Optimization Theory

Lecture 14

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

Stochastic Recursive Gradient Algorithm (SARAH)

Algorithm 1 Stochastic Variance Reduced Gradient

Input: $\mathbf{x}_0$, η, m, S
$\tilde{\mathbf{x}}^{(0)} = \mathbf{x}_0$
for $s = 0, \dots, S-1$
  $\mathbf{v}_0 = \nabla f(\tilde{\mathbf{x}}^{(s)})$
  $\mathbf{x}_0 = \tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{(s)}$
  for $t = 0, \dots, m-1$
    draw i_t from $\{1, \dots, n\}$ uniformly
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$
    $\mathbf{v}_{t+1} = \nabla f_{i_t}(\mathbf{x}_{t+1}) - \nabla f_{i_t}(\mathbf{x}_t) + \mathbf{v}_t$
  end for
  $\tilde{\mathbf{x}}^{(s+1)} = \mathbf{x}_t$ for randomly chosen $t \in \{0, \dots, m-1\}$
end for
Output: $\tilde{\mathbf{x}}^{(S)}$

Stochastic Recursive Gradient Algorithm (SARAH)

SARAH outputs $\tilde{\mathbf{x}}^{(S)}$ satisfying $\mathbb{E}\left\|\nabla f(\tilde{\mathbf{x}}^{(S)})\right\|_2 \leq \epsilon$ within

1. $\mathcal{O}\big((n+\kappa)\log(1/\epsilon)\big)$ IFO complexity for strongly convex objective;
2. $\mathcal{O}\big((n+L/\epsilon^2)\log(1/\epsilon)\big)$ IFO complexity for convex objective.

The more interesting result is in the nonconvex optimization:

1. Cong Fang, Chris Junchi Li, Zhouchen Lin, Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *NeurIPS* 2018.

SGD for Nonconvex Optimization

We consider the stochastic optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \mathbb{E}_{\xi}[F(\mathbf{x};\xi)],$$

where $f(\mathbf{x})$ is L-smooth and lower bounded, and each $F(\mathbf{x};\xi)$ is differentiable.

Suppose there exists $\sigma > 0$ such that $\mathbb{E}\left\|\nabla F(\mathbf{x};\xi) - \nabla f(\mathbf{x})\right\|_2^2 \leq \sigma^2$ for any $\mathbf{x} \in \mathbb{R}^d$. We run SGD iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \cdot \frac{1}{|\mathcal{S}_t|}\sum_{\xi\in\mathcal{S}_t} \nabla F(\mathbf{x}_t;\xi)$$

with $\mathcal{S}_t = \{\xi_1, \dots, \xi_b\}$, where $\xi_i \overset{\text{i.i.d}}{\sim} \mathcal{D}$.

It can find an ϵ-stationary point of $f(\cdot)$ within

$$\mathcal{O}(L\sigma^2\epsilon^{-4})$$

stochastic first-order oracle (SFO) complexity in expectation.

SARAH/SPIDER for Nonconvex Optimization

We consider the L-average smooth function, i.e. there exists $L > 0$ such that

$$\mathbb{E}\left\|\nabla F(\mathbf{x};\xi) - \nabla F(\mathbf{y};\xi)\right\|_2^2 \leq L^2\left\|\mathbf{x} - \mathbf{y}\right\|_2^2$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

The algorithms with stochastic recursive gradient require

$$\mathcal{O}(\sigma^2\epsilon^{-2} + L\sigma^2\epsilon^{-3})$$

SFO complexity to find an ϵ-stationary point.

SARAH/SPIDER for Nonconvex Optimization

We consider the finite-sum problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}).$$

Under the L-average smooth assumption, the algorithms with stochastic recursive gradient require

$$\mathcal{O}(n + L\sqrt{n}\epsilon^{-2})$$

SFO complexity to find an ϵ-stationary point.

Algorithm 2 ProbAbilistic Gradient Estimator (PAGE)

Input: η, T, b_0, b and p.

$\mathcal{S}_0 = \{\xi_1, \dots, \xi_{b_0}\}$ with $\xi_i \overset{\text{i.i.d}}{\sim} \mathcal{D}$

$\mathbf{v}_0 = \frac{1}{b_0} \sum_{\xi \in \mathcal{S}_0} \nabla F(\mathbf{x}_0; \xi)$

for $t = 0, 1, \dots, T$ do

  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$

  draw $\zeta_t \sim \mathrm{Bernoulli}(p)$

  if $\zeta_t = 1$ **then**

    $\mathcal{S}_{t+1} = \{\xi_1, \dots, \xi_{b_0}\}$ where $\xi_i \overset{\text{i.i.d}}{\sim} \mathcal{D}$

    $\mathbf{v}_{t+1} = \frac{1}{b_0} \sum_{\xi \in \mathcal{S}_{t+1}} \nabla F(\mathbf{x}_{t+1}; \xi)$

  else

    $\mathcal{S}_{t+1} = \{\xi_1, \dots, \xi_b\}$ where $\xi_i \overset{\text{i.i.d}}{\sim} \mathcal{D}$

    $\mathbf{v}_{t+1} = \mathbf{v}_t + \frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi))$

  end if

end for

$\mathbf{x}_{\mathrm{out}} = \mathbf{x}_t$ for randomly chosen $t \in \{0, \dots, T-1\}$

Outline