Optimization for Machine Learning

Lecture 9: Faster stochastic methods

6.881: MIT

Suvrit Sra
Massachusetts Institute of Technology

18 Mar, 2021

Stochastic gradient complexity

Method	Assumptions	Batch	Stochastic
Subgradient	convex	$O(1/\sqrt{k})$	$O(1/\sqrt{k})$
Subgradient	strongly cvx	$O(1/k)$	$O(1/k)$

So using stochastic subgradient, solve n times faster.

Stochastic gradient complexity

Method	Assumptions	Batch	Stochastic
Subgradient	convex	$O(1/\sqrt{k})$	$O(1/\sqrt{k})$
Subgradient	strongly cvx	$O(1/k)$	$O(1/k)$

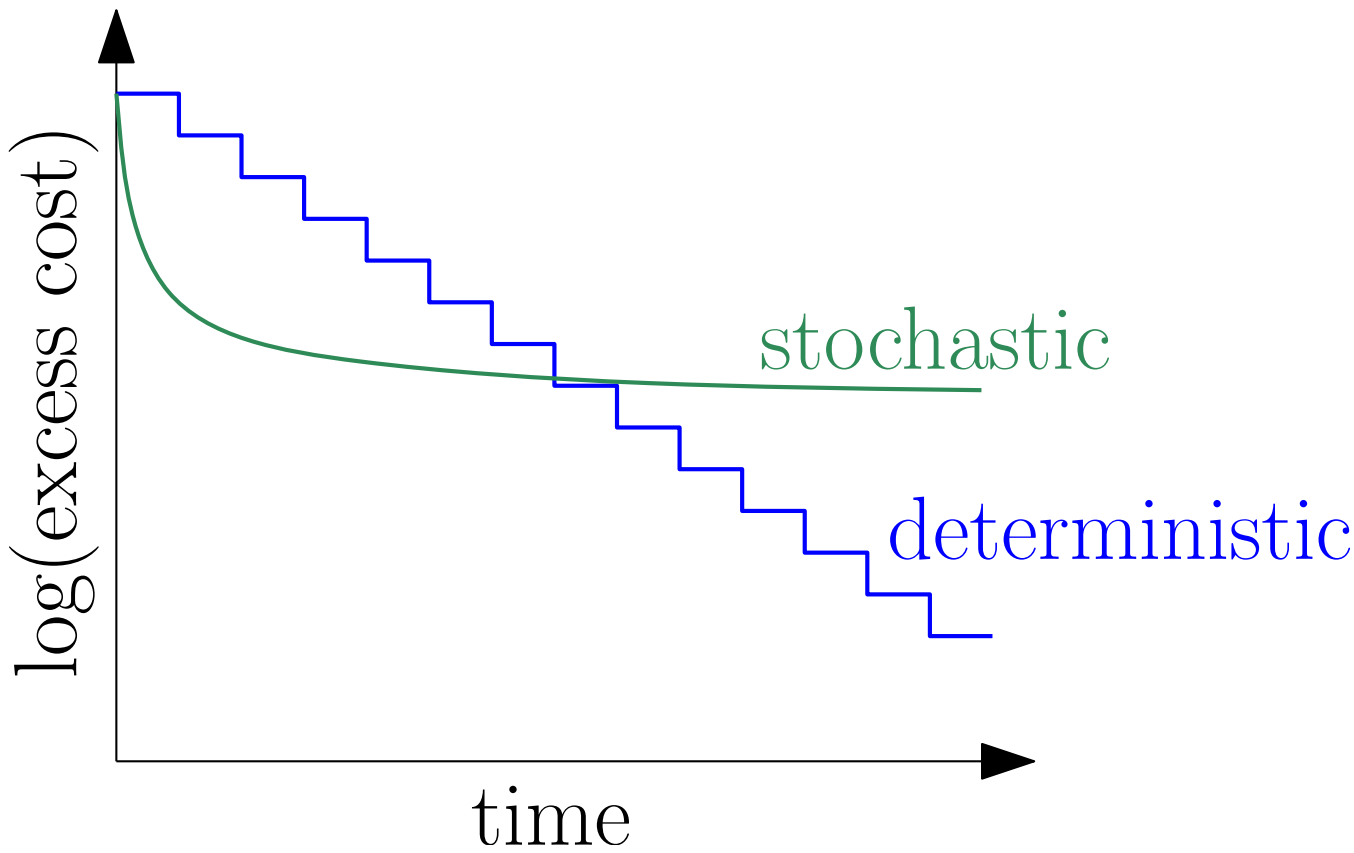
So using stochastic subgradient, solve n times faster.

Method	Assumptions	Batch	Stochastic
Gradient	convex	$O(1/k)$	$O(1/\sqrt{k})$
Gradient	strongly cvx	$O((1-\mu/L)^k)$	$O(1/k)$

For smooth functions there's a big gap!

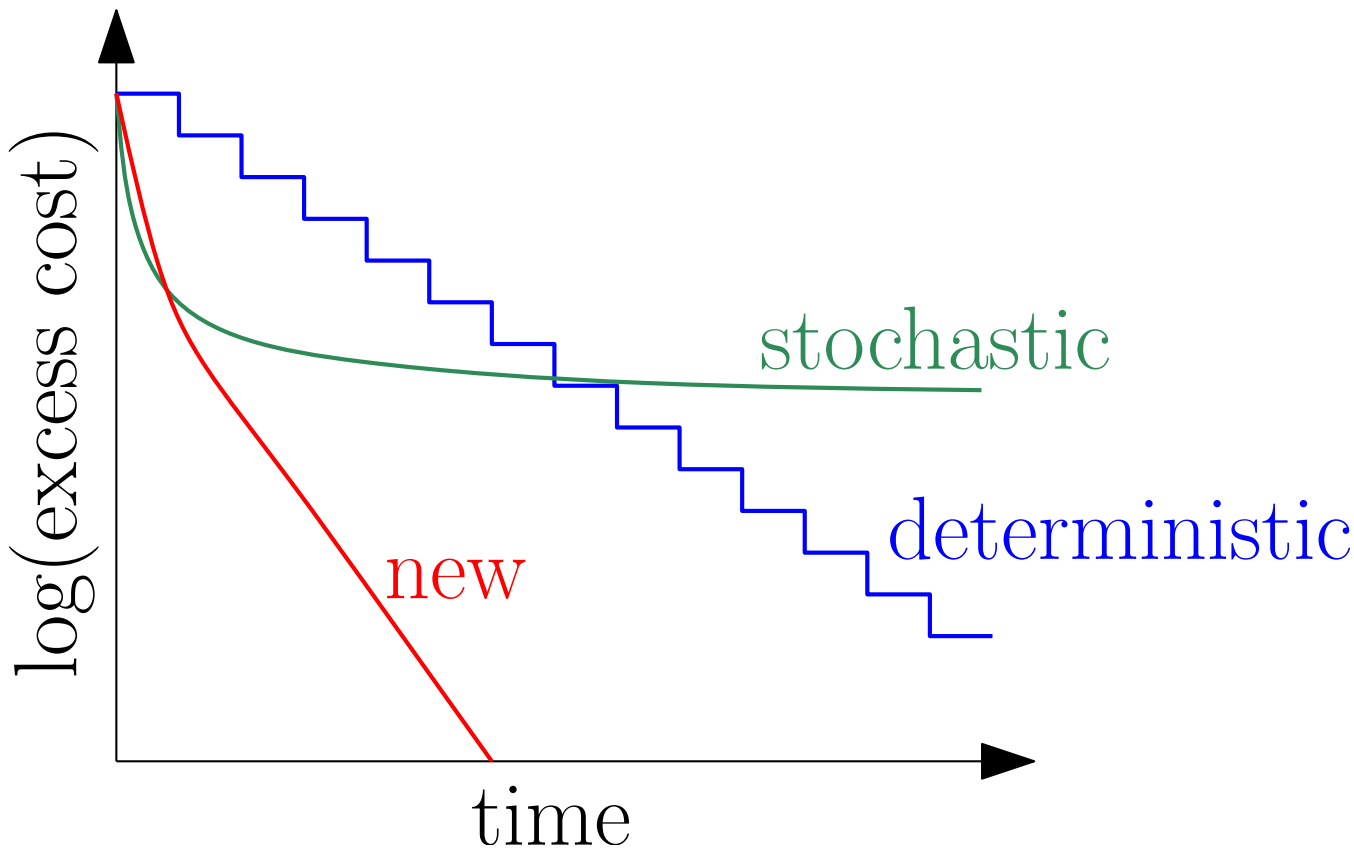
Stochastic vs. deterministic methods

Goal = best of both worlds: Linear rate with $O(d)$ iteration cost
Simple choice of step size



Stochastic vs. deterministic methods

Goal = best of both worlds: Linear rate with $O(d)$ iteration cost
Simple choice of step size



Linearly convergent stochastic methods

- **Many related algorithms**
  - SAG **?**
  - SDCA **?**
  - SVRG **??**
  - MISO **?**
  - Finito **?**
  - SAGA **?**
  - ...
- **Similar rates of convergence and iterations**

Linearly convergent stochastic methods

- **Many related algorithms**
  - SAG **?**
  - SDCA **?**
  - SVRG **??**
  - MISO **?**
  - Finito **?**
  - SAGA **?**
  - . . .
- **Similar rates of convergence and iterations**
- **Different interpretations and proofs / proof lengths**
  - Lazy gradient evaluations
  - Variance reduction

Running-time comparisons (strongly-convex)

- **Assumptions**: $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
  - Each f_i convex L-smooth and f is μ-strongly convex

Running-time comparisons (strongly-convex)

- **Assumptions**: $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
  - Each f_i convex L-smooth and f is μ-strongly convex

Stochastic gradient descent	$d\times$	$\frac{L}{\mu}$	$\times \quad \frac{1}{\varepsilon}$
Gradient descent	$d\times$	$n\frac{L}{\mu}$	$\times \log \frac{1}{\varepsilon}$
Accelerated gradient descent	$d\times$	$n\sqrt{\frac{L}{\mu}}$	$\times \log \frac{1}{\varepsilon}$
SAG/SVRG	$d\times$	$(n+\frac{L}{\mu})$	$\times \log \frac{1}{\varepsilon}$

Running-time comparisons (strongly-convex)

- **Assumptions**: $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
  - Each f_i convex L-smooth and f is μ-strongly convex

Stochastic gradient descent	$d\times$	$\frac{L}{\mu}$	$\times \quad \frac{1}{\varepsilon}$
Gradient descent	$d\times$	$n\frac{L}{\mu}$	$\times \log \frac{1}{\varepsilon}$
Accelerated gradient descent	$d\times$	$n\sqrt{\frac{L}{\mu}}$	$\times \log \frac{1}{\varepsilon}$
SAG/SVRG	$d\times$	$(n + \frac{L}{\mu})$	$\times \log \frac{1}{\varepsilon}$

- **Beating lower bounds ??**: with additional assumptions

  (1) stochastic gradient: exponential rate for finite sums

Running-time comparisons (non-strongly-convex)

- **Assumptions**: $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
  - Each f_i convex L-smooth
  - Ill conditioned problems: f may not be strongly-convex

Stochastic gradient descent	$d\times$	$1/\varepsilon^2$
Gradient descent	$d\times$	n/ε
Accelerated gradient descent	$d\times$	$n/\sqrt{\varepsilon}$
SAG/SVRG	$d\times$	$\sqrt{n}/\varepsilon$

Running-time comparisons (non-strongly-convex)

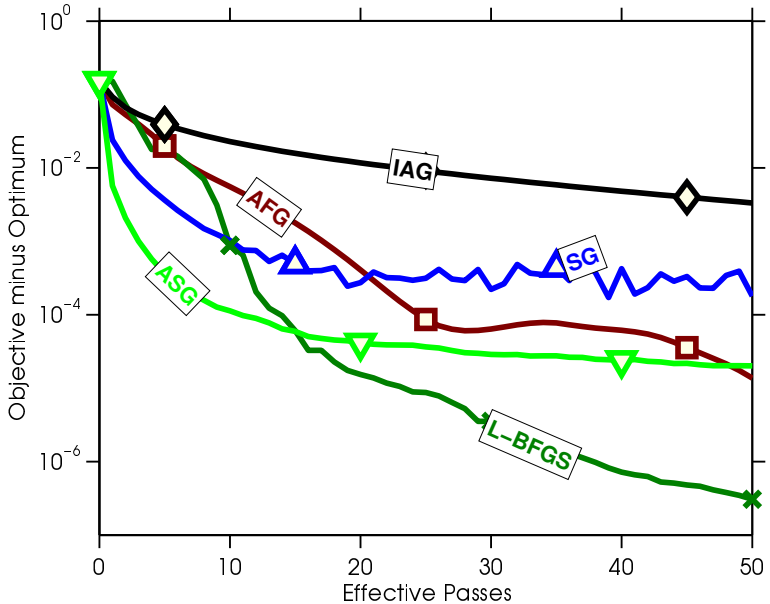
- **Assumptions**: $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
  - Each f_i convex L-smooth
  - Ill conditioned problems: f may not be strongly-convex

Stochastic gradient descent	$d\times$	$1/\varepsilon^2$
Gradient descent	$d\times$	n/ε
Accelerated gradient descent	$d\times$	$n/\sqrt{\varepsilon}$
SAG/SVRG	$d\times$	$\sqrt{n}/\varepsilon$

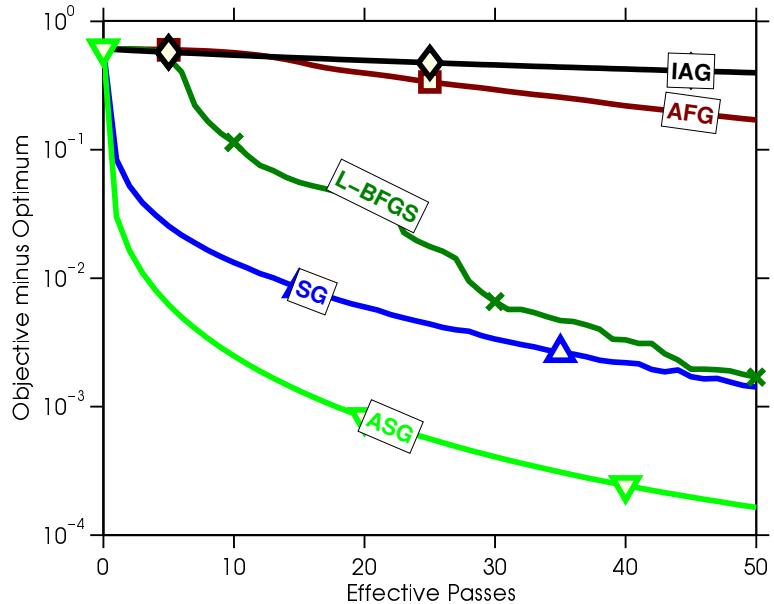
- Adaptivity to potentially hidden strong convexity
- No need to know the local/global strong-convexity constant

Experimental results (logistic regression)

`quantum` dataset
($n = 50\,000, d = 78$)

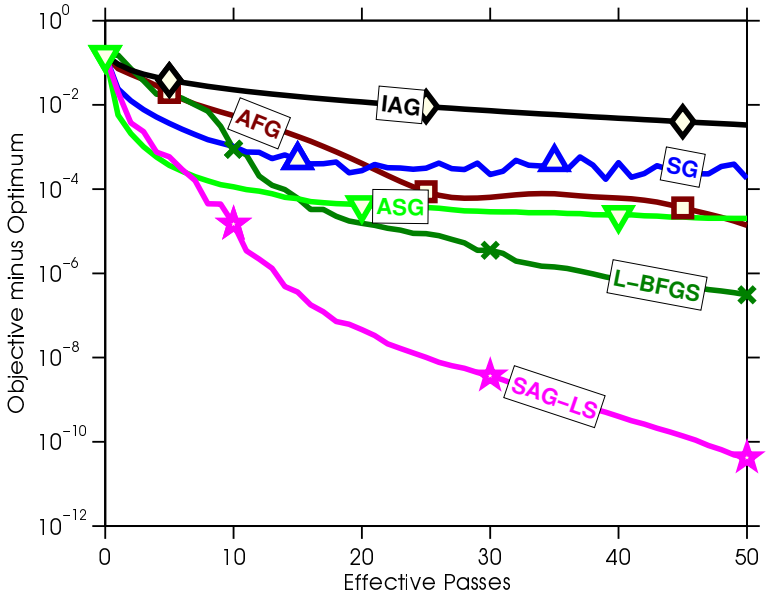


`rcv1` dataset
($n = 697\,641, d = 47\,236$)

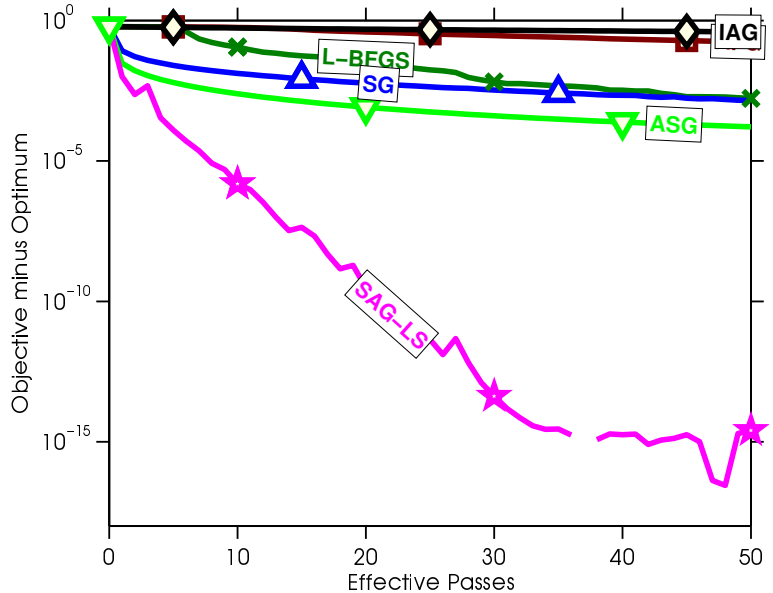


Experimental results (logistic regression)

`quantum` dataset
($n = 50\,000, d = 78$)



`rcv1` dataset
($n = 697\,641, d = 47\,236$)



Stochastic variance reduced gradient (SVRG)

- Initialize $\tilde{x} \in \mathbb{R}^d$
- For $i_{\text{epoch}} = 1$ to # of epochs
  - Compute all gradients $f_i'(\tilde{x})$; store $f'(\tilde{x}) = \frac{1}{n}\sum_{i=1}^n f_i'(\tilde{x})$
  - Initialize $x_0 = \tilde{x}$
  - For $t = 1$ to length of epochs (m)

$$x_t = x_{t-1} - \eta\left[f'(\tilde{x}) + \left(f_{i(t)}'(x_{t-1}) - f_{i(t)}'(\tilde{x})\right)\right]$$

  - Update $\tilde{x} = x_t$
- Output: $\tilde{\theta}$

– two gradient evaluations per inner step

– Two parameters: length of epochs + step-size γ

– Linear convergence rate, simple proof

Key Idea: Variance reduction

Principle: reducing variance of sample of X by using a sample from another random variable Y with known expectation

$$Z_\alpha = \alpha(X - Y) + \mathbb{E}Y$$

- $\mathbb{E}Z_\alpha = \alpha\mathbb{E}X + (1-\alpha)\mathbb{E}Y$
- $\mathrm{var}(Z_\alpha) = \alpha^2\big[\mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X,Y)\big]$
- $\alpha = 1$: no bias, $\alpha < 1$: potential bias (but reduced variance)
- Useful if Y positively correlated with X

Key Idea: Variance reduction

Principle: reducing variance of sample of X by using a sample from another random variable Y with known expectation

$$Z_\alpha = \alpha(X - Y) + \mathbb{E}Y$$

- $\mathbb{E}Z_\alpha = \alpha\mathbb{E}X + (1-\alpha)\mathbb{E}Y$
- $\mathrm{var}(Z_\alpha) = \alpha^2\big[\mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X,Y)\big]$
- $\alpha = 1$: no bias, $\alpha < 1$: potential bias (but reduced variance)
- Useful if Y positively correlated with X

Application to gradient estimation ??

- SVRG: $X = f'_{i(k)}(x_{k-1})$, $Y = f'_{i(k)}(\tilde{x})$, $\alpha = 1$, with $\tilde{\theta}$ stored
- $\mathbb{E}Y = \frac{1}{n}\sum_{i=1}^n f'_i(\tilde{x})$ full gradient at $\tilde{x}$;
  $X - Y = f'_{i(k)}(x_{k-1}) - f'_{i(k)}(\tilde{x})$

SVRG: Convergence

Theorem. Let each $f_i \in C^1_L$ be convex; f be μ-strongly convex. Let m be such that $\rho = \frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1$, then after s epochs

$$\mathbb{E}[f(\tilde{x}_s) - f(x^*)] \leq \rho^s [f(\tilde{x}_0) - f(x^*)].$$

SVRG: Convergence

Theorem. Let each $f_i \in C^1_L$ be convex; f be μ-strongly convex. Let m be such that $\rho = \frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1$, then after s epochs

$$\mathbb{E}[f(\tilde{x}_s) - f(x^*)] \leq \rho^s [f(\tilde{x}_0) - f(x^*)].$$

Proof. Let's consider a proof for a simpler version where

$$\tilde{x}_{s+1} = x_s^j \quad j \in U(0, 1, \ldots, m-1),$$

instead of using $\tilde{x}_{s+1} = x_s^m$ (makes analysis easier).

SVRG: Convergence

Let $g_s^t := \nabla f_{i(t)}(x_s^t) - \nabla f_{i(t)}(\tilde{x}_s) + \nabla f(\tilde{x}_s)$.
As before, using strong convexity, taking suitable expectations

$$\begin{aligned}\mathbb{E}[\|x_s^{t+1} - x^*\|^2] &= \mathbb{E}[\|x_s^t - \eta g_s^t - x^*\|^2] \\ &= \mathbb{E}[\|x_s^t - x^*\|^2] - 2\eta\langle \mathbb{E} g_s^t,\, x_s^t - x^*\rangle + \eta^2 \mathbb{E}[\|g_s^t\|^2]\end{aligned}$$

SVRG: Convergence

Let $g_s^t := \nabla f_{i(t)}(x_s^t) - \nabla f_{i(t)}(\tilde{x}_s) + \nabla f(\tilde{x}_s)$.
As before, using strong convexity, taking suitable expectations

$$\begin{aligned}\mathbb{E}[\|x_s^{t+1} - x^*\|^2] &= \mathbb{E}[\|x_s^t - \eta g_s^t - x^*\|^2] \\ &= \mathbb{E}[\|x_s^t - x^*\|^2] - 2\eta\langle \mathbb{E}g_s^t,\, x_s^t - x^*\rangle + \eta^2\mathbb{E}[\|g_s^t\|^2] \\ &\leq \mathbb{E}[\|x_s^t - x^*\|^2] - 2\eta(f(x_s^t) - f(x^*)) + \eta^2\mathbb{E}[\|g_s^t\|^2]\end{aligned}$$

SVRG: Convergence

Let $g_s^t := \nabla f_{i(t)}(x_s^t) - \nabla f_{i(t)}(\tilde{x}_s) + \nabla f(\tilde{x}_s)$.
As before, using strong convexity, taking suitable expectations

$$\begin{aligned}\mathbb{E}[\|x_s^{t+1} - x^*\|^2] &= \mathbb{E}[\|x_s^t - \eta g_s^t - x^*\|^2] \\ &= \mathbb{E}[\|x_s^t - x^*\|^2] - 2\eta\langle \mathbb{E}g_s^t,\, x_s^t - x^*\rangle + \eta^2\mathbb{E}[\|g_s^t\|^2] \\ &\leq \mathbb{E}[\|x_s^t - x^*\|^2] - 2\eta(f(x_s^t) - f(x^*)) + \eta^2\mathbb{E}[\|g_s^t\|^2]\end{aligned}$$

Key step: Controlling the noise term $\mathbb{E}[\|g_s^t\|^2]$. In particular, we want to bound it in terms of the objective function

SVRG: Key lemma

Lemma $\mathbb{E}[\|g_s^t\|^2] \leq 4L(f(x_s^t) - f(x^*) + f(\tilde{x}_s) - f(x^*))$

Intuition: If $x_s^t \approx \tilde{x}_s \approx x^*$, so $\mathbb{E}[\|g_s^t\|^2] \approx 0$ (reduced variance possible)

SVRG: Key lemma

Lemma $\mathbb{E}[\|g_s^t\|^2] \leq 4L(f(x_s^t) - f(x^*) + f(\tilde{x}_s) - f(x^*))$

Intuition: If $x_s^t \approx \tilde{x}_s \approx x^*$, so $\mathbb{E}[\|g_s^t\|^2] \approx 0$ (reduced variance possible)

Conditional on everything prior to x_s^{t+1}, key lemma yields

$$\mathbb{E}[\|x_s^{t+1} - x^*\|^2] \leq \|x_s^t - x^*\|^2 - 2\eta(f(x_s^t) - f(x^*)) + \eta^2 \mathbb{E}[\|g_s^t\|^2]$$

SVRG: Key lemma

Lemma $\mathbb{E}[\|g_s^t\|^2] \leq 4L(f(x_s^t) - f(x^*) + f(\tilde{x}_s) - f(x^*))$

Intuition: If $x_s^t \approx \tilde{x}_s \approx x^*$, so $\mathbb{E}[\|g_s^t\|^2] \approx 0$ (reduced variance possible)

Conditional on everything prior to x_s^{t+1}, key lemma yields

$$\begin{aligned}\mathbb{E}[\|x_s^{t+1} - x^*\|^2] &\leq \|x_s^t - x^*\|^2 - 2\eta(f(x_s^t) - f(x^*)) + \eta^2\mathbb{E}[\|g_s^t\|^2] \\ &\leq \|x_s^t - x^*\|^2 - 2\eta(f(x_s^t) - f(x^*)) + 4L\eta^2[f(x_s^t) - f(x^*) + f(\tilde{x}_s) - f(x^*)]\end{aligned}$$

SVRG: Key lemma

Lemma $\mathbb{E}[\|g_s^t\|^2] \leq 4L(f(x_s^t) - f(x^*) + f(\tilde{x}_s) - f(x^*))$

Intuition: If $x_s^t \approx \tilde{x}_s \approx x^*$, so $\mathbb{E}[\|g_s^t\|^2] \approx 0$ (reduced variance possible)

Conditional on everything prior to x_s^{t+1}, key lemma yields

$$\begin{aligned}
\mathbb{E}[\|x_s^{t+1} - x^*\|^2] &\leq \|x_s^t - x^*\|^2 - 2\eta(f(x_s^t) - f(x^*)) + \eta^2 \mathbb{E}[\|g_s^t\|^2] \\
&\leq \|x_s^t - x^*\|^2 - 2\eta(f(x_s^t) - f(x^*)) + 4L\eta^2[f(x_s^t) - f(x^*) + f(\tilde{x}_s) - f(x^*)] \\
&= \|x_s^t - x^*\|^2 - 2\eta(1 - 2L\eta)[f(x_s^t) - f(x^*)] + 4L\eta^2[f(\tilde{x}_s) - f(x^*)] \quad (*)
\end{aligned}$$

SVRG: Key lemma

Lemma $\mathbb{E}[\|g_s^t\|^2] \leq 4L(f(x_s^t) - f(x^*) + f(\tilde{x}_s) - f(x^*))$

Intuition: If $x_s^t \approx \tilde{x}_s \approx x^*$, so $\mathbb{E}[\|g_s^t\|^2] \approx 0$ (reduced variance possible)

Conditional on everything prior to x_s^{t+1}, key lemma yields

$$\begin{aligned}
&\mathbb{E}[\|x_s^{t+1} - x^*\|^2] \leq \|x_s^t - x^*\|^2 - 2\eta(f(x_s^t) - f(x^*)) + \eta^2\mathbb{E}[\|g_s^t\|^2] \\
&\leq \|x_s^t - x^*\|^2 - 2\eta(f(x_s^t) - f(x^*)) + 4L\eta^2[f(x_s^t) - f(x^*) + f(\tilde{x}_s) - f(x^*)] \\
&= \|x_s^t - x^*\|^2 - 2\eta(1 - 2L\eta)[f(x_s^t) - f(x^*)] + 4L\eta^2[f(\tilde{x}_s) - f(x^*)] \quad (*)
\end{aligned}$$

Since we set $\tilde{x}_{s+1} = x_s^j \quad j \in U(0, 1, \ldots, m-1)$ we get

$$2\eta(1 - 2L\eta)m\mathbb{E}[f(\tilde{x}_{s+1}) - f(x^*)] = 2\eta(1 - 2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t) - f(x^*)]$$

SVRG Convergence

$$2\eta(1-2L\eta)m\mathbb{E}[f(\tilde{x}_{s+1})-f(x^*)] = 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

SVRG Convergence

$$2\eta(1-2L\eta)m\mathbb{E}[f(\tilde{x}_{s+1})-f(x^*)] = 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^m - x^*\|^2] + 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

SVRG Convergence

$$2\eta(1-2L\eta)m\mathbb{E}[f(\tilde{x}_{s+1})-f(x^*)] = 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^m - x^*\|^2] + 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^0 - x^*\|^2] + 4Lm\eta^2[f(\tilde{x}_s)-f(x^*)] \qquad \text{(using (*))}$$

SVRG Convergence

$$2\eta(1-2L\eta)m\mathbb{E}[f(\tilde{x}_{s+1})-f(x^*)] = 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^m - x^*\|^2] + 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^0 - x^*\|^2] + 4Lm\eta^2[f(\tilde{x}_s)-f(x^*)] \qquad \text{(using (*))}$$

$$= \mathbb{E}[\|\tilde{x}_s - x^*\|^2] + 4Lm\eta^2\mathbb{E}[f(\tilde{x}_s)-f(x^*)]$$

SVRG Convergence

$$2\eta(1-2L\eta)m\mathbb{E}[f(\tilde{x}_{s+1})-f(x^*)] = 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^m - x^*\|^2] + 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^0 - x^*\|^2] + 4Lm\eta^2[f(\tilde{x}_s)-f(x^*)] \qquad \text{(using (*))}$$

$$= \mathbb{E}[\|\tilde{x}_s - x^*\|^2] + 4Lm\eta^2\mathbb{E}[f(\tilde{x}_s)-f(x^*)]$$

$$\leq \frac{2}{\mu}\mathbb{E}[f(\tilde{x}_s)-f(x^*)] + 4Lm\eta^2\mathbb{E}[f(\tilde{x}_s)-f(x^*)]$$

SVRG Convergence

$$2\eta(1-2L\eta)m\mathbb{E}[f(\tilde{x}_{s+1})-f(x^*)] = 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^m - x^*\|^2] + 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}[f(x_s^t)-f(x^*)]$$

$$\leq \mathbb{E}[\|x_{s+1}^0 - x^*\|^2] + 4Lm\eta^2[f(\tilde{x}_s)-f(x^*)] \qquad \text{(using (*))}$$

$$= \mathbb{E}[\|\tilde{x}_s - x^*\|^2] + 4Lm\eta^2\mathbb{E}[f(\tilde{x}_s)-f(x^*)]$$

$$\leq \frac{2}{\mu}\mathbb{E}[f(\tilde{x}_s)-f(x^*)] + 4Lm\eta^2\mathbb{E}[f(\tilde{x}_s)-f(x^*)]$$

Consequently

$$\mathbb{E}[f(\tilde{x}_{s+1})-f(x^*)] \leq \frac{\frac{2}{\mu}+4Lm\eta^2}{2\eta(1-2L\eta)m}\mathbb{E}[f(\tilde{x}_s)-f(x^*)]$$

Which gives us the contraction factor ρ as desired.

References

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014a.

A. Defazio, J. Domke, and T. S. Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In *Proc. ICML*, 2014b.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley & Sons, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer, 2004.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14 (Feb):567–599, 2013.

L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems*, 2013.