Learning from Multiple Distributions: GDRO and MERO

Lijun Zhang

Nanjing University, China

LAMDA & RIKEN-AIP Joint Workshop on Machine Learning

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- Stochastic Approximation of MERO
  - A Multi-Stage SA Approach
  - An Anytime SA Approach
- Conclusion

Statistical Machine Learning

➢ Risk Minimization

$$\min_{\mathbf{w}\in\mathcal{W}} \left\{ R_0(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_0}\left[\ell(\mathbf{w};\mathbf{z})\right] \right\}$$

- $\mathbf{w}$ denotes the learning model, $\mathbf{z}$ is a random sample
- $\mathcal{P}_0$ is a unknown distribution, $\ell(\cdot;\cdot)$ is a loss function

➢ Examples

- SVM $\quad \min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}_{(\mathbf{x},y)\sim\mathcal{P}_0}\left[\max(1-y\mathbf{w}^\top\mathbf{x},0)\right] + \frac{\lambda}{2}\|\mathbf{w}\|^2$
- Linear Regression $\quad \min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}_{(\mathbf{x},y)\sim\mathcal{P}_0}\left[(y-\mathbf{w}^\top x)^2\right]$

Optimization Approaches I

I. Sample Average Approximation (SAA)

I. Empirical Risk Minimization (ERM)

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{w};\mathbf{z}_i)$$

- $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are independently sampled from $\mathcal{P}_0$

☐ Deterministic Optimization

✓ Gradient Descent, Mirror Descent, Newton's method

☐ Stochastic Optimization

✓ Stochastic Gradient Descent, Stochastic Mirror Descent

✓ Variance Reduction (Johnson and Zhang, 2013; Zhang et al., 2013)

Optimization Approaches II

II. Stochastic Approximation (SA)

$$\min_{\mathbf{w}\in\mathcal{W}} \left\{ R_0(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_0}\left[\ell(\mathbf{w};\mathbf{z})\right] \right\}$$

- Stochastic Gradient Descent (SGD)

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\left[\mathbf{w}_t - \eta\nabla\ell(\mathbf{w}_t;\mathbf{z}_t)\right], \quad \mathbf{z}_t \sim \mathcal{P}_0$$

✓ The stochastic gradient is unbiased

$$\mathrm{E}\left[\nabla\ell(\mathbf{w}_t;\mathbf{z}_t)\right] = \nabla R_0(\mathbf{w}_t)$$

At least in theory, we cannot reuse samples!

Optimization Approaches II

II. Stochastic Approximation (SA)

$$\min_{\mathbf{w}\in\mathcal{W}} \left\{ R_0(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_0}\left[\ell(\mathbf{w};\mathbf{z})\right] \right\}$$

- Stochastic Gradient Descent (SGD)

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\left[\mathbf{w}_t - \eta\nabla\ell(\mathbf{w}_t;\mathbf{z}_t)\right], \quad \mathbf{z}_t \sim \mathcal{P}_0$$

- Stochastic Mirror Descent (SMD) (Nemirovski et al., 2009)

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \left\{ \eta\langle\nabla\ell(\mathbf{w}_t;\mathbf{z}_t), \mathbf{w}-\mathbf{w}_t\rangle + B(\mathbf{w},\mathbf{w}_t) \right\}$$

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

✓ SMD becomes SGD when $v(\mathbf{w}) = \|\mathbf{w}\|^2/2$

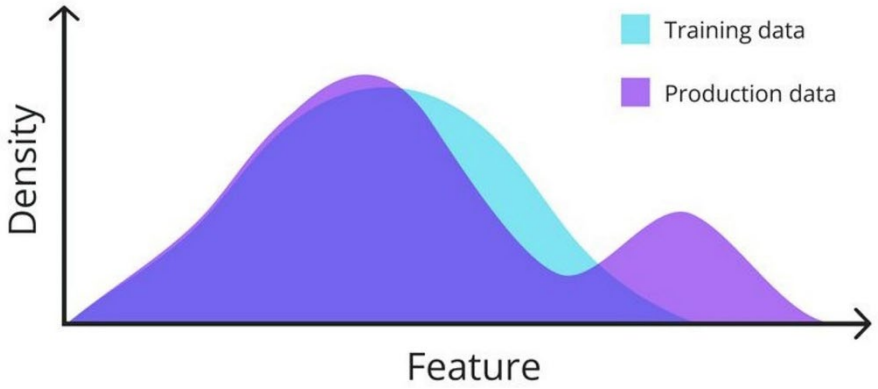
Statistical Machine Learning

- Theoretical Guarantee
  - SAA and SA

$$\underbrace{R_0(\bar{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} R_0(\mathbf{w})}_{\text{Excess Risk}} = O\left(\frac{1}{\sqrt{n}}\right), \quad O\left(\frac{1}{n}\right)$$

- Limitations

Lack robustness when distribution shifts



https://www.nannyml.com/blog/6-ways-to-address-data-distribution-shift

Distributionally Robust Optimization (DRO)

➢ Formulation of DRO

$$\min_{\mathbf{w}\in\mathcal{W}} \sup_{\mathcal{P}\in\mathcal{S}(\mathcal{P}_0)} \left\{ \mathrm{E}_{\mathbf{z}\sim\mathcal{P}} \left[\ell(\mathbf{w};\mathbf{z}) \right] \right\}$$

- $\mathcal{S}(\mathcal{P}_0)$ denotes a set of probability distributions around $\mathcal{P}_0$

➢ A Large Amount of Literature

- Robust optimization (Scarf, 1958; Ben-Tal et al., 2009)
- Asymptotic properties (Duchi and Namkoong, 2021)
- Constructions of the neighborhood (Delage and Ye, 2010; Ben-Tal et al., 2013; Esfahani and Kuhn, 2018)
- Optimization techniques (Namkoong and Duchi, 2016; Levy et al., 2020; Qi et al., 2021; Rafique et al., 2022)

Group DRO (Sagawa et al. 2020)

➢ Formulation: Minimax Risk Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ R_i(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right] \right\}$$

- A finite number of m distributions

☐ A new way for learning from multiple distributions

➢ Advantage: More Robust

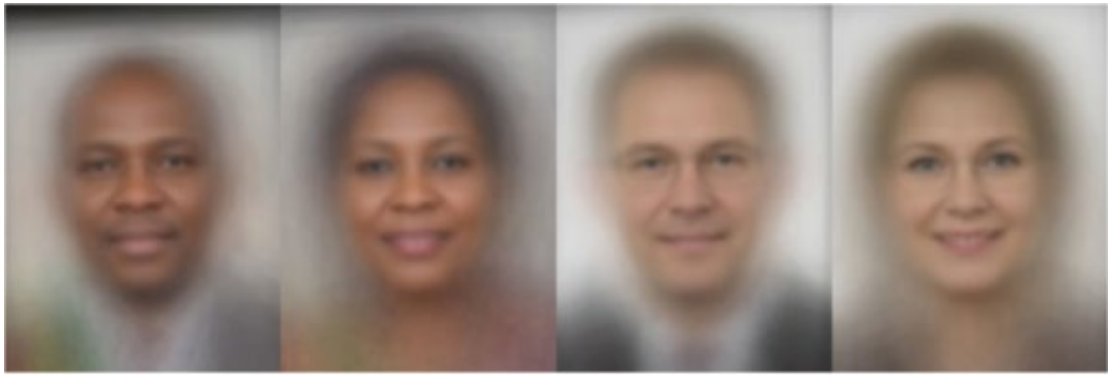
☐ A naïve approach

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{m}\sum_{i=1}^{m} R_i(\mathbf{w})$$

Application: Fairness

➢ Gender Classification (Buolamwini and Gebru 2018)

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



High accuracy for lighter-skinned males, but worse accuracy for darker-skinned females

Solution ⇩

Optimizing performance across all groups

https://stanford-cs221.github.io/autumn2022-extra/modules/machine-learning/group-dro.pdf

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- Stochastic Approximation of MERO
  - A Multi-Stage SA Approach
  - An Anytime SA Approach
- Conclusion

Related Work I

➢ The Seminal Work of Sagawa et al. (ICLR 2020)

Shiori Sagawa*
Stanford University
ssagawa@cs.stanford.edu

Pang Wei Koh*
Stanford University
pangwei@cs.stanford.edu

Tatsunori B. Hashimoto
Microsoft
tahashim@microsoft.com

Percy Liang
Stanford University
pliang@cs.stanford.edu

- Introduce the problem of Group DRO
- Apply stochastic mirror descent (SMD)

for $t = 1, \ldots, T$ **do**
  $g \sim \text{Uniform}(1, \ldots, m)$ // Choose a group g at random
  $x, y \sim P_g$ // Sample x, y from group g
  $q' \leftarrow q^{(t-1)}$; $q'_g \leftarrow q'_g \exp(\eta_q \ell(\theta^{(t-1)}; (x, y)))$ // Update weights for group g
  $q^{(t)} \leftarrow q' / \sum_{g'} q'_{g'}$ // Renormalize q
  $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta_\theta q_g^{(t)} \nabla \ell(\theta^{(t-1)}; (x, y))$ // Use q to update θ
end

- A suboptimal $O(m^2 (\log m) / \epsilon^2)$ sample complexity

Related Work II

➢ The Work of Haghtalab et al. (NeurIPS 2022)

Nika Haghtalab, Michael I. Jordan, and Eric Zhao

University of California, Berkeley

☐ Try to improve the sample complexity by reusing samples

for $a = 1, 2, \ldots, \lceil T/r \rceil$ **do**
  Realize $\xi^{\perp^{(a)}}$ at cost r; // Sample datapoints from every distribution.
  for $t = ar + 1 - r, \ldots, ar$ **do**
    Realize $\xi^{q^{(t)}}$ at cost 1; // Sample from adversary-selected distribution.
    Estimate gradients: $\widehat{g}_+^{(t)} = \widehat{g}_+\left(\xi^{\perp^{(a)}}, p^{(t)}, q^{(t)}\right)$, $\widehat{g}_-^{(t)} = \widehat{g}_-\left(\xi^{q^{(t)}}, p^{(t)}, q^{(t)}\right)$;
    Run Hedge updates: $p^{(t+1)} = \mathcal{Q}_{\text{hedge}}\left(p^{(t)}, \widehat{g}_+^{(t)}\right)$, $q^{(t+1)} = \mathcal{Q}_{\text{hedge}}\left(q^{(t)}, \widehat{g}_+^{(t)}\right)$;
  end for
end for

However, reusing samples introduces a dependence issue, making the analysis invalid.

Related Work III

➢ The Work of Soma et al. (2022)

Tasuku Soma	Khashayar Gatmiry	Stefanie Jegelka
MIT	MIT	MIT
tasuku@mit.edu	gatmiry@mit.edu	stefje@mit.edu

- Utilize online learning to reduce the sample complexity

2: **for** $t = 1, \ldots, T$ **do**
3: Sample $i_t \sim q_t$.
4: Call the stochastic oracle to obtain $z \sim P_{i_t}$.
5: $\theta_{t+1} \leftarrow \text{proj}_\Theta(\theta_t - \eta_{\theta,t} \nabla_\theta \ell(\theta_t; z))$
6: $\nabla\Psi(\tilde{q}_{t+1}) \leftarrow \nabla\Psi(q_t) - \frac{\eta_q}{q_{t,i_t}} \ell(\theta_t; z)\mathbf{e}_{i_t}$; $q_{t+1} \leftarrow \text{argmin}_{q \in Q} D_\Psi(q, \tilde{q}_{t+1})$, where $D_\Psi(x, y) = \Psi(x) - \Psi(y) - \nabla\Psi(x)^\top (y - x)$ is the Bregman divergence with respect to Ψ.
7: **return** $\frac{1}{T}\sum_{t=1}^{T} \theta_t$.

Online Convex Optimization

Multi-armed Bandits (MAB)

- Establish a nearly optimal $O(m (\log m) / \epsilon^2)$ complexity
- Suffer a dependence issue, but can be fixed

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- Stochastic Approximation of MERO
  - A Multi-Stage SA Approach
  - An Anytime SA Approach
- Conclusion

Our Result I (Zhang et al. NeurIPS 2023)

➢ Minimax Risk Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ R_i(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right] \right\}$$

- A finite number of m distributions

Equivalent

➢ Stochastic Convex-Concave Optimization

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

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

☐ Apply stochastic mirror descent (Nemirovski et al., 2009)

Performance Measure

- Stochastic Convex-Concave Optimization

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

- Optimization Error of $(\bar{\mathbf{w}}, \bar{\mathbf{q}})$

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

  - Meaningful for Group DRO

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

$$\leq \max_{\mathbf{q}\in\Delta_m} \sum_{i=1}^{m} q_i R_i(\bar{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{m} \bar{q}_i R_i(\mathbf{w}) = \epsilon_\phi(\bar{\mathbf{w}},\bar{\mathbf{q}})$$

Stochastic Mirror Descent (SMD)

➢ Stochastic Convex-Concave Optimization

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

- Recall that $R_i(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right]$
- Stochastic Gradients at $(\mathbf{w}_t, \mathbf{q}_t)$

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

$$\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t) = [\ell(\mathbf{w}_t; \mathbf{z}_t^{(1)}), \ldots, \ell(\mathbf{w}_t; \mathbf{z}_t^{(m)})]^\top$$

✓ Draw m samples $\mathbf{z}_t^{(i)} \in \mathcal{P}_i,\ i = 1, \ldots, m$

Stochastic Mirror Descent (SMD)

- Update by mirror descent

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \left\{ \eta_w \langle \mathbf{g}_w(\mathbf{w}_t, \mathbf{q}_t), \mathbf{w} - \mathbf{w}_t \rangle + B_w(\mathbf{w}, \mathbf{w}_t) \right\}$$

$$\mathbf{q}_{t+1} = \operatorname*{argmin}_{\mathbf{q}\in\Delta_m} \left\{ \eta_q \langle -\mathbf{g}_q(\mathbf{w}_t, \mathbf{q}_t), \mathbf{q} - \mathbf{q}_t \rangle + B_q(\mathbf{q}, \mathbf{q}_t) \right\}$$

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

- Special cases:

SGD: $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}} \left[\mathbf{w}_t - \eta_w \mathbf{g}_w(\mathbf{w}_t, \mathbf{q}_t) \right]$

Hedge: $q_{t+1,i} = \dfrac{q_{t,i} \exp\left(\eta_q \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)})\right)}{\sum_{j=1}^{m} q_{t,j} \exp\left(\eta_q \ell(\mathbf{w}_t; \mathbf{z}_t^{(j)})\right)}, \ \forall i \in [m]$

Theoretical Guarantee

Theorem 1 *By setting* $\eta_w = D^2\sqrt{\frac{8}{5T(D^2G^2+\ln m)}}$ *and* $\eta_q = (\ln m)\sqrt{\frac{8}{5T(D^2G^2+\ln m)}}$, *with probability at least* $1-\delta$,

$$\epsilon_\phi(\bar{\mathbf{w}}, \bar{\mathbf{q}}) \leq \left(8 + 2\ln\frac{2}{\delta}\right)\sqrt{\frac{10(D^2G^2+\ln m)}{T}} = O\left(\sqrt{\frac{\log m}{T}}\right)$$

- It requires m samples per iteration
- The total sample complexity is $O(m(\log m)/\epsilon^2)$
- Lower bound $\Omega(m/\epsilon^2)$ (Soma et al. 2022)

➢ Credit to Nemirovski et al. (2009, § 3.2)

> **3.2. Application to minimax stochastic problems.** Consider the following minimax stochastic problem:
>
> (3.18) $$\min_{x\in X}\max_{1\leq i\leq m}\{f_i(x) = \mathbb{E}[F_i(x,\xi)]\},$$

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- Stochastic Approximation of MERO
  - A Multi-Stage SA Approach
  - An Anytime SA Approach
- Conclusion

Our Result II (Zhang et al. NeurIPS 2023)

Is it possible to reduce the number of samples per iteration from m to 1?

- The algorithm of Sagawa et al. (ICLR 2020)
  - Apply stochastic mirror descent with 1 sample pe iteration

$$\hat{\mathbf{g}}_w(\mathbf{w}_t, \mathbf{q}_t) = q_{t,i_t} m \nabla \ell(\mathbf{w}_t; \mathbf{z}_t^{(i_t)})$$

$$\hat{\mathbf{g}}_q(\mathbf{w}_t, \mathbf{q}_t) = [0, \ldots, m\ell(\mathbf{w}_t; \mathbf{z}_t^{(i_t)}), \ldots, 0]^\top$$

They are unbiased, but have very large variances.

  - Converge slowly, and have an $O(m^2 (\log m) / \epsilon^2)$ complexiy

Two-player Games

- Stochastic Convex-Concave Optimization

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

- Two-player Games (Rakhlin and Sridharan, 2013)
  - The 1st player minimizes convex functions

$$\sum_{i=1}^{m} q_{1,i} R_i(\mathbf{w}),\ \sum_{i=1}^{m} q_{2,i} R_i(\mathbf{w}),\ \cdots,\ \sum_{i=1}^{m} q_{T,i} R_i(\mathbf{w})$$

  - The 2nd player maximizes linear functions

$$\sum_{i=1}^{m} q_i R_i(\mathbf{w}_1),\ \sum_{i=1}^{m} q_i R_i(\mathbf{w}_2),\ \cdots,\ \sum_{i=1}^{m} q_i R_i(\mathbf{w}_T)$$

Non-oblivious Online Learning

- The 1st player minimizes convex functions

$$\sum_{i=1}^{m} q_{1,i} R_i(\mathbf{w}), \sum_{i=1}^{m} q_{2,i} R_i(\mathbf{w}), \cdots, \sum_{i=1}^{m} q_{T,i} R_i(\mathbf{w})$$

- Non-oblivious online convex optimization (OCO) with stochastic gradients

Stochastic gradients	• We only have stochastic gradients of each online function $\sum_{i=1}^{m} q_{t,i} R_i(\cdot)$
Non-oblivious	• The function $\sum_{i=1}^{m} q_{t,i} R_i(\cdot)$ depends on previous solutions $\mathbf{w}_1, \ldots, \mathbf{w}_{t-1}$

✓ Distinguish our method from that of Soma et al. (2022)

Non-oblivious Online Learning

- The 1st player minimizes convex functions

$$\sum_{i=1}^{m} q_{1,i} R_i(\mathbf{w}), \sum_{i=1}^{m} q_{2,i} R_i(\mathbf{w}), \cdots, \sum_{i=1}^{m} q_{T,i} R_i(\mathbf{w})$$

- Non-oblivious online convex optimization (OCO) with stochastic gradients
- Apply Stochastic Mirror Descent

$$\tilde{\mathbf{g}}_w(\mathbf{w}_t, \mathbf{q}_t) = \nabla \ell(\mathbf{w}_t; \mathbf{z}_t^{(i_t)})$$ → Small variance

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \left\{ \eta_w \langle \tilde{\mathbf{g}}_w(\mathbf{w}_t, \mathbf{q}_t), \mathbf{w} - \mathbf{w}_t \rangle + B_w(\mathbf{w}, \mathbf{w}_t) \right\}$$

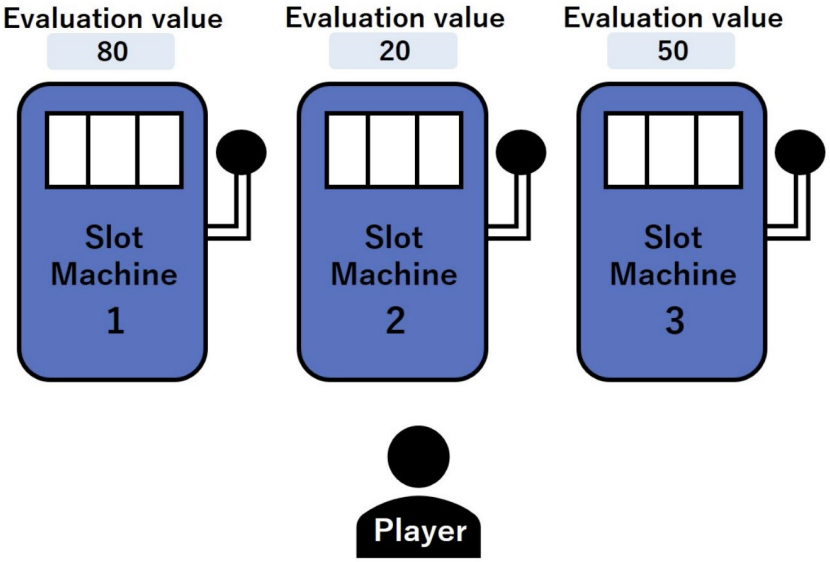
The analysis is significantly different from the traditional SMD

Non-oblivious Online Learning

- The 2nd player maximizes linear functions

$$\sum_{i=1}^{m} q_i R_i(\mathbf{w}_1), \sum_{i=1}^{m} q_i R_i(\mathbf{w}_2), \cdots, \sum_{i=1}^{m} q_i R_i(\mathbf{w}_T)$$

- Non-oblivious multi-armed bandits (MAB) with stochastic rewards



Non-oblivious Online Learning

- The 2nd player maximizes linear functions

$$\sum_{i=1}^{m} q_i R_i(\mathbf{w}_1), \sum_{i=1}^{m} q_i R_i(\mathbf{w}_2), \cdots, \sum_{i=1}^{m} q_i R_i(\mathbf{w}_T)$$

  - Non-oblivious multi-armed bandits (MAB) with stochastic rewards
  - Apply Exp3-IX for non-oblivious MAB (Neu, 2015)

$$\tilde{s}_{t,i} = \frac{1 - \ell(\mathbf{w}_t, \mathbf{z}_t^{(i_t)})}{q_{t,i} + \gamma} \cdot \mathbb{I}[i_t = i]$$

→ Bias-Variance tradeoff

$$\mathbf{q}_{t+1} = \operatorname*{argmin}_{\mathbf{q} \in \Delta_m} \left\{ \eta_q \langle \tilde{\mathbf{s}}_t, \mathbf{q} - \mathbf{q}_t \rangle + B_q(\mathbf{q}, \mathbf{q}_t) \right\}$$

Theoretical Guarantee

Theorem 2 *By setting* $\eta_w = \frac{2D}{G\sqrt{5T}}$, $\eta_q = \sqrt{\frac{\ln m}{mT}}$ *and* $\gamma = \frac{\eta_q}{2}$, *with probability at least* $1-\delta$,

$$\begin{aligned}\epsilon_\phi(\bar{\mathbf{w}}, \bar{\mathbf{q}}) \leq & DG\sqrt{\frac{1}{T}}\left(2\sqrt{5}+8\sqrt{\ln\frac{2}{\delta}}\right)+3\sqrt{\frac{m\ln m}{T}}+\sqrt{\frac{1}{2T}} \\ & +\left(\sqrt{\frac{m}{T\ln m}}+\sqrt{\frac{1}{2T}}+\frac{1}{T}\right)\ln\frac{6}{\delta} \\ = & O\left(\sqrt{\frac{m\log m}{T}}\right)\end{aligned}$$

- It requires 1 samples per iteration
- The total sample complexity is $O(m(\log m)/\epsilon^2)$
- Lower bound $\Omega(m/\epsilon^2)$ (Soma et al. 2022)

Experiments: Convergence Rate

➢ Adult dataset, Logistic loss, 6 Groups

■ SMD(1), $O\left(m\sqrt{\frac{\log m}{T}}\right)$

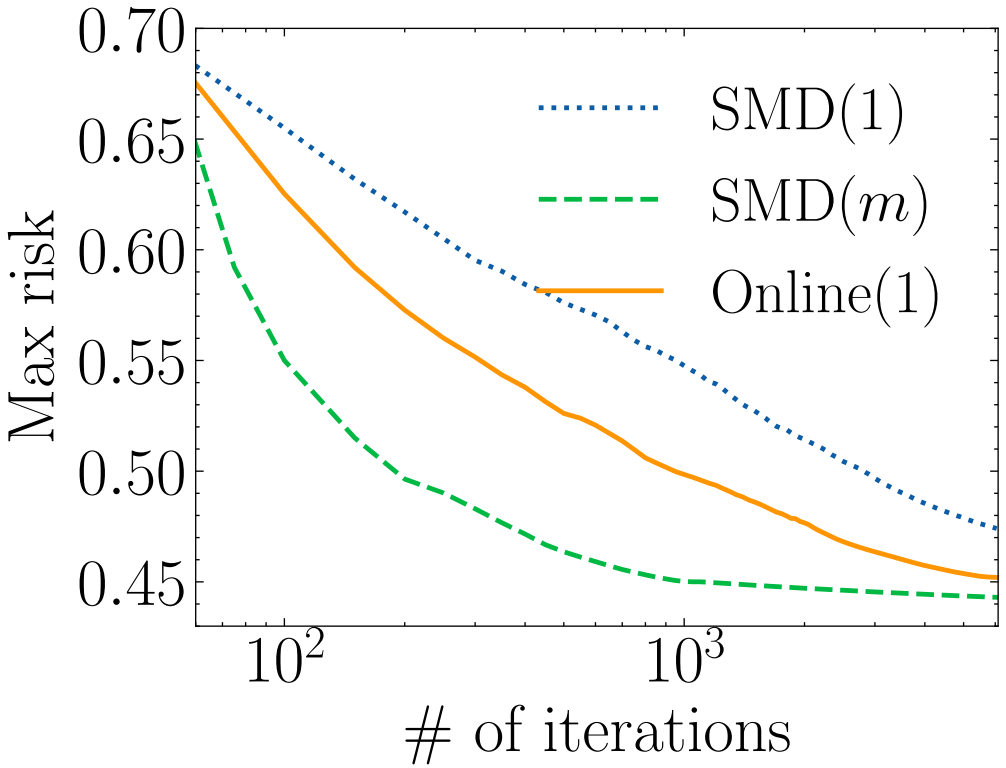
(Sagawa et al. ICLR 2020)

■ SMD(m), $O\left(\sqrt{\frac{\log m}{T}}\right)$

Our Alg. 1

■ Online(1), $O\left(\sqrt{\frac{m(\log m)}{T}}\right)$

Our Alg. 2



Experiments: Sample Complexity

➢ Adult dataset, Logistic loss, 6 Groups

- SMD(1), $O\left(\frac{m^2(\log m)}{\epsilon^2}\right)$

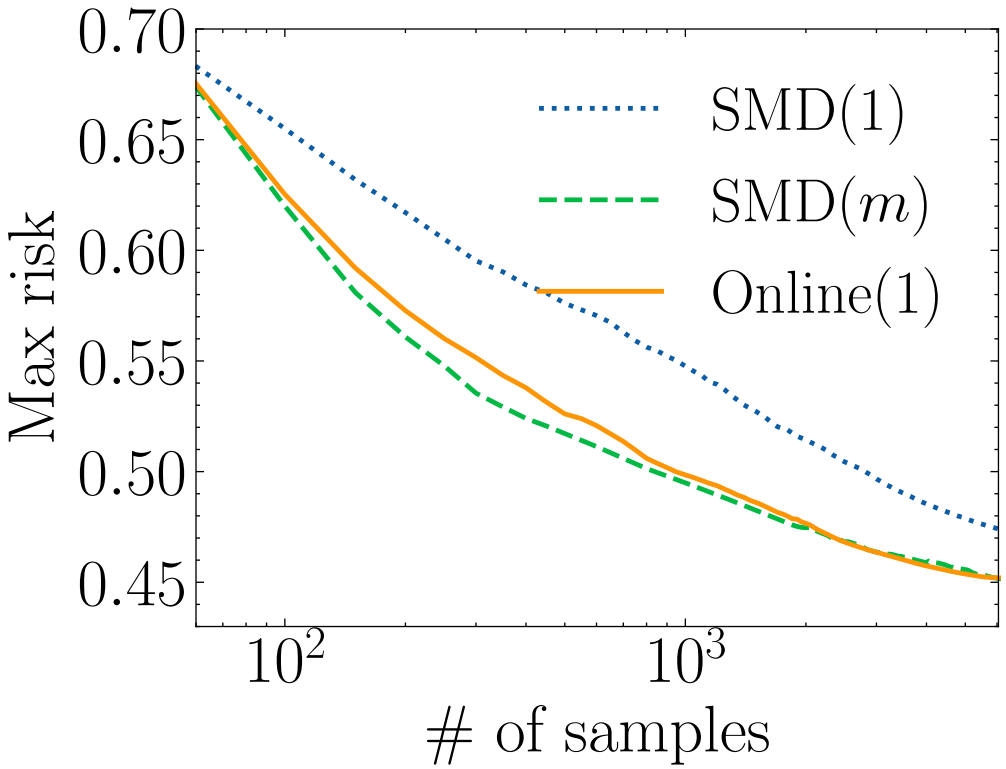
(Sagawa et al. ICLR 2020)

- SMD(m), $O\left(\frac{m(\log m)}{\epsilon^2}\right)$

Our Alg. 1

- Online(1), $O\left(\frac{m(\log m)}{\epsilon^2}\right)$

Our Alg. 2



Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- Stochastic Approximation of MERO
  - A Multi-Stage SA Approach
  - An Anytime SA Approach
- Conclusion

One limitation of GDRO

- Group DRO —— Minimax Risk Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ R_i(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right] \right\}$$

  - A finite number of m distributions

- A Potential Issue of GDRO
  - The max operator is sensitive to outliers
  - The maximal risk can be easily dominated by 1 distribution
    - ✓ Suppose distribution $\mathcal{P}_1$ contains high levels of noise

$$\max_{i\in[m]} R_i(\mathbf{w}) = R_1(\mathbf{w})$$

    - ✓ The remaining $m-1$ distributions are essentially ignored

A Variant of GDRO——MERO

- Group DRO —— Minimax Risk Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ R_i(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right] \right\}$$

  - A finite number of m distributions

- Minimax Excess Risk Optimization (MERO) (Agarwal and Zhang, 2022)

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ \underbrace{\mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right]}_{:=R_i(\mathbf{w})} - \underbrace{\min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right]}_{:=R_i^*} \right\}$$

  - Subtract the intrinsic difficulty of each distribution
  - Suppress the effect of heterogeneous noise

The Optimization of MERO

- Minimax Excess Risk Optimization (MERO)

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

  - Only exist an inefficient algorithm for empirical MERO (Agarwal and Zhang, 2022)

- Stochastic Convex-Concave Optimization

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

  - But R_i^* is unknown

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- Stochastic Approximation of MERO
  - A Multi-Stage SA Approach
  - An Anytime SA Approach
- Conclusion

Our Result I (Zhang et al. ICML 2024)

- A Multi-Stage Stochastic Approximation Approach
  - We first estimate the value of R_i^*, and then solve an approximate problem by replacing R_i^* with its estimation

Stage 1: Minimizing each risk $R_i(\cdot)$

$$\min_{\mathbf{w}\in\mathcal{W}} \left\{R_i(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right]\right\}, \quad \forall i \in [m]$$

- ✓ Runing SMD for T iterations
- ✓ A solution $\bar{\mathbf{w}}^{(i)}$ such that with probability $1-\delta$

$$R_i(\bar{\mathbf{w}}^{(i)}) - R_i^* = O\left(\sqrt{\frac{1}{T}\log\frac{1}{\delta}}\right)$$

- ✓ By union bound, $\max_{i\in[m]}[R_i(\bar{\mathbf{w}}^{(i)}) - R_i^*] = O(\sqrt{(\log m)/T})$

Our Result I (Zhang et al. ICML 2024)

Stage 2: Estimating the value of $R_i(\bar{\mathbf{w}}^{(i)})$

- ✓ Draw T samples $\mathbf{z}_1^{(i)}, \ldots, \mathbf{z}_T^{(i)}$ from distribution $\mathcal{P}_i$
- ✓ Calculate the sample average

$$\widehat{R}_i(\bar{\mathbf{w}}^{(i)}) = \frac{1}{T}\sum_{j=1}^{T} \ell(\bar{\mathbf{w}}^{(i)}; \mathbf{z}_j^{(i)})$$

- ✓ By concentration inequalities and union bound

$$\max_{i\in[m]} |\widehat{R}_i(\bar{\mathbf{w}}^{(i)}) - R_i(\bar{\mathbf{w}}^{(i)})| = O\left(\sqrt{\frac{\log m}{T}}\right)$$

- ✓ As a result

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

Our Result I (Zhang et al. ICML 2024)

Stage 3: Optimizing an approximate problem

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

✓ SMD can be directly applied for T iterations

➢ Theoretical Guarantee

Theorem 3 *After consuming $3mT$ samples, with high probability*

$$\epsilon_\phi(\bar{\mathbf{w}},\bar{\mathbf{q}}) = O\left(\sqrt{\frac{\log m}{T}}\right)$$

- The total sample complexity is $O(m(\log m)/\epsilon^2)$
- It is not an anytime algorithm, because T must be given

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- Stochastic Approximation of MERO
  - A Multi-Stage SA Approach
  - An Anytime SA Approach
- Conclusion

Our Result II (Zhang et al. ICML 2024)

- Stochastic Convex-Concave Optimization

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

- An Anytime Stochastic Approximation Approach
  - Alternate between estimating R_i^* and optimizing the minimax

1. Minimizing each risk $R_i(\cdot)$ by SMD for one step

$$\mathbf{w}_{t+1}^{(i)} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \left\{ \eta_t^{(i)} \langle \nabla \ell(\mathbf{w}_t^{(i)}; \mathbf{z}_t^{(i)}), \mathbf{w} - \mathbf{w}_t^{(i)} \rangle + B_w(\mathbf{w}, \mathbf{w}_t^{(i)}) \right\}$$

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

Our Result II (Zhang et al. ICML 2024)

- Stochastic Convex-Concave Optimization

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

- An Anytime Stochastic Approximation Approach
  - Alternates between estimating R_i^* and optimizing the minimax
  2. Minimizing the problem below by SMD for one step

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

  - ✓ The difference between R_i^* and $R_i(\bar{\mathbf{w}}_t^{(i)})$ is under-control

Our Result II (Zhang et al. ICML 2024)

- Updating $\mathbf{w}_t$ (the same as before)

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

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

- Updating $\mathbf{q}_t$ (different with before)

$$\mathbf{g}_q(\mathbf{w}_t, \mathbf{q}_t)$$

$$= \left[\ell(\mathbf{w}_t; \mathbf{z}_t^{(1)}) - \ell(\bar{\mathbf{w}}_t^{(1)}; \mathbf{z}_t^{(1)}), \ldots, \ell(\mathbf{w}_t; \mathbf{z}_t^{(m)}) - \ell(\bar{\mathbf{w}}_t^{(m)}; \mathbf{z}_t^{(m)}) \right]^\top$$

✓ It is a biased gradient for the original MERO problem

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

Theoretical Guarantee

Theorem 4 *With probability at least* $1-2\delta$,

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

- The convergence rate is almost the same as GDRO
- In the analysis, we need to deal with the biased gradient.
- It can return a solution at any round

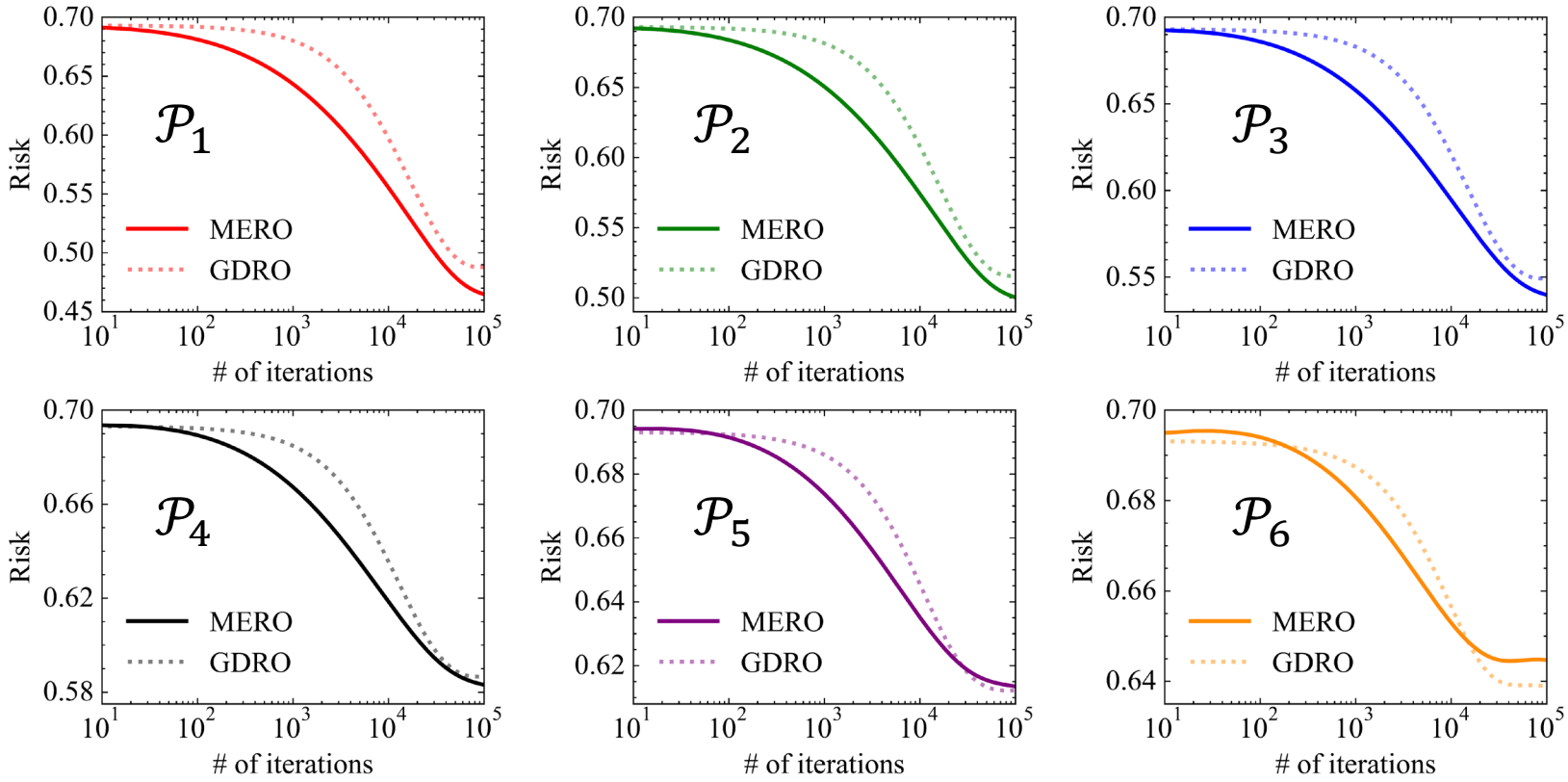
$$\bar{\mathbf{w}}_t = \sum_{j=1}^{t} \frac{\eta_j^w \mathbf{w}_j}{\sum_{k=1}^{t} \eta_k^w}, \text{ and } \bar{\mathbf{q}}_t = \sum_{j=1}^{t} \frac{\eta_j^q \mathbf{q}_j}{\sum_{k=1}^{t} \eta_k^q}$$

➢ Previous two SA approaches for GDRO

- Can be easily modified to be anytime

Experiments: GDRO v.s. MERO

- Synthetic dataset, 6 distributions with different noise



- GDRO performs better on distributions with higher noise

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- Stochastic Approximation of MERO
  - A Multi-Stage SA Approach
  - An Anytime SA Approach
- Conclusion

Conclusion

- GDRO——Minimax Risk Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ R_i(\mathbf{w}) = \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right] \right\}$$

1. Stochastic Mirror Descent, $O(m(\log m)/\epsilon^2)$
2. Non-oblivious Online Learning, $O(m(\log m)/\epsilon^2)$

- MERO—— Minimax Excess Risk Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} \left\{ \underbrace{\mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right]}_{:=R_i(\mathbf{w})} - \underbrace{\min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}_{\mathbf{z}\sim\mathcal{P}_i}\left[\ell(\mathbf{w};\mathbf{z})\right]}_{:=R_i^*} \right\}$$

1. A Multi-Stage SA Approach, $O(m(\log m)/\epsilon^2)$
2. An Anytime SA Approach , $\tilde{O}(m(\log m)/\epsilon^2)$

Future Work

- Investigate the Imbalanced setting
  - Number of samples from different distributions are different
  - Weighted GDRO/MERO (Zhang et al. NeurIPS 2023, ICML 2024)

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

- Investigate the Empirical GDRO/MERO (Yu et al. ICML 2024)

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

- Apply to real-world problems (e.g., training big model)

Reference I

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

Thanks!

- Lijun Zhang, Peng Zhao, Zhen-Hua Zhuang, Tianbao Yang, and Zhi-Hua Zhou. Stochastic Approximation Approaches to Group Distributionally Robust Optimization. In In Advances in Neural Information Processing Systems 36 (NeurIPS), pages 52490–52522, 2023.
- Lijun Zhang, Haomin Bai, Wei-Wei Tu, Ping Yang, and Yao Hu. Efficient Stochastic Approximation of Minimax Excess Risk Optimization. In Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.
- Dingzhi Yu, Yunuo Cai, Wei Jiang, and Lijun Zhang. Efficient Algorithms for Empirical Group Distributional Robust Optimization and Beyond. In Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.
- Herbert Scarf. A min-max solution of an inventory problem. Studies in the Mathematical Theory of Inventory and Production, pages 201–209, 1958.
- Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. Robust Optimization. Princeton University Press, 2009.
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization, 19(4):1574–1609, 2009.

Reference II

- Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research, 58(3):595–612, 2010.
- Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Mathematical Programming, 171:115–166, 2018.
- Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In Advances in Neural Information Processing Systems 33 (NeurIPS), pages 8847–8860, 2020.
- John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. The Annals of Statistics, 49(3):1378 – 1406, 2021.
- Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pages 77 – 91, 2018.
- Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for a class of distributionally robust optimization with non-convex objectives. In Advances in Neural Information Processing Systems 34 (NeurIPS), pages 10067–10080, 2021.

Reference III

- Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex-concave min-max optimization: Provable algorithms and applications in machine learning. Optimization Methods and Software, 37(3):1087–1121, 2022.
- Tasuku Soma, Khashayar Gatmiry, and Stefanie Jegelka. Optimal algorithms for group distributionally robust optimization and beyond. ArXiv e-prints, arXiv:2212.13669, 2022.
- Nika Haghtalab, Michael I. Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In Advances in Neural Information Processing Systems 35 (NeurIPS), pages 406–419, 2022.
- Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. Management Science, 59(2):341–357, 2013.
- Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In Advances in Neural Information Processing Systems 29 (NIPS), pages 2216–2224, 2016.
- Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In Advances in Neural Information Processing Systems 28 (NIPS), pages 3168–3176, 2015.

Reference IV

- Alekh Agarwal and Tong Zhang. Minimax regret optimization for robust machine learning under distribution shift. In Proceedings of 35th Conference on Learning Theory (COLT), pages 2704–2729, 2022.
- Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. Stochastic Systems, 1(1):17–58, 2011.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In International Conference on Learning Representations (ICLR), 2020.
- Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In Advances in Neural Information Processing Systems 26 (NIPS), pages 315–323, 2013.
- Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. In Advance in Neural Information Processing Systems 26 (NIPS), pages 980–988, 2013.
- Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In Advances in Neural Information Processing Systems 26 (NIPS), pages 3066–3074, 2013.