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

Abstract

Technical challenge and related work
• challenge: biased gradient estimator
  – denote \( F(x) := f(g(x)) \) where \( g(x) := E_\xi [g_\xi(x)] \)
  \( F'(x) = [g'_{\xi}(x)]^T f'(g(x)) \)
  – subsampled estimators
  \[ y = \frac{1}{m} \sum_{\xi \in S} g_\xi(x), \quad z = \frac{1}{m} \sum_{\xi \in S} g'_\xi(x) \]
  \( E[y] = g(x) \) and \( E[z] = g'(x) \), but \( E[Z]^T F'(y) \neq F'(x) \)

• related work
  – more general composite stochastic optimization
    (Wang, Fang & Liu 2017; Wang, Liu & Fang 2017; . . .)
  – two-level composite finite-sum: extending SVRG and SAGA

Convergence analysis

Main results

\[
\text{minimize } \psi(x) + r(x) = f(E[g(x)]) + r(x)
\]

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

sample complexities

\[
\begin{align*}
F(\text{x nonconvex}) & \quad \mathcal{O}(\varepsilon^{-3/2}) \\
F(\text{r gradient dominant}) & \quad \mathcal{O}(\varepsilon^{-1} \log(1/\varepsilon)) \\
F(\text{r convex}) & \quad \mathcal{O}(\varepsilon^{-3/2} \log(1/\varepsilon)) \\
F(\text{r convex, r convex}) & \quad \mathcal{O}(\varepsilon^{-1} / n) \\
\end{align*}
\]

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

List of convergence results

step size: \( \eta \sim \frac{1}{\ell} \) where \( \ell_F = \ell_f L_f + \ell_r L_r \)

- finite-sum
  • expectation \( O(\varepsilon^{-3/2}) \)
  • finite-sum \( O(n + \sqrt{n} \varepsilon^{-1}) \)

- \( \Phi \) is \( \mu \)-optimally strongly convex
  • expectation \( O(n(\mu^{-1} \log(1/\varepsilon)) \)
  • finite-sum \( O(n + \frac{\sqrt{n} \varepsilon}{\mu}) \log(1/\varepsilon) \)

Examples

• policy evaluation with linear function approximation
  \[
  \text{minimize } x \in \mathbb{R}^d \quad f(g(x)) + r(x)
  \]
  \( f \); \( g \); \( r \) random, generated by MDP under fixed policy

• risk-averse optimization
  \[
  \text{maximize } \frac{1}{n} \sum_{i=1}^n b_i(x) - \lambda \left( \frac{1}{n} \sum_{i=1}^n h_i(x) - \frac{1}{n} \sum_{i=1}^n b_i(x) \right)^2
  \]
  \( \lambda \); \( \lambda \) variance of rewards (risk)

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

Composite Incremental Variance Reduction (CIVR)

input: \( x_0; \eta > 0; T \geq 1; \{B_t, S_t, B_r\} \) for \( t = 1, \ldots, T \)
for \( t = 0; \ldots, T - 1 \)
  • sample set \( S_t \) with size \( B_t \) and construct the estimates
    \[
    \hat{y}_t = \frac{1}{B_t} \sum_{i \in S_t} g_i(x_t), \quad \hat{z}_t = \frac{1}{B_t} \sum_{i \in S_t} g'_i(x_t)
    \]
  • compute \( \hat{F}(x_t) = (\hat{y}_t) \nabla f'(\hat{y}_t) \) and let \( x_t' = \text{prox}_{\lambda \hat{F}}(x_t - \eta \hat{F}(x_t)) \)

for \( i = 0; \ldots, T - 1 \)
  • sample a set \( S'_{i+1} \) with size \( S_{i+1} \) and construct the estimates
    \[
    \hat{y}'_{i+1} = y_{i+1} + \frac{1}{B_r} \sum_{i \in S_{i+1}} (g_i(x'_i) - g_i(x'_i - 1))
    \]
  • compute \( \hat{F}(x'_i) = (\hat{y}'_{i+1}) \nabla f'(\hat{y}'_{i+1}) \) and let \( x'_i = \text{prox}_{\lambda \hat{F}}(x'_i - \eta \hat{F}(x'_i)) \)
  • set \( x'_0 = x_0 \)

output: `x` sample chosen from \( \{x'_0, \ldots, x'_T\} \)

Experiments on risk-averse optimization

• reduction of gradient norm
  \[
  \| \nabla f(x) \|^2
  \]
  \( f \) and \( f' \) are mean-square Lipschitz with constants \( \ell_f \) and \( L_f \)

• reduction of objective value
  \[
  f(x) - f(x*)
  \]