SAGA

Introduction to Variance Reduction

Yilin Gu

School of Data Science

February 18, 2022

Optimization Problem

Problem 1

$$\min_{x\in\mathbb{R}^n} \; f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

- Each $f_i : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with constant μ.
- Each $\nabla f_i : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant L.

Problem 2

$$\min_{x\in\mathbb{R}^n} \; F(x) := f(x) + h(x)$$

- Function $h : \mathbb{R}^n \to \mathbb{R}$ is convex but potentially nonsmooth.
- The proximal operation of h is easy to compute.

Optimization Problem

Problem 1

$$\min_{x\in\mathbb{R}^n} \; f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x)$$

- Each $f_i : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with constant μ.
- Each $\nabla f_i : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant L.

Problem 2

$$\min_{x\in\mathbb{R}^n} \; F(x) := f(x) + h(x)$$

- Function $h : \mathbb{R}^n \to \mathbb{R}$ is convex but potentially nonsmooth.
- The proximal operation of h is easy to compute.

GD Versus SGD

Optimization problem: $\min_{x\in\mathbb{R}^n} \; f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x)$.

GD

$$\boxed{x^{k+1} = x^k - \gamma \nabla f(x^k)}$$

Pros: Can use constant stepsize γ and achieve linear convergence.

Cons: Evaluation of gradient $\nabla f(x)$ is expensive.

SGD

$$\boxed{x^{k+1} = x^k - \gamma_k \nabla f_j(x^k)}$$

Pros: Evaluate few gradient per iteration.

Cons: Stepsize $\{\gamma_k\}_k$ is decreasing which leads to sublinear convergence.

Motivation: To combine the advantages of both GD and SGD.

GD Versus SGD

Optimization problem: $\min_{x\in\mathbb{R}^n} \; f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x)$.

GD

$$\boxed{x^{k+1} = x^k - \gamma \nabla f(x^k)}$$

Pros: Can use constant stepsize γ and achieve linear convergence.

Cons: Evaluation of gradient $\nabla f(x)$ is expensive.

SGD

$$\boxed{x^{k+1} = x^k - \gamma_k \nabla f_j(x^k)}$$

Pros: Evaluate few gradient per iteration.

Cons: Stepsize $\{\gamma_k\}_k$ is decreasing which leads to sublinear convergence.

Motivation: To combine the advantages of both GD and SGD.

Stochastic Perspective (I)

Define random variable $G : \Omega \to \mathbb{R}^n,\ \Omega := \{1, \cdots, n\}$ as

$$G = \begin{cases} \nabla f_1(x^k), & \text{w.p. } \frac{1}{n} \\ \nabla f_2(x^k), & \text{w.p. } \frac{1}{n} \\ \quad \vdots & \\ \nabla f_n(x^k), & \text{w.p. } \frac{1}{n} \end{cases}$$

Notice that $\mathbb{E}[G] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^k) = \nabla f(x^k)$.

GD

$$\boxed{x^{k+1} = x^k - \gamma \nabla f(x^k) = x^k - \gamma \mathbb{E}[G]}$$

Pros: Can use constant stepsize γ and achieve linear convergence.

Cons: Evaluation of gradient $\mathbb{E}[G]$ is expensive.

Stochastic Perspective (II)

Define random variable $G : \Omega \to \mathbb{R}^n,\ \Omega := \{1, \cdots, n\}$ as

$$G = \begin{cases} \nabla f_1(x^k), & \text{w.p. } \frac{1}{n} \\ \nabla f_2(x^k), & \text{w.p. } \frac{1}{n} \\ \vdots & \\ \nabla f_n(x^k), & \text{w.p. } \frac{1}{n} \end{cases}$$

Note that j is chosen randomly uniformly from $\{1, \cdots, n\}$.

SGD

$$\boxed{x^{k+1} = x^k - \gamma_k \nabla f_j(x^k) = x^k - \gamma_k G}$$

Pros: Evaluate few gradient per iteration.

Cons: Stepsize $\{\gamma_k\}_k$ is decreasing which leads to sublinear convergence.

Variance Reduction

$$\textbf{GD:} \quad x^{k+1} = x^k - \gamma \cdot \mathbb{E}[G]$$

$$\textbf{SGD:} \quad x^{k+1} = x^k - \gamma_k \cdot G$$

Goal: Use some random variable Z to estimate $\mathbb{E}[G] = \nabla f(x^k)$ with less cost and variance so that to use constant stepsize.

$$\textbf{New Algorithm:} \quad x^{k+1} = x^k - \gamma \cdot Z$$

Consider a random variable

$$Z = G - Y + \mathbb{E}[Y]$$

- Z is a unbiased estimator of $\nabla f(x^k)$ because $\mathbb{E}[Z] = \mathbb{E}[G]$.
- The variance of Z diminishes as G and Y become more correlated

$$\mathrm{Var}(Z) = \mathrm{Var}(G) + \mathrm{Var}(Y) - 2\mathrm{Cov}(G, Y).$$

Question: How to choose the random variable Y?

Variance Reduction

$$\textbf{GD:} \quad x^{k+1} = x^k - \gamma \cdot \mathbb{E}[G]$$
$$\textbf{SGD:} \quad x^{k+1} = x^k - \gamma_k \cdot G$$

Goal: Use some random variable Z to estimate $\mathbb{E}[G] = \nabla f(x^k)$ with less cost and variance so that to use constant stepsize.

$$\textbf{New Algorithm:} \quad x^{k+1} = x^k - \gamma \cdot Z$$

Consider a random variable

$$\boxed{Z = G - Y + \mathbb{E}[Y]}$$

- Z is a unbiased estimator of $\nabla f(x^k)$ because $\mathbb{E}[Z] = \mathbb{E}[G]$.
- The variance of Z diminishes as G and Y become more correlated

$$\mathrm{Var}(Z) = \mathrm{Var}(G) + \mathrm{Var}(Y) - 2\mathrm{Cov}(G, Y).$$

Question: How to choose the random variable Y?

Variance Reduction Algorithms

Space Versus Time

Iteration: $x^{k+1} = x^k - \gamma Z$

- The index j is chosen randomly uniformly from set $\{1, \cdots, n\}$.
- **SAGA:** Set the random variable Z as

$$Z = \underbrace{\nabla f_j(x^k)}_{G} - \underbrace{\nabla f_j(\phi_j^k)}_{Y} + \overbrace{\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\phi_i^k)}^{\mathbb{E}[Y]}$$

  Store $\nabla f_j(\phi_j^{k+1}) = \nabla f_j(x^k)$ and $\nabla f_i(\phi_i^{k+1}) = \nabla f_i(\phi_i^k)$ for $i \neq j$.

- **SVRG:** Set the random variable Z as

$$Z = \underbrace{\nabla f_j(x^k)}_{G} - \underbrace{\nabla f_j(\tilde{x})}_{Y} + \overbrace{\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x})}^{\mathbb{E}[Y]}$$

  Compute the full gradient $f(\tilde{x})$ after m iterations.

Variance Reduction Algorithms

Space Versus Time

Iteration: $x^{k+1} = x^k - \gamma Z$

- The index j is chosen randomly uniformly from set $\{1, \cdots, n\}$.
- **SAGA:** Set the random variable Z as

$$Z = \underbrace{\nabla f_j(x^k)}_{G} - \underbrace{\nabla f_j(\phi_j^k)}_{Y} + \overbrace{\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\phi_i^k)}^{\mathbb{E}[Y]}$$

  Store $\nabla f_j(\phi_j^{k+1}) = \nabla f_j(x^k)$ and $\nabla f_i(\phi_i^{k+1}) = \nabla f_i(\phi_i^k)$ for $i \neq j$.

- **SVRG:** Set the random variable Z as

$$Z = \underbrace{\nabla f_j(x^k)}_{G} - \underbrace{\nabla f_j(\tilde{x})}_{Y} + \overbrace{\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x})}^{\mathbb{E}[Y]}$$

  Compute the full gradient $f(\tilde{x})$ after m iterations.

SAGA: Algorithm Framework

At k-th iteration

- Pick j uniformly from $\{1, \cdots, n\}$;
- Update x using $\nabla f_j(\phi_j^k), \nabla f_j(x^k)$ and the table average

$$x^{k+1} = x^k - \gamma \left(\nabla f_j(x^k) - \nabla f_j(\phi_j^k) + \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\phi_i^k) \right)$$

- Update the table by setting $\nabla f_j(\phi_j^{k+1}) = \nabla f_j(x^k)$ and $\forall i \neq j$, $\nabla f_i(\phi_i^{k+1}) = \nabla f_i(\phi_i^k)$;

Example: Consider $n = 3$ and $j = 2$, the update of table:

Old	New
$\nabla f_1(\phi_1)$	$\nabla f_1(\phi_1)$
$\nabla f_2(\phi_2)$	$\nabla f_2(x)$
$\nabla f_3(\phi_3)$	$\nabla f_3(\phi_3)$

Theoretical Result

- Define the Lyapunov function T, where $c = \frac{1}{2\gamma(1-\gamma\mu)n}$

$$T^k := T(x^k, \{\phi_i^k\}_{i=1}^n)$$

$$:= \frac{1}{n}\sum_{i=1}^{n} f_i(\phi_i^k) - \overbrace{\left[f_i(x^*) + \left\langle \nabla f_i(x^*), \phi_i^k - x^* \right\rangle\right]}^{\leq f_i(\phi_i^k)} + c\|x^k - x^*\|^2.$$

- **(Main Theorem)** Run SAGA with constant stepsize $\gamma = \frac{1}{2(\mu n + L)}$,

$$\mathbb{E}[T^{k+1}] \leq (1 - \frac{1}{\kappa}) \cdot T^k,$$

where the constant $\kappa = \frac{1}{\gamma\mu}$.

- **(Corollary)** Since $\|x^k - x^*\|^2 \leq T^k/c$ for all $k \in \mathbb{N}$,

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \frac{(1-\frac{1}{\kappa})^k T^0}{c}.$$

Theoretical Result

- Define the Lyapunov function T, where $c = \frac{1}{2\gamma(1-\gamma\mu)n}$

$$T^k := T(x^k, \{\phi_i^k\}_{i=1}^n)$$

$$:= \frac{1}{n}\sum_{i=1}^n f_i(\phi_i^k) - \overbrace{\left[f_i(x^*) + \left\langle \nabla f_i(x^*), \phi_i^k - x^* \right\rangle\right]}^{\leq f_i(\phi_i^k)} + c\|x^k - x^*\|^2.$$

- **(Main Theorem)** Run SAGA with constant stepsize $\gamma = \frac{1}{2(\mu n + L)}$,

$$\mathbb{E}[T^{k+1}] \leq (1 - \frac{1}{\kappa}) \cdot T^k,$$

where the constant $\kappa = \frac{1}{\gamma\mu}$.

- **(Corollary)** Since $\|x^k - x^*\|^2 \leq T^k/c$ for all $k \in \mathbb{N}$,

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \frac{(1 - \frac{1}{\kappa})^k T^0}{c}.$$