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

\[
\text{minimize } \quad f \left( \frac{1}{n} \sum_{i=1}^{n} g_i(x) \right) + r(x)
\]

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

\[ \text{applications beyond ERM} \]
- reinforcement learning (policy evaluation)
- risk-averse optimization, financial mathematics
- \[ \ldots \]
Composite finite-sum optimization

• problem of focus

\[
\minimize_{x \in \mathbb{R}^d} \quad f \left( \frac{1}{n} \sum_{i=1}^{n} g_i(x) \right) + r(x)
\]

– \( f : \mathbb{R}^p \rightarrow \mathbb{R} \) smooth and possibly nonconvex
– \( g_i : \mathbb{R}^d \rightarrow \mathbb{R}^p \) smooth vector mapping, \( i = 1, \ldots, n \)
– \( r : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\} \) convex but possibly nonsmooth

• extensions for two-level finite-sum problem

\[
\minimize_{x \in \mathbb{R}^d} \quad \frac{1}{m} \sum_{j=1}^{m} f_j \left( \frac{1}{n} \sum_{i=1}^{n} g_i(x) \right) + r(x)
\]

• applications beyond ERM
  – reinforcement learning (policy evaluation)
  – risk-averse optimization, financial mathematics
  – ...
Examples

• policy evaluation with linear function approximation

\[
\begin{align*}
\text{minimize} & \quad \mathbf{E} \left[ A \right] x - \mathbf{E} [ b ]^2 \\
\text{subject to} & \quad x \in \mathbb{R}^d \\
A, \ b \ & \text{random, generated by MDP under fixed policy}
\end{align*}
\]
Examples

• policy evaluation with linear function approximation
  \[
  \min_{x \in \mathbb{R}^d} \| \mathbb{E}[A]x - \mathbb{E}[b] \|^2
  \]
  \(A, b\) random, generated by MDP under fixed policy

• risk-averse optimization
  \[
  \max_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{j=1}^{n} h_j(x) - \lambda \frac{1}{n} \sum_{j=1}^{n} \left( h_j(x) - \frac{1}{n} \sum_{i=1}^{n} h_i(x) \right)^2
  \]
  - average reward
  - variance of rewards (risk)
  - often treated as two-level composite finite-sum optimization
Examples

• policy evaluation with linear function approximation

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

    $A, b$ random, generated by MDP under fixed policy

• risk-averse optimization

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

    $$\text{average reward} \quad \text{variance of rewards (risk)}$$

    – often treated as two-level composite finite-sum optimization

    – simple transformation using $\text{Var}(a) = \mathbb{E}[a^2] - (\mathbb{E}[a])^2$

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

    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

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

    $\mathbb{E}[y] = g(x)$ and $\mathbb{E}[z] = g'(x)$, but $\mathbb{E}[[z]^T f'(y)] \neq F'(x)$
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
    \[
    y = \frac{1}{|S|} \sum_{i \in S} g_i(x), \quad z = \frac{1}{|S|} \sum_{i \in S} g'_i(x), \quad \text{where} \ S \subset \{1, \ldots, n\}
    \]
    \[
    \mathbb{E}[y] = g(x) \quad \text{and} \quad \mathbb{E}[z] = g'(x), \quad \text{but} \ \mathbb{E} \left[ [z]^T f'(y) \right] \neq F'(x)
    \]

• related work
  – more general composite stochastic optimization
    (Wang, Fang & Liu 2017; Wang, Liu & Fang 2017; \ldots )
  – two-level composite finite-sum: extending SVRG
    (Lian, Wang & Liu 2017; Huo, Gu, Liu & Huang 2018; Lin, Fan, Wang & Jordan 2018; \ldots )
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 $\mathbb{E}[\|G(x_t)\|^2] \leq \epsilon$ (with $G = F'$ if $r \equiv 0$)
  - nonconvex smooth $f$ and $g_i$: $O(n + n^{2/3}\epsilon^{-1})$
  - + gradient dominant or strongly convex: $O((n + \kappa n^{2/3}) \log \epsilon^{-1})$

same as SVRG/SAGA for nonconvex finite-sum problems
Main results

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

• sample complexity for \(\mathbb{E}[\|G(x_t)\|^2] \leq \epsilon\) (with \(G = F'\) if \(r \equiv 0\))
  - nonconvex smooth \(f\) and \(g_i\): \(O(n + n^{2/3}\epsilon^{-1})\)
  - + gradient dominant or strongly convex: \(O((n + \kappa n^{2/3}) \log \epsilon^{-1})\)

same as SVRG/SAGA for nonconvex finite-sum problems

• extensions to two-level problem
  - nonconvex smooth \(f\) and \(g_i\): \(O(m + n + (m + n)^{2/3}\epsilon^{-1})\)
    (same as composite-SVRG (Huo et al. 2018))
  - + gradient dominant or optimally strongly convex:
    \[O((m + n + \kappa(m + n)^{2/3}) \log \epsilon^{-1})\]
    (better than composite-SVRG (Lian et al. 2017))
**Composite SAGA algorithm (C-SAGA)**

- **input:** \( x^0 \in \mathbb{R}^d, \alpha_i^0 \) 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 \( S_t \subset \{1, \ldots, n\} \) with \( |S_t| = s \)
  - compute \[
  \begin{align*}
  y_t &= Y_t + \frac{1}{s} \sum_{j \in S_t} (g_j(x^t) - g_j(\alpha_j^t)) \\
  z_t &= Z_t + \frac{1}{s} \sum_{j \in S_t} (g'_j(x^t) - g'_j(\alpha_j^t))
  \end{align*}
  \]
  - \( x^{t+1} = \text{prox}_\eta^r \left( x^t - \eta (z_t^T f'(y_t)) \right) \)
  - update \( \alpha_j^{t+1} = x^t \) if \( j \in S_t \) and \( \alpha_j^{t+1} = \alpha_j^t \) otherwise
  - update \[
  \begin{align*}
  Y_{t+1} &= Y_t + \frac{1}{n} \sum_{j \in S_t} (g_j(x^t) - g_j(\alpha_j^t)) \\
  Z_{t+1} &= Z_t + \frac{1}{n} \sum_{j \in S_t} (g'_j(x^t) - g'_j(\alpha_j^t))
  \end{align*}
  \]
- **output:** randomly choose \( t* \in \{1, \ldots, T\} \) and output \( x^{t*} \)
Convergence analysis

\[ \text{minimize } \quad \mathbf{x} \in \mathbb{R}^{d} \quad f \left( \frac{1}{n} \sum_{i=1}^{n} g_{i}(x) \right) + r(x) \]

- **assumptions**
  - \( f \) is \( \ell_{f} \)-Lipschitz and \( f' \) is \( L_{f} \)-Lipschitz
  - \( g_{i} \) is \( \ell_{g} \)-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

\[
\minimize_{x \in \mathbb{R}^d} \quad f \left( \frac{1}{n} \sum_{i=1}^{n} g_i(x) \right) + r(x)
\]

- assumptions
  - \( f \) is \( \ell_f \)-Lipschitz and \( f' \) is \( L_f \)-Lipschitz
  - \( g_i \) is \( \ell_g \)-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 \( \mathbb{E} \left[ \|G(x_t)\|^2 \right] \leq \epsilon \), where
  \[
  G(x) = \frac{1}{\eta} \left( x - \text{prox}_r^\eta \left( x - \eta F'(x) \right) \right) = F'(x) \text{ if } r \equiv 0
  \]

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

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

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

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

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

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

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

• risk-averse optimization

![](image1.png)

• policy evaluation for MDP

![](image2.png)