Two Recent Results on Stochastic Multi-level Composition Optimization

KRISHNA BALASUBRAMANIAN

DEPARTMENT OF STATISTICS, UC DAVIS

▷ Joint work with:

  ▷ Saeed Ghadimi, University of Waterloo.
  ▷ Anthony Nguyen and Tesi Xiao, UC Davis.

▷ Papers available in arXiv:

  ▷ https://arxiv.org/pdf/2008.10526.pdf (SIOPT, 2022).
  ▷ https://arxiv.org/pdf/2202.04296.pdf (under review).

- Multi-level stochastic composition optimization problem:

$$\min_{x \in X} \left\{ F(x) = f_1 \circ \cdots \circ f_T(x) \right\} \tag{1}$$

- Functions $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}$ for $i = 1, \ldots, T$ are continuously differentiable. Here $d_0 := 1$.

- Feasible set X is either $\mathbb{R}^{d_T}$ or a closed convex constraint set.

- Stochastic setup:
  - $f_i(y) := \mathbb{E}_{\xi_i}[G_i(y, \xi_i)]$ for random vectors $\xi_i \in \mathbb{R}^{\tilde{d}_i}$.
  - When $T = 1$, we have the well-studied standard stochastic optimization or (population) risk minimization problem.

- Simple example for $T = 2$: Minimizing variance instead of expectation.

- Mean-deviation risk-averse optimization is given by the following form

$$\max_x \left\{ \mathbb{E}[U(x,\xi)] - \lambda \mathbb{E}\left[\{ \mathbb{E}[U(x,\xi)] - U(x,\xi) \}^2 \right]^{1/2} \right\}.$$

- As noted in several prior works, the above problem is a stochastic 3-level composition optimization problem with

$$\begin{aligned} f_3 &:= \mathbb{E}[U(x,\xi)] \\ f_2(z,x) &:= \mathbb{E}[\{z - U(x,\xi)\}^2] \\ f_1((y_1,y_2)) &:= y_1 - \sqrt{y_2 + \delta}. \end{aligned}$$

- Sparse additive modeling in non-parametric statistics [Wang et al., 2017].
- Area Under the Precision-Recall Curve (AUPRC) maximization [Qi et al., 2021, Wang et al., 2022, Qiu et al., 2022].
- Bayesian optimization [Astudillo and Frazier, 2021].
- Model-agnostic meta-learning [Chen et al., 2021, Fallah et al., 2021].
- Training Graph Neural Networks [Cong et al., 2020].

- Gradient of $F(x)$ is

$$\nabla F(x) = \nabla f_T(y_T)\nabla f_{T-1}(y_{T-1}) \cdots \nabla f_1(y_1),$$

where ∇f_i denotes the transpose of the Jacobian of f_i, and

$$y_i = f_{i+1} \circ \cdots \circ f_T(x)$$

for $1 \leq i < T$, with $y_T = x$.

- $(y_i)_{1 \leq i \leq T}$ represents the required function values at which to evaluate the Jacobian.

- ▷ Goal: Develop iterative algorithms to solve (1), given noisy evaluations of ∇f_i's and f_i's based on *one sample* of $(\xi_i)_{1 \leq i \leq T}$ per iteration.

- ▷ Challenge: Obtaining gradient estimators in the iterative setting with controlled bias and higher moments becomes non-trivial due to the nested structure.

- Question: Can we obtain level-independent oracle complexity results?
- Motivation:
  - Large deviation results by Ermoliev and Norkin [2013]
  - Central Limit Theorems by Dentcheva et al. [2017]

for Sample-Average Approximation (also called Empirical Risk Minimization in statistics/machine learning) provide required evidence.

Method	Yang et al. [2019]	NASA	LiNASA
Convergence Rate	$\mathcal{O}_T\left(N^{-4/(7+T)}\right)$	$\mathcal{O}_T\left(N^{-1/2}\right)$	
Oracle Complexity	$\mathcal{O}_T\left(1/\epsilon^{(7+T)/2}\right)$	$\mathcal{O}_T\left(1/\epsilon^{6}\right)$	$\mathcal{O}_T\left(1/\epsilon^{4}\right)$
Mini-batch	No	Yes	No
Feasible Set	$X = \mathbb{R}^{d_T}$	(Un)constrained	
Oracle Assumption	Finite 4^{th} moment	Finite 2^{nd} moment	

- Our algorithm is based on the Nested Average Stochastic Approximation (NASA) proposed by Ghadimi et al. [2020] for $T = 2$.

- Zhang and Xiao [2021], Ruszczyński [2021][1] and Chen et al. [2021] also obtained similar level-independent rates. However, they required:
  - a mini-batch of samples with size that scales badly with T [Zhang and Xiao, 2021] (or)
  - stronger smoothness assumptions on the stochastic functions itself [Zhang and Xiao, 2021, Chen et al., 2021] (or)
  - boundedness requirements on the feasible set [Ruszczyński, 2021].

[1] Ruszczyński [2021] also established asymptotic results in the non-smooth case.

Multi-Level NASA

- We use k (as superscript) to represent the iteration index.
- In each iteration, we update a triple $(x^k, \{w_i^k\}_{i=1}^T, z^k)$:
  - x^k – convex combinations of the solutions to gradient-descent subproblem
  - $\{w_i^k\}_{i=1}^T$ – the estimates of inner function values f_i
  - z^k – stochastic gradient of F.

- In each iteration, we perform (projected) gradient descent:

$$u^k = \underset{y \in X}{\operatorname{argmin}} \left\{ \langle z^k, y - x^k \rangle + \frac{\beta}{2} \|y - x^k\|^2 \right\}$$

where x^k is the current iterate and z^k is the stochastic gradient at the current iterate.

- For some parameter τ_k, set:

$$x^{k+1} = (1 - \tau_k)x^k + \tau_k u^k$$

- ▷ How to estimate the stochastic gradient z^k?
- ▷ Recall:

$$\nabla F(x) = \nabla f_T(y_T) \nabla f_{T-1}(y_{T-1}) \cdots \nabla f_1(y_1),$$

where ∇f_i denotes the transpose of the Jacobian of f_i, and

$$y_i = f_{i+1} \circ \cdots \circ f_T(x)$$

for $1 \leq i < T$, with $y_T = x$.

- ▷ The $(y_i)_{1 \leq i \leq T}$ represents the required function values at which to evaluate the Jacobian.

- How to estimate the stochastic gradient z^k:
  - Let w_i^k represent estimates of y_i at iteration k.
  - For each k, with w_i^k being the input, the stochastic oracle outputs:
    - Noisy function values: $G_i^{k+1} \in \mathbb{R}^{d_i}$
    - Noisy Jacobians: $J_i^{k+1} \in \mathbb{R}^{d_i \times d_{i-1}}$

- ▷ The sequences w_i^k is updated as:

$$w_i^{k+1} = (1 - \tau_k) w_i^k + \tau_k \bar{G}_i^{k+1}, \qquad 1 \le i \le T,$$

where

$$\bar{G}_i^{k+1} = \frac{1}{b_k} \sum_{j=1}^{b_k} G_{i,j}^{k+1}.$$

- ▷ The stochastic gradient z^k is updated as:

$$z^{k+1} = (1 - \tau_k) z^k + \tau_k \prod_{i=1}^{T} J_{T+1-i}^{k+1}.$$

Input: Positive integer sequences $\{b_k, \tau_k\}_{k\geq 0}$, step-size parameter β, and initial points $x^0 \in X$, $z^0 \in \mathbb{R}^{d_T}$ and $w_i^0 \in \mathbb{R}^{d_i}$ $\quad 1 \leq i \leq T$, and a probability mass function $P_R(\cdot)$ supported over $\{1, 2, \dots, N\}$, where N is the number of iterations.

0. Generate a random integer number R according to $P_R(\cdot)$.

for $k = 0, 1, 2, \dots, R$ **do**

1. Compute u^k and query the oracle to obtain the stochastic gradients J_i^{k+1}, and function values $G_{i,j}^{k+1}$ at w_{i+1}^k for $i = \{1, \dots, T\}, j = \{1, \dots, b_k\}$.

2. Update x^{k+1}, z^{k+1} and w_i^{k+1}

end for

Output: (x^R, z^R).

- ▷ All functions $f_1, \ldots, f_T$ and their derivatives are Lipschitz continuous.
- ▷ Given $\mathscr{F}_k$, the outputs of the stochastic oracle at each level i, G_i^{k+1} and J_i^{k+1}, are independent.

- For $i \in \{1, \ldots, T\}$, we have the following unbiasedness and bounded moment/variance assumptions.
- Unbiased:
  - $\mathbb{E}[J_i^{k+1} | \mathscr{F}_k] = \nabla f_i(w_{i+1}^k)$
  - $\mathbb{E}[G_i^{k+1} | \mathscr{F}_k] = f_i(w_{i+1}^k)$
- Bounded second-moment/variances:
  - $\mathbb{E}[\| G_i^{k+1} - f_i(w_{i+1}^k) \|^2 | \mathscr{F}_k] < \infty$
  - $\mathbb{E}[\| J_i^{k+1} - \nabla f_i(w_{i+1}^k) \|^2 | \mathscr{F}_k] < \infty$
  - $\mathbb{E}[\| J_i^{k+1} \|^2 | \mathscr{F}_k] < \infty$

- A point $\bar{x}$ is a stationary point of (1) if

$$-\nabla F(\bar{x}) \in N_X(\bar{x})$$

where $N_X(\bar{x})$ stands for the normal cone of X at $\bar{x}$.

- Equivalently, a point $(\bar{x}, \bar{z})$ is a stationary point of (1), if $\bar{u} = \bar{x}$ and $\bar{z} = \nabla F(\bar{x})$, where

$$\bar{u} = \operatorname*{argmin}_{y \in X} \left\{ \langle \bar{z}, y - \bar{x} \rangle + \frac{1}{2} \|y - \bar{x}\|^2 \right\}.$$

- Approximate stationary point:

$$-\nabla F(\bar{x}) \in N_X(\bar{x}) + \mathcal{B}(0, V(\bar{x}, \bar{z})),$$

  where

$$V(\bar{x}, \bar{z}) := \|\bar{u} - \bar{x}\|^2 + \|\bar{z} - \nabla F(\bar{x})\|^2$$

  is our Lyapunov function.

- A pair of points $(\bar{x}, \bar{z})$ generated by the NASA algorithm is called an expected ϵ-stationary pair, if

$$\mathbb{E}[V(\bar{x}, \bar{z})] \leq \epsilon^2,$$

- ▷ Provides unified termination criterion for both the unconstrained and constrained cases.
- ▷ When $X = \mathbb{R}^{d_T}$, $V(\bar{x}, \bar{z})$ provides an upper bound for the $\|\nabla F(\bar{x})\|^2$, because of the fact that $\bar{u} - \bar{x} = \bar{z}$ for unconstrained problems and hence we have

$$V(\bar{x}, \bar{z}) = \|\bar{z}\|^2 + \|\bar{z} - \nabla F(\bar{x})\|^2 \geq \frac{1}{2}\|\nabla F(\bar{x})\|^2.$$

- ▷ For the constrained case, $V(\bar{x}, \bar{z})$ is also related to other popular criterion like *gradient mapping* and *proximal mapping*.

Theorem [BGN22]: Assume that the parameters β, b_k and τ_k are set respectively as:

$$\beta = \mathcal{O}(\sqrt{T}), \quad b_k = \mathcal{O}(\sqrt{N}), \quad \tau_k = \frac{1}{\sqrt{N}}.$$

Then, we have

$$\mathbb{E}[V(x^R, z^R)] \leq \mathcal{O}_T\left(\frac{1}{\sqrt{N}}\right).$$

- To find an ϵ-stationary point, the NASA requires $\mathcal{O}_T(1/\epsilon^4)$ number of iterations.
- The total number of used samples is bounded by

$$\sum_{k=1}^{N} b_k = \mathcal{O}_T\left(1/\epsilon^6\right).$$

- This bound is better than $\mathcal{O}_T\left(1/\epsilon^{(7+T)/2}\right)$ obtained by Yang et al. [2019] when $T > 4$.

Multi-Level Linearized NASA (LiNASA)

- Recall the notation that w_i^k stands for estimates of $y_i = f_{i+1} \circ \cdots \circ f_T(x)$.
- Replace the update rule for w_i^{k+1} with

$$\begin{aligned} w_i^{k+1} &= w_i^k + J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k) + \tau_k(G_i^{k+1} - w_i^k) \\ &= (1-\tau_k)w_i^k + \tau_k G_i^{k+1} + J_i^{k+1}(w_{i+1}^{k+1} - w_{i+1}^k), \end{aligned}$$

- Instead of using the point estimates of f_i's, we use their stochastic linear approximate.

- Linearization technique was used as early as 1980s by Ruszczyński [1987] to handle non-smooth stochastic optimization for $T = 1$.

- More recently, Duchi and Ruan [2018] and Davis and Drusvyatskiy [2019] used other types of linearization for $T = 1$.

Theorem [BGN22]: Assume that the parameters β, b_k and τ_k are set as:

$$\beta = \mathcal{O}(\sqrt{T}), \quad b_k = 1, \quad \tau_k = \frac{1}{\sqrt{N}}.$$

Then, we have

$$\mathbb{E}[V(x^R, z^R)] \leq \mathcal{O}_T\left(\frac{1}{\sqrt{N}}\right).$$

- Result obtained without assuming boundedness of the feasible set or any dependence of the parameter β on Lipschitz constants.
- Indeed, β can be set to any positive number in the order of $\mathcal{O}(\sqrt{T})$, and τ_k depends only on the total number of iterations N.
- This makes LiNASA parameter-free and easy to implement.

- ▷ Note that LiNASA does not use a mini-batch of samples in any iteration, i.e., $b_k = 1$.
- ▷ The total sample complexity of LiNASA for finding an ϵ-stationary point, is hence bounded by

$$\mathcal{O}_T(1/\epsilon^4).$$

- ▷ The above rate is optimal (lower bounds proved for $T = 1$ by Drori and Shamir [2020]).

Projection-Free LiNASA

- Recall that in each iteration we solve:

$$u^k = \operatorname*{argmin}_{y \in X} \left\{ \langle z^k, y - x^k \rangle + \frac{\beta}{2} \|y - x^k\|^2 \right\} \qquad (2)$$

- What if the projection operation is costly ?
- Replace by Frank-Wolfe :

$$u^k = \texttt{Inexact Conditional Gradient}(x^k, z^k, \beta, t_k, \delta).$$

Input: (x, z, β, M, δ)
Set $w^0 = x$.
for $t = 0, 1, 2, \ldots, M$ **do**

1. Find $v^t \in X$ with a quantity $\delta \geq 0$ such that

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

2. Set $u^{t+1} = (1 - \mu_t) u^t + \mu_t v^t$

end for
Output: w^K

- The method only assumes access to a Linear Minization Oracle (LMO).

▷ The FW-gap is defined as

$$g_X(\bar{x}, \nabla F(\bar{x})) := \min_{y \in X} \langle \nabla f(\bar{x}), y - \bar{x} \rangle. \tag{3}$$

▷ Along the trajectory of the algorithm, we show:

$$g_X(x^k, \nabla F(x^k)) \leq V(x^k, z^k).$$

Theorem [XBG22]: Assume that the parameters β, b_k and τ_k are set as:

$$\beta = \mathcal{O}(1), \quad b_k = 1, \quad \tau_k = \frac{1}{\sqrt{N}}, \quad t_k = \sqrt{k}.$$

Then, for the LiNASA+ICG algorithm, we have

$$\mathbb{E}[V(x^R, z^R)] \leq \mathcal{O}_T\left(\frac{1}{\sqrt{N}}\right).$$

- ▷ The total sample complexity and number of calls to the LMO for finding an ϵ-stationary point are bounded respectively by

$$\mathcal{O}_T(\epsilon^{-2}) \qquad \text{and} \qquad \mathcal{O}_T(\epsilon^{-3}).$$

- ▷ The method does not use mini-batches which are common in the analysis of stochastic conditional gradient algorithms.

- Linearization not necessary for $T = 1, 2$.
- While the above results are presented in expectation, one could obtain high-probability results for $T = 1, 2$ with rates depending on the confidence level δ as $\operatorname{poly}\log(1/\delta)$ under sub-Gaussian tail assumptions.
- For the case of $T \geq 1$, we need to derive a Freedman-type martingale concentration inequality for product of random matrices.

Third (Recent) Result

- Consider the case of $T = 1$:

$$\min_{x \in X} \mathbb{E}_\xi[G(x, \xi)]$$

- In each iteration k, instead of an iid sequence, we have samples ξ_k which are drawn from a Markov Chain with state-dependent transition kernel:

$$P_{x^{k-1}}(\xi^k | \xi^{k-1}).$$

- Such a setting arises in strategic classification and reinforcement learning.

▷ Under certain drift condition on the Markov chain, the ASA framework could be extended to this setting:

	iid	Markov
Unconstrained/Projection-Based	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-5})$
Projection-free (sample comp.)	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2.5})$
Projection-free (LMO)	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-5.5})$

▷ Furthermore, under the state-independent Markov chain assumption, we get same rates as the iid setting!

- Stochastic Iterative algorithms are essentially multivariate non-iid sequences (Martingale/Markov Chains/Time-series).
- Huge-literature in the statistics on uncertainty quantification, e.g., online covariance estimation, online bootstrap.
- Some works for the case of $T = 1$ include Anastasiou et al. [2019], Yu et al. [2021], Fang et al. [2018], Zhu et al. [2021].
- Future work: develop methods for $T \geq 1$.

Thank you!

References I

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*, pages 115–137, 2019.

Raul Astudillo and Peter Frazier. Bayesian optimization of function networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

Weilin Cong, Rana Forsati, Mahmut Kandemir, and Mehrdad Mahdavi. Minimal variance sampling with provable guarantees for fast training of graph neural networks. In *KDD*, 2020.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69(4):737–760, 2017.

References II

Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pages 2658–2667. PMLR, 2020.

John Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4): 3229–3259, 2018.

Yuri Ermoliev and Vladimir Norkin. Sample average approximation method for compound stochastic optimization problems. *SIAM Journal on Optimization*, 23(4):2231–2263, 2013.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *Advances in Neural Information Processing Systems*, 34, 2021.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research*, 19(1):3053–3073, 2018.

Saeed Ghadimi, Andrzej Ruszczyński, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

References III

Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in Neural Information Processing Systems*, 34, 2021.

Zi-Hao Qiu, Quanqi Hu, Yongjian Zhong, Lijun Zhang, and Tianbao Yang. Large-scale stochastic optimization of ndcg surrogates for deep learning with provable convergence. *arXiv preprint arXiv:2202.12183*, 2022.

Andrzej Ruszczyński. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12(1):32–49, 1987.

Andrzej Ruszczyński. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM Journal on Control and Optimization*, 59(3):2301–2320, 2021.

Guanghui Wang, Ming Yang, Lijun Zhang, and Tianbao Yang. Momentum accelerates the convergence of stochastic auprc maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 3753–3771. PMLR, 2022.

References IV

Mengdi Wang, Ethan Fang, and Han Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

Shuoguang Yang, Mengdi Wang, and Ethan Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat Erdogdu. An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias. *Advances in Neural Information Processing Systems*, 2021.

Junyu Zhang and Lin Xiao. Multilevel composite stochastic optimization via nested variance reduction. *SIAM Journal on Optimization*, 31(2):1131–1157, 2021.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, pages 1–12, 2021.