535520: Optimization Algorithms
Lecture 7 — Stochastic Gradient Descent and Variance Reduction

Ping-Chun Hsieh (謝秉均)

October 21, 2024

Announcement

- Final project: Please select a paper and fill out the Google form: https://forms.gle/XAAoeWFxi5ZUDhpG7

This Lecture

1. Stochastic Gradient Descent

2. Accelerating SGD: Variance Reduction

- Reading Material:
  - Chapter 4 of the textbook “Optimization Methods for Large-Scale Machine Learning” by Leon Bottou, Frank Curtis, and Jorge Nocedal.
    - Available at https://arxiv.org/abs/1606.04838
  - R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” NIPS 2013

Today's Lecture: Convergence Results of SGD

	Fixed step size $(\eta_t \equiv \eta)$	Diminishing step sizes $(\eta_t = \Theta(1/t))$
μ-strongly convex functions	$\mathbb{E}[F(x_t) - F_*] \leq \frac{\eta L \sigma^2}{2\mu} + (1-\eta\mu)^t \left(F(x_0) - F_*\right)$	$\mathbb{E}[F(x_t) - F_*] = O\left(\frac{1}{t}\right)$
General functions	$\mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} \|\|\nabla F(x_i)\|\|^2\right] \leq \eta L \sigma^2 + \frac{2\left(F(x_0) - F_*\right)}{\eta t}$	$\lim_{t\to\infty} \mathbb{E}\left[\frac{\sum_{i=1}^{t} \eta_i \|\|\nabla F(x_i)\|\|^2}{\sum_{i=1}^{t} \eta_i}\right] = 0$

- For general non-convex functions, SGD converges to a (nearly-)stationary point

Review: Stochastic Optimization in Machine Learning

- Optimization problems in ML arise mainly in two forms:

$$\nabla F(x) = \nabla \mathbb{E}[f(x;\varepsilon)] = \mathbb{E}[\nabla f(x;\varepsilon)]$$

1. Expected risk/loss minimization

$$F^* := \min_{x \in X} F(x), \text{where } F(x) := \mathbb{E}[f(x;\varepsilon)]$$

$\varepsilon \sim D$

where ε is the randomness (possibly unknown) in our problem

2. Empirical risk minimization (aka Finite-sum problem)

$$\nabla F(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(x;d_i)$$

$$F^* := \min_{x \in X} F(x), \text{where } F(x) := \frac{1}{n}\sum_{i=1}^{n} f(x;d_i)$$

where $\{d_i\}_{i=1}^n$ are n random data samples

Stochastic GD for Empirical Risk Minimization

- Idea: Use sampling to estimate expectation [Robbins & Monro, 1951]

At each iteration, we randomly pick an integer $i \in \{1,2,\cdots,n\}$

$$x_{k+1} = x_k - \eta_k \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla f(x_k; d_i) \quad \Rightarrow \quad x_{k+1} = x_k - \eta_k \cdot \nabla f(x_k; d_i)$$

(GD) (SGD)

- The update requires only gradient of one data sample d_i
- $\nabla f(x_k; d_i)$ is an unbiased estimate of $\nabla f(x_k)$, i.e., $\mathbb{E}[\nabla f(x_k; d_i)] = \nabla f(x_k)$

H. Robbins, S. Monro, “A stochastic approximation method,” The Annals of Mathematical Statistics, 1951

Review: Stochastic GD for Expected Risk Minimization

- Idea: Use sampling to estimate expectation [Robbins & Monro, 1951]

$$x_{k+1} = x_k - \eta_k \cdot \mathbb{E}[\nabla f(x_k; \varepsilon_k)] \quad \Rightarrow \quad x_{k+1} = x_k - \eta_k \cdot g(x_k; \varepsilon_k)$$

(GD) (SGD)

- $g(x_k; \varepsilon_k)$ is an unbiased estimate of $\mathbb{E}[\nabla f(x; \varepsilon_k)]$

- Equivalently: $g(x_k; \varepsilon_k) = \mathbb{E}[\nabla f(x_k; \varepsilon_k)] + \underbrace{\varepsilon}$

zero mean noise

- $g(x_k; \varepsilon_k)$ is constructed using one or multiple samples (= mini-batch)

H. Robbins, S. Monro, “A stochastic approximation method,” The Annals of Mathematical Statistics, 1951

Some Mild Technical Assumptions

- Quick recap:

(Objective) $F^* := \min_{x \in X} F(x)$ (where $F(x) = \mathbb{E}[f(x;\varepsilon)]$ or $F(x) := \frac{1}{n}\sum_{i=1}^{n} f(x;d_i)$)

(SGD) $x_{k+1} = x_k - \eta_k \cdot \underbrace{g(x_k;\varepsilon_k)}_{\text{unbiased estimate}}$

- To prove convergence of SGD, consider the following mild technical assumptions:

(A1) $F(x)$ is bounded below, i.e. $F^* > -\infty$

(A2) $F(x)$ is L-smooth

(A3) $g(x_k;\varepsilon_k)$ has bounded variance, i.e.

$$\mathbb{V}\left[g(x_k;\varepsilon_k)\,|\,x_k\right] := \mathbb{E}[\|g(x_k;\varepsilon_k)\|^2|x_k] - \|\mathbb{E}[g(x_k;\varepsilon_k)|x_k]^2\| \leq M + M_V\|\nabla F(x_k)\|^2$$

stochastic gradient

Review: Three Useful Lemmas

Lemma 1: PL conditions for μ-strongly convex functions

$$F(x) - F^* \leq \frac{1}{2\mu}\|\nabla F(x)\|^2$$

Lemma 2: First stochastic descent lemma of SGD

$$\mathbb{E}[F(x_{k+1})|x_k] - F(x_k) \leq -\eta_k \nabla F(x_k)^\top \mathbb{E}[g(x_k;\varepsilon_k)|x_k] + \frac{1}{2}\eta_k^2 L \cdot \mathbb{E}[\|g(x_k;\varepsilon_k)\|^2|x_k]$$

expected improvement

attained improvement by the 1st-order term

effect of 2nd-order term.

If $g(x_k;\varepsilon_k)$ is biased, then this term becomes $-\eta_k \|\nabla F(x_k)\|^2$

If there is stochasticity, then this term becomes $\|\nabla F(x_k)\|^2$

Lemma 3: Second stochastic descent lemma of SGD

$$\mathbb{E}[F(x_{k+1})|x_k] - F(x_k) \leq -\left(1 - \frac{1}{2}\eta_k L(M_v+1)\right)\eta_k\|\nabla F(x_k)\|^2 + \frac{1}{2}\eta_k^2 LM$$

Choose η_k such that RHS is negative

Convergence of SGD: Strong Convexity + Fixed Step Size (SC+F)

We proved the following result in Lecture 6:

Theorem (Convergence of SGD under SC+F):

Suppose the following conditions hold.

(1) Assumptions (A1)-(A3)

(2) $F(x)$ is L-smooth and μ-strongly convex

(3) Step sizes $\eta_k \equiv \eta$ satisfy that $\eta < \dfrac{\mu}{L(M_v + 1)}$

Then, SGD achieves

$$\mathbb{E}[F(x_k)] - F^* \leq \frac{\eta L M}{2\mu} + (1 - \eta\mu)^k \left(F(x_0) - F^* - \frac{\eta L M}{2\mu}\right)$$

A Closer Look at Theorem (SC+F)

$$\mathbb{E}[F(x_k) - F_*] \leq \frac{\eta LM}{2\mu} + (1 - \eta\mu)^k \left(F(x_0) - F_*\right)$$

- What if there is no noise in the estimated gradients?

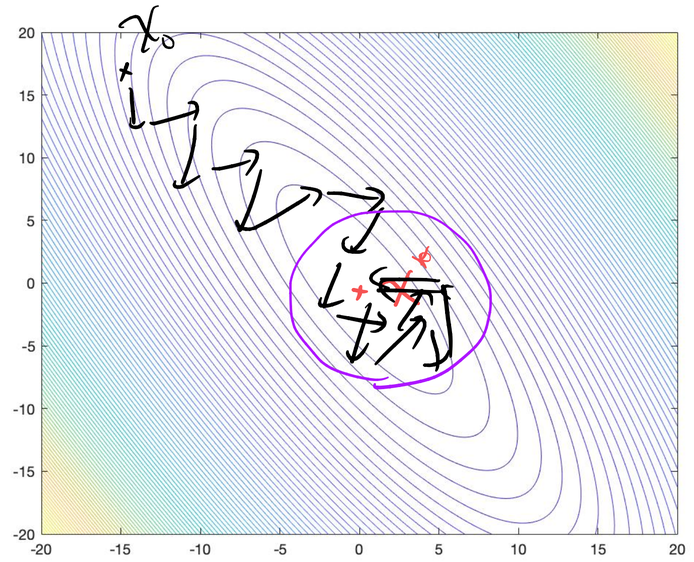
M=0, we have linear convergence.

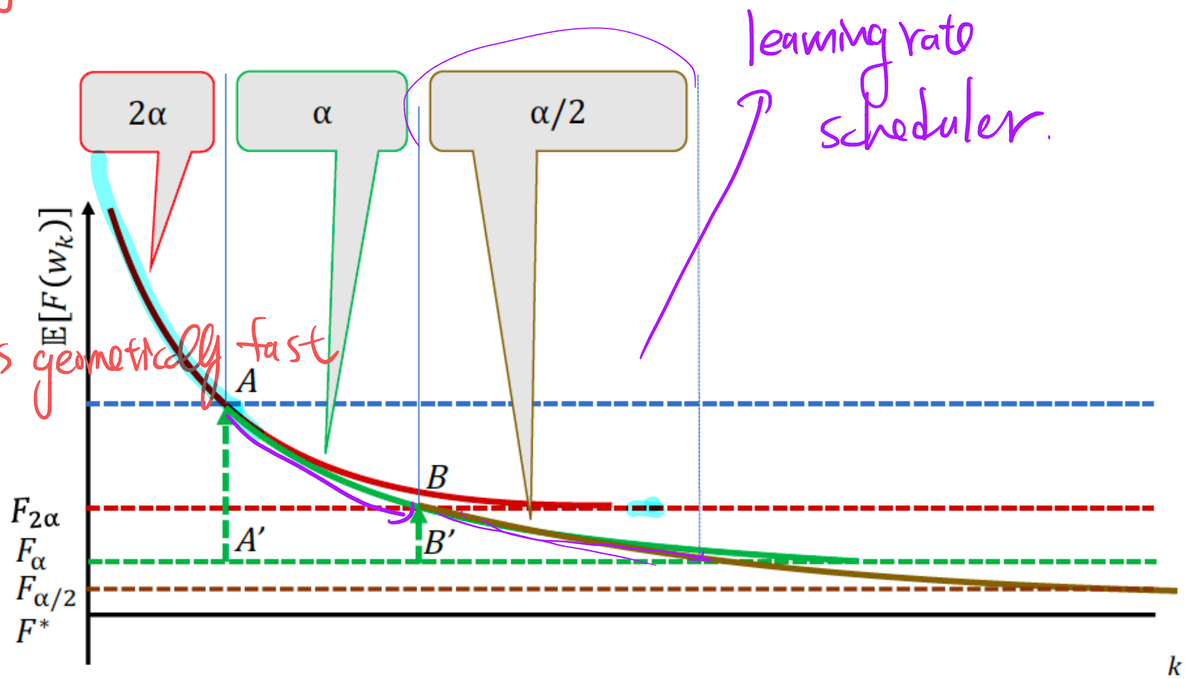
- Why do we need "strong convexity"?

- What's the role of initial condition?

x_0 only takes part in $(1-\eta\mu)^k(F(x_0)-F^*)$, which decreases geometrically fast

- A practical strategy of choosing η in SGD:
  - Run SGD with fixed step sizes
  - Whenever the progress stalls, we reduce the step sizes and proceed with SGD





(Figure Source: [Bottou, Curtis, Nocedal, 2018])

Convergence of SGD: Strong Convexity + Diminishing Step Size (SC+D)

Theorem (Convergence of SGD under SC+D):

Suppose the following conditions hold.

(1) Assumptions (A1)-(A3)

(2) $F(x)$ is L-smooth and μ-strongly convex

(3) Step sizes $\eta_k = \dfrac{\beta}{k+\gamma}$, where $\beta > \dfrac{1}{\mu}$ and $\gamma > 0$ such that $\eta_1 \leq \dfrac{1}{L(M_v+1)}$

Then, SGD achieves $\mathbb{E}[F(x_k)] - F^* \leq \dfrac{\nu}{k+\gamma} = O\left(\dfrac{1}{k}\right)$

where $\nu := \max\left\{\dfrac{\beta^2 LM}{2(\beta\mu - 1)}, (\gamma+1)(F(x_0) - F^*)\right\}$

$\eta_1 \cdot L(M_v+1) \leq 1$

$\Rightarrow \eta_k \cdot L(M_v+1) \leq 1$ for all k

$O\left(\frac{1}{k}\right)$

$\frac{\nu}{k+\gamma} = \varepsilon$

$\Rightarrow k = \frac{\nu - \varepsilon\gamma}{\varepsilon}$

Proof of Theorem (SC+D)

$$E[E[X|Y]] = E[X]$$

Step 1: By the 2nd stochastic descent lemma,

$$E\Big[E[F(x_{k+1})|x_k] - F(x_k)\Big] \le -\left(1 - \frac{1}{2}\eta_k L (M_v+1)\right)\eta_k \|\nabla F(x_k)\|^2 + \frac{1}{2}\eta_k^2 L M$$

$\eta_k L (M_v+1) \le 1$

Law of iterated expectation:

$$\le E\Big[-\frac{1}{2}\eta_k \|\nabla F(x_k)\|^2 + \frac{1}{2}\eta_k^2 L M\Big]$$

$$\le \left(E[F(x_k)] - F^*\right)\cdot 2\mu\cdot\frac{1}{2}\eta_k$$

Step 2: By taking total expectation and subtracting F^*,

$$E[F(x_{k+1}) - F^*] \le (1 - \eta_k \mu)\, E[F(x_k) - F^*] + \frac{1}{2}\eta_k^2 L M \quad -(*)$$

Step 3: We prove $E[F(x_k) - F^*] \le \frac{\nu}{k+\gamma}$ by induction

(Cont.). Want: $\underbrace{E[F(x_k)-F^*] \le \frac{\nu}{k+\gamma}}_{(**)}$, where

$$\nu := \max\left\{\frac{\beta^2 LM}{2(\beta\mu-1)},\ (\gamma+1)\cdot(F(x_0)-F^*)\right\}$$

For $k=1$: By (*),

$$E[F(x_1)-F^*] \le (1-\eta_1\mu)\cdot(F(x_0)-F^*) + \frac{1}{2}\eta_1^2 L\cdot M$$

Suppose (**) holds under k and consider $k+1$:

$$E[F(x_{k+1})-F^*] \le \left(1-\frac{\beta}{\gamma+k}\mu\right)\cdot\frac{\nu}{\gamma+k} + \frac{\beta^2\cdot L\cdot M}{2(\gamma+k)^2}$$

(η_k)

$$= \left(\frac{(\gamma+k)-1}{(\gamma+k)^2}\right)\nu - \underbrace{\left(\frac{\beta\cdot\mu-1}{(\gamma+k)^2}\right)\nu}_{\text{non-positive}} + \frac{\beta^2 L\cdot M}{2(\gamma+k)^2} \le \frac{\nu}{(\gamma+k)+1}$$

since $(\gamma+k)^2 \ge (\gamma+k+1)(\gamma+k-1)$

A Closer Look at Theorem (SC+D)

$$\mathbb{E}[F(x_k) - F_*] \leq \frac{\nu}{k+\gamma} \qquad \nu = \max\left\{\frac{\beta^2 LM}{2(\beta\mu - 1)}, (\gamma+1)(F(x_0) - F(x^*))\right\}$$

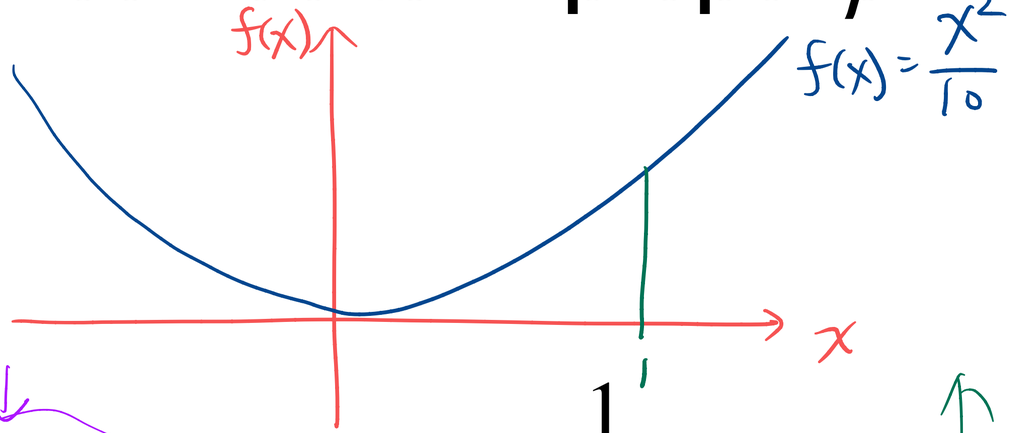
- Why do we need "strong convexity"? Connecting $\|\nabla F(x_k)\|^2$ with sub-optimality gap $(F(x_k) - F^*)$
- Why do we need $\beta > \frac{1}{\mu}$? *For sufficiently fast convergence!*
  - A toy example in [Nemirovski et al., 2009]

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming, SIAM Journal on Optimization, 2009

A Toy Example: Slow Convergence Due to Improper β

- Consider $f(x) = x^2/10$ with $x_0 = 1$
  - What is μ in this example?

$f(x)$ is $\frac{1}{5}$-strongly convex



Suppose we take the step sizes $\eta_k = 1/k$ (i.e., $\beta = 1 < \frac{1}{\mu} = \frac{10}{5}$)

$$x_{k+1} = x_k - \frac{1}{k} f'(x_k) = (1 - \frac{1}{5k}) x_k$$

$$x_k = \prod_{n=0}^{k-1} (1 - \frac{1}{5n}) x_0 > 0.8 \cdot k^{-\frac{1}{5}}$$

$f'(x_k) = \frac{x_k}{5}$

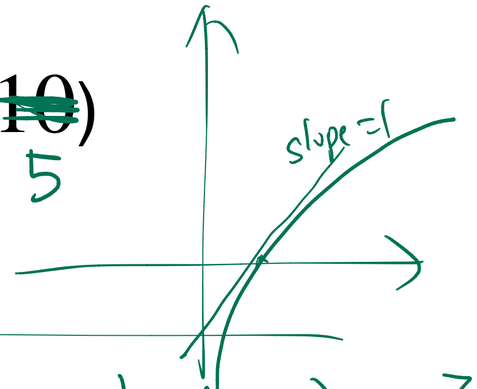
$f(x_k) > \frac{(0.8 \cdot k^{-\frac{1}{5}})^2}{10}$ $\left(1 - \frac{1}{5}\right)\left(1 - \frac{1}{10}\right)\left(1 - \frac{1}{15}\right)(\cdots$

$\log\left(\prod (1 - \frac{1}{5n})\right)$

$= \sum_{n=1}^{k} \log(1 - \frac{1}{5n}) \approx \sum_{n=1}^{k} -\frac{1}{5n} \approx -\frac{1}{5} \log k$

$= \log k^{-\frac{1}{5}}$

$\Rightarrow \prod (1 - \frac{1}{5n}) \approx k^{-\frac{1}{5}}$



$\log(1 - z) \approx -z$

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming, SIAM Journal on Optimization, 2009

Next Question: Is $O(\frac{1}{k})$ optimal?

- (Informal) For the minimization of strongly convex functions, there is no algorithm that can achieve an accuracy better than $O(1/k)$ on performing k queries

Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization

Alekh Agarwal[1] alekh@cs.berkeley.edu

Peter L. Bartlett[1,2,3] peter@berkeley.edu

Pradeep Ravikumar[4] pradeepr@cs.utexas.edu

Martin J. Wainwright[1,2] wainwrig@stat.berkeley.edu

Department of Electrical Engineering and Computer Sciences[1]
Department of Statistics[2]
UC Berkeley, Berkeley, CA

Mathematical Sciences[3]
QUT, Brisbane, Australia

Department of Computer Sciences[4]
UT Austin, Austin, TX

November 22, 2011

Abstract

Relative to the large literature on upper bounds on complexity of convex optimization, lesser attention has been paid to the fundamental hardness of these problems. Given the extensive use of convex optimization in machine learning and statistics, gaining an understanding of these complexity-theoretic issues is important. In this paper, we study the complexity of stochastic convex optimization in an oracle model of computation. We improve upon known results and obtain tight minimax complexity estimates for various function classes.

- Therefore, SGD with step sizes $\Theta(1/k)$ is optimal

Alekh Agarwal @ Google

A. Nemirovski, D. Yudin, “Problem complexity and method efficiency in optimization,” Wiley, 1983

A. Agarwal, P. Barlett, P Ravikumar, M. Wainright, “Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization,” IEEE Transactions on Information Theory, 2011

Comparison: SGD and Batch GD

- Consider strongly convex "empirical risk minimization" with n samples
- To achieve an accuracy of ϵ, we need: ($\kappa := L/\mu$ is the condition number)

	Iteration Complexity	Per-iteration Cost	Total Computation Cost
Batch GD	$\kappa \log \frac{1}{\epsilon}$	n	$n\kappa \log \frac{1}{\epsilon}$
SGD	$\kappa^2 \frac{1}{\epsilon}$	1	$\kappa^2 \frac{1}{\epsilon}$

SGD has an advantage for large n and moderate ϵ ("big data" regime!)

That is, $\frac{1}{\epsilon} < n \log \frac{1}{\epsilon}$

SGD for Non-Convex and Smooth Functions

Convergence of SGD: Non-Convex Functions + Fixed Step Size (NC+F)

Theorem (Convergence of SGD under NC+F):

Suppose the following conditions hold.

(1) Assumptions (A1)-(A3)

(2) $F(x)$ is L-smooth (but not necessarily convex)

(3) Step sizes $\eta_k \equiv \eta$ satisfy that $\eta < \dfrac{\mu}{L(M_v + 1)}$

Then, SGD achieves $\dfrac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1} \|\nabla F(x_k)\|^2\right] \leq \eta L M + \dfrac{2(F(x_0) - F^*)}{K\eta}$

Proof:

Step1: By the 2nd stochastic descent lemma and taking "total expectation",

$$\mathbb{E}[F(x_{k+1})] - \mathbb{E}[F(x_k)]$$

$$\leq -\left(1 - \frac{1}{2}\eta \cdot L \cdot (1 + M_v)\right) \cdot \eta \cdot \mathbb{E}\left[\|\nabla F(x_k)\|^2\right] + \frac{1}{2}\eta^2 \cdot L \cdot M$$

$$\leq -\frac{1}{2}\eta \cdot \mathbb{E}\left[\|\nabla F(x_k)\|^2\right] + \frac{1}{2}\eta^2 \cdot L \cdot M$$

$$\mathbb{E}[F(x_{k+1})] - \mathbb{E}[F(x_k)] \leq \cdots$$

$$\mathbb{E}[F(x_k)] - \mathbb{E}[F(x_{k-1})] \leq \cdots$$

$$\vdots$$

$$+) \quad \mathbb{E}[F(x_1)] - \mathbb{E}[F(x_0)] \leq \cdots$$

Step2: By taking the "telescoping sum", we have

$$\mathbb{E}[F(x_{K+1})] - F(x_0) \leq -\frac{1}{2}\eta \cdot \sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla F(x_k)\|^2\right] + \frac{1}{2}K \cdot \eta^2 \cdot L \cdot M$$

Convergence of SGD: Non-Convex Functions + Diminishing Step Size (NC+D)

Theorem (Convergence of SGD under NC+D):

Suppose the following conditions hold.

(1) Assumptions (A1)-(A3)

(2) $F(x)$ is L-smooth (but not necessarily convex)

(3) Step sizes η_k satisfy $\sum_{k=1}^{\infty} \eta_k = \infty$ and $\sum_{k=1}^{\infty} \eta_k^2 < \infty$

Then, SGD achieves $\lim_{K\to\infty} \mathbb{E}\left[\sum_{k=1}^{K} \eta_k \cdot \|\nabla F(x_k)\|^2\right] < \infty$

and therefore $\lim_{K\to\infty} \mathbb{E}\left[\frac{\sum_{k=1}^{K} \eta_k \|\nabla F(x_k)\|^2}{\sum_{k=1}^{K} \eta_k}\right] = 0$

Ensure that SGD can reach at least a stationary point.

total distance

2nd-order term does NOT explode

"square-summable"

non-summable

Proof: Recall from the 2nd descent lemma

Step 1:

$$E[F(x_{k+1})] - E[F(x_k)]$$

$$\leq -\left(1 - \tfrac{1}{2}\eta_k L \cdot (M_{v+1})\right)\eta_k \cdot E\left[\|\nabla F(x_k)\|^2\right] + \tfrac{1}{2}\eta_k^2 L M$$

$$\leq -\tfrac{1}{2}\eta_k \cdot E\left[\|\nabla F(x_k)\|^2\right] + \tfrac{1}{2}\eta_k^2 L M \quad \cdots (\ )$$

Step 2: By summing the above over $k = 1, \cdots, K$,

$$\underbrace{E[F(x_{K+1})] - E[F(x_1)]}_{\| \; F^* - E[F(x_1)]} \leq -\tfrac{1}{2}\sum_{k=1}^{K} \eta_k \cdot E\left[\|\nabla F(x_k)\|^2\right] + \tfrac{1}{2} L M \cdot \sum_{k=1}^{K} \eta_k^2$$

Recall: Convergence Rates of SGD

- We know the convergence rate of SGD for smooth strongly-convex functions

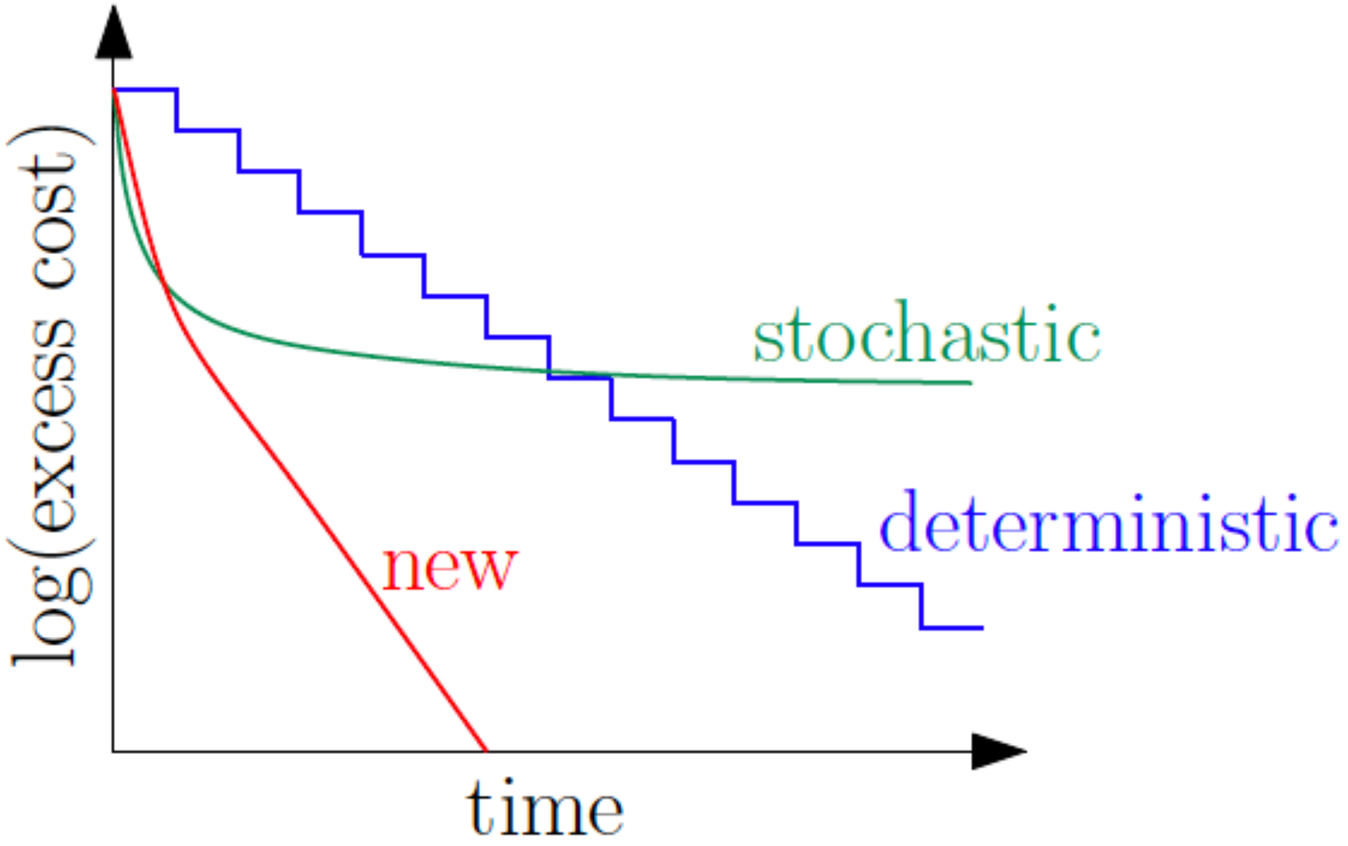
	Fixed step size $(\eta_t \equiv \eta)$	Diminishing step sizes $(\eta_t = \Theta(1/t))$
μ-strongly convex functions	$\mathbb{E}[F(x_t) - F_*] \leq \frac{\eta L \sigma^2}{2\mu} + (1 - \eta\mu)^t \left(F(x_0) - F_*\right)$	$\mathbb{E}[F(x_t) - F_*] = O\left(\frac{1}{t}\right)$

- Any issue with the use of a fixed step size? Wandering around a small neighborhood
- Any issue with the use of $\eta_t = \Theta(1/t)$? Slow convergence

Best of Both World?

Question: Can we have an algorithm that achieves "best of both world"

That is, “linear convergence rate” + “low per-iteration cost”



Next topic: Variance reduction

(Figure Credit: Suvrit Sra)

Variance Reduction for SGD: Finite-Sum Problems

Intuition Behind Variance Reduction

$I_k \sim \text{Unif}(1, n)$.

$\nabla f(x_k; d_{I_k})$.

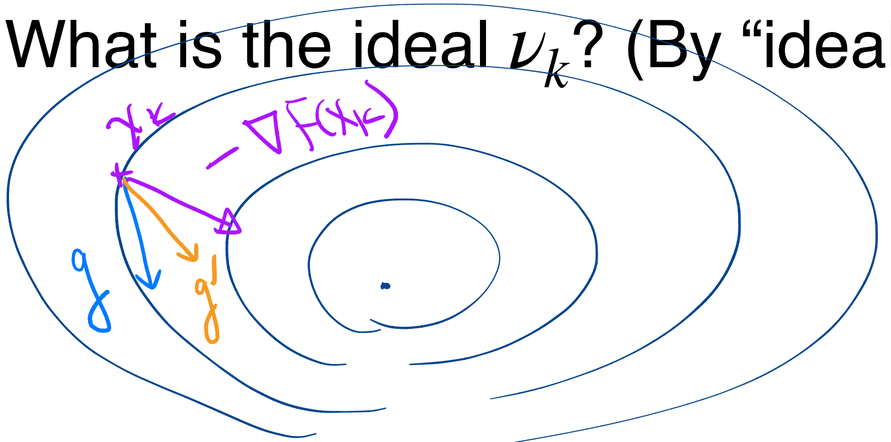
- In vanilla SGD, we use an unbiased estimator $g(x_k, \varepsilon_k)$
  - Question: Can we use $\tilde{g} := g(x_k, \varepsilon_k) + \nu_k$ with $\mathbb{E}[\nu_k] = 0$?

$\mathbb{E}[\tilde{g}] = \mathbb{E}[g(x_k; \varepsilon_k)] = \nabla F(x_k) \Rightarrow$ $\tilde{g}$ is an unbiased estimator ✓

  - Question: What kind of condition do we want about ν_k?

To use $\tilde{g}$ in SGD, we need ν_k to be of bounded variance.

  - Question: What is the ideal ν_k? (By "ideal", we mean zero variance)



Choose ν_k such that $\tilde{g} = \nabla F(x_k)$

That is, $\nu_k = \tilde{g} - g(x_k; \varepsilon_k) = \nabla F(x_k) - g(x_k; \varepsilon_k)$

A Intuitive Idea of Variance Reduction: Gradient Aggregation

- Example: SGD for empirical risk minimization

$$\min_{x\in X} F(x) := \frac{1}{n}\sum_{i=1}^{n} f(x; d_i) \qquad (\{d_i\}_{i=1}^n \text{ are data samples})$$

Vanilla SGD:

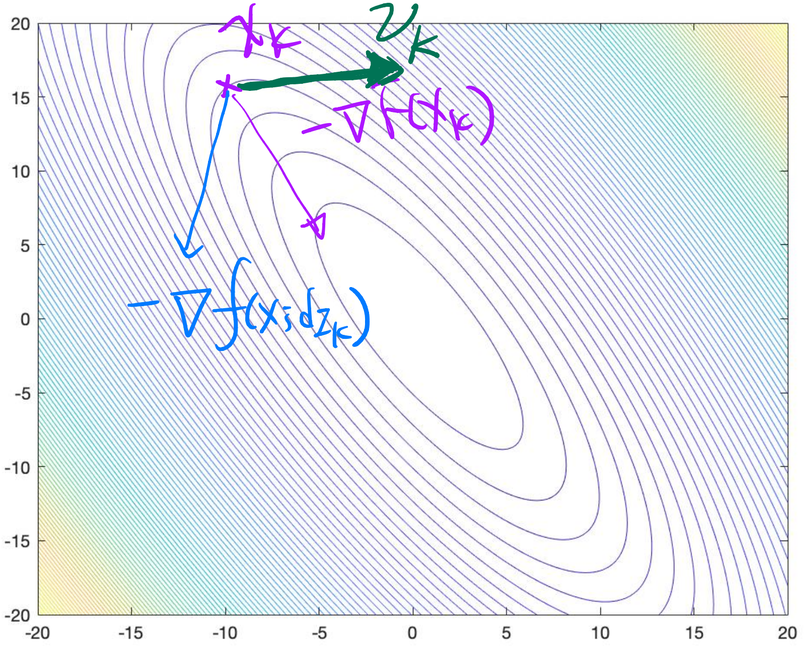
$$x_{k+1} = x_k - \eta \nabla f(x; d_{I_k}), \quad I_k \sim \text{Unif}(1,n)$$

stochastic gradient

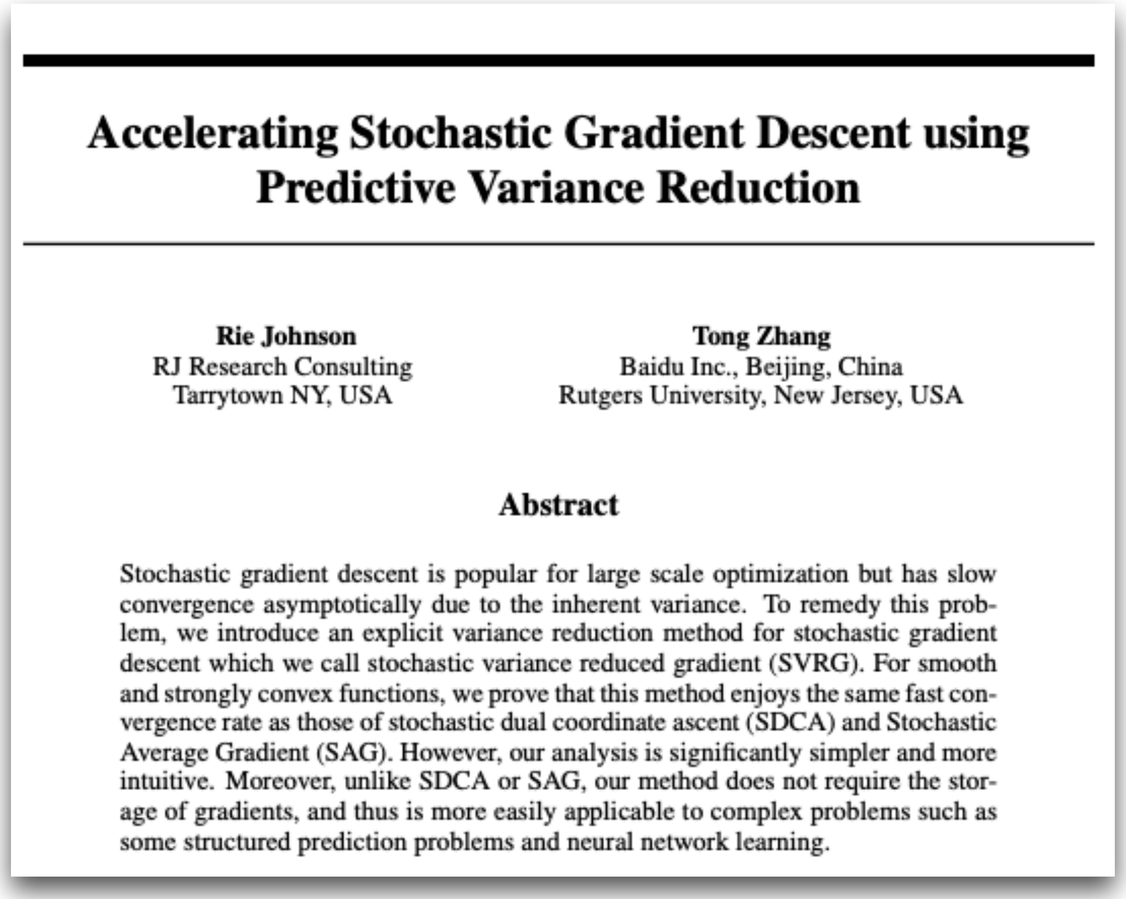
Gradient Aggregation:

$$x_{k+1} = x_k - \eta\left(\nabla f(x; d_{I_k}) - \nu_k\right), \quad I_k \sim \text{Unif}(1,n)$$

where ν_k and $\nabla f(x; d_{I_k})$ are positively correlated (why?)



Stochastic Variance-Reduced Gradient (SVRG)



Accelerating Stochastic Gradient Descent using Predictive Variance Reduction

Rie Johnson
RJ Research Consulting
Tarrytown NY, USA

Tong Zhang
Baidu Inc., Beijing, China
Rutgers University, New Jersey, USA

Abstract

Stochastic gradient descent is popular for large scale optimization but has slow convergence asymptotically due to the inherent variance. To remedy this problem, we introduce an explicit variance reduction method for stochastic gradient descent which we call stochastic variance reduced gradient (SVRG). For smooth and strongly convex functions, we prove that this method enjoys the same fast convergence rate as those of stochastic dual coordinate ascent (SDCA) and Stochastic Average Gradient (SAG). However, our analysis is significantly simpler and more intuitive. Moreover, unlike SDCA or SAG, our method does not require the storage of gradients, and thus is more easily applicable to complex problems such as some structured prediction problems and neural network learning.

- SVRG serves as a simple and intuitive way to achieve gradient aggregation
- This idea has been applied to various other problem settings, including constrained optimization problems and RL

R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” NIPS 2013

Variance-Reduced and Projection-Free Stochastic Optimization

Elad Hazan EHAZAN@CS.PRINCETON.EDU
Princeton University, Princeton, NJ 08540, USA

Haipeng Luo HAIPENGL@CS.PRINCETON.EDU
Princeton University, Princeton, NJ 08540, USA

Abstract

The Frank-Wolfe optimization algorithm has recently regained popularity for machine learning applications due to its projection-free property and its ability to handle structured constraints. However, in the stochastic learning setting, it is still relatively understudied compared to the gradient descent counterpart. In this work, leveraging a recent variance reduction technique, we propose two stochastic Frank-Wolfe variants which substantially improve previous results in terms of the number of stochastic gradient evaluations needed to achieve $1-\epsilon$ accuracy. For example, we improve from $\mathcal{O}(\frac{1}{\epsilon})$ to $\mathcal{O}(\ln\frac{1}{\epsilon})$ if the objective function is smooth and strongly convex, and from $\mathcal{O}(\frac{1}{\epsilon^2})$ to $\mathcal{O}(\frac{1}{\epsilon^{1.5}})$ if the objective function is smooth and Lipschitz. The theoretical improvement is also observed in experiments on real-world datasets for a multiclass classification application.

more (see for example (Hazan & Kale, 2012; Hazan et al., 2012; Jaggi, 2013; Dudik et al., 2012; Zhang et al., 2012; Harchaoui et al., 2015)).

The Frank-Wolfe algorithm (Frank & Wolfe, 1956) (also known as *conditional gradient*) and it variants are natural candidates for solving these problems, due to its projection-free property and its ability to handle structured constraints. However, despite gaining more popularity recently, its applicability and efficiency in the stochastic learning setting, where computing stochastic gradients is much faster than computing exact gradients, is still relatively understudied compared to variants of projected gradient descent methods.

In this work, we thus try to answer the following question: *what running time can a projection-free algorithm achieve in terms of the number of stochastic gradient evaluations and the number of linear optimizations needed to achieve a certain accuracy?* Utilizing Nesterov's acceleration technique (Nesterov, 1983) and the recent variance reduction idea (Johnson & Zhang, 2013; Mahdavi et al., 2013), we

[Hazan and Luo, ICML 2016]

Stochastic Variance-Reduced Policy Gradient

Matteo Papini [*1] **Damiano Binaghi** [*1] **Giuseppe Canonaco** [*1] **Matteo Pirotta** [2] **Marcello Restelli** [1]

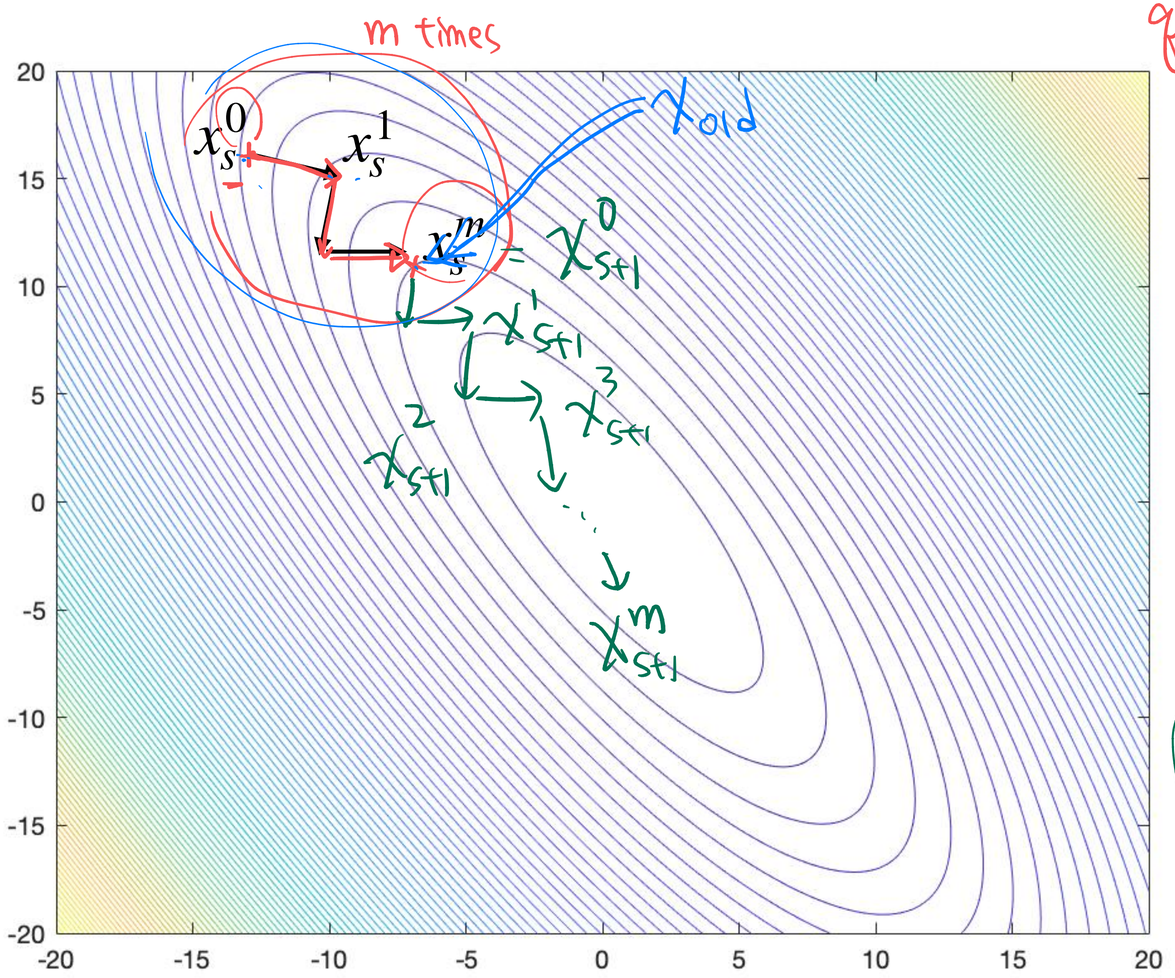
Abstract

In this paper, we propose a novel reinforcement-learning algorithm consisting in a stochastic variance-reduced version of policy gradient for solving Markov Decision Processes (MDPs). Stochastic variance-reduced gradient (SVRG) methods have proven to be very successful in supervised learning. However, their adaptation to policy gradient is not straightforward and needs to account for I) a non-concave objective function; II) approximations in the full gradient computation; and III) a non-stationary sampling process. The result is SVRPG, a stochastic variance-reduced policy gradient algorithm that leverages on importance weights to preserve the unbiasedness of the gradient estimate. Under standard assumptions on the MDP, we provide convergence guarantees for SVRPG with a convergence rate that is linear under increasing batch sizes. Finally, we suggest practical variants of SVRPG, and we empirically evaluate them on continuous MDPs.

a value function, or directly a policy defining the agent's behaviour. Furthermore, when the tasks are characterized by large or continuous state-action spaces, RL needs the powerful function approximators (e.g., neural networks) that are the main subject of study of SL. In a typical SL setting, a performance function $J(\boldsymbol{\theta})$ has to be optimized w.r.t. to model parameters $\boldsymbol{\theta}$. The set of data that are available for training is often a subset of all the cases of interest, which may even be infinite, leading to optimization of finite sums that approximate the expected performance over an unknown data distribution. When generalization to the complete dataset is not taken into consideration, we talk about Empirical Risk Minimization (ERM). Even in this case, stochastic optimization is often used for reasons of efficiency. The idea of stochastic gradient (SG) ascent (Nesterov, 2013) is to iteratively focus on a random subset of the available data to obtain an approximate improvement direction. At the level of the single iteration, this can be much less expensive than taking into account all the data. However, the sub-sampling of data is a source of variance that can potentially compromise convergence, so that per-iteration efficiency and convergence rate must be traded off

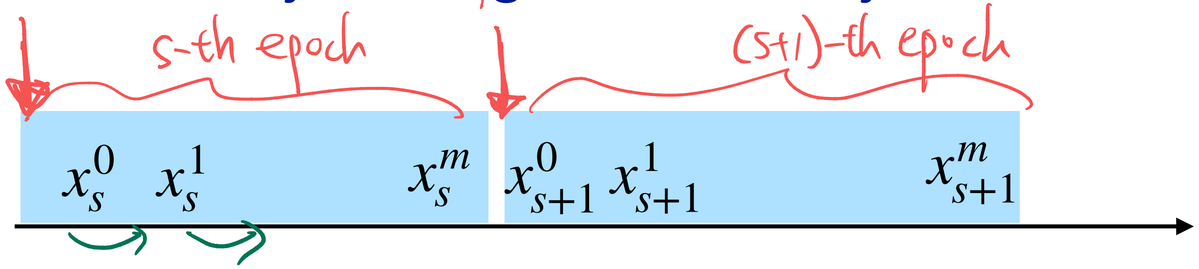
[Papini et al., ICML 2018]

Intuition Behind SVRG



query "true gradient" query "true gradient"

Idea: Query "true gradient" only occasionally



In each stochastic update, use

$$x_s^{k+1} = x_s^k - \eta\left(\nabla f(x_s^k; d_{I_k}) - \nu_k\right), \quad I_k \sim \text{Unif}(1,n)$$

$$\nu_k = \nabla f(x_{old}; d_{I_k}) - \nabla F(x_{old})$$

Pseudo Code of SVRG

Notation: $\nabla f(x; d_i) \equiv \nabla f_i(x)$

Outer loop for snapshots

1: **for** $s = 1, 2, \ldots$ **do**

2: $x_s^{\text{old}} \leftarrow x_s^j$ and compute $\underbrace{\nabla F(x_s^{\text{old}})}_{\text{batch gradient}}$ // update snapshot

($j \sim \text{Unif}(0, \cdots, m-1)$)

3: initialize $x_s^0 \leftarrow x_s^{\text{old}}$

4: **for** $\underbrace{t = 0, \ldots, m-1}_{\text{each epoch contains } m \text{ iterations}}$ **do**

Inner loop for iterate updates

5: choose i_t uniformly from $\{1, \ldots, n\}$, and

$$x_s^{t+1} = x_s^t - \eta\{\underbrace{\nabla f_{i_t}(x_s^t) - \nabla f_{i_t}(x_s^{\text{old}})}_{\text{stochastic gradient}} + \nabla F(x_s^{\text{old}})\}$$

- Question: How many gradient computations are needed under SVRG in one epoch? How about SGD? Per-outer-loop complexity: $(2m + n)$ gradient computations

(Slide Credit: Yuxin Chen)

Comparison: SGD / Batch GD / SVRG

- Consider strongly convex "empirical risk minimization" with n samples

	Iteration Complexity	Per-iteration Cost	Total Computation Cost
Batch GD	$\kappa \log \frac{1}{\epsilon}$	n	$n\kappa \log \frac{1}{\epsilon}$
SGD	$\kappa^2 \frac{1}{\epsilon}$	1	$\kappa^2 \frac{1}{\epsilon}$
SVRG			$(n+\kappa)\log \frac{1}{\epsilon}$

($\kappa := L/\mu$ is the condition number)

Convergence Rate of SVRG

Theorem Suppose the following conditions hold:

(1) All $f_i(x)$ are convex and L-smooth

(2) $F(x)$ is μ-strongly convex

(3). The snapshot period m is sufficiently large such that

$$\alpha = \frac{1}{\mu \cdot \eta \cdot (1-2L\eta) \cdot m} + \frac{2L\eta}{1-2L\eta} < 1$$

Then, we have $E[F(x) - F(x^*)] \leq \alpha^s \cdot (F(x_0) - F(x^*)) = \varepsilon$

A Lemma for Quantifying Suboptimality $F(x) - F(x^*)$

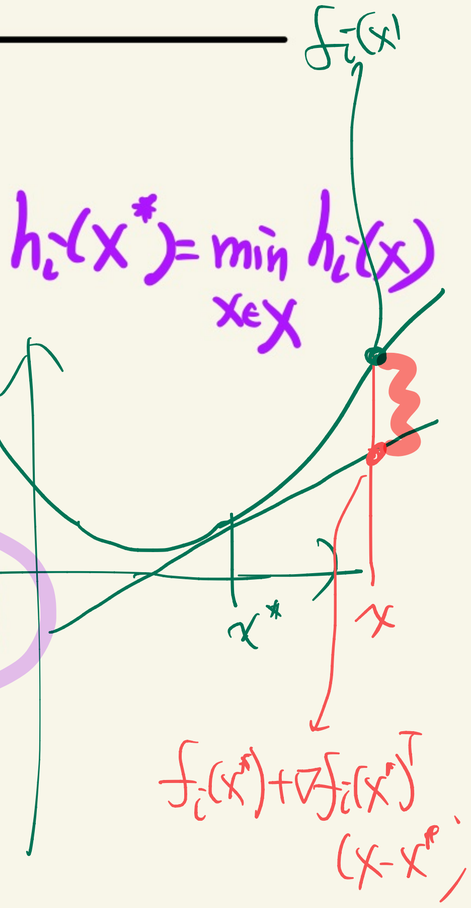
Lemma: $\frac{1}{n}\sum_{i=1}^{n} \| \nabla f_i(x) - \nabla f_i(x^*) \|^2 \leq 2L \cdot (F(x) - F(x^*))$

Proof of Lemma Let us construct a helper function $h_i(x)$ as

Step1: $h_i(x) = f_i(x) - \left(f_i(x^*) + \nabla f_i(x^*)^\top (x - x^*) \right)$ ← Hence, $h_i(x^*) = \min_{x \in X} h_i(x)$

Step2:
$$0 = h_i(x^*) \leq \min_{\eta} \left(h_i(x - \eta \cdot \nabla h_i(x)) \right)$$
$$\leq \min_{\eta} \left(h_i(x) - \eta \cdot \| \nabla h_i(x) \|^2 + \frac{1}{2} L \cdot \eta^2 \cdot \| \nabla h_i(x) \|^2 \right)$$
$$= h_i(x) - \frac{1}{2L} \| \nabla h_i(x) \|^2$$

Smoothness



(Cont.).

Step3: $0 \le h_i(x) - \frac{1}{2L}\|\nabla h_i(x)\|^2$ is equivalent to

$$\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \le 2L\left(f_i(x) - f_i(x^*) - \nabla f_i(x^*)^T(x - x^*)\right) \quad \text{(*)}$$

Step4: By taking the summation of (*) over $i = 1, \cdots, n$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \le 2L\cdot\left(F(x) - F(x^*)\right)$$

A Lemma for Quantifying the Variance of SVRG

Lemma: $$\mathbb{E}\left[\|g_s^t\|^2\right] \le 4L\left(F(x_s^{t-1}) - F(x^*) + F(x_s^{old}) - F(x^*)\right)$$

Question: If $x_s^t \approx x_s^{old} \approx x^*$, then what can we say about $\mathbb{E}\left[\|g_s^t\|^2\right]$?

Proof of Lemma

Recall that $g_s^t = \nabla f_{i_t}(x_s^{t-1}) - \nabla f_{i_t}(x_s^{old}) + \nabla F(x_s^{old})$

$$E[\|g_s^t\|^2] \le 2 \cdot E\left[\|\nabla f_{i_t}(x_s^{t-1}) - \nabla f_{i_t}(x^*)\|^2\right]$$

$$+ 2 \cdot E\left[\left\|\left(\nabla f_{i_t}(x_s^{old}) - \nabla f_{i_t}(x^*)\right) - \underbrace{\nabla F(x_s^{old})}_{=}\right\|^2\right]$$

$$E\left[\nabla f_{i_t}(x_s^{old}) - \nabla f_{i_t}(x^*)\right]$$

$$\le 2 \cdot E\left[\|\nabla f_{i_t}(x_s^{t-1}) - \nabla f_{i_t}(x^*)\|^2\right] + 2 \cdot E\left[\|\nabla f_{i_t}(x_s^{old}) - \nabla f_{i_t}(x^*)\|^2\right] \text{ (why?)}$$

$$\le 4L \cdot \left(\left(F(x_s^{t-1}) - F(x^*)\right) + \left(F(x_s^{old}) - F(x^*)\right)\right) \text{ (why?)}$$

A Useful Property

For any random vector Z, we always have

$$E\left[\|Z - E[Z]\|^2\right] = E\left[\|Z\|^2\right] - \|E[Z]\|^2 \leq E\left[\|Z\|^2\right].$$

Let’s put everything together and prove the convergence of SVRG

Proof of Convergence of SVRG

Step 1:

$$\mathbb{E}\left[\|x_s^t - x^*\|^2 \,\middle|\, x_s^{t-1}\right] = \mathbb{E}\left[\|x_s^{t-1} - \eta \cdot g_s^t - x^*\|^2 \,\middle|\, x_s^{t-1}\right]$$

$$= \|x_s^{t-1} - x^*\|^2 - 2\cdot\eta\cdot(x_s^{t-1} - x^*)^T \cdot \mathbb{E}\left[g_s^t \,\middle|\, x_s^{t-1}\right] + \eta^2 \cdot \mathbb{E}\left[\|g_s^t\|^2 \,\middle|\, x_s^{t-1}\right]$$

$$\leq \|x_s^{t-1} - x^*\|^2 - 2\cdot\eta\cdot\underbrace{(x_s^{t-1} - x^*)\cdot\nabla F(x_s^{t-1})}_{\geq F(x_s^{t-1}) - F(x^*)} + \eta^2 \cdot 4L\Big(F(x_s^{t-1}) - F(x^*) + F(x_s^{old}) - F(x^*)\Big)$$

$$\leq \|x_s^{t-1} - x^*\|^2 - 2\cdot\eta\cdot\Big(F(x_s^{t-1}) - F(x^*)\Big) + \eta^2 4L\Big(F(x_s^{t-1}) - F(x^*) + F(x_s^{old}) - F(x^*)\Big)$$

$$= \|x_s^{t-1} - x^*\|^2 - 2\cdot\eta\cdot(1 - 2L\eta)\cdot\Big(F(x_s^{t-1}) - F(x^*)\Big) + 4\eta^2 L\cdot\Big(F(x_s^{old}) - F(x^*)\Big) \quad -(*)$$

(Cont.)

Step 2: Consider a fixed epoch s

By taking the sum of (*) over $t=1,\cdots,m$ and taking the total expectation,

$$\mathbb{E}\left[\|X_s^m - x^*\|^2\right] + 2\eta(1-2L\eta)m \cdot \mathbb{E}\left[F(x_{s+1}^{old}) - F(x^*)\right]$$

$$\leq \mathbb{E}\left[\|X_s^0 - x^*\|^2\right] + 4Lm\eta^2 \cdot \mathbb{E}\left[F(x_s^{old}) - F(x^*)\right] \cdots (\quad)$$

$$= \mathbb{E}\left[\|X_s^{old} - x^*\|^2\right] + 4Lm\eta^2 \cdot \mathbb{E}\left[F(x_s^{old}) - F(x^*)\right] \cdots (\quad)$$

$$\leq \frac{2}{\mu}\mathbb{E}\left[F(x_s^{old}) - F(x^*)\right] + 4Lm\eta^2 \cdot \mathbb{E}\left[F(x_s^{old}) - F(x^*)\right] \cdots (\quad)$$

$$= \left(\frac{2}{\mu} + 4Lm\eta^2\right) \cdot \mathbb{E}\left[F(x_s^{old}) - F(x^*)\right]$$

(Cont.).

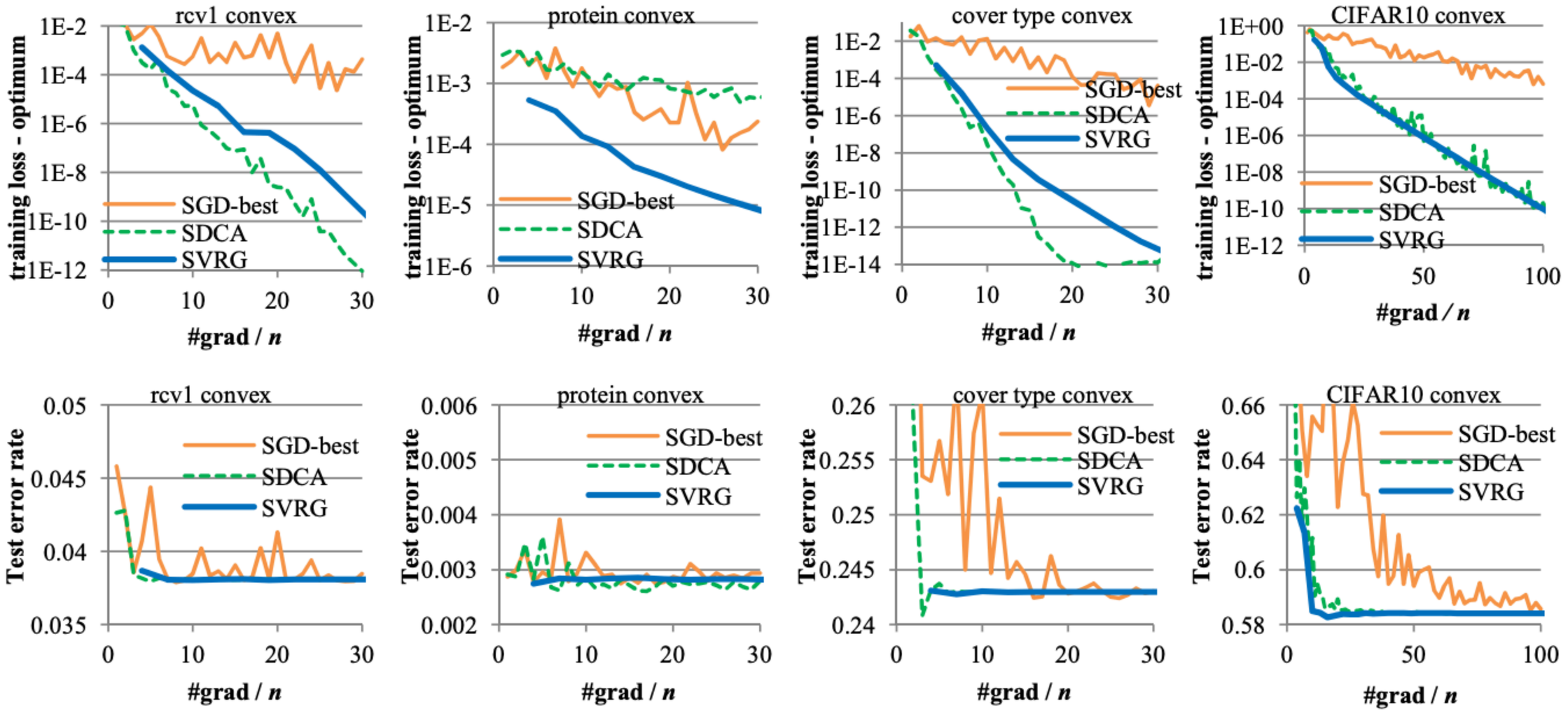
Step3: Consequently, we have

$$\mathbb{E}\left[F(x_{s+1}^{old}) - F(x^*)\right]$$

$$\leq \frac{\frac{2}{\mu} + 4Lm\eta^2}{2\eta(1-2L\eta)m} \cdot \mathbb{E}\left[F(x_s^{old}) - F(x^*)\right] - \underbrace{\frac{1}{2\eta(1-2L\eta)m}\mathbb{E}\left[\|x_s^m - x^*\|^2\right]}_{\geq 0}$$

$$\leq \left(\frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta}\right) \cdot \mathbb{E}\left[F(x_s^{old}) - F(x^*)\right]$$

Numerical Comparison: SGD and SVRG



R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction, NIPS 2013

A General Recipe of Variance Reduction

- Suppose we'd like estimate $E[X]$ (where X is a random variable)
- A general VR approach: If we could compute efficiently $E[Y]$ for another random variable Y that is highly correlated with X, then we could use an estimator as

$$\theta_\alpha = \alpha(X - Y) + E[Y]$$

- Question: Expected value and variance of θ_α?

$$\mathbb{E}[\theta_\alpha] = \alpha\mathbb{E}[X] + (1 - \alpha)\mathbb{E}[Y]$$

$$\mathrm{Var}(\theta_\alpha) = \alpha^2(\mathrm{Var}(X) + \mathrm{Var}(X) - 2\mathrm{Cov}(X, Y))$$

- By varying α from 0 to 1, what do we have regarding variance and bias?
- How to interpret SVRG?

A General Recipe of Variance Reduction: SAGA vs SVRG

SVRG $\quad x_s^{k+1} = x_s^k - \eta\Big(\nabla f(x_s^k; d_{I_k}) - \nabla f(x_{old}; d_{I_k}) - \nabla F(x_{old})\Big), \quad I_k \sim \text{Unif}(1,n)$

SAGA $\quad x^{k+1} = x^k - \eta\Big(\nabla f(x^k; d_{I_k}) - \nabla f(\phi_{I_k}^k; d_{I_k}) - \frac{1}{n}\sum_{i=1}^{n} \nabla f(\phi_i^k; d_i)\Big), \quad I_k \sim \text{Unif}(1,n)$

where $\nabla f(\phi_i^k; d_i)$ are past stochastic gradients stored in the buffer

---

General recipe: $\theta_\alpha = \alpha(X - Y) + E[Y]$. What are X, Y, α in. SAGA?

A. Defazio, F. Bach, and Simon Lacoste-Julien, “SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives,” NIPS 2017

Many More VR Methods

- SAG (**S**tochastic **A**veraged **G**radient)
- SAGA [Defazio, Bach, and Lacoste-Julien, NIPS 2014]
- SARAH (**S**toch**A**stic **R**ecursive gr**A**dient algorit**H**m) [Nguyen, Liu, Scheinberg, Takac, ICML 2017]
- SDCA (**S**tochastic **D**ual **C**oordinate **A**scent) [Johnson and Zhang, NIPS 2013; Shalev-Shwartz, ICML 2016]
- And more!

More VR Methods: SARAH

Question: Can we do VR without taking a snapshot?

- **S**toch**A**stic **R**ecursive gr**A**dient algorit**H**m (SARAH)

$$x^{t+1} = x^t - \eta g^t$$

$$g^t = \left(\nabla f_{I_t}(x^t) - \nabla f_{I_t}(x^{t-1}\right) + g^{t-1}$$

Idea: Recursive updates of gradient estimates

- Question: Why is this reasonable? Any issue?

L. Nguyen, J. Liu, K. Scheinberg, M. Takac, “SARAH: A novel method for machine learning problems using stochastic recursive gradient,” ICML 2017

Pseudo Code of SARAH

Notation: $\nabla f(x; d_i) \equiv \nabla f_i(x)$

1: **for** $s = 1, 2, \ldots, S$ **do** — Outer loop for periodic resets

2: $\quad \boldsymbol{x}_s^0 \leftarrow \boldsymbol{x}_{s-1}^{m+1}$, and compute $\underbrace{\boldsymbol{g}_s^0 = \nabla F(\boldsymbol{x}_s^0)}_{\text{batch gradient}}$ // restart $\boldsymbol{g}$ anew

3: $\quad \boldsymbol{x}_s^1 = \boldsymbol{x}_s^0 - \eta \boldsymbol{g}_s^0$

4: $\quad$ **for** $t = 1, \ldots, m$ **do** — Inner loop for recursive updates

5: $\quad\quad$ choose i_t uniformly from $\{1, \ldots, n\}$

6: $\quad\quad \boldsymbol{g}_s^t = \underbrace{\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{t-1})}_{\text{stochastic gradient}} + \boldsymbol{g}_s^{t-1}$

7: $\quad\quad \boldsymbol{x}_s^{t+1} = \boldsymbol{x}_s^t - \eta \boldsymbol{g}_s^t$

- Question: Why do we need to reset g^t periodically?

(Slide Credit: Yuxin Chen)

Next Lecture: How to accelerate SGD for expected loss?