Differentially Private Worst-group Risk Minimization

Xinyu Zhou [1] **Raef Bassily** [2]

Abstract

We initiate a systematic study of worst-group risk minimization under (ϵ, δ)-differential privacy (DP). The goal is to privately find a model that approximately minimizes the maximal risk across p sub-populations (groups) with different distributions, where each group distribution is accessed via a sample oracle. We first present a new algorithm that achieves excess worst-group population risk of $\tilde{O}(\frac{p\sqrt{d}}{K\epsilon} + \sqrt{\frac{p}{K}})$, where K is the total number of samples drawn from all groups and d is the problem dimension. Our rate is nearly optimal when each distribution is observed via a fixed-size dataset of size K/p. Our result is based on a new stability-based analysis for the generalization error. In particular, we show that Δ-uniform argument stability implies $\tilde{O}(\Delta + \frac{1}{\sqrt{n}})$ generalization error w.r.t. the worst-group risk, where n is the number of samples drawn from each sample oracle. Next, we propose an algorithmic framework for worst-group population risk minimization using any DP online convex optimization algorithm as a subroutine. Hence, we give another excess risk bound of $\tilde{O}\left(\sqrt{\frac{d^{1/2}}{\epsilon K}} + \sqrt{\frac{p}{K\epsilon^2}} + \sqrt{\frac{p}{K}}\right)$. Assuming the typical setting of $\epsilon = \Theta(1)$, this bound is more favorable than our first bound in a certain range of p as a function of K and d. Finally, we study differentially private worst-group *empirical* risk minimization in the offline setting, where each group distribution is observed by a fixed-size dataset. We present a new algorithm with nearly optimal excess risk of $\tilde{O}(\frac{p\sqrt{d}}{K\epsilon})$.

[1]Department of Computer Science & Engineering, The Ohio State University [2]Department of Computer Science & Engineering and the Translational Data Analytics Institute (TDAI), The Ohio State University. Correspondence to: Xinyu Zhou <zhou.3542@buckeyemail.osu.edu>.

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

1. Introduction

Multi-distribution learning has gained increasing attention due to its close connection to robustness and fairness in machine learning. Consider p groups (or, sub-populations), where each group is associated with some unknown data distribution. In the multi-distribution learning paradigm, the learner tries to find a model that minimizes the maximal population risk across all groups. Let D_i denote distribution of group $i \in [p]$ and $\ell : \mathcal{W} \times \mathcal{Z}$ be a loss function over the parameter space $\mathcal{W}$ and data domain $\mathcal{Z}$, our objective of minimizing the worst-group population risk can be formulated as a minimax stochastic optimization problem written as

$$\min_{w\in\mathcal{W}} \max_{i\in[p]} \{L_{D_i}(w) \coloneqq \mathbb{E}_{z\sim D_i}\ell(w,z)\} \tag{1}$$

When each group distribution D_i is observed by a dataset S_i, we also define the worst-group empirical risk minimization problem as

$$\min_{w\in\mathcal{W}} \max_{i\in[p]} \left\{ L_{S_i}(w) \coloneqq \frac{1}{|S_i|} \sum_{z\sim S_i} \ell(w,z) \right\} \tag{2}$$

The worst-group risk minimization problems (1) and (2) have wide applications in a variety of learning scenarios. In the regime of robust learning, the objective represents a class of problems named as group distributionally robust optimization (Group DRO) (Soma et al., 2022; Zhang et al., 2023; Sagawa et al., 2019). This formulation also captures the mismatch between distributions from different domains (or, sub-populations). In this type of scenarios, instead of assuming a fixed target distribution, this formulation can be used to find a model that can work well for any possible target distribution formed as a mixture of the distributions from different domains/sub-populations. From the perspective of learning with fairness, each group may represent a protected class or a demographic group. Minimizing the worst-group risk leads to the notions of good-intent fairness (Mohri et al., 2019) or min-max group fairness (Abernethy et al., 2022; Diana et al., 2021; Papadaki et al., 2022). By making sure the worst-off group performs as good as possible, the objective prevents the learner from overfitting to certain groups at the cost of others. Moreover, the objectives are also connected to other learning applications such as collaborative learning (Blum et al., 2017) and agnostic fed-

erated learning (Mohri et al., 2019) where the output model is optimized to perform well across multiple distributions.

Meanwhile, machine learning algorithms typically depend on large volumes of data, which may pose serious privacy risk, particularly in certain privacy-sensitive applications like healthcare and finance. Therefore, it is important that we provide a strong and provable privacy guarantee for such algorithms to ensure that the sensitive information always remains private during the learning process. Although there has been a significant progress in understanding worst-group risk minimization in the non-private setting (Soma et al., 2022; Haghtalab et al., 2022; Zhang et al., 2023), this problem has not been formally studied in the context of learning with provable privacy guarantees.

Motivated by this, in this paper, we initiate a formal study of worst-group risk minimization under (ϵ, δ)-differential privacy (DP). We consider the setting where the loss function is convex and Lipschitz over a compact parameter space $\mathcal{W} \subset \mathbb{R}^d$. For the sample access model, each group distribution is accessed via a sample oracle and the learner is able to sample new data points from any group distribution via its sample oracle as needed during the learning process.

1.1. Contribution

As we mentioned earlier, we let p denote the number of groups , K denote the total number of samples drawn from all groups and d denote the problem dimension.

We first provide a new algorithm adapted from the phased-ERM approach in (Feldman et al., 2020) with the excess worst-group population risk of $\tilde{O}\left(\sqrt{\frac{p}{K}} + \frac{p\sqrt{d}}{K\epsilon}\right)$. The first term in this bound matches the optimal non-private bound (Soma et al., 2022) and the second term represents the cost of privacy. Our rate is optimal up to logarithmic factors in the offline setting where each distribution is observed via a a fixed-size dataset of size K/p. The utility guarantee of our algorithm relies on a new stability based argument to establish the generalization error bound. Roughly speaking, we show that Δ-*uniform argument stability* (see Definition 1) with respect to the parameter w implies a $\tilde{O}(\Delta + 1/\sqrt{n})$ generalization error where n is the number of samples drawn from each sample oracle. This is distinct from existing generalization results based on uniform convergence (Mohri et al., 2019; Abernethy et al., 2022), which lead to suboptimal rates for general convex losses. On the other hand, our result also circumvents a known bottleneck in deriving generalization guarantees using argument uniform stability in the stochastic saddle-point problems. In particular, in the L_2/L_2 setting, the best known generalization guarantee for a Δ-uniform argument stable algorithm is $\sqrt{\Delta}$ generalization error based on the analysis of the strong duality gap (Ozdaglar et al., 2022).

Our algorithm is akin to the Phased ERM approach of (Feldman et al., 2020), which was introduced to solve private stochastic convex optimization. We repurpose this approach for the private stochastic minimax optimization. In particular, we define a sequence of regularized minimax problems, which are solved iteratively using any generic non-private solver. Our version of this algorithm requires different tuning of parameters across iterations and, more crucially, a new analysis that utilizes the uniform argument stability of the solutions of the regularized minimax problems.

As a side product, our stability result naturally leads to a new non-private regularization based method which matches the non-private lower bound in (Soma et al., 2022) up to logarithmic factors. This may be of independent interest because existing non-private methods with (nearly) optimal rate rely on single-pass first order method rather than regularization.

Next, we give a framework to minimize the worst-group population risk using a black-box access to any DP online convex optimization (DP OCO) algorithm as a subroutine. Our framework leverages the idea in (Soma et al., 2022; Haghtalab et al., 2022) on casting the worst-group risk minimization objective in a zero-sum game and decomposing the excess worst-group risk into the expected regret of both players. By instantiating our framework with the DP-FTRL algorithm from (Kairouz et al., 2021), we obtain another bound on the excess worst-group population risk of $\tilde{O}\left(\sqrt{\frac{d^{1/2}}{\epsilon K}} + \sqrt{\frac{p}{K\epsilon^2}} + \sqrt{\frac{p}{K}}\right)$. Assuming the typical setting for the privacy parameter ϵ, namely, $\epsilon = \Theta(1)$, this bound is more favorable than our bound based on the Phased ERM approach in a certain range of p as a function of K and d, particularly, when $p \geq \max\{\sqrt{d}, K/d\}$ or $\sqrt{\frac{K}{d^{1/2}}} \leq p \leq \sqrt{d}$.

Finally, we investigate the worst-group *empirical* risk minimization problem (2) in the *offline setting*, where each distribution D_i is observed by a fixed-size dataset S_i. We present a new algorithm based on a private version of the multiplicative group reweighing scheme (Abernethy et al., 2022; Diana et al., 2021). We show that our algorithm achieves a nearly optimal excess worst-group empirical risk of $\tilde{O}\left(\frac{p\sqrt{d}}{K\epsilon}\right)$.

1.2. Related Work

The worst-group risk minimization is closely related to the group DRO problem. In (Sagawa et al., 2019), the authors consider solving the empirical objective in the offline setting and propose an online algorithm based on the stochastic mirror descent (SMD) method from (Nemirovski et al., 2009). The convergence rate achieved in (Sagawa et al., 2019) is suboptimal because of the high variance of the gradient estimators. On the other hands, several works (Haghtalab et al.,

2022; Soma et al., 2022; Zhang et al., 2023) has studied the group DRO problem in the stochastic oracle setting and achieves an excess worst-group population risk of $\tilde{O}\left(\sqrt{\frac{p}{K}}\right)$ The basic idea behind the methods used in these works is to cast the objective into a zero-sum game based on the observation that the optimality gap depends on the expected regret of the max and min players. The authors in (Soma et al., 2022) demonstrate the tightness of this rate by proving a matching lower bound result.

The worst-group risk minimization problem is also extensively studied under the notion of min-max group fairness. Reference (Diana et al., 2021) studies learning with min-max group fairness and propose a multiplicative reweighting based method. However, their method assumes the access to a weighted empirical risk minimization oracle which may not be realistic in practice. Meanwhile, reference (Abernethy et al., 2022) studies a similar problem and present more efficient gradient based methods to achieve the min-max group fairness based on the idea of active sampling. The problem of min-max group fairness is also studied in (Mohri et al., 2019) under the federated learning setting. However, all these works do not provide any privacy guarantees.

Our work is also related to the line of works on differentially private stochastic saddle point (DP-SSP) problem. References (Yang et al., 2022; Zhang et al., 2022) have presented algorithms with the optimal rate for the weak duality gap, while the optimal rate for the strong duality gap is achieved in (Bassily et al., 2023). However, these works only consider the objective with L_2/L_2 geometry, directly applying the results from existing DP-SSP works will lead to a sub-optimal convergence rate due to the dependence of p on the L_2 Lipschtiz constant for the model parameters.

2. Preliminaries

We consider p groups, where each group $i \in [p]$ corresponds to a distributions D_i. Given a parameter space $\mathcal{W} \in \mathcal{R}^d$ and a loss function ℓ, we let $\ell(w, z)$ be the loss incurred by a weight vector w on a data point z. We assume the L_2 diameter of $\mathcal{W}$, $\|\mathcal{W}\|_2$, is bounded by M. Throughout this paper, we assume that the loss function ℓ is convex, L-Lipschitz and bounded by $[0, B]$ unless stated otherwise . The population loss of distribution D_i is written as $L_{D_i}(w) = \mathbb{E}_{z\sim D_i}\ell(w, z)$. When each group distribution D_i is observed by a dataset S_i with i.i.d samples drawn from D_i, we denote the empirical loss evaluated on S_i as $L_{S_i}(w) = \frac{1}{|S_i|}\sum_{z\in S_i}\ell(w, z)$.

Sample oracle: A sample oracle $\mathcal{C}_i$ w.r.t distribution D_i, $i \in [p]$ returns an i.i.d sample z drawn from D_i. We denote the collection of sample oracles from all groups as $\tilde{\mathcal{C}} = \{\mathcal{C}_1, \dots \mathcal{C}_p\}$.

α-saddle point: We say $(\tilde{x}, \tilde{y})$ is an α-saddle point of a minimax problem $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \phi(x, y)$ if $\phi(\tilde{x}, y) - \phi(x, \tilde{y}) \leq \alpha$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. When $\alpha = 0$, we simply call $(\tilde{x}, \tilde{y})$ a saddle point of ϕ.

Worst-group population risk minimization: Given p groups that each of them is with a distribution D_i, we define the expected (population) worst-group risk of model w as the maximal population loss of w across all group distributions written as

$$R(w) = \max_{i\in[p]} L_{D_i}(w)$$

Our objective of minimizing the worst-group risk is therefore written as

$$\min_{w\in\mathcal{W}} \max_{i\in[p]} L_{D_i}(w) \tag{3}$$

Denote $\Delta_p = \{\lambda \in [0, 1]^p : \|\lambda\|_1 = 1\}$ as the probability simplex over p dimensions and $\phi(w, \lambda) = \sum_{i=1}^p \lambda_i L_{D_i}(w)$.

Then the objective in (3) can be equivalently written as

$$\min_{w\in\mathcal{W}} \max_{\lambda\in\Delta_p} \phi(w, \lambda) \tag{4}$$

The equivalence is based on the fact that for any $w \in \mathcal{W}$, $\phi(w, \lambda)$ is maximized by a λ with all dimensions being zero except the one with the highest $L_{D_i}(w)$. We also define the *excess worst-group population risk* of $w \in \mathcal{W}$ as

$$\mathcal{E}(w; \{D_i\}_{i=1}^p) \triangleq \max_{i\in[p]} L_{D_i}(w) - \min_{\tilde{w}\in\mathcal{W}} \max_{i\in[p]} L_{D_i}(\tilde{w})$$

Differential Privacy (Dwork et al., 2006): A mechanism $\mathcal{M}$ is (ϵ, δ)-DP if for any two neighboring datasets S and S' that differ on single *data point* and any output set $\mathcal{O}$, we have

$$P\left(M(S) \in \mathcal{O}\right) \leq e^{\epsilon} P\left(M(S') \in \mathcal{O}\right) + \delta$$

We focus on record-level privacy throughout the paper. We say that a pair of sequences of sampled data points $\{z_1, \dots z_T\}$ and $\{z'_1, \dots z'_T\}$ are neighbors if, for some $j \in [T]$, $z_i = z'_i$ for all $i \in [T] \setminus \{j\}$ and $z_j \neq z'_j$.

Stability: Our analysis depends on the notion of uniform argument stability defined as follows

Definition 1. Consider a randomized algorithm $\mathcal{A}$ that takes a collection of datasets $\tilde{S}$ as input and outputs $(\mathcal{A}_w(\tilde{S}), \mathcal{A}_\lambda(\tilde{S}))$ as the solution to a minimax problem $\min_{w\in\mathcal{W}} \max_{\lambda\in\Delta} \phi(w, \lambda)$. Then $\mathcal{A}$ is said to attain Δ-uniform argument stability with respect to w if for any pairs of adjacent $\tilde{S}$ and $\tilde{S}'$ that differs in a single data point, we have

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathcal{A}_w(\tilde{S}) - \mathcal{A}_w(\tilde{S}')\right\|\right] \leq \Delta$$

3. Minimax Phased ERM

In this section, we propose an algorithm attaining an excess worst-group population risk of $\tilde{O}\left(\sqrt{\frac{p}{K}} + \frac{p\sqrt{d}}{K\epsilon}\right)$, where p and K denote the number of groups and the total number of samples drawn from all groups, respectively. Our algorithm involves iteratively solving a sequence of regularized minimax objectives. In iteration t, given the previous iterate w_{t-1} and a dataset collection $\{S_i^t\}_{i=1}^p$ where S_i^t is the dataset formed by the data points drawn from the sample oracle $\mathcal{C}_i$ at current iteration, we find an approximate saddle point $(\tilde{w}_t, \tilde{\lambda}_t)$ of the regularized objective

$$F_t(w, \lambda) = \sum_{i=1}^p \lambda_i L_{S_i^t}(w) + \mu_w^t \|w - w_{t-1}\|_2^2 - \mu_\lambda^t \sum_{j=1}^p \lambda_j \log \lambda_j$$

and output $\tilde{w}_t$. We can show that $\tilde{w}_t$ has low sensitivity w.r.t. the collection of samples drawn from all groups , and hence, we can use the standard Gaussian mechanism to ensure that the computation of $\tilde{w}_t$ is differentially private.

Proving the utility guarantee of our algorithm involves new stability based argument. First, given a dataset collection $\tilde{S} = \{S_1, \dots S_p\}$ where each $S_i \sim D_i^n$ is a dataset formed by n samples drawn from the sample oracle $\mathcal{C}_i$. we show that an algorithm that outputs an approximate saddle point $(\tilde{w}, \tilde{\lambda})$ for the regularized objective $F(w, \lambda) = \sum_{i=1}^p \lambda_i L_{S_i}(w) + \frac{\mu_w}{2}\|w - w'\|_2^2 - \mu_\lambda \sum_{j=1}^p \lambda_j \log \lambda_j$ has uniform argument stability of $O\left(\frac{L}{n\mu_w} + \frac{B}{n\sqrt{\mu_w \mu_\lambda}}\right)$ with respect to $\tilde{w}$. Then we prove that Δ-uniform argument stability of $\tilde{w}$ leads to a $\tilde{O}\left(\Delta + \frac{B}{\sqrt{n}}\right)$ generalization error bound.

Our stability-based generalization argument is distinct from existing generalization analysis based on uniform convergence (Mohri et al., 2019; Abernethy et al., 2022), which leads to sub-optimal excess worst-group risk for general convex losses even in the non-private regime. On the other hand, our result circumvents a known bottleneck in the relationship between stability and generalization in stochastic minimax optimization (also, known as Stochastic Saddle Point (SSP) problems). In particular, in the L_2/L_2 setting of the SSP problem, the best known generalization guarantee using a Δ-argument stable algorithm is $\sqrt{\Delta}$ generalization error based on the analysis of the strong duality gap (Ozdaglar et al., 2022).

3.1. Algorithm description

Our algorithm relies on a non-private subroutine to solve a regularized minimax problem. More concretely, let $\tilde{S} = \{S_1, \dots S_p\}$ be a collection of datasets. Let μ_w, $\mu_\lambda > 0$ and $w' \in \mathcal{W}$. Let $\mathcal{A}_{emp}(\tilde{S}, \mu_w, \mu_\lambda, w', \alpha)$ be an empirical minimax solver that computes an α-saddle point $(\tilde{w}, \tilde{\lambda})$ of

$$\min_{w \in \mathcal{W}} \max_{\lambda \in \Delta_p} \left\{ F(w, \lambda) := \sum_{i=1}^p \lambda_i L_{S_i}(w) + \frac{\mu_w}{2}\|w - w'\|_2^2 - \mu_\lambda \sum_{j=1}^p \lambda_j \log \lambda_j \right\} \tag{5}$$

and outputs $\tilde{w}$.

The formal description of our algorithm is given in Algorithm 1. Our algorithm is inspired by the Phased ERM approach of (Feldman et al., 2020), which was used to attain the optimal rate for the simpler problem of differentially private stochastic convex optimization. We repurpose the Phased ERM approach for the stochastic minimax problem. Our algorithm involves several modifications to this approach that enable us to attain strong guarantees for the worst-group population risk minimization problem. Crucially, we provide a new analysis for our algorithm that utilizes the argument uniform stability of the regularized minimax problems (see, Lemma 4).

In Algorithm 1, a sequence of regularized empirical minimax problems are defined and solved iteratively using the non-private minimax solver $\mathcal{A}_{emp}$ (defined earlier in this section). The total number of such problems (the number of rounds in the algorithm) T is logarithmic in K/p. In round $t \in [T]$, we first sample a dataset of size n_t from each distribution to construct a new dataset collection $\tilde{S}_t$. The center of the regularization term is chosen to be $w' = w_{t-1}$, i.e., the previous iterate. The settings of the regularization parameters in each round are carefully chosen to ensure convergence to a nearly optimal rate. The regularization parameter μ_λ is fixed across all rounds whereas μ_w is doubled with every round. Our main result is stated below.

Theorem 2. *Algorithm 1 is (ϵ, δ)-differentially private. Let $\eta = \frac{M}{D} \min\left\{\frac{\epsilon}{\sqrt{72d\log(\frac{K}{p})\log(\frac{1}{\delta})}}, \frac{\sqrt{p}}{\log^{\frac{3}{4}}(K)\sqrt{K}}\right\}$. We have*

$$\mathbb{E}\left[\max_{i \in [p]} L_{D_i}(w_T)\right] - \min_{w \in \mathcal{W}} \max_{i \in [p]} L_{D_i}(w) = O\left(MD\left(\log^{\frac{11}{4}}(K)\sqrt{\frac{p}{K}} + \frac{\log^{\frac{5}{2}}(K) p\sqrt{d\log(1/\delta)}}{K\epsilon}\right)\right),$$

where w_T is the output of Algorithm 1 and the expectation is taken over the sampled data points and the algorithm's inner randomness.

Remarks:

- One can show that the rate in Theorem 2 is nearly optimal in the offline setting considered in Section 5, where each distribution is observed by a fixed-size dataset of size K/p. A lower bound instance of $\Omega\left(\frac{p\sqrt{d}}{K\epsilon} + \sqrt{\frac{p}{K}}\right)$ can

Algorithm 1 Minimax Phased ERM

Input Sample oracles $\{\mathcal{C}_1, \dots \mathcal{C}_p\}$, step size η, $D = \max\{L, B\}$, total number of samples drawn from all groups K, non-private empirical solver A_{emp}.
Initialize w_0 arbitrary parameter vector in $\mathcal{W}$.
Set $n = K/p$ and $T = \log_2(n)$.
for $t = 1, \dots, T$ **do**
  Let $n_t = n/T$, $\eta_t = \eta 2^{-t}$, $\mu_w^t = 1/(\eta_t n_t)$, $\mu_\lambda^t = 1/(\eta n)$.
  For each $i \in [p]$, sample $S_i^t \sim D_i^{n_t}$ using the sample oracle $\mathcal{C}_i$. Denote $\tilde{S}^t = \{S_1^t, \dots S_p^t\}$.
  Let $\alpha_t = \frac{L^2}{8n_t^2\mu_w^t} + \frac{B^2}{8n_t^2\mu_\lambda^t}$.
  Let $\tilde{w}_t = \mathcal{A}_{emp}(\tilde{S}^t, \mu_w^t, \mu_\lambda^t, w_{t-1}, \alpha_t)$.
  Set $w_t = \tilde{w}_t + \xi_t$ where $\xi_t \sim \mathcal{N}(0, \sigma_t I_d)$ and $\sigma_t = 6D\sqrt{2\log_2(n)\log(1/\delta)\eta\eta_t}/\epsilon$.
end for
Output: w_T.

be constructed in this case. More details are given in Appendix D. Also, note that when $d = O(K/p)$, the rate in Theorem 2 matches the the non-private lower bound $\Omega\left(\sqrt{\frac{p}{K}}\right)$ up to logarithmic factors.

- In Appendix A.1, we give an instantiation of $\mathcal{A}_{emp}$, which is an iterative algorithm for solving the regularized minimax problem in (5) with a convergence rate of $\tilde{O}(1/N)$, where N is the number of iterations.

The privacy and utility guarantees in Theorem 2 rely on a stability-based result of the saddle point of the regularized objective (5), which we present in the following lemma.

Lemma 3. *Consider two neighboring dataset collections $\tilde{S} = \{S_1, \dots S_p\} \in \mathcal{Z}^{n\times p}$ and $\tilde{S}' = \{S_1', \dots S_p'\} \in \mathcal{Z}^{n\times p}$ that differs in a single data point. Let $(\tilde{w}, \tilde{\lambda})$ and $(\tilde{w}', \tilde{\lambda}')$ be an α-saddle point of $F(w, \lambda)$ in (5) when the dataset collections are $\tilde{S}$ and $\tilde{S}'$, respectively. By letting $\alpha \le \frac{L^2}{8n^2\mu_w} + \frac{B^2}{8n^2\mu_\lambda}$, we have*

$$\|\tilde{w} - \tilde{w}'\|_2 \le \frac{3}{n}\left(\frac{L}{\mu_w} + \frac{B}{\sqrt{\mu_w\mu_\lambda}}\right)$$

The proof of Lemma 3 follows from Lemma 2 in (Zhang et al., 2021) as well as the observation that $F(w, \lambda)$ can be equivalently written in the finite-sum form composed with regularization terms. We give full details in Appendix A.3.

Now we are ready to present our main technical lemma, which shows that an approximate saddle point of the regularized objective $F(w, \lambda)$ in (5) gives a good solution to the worst-group population risk minimization problem. We will sketch the proof of this lemma and defer the full proof to Appendix A.4.

Lemma 4. *Let $(\tilde{w}, \tilde{\lambda})$ be an α-saddle point of $F(w, \lambda)$ in (5) with $\alpha \le \frac{L^2}{8n^2\mu_w} + \frac{B^2}{8n^2\mu_\lambda}$. For any $w \in \mathcal{W}$,*

$$\mathbb{E}\left[\max_{i\in[p]} L_{D_i}(\tilde{w})\right] - \max_{i\in[p]} L_{D_i}(w)$$
$$= O\Bigg(\mu_w \|w - w'\|_2^2 + \mu_\lambda \log p$$
$$+ \left(\frac{L^2}{n\mu_w} + \frac{LB}{n\sqrt{\mu_w\mu_\lambda}}\right)\log(n)\log(np) + \frac{B\sqrt{\log(pn)}}{\sqrt{n}}\Bigg),$$

where the expectation is taken over the randomness in the datasets $\tilde{S}$.

Proof. (sketch) For simplicity, we let $\alpha = 0$, which means $(\tilde{w}, \tilde{\lambda})$ is the exact saddle point of $F(w, \lambda)$. In particular, by the definition of saddle point, one can show that for any $w \in \mathcal{W}$

$$\max_{i\in[p]} L_{S_i}(\tilde{w}) - \hat{\phi}(w, \tilde{\lambda}) \le \frac{\mu_w}{2}\|w - w'\|_2^2 + \mu_\lambda \log p$$

where $\hat{\phi}(w, \lambda) = \sum_{i=1}^p \lambda_i L_{S_i}(w)$.

By Lemma 3 and Lipschitzness, for any given $i \in [p]$, computing $\tilde{w}$ has a uniform stability of $\gamma = \frac{3}{n}\left(\frac{L^2}{\mu_w} + \frac{LB}{\sqrt{\mu_w\mu_\lambda}}\right)$ with respect to the change of datapoint in S_i. By Theorem 1.1 in (Feldman & Vondrak, 2019), we have with high probability over the sampling of $\tilde{S}$,

$$|L_{S_i}(\tilde{w}) - L_{D_i}(\tilde{w})| = \tilde{O}\left(\gamma + \frac{B}{\sqrt{n}}\right)$$

Then we can use union bound across all $i \in [p]$ and obtain with high probability

$$|L_{S_i}(\tilde{w}) - L_{D_i}(\tilde{w})| = \tilde{O}\left(\gamma + \frac{B}{\sqrt{n}}\right) \quad \forall i \in [p]$$

In particular, we have with high probability

$$|\max_{i\in[p]} L_{D_i}(\tilde{w}) - \max_{i\in[p]} L_{S_i}(\tilde{w})| = \tilde{O}\left(\gamma + \frac{B}{\sqrt{n}}\right)$$

Meanwhile, we can show that with high probability, $|\phi(w, \tilde{\lambda}) - \hat{\phi}(w, \tilde{\lambda})| = \tilde{O}\left(\frac{B}{\sqrt{n}}\right)$ for any given $w \in \mathcal{W}$. Finally, we put everything together and have with high probability

$$\max_{i\in[p]} L_{D_i}(\tilde{w}) - \max_{i\in[p]} L_{D_i}(w)$$
$$\le \max_{i\in[p]} L_{D_i}(\tilde{w}) - \phi(w, \tilde{\lambda})$$
$$\le \max_{i\in[p]} L_{S_i}(\tilde{w}) - \hat{\phi}(w, \tilde{\lambda}) + \tilde{O}\left(\gamma + \frac{B}{\sqrt{n}}\right)$$
$$= \tilde{O}\left(\mu_w \|w - w'\|_2^2 + \mu_\lambda + \gamma + \frac{B}{\sqrt{n}}\right)$$

Replacing γ with $\frac{3}{n}\left(\frac{L^2}{\mu_w}+\frac{LB}{\sqrt{\mu_w\mu_\lambda}}\right)$ and we obtain the desired result. □

Remark: As a side product of Lemma 4, we can sample a new dataset collection $\tilde{S} = \{S_1, \dots S_p\}$ where $S_i \sim D_i^{K/p}$. By letting $\mu_w = \frac{D}{M}\sqrt{\frac{p}{K}}\sqrt{\log(K/p)\log(K)}$, $\mu_\lambda = D\sqrt{\frac{p\log(K/p)\log(K)}{K\log p}}$ and w' be any arbitrary parameter in $\mathcal{W}$, solving the regularized objective (5) with $\tilde{S}$ gives an excess worst-group population risk of $\tilde{O}\left(\sqrt{\frac{p}{K}}\right)$, which matches the non-private lower bound in (Soma et al., 2022) up to logarithmic factors. Prior methods (nearly) matching the non-private lower bound depend on making one pass over the sampled data points to directly optimize the population objective, our regularization-based method however does not require such one pass structure.

Now, we are ready to give a proof sketch of Theorem 2. The full proof is deferred to Appendix A.5.

Proof. (Proof Sketch of Theorem 2)

Privacy: At iteration t, by the stability argument in Lemma 3, we have

$$\begin{aligned}\left\|\tilde{w_t}-\tilde{w_t}'\right\|_2 &\le \frac{3L}{n_t\mu_w^t}+\frac{3B}{n_t\sqrt{\mu_w^t\mu_\lambda^t}}\\ &= 3L\eta_t+\frac{3B\sqrt{n_t n\eta_t\eta}}{n_t}\\ &\le 6D\sqrt{(\log_2 n)\eta_t\eta}\end{aligned}$$

where $\tilde{w_t}$ and $\tilde{w_t}'$ are outputs from neighboring dataset collections $\tilde{S}^t$ and $\tilde{S}^{t'}$ that differ in one datapoint. Since the sampled minibatches $\tilde{S}^t$ are disjoint across iterations, the privacy of Algorithm 1 follows from the privacy guarantee of the Gaussian mechanism and parallel composition.

Utility: Recall that $R(w) = \max_{i\in[p]} L_{D_i}(w)$. By Lemma 4, we have

$$\begin{aligned}&\mathbb{E}[R(\tilde{w_t})-R(\tilde{w_{t-1}})]\\ &= \tilde{O}\left(\frac{\mathbb{E}[\|\xi_{t-1}\|_2^2]}{n_t\eta_t}+\frac{1}{\eta n}+D^2\sqrt{\eta\eta_t}+\frac{B}{\sqrt{n_t}}\right)\end{aligned}$$

.

Also, we have

$$\mathbb{E}\|\xi_t\|_2^2 = d\sigma_t^2 = 72dD^2\log n\log(1/\delta)2^{-t}\eta^2/\epsilon^2$$

Therefore, as long as

$$\eta \le \frac{M\epsilon}{D\sqrt{72d\log n\log(1/\delta)}}$$

We will have

$$\mathbb{E}\|\xi_t\|_2^2 \le 2^{-t}M^2 \text{ and } \mathbb{E}\|\xi_t\|_2 \le \sqrt{2}^{-t}M$$

Let $\tilde{w}_0 = w^*$ and $\xi_0 = w_0 - w^*$. Then $\|\xi_0\|_2 \le M$. Hence,

$$\begin{aligned}&\mathbb{E}[R(w_T)]-R(w^*)\\ &= \sum_{t=1}^{T}\mathbb{E}[R(\tilde{w_t})-R(\tilde{w_{t-1}})]+\mathbb{E}[R(w_T)-R(\hat{w}_T)]\\ &\le \sum_{t=1}^{T}\tilde{O}\left(\frac{\mathbb{E}[\|\xi_{t-1}\|_2^2]}{n\eta_t}+\frac{1}{\eta n}+D^2\sqrt{\eta\eta_t}+\frac{B}{\sqrt{n}}\right)\\ &\quad + L\mathbb{E}[\|\xi_T\|_2]\end{aligned}$$

The inequality holds since $R(\cdot)$ is L-Lipschitz and $n_t = n/\log_2(n)$.

We have

$$\frac{\mathbb{E}[\|\xi_{t-1}\|_2^2]}{n_t\eta_t} \le \frac{2^{-(t-1)}M^2\log_2 n}{n2^{-t}\eta} = \frac{2(\log_2 n)M^2}{\eta n}$$

and $\mathbb{E}[\|\xi_T\|_2] \le M\sqrt{2}^{-\log_2 n} = \frac{M}{\sqrt{n}}$.

Therefore,

$$\begin{aligned}&\mathbb{E}[R(w_T)]-R(w^*)\\ &\le \sum_{t=1}^{T}\tilde{O}\left(\frac{M^2}{\eta n}+D^2\eta+\frac{B}{\sqrt{n}}\right)+\frac{ML}{\sqrt{n}}\\ &= \tilde{O}\left(\frac{MD}{\sqrt{n}}+\frac{MD\sqrt{d}}{n\epsilon}\right)\end{aligned}$$

Replace n with K/p and we get the desired result. □

4. Worst-group Risk Minimization using DP OCO Algorithm

In this section, we give a framework to directly minimize the expected worst-group loss in (4) using any DP online convex optimization (OCO) algorithm as a subroutine.

Our framework leverages the idea from (Haghtalab et al., 2022; Soma et al., 2022; Zhang et al., 2023) of casting the objective into a two-player zero-sum game. One can show that the excess expected worst-group risk is bounded by the sum of the expected regret bounds of the min-player (the w-player) and max-player (the λ-player) using stochastic gradient estimation. Therefore, given any DP-OCO algorithm $\mathcal{Q}_-$, our frameworks proceed by letting the w-player run $\mathcal{Q}_-$ with an estimate of the loss function $\phi(w, \lambda_t)$. Meanwhile, we instantiate the λ-player as EXP3 (Auer et al., 2002), an adversarial multi-armed bandit algorithm and feed it with a privatized estimator of $\nabla_\lambda\phi(w_t, \lambda_t)$.

An online convex optimization algorithm interacts with an adversary for T rounds. In each round, the algorithm picks a vector w_t and then the adversary chooses a convex loss function ℓ_t. The performance of an OCO algorithm $\mathcal{Q}$ is

therefore measured by the regret defined as

$$r_{\mathcal{Q}}(T) = \frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(w_t) - \min_{w^*\in\mathcal{W}}\sum_{t=1}^{T}\ell_t(w^*)\right]$$

Definition 5. An online algorithm $\mathcal{M}$ is (ϵ,δ)-DP if for any two sequence of loss functions $S = (\ell_1,\dots\ell_T)$ and $S' = (\ell'_1,\dots\ell'_T)$ that differs in one loss function, we have

$$P(\mathcal{M}(S)\in O) \le e^{\epsilon}P(\mathcal{M}(S')\in O) + \delta$$

Suppose we have a DP OCO algorithm $\mathcal{Q}_-$ with regret $r_{\mathcal{Q}_-}(T)$ that observes $l_t = \ell(\cdot, x_t)$ for some data point x_t in each iteration, the formal description of our framework is given in Algorithm 2. We give the privacy and utility guarantee of Algorithm 2 in the following theorems.

Algorithm 2 Worst-group risk minimization using DP-OCO algorithm

Input Sample oracles $\{\mathcal{C}_1,\dots\mathcal{C}_p\}$, DP OCO algorithm $\mathcal{Q}_-$, learning rate η, privacy parameters ϵ,δ.
Initialize $w_1\in\mathcal{W}$ and $\lambda_1 = (1/p,\dots 1/p)\in[0,1]^p$.
Set $U = B + \frac{2B}{\epsilon}\log(T)$.
for $t = 1,\dots,T-1$ **do**
  Sample $i_t\sim\lambda_t$.
  Sample $x_t^-, x_t^+\sim D_{i_t}$ using the sample oracle $\mathcal{C}_{i_t}$.
  Update $w_{t+1} = \mathcal{Q}_-(w_{1:t}, x_t^-)$.
  Compute $\tilde{\ell}_t = U - \ell(w_t, x_t^+) + y_t,\ y_t\sim \mathrm{Lap}(B/\epsilon)$.
  Update $\lambda_{t+1,i_t} = \lambda_{t,i_t}\exp\left(\frac{-\eta\tilde{\ell}_t}{\lambda_{t,i_t}}\right)$.
  Normalize $\lambda_{t+1} = \frac{\lambda_{t+1}}{\|\lambda_{t+1}\|_1}$.
end for
Output $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T}w_t$, $\bar{\lambda}_T = \frac{1}{T}\sum_{t=1}^{T}\lambda_t$.

Theorem 6. *(Privacy Guarantee) Algorithm 2 is (ϵ,δ)-differentially private.*

Proof. The sampled dataset can be divided into two disjoint sets. $S = \{S_-, S_+\}$ where $S_- = \{x_1^-,\dots x_T^-\}$ and $S_+ = \{x_1^+,\dots x_T^+\}$. Now, we proceed inductively. When $t = 1$, it is easy to see that the generation of (w_1,λ_1) is (ϵ,δ)-DP.

Now we assume that the generation of $\{(w_i,\lambda_i)\}_{i=1}^t$ is (ϵ,δ)-DP for iteration t. At iteration $t+1$, we observe that the computation of w_{t+1} only depends on S_- and $\{w_i\}_{i=1}^t$, hence w_{t+1} can be generated under (ϵ,δ)-DP constraint due to the privacy guarantee of $\mathcal{Q}_-$.

Furthermore, since ℓ is uniformly bounded by B, then clearly the sensitivity of $\ell(w_t,\cdot)$ w.r.t. replacing one data point in the input sequence is bounded by B. Hence, by the properties of the Laplace mechanism, $\tilde{\ell}_t$ is also generated in $(\epsilon,0)$-DP manner. Given that λ_{t+1} depends on only $\tilde{\ell}_t$ and λ_t and since DP is closed under postprocessing, computing λ_{t+1} is (ϵ,δ)-DP. Since in iteration t, two different data points (namely, x_t^- and x_t^+) are used to generate w_{t+1} and λ_{t+1}, then by parallel composition (and given the induction hypothesis), generating (w_{t+1},λ_{t+1}) is (ϵ,δ)-DP. □

Theorem 7. *(Utility Guarantee) Suppose the regret of $\mathcal{Q}_-$ is denoted as $r_{\mathcal{Q}_-}(\cdot)$. By setting $\eta = \sqrt{\frac{\ln(p)}{pTU^2}}$, we have*

$$\mathbb{E}\left[\max_{i\in[p]}L_{D_i}(\bar{w}_T)\right] - \min_{w\in\mathcal{W}}\max_{i\in[p]}L_{D_i}(w)$$
$$= r_{\mathcal{Q}_-}(K/2) + O\left(\left(B + \frac{B\log(K)}{\epsilon}\right)\sqrt{\frac{p\log(p)}{K}}\right)$$

where $\bar{w}_T$ is the output of Algorithm 2 and the expectation is over the sampling of data points and the algorithm's randomness.

The proof of Theorem 7 can be found in Appendix B.1. In particular, by instantiating the DP OCO algorithm $\mathcal{Q}_-$ with the DP-FTRL algorithm (Kairouz et al., 2021), we have the following corollary.

Corollary 8. *Let $\mathcal{Q}_-$ in Algorithm 2 be DP-FTRL algorithm from (Kairouz et al., 2021). By plugging the regret bound of DP-FTRL into Theorem 7, we have*

$$\mathbb{E}\left[\max_{i\in[p]}L_{D_i}(\bar{w}_T)\right] - \min_{w\in\mathcal{W}}\max_{i\in[p]}L_{D_i}(w)$$
$$= O\left(\frac{ML}{\sqrt{K}} + ML\sqrt{\frac{d^{\frac{1}{2}}\log^2(1/\delta)\log(K)}{\epsilon K}}\right.$$
$$\left. + \left(B + \frac{B\log(K)}{\epsilon}\right)\sqrt{\frac{p\log(p)}{K}}\right),$$

where the expectation is over the sampling of data points and the algorithm's randomness.

Remark: Corollary 8 demonstrates an excess expected worst-group risk of $\tilde{O}\left(\sqrt{\frac{d^{1/2}}{\epsilon K}} + \sqrt{\frac{p}{K\epsilon^2}} + \sqrt{\frac{p}{K}}\right)$. For the common case of $\epsilon = \Theta(1)$, the rate in Corollary 8 matches the lower bound of $\Omega\left(\sqrt{\frac{p}{K}}\right)$ shown in (Soma et al., 2022) up to logarithmic factors in the regime of $d = \tilde{O}\left(p^2\right)$.

5. Private Worst-group Empirical Risk Minimization

In this section, we consider the *offline setting* where each distribution D_i is observed by a fixed-size dataset S_i with i.i.d samples drawn from D_i. In the offline setting, the learner will take the dataset collection $\tilde{S} = \{S_1,\dots S_p\}$ as input instead of querying the sample oracles directly. This

setting captures scenarios where each group is represented by a dataset whose size is fixed beforehand.

In particular, we denote the dataset collection as $\tilde{S} = \{S_1, \dots S_p\}$. Without loss of generality, we assume all datasets S_i have same size, namely $|S_i| = n$ for all $i \in [p]$. Note that n can also be written as $n = K/p$ in this case. We first define the worst-group empirical risk minimization problem as follows.

Worst-group empirical risk minimization: Given a dataset collection $\tilde{S} = \{S_1, \dots S_p\} \in \mathcal{Z}^{n \times p}$, we denote $\hat{\phi}(w, \lambda) = \sum_{i=1}^{p} \lambda_i L_{S_i}(w)$. The worst-group empirical loss can be expressed as $\max_{i \in [p]} L_{S_i}(w) = \max_{\lambda \in \Delta_p} \hat{\phi}(w, \lambda)$. The empirical objective is hence written as

$$\min_{w \in \mathcal{W}} \max_{\lambda \in \Delta_p} \hat{\phi}(w, \lambda) \tag{6}$$

We define the *excess worst-group empirical risk* of $w \in \mathcal{W}$ as $\hat{\mathcal{E}}(w; \{S_i\}_{i=1}^p) \triangleq \max_{i \in [p]} L_{S_i}(w) - \min_{\tilde{w} \in \mathcal{W}} \max_{i \in [p]} L_{S_i}(\tilde{w})$.

Next, we will present a method described in Algorithm 3 to solve the empirical objective (6) under the differential privacy constraint. Algorithm 3 is based on a private version of the multiplicative reweighting scheme (Abernethy et al., 2022; Diana et al., 2021) and attains nearly optimal excess empirical worst-group risk.

5.1. Private Multiplicative Group Reweighting

Here, we provide the details of our algorithm described formally in Algorithm 3. In this algorithm, we maintain a sampling weight vector $\lambda \in [0, 1]^p$ for all datasets $\{S_i\}_{i=1}^p$. In each iteration, we first sample a dataset according to λ. Next, we sample a mini-batch from this dataset and update the model parameters using Noisy-SGD. We then update the group weight vector λ using the multiplicative weights approach based on privatized loss values computed over the datasets at the model parameters. We present the privacy and utility guarantee of Algorithm 3 in the following theorems.

Theorem 9. *(Privacy guarantee) In Algorithm 3, let* $\tau = \frac{cBp}{K\epsilon}\sqrt{T\log(1/\delta)}$ *and* $\sigma^2 = \frac{cTp^2L^2\log(T/\delta)\log(1/\delta)}{K^2\epsilon^2}$ *for some universal constant c. Then, Algorithm 3 is* (ϵ, δ)*-DP.*

Theorem 10. *(Convergence guarantee) There exist settings of* $T = O\left(\frac{(ML+B\sqrt{\log(p)})K^2\epsilon^2}{GBdp^2\log(1/\delta)}\right)$, $\eta_w = O\left(\frac{M^2}{T(G^2+d\sigma^2)}\right)$ *and* $\eta_\lambda = O\left(\sqrt{\frac{\log(p)}{U^2T}}\right)$, *where* $U = O\left(B + \frac{Bp}{K\epsilon}\sqrt{T\log(1/\delta)}\log(KT)\right)$ *such that*

$$\begin{aligned}
&\mathbb{E}\left[\max_{i \in [p]} L_{S_i}(\bar{w}_T)\right] - \min_{w \in \mathcal{W}} \max_{i \in [p]} L_{S_i}(w) \\
&= O\Bigg(\frac{MLp\sqrt{d\log(1/\delta)\log(K/\delta)}}{K\epsilon} \\
&\quad + \frac{Bp\sqrt{\log(p)\log(1/\delta)}\log(\frac{K\epsilon}{d\log(1/\delta)})}{K\epsilon}\Bigg),
\end{aligned}$$

where the expectation is over the algorithm's randomness.

Remarks:

- The convergence rate in Theorem 10 is nearly optimal up to logarithmic factors. A lower bound instance can be created to reduce the original problem to a DP-ERM problem with single dataset, which leads to a lower bound of $\Omega\left(\frac{p\sqrt{d}}{K\epsilon}\right)$(Bassily et al., 2014). A more detailed argument for the lower bound instance can be found in Appendix D.
- (Abernethy et al., 2022) proposes a gradient method based on the active group selection scheme. One can also design a private algorithm based on this scheme and achieve a suboptimal rate of $\tilde{O}\left(\sqrt{\frac{MLBp}{K\epsilon}} + \frac{MLp\sqrt{d}}{K\epsilon}\right)$. We defer the details to Appendix C.2.
- Our results for the offline setting can be readily extended to the case where the dataset sizes are non-uniform. In particular, when different groups are associated with datasets

Algorithm 3 Noisy SGD with Multiplicative Group Reweighting (Noisy-SGD-MGR)

Input: Collection of datasets $\tilde{S} = \{S_1, \dots S_p\} \in \mathcal{Z}^{n \times p}$, mini-batch size m, # iterations T, learning rates η_w and η_λ, privacy parameters ϵ, δ, and noise scales σ^2, τ.
Initialize $w_1 \in \mathcal{W}$ and $\lambda_1 = \left(\frac{1}{p}, \dots \frac{1}{p}\right) \in [0, 1]^p$.
for $t = 1, \dots, T-1$ **do**
  Sample $i_t \sim \lambda_t$.
  Sample $B_t = \{z_1, \dots z_m\}$ from S_{i_t} uniformly with replacement.
  Update the model as follows

$$w_{t+1} = \text{Proj}_{\mathcal{W}}\left(w_t - \eta_w\left(\frac{1}{m}\sum_{z \in B_t} \nabla \ell(w_t, z) + G_t\right)\right)$$

  where $G_t \sim \mathcal{N}(0, \sigma^2 I_d)$.
  Compute $L_t = [-L_{S_i}(w_t) + y_{i,t}]_{i=1}^q$, $y_{i,t} \overset{\text{iid}}{\sim} \text{Lap}(\tau)$.
  Update the weights $\tilde{\lambda}_{t+1}^i = \lambda_t^i \exp(-\eta_\lambda L_t^i)$, $\forall\, i \in [p]$.
  Normalize $\lambda_{t+1} = \frac{\tilde{\lambda}_{t+1}}{\|\tilde{\lambda}_{t+1}\|_1}$.
end for
Output: $\bar{w}_T = \frac{1}{T}\sum_{t=1}^T w_t$, $\bar{\lambda}_T = \frac{1}{T}\sum_{t=1}^T \lambda_t$.

of different sizes, we find the group with the minimum dataset size, which we denote as n_{min}. For the worst-group population risk, we can simply replace n with n_{min} in Algorithm 1 and sample without replacement from the dataset of each group. This gives us an excess worst-group population risk of $\tilde{O}(\frac{\sqrt{d}}{n_{min}\epsilon} + \frac{1}{\sqrt{n_{min}}})$. For the worst-group empirical risk, we can let the added noise scale with n_{min} instead of n in Algorithm 3, which gives an excess worst-group empirical risk of $\tilde{O}(\frac{\sqrt{d}}{n_{min}\epsilon})$. Both rates can be shown to be nearly optimal by constructing a matching lower bound instance (similar to our construction in Appendix D) in which the group with the minimum dataset size has a higher population or empirical risk than the other groups.

6. Conclusion

We presented differentially private algorithms for solving the worst-group risk minimization problem. In particular, we gave two upper bounds on the excess worst-group population risk, one of them is tight in the offline setting and we established the optimal rate for the excess worst-group empirical risk. Establishing the optimal rate for the worst-group population risk in general is left as an interesting open problem for future work.

Acknowledgements

This work is supported by NSF Award 2112471 and NSF CAREER Award 2144532.

Impact Statement

This work aims at advancing the theoretical understanding of privacy-preserving machine learning, which is an area of clear and well-established societal impacts. There is nothing on the societal impact of this work which we feel must be specifically highlighted here.

References

Abernethy, J. D., Awasthi, P., Kleindessner, M., Morgenstern, J., Russell, C., and Zhang, J. Active sampling for min-max fairness. In *International Conference on Machine Learning*, volume 162, 2022.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Bassily, R., Guzmán, C., and Menart, M. Differentially private algorithms for the stochastic saddle point problem with optimal rates for the strong gap. *arXiv preprint arXiv:2302.12909*, 2023.

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

Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 66–76, 2021.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279. PMLR, 2019.

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Haghtalab, N., Jordan, M., and Zhao, E. On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 35: 406–419, 2022.

Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pp. 5213–5225. PMLR, 2021.

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

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.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Ozdaglar, A., Pattathil, S., Zhang, J., and Zhang, K. What is a good metric to study generalization of minimax learners? *Advances in Neural Information Processing Systems*, 35:38190–38203, 2022.

Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. Minimax demographic group fairness in federated learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 142–159, 2022.

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. *arXiv preprint arXiv:1911.08731*, 2019.

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

Yang, Z., Hu, S., Lei, Y., Vashney, K. R., Lyu, S., and Ying, Y. Differentially private sgda for minimax problems. In *Uncertainty in Artificial Intelligence*, pp. 2192–2202. PMLR, 2022.

Zhang, J., Hong, M., Wang, M., and Zhang, S. Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pp. 568–576. PMLR, 2021.

Zhang, L., Thekumparampil, K. K., Oh, S., and He, N. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *Advances in Neural Information Processing Systems*, 35:35174–35187, 2022.

Zhang, L., Zhao, P., Yang, T., and Zhou, Z.-H. Stochastic approximation approaches to group distributionally robust optimization. *arXiv preprint arXiv:2302.09267*, 2023.

A. Missing Proofs in Section 3

A.1. Non-private algorithm for the regularized objective

Here, we provide an non-private algorithm in solving the regularized minimax problem in (5) written as

$$\min_{w\in\mathcal{W}}\max_{\lambda\in\Delta_p} F(w,\lambda) := \sum_{i=1}^{p}\lambda_i L_{S_i}(w) + \frac{\mu_w}{2}\|w-w'\|^2 - \mu_\lambda\sum_{j=1}^{p}\lambda_j\log\lambda_j$$

with an convergence rate of $\tilde{O}\left(1/N\right)$ where N is the number of iterations.

Algorithm 4 Minimax Optimization for SC-SC Objective

Input: Collection of datasets $S=\{S_1,\dots S_p\}$, regularization parameters μ_w, $\mu_\lambda>0$.
Init: $w_1\in\mathcal{W}$.
for $t=1,\dots,N-1$ **do**
  Compute $\tilde{\lambda}_t=\left[\exp(L_{S_1}(w_t)/\mu_\lambda),\dots\exp(L_{S_p}(w_t)/\mu_\lambda)\right]$.
  Let $\lambda_t=\frac{\tilde{\lambda}_t}{\|\lambda_t\|_1}$.
  Compute $\nabla_t=\sum_{i=1}^{p}\lambda_t^i\nabla L_{S_i}(w_t)+\mu_w(w_t-w')$.
  Update $w_{t+1}=\text{Proj}_{\mathcal{W}}(w_t-\eta_t\nabla_t)$.
end for
Output: $\bar{w}_N=\frac{1}{N}\sum_{t=1}^{N}w_t$ and $\bar{\lambda}_N=\frac{1}{N}\sum_{t=1}^{N}\lambda_t$.

We now provide the convergence guarantee on Algorithm 4.

Theorem 11. *Let $\eta_t=\frac{1}{\mu_w t}$, we have*

$$\max_{\lambda\in\Delta_p}F(\bar{w}_N,\lambda)-\min_{w\in\mathcal{W}}F(w,\bar{\lambda}_N)=O\left(\frac{\ln N}{N}\left(\frac{L^2}{\mu_w}+M^2\mu_w\right)\right)$$

Proof. An important observation here is that in λ_t is the best response of function $F(w_t,\cdot)$, that is,

$$\lambda_t=\arg\min_{\lambda\in\Delta_p}F(w_t,\lambda)$$

Then by Corollary 11.16 in (Orabona, 2019), we have

$$\max_{\lambda\in\Delta_p}F(\bar{w}_N,\lambda)-\min_{w\in\mathcal{W}}F(w,\bar{\lambda}_N)\le\frac{\text{Regret}_N^w}{N} \tag{7}$$

where $\text{Regret}_N^w=\sum_{t=1}^{N}l_t(w_t)-\min_{w\in\mathcal{W}}\sum_{t=1}^{N}l_t(w)$ and $l_t(w)=F(w,\lambda_t)$. Since w is updated using online gradient descent and $l_t(w)$ is strongly convex, by Corollary 4.9 in (Orabona, 2019) and $\|\nabla_t\|_2\le G+\mu_w B$, we have

$$\text{Regret}_N^w=O\left(\frac{(L+\mu_w M)^2}{\mu_w}(1+\ln N)\right)=O\left(\ln N\left(\frac{L^2}{\mu_w}+M^2\mu_w\right)\right) \tag{8}$$

Combining equations (7) and (8), we get the desired result. □

A.2. Auxiliary Lemma

Lemma 12. *Let $(\hat{w},\hat{\lambda})$ be the exact saddle point of (5) and $(\tilde{w},\tilde{\lambda})$ be an α-saddle point of (5). Then we have*

$$\|\hat{w}-\tilde{w}\|_2\le\sqrt{\frac{2\alpha}{\mu_w}}$$

Proof. Since $(\tilde{w}, \tilde{\lambda})$ is an α-saddle point of $F(w, \lambda)$, then

$$\begin{aligned} F(\tilde{w}, \hat{\lambda}) - F(\hat{w}, \tilde{\lambda}) &\leq \alpha \\ \implies F(\tilde{w}, \hat{\lambda}) - F(\hat{w}, \hat{\lambda}) + F(\hat{w}, \hat{\lambda}) - F(\hat{w}, \tilde{\lambda}) &\leq \alpha \end{aligned}$$

Since $(\hat{w}, \hat{\lambda})$ is the saddle point, then

$$F(\hat{w}, \hat{\lambda}) - F(\hat{w}, \tilde{\lambda}) \geq 0$$

Therefore,

$$F(\tilde{w}, \hat{\lambda}) - F(\hat{w}, \hat{\lambda}) \leq \alpha$$

Also we have $\hat{w} = \arg\min_{w \in \mathcal{W}} F(w, \hat{\lambda})$ and $F(\cdot, \hat{\lambda})$ is a μ_w-strongly convex function, then

$$\begin{aligned} \mu_w/2 \left\| \hat{w} - \tilde{w} \right\|_2^2 &\leq F(\tilde{w}, \hat{\lambda}) - F(\hat{w}, \hat{\lambda}) \leq \alpha \\ \implies \left\| \hat{w} - \tilde{w} \right\|_2 &\leq \sqrt{\frac{2\alpha}{\mu_w}} \end{aligned}$$

□

A.3. Proof of Lemma 3

Proof. Note that the empirical objective $\hat{\phi}(w, \lambda) = \sum_{i=1}^{p} \lambda_i L_{S_i}(w)$ can also be written in a finite sum form. For any $i \in [p]$, we denote z_i^j be the j_{th} datapoint of dataset S_i and define batched node $\xi^j = (z_1^j, \dots z_p^j)$. Hence $\tilde{S}$ can be expressed as $\tilde{S} = \{\xi^1, \dots \xi^n\}$. We also define a new loss function $f(w, \lambda, \xi)$ where $\xi = \{z_1, \dots z_p\}$ as

$$f(w, \lambda, \xi) = \sum_{i=1}^{p} \lambda_i \ell(w, z_i)$$

Then it is easy to show that

$$\hat{\phi}(w, \lambda) = \frac{1}{n} \sum_{j=1}^{n} f(w, \lambda, \xi^j) \tag{9}$$

Suppose $\tilde{S} = \{\xi_1, \dots \xi_i, \dots \xi_n\}$ and $\tilde{S}' = \{\xi_1, \dots \xi_i', \dots \xi_n\}$ where ξ_i and ξ_i' differ with single datapoint z_i^j for some $j \in [n]$. By equation (9), we have

$$\begin{aligned} \hat{\phi}(w, \lambda) &= \frac{1}{n} \sum_{\xi \in \tilde{S}} f(w, \lambda, \xi) \\ \hat{\phi}'(w, \lambda) &= \frac{1}{n} \sum_{\xi \in \tilde{S}'} f(w, \lambda, \xi) \end{aligned}$$

We also let $\Psi(w, \lambda) = \frac{\mu_w}{2} \left\| w - w' \right\|_2^2 - \mu_\lambda \sum_{j=1}^{p} \lambda_j \log \lambda_j$.

Let $(\hat{w}, \hat{\lambda})$ and $(\hat{w}', \hat{\lambda}')$ be the exact saddle point of $\hat{\phi}(w, \lambda) + \Psi(w, \lambda)$ and $\hat{\phi}'(w, \lambda) + \Psi(w, \lambda)$ respectively.

It is easy to show that $\left\| \nabla_w f(w, \lambda, \xi) \right\|_2 \leq L$ and $\left\| \nabla_\lambda f(w, \lambda, \xi) \right\|_\infty \leq B$ for any ξ. Also, $\hat{\phi}(w, \lambda) + \Psi(w, \lambda)$ is μ_w-strongly convex w.r.t $\left\| \cdot \right\|_2$ and μ_λ-strongly concave w.r.t $\left\| \cdot \right\|_1$. Then using Lemma 2 from (Zhang et al., 2021), we obtain

$$\sqrt{\mu_w \left\| \hat{w} - \hat{w}' \right\|_2^2 + \mu_\lambda \left\| \hat{\lambda} - \hat{\lambda}' \right\|_1^2} \leq \frac{2}{n} \sqrt{\frac{L^2}{\mu_w} + \frac{B^2}{\mu_\lambda}}$$

Furthermore, we have

$$\|\hat{w}-\hat{w}'\|_2 \le \sqrt{\mu_w \|\hat{w}-\hat{w}'\|_2^2 + \mu_\lambda \left\|\hat{\lambda}-\hat{\lambda}'\right\|_1^2} / \sqrt{\mu_w}$$
$$\le \frac{1}{n}\sqrt{\frac{4L^2}{\mu_w^2} + \frac{4B^2}{\mu_\lambda \mu_w}} \le \frac{2}{n}\left(\frac{L}{\mu_w} + \frac{B}{\sqrt{\mu_w \mu_\lambda}}\right)$$

Meanwhile, by Lemma 12, we also have

$$\|\tilde{w}-\hat{w}\|_2 \le \sqrt{\frac{2\alpha}{\mu_w}} \le \frac{1}{2n}\left(\frac{L}{\mu_w} + \frac{B}{\sqrt{\mu_w \mu_\lambda}}\right)$$

and

$$\|\tilde{w}'-\hat{w}'\|_2 \le \sqrt{\frac{2\alpha}{\mu_w}} \le \frac{1}{2n}\left(\frac{L}{\mu_w} + \frac{B}{\sqrt{\mu_w \mu_\lambda}}\right)$$

Putting every thing together, we have

$$\|\tilde{w}-\tilde{w}'\|_2 \le \frac{3}{n}\left(\frac{L}{\mu_w} + \frac{B}{\sqrt{\mu_w \mu_\lambda}}\right)$$

□

A.4. Proof of Lemma 4

Proof. Let $R(w) = \max_{i\in[p]} L_{D_i}(w)$ and $\hat{\phi}(w,\lambda) = \sum_{i\in[p]} \lambda_i L_{S_i}(w)$. Denote $(\hat{w},\hat{\lambda})$ the exact saddle point of $F(w,\lambda)$. By the definition of saddle point, we have for any $w \in \mathcal{W}$ and $\lambda \in \Delta_p$,

$$\hat{\phi}(\hat{w},\lambda) + \frac{\mu_w}{2}\|\hat{w}-w'\|_2^2 - \mu_\lambda \sum_{j=1}^p \lambda_j \log \lambda_j$$
$$-\left(\hat{\phi}(w,\hat{\lambda}) + \frac{\mu_w}{2}\|w-w'\|_2^2 - \mu_\lambda \sum_{j=1}^p \hat{\lambda}_j \log \hat{\lambda}_j\right) \le 0$$
$$\implies \hat{\phi}(\hat{w},\lambda) - \hat{\phi}(w,\hat{\lambda}) \le$$
$$\frac{\mu_w}{2}\|w-w'\|_2^2 - \mu_\lambda \sum_{j=1}^p \hat{\lambda}_j \log \hat{\lambda}_j - \left(\frac{\mu_w}{2}\|\hat{w}-w'\|_2^2 - \mu_\lambda \sum_{j=1}^p \lambda_j \log \lambda_j\right)$$
$$\le \frac{\mu_w}{2}\|w-w'\|_2^2 - \mu_\lambda \sum_{j=1}^p \hat{\lambda}_j \log \hat{\lambda}_j$$
$$\le \frac{\mu_w}{2}\|w-w'\|_2^2 + \mu_\lambda \log p$$

In particular, we have for any $w \in \mathcal{W}$

$$\max_{i\in[p]} L_{S_i}(\hat{w}) - \hat{\Phi}(w,\hat{\lambda}) \le \frac{\mu_w}{2}\|w-w'\|_2^2 + \mu_\lambda \log p$$

By Lemma 3, for any fixed $i \in [p]$, computing $\hat{w}$ has a uniform stability of $\gamma = \frac{2L^2}{n\mu_w} + \frac{2LB}{n\sqrt{\mu_w \mu_\lambda}}$ with respect to the change of datapoint in dataset S_i. Therefore, by fixing the randomness of $\tilde{S}_{-i} = \{S_1, \dots S_{i-1}, S_{i+1}, \dots S_p\}$ and Theorem 1.1 from (Feldman & Vondrak, 2019), we have

$$P_{S_i \sim D_i^n}\left(|L_{S_i}(\hat{w}) - L_{D_i}(\hat{w})| \ge c\left(\gamma \log(n)\log(n/\beta) + \frac{B\sqrt{\log 1/\beta}}{\sqrt{n}}\right)\right) \le \beta$$

Then we can release the randomness of $\tilde{S}_{-i}$ and obtain

$$P_{\tilde{S}}\left(|L_{S_i}(\hat{w}) - L_{D_i}(\hat{w})| \geq c\left(\gamma \log(n)\log(n/\beta) + \frac{B\sqrt{\log 1/\beta}}{\sqrt{n}}\right)\right) \leq \beta$$

Using union bound across all $i \in [p]$, we obtain with probability over $1-\beta$, for all $i \in [p]$

$$|L_{S_i}(\hat{w}) - L_{D_i}(\hat{w})| = O\left(\left(\frac{L^2}{n\mu_w} + \frac{LB}{n\sqrt{\mu_w\mu_\lambda}}\right)\log(n)\log(np/\beta) + \frac{B\sqrt{\log(p/\beta)}}{\sqrt{n}}\right)$$

Therefore, with probability $1-\beta$, we have

$$\begin{aligned}
&|\max_{i\in[p]} L_{D_i}(\hat{w}) - \max_{i\in[p]} L_{S_i}(\hat{w})| \\
&= O\left(\left(\frac{L^2}{n\mu_w} + \frac{LB}{n\sqrt{\mu_w\mu_\lambda}}\right)\log(n)\log(np/\beta) + \frac{B\sqrt{\log(p/\beta)}}{\sqrt{n}}\right)
\end{aligned}$$

Also, since w is independent of the dataset, with probability over $1-\beta$

$$|L_{S_i}(w) - L_{D_i}(w)| = O\left(\frac{B\log(p/\beta)}{\sqrt{n}}\right) \quad \forall i \in [p]$$

and

$$|\Phi(w,\hat{\lambda}) - \hat{\Phi}(w,\hat{\lambda})| = O\left(\frac{B\log(p/\beta)}{\sqrt{n}}\right)$$

After putting everything together, we obtain with probability over $1-2\beta$

$$\begin{aligned}
&R(\hat{w}) - R(w) = \max_{i\in[p]} L_{D_i}(\hat{w}) - \max_{i\in[p]} L_{D_i}(w) \\
&\leq \max_{i\in[p]} L_{D_i}(\hat{w}) - \phi(w,\hat{\lambda}) \\
&\leq \max_{i\in[p]} L_{S_i}(\hat{w}) - \hat{\phi}(w,\hat{\lambda}) + O\left(\left(\frac{L^2}{n\mu_w} + \frac{LB}{n\sqrt{\mu_w\mu_\lambda}}\right)\log(n)\log(np/\beta) + \frac{B\sqrt{\log(p/\beta)}}{\sqrt{n}}\right) \\
&= O\left(\frac{\mu_w}{2}\|w-w'\|_2^2 + \mu_\lambda \log p + \left(\frac{L^2}{n\mu_w} + \frac{LB}{n\sqrt{\mu_w\mu_\lambda}}\right)\log(n)\log(np/\beta) + \frac{B\sqrt{\log(p/\beta)}}{\sqrt{n}}\right)
\end{aligned}$$

Choosing $\beta = 1/n$ and taking expectation over the dataset collection $\tilde{S}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[R(\hat{w})\right] - R(w) \\
&= O\left(\frac{\mu_w}{2}\|w-w'\|_2^2 + \mu_\lambda \log p + \left(\frac{L^2}{n\mu_w} + \frac{LB}{n\sqrt{\mu_w\mu_\lambda}}\right)\log(n)\log(np) + \frac{B\sqrt{\log(pn)}}{\sqrt{n}}\right)
\end{aligned} \tag{10}$$

By Lemma 12, we have

$$\|\tilde{w} - \hat{w}\|_2 \leq \sqrt{\frac{2\alpha}{\mu_w}} \leq \frac{1}{2n}\left(\frac{L}{\mu_w} + \frac{B}{\sqrt{\mu_w\mu_\lambda}}\right)$$

Given the fact that $R(\cdot)$ is L-Lipschitz, we have

$$|R(\tilde{w}) - R(\hat{w})| \leq L\|\tilde{w} - \hat{w}\|_2 = O\left(\frac{L^2}{n\mu_w} + \frac{LB}{n\sqrt{\mu_w\mu_\lambda}}\right) \tag{11}$$

Combining equations (10) and (11), we have

$$\begin{aligned}&\mathbb{E}\left[R(\tilde{w})\right] - R(w)\\&= O\left(\frac{\mu_w}{2}\left\|w - w'\right\|_2^2 + \mu_\lambda \log p + \left(\frac{L^2}{n\mu_w} + \frac{LB}{n\sqrt{\mu_w\mu_\lambda}}\right)\log(n)\log(np) + \frac{B\sqrt{\log(pn)}}{\sqrt{n}}\right)\end{aligned}$$

□

A.5. Proof of Theorem 2

Proof. **Privacy:** At iteration t, by the stability argument in Lemma 3, we have

$$\begin{aligned}\left\|\tilde{w}_t - \tilde{w}_t{}'\right\|_2 &\leq \frac{3L}{n_t\mu_w^t} + \frac{3B}{n_t\sqrt{\mu_w^t\mu_\lambda^t}}\\&= 3L\eta_t + \frac{3B\sqrt{n_t n\eta_t\eta}}{n_t}\\&\leq 6D\sqrt{(\log_2 n)\eta_t\eta}\end{aligned}$$

where $\tilde{w}_t$ and $\tilde{w}_t{}'$ are outputs from neighboring dataset collections $\tilde{S}^t$ and $\tilde{S}^{t'}$ that differ in one datapoint. Since the sampled minibatches $\tilde{S}^t$ are disjoint across iterations, the privacy of Algorithm 1 follows from the privacy guarantee of the Gaussian mechanism and parallel composition.

Utility: By Lemma 4, we have

$$\begin{aligned}&\mathbb{E}[R(\tilde{w}_t) - R(\tilde{w}_{t-1})]\\&= O\left(\frac{\mathbb{E}[\|\xi_{t-1}\|_2^2]}{n_t\eta_t} + \frac{\log p}{\eta n} + D^2\sqrt{\eta\eta_t}\log^{1.5} n\log(np) + \frac{B\sqrt{\log(pn)}}{\sqrt{n_t}}\right)\end{aligned}$$

where $R(w) = \max_{i\in[p]} L_{D_i}(w)$.

Also, we have

$$\mathbb{E}\left\|\xi_t\right\|_2^2 = d\sigma_t^2 = 72dD^2\log(n)\log(1/\delta)2^{-t}\eta^2/\epsilon^2$$

Therefore, as long as

$$\eta \leq \frac{M\epsilon}{D\sqrt{72d\log(n)\log(1/\delta)}}$$

We will have

$$\mathbb{E}\left\|\xi_t\right\|_2^2 \leq 2^{-t}M^2 \text{ and } \mathbb{E}\left\|\xi_t\right\|_2 \leq \sqrt{2}^{-t}M$$

Let $\tilde{w}_0 = w^*$ and $\xi_0 = w_0 - w^*$. Then $\|\xi_0\|_2 \leq M$. Hence

$$\begin{aligned}&\mathbb{E}[R(w_T)] - R(w^*) = \sum_{t=1}^{T}\mathbb{E}[R(\tilde{w}_t) - R(\tilde{w}_{t-1})] + \mathbb{E}[R(w_T) - R(\hat{w}_T)]\\&\leq \sum_{t=1}^{T} O\left(\frac{\mathbb{E}[\|\xi_{t-1}\|_2^2]}{n_t\eta_t} + \frac{\log p}{\eta n} + D^2\sqrt{\eta\eta_t}\log^{1.5} n\log(np) + \frac{B\sqrt{\log(n)\log(pn)}}{\sqrt{n}}\right)\\&\quad + L\mathbb{E}[\|\xi_T\|_2] \quad (R(\cdot) \text{ is } L\text{-Lipschitz.})\end{aligned}$$

We have

$$\frac{\mathbb{E}[\|\xi_{t-1}\|_2^2]}{n_t\eta_t} \leq \frac{2^{-(t-1)}M^2\log_2 n}{n2^{-t}\eta} = \frac{2(\log_2 n)M^2}{\eta n}$$

and $\mathbb{E}[\|\xi_T\|_2] \leq M\sqrt{2}^{-\log_2 n} = \frac{M}{\sqrt{n}}$.

Therefore,

$$
\begin{aligned}
&\mathbb{E}[R(w_T)] - R(w^*) \\
&\leq \sum_{t=1}^{T} O\left(\frac{\log(n)M^2}{\eta n} + \frac{\log(p)}{\eta n} + D^2\eta\log^{1.5}(n)\log(np) + \frac{B\sqrt{\log(n)\log(pn)}}{\sqrt{n}}\right) \\
&\quad + \frac{ML}{\sqrt{n}} \\
&= \log n \cdot O\left(\frac{(\log(np))M^2}{\eta n} + D^2\eta\log^{2.5}(np) + \frac{B\sqrt{\log(n)\log(np)}}{\sqrt{n}}\right) \\
&\quad + \frac{ML}{\sqrt{n}} \\
&= O\left(\frac{MD\log^{11/4}(np)}{\sqrt{n}} + (MD\log^{5/2}(np))\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)
\end{aligned}
$$

Replacing n with K/p gives the desired result. □

B. Missing Proofs in Section 4

B.1. Proof of Theorem 7

Proof. It is easy to see that the objective in (3) is equivalent to

$$\min_{w\in\mathcal{W}} \max_{\lambda\in\Delta_p} \sum_{i\in[p]} \lambda_i L_{D_i}(w)$$

where $\Delta_p = \{\lambda \in [0,1]^p : \|\lambda\|_1 = 1\}$ the probability simplex over p users. Recall $\phi(w,\lambda) = \sum_{i\in[p]} \lambda_i L_{D_i}(w)$, then we have

$$
\begin{aligned}
&\max_{\lambda\in\Delta_p} \phi(\bar{w}_T,\lambda) - \min_{w\in\mathcal{W}} \max_{\lambda\in\Delta_p} \phi(w,\lambda) \\
&= \max_{\lambda\in\Delta_p} \phi(\bar{w}_T,\lambda) - \max_{\lambda\in\Delta_p} \min_{w\in\mathcal{W}} \phi(w,\lambda) \\
&\leq \frac{1}{T} \max_{w\in\mathcal{W},\lambda\in\Delta_p} \left\{\sum_{t=1}^{T} \phi(w_t,\lambda) - \phi(w,\lambda_t)\right\} \\
&= \frac{1}{T} \max_{w\in\mathcal{W},\lambda\in\Delta_p} \left\{\sum_{t=1}^{T} (\phi(w_t,\lambda) - \phi(w_t,\lambda_t) + \phi(w_t,\lambda_t) - \phi(w,\lambda_t))\right\} \\
&= \frac{1}{T}\left[\sum_{t=1}^{T} \phi(w_t,\lambda_t) - \min_{w\in\mathcal{W}} \sum_{t=1}^{T} \phi(w,\lambda_t)\right] \quad (A) \\
&\quad + \frac{1}{T}\left[\sum_{t=1}^{T} -\phi(w_t,\lambda_t) - \min_{\lambda\in\Delta_p} \sum_{t=1}^{T} (-\phi(w_t,\lambda))\right] \quad (B)
\end{aligned}
$$

We address the term A first. By fixing the randomness of w_t and λ_t for some iteration t, we have $\mathbf{E}_{x_t^-}\ell(w_t, x_t^-) = \phi(w_t,\lambda_t)$. By extending the same argument to all $t \in [T]$, we obtain for any w,

$$\frac{1}{T}\mathbf{E}\left[\sum_{t=1}^{T} (\phi(w_t,\lambda_t) - \phi(w,\lambda_t))\right] = \frac{1}{T}\mathbf{E}\left[\sum_{t=1}^{T} \ell(w_t,x_t^-) - \ell(w,x_t^-)\right]$$

By the regret guarantee of $\mathcal{Q}_-$, we have for any sequence $x_1^-, \ldots x_T^-$

$$\frac{1}{T}\mathbf{E}\left[\sum_{t=1}^{T} \ell(w_t,x_t^-) - \ell(w,x_t^-)\right] \leq r_{Q_-}(T)$$

where the expectation is taken from the randomness over the algorithm. Therefore, we have

$$\mathbf{E}[A] \leq r_{Q_-}(T) \tag{12}$$

We then address the term B. We first define $\ell'(w, x)$ as the difference between some fixed constant U and the original loss function $\ell(w, x)$, that is, $\ell'(w, x) = U - \ell(w, x)$. Similarly the vector of p population risks over the new loss function ℓ' is therefore written as $g(w) = [L'_{D_1}(w), \dots L'_{D_p}(w)]$. Therefore, the term B can be equivalently written as

$$B = \frac{1}{T}\left[\sum_{t=1}^{T}\langle g(w_t), \lambda_t\rangle - \min_{\lambda\in\Delta^p}\sum_{t=1}^{T}\langle g(w_t), \lambda\rangle\right]$$

Notice that at each iteration t, we have

$$\nabla_\lambda\langle g(w_t), \lambda\rangle = g(w_t)$$

We construct a unbiased estimator of $g(w_t)$ as

$$g'(w_t) = \begin{cases} \frac{\tilde{\ell}_t}{\lambda_t^i} & i = j_t \\ 0 & o.w. \end{cases}$$

By letting $U = B + \frac{2B}{\epsilon}\log(T)$, we have we have w.p. over $1 - \frac{1}{T}$, $\max_{t\in[T]}|y_t| \leq \frac{2B}{\epsilon}\log(T)$, which we denote as event E. Conditioned on E, we have $\tilde{l}_t \in [0, 2U]$ for any $t \in [T]$.

Observing that for any fixed w_t, since $x_t^+ \sim D_{i_t}$, we obtain

$$\mathbf{E}[g'(w_t)|E] = g(w_t)$$

We now establish the bound on term B

$$\begin{aligned}&\mathbf{E}\left[\sum_{t=1}^{T}\langle g(w_t), \lambda_t\rangle - \sum_{t=1}^{T}\langle g(w_t), \lambda\rangle\right]\\ &= \mathbf{E}\left[\sum_{t=1}^{T}\langle g'(w_t), \lambda_t\rangle - \sum_{t=1}^{T}\langle g'(w_t), \lambda\rangle\right]\end{aligned}$$

Since $\|g'(w_t)\|_\infty \leq 2U$ conditioned on event E, by setting $\eta = \sqrt{\frac{\ln(p)}{pTU^2}}$ and invoking the regret bound of EXP3, we obtain

$$\begin{aligned}\mathbf{E}[B|E] &= \frac{1}{T}\mathbf{E}\left[\sum_{t=1}^{T}\langle g(w_t), \lambda_t\rangle - \sum_{t=1}^{T}\langle g(w_t), \lambda\rangle\right]\\ &= O\left(U\sqrt{\frac{p\log(p)}{T}}\right)\end{aligned}$$

Then we take the expectation on E and obtain

$$\mathbf{E}[B] = P(E)\mathbf{E}[B|E] + (1 - P(E))\mathbf{E}[B|E^c] = O\left(U\sqrt{\frac{p\log(p)}{T}}\right)$$

Finally we plug in the value of U,

$$\mathbf{E}[B] = O\left(\left(B + \frac{B\log(T)}{\epsilon}\right)\sqrt{\frac{p\log(p)}{T}}\right) \tag{13}$$

We combine equations (12) and (13) to get the desired result. □

C. Missing Proofs from Section 5

C.1. Proof of Theorem 9 and 10

Proof. (Proofs of Theorem 9) Note that given two neighboring data collections $\tilde{S} = \{S_1, \dots S_p\}$ and $\tilde{S}' = \{S'_1, \dots S'_p\}$ that differ in one datapoint of j_{th} dataset for some $j \in [p]$. In iteration t, if $i_t \neq j$, then the output distribution of w_{t+1} is same for $\tilde{S}$ and $\tilde{S}'$. Otherwise, generating w_{t+1} satisfies $(\frac{\epsilon}{c\sqrt{T\log(1/\delta)}}, \frac{\delta}{2T})$-DP based on the property of Gaussian mechanism and privacy amplification by subsampling. Therefore, overall generating w_{t+1} is $(\frac{\epsilon}{c\sqrt{T\log(1/\delta)}}, \frac{\delta}{2T})$-DP.

Meanwhile, generating λ_{t+1} is $(\frac{\epsilon}{c\sqrt{T\log(1/\delta)}}, 0)$-DP based on the property of Laplace mechanism. Using basic composition, we have generating the pair (w_{t+1}, λ_{t+1}) is $(\frac{2\epsilon}{c\sqrt{T\log(1/\delta)}}, \frac{\delta}{2T})$-DP. Hence, by strong composition for $t \in [T]$, we obtain the algorithm is (ϵ, δ)-DP. □

Proof. (Proof of Theorem 10) Recall that $n = K/p$. We have

$$
\begin{aligned}
&\max_{\lambda\in\Delta_p} \hat{\phi}(\bar{w}_T, \lambda) - \min_{w\in\mathcal{W}} \max_{\lambda\in\Delta_p} \hat{\phi}(w, \lambda) \\
&\leq \max_{\lambda\in\Delta_p} \hat{\phi}(\bar{w}_T, \lambda) - \min_{w\in\mathcal{W}} \hat{\phi}(w, \bar{\lambda}_T) \\
&\leq \max_{\lambda\in\Delta_p, w\in\mathcal{W}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \Big(\hat{\phi}(w_t, \lambda) - \hat{\phi}(w, \lambda_t) \Big) \right\} \\
&= \max_{\lambda\in\Delta_p, w\in\mathcal{W}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \Big(\hat{\phi}(w_t, \lambda) + \hat{\phi}(w_t, \lambda_t) - \hat{\phi}(w_t, \lambda_t) - \hat{\phi}(w, \lambda_t) \Big) \right\} \\
&= \frac{1}{T}\left[\sum_{t=1}^{T} \hat{\phi}(w_t, \lambda_t) - \min_{w\in\mathcal{W}} \sum_{t=1}^{T} \hat{\phi}(w, \lambda_t) \right] \quad (A) \\
&\quad + \frac{1}{T}\left[\sum_{t=1}^{T} -\hat{\phi}(w_t, \lambda_t) - \min_{\lambda\in\Delta_p} \sum_{t=1}^{T} (-\hat{\phi}(w_t, \lambda)) \right] \quad (B)
\end{aligned}
$$

Next, we will bound the terms A and B separately. We address term A first. In particular, for any $w \in \mathcal{W}$, we can rewrite the expectation of term A as

$$
\begin{aligned}
\mathbb{E}[A] &= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \Big(\hat{\phi}(w_t, \lambda_t) - \hat{\phi}(w, \lambda_t) \Big) \right] \\
&\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla\hat{\phi}(w_t, \lambda_t), w_t - w \rangle \right]
\end{aligned}
$$

For any iteration t, by fixing the randomness up to w_t and λ_t and letting $\nabla_t = \frac{1}{m}\sum_{z\in B_t} \nabla\ell(w_t, z) + G_t$ be the gradient update in step 6 in Algorithm 3, we have

$$
\mathbb{E}[\nabla_t] = \nabla\hat{\phi}(w_t, \lambda_t)
$$

We can extend the same augment for all $t \in [T]$ and obtain

$$\begin{aligned}
\mathbb{E}[A] &\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla\hat{\phi}(w_t,\lambda_t), w_t - w\rangle\right] \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla_t, w_t - w\rangle\right] \\
&\leq \frac{M^2}{2T\eta} + \frac{\eta}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_t\|^2\right] \\
&\leq \frac{\eta L^2}{2} + \frac{M^2}{2T\eta} + \frac{d\sigma^2\eta}{2}
\end{aligned}$$

where the first equality holds by standard convex analysis.

Meanwhile, we let $U = B + \frac{8B}{n\epsilon}\sqrt{T\log(1/\delta)}\log(pnT/2)$, and denote $L'_t = [U - L_{S_i}(w_t) + y_{i,t}]_{i=1}^{p}$. By using a union bound, we have $\|L'(t)\|_\infty \leq 2U$ for all $t \in [T]$ with probability over $1 - \frac{1}{n}$, which we denote as event E. Conditioned on event E, our update rule for λ can be seen as applying the Hedge algorithm with L'_t as the loss input and the term B is formulated as its corresponding regret bound. Therefore, by the regret bound of Hedge , we obtain

$$\mathbb{E}[B|E] = O\left(\|L_t\|_\infty\sqrt{\frac{\log(p)}{T}}\right) = O\left(B\sqrt{\frac{\log(p)}{T}} + \frac{B\sqrt{\log(p)\log(1/\delta)}\log(pnT)}{n\epsilon}\right)$$

Then we can take the expectation over E and have

$$\mathbb{E}[B] = P(E)\mathbb{E}[B|E] + P(E^c)\mathbb{E}[B|E^c] = O\left(B\sqrt{\frac{\log(p)}{T}} + \frac{B\sqrt{\log(p)\log(1/\delta)}\log(pnT)}{n\epsilon}\right)$$

By combining the bounds of $\mathbb{E}[A]$ and $\mathbb{E}[B]$ and plugging into the values of T, η_w and η_λ, we obtain the desired result. □

C.2. Private Active Group Selection

Here we present another algorithm built upon the active group selection scheme described in Algorithm 5. Algorithm 5 is different from Algorithm 3 on the group selection method. Instead of maintaining a weight vector to sample the dataset as in Algorithm 3, we select the dataset with highest loss in each iteration. To maintain privacy, we resort to the report-noisy-max mechanism for the group selection. After such (approximately) worst-off group is selected, we sample a mini-batch from it and use Noisy-SGD to update the model parameters.

Algorithm 5 Noisy SGD with active group selection (Noisy-SGD-AGS)

InputCollection of datasets $S = \{S_1, \ldots S_p\} \in \mathcal{Z}^{n\times p}$, mini-batch size m, # iterations T, learning rate η
for $t = 1, \ldots, T-1$ **do**
  Compute $i_t = \arg\max_{i\in[p]} L_{S_i}(w_t) + y_{i,t}$ where $y_{i,t} \overset{\text{iid}}{\sim} \text{Lap}(\tau)$.
  Sample $B_t = \{z_1, \ldots z_m\}$ from S_{i_t} uniformly with replacement.
  Update the model by

$$w_{t+1} = \text{Proj}_{\mathcal{W}}\left(w_t - \eta\cdot\left(\frac{1}{m}\sum_{z\in B_t}\nabla\ell(w_t, z) + G_t\right)\right).$$

  where $G_t \sim \mathcal{N}(0, \sigma^2 I_d)$.
end for
Output $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$

Theorem 13. *(Privacy guarantee) Algorithm 5 is (ϵ, δ)-DP with $\tau = \frac{cBp}{K\epsilon}\sqrt{T\log(1/\delta)}$ and $\sigma^2 = \frac{cTL^2q^2\log(K/\delta)\log(1/\delta)}{K^2\epsilon^2}$ for some universal constant c.*

.

Theorem 14. *(Convergence rate) with probability over $1-\kappa$, by letting $T = \frac{MLK\epsilon}{16Bp\sqrt{\log(1/\delta)}}$ and $\eta = \frac{M}{\sqrt{T(G^2+d\sigma^2)}}$, we have*

$$\mathbb{E}[\max_{i\in[p]} L_{S_i}(\bar{w}_T)] - \min_{w\in\mathcal{W}} \max_{i\in[p]} L_{S_i}(w) = O\left(\left(\sqrt{\frac{MLBp}{K\epsilon}} + \frac{MLp\sqrt{d}}{K\epsilon}\right)\cdot \mathrm{polylog}(p,\delta,\kappa,n)\right).$$

Proof. By the union bound, we have with probability over $1-\kappa$

$$|y_{i,t}| \le \frac{8B}{n\epsilon}\sqrt{T\log(1/\delta)}\log(pT/2\kappa)$$

for all $i \in [p]$ and $t \in [T]$. We denote this as event A and condition on A. By denoting $R(\kappa,T) = \frac{8B}{n\epsilon}\sqrt{T\log(1/\delta)}\log(pT/2\kappa)$, we have

$$\begin{aligned}
&\mathbb{E}[\max_{i\in[q]} L_{S_i}(\bar{w}_T)] \\
&\le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\max_{i\in[q]} L_{S_i}(w_t)\right] \quad \text{(convexity of } \max_i L_{S_i}) \\
&\le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} L_{S_{i_t}}(w_t)\right] + R(\kappa,T) \quad \text{(Event A)}
\end{aligned}$$

For any fixed w^*, by the convexity of the loss function, we obtain

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(L_{S_{i_t}}(w_t) - L_{S_{i_t}}(w^*))\right] \le \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla L_{S_{i_t}}(w_t), w_t - w^*\rangle\right]$$

Denote $\nabla_t = \frac{1}{m}\sum_{z\in B_t}\nabla\ell(w_t,z) + G_t$ where B_t is the mini-batch uniformly sampled from S_{i_t} and $G_t \sim \mathcal{N}(0,\sigma^2 I_d)$. If we fix the randomness of w_t and i_t, it is easy to see that

$$\mathbb{E}[\nabla_t] = \nabla L_{S_{i_t}}(w_t)$$

where the randomness comes from the datapoint sampling and added Gaussian noise.

By releasing the randomness of w_t and i_t and extending the same analysis to all $t \in [T]$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(L_{S_{i_t}}(w_t) - L_{S_{i_t}}(w^*)\right)\right] &\le \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla L_{S_{i_t}}(w_t), w_t - w^*\rangle\right] \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla_t, w_t - w^*\rangle\right] \\
&\le \frac{M^2}{2T\eta} + \frac{\eta}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_t\|^2\right] \\
&\le \frac{\eta L^2}{2} + \frac{M^2}{2T\eta} + \frac{d\sigma^2\eta}{2}
\end{aligned}$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}[\max_{i\in[q]} L_{S_i}(\bar{w}_T)] \\
&\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} L_{S_{i_t}}(w_t)\right] + R(\kappa, T) \quad \text{(Event A)} \\
&\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} L_{S_{i_t}}(w^*)\right] + \frac{\eta L^2}{2} + \frac{M^2}{2T\eta} + \frac{d\sigma^2\eta}{2} + R(\kappa, T) \\
&\leq \max_{i\in[q]} L_{S_i}(w^*) + \frac{\eta L^2}{2} + \frac{M^2}{2T\eta} + \frac{d\sigma^2\eta}{2} + R(\kappa, T)
\end{aligned}
$$

By plugging into the value of η, σ^2 and T and replacing n with K/p, we get the desired result. □

D. Lower Bounds for the Offline Setting

Consider a L-Lipschitz convex loss function $\ell(w, z)$ bounded by $[0, B]$ for any $w \in \mathcal{W}$ and $z \in \mathcal{Z}$. We create a new loss function $\ell'(w, (z, y)) = \ell(w, z) + y$ where $y \in [0, B]$. It is easy to see that $\ell'(\cdot, (z, y))$ is also convex and L-Lipschitz.

Empirical case We let $y = B$ for all datapoints in $S_1 = \{(z_1, y_1), \ldots (z_n, y_n)\}$ and $y = 0$ for all datapoints in S_i with $i \neq 1$. Therefore, it is easy to see that for any $w \in \mathcal{W}$ and $i \in [p]$

$$L'_{S_1}(w) \geq L'_{S_i}(w)$$

where $L'_{S_i}(w) = \frac{1}{n}\sum_{(z,y)\in S_i} \ell'(w, (z, y))$.

Then we have

$$\min_{w\in\mathcal{W}} \max_{i\in[p]} L'_{S_1}(w) = \min_{w\in\mathcal{W}} L'_{S_1}(w)$$

Then the original worst-group empirical risk minimization problem is reduced to single-distribution DP-ERM problem which is known to be lower bounded by $\Omega\left(\frac{\sqrt{d}}{n\epsilon}\right)$ for Lipschitz convex loss (Bassily et al., 2014). Therefore, a lower bound of the excess worst-group empirical risk is also $\Omega\left(\frac{\sqrt{d}}{n\epsilon}\right)$. Since $n = K/p$, this lower bound can also be written as $\Omega\left(\frac{p\sqrt{d}}{K\epsilon}\right)$.

Population case We let $y = B$ when $(z, y) \sim D_1$ and $y = 0$ when $(z, y) \sim D_i$ with $i \neq 1$. The distribution of z is arbitrary. Therefore, we have for any $w \in \mathcal{W}$ and $i \in [p]$

$$L'_{D_1}(w) \geq L'_{D_i}(w)$$

and

$$\min_{w\in\mathcal{W}} \max_{i\in[p]} L'_{D_i}(w) = \min_{w\in\mathcal{W}} L'_{D_1}(w)$$

Similar to the empirical case, we can reduce the worst-group population risk minimization problem to single-distribution DP-SCO problem which is lower bounded by $\Omega\left(\frac{\sqrt{d}}{n\epsilon} + \frac{1}{\sqrt{n}}\right)$ (Bassily et al., 2019). Therefore a lower bound of the population excess worst-group risk is also $\Omega\left(\frac{\sqrt{d}}{n\epsilon} + \frac{1}{\sqrt{n}}\right)$. Since $n = K/p$, this lower bound can be also written as $\Omega\left(\frac{p\sqrt{d}}{K\epsilon} + \sqrt{\frac{p}{K}}\right)$.