SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path Integrated Differential Estimator

Cong Fang

Chris Junchi Li

Zhouchen Lin

Tong Zhang

Problem

We consider the following non-convex problem:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \qquad f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \tag{**}$$

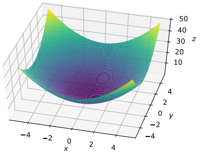
Study both finite-sum case (n is finite) and online case (n is ∞).

- ϵ-*approximate first-order stationary point*, or simply an FSP, if

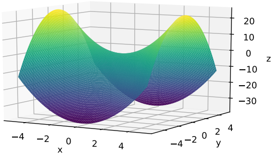
$$\|\nabla f(x)\| \le \epsilon \tag{0.1}$$

- (ϵ, δ)-*approximate second-order stationary point*, or simply an SSP, if

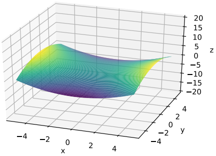
$$\|\nabla f(x)\| \le \epsilon, \qquad \lambda_{\min}\left(\nabla^2 f(x)\right) \ge -\mathcal{O}(\sqrt{\varepsilon}) \tag{0.2}$$



Local Minimizer



Conspicuous Saddle



SSP

Comparison of Existing Methods

	Algorithm		Online	Finite-Sum
First-order Stationary Point	GD / SGD	(Nesterov,2004)	ε^{-4}	$n\varepsilon^{-2}$
	SVRG / SCSG	(Allen-Zhu, Hazan, 2016) (Reddi et al., 2016) (Lei et al., 2017)	$\varepsilon^{-3.333}$	$n + n^{2/3}\varepsilon^{-2}$
	SNVRG	(Zhou et al., 2018)	ε^{-3}	$n + n^{1/2}\varepsilon^{-2}$
	SPIDER-SFO	(this work)	ε^{-3}	$n + n^{1/2}\varepsilon^{-2}$ Δ
Second-order Stationary Point (Hessian-Lipschitz Required)	Perturbed GD / SGD	(Ge et al.,2015) (Jin et al.,2017b)	$poly(d)\varepsilon^{-4}$	$n\varepsilon^{-2}$
	NEON+GD / NEON+SGD	(Xu et al.,2017) (Allen-zhu, Li,2017b)	ε^{-4}	$n\varepsilon^{-2}$
	AGD	(Jin et al.,2017b)	N/A	$n\varepsilon^{-1.75}$
	NEON+SVRG / NEON+SCSG	(Allen-Zhu, Hazan, 2016) (Reddi et al.,2016) (Lei et al.,2017)	$\varepsilon^{-3.5}$ ($\varepsilon^{-3.333}$)	$n\varepsilon^{-1.5} + n^{2/3}\varepsilon^{-2}$
	NEON+FastCubic/CDHS	(Agarwal et al.,2017) (Carmon et al.,2016) (Tripuraneni et al.,2017)	$\varepsilon^{-3.5}$	$n\varepsilon^{-1.5} + n^{3/4}\varepsilon^{-1.75}$
	NEON+Natasha2	(Allen-Zhu, 2017) (Xu et al., 2017) (Allen-Zhu, Li, 2015)	$\varepsilon^{-3.5}$ ($\varepsilon^{-3.25}$)	$n\varepsilon^{-1.5} + n^{2/3}\varepsilon^{-2}$
	SPIDER-SFO^{+}	(this work)	ε^{-3}	$n^{1/2}\epsilon^{-2}$ $(n \geq \epsilon^{-1})$

Example: Algorithm for Searching FSP in Expectation

Algorithm 1 SPIDER-SFO in Expectation: Input $\mathbf{x}^0$, q, S_1, S_2, n_0, ϵ (For a finding FSP)

1: **for** $k = 0$ to K **do**
2: **if** $\mod(k, q) = 0$ **then**
3: Draw S_1 samples (or compute the full gradient for the finite-sum case), $\mathbf{v}^k = \nabla f_{\mathcal{S}_1}(\mathbf{x}^k)$
4: **else**
5: Draw S_2 samples, and let $\mathbf{v}^k = \nabla f_{\mathcal{S}_2}(\mathbf{x}^k) - \nabla f_{\mathcal{S}_2}(\mathbf{x}^{k-1}) + \mathbf{v}^{k-1}$
6: **end if**
7: $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta^k \mathbf{v}^k$ where $\eta^k = \min\left(\frac{\epsilon}{L n_0 \|\mathbf{v}^k\|}, \frac{1}{2 L n_0}\right)$
8: **end for**
9: Return $\tilde{\mathbf{x}}$ chosen uniformly at random from $\{\mathbf{x}^k\}_{k=0}^{K-1}$

- We prove the stochastic gradient costs to find an approximate FSP is both upper and lower bounded by $\mathcal{O}\left(n^{1/2}\epsilon^{-2}\right)$ under certain conditions
- A similar complexity has also been obtain by Zhou et al., (2018)

Stochastic Path-Integrated Differential Estimator: Core Idea

Observe a sequence $\widehat{x}_{0:K} = \{\widehat{x}_0, \ldots, \widehat{x}_K\}$, the goal is to dynamically track for a quantity $Q(x)$. For $Q(\widehat{x}^k)$ for $k = 0, 1, \ldots, K$

- Initial estimate $\widetilde{Q}(\widehat{x}^0) \approx Q(\widehat{x}^0)$
- Unbiased estimate $\xi_k(\widehat{x}_{0:k})$ of $Q(\widehat{x}^k) - Q(\widehat{x}^{k-1})$ such that for each $k = 1, \ldots, K$
$$\mathbb{E}\left[\xi_k(\widehat{x}_{0:k}) \mid \widehat{x}_{0:k}\right] = Q(\widehat{x}^k) - Q(\widehat{x}^{k-1})$$
- Integrate the stochastic differential estimate as
$$\widetilde{Q}(\widehat{x}_{0:K}) := \widetilde{Q}(\widehat{x}^0) + \sum_{k=1}^{K} \xi_k(\widehat{x}_{0:k}) \tag{0.3}$$
- Call estimator $\widetilde{Q}(\widehat{x}_{0:K})$ the Stochastic Path-Integrated Differential EstimatoR, or SPIDER for brevity
- Example: $Q(x)$ is picked as $\nabla f(x)$ (or $f(x)$)

A similar idea, named SARAH, has been proposed by Nguyen et al. (2017)

Summary and Extension

Summary:

(i) Proposed SPIDER technique for tracking:
- Avoidance of excessive access of oracles and reduction of time complexity
- Potential application in many stochastic estimation problems

(ii) Proposed SPIDER-SFO algorithms for first-order non-convex optimization
- Achieves $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ rate for finding ε-FSP in expectation
- Proved that SPIDER-SFO matches the lower bound in the finite-sum case (Carmon et al. 2017)

Extension in the long version: https://arxiv.org/pdf/1807.01695.pdf

(i) Obtain high-probability results for SPIDER-SFO

(ii) Proposed SPIDER-SFO$^+$ algorithms for first-order non-convex optimization
- Achieves $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ rate for finding $(\varepsilon, \mathcal{O}(\sqrt{\varepsilon}))$-SSP

(iii) Proposed SPIDER-SZO algorithm for zeroth-order non-convex optimization
- Achieves an improved rate of $\mathcal{O}(d\varepsilon^{-3})$

Thank you!

Welcome to Poster #49 in Room 210 & 230 AB today!