A Projection-free Algorithm for Constrained Stochastic Composition Optimization

Speaker: Tesi Xiao

PhD Candidicate in Statistics

Department of Statistics, UC Davis

NeurIPS, 2022

Joint work with

Krishnakumar Balasubramanian
Department of Statistics
UC Davis

Saeed Ghadimi
Department of Management Sciences
University of Waterloo

Problem

Consider the following multi-level composition optimization problem:

$$\min_{x \in \mathcal{X}} \quad F(x) := f_1 \circ \cdots \circ f_T(x), \tag{1}$$

where

- $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}, i = 1, ..., T$ are continuously differentiable $(d_0 = 1)$;
- F is bounded below by $F^\star > -\infty$;
- $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set
- $F(x)$ is possibly nonconvex

Setting

Our goal is to design online projection-free algorithms solving the above optimization problem, given access to noisy evaluations of ∇f_i's and f_i's.

- nonconvex + multi-level
- fully online manner: one sample, no min-batch
- projection-free algorithm: conditonal gradient based methods
- stochastic setting: only Stochatic Zeroth/First-order Oracle (SZO/SFO) is accessible

Challenges

Consider solving the two-level stochastic composition optimization

$$\min_{x \in \mathcal{X}} \quad F(x) := f_1(f_2(x)), \tag{2}$$

given access to noisy evaluations of ∇f_1, f_2, and ∇f_2.

- Vanilla SGD performs poorly due to the biasedness:

$$\mathbb{E}[\widetilde{\nabla f}_1(\widetilde{f}_2(x)) \cdot \widetilde{\nabla f}_2(x)] \quad \neq \quad \nabla f_1(f_2(x)) \cdot \nabla f_2(x) = \nabla F(x)$$

  mini-batch stochastic gradient estimators lead to oracle complexities that depend exponentially on T.
- Most existing projection-free algorithms require increasing order of mini-batches [1]; some recent one-sample variants require stronger assumptions or are not in the fully online manner [2].

[1] [LZ16, RSPS16, HL16, QLX18, YSC19]
[2] [ZSM+20, ABTR21]

Our Method: Moving Average Estimator

Auxiliary sequences cumulatively estimate the inner function values

$$u_i^k \longrightarrow f_i(u_{i+1}^k), \quad i = 1, \ldots, T, \quad (u_{T+1}^k = x^k)$$

and the gradient of $F(x)$

$$z_k \longrightarrow \nabla F(x^k).$$

E.g., $T = 2$: for some $\tau_k \in [0, 1)$

$$\begin{aligned} u^{k+1} &= (1 - \tau_k)u^k + \tau_k \widetilde{f}_2(x^k) \\ z^{k+1} &= (1 - \tau_k)z^k + \tau_k \widetilde{\nabla f_2}(x^k)^\top \widetilde{\nabla f_1}(u^k) \end{aligned}$$

The idea is also referred to the Averaged Stochastic Approxmiation (ASA) and Dual Averaging.

Our Method: Conditional Gradient Sliding

The projection step at the iterate x^k with the gradient estimate z^k and stepsize $1/\beta$,

$$\tilde{x} = \mathsf{Proj}_{\mathcal{X}}\left(x^k - \frac{1}{\beta} z^k\right),$$

can be written in the form of

$$\underset{\tilde{x} \in \mathcal{X}}{\arg\min} \left\{ \langle z^k, \tilde{x} \rangle + \frac{\beta}{2} \|\tilde{x} - x^k\|^2 \right\},$$

which is a constrained quadratic minimization problem that can be solved by iteratively running Frank-Wolfe method with the exact line search.

Solving projection subproblems via the Frank-Wolfe algorithm is known as conditional gradient sliding.

Frank-Wolfe method with the exact line search

Algorithm 2 Inexact Conditional Gradient Method (ICG)

Input: (x, z, β, M, δ)
Set $w^0 = x$.
for $t = 0, 1, 2, \ldots, M$ **do**
1. Find $v^t \in \mathcal{X}$ with a quantity $\delta \geq 0$ such that

$$\langle z + \beta(w^t - x), v^t \rangle \leq \min_{v \in \mathcal{X}} \langle z + \beta(w^t - x), v \rangle + \frac{\beta D_{\mathcal{X}}^2 \delta}{t+2}.$$

2. Set $w^{t+1} = (1 - \mu_t) w^t + \mu_t v^t$ with $\mu_t = \min \left\{ 1, \frac{\langle \beta(x - w^t) - z, v^t - w^t \rangle}{\beta \| v^t - w^t \|^2} \right\}$.

end for
Output: w^M

Remark

The exact solution to the linear minimization problem is not required.

Our Algorithm: Linearized NASA with ICG Method

1. Update the solution:

$$\tilde{y}^k = \texttt{ICG}(x^k, z^k, \beta_k, t_k, \delta),$$
$$x^{k+1} = x^k + \tau_k(\tilde{y}^k - x^k),$$

and compute stochastic Jacobians J_i^{k+1}, and function values G_i^{k+1} at u_{i+1}^k for $i = 1, \ldots, T$.

2. Update average gradients z and function value estimates u_i for each level $i = 1, \ldots, T$

$$z^{k+1} = (1-\tau_k)z^k + \tau_k \prod_{i=1}^{T} J_{T+1-i}^{k+1},$$
$$u_i^{k+1} = (1-\tau_k)u_i^k + \tau_k G_i^{k+1} + \langle J_i^{k+1}, u_{i+1}^{k+1} - u_{i+1}^k \rangle.$$

Linearization helps to get rid of level-dependent batch size

Notions of Stationarity

Definition

A point $\bar{x} \in \mathcal{X}$ generated by an algorithm is called an ϵ-stationary point in terms of GM, if we have $\mathbb{E}[\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\|^2] \leq \epsilon$. A point $\bar{x} \in \mathcal{X}$ generated by an algorithm is called an ϵ-stationary point in terms of FW-gap, if we have $\mathbb{E}[g_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}))] \leq \epsilon$.

- Gradient Mapping (GM):

$$\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta) := \beta \left(\bar{x} - \Pi_{\mathcal{X}} \left(\bar{x} - \frac{1}{\beta} \nabla F(\bar{x}) \right) \right)$$

- Frank-Wolfe Gap:

$$g_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x})) := \max_{y \in \mathcal{X}} \langle \nabla F(\bar{x}), \bar{x} - y \rangle.$$

Proposition (Translation)

- $\|\mathcal{G}_{\mathcal{X}}(x, \nabla F(x), \beta)\|^2 \leq g_{\mathcal{X}}(x, \nabla F(x)), \forall x \in \mathcal{X}$.
- *Under regular conditions: (i)* $\mathcal{X} \subset \mathbb{R}^d$ *is convex and closed with diameter* $D_{\mathcal{X}} > 0$*; (ii)* $f_1, \ldots, f_T$ *and their derivatives are Lipschitz continuous, we have* $g_{\mathcal{X}}(x, \nabla F(x)) \leq \left[(1/\beta) \prod_{i=1}^{T} L_{f_i} + D_{\mathcal{X}} \right] \|\mathcal{G}_{\mathcal{X}}(x, \nabla F(x), \beta)\|$.

Main Results

Theorem

Under regular conditions:

- $\mathcal{X} \subset \mathbb{R}^d$ *is convex and closed with diameter* $D_{\mathcal{X}} > 0$*;*
- $f_1, \ldots, f_T$ *and their derivatives are Lipschitz continuous;*
- J_i^k, G_i^k*'s are unbiased, mutually independent, and have bounded second moment.*

Let $\{x^k, z^k, \{u_i^k\}_{1 \le i \le T}\}_{k \ge 0}$ *be the sequence generated by* ***LiNASA+ICG*** *with* $N \ge 1, \tau_0 = 1, t_0 = 0$ *and*

$$\beta_k \equiv \beta > 0, \quad \tau_k = \frac{1}{\sqrt{N}}, \quad t_k = \lceil \sqrt{k} \rceil, \quad \forall k \ge 1,$$

we have $\mathbb{E}\left[\|\mathcal{G}_{\mathcal{X}}(x, \nabla F(x), \beta)\|^2\right] \le \mathcal{O}_T(N^{-1/2})$,

$$\mathbb{E}\left[\|f_i(u_{i+1}^R) - u_i^R\|^2\right] \le \mathcal{O}_T(N^{-1/2}),\ 1 \le i \le T,\ u_{T+1} = x$$

The random integer number R *is uniformly distributed over* $\{1, \ldots, N\}$*.*

Main Results

Table: Complexity results for stochastic conditional gradient type algorithms to find an ϵ-stationary solution in the nonconvex setting.

Algorithm	Criterion	# of levels	Batch size	SFO	LMO
SPIFER-SFW [YSC19]	FW-gap (GM)	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
1-SFW [ZSM+20]	FW-gap (GM)	1	1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
SCFW [ABTR21]	FW-gap (GM)	2	1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
SCGS [QLX18]	GM	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
SGD+ICG [BG21]	GM	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
LiNASA+ICG	GM	T	1	$\mathcal{O}_T(\epsilon^{-2})$	$\mathcal{O}_T(\epsilon^{-3})$

$\mathcal{O}_T$ hides constants in T.

Existing one-sample based stochastic conditional gradient algorithms are either (i) not applicable to the case of general $T > 1$, or (ii) require strong assumptions [ZSM+20], or (iii) are not truly online [ABTR21]. The results in [BG21] are actually presented for the zeroth-order setting; however the above stated first-order complexities follow immediately.

High-probability Results for $T = 1$

- No existing work present high-probability results for nonconvex constrainted stochastic optimization problems.
- [MDB21] identify the technical difficulities of obtaining high-probability results of projected SGD in the non-convex setting.

Algorithm: ASA+ICG

Update the solution:

$$\tilde{y}^k = \texttt{ICG}(x^k, z^k, \beta_k, t_k, \delta),$$
$$x^{k+1} = x^k + \tau_k(\tilde{y}^k - x^k).$$

Update the average gradient:

$$z^{k+1} = (1 - \tau_k)z^k + \tau_k J_1^{k+1}$$

High-probability Results for $T = 1$

Definition

A point $\bar{x} \in \mathcal{X}$ generated by our algorithm is called an (ϵ, δ)-stationary point, if we have $\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\|^2 \leq \epsilon$ with probability $1 - \delta$.

Assumption

Let $\Delta^{k+1} = \nabla F(x^k) - J_1^{k+1}$ *for* $k \geq 0$. *For each* k, *given* $\mathscr{F}_k$ *we have* $\mathbb{E}[\Delta^{k+1} | \mathscr{F}_k] = 0$ *and* $\|\Delta^{k+1}\| \mid \mathscr{F}_k$ *is* K*-sub-Gaussian.*

Theorem

Let $\tau_0 = 1, t_0 = 0$, $\tau_k = \frac{1}{\sqrt{N}}, t_k = \lceil \sqrt{k} \rceil, \forall k \geq 1$, *where* N *is the total number of iterations. Let* $T = 1$ *and let* $\{x^k, z^k\}_{k \geq 0}$ *be the sequence generated by* `ASA+ICG` *with* $\beta_k \equiv \beta > 0$. *Then, under above assumptions, we have* $\forall N \geq 1, \delta > 0$, *with probability at least* $1 - \delta$,

$$\min_{k=1,\ldots,N} \left\| \mathcal{G}_{\mathcal{X}}(x^k, \nabla F(x^k), \beta) \right\|^2 \leq \mathcal{O}\left(\frac{K^2 \log(1/\delta)}{\sqrt{N}} \right)$$

Therefore, the number of calls to SFO and LMO to get an (ϵ, δ)*-stationary point is upper bounded by* $\mathcal{O}(\epsilon^{-2} \log^2(1/\delta)), \mathcal{O}(\epsilon^{-3} \log^3(1/\delta))$ *respectively.*

Conclusion

1. `LiNASA+ICG` is completely parameter-free for any $T \geq 1$:
   - arbitrary step size $\beta > 0$;
   - sliding parameter $\tau_k = \frac{1}{\sqrt{N}}$, N is the total number of iterations;
   - number of CG updates $t_k = \lceil \sqrt{k} \rceil$, i.e., accurate ICG solutions are not required for all iterations.

2. $T = 1$, we provide the fisrt high-probability results for nonconvex constrained stochastic optimization.

Thanks for Listening!

Reference I

[ABTR21] Zeeshan Akhtar, Amrit Singh Bedi, Srujan Teja Thomdapu, and Ketan Rajawat. Projection-Free Algorithm for Stochastic Bi-level Optimization. *arXiv preprint arXiv:2110.11721*, 2021.

[BG21] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-Order Nonconvex Stochastic Optimization: Handling Constraints, High Dimensionality, and Saddle Points. *Foundations of Computational Mathematics*, pages 1–42, 2021.

[HL16] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

[LZ16] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[MDB21] Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High-probability convergence bounds for non-convex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2021.

[QLX18] Chao Qu, Yan Li, and Huan Xu. Non-convex conditional gradient sliding. In *International Conference on Machine Learning*, pages 4208–4217. PMLR, 2018.

Reference II

[RSPS16] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1244–1251. IEEE, 2016.

[YSC19] Alp Yurtsever, Suvrit Sra, and Volkan Cevher. Conditional gradient methods via stochastic path-integrated differential estimator. In *International Conference on Machine Learning*, pages 7282–7291. PMLR, 2019.

[ZSM+20] Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One-sample Stochastic Frank-Wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020.