Differentially Private Worst-group Risk Minimization

Xinyu Zhou, Raef Bassily

The Ohio State University

ICML 2024

Worst-group Risk Minimization

Given p distributions $\{D_1,\ D_2, \ldots D_p\}$, the worst-group **population** risk minimization problem is defined as

$$\min_{w\in W}\max_{i\in[p]} \{L_{D_i}(w) \triangleq \mathrm{E}_{z\sim D_i}\ell(w,z)\}$$

Equivalently $$\min_{w\in W}\max_{\lambda\in\Delta_p} \left\{\phi(w,\lambda) \triangleq \sum_{i\in[p]} \lambda_i\, L_{D_i}(w)\right\}$$

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

$$\min_{w\in W}\max_{i\in[p]} \{L_{S_i}(w) \triangleq \frac{1}{|S_i|} \sum_{z\sim S_i} \ell(w,z)\}$$

Equivalently $$\min_{w\in W}\max_{\lambda\in\Delta_p} \left\{\hat{\phi}(w,\lambda) \triangleq \sum_{i\in[p]} \lambda_i\, L_{S_i}(w)\right\}$$

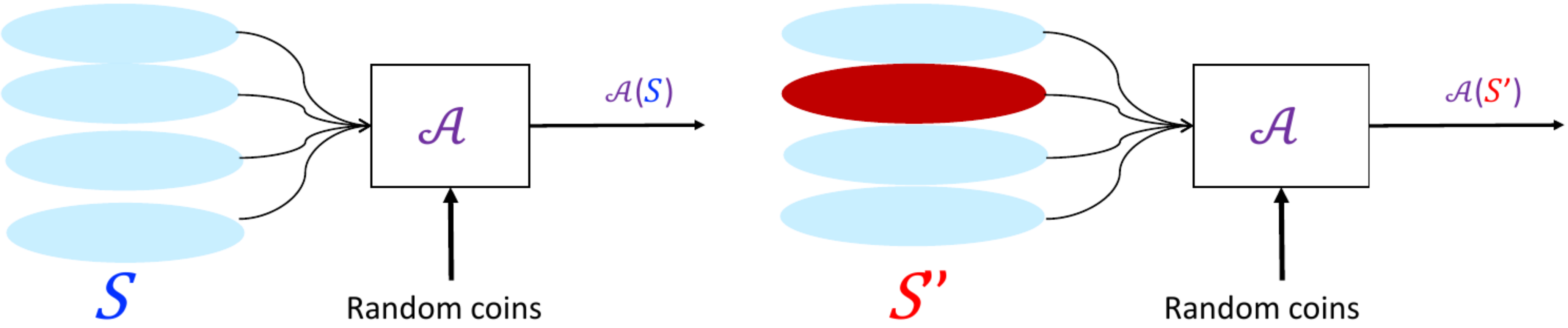
- $\ell(w,z)$ is a convex, L-Lipschitz loss function bounded by B
- Each group distribution is accessed via a **sample oracle**

Applications

- Robust Learning
  - find a model that works well for all group distributions
- Learning with fairness
  - prevent the learner from overfitting to certain groups at the cost of others
- Collaborative learning, Agnostic federated learning

And more...

Differential Privacy (DP)



Differential Privacy (DP)

A randomized algorithm A is said to be (ϵ, δ)-DP if for any pair of datasets S and S' differing in one point and any event O in the range of A, it holds that

$$P(A(S) \in O) \leq e^{\epsilon} P(A(S') \in O) + \delta$$

Objective

- We study the worst-group risk minimization problem under (ϵ, δ)-DP
- Given W, a convex and compact subset of $\mathbb{R}^d$, the goal is to privately find a model $w \in W$ with small
  - excess worst-group **population** risk

$$\mathcal{E}\left(w, \{D_i\}_{i=1}^{p}\right) = \max_{i\in[p]} L_{D_i}(w) - \min_{\widetilde{w}\in W} \max_{i\in[p]} L_{D_i}(\widetilde{w})$$

  - excess worst-group **empirical** risk

$$\widehat{\mathcal{E}}\left(w, \{S_i\}_{i=1}^{p}\right) = \max_{i\in[p]} L_{S_i}(w) - \min_{\widetilde{w}\in W} \max_{i\in[p]} L_{S_i}(\widetilde{w})$$

Contributions

- We give two algorithms for *DP worst-group population risk minimization*
  - **Minimax phased ERM** that attains $\tilde{O}\left(\sqrt{\frac{p}{K}} + \frac{p\sqrt{d}}{K\epsilon}\right)$ rate.
    - This rate is **optimal** in the offline setting.
  - **DP-OCO approach** that attains $\tilde{O}\left(\sqrt{\frac{p}{K}} + \sqrt{\frac{p}{K\epsilon^2}} + \sqrt{\frac{d^{1/2}}{K\epsilon}}\right)$ rate.
- We give an algorithm for *DP worst-group empirical risk minimization* that attains **nearly optimal** rate of $\tilde{O}\left(\frac{p\sqrt{d}}{K\epsilon}\right)$.

K: total number of samples from all groups
p: number of groups
d: problem dimension

Minimax Phased ERM – Stability Lemma

Regularized ERM objective: given arbitrary $w' \in W$ and dataset collection $\{S_i\}_{i=1}^{p} \in Z^{n \times p}$

$$F(w,\lambda) = \sum_{i=1}^{p} \lambda_i L_{S_i}(w) + \frac{\mu_w}{2}||w - w'||^2 - \mu_\lambda \sum_{i=1}^{p} \lambda_i \log \lambda_i$$

Stability lemma:

Let $(\tilde{w}, \tilde{\lambda})$ be the saddle point of $F(w,\lambda)$. For any $w \in W$, we have

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

Minimax Phased ERM - Overview

- Set $n = K/p, T = \log(n)$, and $\eta = \tilde{O}\left(\min\{\epsilon/\sqrt{d}, \sqrt{p/K}\}\right)$
- At iteration $t = 1, \ldots T$:
  1. Let $n_t = n/T, \eta_t = \eta 2^{-t}, \mu_w^t = 1/(\eta_t n_t)$ and $\mu_\lambda^t = 1/(\eta n)$
  2. Sample $\tilde{S}_t = \{S_1, \ldots S_p\} \in Z^{n_t \times p}$ from the sample oracles
  3. Solve for the approximate saddle point $(\tilde{w}_t, \tilde{\lambda}_t)$ of
  $$F_t(w, \lambda) = \sum_{i=1}^{p} \lambda_i L_{S_i}(w) + \frac{\mu_w^t}{2} ||w - w_{t-1}||^2 - \mu_\lambda^t \sum_{i=1}^{p} \lambda_i log \lambda_i$$
  4. Obtain $w_t = \tilde{w}_t + N(0, \sigma_t^2 I)$ with $\sigma_t = O\left(\frac{\sqrt{\log(n)\log(1/\delta)}\eta\eta_t}{\epsilon}\right)$
- Output w_T

With properly chosen parameters, the algorithm is (ϵ, δ)-DP and we have

$$\mathrm{E}\left[\max_{i\in[p]} L_{D_i}(w_T)\right] - \min_{w\in W}\max_{i\in[p]} L_{D_i}(w) = \tilde{O}\left(\sqrt{\frac{p}{K}} + \frac{p\sqrt{d}}{K\epsilon}\right)$$

Optimal in the offline setting

Optimal non-private rate

Cost of privacy

DP-OCO Based Algorithm

Algorithm overview:

- Cast the objective $\min_{w \in W} \max_{\lambda \in \Delta} \phi(w, \lambda)$ into a two-player zero-sum game.
- min-player: any generic DP-OCO algorithm.
- max-player: adversarial multi-armed bandit algorithm (EXP3 [2]) with privatized gradient estimate.
- One can show that the expected excess risk is bounded by the sum of the regrets of both players.

By instantiating the DP-OCO algorithm with DP-FTRL in [3], our algorithm is (ϵ, δ)-DP and

$$\mathrm{E}\left[\max_{i \in [p]} L_{D_i}(w_T)\right] - \min_{w \in W} \max_{i \in [p]} L_{D_i}(w) = \tilde{O}\left(\sqrt{\frac{p}{K}} + \sqrt{\frac{p}{K\epsilon^2}} + \sqrt{\frac{d^{1/2}}{K\epsilon}}\right)$$

- Match the non-private optimal rate when $d = \tilde{O}(p^2)$.

Worst-group Empirical risk minimization

Based on a private version of the multiplicative group reweighting method [4].

- At iteration $t = 1, \ldots T$:
  1. Sample $i_t \sim \lambda_t$ and a minibatch B_t from S_{i_t}.
  2. Update $w_{t+1} = \mathrm{Noisy}y\mathrm{SGD}(w_t, B_t)$
  3. Privatized losses $L_t = \left\{ L_{S_t}(w_t) + \mathrm{Lap}\left(\frac{p}{K\epsilon}\sqrt{T\log(1/\delta)}\right)\right\}$
  4. Update $\lambda_{t+1} = \lambda_t \exp(-\eta L_t)$
- Output $\overline{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$

Excess worst-group empirical risk: $\tilde{O}\left(\frac{p\sqrt{d}}{K\epsilon}\right)$

- The rate is nearly **optimal**.

Reference:

[1] 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.

[2] 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.

[3] 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.

[4] 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.

Thanks!