First and Zeroth-order algorithms for Nonconvex Multi-level Stochastic Composition Optimization

Anthony Nguyen

Department of Mathematics
University of California, Davis

PhD advisor: Professor Krishnakumar Balasubramanian
September 24, 2019

Outline

Outline

1. Overview of Contributions
2. Motivating application in compressive sensing
3. Algorithm Analysis for Two Level Algorithm
4. Future work

Specific case: Stochastic optimization problem

- Stochastic optimization problem:

$$F^* = \min_{x \in \mathbb{R}^{d_1}} \{F(x) := f_1(x) = \mathbb{E}[G_1(x, \xi_1)]\}$$

- Function $f_1 : \mathbb{R}^{d_1} \to \mathbb{R}$ and Random variable $\xi_1 \in \mathbb{R}^{\tilde{d}_1}$.
- **Goal:** Estimate F^* given access to noisy unbiased approximation of $f_1(x)$, $\nabla f_1(x)$ [4].

Workhorse of Stochastic Optimization

- A popular method to solve this problem is the stochastic gradient descent (SGD) [8]:

$$x^{k+1} = x^k - h_k J(x^k), \quad h_k > 0$$

where $J(x^k)$ is a noisy unbiased estimate of $\nabla f_1(x^k)$.

Zeroth Order Setting

- If we can't get noisy unbiased estimates of $\nabla f_1(x)$ but just the noisy estimate of $f_1(x)$, what can be done?
- Define zeroth-order stochastic gradient:

$$(G_1)_\mu(x^k, \xi_k, v_k) = \frac{G_1(x^k + \mu v_k, \xi_k) - G_1(x^k, \xi_k)}{\mu} v_k$$

  and the smoothed function $(f_1)_\varrho(x) = \mathbb{E}_v[f_1(x + \varrho v)]$ of f_1 where $v, v_k \sim \mathcal{N}(0, I_{d_1 \times d_1})$, $\varrho \in (0, \infty)$ [5].
- $(G_1)_\mu$ is a noisy unbiased estimate of $\nabla (f_1)_\varrho$ and biased estimate of ∇f_1; we now have a gradient free zeroth order method [4].

Complexity Results for one level nested problem

After N number of iterations, we have

1-Level Nested Problem	1st Order	Zeroth Order
Nonconvex[1]	$\mathcal{O}\left(N^{-1/2}\right)$ [7]	$\mathcal{O}\left(\sqrt{\frac{d_1}{N}}\right)$ [4]
Convex[2]	$\mathcal{O}\left(N^{-1/2}\right)$ [7]	$\mathcal{O}\left(\sqrt{\frac{d_1}{N}}\right)$ [4]
Strongly Convex	$\mathcal{O}\left(N^{-1}\right)$ [7]	NA

[1] Metric for convergence rate for nonconvex problems is $\mathbb{E}[\|\nabla F(x^R)\|^2]$

[2] Metric for convergence rate for convex (μ-strongly convex) problems is $\mathbb{E}[F(x^R) - F^*]$

Two Level Nested Problem

- Want to solve

$$F^* = \min_{x \in X \subseteq \mathbb{R}^{d_2}} \{F(x) = f_1 \circ f_2(x)\} \tag{1}$$

  where X is closed and convex [1].
- Functions $f_1 : \mathbb{R}^{d_1} \to \mathbb{R}$ and $f_2 : X \subseteq \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$; $f_1(x) = \mathbb{E}_{\xi_1}[G_1(x, \xi_1)]$, $f_2(x) = \mathbb{E}_{\xi_2}[G_2(x, \xi_2)]$, random variables $\xi_1 \in \mathbb{R}^{\tilde{d}_1}, \xi_2 \in \mathbb{R}^{\tilde{d}_2}$ respectively [1]. We assume ξ_1, ξ_2 are independent.
- **Goal:** Estimate F^* given access to noisy unbiased approximation to $f_2(x), \nabla f_2(x), \nabla f_1(f_2(x))$. [3]

[3] When we talk about ∇f_2, we mean the Jacobian of f_2. This notation is for simplification purposes.

Complexity Results for two level nested problem

2-Level Nested Problem	1st Order	Zeroth Order
Nonconvex	$\mathcal{O}\left(N^{-4/9}\right)$ [7]	?
Convex	$\mathcal{O}\left(N^{-4/9}\right)$ [7]	NA
Strongly Convex	$\mathcal{O}\left(N^{-4/5}\right)$ [7]	NA

Complexity Results for two level nested problem

Our contribution for the zeroth order setting.

2-Level Nested Problem	1st Order	Zeroth Order
Nonconvex	$\mathcal{O}\left(N^{-1/2}\right)$ [1]	$\mathcal{O}\left(\sqrt{\frac{d_1^3 d_2}{N}}\right)$
Convex	$\mathcal{O}\left(N^{-4/9}\right)$ [7]	NA
Strongly Convex	$\mathcal{O}\left(N^{-4/5}\right)$ [7]	NA

Contribution 1: Zeroth Order Algorithm

- We have only noisy unbiased approximations to $f_1(x), f_2(x)$; consider stochastic gradients to get a noisy unbiased approximations to $\nabla(f_1)_\varrho, \nabla(f_2)_\varrho$ where $\varrho \in (0,\infty)$.
- Can't solve $\min_{x\in X}\{F(x) = f_1 \circ f_2(x)\}$ directly.
- Smooth functions f_1, f_2 and solve the perturbed problem $\min_{x\in X}\{h(x) := (f_1)_\varrho \circ (f_2)_\varrho(x)\}$.

T-level nested problem

- Given $F(x) = f_1 \circ \cdots \circ f_T(x)$ with $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}$ for $i = 1, \ldots, T$ with $d_0 = 1$.
- **Goal** is to solve

$$F^* = \min_{x \in X \subseteq \mathbb{R}^{d_T}} \{F(x) = f_1 \circ \cdots \circ f_T(x)\}$$

- Functions $f_i(x) = \mathbb{E}_{\xi_i}[G_i(x, \xi_i)]$, random variables $\xi_i \in \mathbb{R}^{\tilde{d}_i}$.
- $\xi_1, \ldots, \xi_T$ are assumed to be independent.

T-level nested problem

- **Goal:** Estimate F^* given access to noisy unbiased approximation to $\nabla f_1(x)$, $\nabla f_2(x), \ldots, \nabla f_T(x)$, $f_2(x), \ldots, f_T(x)$ for all x.

Complexity Results for T level nested problem

T-Level Nested Problem	1st Order	Zeroth Order
Nonconvex	$\mathcal{O}\left(N^{-4/(7+T)}\right)$ [7]	?
Convex	$\mathcal{O}\left(N^{-4/(7+T)}\right)$ [7]	NA
Strongly Convex	$\mathcal{O}\left(N^{-4/(3+T)}\right)$ [7]	NA

Contributions 2 and 3: T-level algorithms

T-Level Nested Problem	1st Order	Zeroth Order
Nonconvex	$\mathcal{O}\left(T^2 N^{-1/2}\right)$	$\mathcal{O}\left(T^2\sqrt{\frac{(d_1 \cdots d_{T-1})^3 d_T}{N}}\right)$
Convex	$\mathcal{O}\left(N^{-4/(7+T)}\right)$ [7]	NA
Strongly Convex	$\mathcal{O}\left(N^{-4/(3+T)}\right)$ [7]	NA

Outline

Linear Compressed Sensing

- We want to *reconstruct* an unknown vector $x^* \in \mathbb{R}^d$ after observing $m < d$ linear measurements y_i with added noise $\eta \in \mathbb{R}^m$:

$$y = Ax^* + \eta,$$

  where $A \in \mathbb{R}^{m \times d}$ is a measurement matrix (entries of A can be subgaussian) [2].
- This system is *underdetermined*, so recovering x^* is impossible unless we impose the structure of the unknown vector x^*.
- The most common assumption we can make is that the vector x^* is k-sparse.
- The sample complexity is determined by k (i.e., k determines m).

Generative Model from T-layer neural network

- Assumption on x^* from [3]: There exist a latent vector $z^* \in \mathbb{R}^k$ and a neural network $G : \mathbb{R}^k \to \mathbb{R}^d$ such that $x^* = G(z^*)$
- Next, the function G is a T-layer neural network using **Rectified Linear Unit (ReLU)** activations; this is a function $G : \mathbb{R}^k \to \mathbb{R}^d$, assuming $k \ll d$ [3].
- To retrieve signal x^*, define the **loss function** to be

$$\text{loss}(z) = \| AG(z) - y \|^2$$

- For $\hat{z} \in \underset{z}{\operatorname{argmin}} \| AG(z) - y \|^2$, the reconstruction of x^* is $G(\hat{z})$ [3].

Rectified Linear Unit Neural Network

- The function $G : \mathbb{R}^k \to \mathbb{R}^d$ is a fixed **ReLU** neural network of the form [3]:

$$G(x) = \sigma \circ (W_1 \sigma \circ (W_2 \cdots \sigma \circ (W_T x))),$$

- We consider when the number of layers T is smaller than d, $W_i \in \mathbb{R}^{d_{i-1} \times d_i}$ for $i = 1, \ldots, T$ with $d_0 = d$, $d_T = k$, and $d_i \le d$.
- $x^* := G(z^*)$ is bounded with respect to the euclidean norm, and $\{(a_i, y_i)\}_{i=1}^m$ be i.i.d.s where a_i's are the ith row of A [3].

Details on the structure of the compressed sensing problem

- **Example:** Random matrix W_i with entries i.i.d.s drawn from $\mathcal{N}(0, 1/d_i)$ for $i = 1, \ldots, T$ and $\sigma(x) = \max(x, 0)$ [3].
- G is not smooth by the presence of σ, but we can approximate σ by $\sigma(x) \approx \frac{\sqrt{x^2+\epsilon^2}+x}{2}$ for $\epsilon \ll 1$.
- Result presented later assumes our objective function F is smooth, but we can smooth G.

Outline

Theory on Existence of a solution

Theorem (Minimum and Normal Cones [10])

If a point $\hat{x} \in X$ is a local minimum of problem (1)*, then*

$$-\nabla F(\hat{x}) \in \mathcal{N}_X(\hat{x}) := \{g \mid g^T(w - \hat{x}) \leq 0 \text{ for all } w \in X\} \quad (2)$$

where $\mathcal{N}_X(\hat{x})$ denotes the normal cone to X at the point $\hat{x}$. Furthermore, if $F = f_1 \circ f_2$ is convex, then every point $\hat{x}$ satisfying (2) *is the global minimum of the problem* (1).

Introducing Nested Averaged Stochastic Approximation (NASA)

- The algorithm produces three random sequences: **approximate solutions** $\{x^k\}$, **average gradients** $\{z^k\}$, and **average** f_2**-values** $\{u^k\}$
- These random sequences are defined on a certain probability space $(\Omega, \mathscr{F}, P)$.
- We let $\mathscr{F}_k = \sigma(\{x^0, \ldots, x^k, z^0, \ldots, z^k, u^0, \ldots, u^k\})$ be the sigma algebra generated by these sequences.

First Order Stochastic Oracle

- For each $k \geq 0$, the **stochastic oracle** returns random vectors $G^{k+1} \in \mathbb{R}^{d_1}, s^{k+1} \in \mathbb{R}^{d_1}$, and a random matrix $J^{k+1} \in \mathbb{R}^{d_1 \times d_2}$, such that s^{k+1}, J^{k+1} are *conditionally independent* given $\mathscr{F}_k$, and
-

$$\begin{aligned} \mathbb{E}[G^{k+1}|\mathscr{F}_k] &= f_2(x^{k+1}), \quad \mathbb{E}[\|G^{k+1} - f_2(x^{k+1})\|^2|\mathscr{F}_k] \leq \sigma_G^2, \\ \mathbb{E}[J^{k+1}|\mathscr{F}_k] &= \nabla f_2(x^{k+1}), \quad \mathbb{E}[\|J^{k+1}\|^2|\mathscr{F}_k] \leq \sigma_J^2, \\ \mathbb{E}[s^{k+1}|\mathscr{F}_k] &= \nabla f_1(u^k), \quad \mathbb{E}[\|s^{k+1}\|^2|\mathscr{F}_k] \leq \sigma_s^2, \end{aligned}$$

NASA Algorithm

- **Input:** $x^0 \in X \subseteq \mathbb{R}^{d_2}$, $z^0 \in \mathbb{R}^{d_2}, u^0 \in \mathbb{R}^{d_1}, a > 0, b > 0.$
- **Step 0:** Set $k = 0$
- **Step 1:** Pick $\beta_k > 0$ and stepsize $\tau_k \in (0, \frac{1}{a}]$, compute

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

  and set

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

- **Step 2:** Call the stochastic oracle to obtain s^{k+1} at u^k, G^{k+1} and J^{k+1} at x^{k+1}, and update the running averages as

$$z^{k+1} = (1 - a\tau_k) z^k + a\tau_k s^{k+1} J^{k+1},$$
$$u^{k+1} = (1 - b\tau_k) u^k + b\tau_k G^{k+1}.$$

- **Step 3:** Increment k and go to Step 1.

Introducing a special Lyaponuv function

- We define a **Lypaponuv function**

$$V(x,z) = \|\Pi_X(x-z) - x\|^2 + \|z - \nabla F(x)\|^2 \tag{3}$$

  the operation of the orthogonal projection on the set X. This *measures the violation of the optimality condition* $-\nabla F(\hat{x}) \in \mathcal{N}_X(\hat{x})$.
- For simplicity, we look at the **unconstrained problem**; this reduces (3) to

$$V(x,z) = \|z\|^2 + \|z - \nabla F(x)\|^2 \tag{4}$$

Approximate Stationary Point

Theorem (Sufficient and Necessary Condition)

$$-z \in \mathcal{N}_X(x) \Leftrightarrow \Pi_X\left(x - \frac{z}{\beta}\right) = x \quad \textit{where } \beta > 0 \tag{5}$$

- This theorem is crucial in establishing an approximate stationary point since $V(x,z) < \epsilon$ implies that $z \approx \nabla F(x)$, so $V(x,z) \approx \|\nabla F(x)\|^2 < \epsilon$.

Convergence Analysis of the first order NASA

- Assume $f_1, f_2, \nabla f_1, \nabla f_2$ are Lipschitz continuous with Lipschitz constants denoted by $L_{f_1}, L_{f_2}, L_{\nabla f_1}, L_{\nabla f_2}$. Then F and ∇F are Lipschitz continuous.
- Define $\eta(x,z) = \min_{y\in X}\left\{\langle z, y-x\rangle + \frac{\beta}{2}\|y-x\|^2\right\}$. Then

$$L_{\nabla\eta} = 2\sqrt{(1+\beta)^2 + \left(1+\frac{1}{2\beta}\right)^2}$$

- Define the **merit function**:

$$W(x,z,u) = a(F(x)-F^*) - \eta(x,z) + \frac{\gamma}{2}\|f_2(x)-u\|^2, \tag{6}$$

where $\gamma > 0$ and $F^* := \inf_{x\in X} F(x)$.

Useful Bounds for NASA Convergence

Proposition (Ghadimi, Ruszczynski, and Wang [1])

Suppose $\tau_0 = \frac{1}{a}$ and assumption on the stochastic oracle holds, then

$$\beta_k^2 \mathbb{E}[\|y^k - x^k\|^2 | \mathscr{F}_{k-1}] \leq \mathbb{E}[\|z^k\|^2 | \mathscr{F}_{k-1}] \leq \sigma_J^2 \sigma_s^2 \quad \forall k \geq 1; \tag{7}$$

We have

$$\mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathscr{F}_k] \leq 4\sigma_J^2 \sigma_s^2 \tau_k^2 \tag{8}$$

We define

$$\sigma^2 := \frac{1}{2}\left([L_{\nabla F} + L_{\nabla \eta} + \gamma L_{f_2}^2 + 2aL_{f_2}^2 L_{\nabla f_1}] \frac{\sigma_J^2 \sigma_s^2}{\beta^2} + b^2 \gamma \sigma_{f_2}^2 + 4L_{\nabla \eta} \sigma_J^2 \sigma_s^2 \right) \in \mathscr{O}(1)$$

which we will utilize in the upcoming theorem.

Bound on the Lyaponuv Function

- Now we're ready to upper bound the Lyaponuv function as follows:

$$V(x^k, z^k) \leq \max(1, \beta_k^2)\|y^k - x^k\|^2 + \|z^k - \nabla F(x^k)\|^2. \tag{9}$$

- Equation (7) gives us an upper bound on the first term in (9). The following theorem will allow us to bound the second term. The idea is to define a random integer variable $R \in \{0, 1, \dots, N-1\}$ with probability mass function

$$P[R = k] = \frac{\tau_k}{\sum_{j=0}^{N-1} \tau_j} \tag{10}$$

  to bound $\mathbb{E}[V(x^R, z^R)]$.

- We define a notation for the next theorem:

$$\Gamma_1 := \begin{cases} 1, & \tau_0 = 1/a \\ 1 - a\tau_0, & \tau_0 < 1/a, \end{cases} \quad \Gamma_k := \Gamma_1 \prod_{i=1}^{k-1}(1 - a\tau_i) \quad \forall k \geq 2,$$

Error Complexity of the Approximate Stationary Point

Theorem (Ghadimi, Ruszczynski, and Wang [1])

Let $\beta_k := \beta > 0$ for all $k \geq 0$ and assume a, b, c, γ are chosen such that $2(a\beta - c)(\gamma b - 2c) \geq L_{f_2}^2 (aL_{\nabla f_1} + \gamma)^2$ and

$$\sum_{i=k+1}^{N} \tau_i \Gamma_i \leq \bar{c}\Gamma_{k+1} \quad \forall k \geq 0 \text{ and } \forall N \geq 1,$$

where $\bar{c}$ is a positive constant. Then for $N \geq 1$, we have

$$\sum_{k=1}^{N} \tau_k \mathbb{E}[\|\nabla F(x^k) - z^k\|^2 | \mathscr{F}_{k-1}] \leq a\bar{c}\left(\frac{1}{c}\max(L_1, L_2)\sigma^2 + 4a\sigma_J^2\sigma_s^2\right)\left(\sum_{k=0}^{N-1} \tau_k^2\right) + \frac{a\bar{c}}{c}\max(L_1, L_2) W(x^0, z^0, u^0),$$

$$L_1 := \frac{2L_{\nabla F}^2}{a^2} + 4L_{f_2}^4 L_{\nabla f_1}^2, \quad L_2 := 4L_{f_2}^2 L_{\nabla f_1}^2.$$

Error Complexity Cont.

Theorem (Ghadimi, Ruszczynski, and Wang [1])

Pick $a = b = 1, \bar{c} = 1,\ \beta_k := \beta = \left(\frac{(1+\alpha)^2}{\alpha} L_{f_2}^2 + \frac{\alpha}{4}\right) L_{\nabla f_1} \quad \forall k \geq 0,$ *for some* $\alpha > 0,$ *and* $\tau_0 = 1, \tau_k \equiv \frac{1}{\sqrt{N}} \quad \forall k = 1, \ldots, N-1,$ *then*

$$\mathbb{E}[V(x^R, z^R)] \leq \frac{4}{\sqrt{N}}\left(\frac{2}{\alpha L_{\nabla F}}[\max(L_1, L_2) + \max(1, \beta^2)][W(x^0, z^0, u^0) + \sigma^2] + \sigma_J^2 \sigma_s^2 + \|\nabla F(x^0) - z^0\|^2\right) \in \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

and

$$\mathbb{E}[\| f_2(x^R) - u^R \|^2] \leq \frac{W(x^0, z^0, u^0) + 2\sigma^2}{\alpha L_{\nabla F} \sqrt{N}} \in \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

Error Complexity Cont.

Sketch Proof.

Define

$$\begin{aligned}
\Delta_i^F &:= \nabla f_1(u^i)\nabla f_2(x^{i+1}) - s^{i+1}J^{i+1} \\
e_i &:= \frac{(1-a\tau_i)}{a\tau_i}[\nabla F(x^{i+1}) - \nabla F(x^i)] + \nabla F(x^{i+1}) - \nabla f_1(u^i)\nabla f_2(x^{i+1}) \\
\delta_i &:= \langle (1-a\tau_i)[\nabla F(x^i) - z^i] + a\tau_i e_i, \Delta_i^F \rangle.
\end{aligned}$$

We have

$$\|\nabla F(x^k) - z^k\|^2 \le \Gamma_k \left[\sum_{i=0}^{k-1} \left(\frac{a\tau_i}{\Gamma_{i+1}}\|e_i\|^2 + \frac{a^2\tau_i^2}{\Gamma_{i+1}}\|\Delta_i^F\|^2 + \frac{2a\tau_i\delta_i}{\Gamma_{i+1}} \right) \right]$$

Compute $\sum_{k=1}^N \tau_k \mathbb{E}[\|\nabla F(x^k) - z^k\|^2 | \mathscr{F}_{k-1}]$ and upper bound terms. □

Background to Zeroth Order Algorithm

- Introduce two smoothed versions of f_1, f_2: $(f_1)_\varrho(x) = \mathbb{E}_{v_1}[f_1(x+\varrho v_1)]$ and $(f_2)_\varrho(x) = \mathbb{E}_{v_2}[f_2(x+\varrho v_2)]$, respectively; where $v_1 \sim \mathcal{N}(0, I_{d_1 \times d_1})$, $v_2 \sim \mathcal{N}(0, I_{d_2 \times d_2})$, and $\varrho \in (0, \infty)$.
- Introduce **stochastic gradients** of $(f_1)_\varrho, (f_2)_\varrho$ by

$$E_\varrho(x, \xi, v_1) = \left[\frac{G_1(x+\varrho v_1, \xi) - G_1(x, \xi)}{\varrho}\right] v_1$$

$$[D_\varrho(x, \xi, v_2)]_j = \left[\frac{(G_2)_j(x+\varrho v_2, \xi) - (G_2)_j(x, \xi)}{\varrho}\right] v_2 \quad \text{for } j = 1, \ldots, d_1,$$

respectively [5].

Results about Stochastic Gradients

- By [9] and Jensen's inequality, it can be shown that

$$\mathbb{E}_{\xi, v_1}[E_\varrho(x, \xi, v_1)] = \nabla (f_1)_\varrho(x) \in \mathbb{R}^{d_1}$$
$$\mathbb{E}_{\xi, v_2}[D_\varrho(x, \xi, v_2)] = \nabla (f_2)_\varrho(x) \in \mathbb{R}^{d_1 \times d_2}$$

and

$$\|\nabla (f_1)_\varrho(x)\|^2 = \|\mathbb{E}_{\xi, v_1}[E_\varrho(x, \xi, v_1)]\|^2 \leq \frac{\varrho^2}{2} L^2_{\nabla f_1}(d_1+6)^3 + 2(d_1+4)L^2_{f_1} =: \sigma_s^2$$
$$\|\nabla (f_2)_\varrho(x)\|^2 = \|\mathbb{E}_{\xi, v_2}[D_\varrho(x, \xi, v_2)]\|^2 \leq d_1 \left[\frac{\varrho^2}{2} L^2_{\nabla f_2}(d_2+6)^3 + 2(d_2+4)L^2_{f_2}\right] =: \sigma_J^2$$

Zeroth Order Stochastic Oracle

- For each $k \geq 0$, the **stochastic oracle** returns $G^{k+1}, G_\varrho^{k+1} \in \mathbb{R}^{d_1}$, $J_\varrho^{k+1} \in \mathbb{R}^{d_1 \times d_2}$, $s_\varrho^{k+1} \in \mathbb{R}^{d_1}$ such that
- $$\begin{aligned} &\mathbb{E}[G^{k+1} | \mathscr{F}_k] = f_2(x^{k+1}), \quad \mathbb{E}[G_\varrho^{k+1} | \mathscr{F}_k] = (f_2)_\varrho(x^{k+1}) \\ &\mathbb{E}[\| G^{k+1} - f_2(x^{k+1}) \|^2 | \mathscr{F}_k] \leq \sigma_G^2 \\ &\mathbb{E}[J_\varrho^{k+1} | \mathscr{F}_k] = \nabla (f_2)_\varrho(x^{k+1}), \quad \mathbb{E}[\| J_\varrho^{k+1} \|^2 | \mathscr{F}_k] \leq \sigma_J^2, \\ &\mathbb{E}[s_\varrho^{k+1} | \mathscr{F}_k] = \nabla (f_1)_\varrho(u^k), \quad \mathbb{E}[\| s_\varrho^{k+1} \|^2 | \mathscr{F}_k] \leq \sigma_s^2. \end{aligned}$$

Zeroth Order NASA Algorithm

- **Input:** $x^0 \in X \subseteq \mathbb{R}^{d_2}$, $z^0 \in \mathbb{B}(0, \sigma_J \sigma_s) \subseteq \mathbb{R}^{d_2}, u^0 \in \mathbb{R}^{d_1}, a > 0, b > 0, \varrho > 0.$
- **Step 0.** Set $k = 0$
- **Step 1.** Pick $\beta_k > 0$ and stepsize $\tau_k \in (0, \frac{1}{a}]$, compute

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

  and set

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

- **Step 2.** Call the stochatic oracle to obtain s_ϱ^{k+1} at u^k, G_ϱ^{k+1} and J_ϱ^{k+1} at x^{k+1}, and update the running averages as

  $$z^{k+1} = (1 - a\tau_k) z^k + a\tau_k s_\varrho^{k+1} J_\varrho^{k+1},$$
  $$u^{k+1} = (1 - b\tau_k) u^k + b\tau_k G_\varrho^{k+1}.$$

- **Step 3.** Increment k and go to Step 1.

Sample Complexity of the Zeroth Order Algorithm

- Define the modified Lyaponuv function

$$V_1(x,z) = \|z\|^2 + \|z - \nabla((f_1)_\varrho \circ (f_2)_\varrho)(x)\|^2.$$

- We pick $a = b = 1$, $0 < \varrho \leq \frac{1}{\max\{d_1,d_2\}^2}$, $\beta_k := \beta = \left(\frac{(1+\alpha)^2}{\alpha} L^2_{(f_2)_\varrho} + \frac{\alpha}{4}\right) L_{\nabla(f_1)_\varrho}$ for some $\alpha > 0$, and $\tau_k \equiv \frac{1}{\sqrt{d_1 d_2 N}} \quad \forall k = 0, 1, \ldots, N-1$. Then

$$\mathbb{E}[V_1(x^R, z^R)] \leq \mathcal{O}\left(\sqrt{\frac{d_1^3 d_2}{N}}\right).$$

Sample Complexity of the Zeroth Order Algorithm

- How close is our solution x^k for the perturbed problem to the original problem?

Theorem (Our contribution)

The sequence generated zeroth order algorithm achieves the following bound:

$$\mathbb{E}[V(x^R, z^R)] \leq 4\sqrt{\frac{d_1^3 d_2}{N}} + 2\varrho^2[d_2 L_{\nabla F}^2 + 2d_1 L_{f_2}^2 L_{\nabla f_1}^2]$$

Sample Complexity of the Zeroth Order Algorithm

Sketch Proof.

Define $V_2(x,z) := \|z\|^2 + \|z - \nabla f_1 \circ (f_2)_\varrho(x)\|^2$. Then

$$V(x,z) \leq 2(V_2(x,z) + \varrho^2 d_2 L_{\nabla F})$$

Furthermore, one can show

$$V_2(x,z) \leq 2(V_1(x,z) + L_{f_2}^2 L_{\nabla f_1}^2 \varrho^2 d_1)$$

Using $\mathbb{E}[V_1(x^R, z^R)] \leq \mathcal{O}\left(\sqrt{\frac{d_1^3 d_2}{N}}\right)$ and these two relations, we have our conclusion. $\square$

T-Level First Order Algorithm Theorem

Theorem (Our Contribution)

Suppose the stochastic oracle assumption and Lipschitz continuity are satisfied and let $\{x^k, z^k, y^k, w_1^k, \ldots, w_{T-1}^k\}_{k\geq 0}$ *be the sequence generated by the* T*-stage NASA algorithm. Let*

$$L_j = C_j T^2 \quad \textit{for } j = 0, \ldots, T-1 \quad \textit{and } C_j > 0.$$

T-Level First Order Algorithm Theorem (Cont.)

Theorem (Cont.)

We pick

$$\begin{aligned}
M_1 &= \max\{L_{f_{j+1}\circ\cdots\circ f_T} \,|\, j = 1,\ldots,T-1\},\\
M_2 &= \max\{1, L_{\nabla f_1}, L_{f_1\circ\cdots\circ f_{j-1}} \,|\, j = 2,\ldots,T-1\},\\
a &= b_j = \bar{c} = 1, \qquad \textit{for } j = 1,\ldots,T-1,\\
\gamma_j &= 4c = 2\alpha M_2, \qquad \textit{for } j = 1,\ldots,T-1,\\
\beta_k &= \left(\frac{(T-1)}{2}\frac{(1+2\alpha)^2}{\alpha}M_1^2 + \frac{\alpha}{2}\right)M_2 \qquad \textit{where for some } \alpha > 0
\end{aligned}$$

, and $\tau_0 = 1, \tau_k = \frac{1}{\sqrt{N}}$, *for* $k = 1,\ldots,N-1$. *Then*

$$\begin{aligned}
\mathbb{E}[V(x^R, z^R)] \leq \frac{4}{\sqrt{N}}\Big(&\frac{2}{\alpha M_2}[\max(L_0,\ldots,L_{T-1}) + \max(1,\beta^2)][\sigma^2\\
&+ W(x^0, z^0, \overline{w^0})] + 2(\sigma_{J_1}\cdots\sigma_{J_T})^2 + \|\nabla F(x^0) - z^0\|^2\Big)
\end{aligned}$$

T-Level Zeroth Order Algorithm

Theorem (Our Contribution)

We pick $L_0, \ldots, L_{T-1}, M_1, M_2, a, b_j, \bar{c}, \gamma_j, \beta_k$ *be as before and pick* $\tau_k = \frac{1}{\sqrt{N d_1 \cdots d_T}}$ *for* $k = 0, \ldots, N-1$ *and* $0 < \varrho \leq \frac{1}{\max\{d_1, \ldots, d_T\}^2}$*. Then*

$$\mathbb{E}[\tilde{V}(x^R, z^R)] \leq \mathcal{O}\left(2^T \sqrt{\frac{(d_1 \cdots d_{T-1})^3 d_T}{N}}\right)$$

where $\tilde{V}(x,z) = \|z\|^2 + \|z - \nabla((f_1)_\varrho \cdots (f_T)_\varrho)(x)\|^2$*.*

Discussion on T-Level Algorithms

- For the zeroth order algorithm, we need to introduce the stochastic gradients of $(f_i)_\varrho$ for $1 \leq i \leq T$ and make appropriate assumptions on the zeroth order stochastic oracle.
- The analysis for the T-level algorithm is more technical and involved than the two level nested problem.

Outline

Future direction

- Run numerical simulations of our zeroth and first order algorithm to the nested problem.
- Based on the current results on the error complexity for these nested problems, not much is known about the convex and strongly convex cases.
- The two-level nested stochastic optimization problem could be solved using the sample average approximation method where it is assumed that ξ_1, ξ_2 are independent [11]. What if ξ_1, ξ_2 were dependent? What changes structurally in the NASA implementation?

References

1. Ghadimi, Saeed, Andrzej Ruszczynski, and Mengdi Wang. "A Single Time-Scale Stochastic Approximation Method for Nested Stochastic Optimization." arXiv preprint arXiv:1812.01094 (2018).
2. Bora, Ashish, et al. "Compressed sensing using generative models." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
3. Qiu, Shuang, Xiaohan Wei, and Zhuoran Yang. "Robust One-Bit Recovery via ReLU Generative Networks: Improved Statistical Rates and Global Landscape Analysis." arXiv preprint arXiv:1908.05368 (2019).
4. Ghadimi, Saeed, and Guanghui Lan. "Stochastic first-and zeroth-order methods for nonconvex stochastic programming." SIAM Journal on Optimization 23.4 (2013): 2341-2368.

References

5. Balasubramanian, Krishnakumar, and Saeed Ghadimi. "Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality, and saddle-points." arXiv preprint arXiv:1809.06474 (2019): 651-676.
6. Wang, Mengdi, Ethan X. Fang, and Han Liu. "Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions." arXiv preprint arXiv:1411.3803 (2014).
7. Yang, Shuoguang, Mengdi Wang, and Ethan X. Fang. "Multilevel stochastic gradient methods for nested composition optimization." SIAM Journal on Optimization 29.1 (2019): 616-659.
8. Rakhlin, Alexander, Ohad Shamir, and Karthik Sridharan. "Making gradient descent optimal for strongly convex stochastic optimization." arXiv preprint arXiv:1109.5647 (2011).

References

9. Nesterov, Yurii, and Vladimir Spokoiny. "Random gradient-free minimization of convex functions." Foundations of Computational Mathematics 17.2 (2017): 527-566.
10. Ruszczynski, Andrzej P., and Andrzej Ruszczynski. Nonlinear optimization. Vol. 13. Princeton university press, 2006.
11. Hu, Yifan, Xin Chen, and Niao He. "Sample Complexity of Sample Average Approximation for Conditional Stochastic Optimization." arXiv preprint arXiv:1905.11957 (2019).

Thanks for your attention!
Any questions?

References

- Ghadimi, Saeed, Andrzej Ruszczynski, and Mengdi Wang. "A Single Time-Scale Stochastic Approximation Method for Nested Stochastic Optimization." arXiv preprint arXiv:1812.01094 (2018).
- Bora, Ashish, et al. "Compressed sensing using generative models." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
- Qiu, Shuang, Xiaohan Wei, and Zhuoran Yang. "Robust One-Bit Recovery via ReLU Generative Networks: Improved Statistical Rates and Global Landscape Analysis." arXiv preprint arXiv:1908.05368 (2019).
- Ghadimi, Saeed, and Guanghui Lan. "Stochastic first-and zeroth-order methods for nonconvex stochastic programming." SIAM Journal on Optimization 23.4 (2013): 2341-2368.
- Balasubramanian, Krishnakumar, and Saeed Ghadimi. "Zeroth-order