A Projection-free Algorithm for Constrained Stochastic Multi-level Composition Optimization

Tesi Xiao[1], Krishnakumar Balasubramanian[1], Saeed Ghadimi[2]
[1]Department of Statistics, Univeristy of California, Davis
[2]Department of Management Sciences, University of Waterloo

Introduction

We consider the following multi-level composition optimization problem:

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

where $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}, i = 1, ..., T$ are continuously differentiable $(d_0 = 1)$, $F(x)$ is possibly nonconvex and bounded below by $F^\star > -\infty$, and $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set. Our goal is to design online projection-free algorithms solving the above problem, given access to noisy evaluations of ∇f_i's and f_i's.

Algorithm

Linearized Nested Averaged Stochastic Approximation with Inexact Conditional Gradient Methods (`LiNASA+ICG`)

Input: $x^0 \in \mathcal{X}$, $z^0 = 0 \in \mathbb{R}^d$, $u_i^0 \in \mathbb{R}^{d_i}, i = 1, \ldots, T$, $\beta_k > 0$, $t_k > 0$, $\tau_k \in (0, 1]$, $\delta \geq 0$.
for $k = 0, 1, 2, \ldots, N$ **do**
1. Update the solution:

$$\underbrace{\tilde{y}^k = \texttt{ICG}(x^k, z^k, \beta_k, t_k, \delta),}_{\text{inexact solution of the projection step by Frank-Wolfe methods}} \qquad x^{k+1} = x^k + \tau_k(\tilde{y}^k - x^k),$$

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 $i = 1, \ldots, T$

$$z^{k+1} = (1 - \tau_k) z^k + \tau_k \underbrace{\prod_{i=1}^{T} J_{T+1-i}^{k+1}}_{\text{biased gradient obtained by chain rule}},$$

$$u_i^{k+1} = (1 - \tau_k) u_i^k + \tau_k G_i^{k+1} + \underbrace{\langle J_i^{k+1}, u_{i+1}^{k+1} - u_{i+1}^k \rangle}_{\text{linearization helps to get rid of level-dependent batch size}}.$$

end for
Output: $(x^R, z^R, u_1^R, \cdots, u_T^R)$, where R is uniformly distributed over $\{1, 2, \ldots, N\}$

Procedure $\texttt{ICG}(x, z, \beta, M, \delta)$
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

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

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

Theoretical Results

Measure of Non-stationarity

- Gradient Mapping (GM): $\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta) := \beta \left(\bar{x} - \Pi_{\mathcal{X}} \left(\bar{x} - \frac{1}{\beta} \nabla F(\bar{x})\right)\right)$
  A point $\bar{x} \in \mathcal{X}$ generated by an algorithm is called an ϵ-stationary point in terms of GM, if we have $\mathbb{E}[\|\mathcal{G}_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}), \beta)\|^2] \leq \epsilon$.
- Frank-Wolfe Gap: $g_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x})) := \max_{y \in \mathcal{X}} \langle \nabla F(\bar{x}), \bar{x} - y \rangle$
  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 $\mathbb{E}[g_{\mathcal{X}}(\bar{x}, \nabla F(\bar{x}))] \leq \epsilon$.

Main Results

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

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

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

$$\mathbb{E}\left[\|\mathcal{G}_{\mathcal{X}}(x^R, \nabla F(x^R), \beta)\|^2\right] \leq \mathcal{O}_T\left(1/\sqrt{N}\right).$$

High-probability Bound for $T = 1$

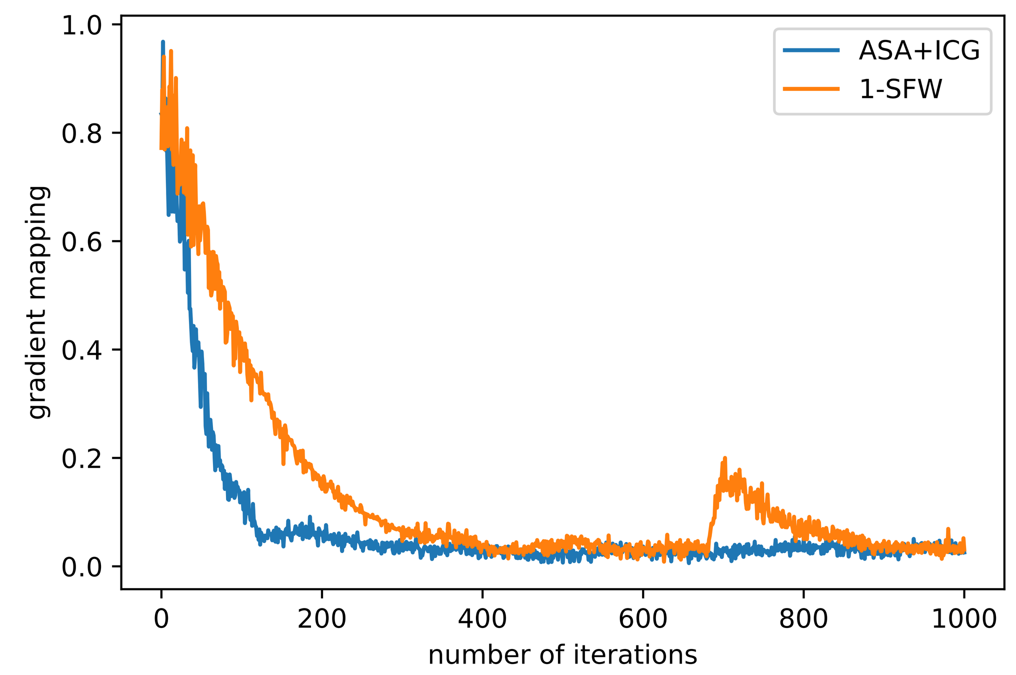
Let $\Delta^{k+1} = \nabla F(x^k) - J_1^{k+1}$ for $k \geq 0$. For each k, given $\mathscr{F}_k$ we have $\mathbb{E}[\Delta^{k+1} | \mathscr{F}_k] = 0$ and $\|\Delta^{k+1}\| \mid \mathscr{F}_k$ is K-sub-Gaussian. Let $\tau_0 = 1, t_0 = 0$, $\tau_k = \frac{1}{\sqrt{N}}, t_k = \lceil \sqrt{k} \rceil, \forall k \geq 1$. 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 $1 - \delta$,

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

Experimental Results

To recover a low-rank matrix B from the following matrix-valued single-index model with low-rank constraints: $y = |\langle A, B^\star \rangle_F|^2 + \epsilon, \text{rank}(B^\star) \leq s$, , one can optimize the mean squared loss with nuclear norm constraint, in which the Frank-Wolfe update is much cheaper than the projection operator especially with large-scale matrices.

$$\min \ F(B) = \mathbb{E}_{A,\epsilon}\left[(y - |\langle A, B \rangle_F|^2)^2\right] \quad \text{s.t. } \|B\|_\star \leq s.$$



Contributions

Complexity results for stochastic conditional gradient type algorithms to find an ϵ-stationary solution in the nonconvex setting. (**SFO**: Stochastic First-order Oracle; **LMO**: Linear Minimization Oracle)

Algorithm	Criterion	# of levels	Batch size	SFO	LMO
SPIFER-SFW [4]	FW-gap (GM)	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
1-SFW [5]	FW-gap (GM)	1	1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
SCFW [1]	FW-gap (GM)	2	1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$
SCGS [3]	GM	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
SGD+ICG [2]	GM	1	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
LiNASA+ICG	GM	T	1	$\mathcal{O}_T(\epsilon^{-2})$	$\mathcal{O}_T(\epsilon^{-3})$

- 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 [5], or (iii) are not truly online [1]
- `LiNASA+ICG` is completely parameter-free for any $T \geq 1$
- $T = 1$, we provide the fisrt high-probability results for nonconvex constrained stochastic optimization

References

[1] Z. Akhtar, A. S. Bedi, S. T. Thomdapu, and K. Rajawat. Projection-Free Algorithm for Stochastic Bi-level Optimization. *arXiv preprint arXiv:2110.11721*, 2021.

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

[3] C. Qu, Y. Li, and H. Xu. Non-convex conditional gradient sliding. In *International Conference on Machine Learning*, pages 4208–4217. PMLR, 2018.

[4] A. Yurtsever, S. Sra, and V. Cevher. Conditional gradient methods via stochastic path-integrated differential estimator. In *International Conference on Machine Learning*, pages 7282–7291. PMLR, 2019.

[5] M. Zhang, Z. Shen, A. Mokhtari, H. Hassani, and A. Karbasi. One-sample Stochastic Frank-Wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020.