Group Distributionally Robust Optimization

Lijun Zhang
Nanjing University, China

CSIAM-BDAI 2023

(Updated at 2023.11.21)

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- GDRO with Imbalanced Data
  - Stochastic Mirror Descent with Non-uniform Sampling
  - Stochastic Mirror-Prox Algorithm with Mini-batches
- 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
$$\min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}_{(\mathbf{x},y)\sim\mathcal{P}_0}\left[\max(1-y\mathbf{w}^\top\mathbf{x},0)\right] + \frac{\lambda}{2}\|\mathbf{w}\|^2$$
  - Linear Regression
$$\min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}_{(\mathbf{x},y)\sim\mathcal{P}_0}\left[(y-\mathbf{w}^\top x)^2\right]$$

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 $\nu(\mathbf{w}) = \|\mathbf{w}\|^2/2$

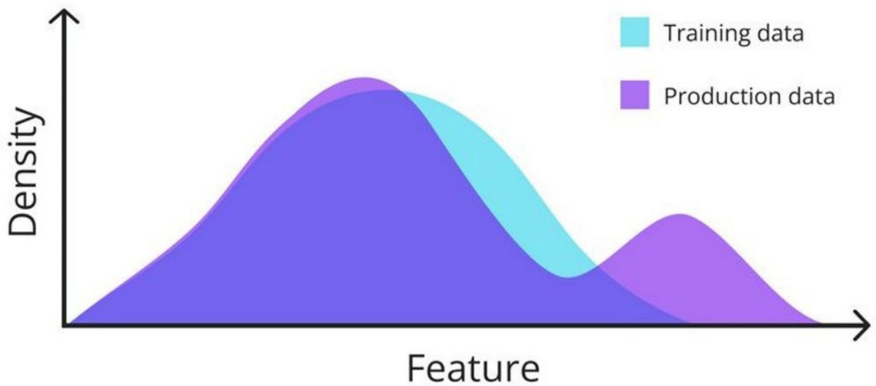
Statistical Machine Learning

- 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 Vast 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

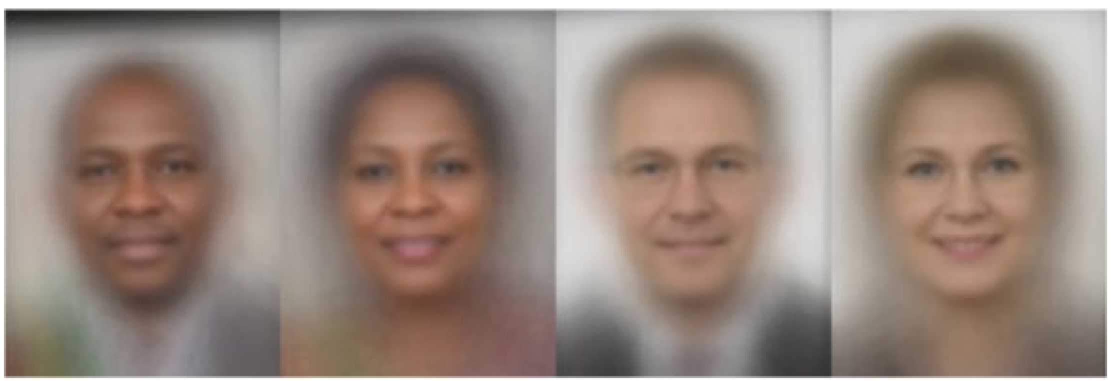
$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{m}\sum_{i=1}^{m} R_i(\mathbf{w})$$

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



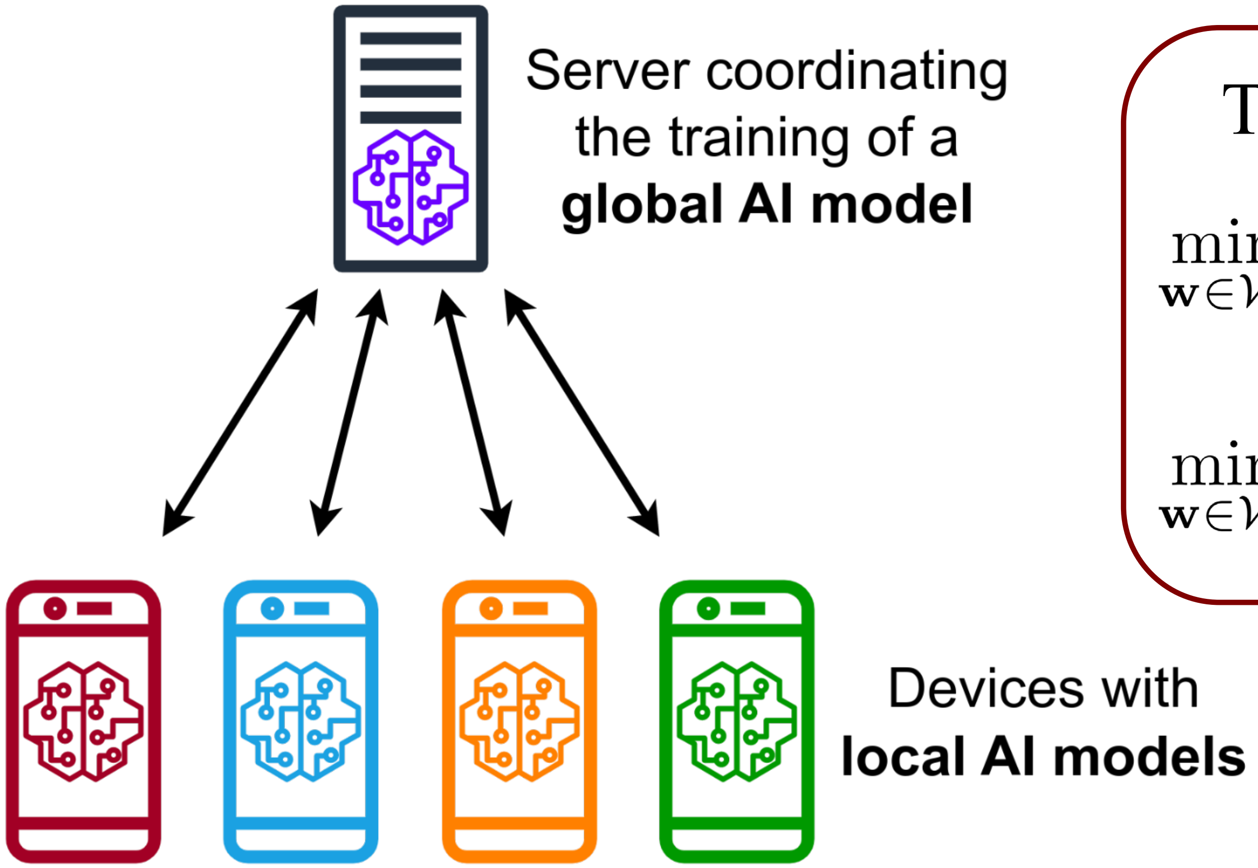
Solution ⇩

Optimizing performance across all groups

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

Application: Federated Learning

- A Single Model facing Multiple Distributions



Two Choices

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{i\in[m]} R_i(\mathbf{w})$$

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

https://en.wikipedia.org/wiki/Federated_learning

Outline

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

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- GDRO with Imbalanced Data
  - Stochastic Mirror Descent with Non-uniform Sampling
  - Stochastic Mirror-Prox Algorithm with Mini-batches
- 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, ..., ⌈T/r⌉ do
  Realize ξ^{⊥(a)} at cost r;      // Sample datapoints from every distribution.
  for t = ar + 1 − r, ..., ar do
    Realize ξ^{q^(t)} at cost 1;   // Sample from adversary-selected distribution.
    Estimate gradients: ĝ_+^(t) = ĝ_+(ξ^{⊥(a)}, p^(t), q^(t)),  ĝ_-^(t) = ĝ_-(ξ^{q^(t)}, p^(t), q^(t));
    Run Hedge updates: p^(t+1) = Q_hedge(p^(t), ĝ_+^(t)), q^(t+1) = Q_hedge(q^(t), ĝ_+^(t));
  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
MIT
tasuku@mit.edu

Khashayar Gatmiry
MIT
gatmiry@mit.edu

Stefanie Jegelka
MIT
stefje@mit.edu

- Utilize online learning to reduce the sample complexity

for $t = 1, \ldots, T$ **do**
    Sample $i_t \sim q_t$.
    Call the stochastic oracle to obtain $z \sim P_{i_t}$.
    $\theta_{t+1} \leftarrow \mathrm{proj}_\Theta(\theta_t - \eta_{\theta,t} \nabla_\theta \ell(\theta_t; z))$
    $\nabla\Psi(\tilde{q}_{t+1}) \leftarrow \nabla\Psi(q_t) - \frac{\eta_q}{q_{t,i_t}} \ell(\theta_t; z) \mathbf{e}_{i_t}$; $q_{t+1} \leftarrow \mathrm{argmin}_{q \in Q} D_\Psi(q, \tilde{q}_{t+1})$, where $D_\Psi(x, y) = \Psi(x) - \Psi(y) - \nabla\Psi(x)^\top (y - x)$ is the Bregman divergence with respect to Ψ.
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
- GDRO with Imbalanced Data
  - Stochastic Mirror Descent with Non-uniform Sampling
  - Stochastic Mirror-Prox Algorithm with Mini-batches
- 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:

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

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

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:

$$\min_{x\in X}\max_{1\leq i\leq m}\{f_i(x) = \mathbb{E}[F_i(x,\xi)]\}, \quad (3.18)$$

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- GDRO with Imbalanced Data
  - Stochastic Mirror Descent with Non-uniform Sampling
  - Stochastic Mirror-Prox Algorithm with Mini-batches
- 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

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

- 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 $\mathbf{w}_1, \ldots, \mathbf{w}_{t-1}$

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

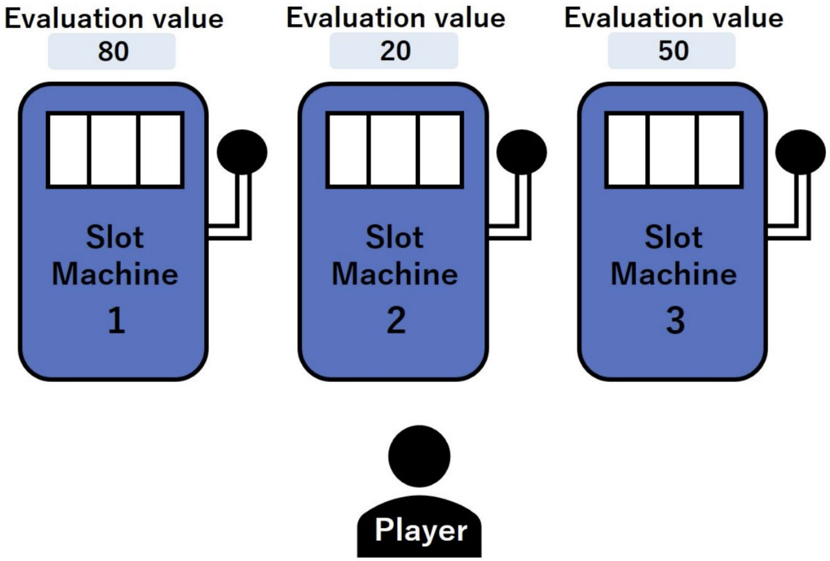
$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \left\{ \eta_w \langle \tilde{\mathbf{g}}_w(\mathbf{w}_t, \mathbf{q}_t), \mathbf{w} - \mathbf{w}_t \rangle + B_w(\mathbf{w}, \mathbf{w}_t) \right\}$$

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



http://www.apsipa.org/proceedings/2021/pdfs/0001899.pdf

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)

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

→ 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* $1-\delta$,

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

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

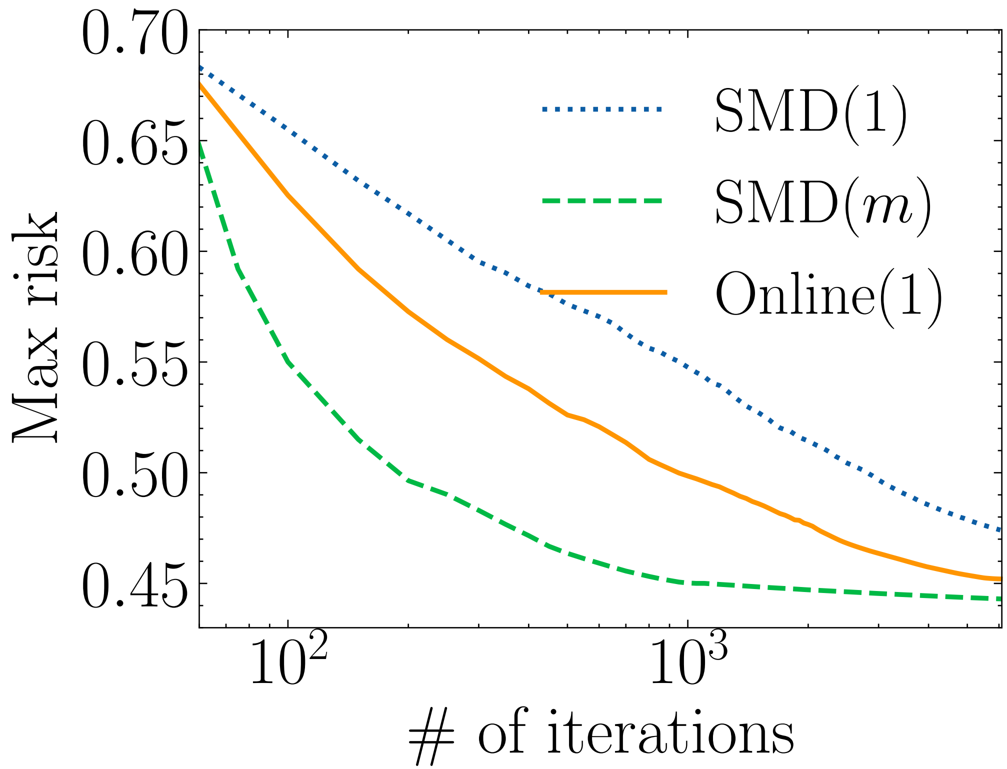
(Sagawa et al. ICLR 2020)

- 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

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

- Adult dataset, Logistic loss, 6 Groups

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

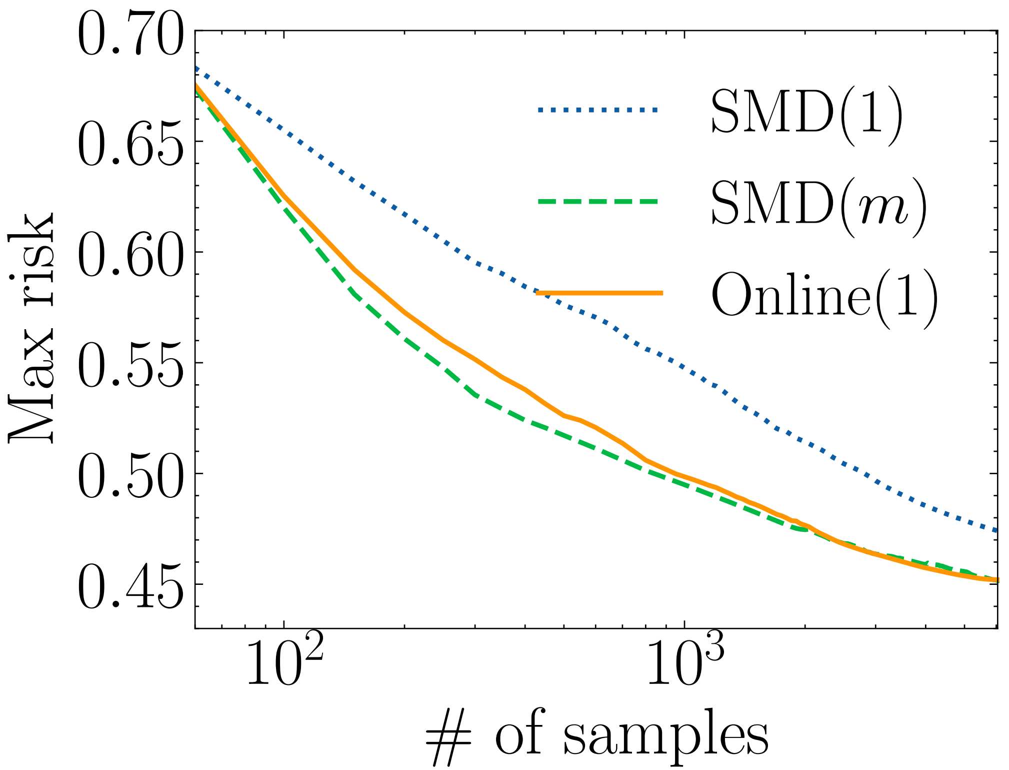
(Sagawa et al. ICLR 2020)

■ 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

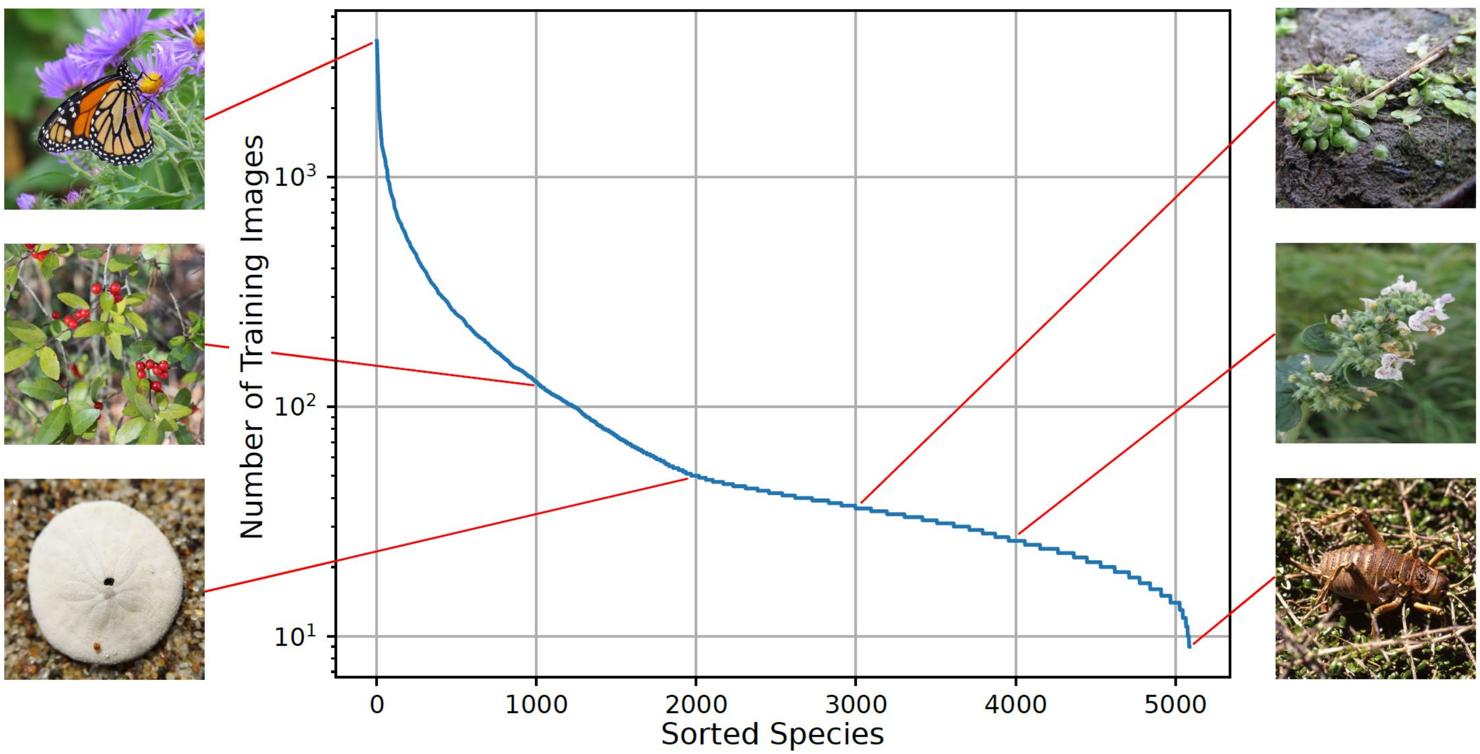


Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- GDRO with Imbalanced Data
  - Stochastic Mirror Descent with Non-uniform Sampling
  - Stochastic Mirror-Prox Algorithm with Mini-batches
- Conclusion

Imbalanced datasets

- iNaturalist dataset, consisting of 859,000 images from over 5,000 different species (Horn et al, 2018)



Distribution of training images per species

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

It draws **the same number** of samples from every distribution.

GDRO Under Imbalanced Setting

- n_i be the number of samples can be drawn from $\mathcal{P}_i$

$$n_1 \geq n_2 \geq \cdots \geq n_m$$

- A naive **baseline**: running SMD for n_m rounds

n_m rounds

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

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

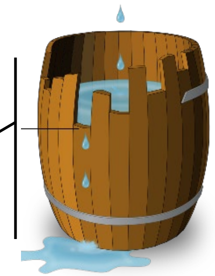
$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \left\{ \eta_w \langle \mathbf{g}_w(\mathbf{w}_t, \mathbf{q}_t), \mathbf{w} - \mathbf{w}_t \rangle + B_w(\mathbf{w}, \mathbf{w}_t) \right\}$$

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

Limitations of Baseline

1. The optimization error is determined by n_m
   - According to Theorem 1, we have

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



Barrel Effect

2. A large amount of samples are wasted
   - For distribution $\mathcal{P}_1$, $n_1 - n_m$ samples are wasted
   - For distribution $\mathcal{P}_2$, $n_2 - n_m$ samples are wasted

   ……

   - For distribution $\mathcal{P}_{m-1}$, $n_{m-1} - n_m$ samples are wasted

Outline

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

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- GDRO with Imbalanced Data
  - Stochastic Mirror Descent with Non-uniform Sampling
  - Stochastic Mirror-Prox Algorithm with Mini-batches
- Conclusion

Our Result III (Zhang et al. NeurIPS 2023)

- Applying Non-uniform Sampling
  - Run n_1 iterations, and draw a sample from $\mathcal{P}_i$ with probability $p_i = n_i / n_1$

$$\text{Expected \# of Samples: } n_1 \cdot \frac{n_i}{n_1} = n_i$$

- Updating according to SMD
  - Construct stochastic gradients

$$\mathbf{g}_w(\mathbf{w}_t, \mathbf{q}_t) = \sum_{i \in C_t} \frac{q_{t,i}}{p_i} \nabla \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)})$$

$$[\mathbf{g}_q(\mathbf{w}_t, \mathbf{q}_t)]_i = \begin{cases} \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)}) / p_i, & i \in C_t \\ 0, & \text{otherwise} \end{cases}$$

It yields very slow convergence due to the large variance caused by $1/p_m = n_1/n_m$.

✓ C_t is the set of indexes of selected distributions

Our Result III (Zhang et al. NeurIPS 2023)

- Applying Non-uniform Sampling
  - Run n_1 iterations, and draw a sample from $\mathcal{P}_i$ with probability $p_i = n_i / n_1$

$$\text{Expected \# of Samples: } n_1 \cdot \frac{n_i}{n_1} = n_i$$

- Updating according to SMD
  - Construct stochastic gradients

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

$$[\mathbf{g}_q(\mathbf{w}_t, \mathbf{q}_t)]_i = \begin{cases} \ell(\mathbf{w}_t; \mathbf{z}_t^{(i)}), & i \in C_t \\ 0, & \text{otherwise} \end{cases}$$

✓ C_t is the set of indexes of selected distributions

Our Result III (Zhang et al. NeurIPS 2023)

- A Weighted GDRO Problem

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

  - The more the number of samples, the larger the weights

- Updating according to SMD
  - Construct stochastic gradients

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

$$[\mathbf{g}_q(\mathbf{w}_t,\mathbf{q}_t)]_i = \begin{cases} \ell(\mathbf{w}_t;\mathbf{z}_t^{(i)}), & i \in C_t \\ 0, & \text{otherwise} \end{cases}$$

    - ✓ C_t is the set of indexes of selected distributions

Advantages of Weighted GDRO

- A Weighted GDRO Problem

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

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

$$\max_{i\in[m]} p_i R_i(\bar{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \varphi(\mathbf{w},\mathbf{q}) \le \epsilon_\varphi(\bar{\mathbf{w}},\bar{\mathbf{q}})$$

- Risk of Each Distribution

$$R_i(\bar{\mathbf{w}}) \le \frac{1}{p_i} \min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \varphi(\mathbf{w},\mathbf{q}) + \frac{1}{p_i}\epsilon_\varphi(\bar{\mathbf{w}},\bar{\mathbf{q}})$$

More Samples → Larger Weights → Faster Rates

Theoretical Guarantee

Theorem 3 *By setting* $\eta_w = D^2\sqrt{\frac{8}{5n_1(D^2G^2+\ln m)}}$ *and* $\eta_q = (\ln m)\sqrt{\frac{8}{5n_1(D^2G^2+\ln m)}}$, *with probability at least* $1-\delta$,

$$R_i(\bar{\mathbf{w}}) - \frac{1}{p_i}p_\varphi^* \leq \mu(\delta)\frac{\sqrt{10(D^2G^2+\ln m)n_1}}{n_i}$$

$$= O\left(\frac{\sqrt{n_1 \log m}}{n_i}\right)$$

Distribution-dependent

- The $O\left(\frac{\sqrt{n_1\log m}}{n_i}\right)$ rate is better than Baseline's $O\left(\sqrt{\frac{\log m}{n_m}}\right)$ rate when $n_i \geq \sqrt{n_1 n_m}$
- For distributions $\mathcal{P}_1$, the rate $O\left(\sqrt{\frac{\log m}{n_1}}\right)$ is nearly optimal

Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- GDRO with Imbalanced Data
  - Stochastic Mirror Descent with Non-uniform Sampling
  - Stochastic Mirror-Prox Algorithm with Mini-batches
- Conclusion

Our Result IV (Zhang et al. NeurIPS 2023)

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

- Applying Mini-batches
  - Run n_m iterations, and draw n_i/n_m sample from $\mathcal{P}_i$

$$\text{\# of Samples: } n_m \cdot \frac{n_i}{n_m} = n_i$$

- Stochastic Gradients with Elements Having Small Variance

$$\mathbf{g}_w(\mathbf{w}_t, \mathbf{q}_t) = \sum_{i=1}^{m} q_{t,i} \left(\frac{n_m}{n_i} \sum_{j=1}^{n_i/n_m} \nabla \ell(\mathbf{w}_t; \mathbf{z}_t^{(i,j)}) \right)$$

$$\mathbf{g}_q(\mathbf{w}_t, \mathbf{q}_t) = \left[\frac{n_m}{n_1} \sum_{j=1}^{n_1/n_m} \ell(\mathbf{w}_t; \mathbf{z}_t^{(1,j)}), \ldots, p_m \ell(\mathbf{w}_t; \mathbf{z}_t^{(m)}) \right]^\top$$

Two Challenges

1. The performance of SMD does not depend on variance

➢ Stochastic Mirror-Prox Algorithm (SMPA) (Juditsky et al., 2011)

- Basically, it performs SMD twice in each iteration
- The convergence rate depends on the variance

2. The whole gradient still have a large variance

➢ A Weighted GDRO Problem

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

- Set larger weights for distributions with smaller variance

Advantages of Weighted GDRO

- A Weighted GDRO Problem

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

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

$$\max_{i\in[m]} p_i R_i(\bar{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \varphi(\mathbf{w},\mathbf{q}) \leq \epsilon_\varphi(\bar{\mathbf{w}}, \bar{\mathbf{q}})$$

- Risk of Each Distribution

$$R_i(\bar{\mathbf{w}}) \leq \frac{1}{p_i} \min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{q}\in\Delta_m} \varphi(\mathbf{w},\mathbf{q}) + \frac{1}{p_i}\epsilon_\varphi(\bar{\mathbf{w}}, \bar{\mathbf{q}})$$

More Samples → Smaller Variances → Larger Weights → Faster Rates

Theoretical Guarantee

Theorem 4 *By setting* $p_i = \frac{1/\sqrt{n_m}+1}{1/\sqrt{n_m}+\sqrt{n_m/n_i}}$, *with high probability*

$$R_i(\bar{\mathbf{w}}) - \frac{1}{p_i} p_\varphi^* = O\left(\left(\frac{1}{n_m} + \frac{1}{\sqrt{n_i}}\right)\sqrt{\kappa + \ln^2 m}\right)$$

Distribution -dependent

- ◻ A fast $O\left(\frac{\log m}{\sqrt{n_i}}\right)$ rate for distributions $\mathcal{P}_i$ such that $n_i \leq n_m^2$
- ◻ In contrast, the rate of Baseline is $O\left(\sqrt{\frac{\log m}{n_m}}\right)$
- ◻ A fast $O\left(\frac{\log m}{n_m}\right)$ rate for distributions $\mathcal{P}_i$ such that $n_i \geq n_m^2$
  - ✓ There exists a performance limit

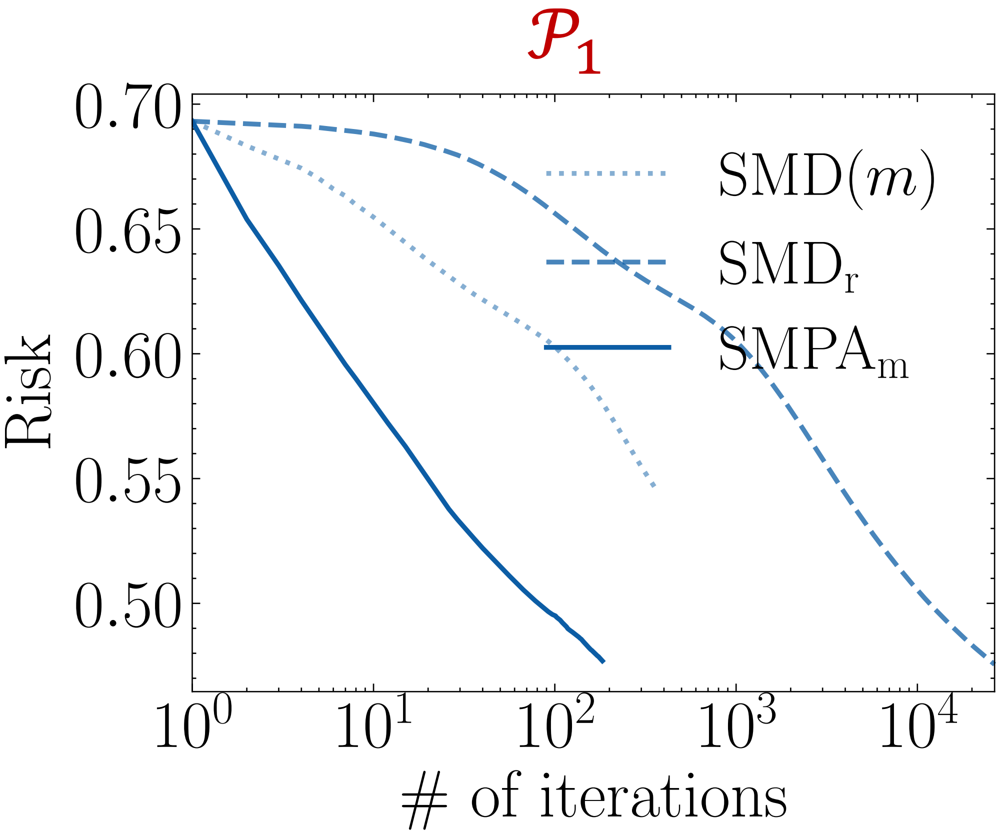
Experiments: Convergence Rate

- Adult dataset, Logistic loss, 6 Groups
- # of Samples: 26656, 11518, 1780, 1720, 998, and 364

■ SMD(m), $O\left(\sqrt{\frac{\log m}{n_6}}\right)$
Our Alg. 1

■ SMD_r, $O\left(\sqrt{\frac{\log m}{n_1}}\right)$ ✓
Our Alg. 3

■ SMPA_m, $O\left(\frac{\log m}{\sqrt{n_1}}\right)$ ✓
Our Alg. 4



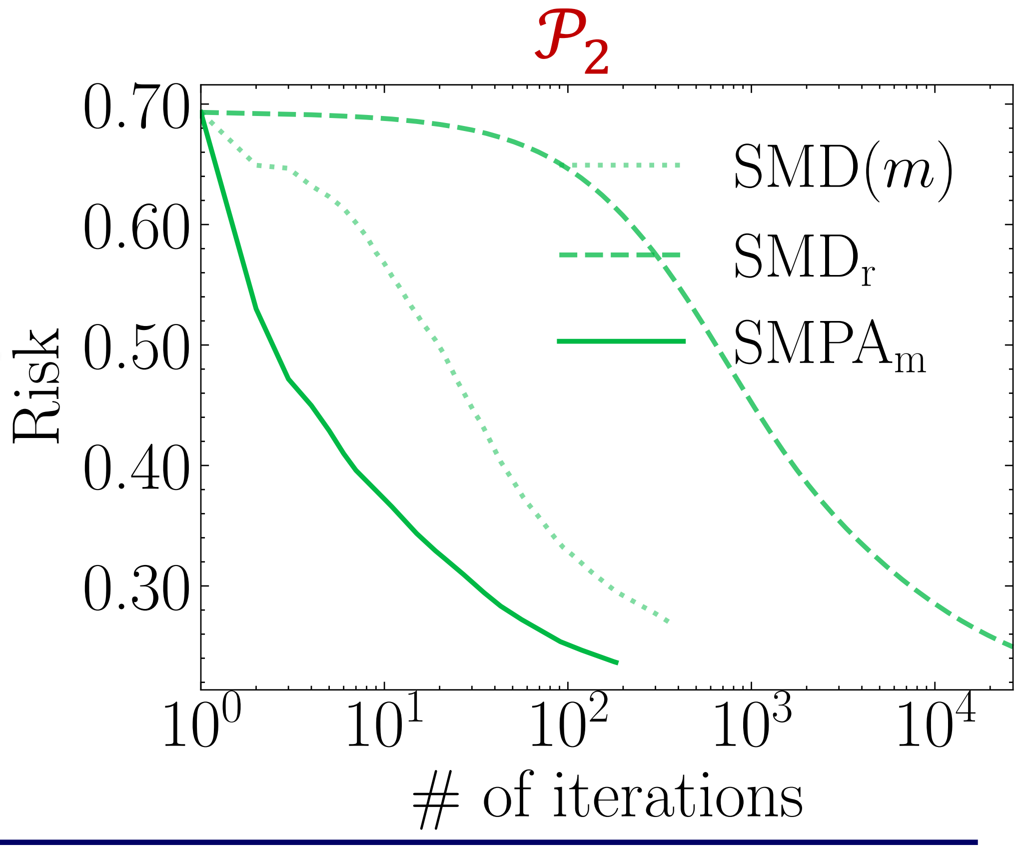
Experiments: Convergence Rate

- Adult dataset, Logistic loss, 6 Groups
- # of Samples: 26656, 11518, 1780, 1720, 998, and 364

- SMD(m), $O\left(\sqrt{\frac{\log m}{n_6}}\right)$
Our Alg. 1

- SMD_r, $O\left(\frac{\sqrt{n_1 \log m}}{n_2}\right)$ ✓
Our Alg. 3

- SMPA_m, $O\left(\frac{\log m}{\sqrt{n_2}}\right)$ ✓
Our Alg. 4



Experiments: Convergence Rate

http://lamda.nju.edu.cn

- Adult dataset, Logistic loss, 6 Groups
- # of Samples: 26656, 11518, 1780, 1720, 998, and 364

■ SMD(m), $O\left(\sqrt{\frac{\log m}{n_6}}\right)$

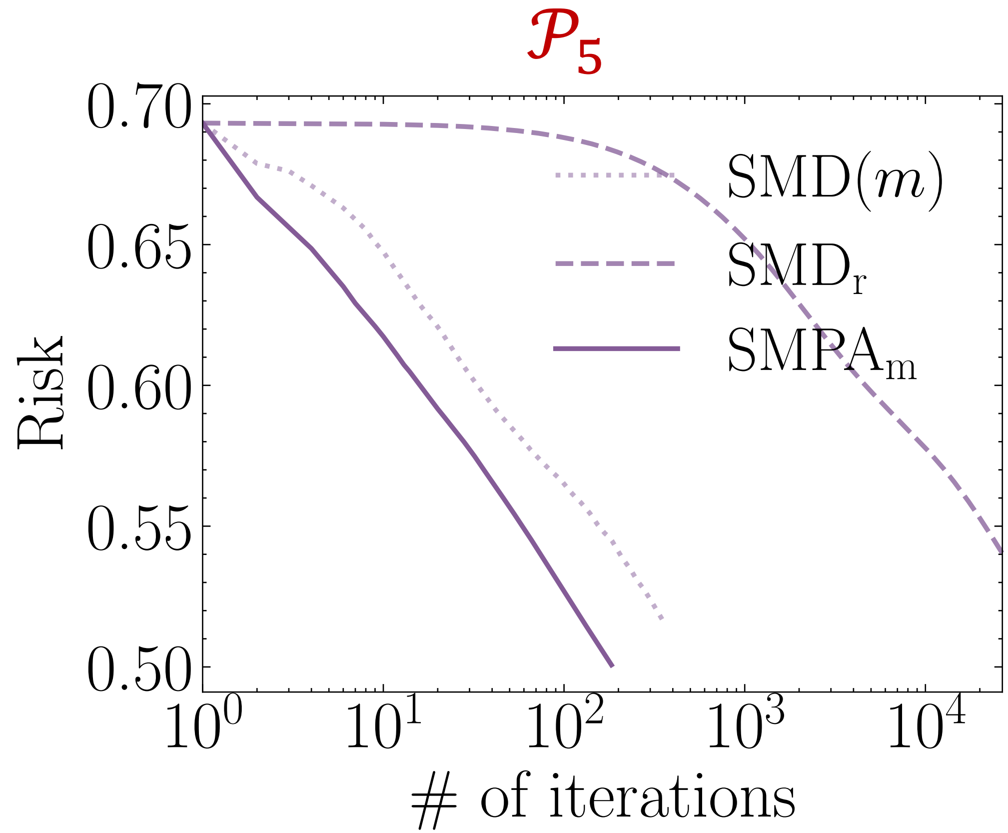
Our Alg. 1

■ SMD_r, $O\left(\frac{\sqrt{n_1}\log m}{n_5}\right)$

Our Alg. 3

■ SMPA_m, $O\left(\frac{\log m}{\sqrt{n_5}}\right)$

Our Alg. 4



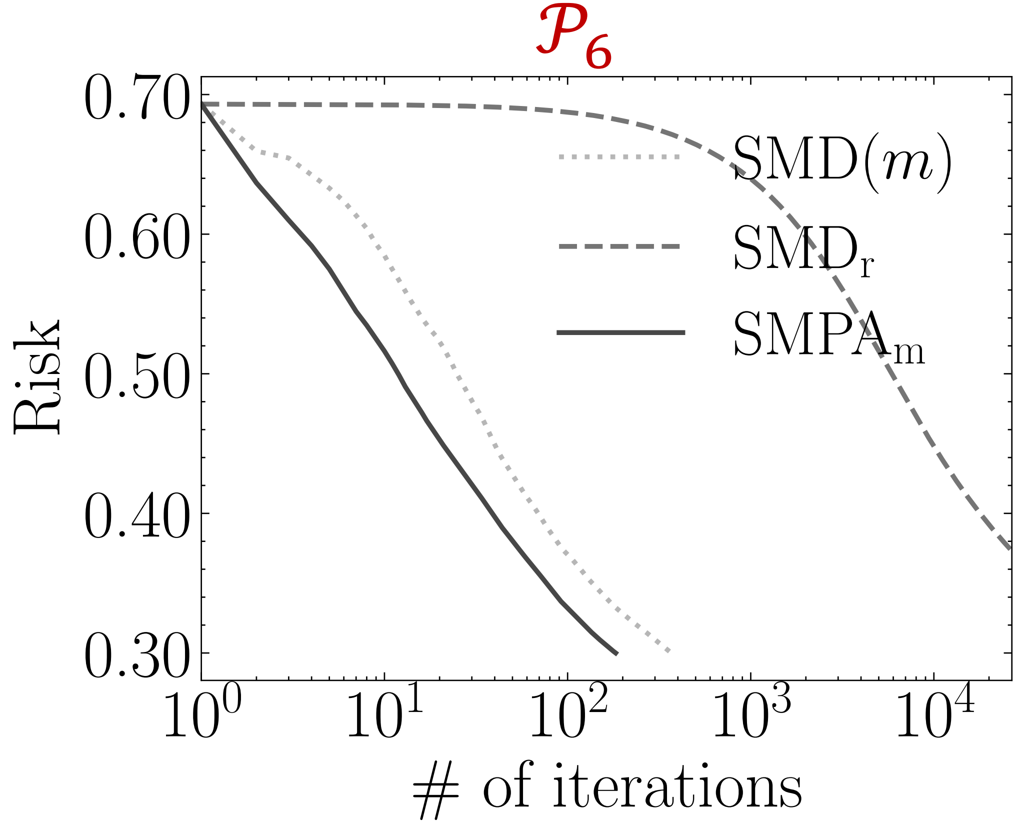
Experiments: Convergence Rate

- Adult dataset, Logistic loss, 6 Groups
- # of Samples: 26656, 11518, 1780, 1720, 998, and 364

- SMD(m), $O\left(\sqrt{\frac{\log m}{n_6}}\right)$ ✓
Our Alg. 1

- SMD_r, $O\left(\frac{\sqrt{n_1}\log m}{n_6}\right)$
Our Alg. 3

- SMPA_m, $O\left(\frac{\log m}{\sqrt{n_6}}\right)$ ✓
Our Alg. 4



Outline

- Introduction
- Related Work
- Stochastic Approximation of GDRO
  - Stochastic Mirror Descent
  - Non-oblivious Online Learning
- GDRO with Imbalanced Data
  - Stochastic Mirror Descent with Non-uniform Sampling
  - Stochastic Mirror-Prox Algorithm with Mini-batches
- 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, $O(m(\log m)/\epsilon^2)$

- GDRO with Imbalanced Data

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

1. Stochastic Mirror Descent with Non-uniform Sampling
2. Stochastic Mirror-Prox Algorithm with Mini-batches

- Distribution-dependent Convergence Rates

Future Work

- More Investigations of the Imbalanced Scenario
  - Understand the red terms below

$$R_i(\bar{\mathbf{w}}) - \frac{1}{p_i} p_\varphi^* = O(\cdot)$$

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

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

  - Subtracting the intrinsic difficulty of each distribution
  - Efficient stochastic algorithms (Zhang et al. 2023)

Reference I

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

- 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), 2023.
- Lijun Zhang and Wei-Wei Tu. Efficient Stochastic Approximation of Minimax Excess Risk Optimization. ArXiv e-prints, arXiv:2306.00026, 2023.
- 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.
- 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.

Reference II

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

- 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.
- 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.

Reference III

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

- 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.
- 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.

Reference IV

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

- 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.
- Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8769–8778, 2018