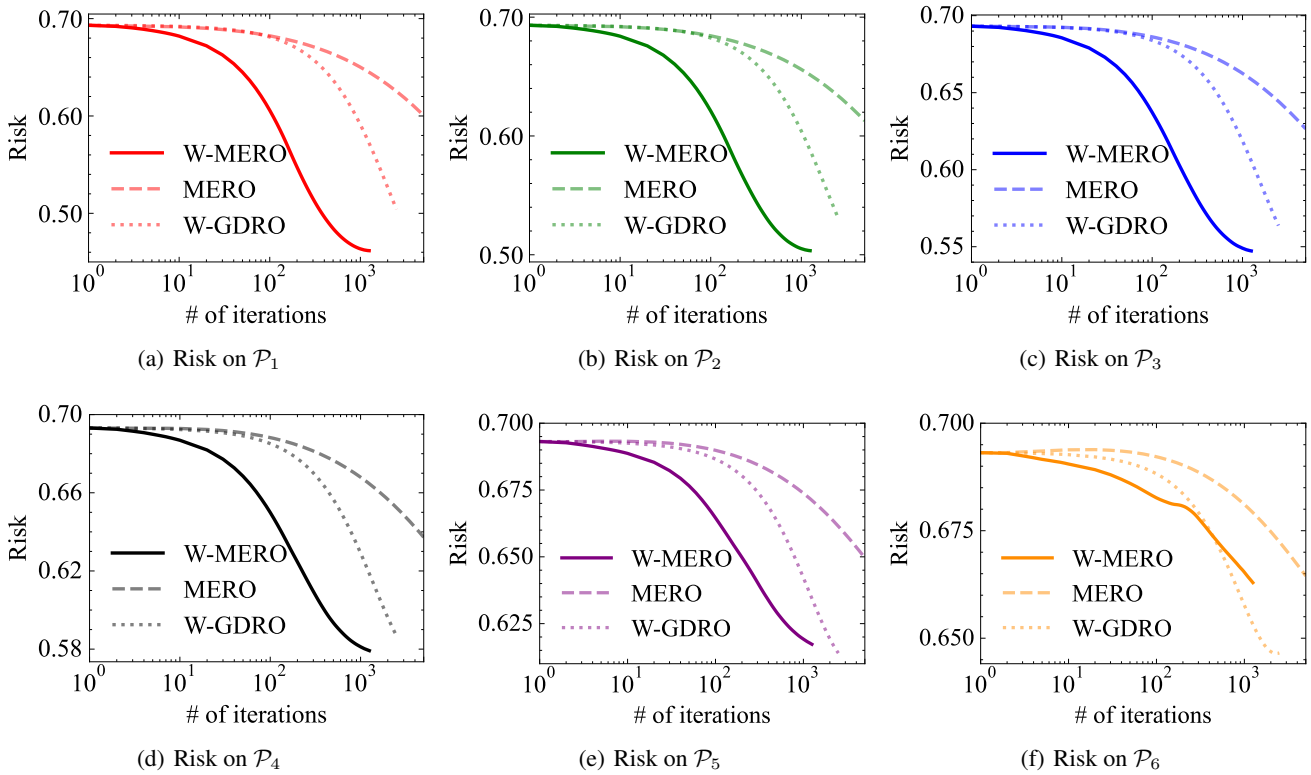
Figure 7. The MWER versus the running time on the *imbalanced* synthetic dataset.

Table 2. Running times of W-MERO and EW-MERO on the *imbalanced* synthetic dataset.

Algorithm	MWER Value	Times (s)
W-MERO	0.12	20.3
EW-MERO	0.12	78.5

C.3. Experiments on Imbalanced Data

For experiments involving imbalanced data, we fix the number of samples per distribution, with the quantity varying among distributions. Mirroring the setup of Zhang et al. (2023), we designate the sample size for the synthetic dataset as $n_i = 5000 \times (7 - i)$, generating each sample as before. Pertaining to the Adult dataset, we randomly extract 364 samples from each group, reserving them for subsequent risk assessment. The remaining number of samples across the 6 groups is $\{26656, 11519, 1780, 1720, 999, 364\}$. Furthermore, each sample within the groups is processed only *once* to simulate the imbalanced scenario. In this way, $\mathcal{P}_i$ corresponds to the (unknown) underlying distribution from which the samples in the i-th group are drawn.



(a) Risk on $\mathcal{P}_1$ (b) Risk on $\mathcal{P}_2$ (c) Risk on $\mathcal{P}_3$

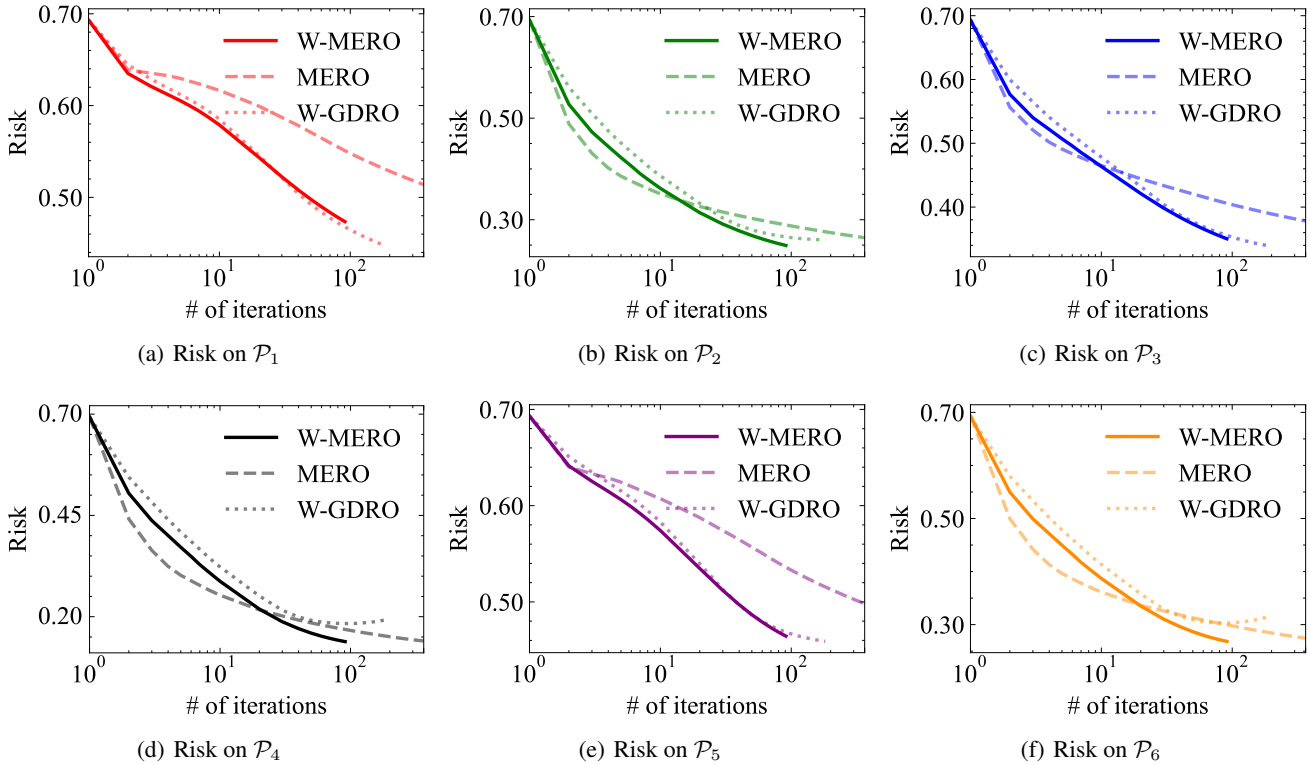
(d) Risk on $\mathcal{P}_4$ (e) Risk on $\mathcal{P}_5$ (f) Risk on $\mathcal{P}_6$

Figure 8. Individual risk versus the number of iterations on the *imbalanced* synthetic dataset.

Note that the optimization procedure of Agarwal & Zhang (2022) is also applicable for minimizing the empirical counterpart of weighted MERO. Therefore, we initially conduct a comparison between our two-stage approach in Section 4.2, and their method. For ease of reference, we label our algorithm and theirs as W-MERO and EW-MERO, respectively, to emphasize that the former is designed for weighted MERO, while the latter focuses on the empirical variant. In Fig. 7, we illustrate how the maximal weighted excess risk (MWER), denoted as $\max_{i\in[m]}\{p_i[R_i(\mathbf{w}) - R_i^*]\}$, changes with respect to the running time. Acknowledging that both algorithms have an initialization phase, so their curves do not start from zero. Consistent with previous experiments in Fig. 1, our W-MERO converges more rapidly than EW-MERO.[2] For a detailed view, the precise times required for W-MERO and EW-MERO to reach a certain MWER value are listed in Table 2.

Next, we examine the effectiveness of our W-MERO in handling imbalanced scenarios. To this end, we compare it with the original MERO—running Algorithm 1 for n_m iterations. Recall that W-MERO reduces the excess risk of the i-th distribution at an $O((\log m)/\sqrt{n_i})$ rate, and MERO attains an $\widetilde{O}(\sqrt{(\log m)/n_m})$ rate for all distributions. Additionally, we include the result of weighted GDRO (W-GDRO) (Zhang et al., 2023, Algorithm 4) to reiterate the distinction between risk minimization and excess risk minimization. Experimental results on the synthetic dataset are provided in Fig. 8, and consistent with our theoretical expectations. Specifically, the final risk of W-MERO closely approaches that of MERO on distribution $\mathcal{P}_6$, which holds the smallest sample size, and W-MERO surpasses MERO on all other distributions. Moreover, the larger the number of samples is, the more pronounced the advantage of W-MERO becomes. In line with the experiments in Fig. 3, W-GDRO performs well on the last two distributions, characterized by their significantly high noise levels. We present the outcomes on the Adult dataset in Fig. 9, and observe analogous patterns.

[2]In this experiment, the imbalanced Adult dataset is omitted. This is because the limited number of samples prevents an accurate estimation of the value of R_i^*.



(a) Risk on $\mathcal{P}_1$ (b) Risk on $\mathcal{P}_2$ (c) Risk on $\mathcal{P}_3$

(d) Risk on $\mathcal{P}_4$ (e) Risk on $\mathcal{P}_5$ (f) Risk on $\mathcal{P}_6$

Figure 9. Individual risk versus the number of iterations on the *imbalanced* Adult dataset.

D. Supporting Lemmas

D.1. Proof of Lemma 4.2

From the definition of $\varphi(\cdot,\cdot)$ and $\widehat{\varphi}(\cdot,\cdot)$, for any $\mathbf{q}$, we have

$$\operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \varphi(\mathbf{w},\mathbf{q}) = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \widehat{\varphi}(\mathbf{w},\mathbf{q}) \tag{74}$$

and for any $(\mathbf{w},\mathbf{q})$

$$\left|\varphi(\mathbf{w},\mathbf{q}) - \widehat{\varphi}(\mathbf{w},\mathbf{q})\right| = \left|\sum_{i=1}^m q_i p_i \left[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*\right]\right| \leq \max_{i\in[m]} \left\{p_i\left[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*\right]\right\}. \tag{75}$$

Let $\widehat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \varphi(\mathbf{w},\bar{\mathbf{q}})$ and $\widehat{\mathbf{q}} = \operatorname{argmax}_{\mathbf{q}\in\Delta_m} \varphi(\bar{\mathbf{w}},\mathbf{q})$. Then, we have

$$\begin{aligned}
\epsilon_{\varphi}(\bar{\mathbf{w}},\bar{\mathbf{q}}) &= \max_{\mathbf{q}\in\Delta_m} \varphi(\bar{\mathbf{w}},\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \varphi(\mathbf{w},\bar{\mathbf{q}}) = \varphi(\bar{\mathbf{w}},\widehat{\mathbf{q}}) - \varphi(\widehat{\mathbf{w}},\bar{\mathbf{q}}) \\
&\overset{(75)}{\leq} \widehat{\varphi}(\bar{\mathbf{w}},\widehat{\mathbf{q}}) - \widehat{\varphi}(\widehat{\mathbf{w}},\bar{\mathbf{q}}) + 2\max_{i\in[m]} \left\{p_i\left[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*\right]\right\} \\
&\leq \max_{\mathbf{q}\in\Delta_m} \widehat{\varphi}(\bar{\mathbf{w}},\mathbf{q}) - \widehat{\varphi}(\widehat{\mathbf{w}},\bar{\mathbf{q}}) + 2\max_{i\in[m]} \left\{p_i\left[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*\right]\right\} \\
&\overset{(74)}{=} \max_{\mathbf{q}\in\Delta_m} \widehat{\varphi}(\bar{\mathbf{w}},\mathbf{q}) - \min_{\mathbf{w}\in\mathcal{W}} \widehat{\varphi}(\mathbf{w},\bar{\mathbf{q}}) + 2\max_{i\in[m]} \left\{p_i\left[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*\right]\right\} \\
&= \epsilon_{\widehat{\varphi}}(\bar{\mathbf{w}},\bar{\mathbf{q}}) + 2\max_{i\in[m]} \left\{p_i\left[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*\right]\right\}.
\end{aligned}$$

D.2. Proof of Lemma B.3

We create a virtual sequence by performs SMD with $\mathrm{E}_{t-1}\left[\mathbf{g}(\mathbf{w}_t, \mathbf{q}_t)\right] - \mathbf{g}(\mathbf{w}_t, \mathbf{q}_t)$ as the gradient:

$$\mathbf{y}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \left\{\eta_j\big\langle \mathrm{E}_{t-1}[\mathbf{g}(\mathbf{w}_t, \mathbf{q}_t)] - \mathbf{g}(\mathbf{w}_t, \mathbf{q}_t), \mathbf{x} - \mathbf{y}_t\big\rangle + B(\mathbf{x}, \mathbf{y}_t)\right\} \tag{76}$$

where $\mathbf{y}_1 = \mathbf{x}_1$. Then, we further decompose the error term as

$$\begin{aligned}
&\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \sum_{j=1}^{t} \eta_j\big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\big\rangle \\
\leq &\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \underbrace{\sum_{j=1}^{t} \eta_j\big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{y}_j - \mathbf{x}\big\rangle}_{:=A} \\
&+ \underbrace{\sum_{j=1}^{t} \eta_j\big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{y}_j\big\rangle}_{:=B}.
\end{aligned} \tag{77}$$

To bound term A, we repeat the analysis of (60), and have

$$\begin{aligned}
&\eta_j\big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{y}_j - \mathbf{x}\big\rangle \\
\leq &B(\mathbf{x}, \mathbf{y}_j) - B(\mathbf{x}, \mathbf{y}_{j+1}) + \frac{\eta_j^2}{2}\left\|\mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right\|_*^2 \\
\leq &B(\mathbf{x}, \mathbf{y}_j) - B(\mathbf{x}, \mathbf{y}_{j+1}) + 2M^2\eta_j^2
\end{aligned} \tag{78}$$

where the last step makes use of the following inequality

$$\begin{aligned}
&\left\|\mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right\|_* \leq \left\|\mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right]\right\|_* + \|\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\|_* \\
\leq &\mathrm{E}_{j-1}\left[\|\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\|_*\right] + \|\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\|_* \overset{(56)}{\leq} 2M.
\end{aligned} \tag{79}$$

Summing (78) over $j = 1, \ldots, t$, we have

$$\begin{aligned}
A = &\sum_{j=1}^{t} \eta_j\big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{y}_j - \mathbf{x}\big\rangle \\
\leq &B(\mathbf{x}, \mathbf{y}_1) + 2M^2\sum_{j=1}^{t}\eta_j^2 \overset{(51)}{\leq} 1 + 2M^2\sum_{j=1}^{t}\eta_j^2.
\end{aligned} \tag{80}$$

To bound term B in (77), we define

$$\delta_j = \eta_j\big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{y}_j\big\rangle.$$

As $\mathbf{x}_j$ and $\mathbf{y}_j$ are independent of the random samples $\mathbf{z}_j^{(1)}, \ldots, \mathbf{z}_j^{(m)}$ used to construct $\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)$ in (55), $\delta_1, \ldots, \delta_t$ forms a martingale difference sequence. Consequently, we have

$$\mathrm{E}[B] = \mathrm{E}\left[\sum_{j=1}^{t}\delta_j\right] = 0. \tag{81}$$

Taking expectations over both sides of (77), we have

$$\begin{aligned}
&\mathrm{E}\left[\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} \sum_{j=1}^{t} \eta_j\big\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{x}\big\rangle\right] \\
\leq &\mathrm{E}\left[\max_{\mathbf{x}\in\mathcal{W}\times\Delta_m} A\right] + \mathrm{E}[B] \overset{(80),(81)}{\leq} 1 + 2M^2\sum_{j=1}^{t}\eta_j^2
\end{aligned}$$

which proves (62).

To establish a high probability bound, we follow the analysis in Section B.1.2 and utilize Lemma B.1 to bound B. To this end, we first show that $|\delta_j|$ is bounded:

$$\begin{aligned}|\delta_j| =& \left|\eta_j \left\langle \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j), \mathbf{x}_j - \mathbf{y}_j \right\rangle\right| \\ \leq& \eta_j \left\|\mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right] - \mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right\|_* \|\mathbf{x}_j - \mathbf{y}_j\| \\ \overset{(79)}{\leq}& 2M\eta_j \left(\|\mathbf{x}_j - \mathbf{x}_1\| + \|\mathbf{x}_1 - \mathbf{y}_j\|\right) \\ \leq& 2M\eta_j \left(\sqrt{2B(\mathbf{x}_j, \mathbf{x}_1)} + \sqrt{2B(\mathbf{y}_j, \mathbf{x}_1)}\right) \overset{(51)}{\leq} 4\sqrt{2}M\eta_j.\end{aligned}$$

From Lemma B.1 and the union bound, with probability at least $1-\delta$, we have

$$B = \sum_{j=1}^{t} \delta_j \leq 8M \sqrt{\sum_{j=1}^{t} \eta_j^2 \left(\ln \frac{2t^2}{\delta}\right)}, \quad \forall t \in \mathbb{Z}_+. \tag{82}$$

We obtain (63) by substituting (80) and (82) into (77).

D.3. Proof of Lemma B.4

First, we have

$$\begin{aligned}&\left\langle F(\mathbf{w}_j, \mathbf{q}_j) - \mathrm{E}_{t-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right], \mathbf{x}_j - \mathbf{x}\right\rangle \\ \overset{(57)}{=}& \left\langle \left[0, -\left[R_1(\bar{\mathbf{w}}_j^{(1)}) - R_1^*, \ldots, R_m(\bar{\mathbf{w}}_j^{(m)}) - R_m^*\right]^\top\right], \mathbf{x}_j - \mathbf{x}\right\rangle \\ =& -\left\langle \left[R_1(\bar{\mathbf{w}}_j^{(1)}) - R_1^*, \ldots, R_m(\bar{\mathbf{w}}_j^{(m)}) - R_m^*\right], \mathbf{q}_j - \mathbf{q}\right\rangle \\ \leq& \left\|\left[R_1(\bar{\mathbf{w}}_j^{(1)}) - R_1^*, \ldots, R_m(\bar{\mathbf{w}}_j^{(m)}) - R_m^*\right]\right\|_\infty \|\mathbf{q}_j - \mathbf{q}\|_1 \\ \leq& 2\left\|\left[R_1(\bar{\mathbf{w}}_j^{(1)}) - R_1^*, \ldots, R_m(\bar{\mathbf{w}}_j^{(m)}) - R_m^*\right]\right\|_\infty\end{aligned}$$

which implies

$$\begin{aligned}&\max_{\mathbf{x} \in \mathcal{W} \times \Delta_m} \sum_{j=1}^{t} \eta_j \left\langle F(\mathbf{w}_j, \mathbf{q}_j) - \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right], \mathbf{x}_j - \mathbf{x}\right\rangle \\ \leq& \sum_{j=1}^{t} 2\eta_j \left\|\left[R_1(\bar{\mathbf{w}}_j^{(1)}) - R_1^*, \ldots, R_m(\bar{\mathbf{w}}_j^{(m)}) - R_m^*\right]\right\|_\infty.\end{aligned} \tag{83}$$

From (19) in Theorem 3.6, we know that with probability at least $1-\delta$,

$$\left\|\left[R_1(\bar{\mathbf{w}}_j^{(1)}) - R_1^*, \ldots, R_m(\bar{\mathbf{w}}_j^{(m)}) - R_m^*\right]\right\|_\infty \leq \frac{DG\left[3 + \ln j + 16\sqrt{(1+\ln j)\ln(2mj^2/\delta)}\right]}{4(\sqrt{j+1}-1)} \tag{84}$$

for all $j \in \mathbb{Z}_+$. We obtain (65) by substituting (84) into (83).

The proof of the expectation bound is more involved. Taking expectations over (83), we have

$$\begin{aligned}&\mathrm{E}\left[\max_{\mathbf{x} \in \mathcal{W} \times \Delta_m} \sum_{j=1}^{t} \eta_j \left\langle F(\mathbf{w}_j, \mathbf{q}_j) - \mathrm{E}_{j-1}\left[\mathbf{g}(\mathbf{w}_j, \mathbf{q}_j)\right], \mathbf{x}_j - \mathbf{x}\right\rangle\right] \\ \leq& \sum_{j=1}^{t} 2\eta_j \mathrm{E}\left[\left\|\left[R_1(\bar{\mathbf{w}}_j^{(1)}) - R_1^*, \ldots, R_m(\bar{\mathbf{w}}_j^{(m)}) - R_m^*\right]\right\|_\infty\right].\end{aligned} \tag{85}$$

Then, one may attempt to make use of the expectation bound (33) in Theorem 3.6. However, due to the presence of the infinity norm in (85), it is difficult to obtain a tight upper bound. As an alternative, we will exploit the fact that $R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i^*$ is sub-Gaussian (Vershynin, 2018), for all $i \in [m]$, $t \in \mathbb{Z}_+$.

Recall the martingale difference sequence $\delta_1^{(i)}, \ldots, \delta_t^{(i)}$ in (37). From (43) and Lemma B.1, we have

$$\Pr\left[\sum_{j=1}^{t} \delta_j^{(i)} > x\right] \leq \exp\left(-\frac{x^2}{64D^2G^2\sum_{j=1}^{t}(\eta_j^{(i)})^2}\right). \tag{86}$$

From (36), we have

$$\begin{aligned}
&\Pr\left[\left(R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)})\right)\left(\sum_{j=1}^{t}\eta_j^{(i)}\right) - \left(D^2 + \frac{G^2}{2}\sum_{j=1}^{t}(\eta_j^{(i)})^2\right) > x\right] \\
\leq &\Pr\left[\sum_{j=1}^{t}\delta_j^{(i)} > x\right] \overset{(86)}{\leq} \exp\left(-\frac{x^2}{64D^2G^2\sum_{j=1}^{t}(\eta_j^{(i)})^2}\right)
\end{aligned}$$

which implies

$$\begin{aligned}
&\Pr\left[R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)}) - \frac{1}{\sum_{j=1}^{t}\eta_j^{(i)}}\left(D^2 + \frac{G^2}{2}\sum_{j=1}^{t}(\eta_j^{(i)})^2\right) > x\right] \\
\leq &\exp\left(-\frac{x^2\left(\sum_{j=1}^{t}\eta_j^{(i)}\right)^2}{64D^2G^2\sum_{j=1}^{t}(\eta_j^{(i)})^2}\right).
\end{aligned}$$

Since $\eta_j^{(1)} = \cdots = \eta_j^{(m)}$, we can invoke the following lemma to bound the expectation of $\max_{i\in[m]}\{R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)})\}$.

Lemma D.1. *Suppose there are m non-negative random variables X_i such that*

$$\Pr\left[X_i \geq \mu + t\right] \leq \exp\left(-\frac{t^2}{\sigma^2}\right)$$

for all $i \in [m]$. Then, we have

$$\mathrm{E}\left[\max_{i\in[m]} X_i\right] \leq \mu + \sigma\sqrt{2\ln m} + \frac{\sigma\sqrt{2\pi}}{2}.$$

From the above lemma, we have

$$\begin{aligned}
&\mathrm{E}\left[\max_{i\in[m]}\left\{R_i(\bar{\mathbf{w}}_t^{(i)}) - R_i(\mathbf{w}_*^{(i)})\right\}\right] \\
\leq &\frac{1}{\sum_{j=1}^{t}\eta_j^{(i)}}\left(D^2 + \frac{G^2}{2}\sum_{j=1}^{t}(\eta_j^{(i)})^2\right) + \left(\sqrt{2\ln m} + \frac{\sqrt{2\pi}}{2}\right)\sqrt{\frac{64D^2G^2\sum_{j=1}^{t}(\eta_j^{(i)})^2}{\left(\sum_{j=1}^{t}\eta_j^{(i)}\right)^2}} \\
= &\frac{2D^2 + G^2\sum_{j=1}^{t}(\eta_j^{(i)})^2 + 16DG(\sqrt{\ln m} + \sqrt{\pi}/2)\sqrt{2\sum_{j=1}^{t}(\eta_j^{(i)})^2}}{2\sum_{j=1}^{t}\eta_j^{(i)}} \\
= &\frac{2DG + DG\sum_{j=1}^{t}\frac{1}{j} + 16DG(\sqrt{\ln m} + \sqrt{\pi}/2)\sqrt{2\sum_{j=1}^{t}\frac{1}{j}}}{2\sum_{j=1}^{t}\frac{1}{\sqrt{j}}} \\
\overset{(40)}{\leq} &\frac{DG\left[3 + \ln t + 16(\sqrt{\ln m} + \sqrt{\pi}/2)\sqrt{2(1+\ln t)}\right]}{4(\sqrt{t+1} - 1)}.
\end{aligned} \tag{87}$$

We obtain (64) by substituting (87) into (85) and noticing $\sqrt{\pi}/2 \leq 1$.

D.4. Proof of Lemma D.1

First, we have

$$\Pr\left[\max_{i\in[m]} X_i \geq \mu + t\right] \leq \sum_{i=1}^{m} \Pr\left[X_i \geq \mu + t\right] \leq m \exp\left(-\frac{t^2}{\sigma^2}\right).$$

To simplify the notation, we define $X = \max_{i\in[m]} X_i$. Since X is non-negative, we have

$$\begin{aligned}
\mathrm{E}\left[X\right] \leq & \int_0^\infty \Pr\left[X \geq x\right] dx = \int_0^{\mu+\sigma\sqrt{2\ln m}} \Pr\left[X \geq x\right] dx + \int_{\mu+\sigma\sqrt{2\ln m}}^\infty \Pr\left[X \geq x\right] dx \\
\leq & \mu + \sigma\sqrt{2\ln m} + \int_{\mu+\sigma\sqrt{2\ln m}}^\infty m \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) dx \\
= & \mu + \sigma\sqrt{2\ln m} + \int_{\mu+\sigma\sqrt{2\ln m}}^\infty m \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx.
\end{aligned}$$

When $x \geq \mu + \sigma\sqrt{2\ln m}$, we have

$$m \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \leq 1.$$

Thus

$$\begin{aligned}
\mathrm{E}\left[X\right] \leq & \mu + \sigma\sqrt{2\ln m} + \int_{\mu+\sigma\sqrt{2\ln m}}^\infty \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
\leq & \mu + \sigma\sqrt{2\ln m} + \int_{\mu}^\infty \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \mu + \sigma\sqrt{2\ln m} + \frac{\sigma\sqrt{2\pi}}{2}.
\end{aligned}$$