A Stochastic Composite Gradient Method with Incremental Variance Reduction

Junyu Zhang (University of Minnesota)
and
Lin Xiao (Microsoft Research)

Neural Information Processing Systems (NeurIPS)

Vancouver, Canada
December 8-14 , 2019

Composite stochastic optimization

- composition with exepctation

$$\underset{x\in\mathbf{R}^d}{\text{minimize}} \quad f\left(\mathbf{E}_\xi[g_\xi(x)]\right) + r(x)$$

  - $f : \mathbf{R}^p \to \mathbf{R}$ smooth and possibly nonconvex
  - $g_\xi : \mathbf{R}^d \to \mathbf{R}^p$ smooth vector mapping for every ξ
  - $r : \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ convex but possibly nonsmooth
- composition with finite sum

$$\underset{x\in\mathbf{R}^d}{\text{minimize}} \quad f\left(\frac{1}{n}\sum_{i=1}^n g_i(x)\right) + r(x)$$

- applications beyond ERM
  - policy evaluation in reinforcement learning
  - risk-averse optimization, financial mathematics
  - ...

Examples

- policy evaluation with linear function approximation

$$\text{minimize}_{x\in\mathbf{R}^d} \left\|\mathbf{E}[A]x - \mathbf{E}[b]\right\|^2$$

  A, b random, generated by MDP under fixed policy
- risk-averse optimization

$$\underset{x\in\mathbf{R}^d}{\text{maximize}} \quad \underbrace{\frac{1}{n}\sum_{j=1}^n h_j(x)}_{\text{average reward}} - \lambda \underbrace{\frac{1}{n}\sum_{j=1}^n\left(h_j(x) - \frac{1}{n}\sum_{i=1}^n h_i(x)\right)^2}_{\text{variance of rewards (risk)}}$$

  - often treated as two-level composite finite-sum optimization
  - simple transformation using $\mathbf{Var}(a) = \mathbf{E}[a^2] - (\mathbf{E}[a])^2$

$$\underset{x\in\mathbf{R}^d}{\text{maximize}} \quad \frac{1}{n}\sum_{j=1}^n h_j(x) - \lambda\left(\frac{1}{n}\sum_{j=1}^n h_j^2(x) - \left(\frac{1}{n}\sum_{i=1}^n h_i(x)\right)^2\right)$$

  actually a one-level composite finite-sum problem

Technical challenge and related work

- challenge: biased gradient estimator
  - denote $F(x) := f(g(x))$ where $g(x) := \mathbf{E}_\xi[g_\xi(x)]$

$$F'(x) = [g'(x)]^T f'(g(x))$$

  - subsampled estimators

$$y = \frac{1}{|\mathcal{S}|}\sum_{\xi\in\mathcal{S}} g_\xi(x), \quad z = \frac{1}{|\mathcal{S}|}\sum_{\xi\in\mathcal{S}} g'_\xi(x)$$

$$\mathbf{E}[y] = g(x) \text{ and } \mathbf{E}[z] = g'(x), \text{ but } \mathbf{E}\left[[z]^T f'(y)\right] \neq F'(x)$$

- related work
  - more general composite stochastic optimization
    (Wang, Fang & Liu 2017; Wang, Liu & Fang 2017; ...)
  - two-level composite finite-sum: extending SVRG and SAGA
    (Lian, Wang & Liu 2017; Huo, Gu, Liu & Huang 2018; Lin, Fan, Wang & Jordan 2018; Zhang & Xiao 2019, ...)

Main results

$$\underset{x}{\text{minimize}} \quad \Psi(x) + r(x) \triangleq f\left(\mathbf{E}_\xi[g_\xi(x)]\right) + r(x)$$

idea: use SARAH/SPIDER estimator for both $g(x)$ and $g'(x)$

sample complexities

	assumptions (common: f and g_ξ Lipschitz and smooth, thus F smooth)		
	F nonconvex, r convex	F ν-gradient dominant, $r \equiv 0$	F convex, r convex, Φ μ-optimally strongly convex
E	$\mathcal{O}(\epsilon^{-3/2})$	$\mathcal{O}\left((\nu\epsilon^{-1})\log\epsilon^{-1}\right)$	$\mathcal{O}\left((\mu^{-1}\epsilon^{-1})\log\epsilon^{-1}\right)$
$\sum$	$\mathcal{O}(\min\{\epsilon^{-3/2}, n^{1/2}\epsilon^{-1}\})$	$\mathcal{O}\left((n+\nu n^{1/2})\log\epsilon^{-1}\right)$	$\mathcal{O}\left((n+\mu^{-1}n^{1/2})\log\epsilon^{-1}\right)$

- same as best complexities for problems without composition
- lower bound $\mathcal{O}(\min\{\epsilon^{-3/2}, n^{1/2}\epsilon^{-1}\})$ (Fang, Li, Lin & Zhang 2018)
- composition (biased estimator) does not incur higher complexity

Composite Incremental Variance Reduction (CIVR)

input: x_0^1, $\eta > 0$, $T \geq 1$, $\{\tau_t, B_t, S_t\}$ for $t = 1, \dots, T$
for $t = 0, \dots, T-1$
- sample set $\mathcal{B}_t$ with size B_t and construct the estimates

$$y_0^t = \frac{1}{B_t}\sum_{\xi\in\mathcal{B}_t} g_\xi(x_0^t), \qquad z_0^t = \frac{1}{B_t}\sum_{\xi\in\mathcal{B}_t} g'_\xi(x_0^t)$$

- compute $\tilde{\nabla}F(x_0^t) = (z_0^t)^T f'(y_0^t)$ and let $x_1^t = \mathbf{prox}_r^\eta\left(x_0^t - \eta\tilde{\nabla}F(x_0^t)\right)$
- **for** $i = 0, \dots, \tau_t - 1$
  - sample a set $\mathcal{S}_i^t$ with size S_t and construct the estimates

$$y_i^t = y_{i-1}^t + \frac{1}{S_t}\sum_{\xi\in\mathcal{S}_i^t}\left(g_\xi(x_i^t) - g_\xi(x_{i-1}^t)\right)$$
$$z_i^t = z_{i-1}^t + \frac{1}{S_t}\sum_{\xi\in\mathcal{S}_i^t}\left(g'_\xi(x_i^t) - g'_\xi(x_{i-1}^t)\right)$$

  - compute $\tilde{\nabla}F(x_i^t) = (z_i^t)^T f'(y_i^t)$ and $x_{i+1}^t = \mathbf{prox}_r^\eta\left(x_i^t - \eta\tilde{\nabla}F(x_i^t)\right)$
- set $x_0^{t+1} = x_{\tau_t}^t$.

output: $\bar{x}$ randomly chosen from $\left\{x_i^t\right\}_{i=0,\dots,\tau_t-1}^{t=1,\dots,T}$

Convergence analysis

$$\underset{x\in\mathbf{R}^d}{\text{minimize}} \quad \underbrace{f\left(\mathbf{E}_\xi[g_\xi(x)]\right)}_{F(x)} + r(x)$$

- **assumptions**
  - f is ℓ_f-Lipschitz and f' is L_f-Lipschitz
  - r convex but can be non-smooth
  - g_ξ and g'_i are mean-square Lipschitz with constants ℓ_g and L_g

$$\mathbf{E}\left[\|g_\xi(x) - g_\xi(y)\|^2\right] \leq \ell_g^2\|x-y\|^2$$
$$\mathbf{E}\left[\|g'_\xi(x) - g'_\xi(y)\|^2\right] \leq L_g^2\|x-y\|^2$$

  - g_ξ and g'_ξ have bounded variance

$$\mathbf{E}\left[\|g_\xi(x) - g(y)\|^2\right] \leq \sigma_g^2, \qquad \mathbf{E}\left[\|g'_\xi(x) - g'(y)\|^2\right] \leq \sigma_{g'}^2$$

- **sample complexity** for $\mathbf{E}\left[\|\mathcal{G}(x_t)\|^2\right] \leq \epsilon$, where

$$\mathcal{G}(x) = \frac{1}{\eta}\left(x - \mathbf{prox}_r^\eta\left(x - \eta F'(x)\right)\right) \quad = F'(x) \text{ if } r \equiv 0$$

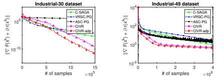
List of convergence results

step size: $\eta \sim \frac{1}{L_F}$ where $L_F = \ell_g^2 L_f + \ell_f L_g$

structure	T	τ_t	S_t	B_t	sample complexity
expectation	$\frac{1}{\sqrt{\epsilon}}$	$\frac{1}{\sqrt{\epsilon}}$	$\frac{1}{\sqrt{\epsilon}}$	$\frac{\sigma_0^2}{\epsilon}$	$\mathcal{O}(\epsilon^{-3/2})$
finite-sum	$\frac{1}{\sqrt{n}\epsilon}$	$\sqrt{n}$	$\sqrt{n}$	n	$\mathcal{O}(n + \sqrt{n}\epsilon^{-1})$
F is ν-gradient dominant					
expectation	$\frac{16\nu\sqrt{\epsilon}}{\eta}$	$\frac{1}{\sqrt{\epsilon}}$	$\frac{1}{\sqrt{\epsilon}}$	$\frac{12\nu\sigma_0^2}{\epsilon}$	$\mathcal{O}\left(\nu\epsilon^{-1}\log\epsilon^{-1}\right)$
finite-sum	$\frac{16\nu}{\sqrt{n}\eta}$	$\sqrt{n}$	$\sqrt{n}$	n	$\mathcal{O}\left(\left(n + \frac{\nu\sqrt{n}}{\eta}\right)\log\epsilon^{-1}\right)$
Φ is μ-optimally strongly convex					
expectation	$\frac{5\sqrt{\epsilon}}{\mu\eta}$	$\frac{1}{\sqrt{\epsilon}}$	$\frac{1}{\sqrt{\epsilon}}$	$\frac{9\sigma_0^2}{2\mu\epsilon}$	$\mathcal{O}\left(\mu^{-1}\epsilon^{-1}\log\epsilon^{-1}\right)$
finite-sum	$\frac{5}{\sqrt{n}\mu\eta}$	$\sqrt{n}$	$\sqrt{n}$	n	$\mathcal{O}\left(\left(n + \frac{\sqrt{n}}{\mu\eta}\right)\log\epsilon^{-1}\right)$

Experiments on risk-averse optimization

- reduction of gradient norm



- reduction of objective value