An Online Method for A Class of Distributionally Robust Optimization with Non-Convex Objectives

Qi Qi*† , Zhishuai Guo*†, Yi Xu‡, Rong Jin‡, Tianbao Yang†

† Department of Computer Science, University of Iowa, Iowa City, IA 52242 ‡ Machine Intelligence Technology, Alibaba Group

{qi-qi, zhishuai-guo, tianbao-yang}@uiowa.edu,{yixu, jinrong.jr}@alibaba-inc.com * Equal contribution

Background and Problem Definition

Distributionally Robust Optimization (DRO):

- Problem Formulation:

$$\min_{\mathbf{w}\in R^d}\max_{\mathbf{p}\in\Delta_n}\left\{F_p(\mathbf{w})=\sum_{i=1}^n p_i\ell(\mathbf{w};\mathbf{z}_i)-h(\mathbf{p},\mathbf{1}/n)+r(\mathbf{w})\right\}, \quad (1)$$

1. where $\Delta_n=\{\mathbf{p}\in R^n:\sum_i p_i=1, p_i\geq 0\}$.
2. $\ell(\mathbf{w};\mathbf{z})$: denotes a loss function on data $\mathbf{z}=(\mathbf{x},y_l)\sim\mathbf{D}$,
3. $h(\mathbf{p},\mathbf{1}/n)$: a divergence measure
4. $r(\mathbf{w})$ is convex regularizer of $\mathbf{w}$

Non-Convex Concave Min-Max Optimization

Related Work

Primal-Dual algorithms:

- PG-SMD [1], Epoch Primal-Dual SGD [2], $O(1/\varepsilon^4)$
- Stoc-AGDA [3], PE-SGDA [4]
  - $O(1/\mu^2\varepsilon)$ by leveraging PL Condition

Deficiencies:

- $O(n)$ computational cost per $\mathbf{p}$ variable updates.

Note: As the computational cost is intolerable for updating $\mathbf{p}$ variable when the data size n is huge, Primal-Dual algorithms are not as favorable as SGD-type algorithms for DL.

Research Question

Can we design an efficient algorithm that is independent of data size n, has faster convergence rates and is also applicable to deep neural network training? Yes

Compositional Equivalence of $F_{\mathbf{p}}(\mathbf{w})$

Considering the KL-divergence measure, e.g., $h(\mathbf{p},\mathbf{1}/n)=-\sum_{i=1}^n p_i\log(np_i)$

$$\min_{\mathbf{w}\in\mathbb{R}^d}\left\{F_{dro}(\mathbf{w})=\lambda\log\left(\mathbb{E}_{\mathbf{z}}\exp\left(\frac{\ell(\mathbf{w};\mathbf{z})}{\lambda}\right)\right)+r(\mathbf{w})\right\}. \quad (2)$$

General Compositional Optimization

Assumptions

Assumption 1: (Non-convex Setting) Let C_f, L_f, C_g and L_g be positive constants. Assume that

(a) $f:\mathbb{R}^p\to\mathbb{R}$ is a C_f-Lipschitz function and its gradient ∇f is L_f-Lipschitz.
(b) $g_{\mathbf{z}}:\mathbb{R}^d\to\mathbb{R}^p$ satisfies $\mathbb{E}\|g_{\mathbf{z}}(\mathbf{w}_1)-g_{\mathbf{z}}(\mathbf{w}_2)\|^2]\leq C_g^2\|\mathbf{w}_1-\mathbf{w}_2\|^2$ for any $\mathbf{w}_1,\mathbf{w}_2$ and its Jacobian $\nabla g_{\mathbf{z}}$ satisfies $\mathbb{E}[\|\nabla g_{\mathbf{z}}(\mathbf{w}_1)-\nabla g_{\mathbf{z}}(\mathbf{w}_2)\|^2]\leq L_g^2\|\mathbf{w}_1-\mathbf{w}_2\|^2$.
(c) $r:R^d\to\mathbb{R}\cup\{\infty\}$ is a convex and lower-semicontinuous function.
(d) $F_*=\inf_{\mathbf{w}}F(\mathbf{w})\geq-\infty$ and $F(\mathbf{w}_1)-F_*\leq\Delta_F$ for the initial solution $\mathbf{w}_1$.
(e) Let σ_g and $\sigma_{g'}$ be positive constants and $\sigma^2=\sigma_g^2+\sigma_{g'}^2$. Assume that

$$\mathbb{E}_{\mathbf{z}}[\|g_{\mathbf{z}}(\mathbf{w})-g(\mathbf{w})\|^2]\leq\sigma_g^2,\ \mathbb{E}_{\mathbf{z}}[\|\nabla g_{\mathbf{z}}(\mathbf{w})-\nabla g(\mathbf{w})\|^2]\leq\sigma_{g'}^2.$$

Assumption 2: (PL Condition Setting) When $r(\mathbf{w})$ is a smooth function, $F(\mathbf{w})$ satisfies the μ-PL condition if there exists $\mu>0$ such that

$$2\mu(F(\mathbf{w})-\min_{\mathbf{w}\in\mathbb{R}^d}F(\mathbf{w}))\leq\|\nabla F(\mathbf{w})\|^2. \quad (3)$$

Literature

[1] Rafique, H., Liu, M., Lin, Q., Yang, T. (1810). Non-convex min–max optimization: provable algorithms and applications in machine learning (2018).
[2] Yan, Y., Xu, Y., Lin, Q., Liu, W., Yang, T. (2020). Sharp analysis of epoch stochastic gradient descent ascent methods for min-max optimization.

Proposed Stochastic Algorithms: COVER & RECOVER

Proximal Gradient Convergence Measure:

$$\mathscr{G}_\eta(\mathbf{w})=\frac{1}{\eta}(\mathbf{w}-\text{prox}_r^\eta(\mathbf{w}-\eta\nabla g(\mathbf{w})^\top\nabla f(g(\mathbf{w})))).$$

When $r=0$, the proximal gradient reduces to the standard gradient measure, i.e., $\mathscr{G}_\eta(\mathbf{w})=\nabla F(\mathbf{w})$.

Sample Complexity: The **sample complexity** is defined as the number of samples $\mathbf{z}$ in order to achieve $\mathbb{E}[\|\mathscr{G}_\eta(\mathbf{w})\|^2]\leq\varepsilon$ for a certain $\eta>0$ or $\mathbb{E}[F(\mathbf{w})-F_*]\leq\varepsilon$.

Algorithm 1 COVER $(\mathbf{w}_1,u_1,\mathbf{v}_1,\{\eta_t\},T,\text{PL}=\text{False})$

Let $a_t=c\eta_t^2$
if not PL **then**
  Draw a samples $\mathbf{z}$ and construct the estimates: $u_1=g_{\mathbf{z}}(\mathbf{w}_1),\ \mathbf{v}_1=\nabla g_{\mathbf{z}}(\mathbf{w}_1)$
end if
for $t=1,\ldots,T-1$ **do**
  $\mathbf{w}_{t+1}\leftarrow\text{prox}_r^{\eta_t}(\mathbf{w}_t-\eta_t\mathbf{v}_t^\top\nabla f(u_t))$
  Draw a samples $\mathbf{z}_{t+1}$, and update

$$u_{t+1}=g_{\mathbf{z}_{t+1}}(\mathbf{w}_{t+1})+(1-a_{t+1})(u_t-g_{\mathbf{z}_{t+1}}(\mathbf{w}_t))$$
$$\mathbf{v}_{t+1}=\nabla g_{\mathbf{z}_{t+1}}(\mathbf{w}_{t+1})+(1-a_{t+1})(\mathbf{v}_t-\nabla g_{\mathbf{z}_{t+1}}(\mathbf{w}_t))$$

end for
Return: $(\mathbf{w}_\tau,u_\tau,\mathbf{v}_\tau)$ for randomly selected $\tau\in\{1,\ldots,T\}$.

Algorithm 2 RECOVER$(\mathbf{w}_0,\varepsilon_0,c)$

Initialization: Draw a sample $\mathbf{z}_0$ and construct the estimates $u_0=g_{\mathbf{z}_0}(\mathbf{w}_0),\ \mathbf{v}_0=\nabla g_{\mathbf{z}_0}(\mathbf{w}_0)$
for $k=1,\ldots,K$ **do**
  $(\mathbf{w}_k,u_k,\mathbf{v}_k)=\text{COVER}(\mathbf{w}_{k-1},u_{k-1},\mathbf{v}_{k-1},\eta_k,T_k,\text{True})$
  change η_k,T_k according to Theorem RECOVER
end for
Return: $\mathbf{w}_K$

Theorem COVER (Non-convex)

Assume Assumption 1 holds, for any $C>0$, $k=\frac{C\sigma^{2/3}}{L}$, $c=128L+\sigma^2/(7Lk^3)$, $w=\max((16Lk^3),2\sigma^2,(\frac{ck}{4L})^3)$, and $\eta_t=k/(w+\sigma^2t)^{1/3}$. The output of COVER satisfies

$$\mathbb{E}[\|\mathscr{G}_{\eta_{t_*}}(\mathbf{w}_{t_*})\|^2]\leq\widetilde{O}\left(\frac{\Delta_F}{T^{2/3}}+\frac{\sigma^2}{T^{2/3}}\right). \quad (4)$$

where t_* is sampled from $\{1,\ldots,T\}$.

Theorem RECOVER (PL Condition)

Assume that assumption 1, 2 hold and define constants $\varepsilon_1=\frac{c^2\sigma^2}{64\mu L^4}$ and $\varepsilon_k=\varepsilon_1/2^{k-1}$. By setting $\eta_k=\min\{\frac{\sqrt{\mu\varepsilon_k}L}{2c\sigma},\frac{1}{16L}\}$, $T_k=O(\max\{\frac{96c\sigma}{\mu^{3/2}\sqrt{\varepsilon_k}L},\frac{2c^2\sigma^2}{\mu L^2\varepsilon_k},\frac{\Delta_F}{\sigma^2}\})$, $c=104L^2$, then after $K=O(\log(\varepsilon_1/\varepsilon))$ stages, the output of RECOVER satisfies $\mathbb{E}[F(\mathbf{w}_K)-F_*]\leq\varepsilon$.

[3] Yang, J., Kiyavash, N., He, N. (2020). Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems.
[4] Guo, Z., Yuan, Z., Yan, Y., Yang, T. (2020). Fast Objective Duality Gap Convergence for Nonconvex-Strongly-Concave Min-Max Problems.
[5] Wang, M., Liu, J., Fang, E. X. (2017). Accelerating stochastic composition optimization. Journal of Machine Learning Research.
[6] Zhang, J., Xiao, L. (2019). A stochastic composite gradient method with incremental variance reduction. Advances in Neural Information Processing Systems

Summary of properties of state-of-the-art algorithms for solving our DRO problem. The sample complexity is measured in terms of finding an ε-stationary point w/o PL condition, i.e., $\|\nabla F(\mathbf{w})\|^2\leq\varepsilon$, or achieving ε-objective gap, i.e, $F(\mathbf{w})-\min_{\mathbf{w}}F(\mathbf{w})\leq\varepsilon$ with PL condition. $\widetilde{O}$ omits a logarithmic dependence over ε. n represents the size of datasets for a finite sum problem, d denotes the dimension of $\mathbf{w}$. GDS represents whether the step size is geometrically decreased.

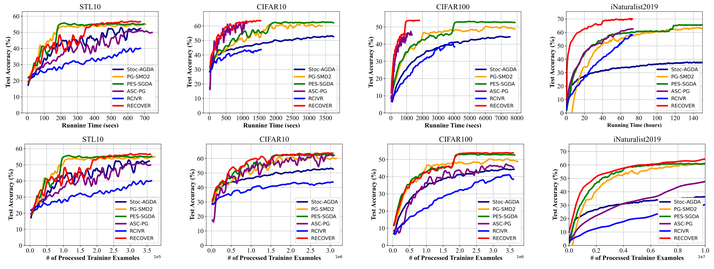
Settings / Cost	Algorithms / Style	Sample Complexity	batch size	GDS η	Memory	
w/o PL	PG-SMD2 [1]	$O(n/\varepsilon+1/\varepsilon^2)$	$O(1)$	x	$O(n+d)$	Primal-Dual
	ASC-PG [5]	$O(1/\varepsilon^2)$	$O(1)$	x	$O(d)$	Compositional
	CIVR [6]	$O(1/\varepsilon^{3/2})$	$O(1/\varepsilon)$	x	$O(d)$	Compositional
	COVER (This paper)	$\widetilde{O}(1/\varepsilon^{3/2})$	$O(1)$	x	$O(d)$	Compositional
w/ PL	Stoc-AGDA [3]	$O(1/\mu^2\varepsilon)$	$O(1)$	x	$O(n+d)$	Primal-Dual
	PES-SGDA [4]	$O(1/\mu^2\varepsilon)$	$O(1)$	✓	$O(n+d)$	Primal-Dual
	RCIVR [6]	$\widetilde{O}(1/\mu\varepsilon)$	$O(1/\varepsilon)$	x	$O(d)$	Compositional
	RECOVER (This paper)	$O(1/\mu\varepsilon)$	$O(1)$	✓	$O(d)$	Compositional

Emperical Studies on Multi-class Imbalance Deep Learning Tasks

Datasets:

	First Half	Last Half	batch	Classes	Size	Network Arch
STL10	100	500	32	10	5000	Resnet20
Clifar10	100	5000	128	10	50000	Resnet20
CIFAR100	100	500	128	100	50000	Resnet20
iNaturalist 2019	Practical Imbalanced Datatset		64	1010	265,213	Inception-V3

Evaluation: **Testing accuracy vs GPU time**
Objective: $\lambda=5$, $\ell(w,z)$ cross entropy loss.



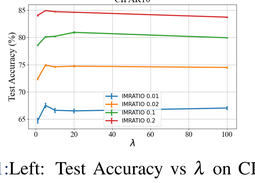
RECOVER optimized DRO vs SGD optimized ERM

We compare the test accuracy learned by optimizing DRO using RECOVER and optimizing ERM using SGD on the imbalanced datasets: STL10, CIFAR10, CIFAR100, with four imbalance ratio $\rho=\{0.02,0.05,0.1,0.2\}$.
Table:Test accuracy (%), mean (std), of SGD for ERM and RECOVER for DRO over 5 independent runs.

IMRATIO	STL10		CIFAR10		CIFAR100	
	SGD	RECOVER	SGD	RECOVER	SGD	RECOVER
0.02	37.97 (0.78)	**38.08 (0.59)**	65.36(0.64)	**66.14 (0.48)**	38.99 (0.62)	**39.45 (0.56)**
0.05	41.12 (0.94)	**42.68 (0.60)**	74.74 (0.71)	**75.90 (0.33)**	45.79 (0.69)	**44.47 (0.66)**
0.1	46.03 (0.96)	**48.94 (0.86)**	79.32 (0.42)	**80.93 (0.31)**	49.45 (0.5)	**50.84 (0.86)**
0.2	51.75 (1.14)	**56.06 (1.26)**	84.84 (0.51)	**85.93 (0.14)**	55.80 (0.74)	**56.90 (0.42)**

Efficient as Fine-tune Methods

Dataset: ImageNet-LT, Places-LT **Models:** ResNet50, ResNet152



Model	ImageNet-LT	Places-LT
Pretrained	40.50	23.28
CE (SGD)	41.29 (3e-3)	27.47 (1e-3)
Focal (SGD)	41.10 (2e-2)	27.64 (6e-3)
DRO (RECOVER)	**42.30** (4e-4)	**28.75** (4e-5)

Figure 1:Left: Test Accuracy vs λ on CIFAR10 data; Right: Test accuracy (%) of finetuned models by different methods.

Code can be found at: `https://github.com/qiqi-helloworld/RECOVER`