A Composite Randomized Incremental Gradient Method

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

International Conference on Machine Learning (ICML)

Long Beach, California
June 11, 2019

Composite finite-sum optimization

- problem of focus

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

  - $f : \mathbf{R}^p \to \mathbf{R}$ smooth and possibly nonconvex
  - $g_i : \mathbf{R}^d \to \mathbf{R}^p$ smooth vector mapping, $i = 1, \ldots, n$
  - $r : \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ convex but possibly nonsmooth

Composite finite-sum optimization

- problem of focus

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

  - $f : \mathbf{R}^p \to \mathbf{R}$ smooth and possibly nonconvex
  - $g_i : \mathbf{R}^d \to \mathbf{R}^p$ smooth vector mapping, $i = 1, \ldots, n$
  - $r : \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ convex but possibly nonsmooth

- extensions for two-level finite-sum problem

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

- applications beyond ERM
  - reinforcement learning (policy evaluation)
  - 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

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

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 $g(x) := \frac{1}{n}\sum_{i=1}^{n} g_i(x)$

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

  - subsampled estimators

$$y = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} g_i(x), \quad z = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} g_i'(x), \quad \text{where } \mathcal{S} \subset \{1,\ldots,n\}$$

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

Technical challenge and related work

- challenge: biased gradient estimator
  - denote $F(x) := f(g(x))$ where $g(x) := \frac{1}{n}\sum_{i=1}^{n} g_i(x)$

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

  - subsampled estimators

$$y = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} g_i(x), \quad z = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} g_i'(x), \quad \text{where } \mathcal{S} \subset \{1,\ldots,n\}$$

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

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

Main results

- composite-SAGA: single loop vs double loops of composite-SVRG

Main results

- composite-SAGA: single loop vs double loops of composite-SVRG
- sample complexity for $\mathbf{E}\left[\|\mathcal{G}(x_t)\|^2\right] \leq \epsilon$ (with $\mathcal{G} = F'$ if $r \equiv 0$)
  - nonconvex smooth f and g_i: $O\left(n + n^{2/3}\epsilon^{-1}\right)$
  - $+$ gradient dominant or strongly convex: $O\left((n+\kappa n^{2/3}) \log \epsilon^{-1}\right)$

  same as SVRG/SAGA for nonconvex finite-sum problems
  (Allen-Zhu & Hazan 2016; Reddi et al. 2016; Let et al. 2017)

Main results

- composite-SAGA: single loop vs double loops of composite-SVRG
- sample complexity for $\mathbf{E}\left[\|\mathcal{G}(x_t)\|^2\right] \leq \epsilon$ (with $\mathcal{G} = F'$ if $r \equiv 0$)
  - nonconvex smooth f and g_i: $O\left(n + n^{2/3}\epsilon^{-1}\right)$
  - $+$ gradient dominant or strongly convex: $O\left((n+\kappa n^{2/3}) \log \epsilon^{-1}\right)$

  same as SVRG/SAGA for nonconvex finite-sum problems
  (Allen-Zhu & Hazan 2016; Reddi et al. 2016; Let et al. 2017)

- extensions to two-level problem
  - nonconvex smooth f and g_i: $O\left(m + n + (m+n)^{2/3}\epsilon^{-1}\right)$
    (same as composite-SVRG (Huo et al. 2018))
  - $+$ gradient dominant or optimally strongly convex:

$$O\left((m + n + \kappa(m+n)^{2/3}) \log \epsilon^{-1}\right)$$

    (better than composite-SVRG (Lian et al. 2017))

Composite SAGA algorithm (C-SAGA)

- **input:** $x^0 \in \mathbf{R}^d$, α_i^0 for $i = 1, \ldots, n$, and step size $\eta > 0$
- initialize $Y_0 = \frac{1}{n}\sum_{i=1}^n g_i(\alpha_i^0)\,, \quad Z_0 = \frac{1}{n}\sum_{i=1}^n g_i'(\alpha_i^0)$
- **for** $t = 0, \ldots, T-1$
  - sample with replacement $\mathcal{S}_t \subset \{1, \ldots, n\}$ with $|\mathcal{S}_t| = s$
  - compute $\begin{cases} y_t = Y_t + \frac{1}{s}\sum_{j\in\mathcal{S}_t}(g_j(x^t) - g_j(\alpha_j^t)) \\ z_t = Z_t + \frac{1}{s}\sum_{j\in\mathcal{S}_t}(g_j'(x^t) - g_j'(\alpha_j^t)) \end{cases}$
  - $x^{t+1} = \mathbf{prox}_r^\eta\Big(x^t - \eta\big(z_t^T f'(y_t)\big)\Big)$
  - update $\alpha_j^{t+1} = x^t$ if $j \in \mathcal{S}_t$ and $\alpha_j^{t+1} = \alpha_j^t$ otherwise
  - update $\begin{cases} Y_{t+1} = Y_t + \frac{1}{n}\sum_{j\in\mathcal{S}_t}(g_j(x^t) - g_j(\alpha_j^t)) \\ Z_{t+1} = Z_t + \frac{1}{n}\sum_{j\in\mathcal{S}_t}(g_j'(x^t) - g_j'(\alpha_j^t)) \end{cases}$
- **output:** randomly choose $t_* \in \{1, \ldots, T\}$ and output x^{t_*}

Convergence analysis

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

- **assumptions**
  - f is ℓ_f-Lipschitz and f' is L_f-Lipschitz
  - g_i is ℓ_g-Lipschitz and g_i' is L_g-Lipschitz, $i = 1, \ldots, n$
  - r convex but can be non-smooth

  implication: F' is L_F-Lipschitz with $L_F = \ell_g^2 L_f + \ell_f L_g$

Convergence analysis

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

- **assumptions**
  - f is ℓ_f-Lipschitz and f' is L_f-Lipschitz
  - g_i is ℓ_g-Lipschitz and g_i' is L_g-Lipschitz, $i = 1, \ldots, n$
  - r convex but can be non-smooth

  implication: F' is L_F-Lipschitz with $L_F = \ell_g^2 L_f + \ell_f L_g$

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

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

  - if $s = 1$ and $\eta = O\big(1/(nL_F)\big)$, then complexity $O\big(n/\epsilon\big)$
  - if $s = n^{2/3}$ and $\eta = O\big(1/L_F\big)$, then complexity $O\big(n + n^{2/3}/\epsilon\big)$

Linear convergence results

- **gradient-dominant functions**
  - assumption: $r \equiv 0$ and $F(x) := f\left(\frac{1}{n}\sum_{i=1}^{n} g_i(x)\right)$ satisfies

$$F(x) - \inf_y F(y) \leq \frac{\nu}{2}\|F'(x)\|^2, \quad \forall x \in \mathbf{R}^d$$

  - if $s = n^{2/3}$ and $\eta = O(1/L_F)$, complexity $O\big((n+\nu n^{2/3})\log \epsilon^{-1}\big)$

- **optimally strongly convex functions**
  - assumption: $\Phi(x) := F(x) + r(x)$ satisfies

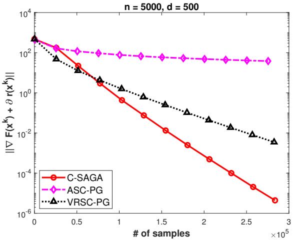
$$\Phi(x) - \Phi(x_\star) \geq \frac{\mu}{2}\|x - x_\star\|^2, \quad \forall x \in \mathbf{R}^d$$

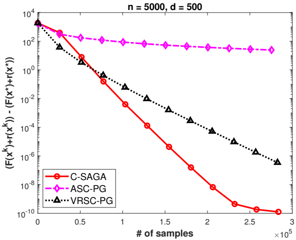
  - if $s = n^{2/3}$ and $\eta = O(1/L_F)$, complexity $O\big((n+\mu^{-1} n^{2/3})\log \epsilon^{-1}\big)$

- extension to two-level case: $O\big((m + n + \kappa(m+n)^{2/3})\log \epsilon^{-1}\big)$

Experiments

- risk-averse optimization





- policy evaluation for MDP