Stochastic Variance-Reduced Optimization

Qincheng Lu

Nov 21, 2019

Finite Sum Minimization Problem

$$\min_{x \in \mathbb{R}^d} := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \Rightarrow \min_{x \in \mathbb{R}^d} := \{\varphi(x) + \frac{1}{n} \sum_{i=1}^{n} f_i(x)\}$$

- $\varphi(x)$ - a convex regularizer
- $f_i(x)$ - a convex loss function

Eg: $a_i \in \mathbb{R}^d, b_i \in \{\pm 1\}$

- Ridge Regression $\frac{\lambda}{2} \|x\|^2 + \frac{1}{2n} \sum_{i=1}^{n} (a_i^T x - b_i)^2$
- Lasso Regression $\lambda \|x\|_1 + \frac{1}{2n} \sum_{i=1}^{n} (a_i^T x - b_i)^2$
- SVM $\frac{\lambda}{2} \|x\|^2 + \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - b_i \cdot a_i^T x\}$
- Logistic Regression $\lambda \|x\|_1 + \frac{1}{n} \sum_{i=1}^{n} log(1 + e^{-b_i \cdot a_i^T x})$
- Regularized Generalized Linear Model
- ...

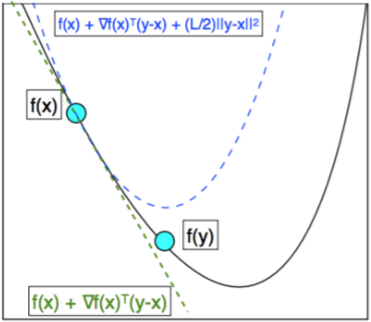
Observation

$$\min_{x \in \mathbb{R}^d} := \{\varphi(x) + \frac{1}{n} \sum_{i=1}^{n} f_i(a_i^T x)\}$$

Complexity Analysis

***L*-Smooth:** $f(y) \le f(x) + \nabla f(x)^\intercal (y-x) + \frac{L}{2}\|y-x\|^2$

For convex f, this is equivalent to saying f has L-Lipschitz continuous gradient: $\|\nabla f(y) - \nabla f(x)\| \le L\|y-x\|$



Complexity Analysis

Use update rule $x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$ and L-smooth assumption:

$\Rightarrow f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^\intercal (x^{k+1} - x^k) + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2$

$\Rightarrow f(x^{k+1}) \leq f(x^k) - \frac{1}{L}\nabla f(x^k)^\intercal \nabla f(x^k) + \frac{L}{2}\left\|\frac{1}{L}\nabla f(x^k)\right\|^2$

$\Rightarrow f(x^{k+1}) \leq f(x^k) - \frac{1}{L}\left\|\nabla f(x^k)\right\|^2 + \frac{1}{2L}\left\|\nabla f(x^k)\right\|^2$

Bound on guaranteed progress with $\frac{1}{L}$ as step-size

$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\left\|\nabla f(x^k)\right\|^2$

$\Rightarrow \left\|\nabla f(x^k)\right\|^2 \leq 2L\left[f(x^k) - f(x^{k+1})\right]$

Complexity Analysis

μ-Strongly Convex: $f(y) \geq f(x) + \nabla f(x)^{\intercal}(y-x) + \frac{\mu}{2}\|y-x\|^2$

It impelies Polyak-Łojasiewicz (PL) inequality:
$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$ where f^* is the optimal value

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\left\|\nabla f(x^k)\right\|^2 \leq f(x^k) - \frac{\mu}{L}(f(x^k) - f^*)$$

$$f(x^{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)(f(x^k) - f^*) \leq \left(1 - \frac{\mu}{L}\right)^{k+1}[f(x^0) - f^*]$$

Since $\mu \leq L$ and $\left(1 - \frac{\mu}{L}\right)^k \leq e^{-k\frac{\mu}{L}}$

We get $f(x^k) - f^* \leq Me^{-k\frac{\mu}{L}}$

$(f(x^0) - f^* = M$ as a constant term$)$

Complexity Analysis

$f(x^k) - f^* \leq Me^{-k\frac{\mu}{L}}$, $f(x^0) - f^* = M$ is a constant

Stopping criteria: $f(x^k) - f^* \leq \epsilon$

$Me^{-k\frac{\mu}{L}} \leq \epsilon \Rightarrow$ Convergence at k^{th} iteration

$Me^{-k\frac{\mu}{L}} \leq \epsilon \Rightarrow k \geq \frac{L}{\mu} log \frac{M}{\epsilon} = O(log \frac{1}{\epsilon})$

Iteration Complexity

The smallest k such that we are within ϵ

Linear Convergence

To get $f(x^k) - f^* \leq \epsilon$, needs $O(log \frac{1}{\epsilon})$ iterations
Without μ-Strongly Convex assumption, this number is $O(\frac{1}{\epsilon})$

For $t = log \frac{1}{\epsilon}$
To get 1 digit, need 2.302585 iterations
To get 10 digit, need 23.02585 iterations

$$x^{k+1} = x^k - \overset{\text{Stepsize}}{\alpha}\, \underset{\substack{\text{Unbiased Estimator of the Gradient} \\ \mathbb{E}[g^k] = \nabla f(x^k)}}{g^k}$$

Variance Matter: $\mathbb{V}[g^k] = \mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right]$

GD: $g^k = \nabla f(x^k) \Rightarrow \mathbb{V}[g^k] = 0$

SGD: $g^k = \nabla f_i(x^k) \Rightarrow \mathbb{V}[g^k] \neq 0$

Convergence Behaviour of SGD

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^\intercal (x^{k+1} - x^k) + \frac{L}{2} \left\| x^{k+1} - x^k \right\|^2$$

SGD update rule: $x^{k+1} = x^k - \alpha_k \nabla f_{i_k}(x^k)$

$$\Rightarrow f(x^{k+1}) \leq f(x^k) - \alpha_k \nabla f(x^k)^\intercal \nabla f_{i_k}(x^k) + \alpha_k^2 \frac{L}{2} \left\| \nabla f_{i_k}(x^k) \right\|^2$$

Take the expectation for i_k with $\mathbb{E}[\nabla f_{i_k}(x^k)] = \nabla f(x^k)$ (unbiased)

$$\Rightarrow \mathbb{E}[f(x^{k+1})] \leq f(x^k) - \alpha_k \left\| \nabla f(x^k) \right\|^2 + \alpha_k^2 \frac{L}{2} \left\| \nabla f_{i_k}(x^k) \right\|^2$$

With strong convexity assumption:

$$\Rightarrow \mathbb{E}[f(x^{k+1})] - f^* \leq (1 - \alpha_k \mu)[f(x^k) - f^*] + \alpha_k^2 \frac{L}{2} \left\| \nabla f_{i_k}(x^k) \right\|^2$$

"converge" to a ball with radius proportional to α_k

To get convergence, we need a decreasing step size

Variance Reduced Method

Basic idea: find a better gradient estimator

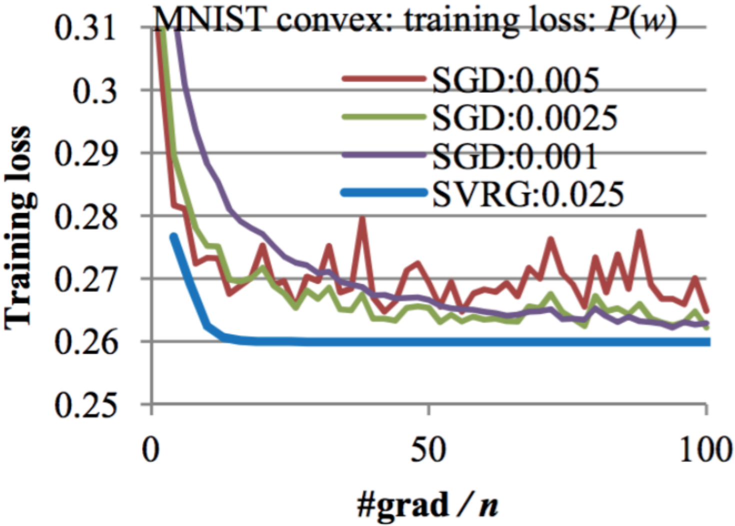
An example: SVRG [Johnson, Zhang (2013)]

- chose an epoch which contains m iterations ($m = 2n$)
- $\widetilde{x} = x_0 = x_m = x_{2m}\ldots$ is the snap shot point
- Calculate full gradient at snap shot point $\widetilde{g} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\widetilde{x})$
- Use gradient estimator at follow up iteration
  $g^k = \widetilde{g} - \nabla f_i(\widetilde{x}) + \nabla f_i(x_k)$
- Unbiased since $\mathbb{E}[g^k] = \mathbb{E}[\widetilde{g} - \nabla f_i(\widetilde{x})] + \mathbb{E}[\nabla f_i(x_k)] = \nabla f(x_k)$
- $\left\|g^k - \nabla f(x_k)\right\|^2$ approaches 0

Other methods like SAG [LeRoux, Schmidt, Bach (2012)]
SAGA [Defazio, Bach, LacosteJulien (2014)]

Experiment

Multiclass Logistic Regression on MNIST



Roadmap

- Fenchel Dual $g^*(y) := \max_{x \in \mathbb{R}^d} \{y^\intercal x - g(x)\}$
- $\boxed{\text{Primal: } \min_{x \in \mathbb{R}^d} \{\varphi(x) + \frac{1}{n} \sum_{i=1}^n f_i(a_i^\intercal x)\}}$
- Primal-Dual: $\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^n} \{\varphi(x) - \frac{1}{n} \sum_{i=1}^n f_i^*(y_i) + \frac{1}{n} y^\intercal A x\}$
- Dual: $-\min_{y \in \mathbb{R}^n} \{\varphi^*(-\frac{1}{n} A^\intercal y) + \frac{1}{n} \sum_{i=1}^n f_i^*(y_i)\}$

Thank You