COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

A Stochastic Algorithm for Sinkhorn Distance-Regularized DRO

Yufeng Yang   Yi Zhou   Zhaosong Lu

Texas A&M University, College Station; University of Minnesota, Twin Cities

ISyE

Motivation of Study

Why Distributionally Robust Optimization(DRO)

DRO improves model robustness against distribution shift, which has important applications many several ML fields, including

- **Adversarial attack**: Gradient Attack will cause distribution shift over training data
- **Self-Supervised Learning**: Wrong selection of negative samples will cause distribution shift in embedded image-text pairs (i.e., CLIP model).
- **Reinforcement Learning**: Environment is subject to change, need to force policy shift for safety issue in real applications.

In this work, we study the information-divergence regularized DRO problem

$$\min_{x\in\mathbf{R}^d}\sup_{\mathbb{Q}}\Big\{\mathbb{E}_{\xi\sim\mathbb{Q}}\big[\ell(x;\xi)\big]-\lambda W_\varepsilon(\mathbb{P},\mathbb{Q})\Big\}, \tag{1}$$

$\ell(x;\xi)$ represents loss function under shifted distribution $\mathbb{Q}$, $W_\varepsilon(\mathbb{P},\mathbb{Q})$ represents information divergence among nominal distribution $\mathbb{P}$ and shifted distribution $\mathbb{Q}$.
Challenges: $\sup_{\mathbb{Q}}$ is maximized over distribution $\to$ Hard to find explicit $\mathbb{Q}^*$ in practice.

Choice of $W_\varepsilon(\mathbb{P},\mathbb{Q})$: Generalized Sinkhorn Distance

Denote $\Gamma(\mathbb{P},\mathbb{Q})$ as the set of joint distributions that have marginal distributions $\mathbb{P},\mathbb{Q}$. For a fixed regularization parameter $\varepsilon>0$ and a cost metric $c:\Omega\times\Omega\to\mathbf{R}$, the generalized Sinkhorn distance is defined as

$$W_\varepsilon(\mathbb{P},\mathbb{Q})=\inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})}\Big\{\mathbb{E}_{(\zeta,\xi)\sim\gamma}\big[c(\zeta,\xi)\big]+\varepsilon D_f(\gamma\,|\,\mathbb{P}\otimes\nu)\Big\},$$

where D_f corresponds to the f-divergence, that is,

$$D_f(\gamma\,|\,\mathbb{P}\otimes\nu)=\int f\Big(\frac{\mathrm{d}\gamma(\zeta,\xi)}{\mathrm{d}\mathbb{P}(\zeta)\mathrm{d}\nu(\xi)}\Big)\mathrm{d}\nu(\xi)\mathrm{d}\mathbb{P}(\zeta).$$

And $\frac{\mathrm{d}\gamma(\zeta,\xi)}{\mathrm{d}\mathbb{P}(\zeta)\mathrm{d}\nu(\xi)}$ represents density ratio of γ with respect to $\mathbb{P}\otimes\nu$ evaluated at (ζ,ξ).

Why Generalized Sinhkhorn Distance?

- **vs. KL**: 1. Symmetric; 2. Allows sample to have different probability support.
- **vs. Wasserstein Distance** Convex Programming $\to$ easier to solve.
- **vs. Original Sinkhorn Distance** f-divergence is more general than KL-divergence.

Our Contributions

TL; DR

- **Generalize** Sinkhorn distance based on the class of f-divergence measures, which allows to use a broader range of divergences to model the ambiguity set.
- **Derive** an equivalent dual formulation with strong duality guarantee. The dual formulation shares novel structures, but it can be solved efficiently using nested stochastic programming.
- **Design** a Nested-SGD algorithm with convergence guarantee, which enables to solve large-scale problems.

Ghadimi, S. and Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Jin, J., Zhang, B., Wang, H., and Wang, L. (2021). Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Advances in Neural Information Processing Systems*, pages 2771–2782. Curran Associates, Inc.

Dual Problem Formulation, Assumptions and Structures

Denote $\gamma_\zeta(\xi)$ as conditional probability over ξ. We decompose the joint distribution as $\gamma(\zeta,\xi)=\gamma_\zeta(\xi)\mathbb{P}(\zeta)$ From principle of interchangeability, the primal problem in (1) can be rewritten as

$$\min_{x\in\mathbf{R}^d}\mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\sup_{\gamma_\zeta}\Big(\mathbb{E}_{\xi\sim\gamma_\zeta}\big[\ell(x;\xi)-\lambda c(\zeta,\xi)\big]-\lambda\varepsilon D_f(\gamma_\zeta\,|\,\nu)\Big)\Big]. \tag{2}$$

By inverse C.D.F sampling, the inner supremum term $\sup_{\gamma_\zeta}(\cdot)$ has the following equivalent dual formulation

$$\underbrace{\min_{\eta\in\mathbf{R}}\Big\{L_\zeta(x,\eta):=\lambda\varepsilon\mathbb{E}_{\xi\sim\nu}\Big[f^*\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\varepsilon}\Big)\Big]+\eta\Big\}}_{L_\zeta(x,\eta_x^*(\zeta))}, \tag{3}$$

where ν is the reference measure of $\xi\sim\mathbb{Q}$; η is dual variable, f^* denotes the conjugate function of f and $\eta_x^*(\zeta)\in\arg\min_\eta L_\zeta(x,\eta)$.

For simplicity, denote

$$\Psi_\zeta(x):=L_\zeta(x,\eta_x^*(\zeta)),\quad L_{\zeta,\xi}(x,\eta):=\lambda\varepsilon f^*\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\varepsilon}\Big)+\eta.$$

Then, the **dual problem** of (1) can be written as the following problem

$$\min_{x\in\mathbf{R}^d}\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big],\ \text{where}\ \Psi_\zeta(x)=L_\zeta(x,\eta^*(\zeta)). \tag{4}$$

Challenges:

- **Double Expectation over different probability measure, $\xi\sim\nu$ and $\zeta\sim\mathbb{P}\to$** Nested Structure!
- **In-context inner minimizer $\eta^*(\zeta)$ subject to change with each ζ.$\to$** Sample Inefficiency!

In this work, we adapt following assumptions

- **Lipschitz Continuous and Smooth** $\ell(\cdot;\xi)$ For every ξ, $\ell(\cdot;\xi)$ is G-Lipschitz continuous, and $\ell(\cdot;\xi)$ is differentiable and L-smooth.
- **Smoothness of f^* Function** $f^*(\cdot)$ is differentiable and M-smooth.
- **Bounded Variance of ℓ** For every x, the variance of $\ell(x;\cdot)$ is bounded by σ^2.
- **Bounded Variance of c** For every ζ, the variance of $c(\zeta,\cdot)$ is bounded by δ^2. And for every ξ, the variance of $c(\cdot,\xi)$ is bounded by δ^2.

Why Dual Formulation can be Solved by Nested Stochastic Programming? Two Fundamental Conclusions!

- **Gradient Equivalence between $\nabla\Psi_\zeta(x)$ and $\nabla_1 L_\zeta(x,\eta_x^*(\zeta))$(Jin et al., 2021)** Let Assumptions hold and consider any fixed x and ζ. Then, the function $\Psi_\zeta(x)$ is differentiable and satisfies $\nabla\Psi_\zeta(x)=\nabla_1 L_\zeta(x,\eta_x^*(\zeta))$, where $\eta_x^*(\zeta)\in\arg\min_\eta L_\zeta(x,\eta)$.
- **Approximation Error Relationship** Suppose we obtain x and $\eta_x(\zeta)$ such that the gradient taken over second argument satisfies

$$\big|\nabla_2 L_\zeta(x,\eta_x(\zeta))\big|\le\varepsilon_1. \tag{5}$$

Then, for any ζ, the gradient taken over first argument satisfies

$$\|\nabla\Psi_\zeta(x)-\nabla_1 L_\zeta(x,\eta_x(\zeta))\|\le G\varepsilon_1. \tag{6}$$

Conclusion: As long as $\eta_x^*(\zeta)$ is near-optimal, we can guarantee $\nabla_1 L_\zeta(x,\eta_x(\zeta))$ approximate $\nabla\Psi_\zeta(x)$ with controllable error!

Proposed Algorithms, Properties and Convergence

Algorithm 1 Nested-SGD for solving $\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)]$

Input: $T\in\mathbb{N}$, initialization x_0, η_0, learning rate γ_t
for $t=0$ **to** $T-1$ **do**
  Sample $\{\zeta\}$ and $\{\xi\}_{B_1}$ with batch size B_1
  Construct estimator $\eta_{x_t}(\zeta)$ via Algorithm 2
  Compute gradient estimator $\hat{g}_t^B$ for $\nabla\Psi_\zeta(x)$
  Update $x_{t+1}=x_t-\gamma_t\hat{g}_t^B$
end for
Output: $x_{\bar{t}}$, where $\bar{t}$ is sampled from $\{0,\dots,T-1\}$ uniformly at random

Algorithm 2 Construct Estimator $\eta_x(\zeta)$

Input: $D\in\mathbb{N}$, learning rate α_d
for $d=0$ **to** $D-1$ **do**
  Utilize the ζ sampled in Algorithm 1
  Sample $\{\xi\}_{B_2}$ with batch size B_2
  Compute gradient estimator v_d^B for $\nabla_2 L_{\zeta,\xi}(x,\eta)$
  Update $\eta_{x_t}^{d+1}(\zeta)=\eta_{x_t}^d(\zeta)-\alpha_d v_d^B$
end for
Output: $\eta_{x_t}^{\bar{d}}(\zeta)$, where $\bar{d}\in\{0,\dots,D-1\}$ corresponds to the index with minimal gradient norm

- **Directional Smoothness**: For variable x and η, the following smoothness conditions hold. For any x,x', it holds that

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big\|\nabla\Psi_\zeta(x)-\nabla_1 L_\zeta(x',\eta_x^*(\zeta))\big\|^2\le K^2\|x-x'\|^2, \tag{7}$$

where $K=G^2(\lambda\varepsilon)^{-1}M+L$.
For any x and any η,η', it holds that

$$\mathbb{E}_{\xi\sim\nu}\big\|\nabla_2 L_{\zeta,\xi}(x,\eta)-\nabla_2 L_{\zeta,\xi}(x,\eta')\big\|^2\le K'^2\|\eta-\eta'\|^2, \tag{8}$$

where $K'=M(\lambda\varepsilon)^{-1}$.

- **Affine Bounded Variance:** For mini-batch gradient estimator $\hat{g}_t^B$ used in Algorithm 1, it satisfies

$$\mathbb{E}_{\zeta\sim\mathbb{P},\xi_B\sim\nu}\big\|\hat{g}_t^B\big\|^2\le R_{B_1}+\frac{8G^2\varepsilon_1^2}{B_1}+\big\|\nabla_1\mathbb{E}_{\zeta\sim\mathbb{P}}\big[L_\zeta(x_t,\eta_{x_t}(\zeta))\big]\big\|^2, \tag{9}$$

where $R_{B_1}=O\big(\frac{G^2+G^2M^2(\lambda\varepsilon)^{-2}\sigma^2}{B_1}+G^2M^2\varepsilon^{-2}\delta^{-2}\big)$.
For mini-batch gradient estimator v_d^B used in Algorithm 2, it satisfies

$$\mathbb{E}_{\xi_B\sim\nu}\big\|v_d^B\big\|^2\le\frac{R_2}{B_2}+\big\|\nabla_2 L_\zeta(x_t,\eta_{x_t}^d(\zeta))\big\|^2, \tag{10}$$

where $R_2=2M^2(\lambda\varepsilon)^{-2}(\sigma^2+\lambda^2\delta^2)$.

Convergence of Main Algorithm

Let Assumptions hold. Denote $\Delta=\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_0)\big]-\inf_x\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$. Run Nested-SGD Algorithm 1 for T iterations with learning rate $\gamma_t=\min\big\{\frac{1}{3K},\sqrt{\frac{2\Delta}{KR_{B_1}T}}\big\}$ and error threshold $\varepsilon_1(t)=\Theta(G^{-1}T^{-\frac{1}{2}})$ for all t. Then, the convergence result is

$$\mathbb{E}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\le O\Big(\sqrt{\frac{\Delta KR_{B_1}}{T}}\Big)+O\Big(\frac{\Delta K}{T}\Big)+O\Big(\frac{B_1^{-1}\sqrt{\Delta K/R_{B_1}}}{T^{3/2}}\Big). \tag{11}$$

Moreover, to achieve $\mathbb{E}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|\le\delta_1$, choose $B_1=\Theta(1)$, then the sample complexity of Algorithm 1 is $\Omega(\Delta KR_{B_1}\delta_1^{-4})$.

For Algorithm 2(1-dimension stochastic programming), the convergence analysis follows the standard analysis (Ghadimi and Lan, 2013). The difference is we use $B_2=\Theta(\varepsilon^{-2})$ mini-batch size to ensure convergence.

yufeng.yang@tamu.edu