Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 7: Stochastic gradient methods

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2017)

License Information for Mathematics of Data Slides

Outline

- This class
  1. Stochastic gradient methods
  2. Stochastic gradient methods with averaging
  3. Accelerated stochastic gradient methods
  4. Stochastic variance reduced gradient methods
- Next class
  1. Composite convex minimization

Recommended reading materials

1. V. Cevher; S. Becker, and M. Schmidt. Convex optimization for big data. *IEEE Signal Process. Mag.*, vol. 31, pp. 32–43, 2014.
2. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, vol. 19, pp. 1574–1609, 2008.
3. L. Xiao and T. Zhang, A proximal stochastic gradient method with progressive variance reduction, *SIAM J. Optim.*, vol. 24, pp. 2057–2075, 2014.

What is this class about?

Recall: Gradient method

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k)$$

where γ_k is a step-size to be chosen so that $\mathbf{x}^k$ converges to $\mathbf{x}^\star$.

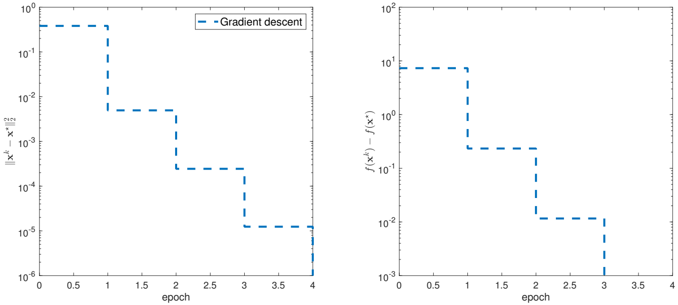
What is this class about?

Recall: Gradient method

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k)$$

where γ_k is a step-size to be chosen so that $\mathbf{x}^k$ converges to $\mathbf{x}^\star$.



- Least squares: $\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, where $\mathbf{A} \in \mathbb{R}^{n\times p}$.
- 1 epoch means 1 'pass' over the full gradient, i.e., 1 epoch $= n$.

What is this class about?

Stochastic gradient method

Let $G(\mathbf{x}^k, \theta_k)$ be an *unbiased estimate* of the gradient $\nabla f(\mathbf{x}^k)$, i.e.,

$$\mathbb{E}_{\theta_k}[G(\mathbf{x}^k, \theta_k)] = \nabla f(\mathbf{x}^k).$$

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k).$$

What is this class about?

Stochastic gradient method

Let $G(\mathbf{x}^k, \theta_k)$ be an *unbiased estimate* of the gradient $\nabla f(\mathbf{x}^k)$, i.e.,

$$\mathbb{E}_{\theta_k}[G(\mathbf{x}^k, \theta_k)] = \nabla f(\mathbf{x}^k).$$

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k).$$

Claim: The stochastic gradient computation can be super cheap!

What is this class about?

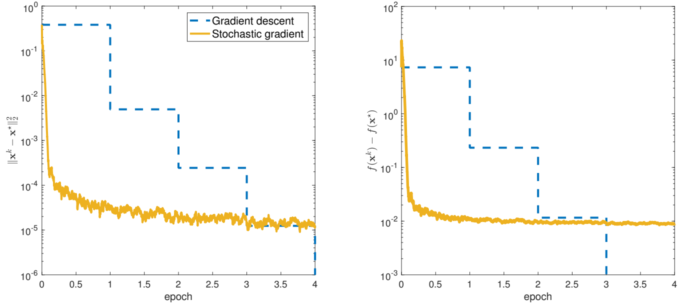
Stochastic gradient method

Let $G(\mathbf{x}^k, \theta_k)$ be an *unbiased estimate* of the gradient $\nabla f(\mathbf{x}^k)$, i.e.,

$$\mathbb{E}_{\theta_k}[G(\mathbf{x}^k, \theta_k) := \nabla f_i(\mathbf{x}^k)] = \nabla f(\mathbf{x}^k).$$

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k).$$



Example: Large scale optimization

Convex optimization with finite sums

$$\arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\}.$$

Example: Large scale optimization

Convex optimization with finite sums

$$\arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\}.$$

Gradient descent method (GD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\mathbf{x}^k).$$

The computational cost of the deterministic gradient method per iteration is proportional to n. Hence, it can be expensive for a large n.

Example: Statistical learning with ERM

Recall:

A basic statistical learning model [1]

A statistical learning model consists of the following three elements.

1. A sample of i.i.d. random variables $(\mathbf{a}_j, b_j) \in \mathcal{A} \times \mathcal{B}$, $j = 1, \ldots, n$, following an *unknown* probability distribution $\mathbb{P}$.
2. A class (set) $\mathcal{F}$ of functions $f : \mathcal{A} \to \mathcal{B}$.
3. A loss function $L : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$.

Example: Statistical learning with ERM

Recall:

A basic statistical learning model [1]

A statistical learning model consists of the following three elements.

1. A sample of i.i.d. random variables $(\mathbf{a}_j, b_j) \in \mathcal{A} \times \mathcal{B}$, $j = 1, \ldots, n$, following an *unknown* probability distribution $\mathbb{P}$.
2. A class (set) $\mathcal{F}$ of functions $f : \mathcal{A} \to \mathcal{B}$.
3. A loss function $L : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$.

Definition (Risk)

Let $(\mathbf{a}, b)$ follow the probability distribution $\mathbb{P}$ and be independent of $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$. Then, the risk corresponding to any $f \in \mathcal{F}$ is its expected loss:

$$R(f) := \mathbb{E}_{(\mathbf{a},b)}\left[L(f(\mathbf{a}), b)\right].$$

Statistical learning seeks to find a $f^\star \in \mathcal{F}$ that minimizes the risk, i.e., it solves

$$f^\star \in \arg\min_{f \in \mathcal{F}} R(f).$$

Many problems in machine learning cast into this formulation!

Recall: Empirical risk minimization (ERM)

- By the law of large numbers, we can expect that for each $f \in \mathcal{F}$,

$$R(f) := \mathbb{E}\left[L(f(\mathbf{a}), b)\right] \approx \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{a}_j), b_j)$$

when n is large enough, with high probability.

Empirical risk minimization (ERM) [1]

We approximate $f^\star$ by minimizing the *empirical average of the loss* instead of the risk.

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \left\{ R_n(f) := \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{a}_j), b_j) \right\}.$$

Recall: Empirical risk minimization (ERM)

- By the law of large numbers, we can expect that for each $f \in \mathcal{F}$,

$$R(f) := \mathbb{E}\left[L(f(\mathbf{a}), b)\right] \approx \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{a}_j), b_j)$$

when n is large enough, with high probability.

Empirical risk minimization (ERM) [1]

We approximate $f^\star$ by minimizing the *empirical average of the loss* instead of the risk.

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \left\{ R_n(f) := \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{a}_j), b_j) \right\}.$$

Least squares

Recall that the LS estimator is given by

$$\hat{\mathbf{x}}_{\mathsf{LS}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\} = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{j=1}^{n} (b_j - \langle \mathbf{a}_j, \mathbf{x}\rangle)^2 \right\},$$

where we define $\mathbf{b} := (b_1, \ldots, b_n)^T$ and $\mathbf{a}_j^T$ to be the j-th row of $\mathbf{A}$.

Recall: Empirical risk minimization (ERM)

- By the law of large numbers, we can expect that for each $f \in \mathcal{F}$,

$$R(f) := \mathbb{E}\left[L(f(\mathbf{a}), b)\right] \approx \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{a}_j), b_j)$$

when n is large enough, with high probability.

Empirical risk minimization (ERM) [1]

We approximate $f^\star$ by minimizing the *empirical average of the loss* instead of the risk.

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \left\{ R_n(f) := \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{a}_j), b_j) \right\}.$$

SVM

Recall the unconstrained SVM formulation

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \frac{1}{n}\sum_{j=1}^{n} \max\left\{1 - b_j\langle \mathbf{a}_j, \mathbf{x}\rangle, 0\right\} + \lambda\|\mathbf{x}\|_2^2 \right\}$$

where $\mathbf{b} := (b_1, \ldots, b_n)^T \in \{-1, 1\}^n$.

Recall: Empirical risk minimization (ERM)

- By the law of large numbers, we can expect that for each $f \in \mathcal{F}$,

$$R(f) := \mathbb{E}\left[L(f(\mathbf{a}), b)\right] \approx \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{a}_j), b_j)$$

when n is large enough, with high probability.

Empirical risk minimization (ERM) [1]

We approximate $f^\star$ by minimizing the *empirical average of the loss* instead of the risk.

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \left\{ R_n(f) := \frac{1}{n}\sum_{j=1}^{n} L(f(\mathbf{a}_j), b_j) \right\}.$$

Logistic regression

Recall the logistic regression formulation

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x},\mu} \left\{ \frac{1}{n}\sum_{j=1}^{n} \log\left(1 + e^{-b_j(\langle \mathbf{x}, \mathbf{a}_j\rangle + \mu)}\right) : \mathbf{x} \in \mathbb{R}^p, \mu \in \mathbb{R} \right\}$$

where $\mathbf{b} := (b_1, \ldots, b_n)^T \in \{-1, 1\}^n$.

*Motivation: Statistical learning with streaming data (self-study)

Recall that statistical learning seeks to find a $f^\star \in \mathcal{F}$ that minimizes the *expected* risk,

$$f^\star \in \arg\min_{f \in \mathcal{F}} \left\{ R(f) := \mathbb{E}_{(\mathbf{a},b)} \left[L(f(\mathbf{a}), b) \right] \right\}, \qquad .$$

In practice, data can arrive in a *streaming* way.

Example: Markowitz portfolio optimization

$$f^\star := \min_{\mathbf{x} \in \mathcal{X}} \left\{ \mathbb{E}\left[|\rho - \langle \mathbf{x}, \theta_t \rangle|^2 \right] \right\}$$

- $\rho \in \mathbb{R}$ is the desired return.
- $\mathcal{X}$ is intersection of the standard simplex and the constraint: $\langle \mathbf{x}, \mathbb{E}[\theta_t] \rangle \geq \rho$.

Gradient method

$$f^{k+1} = f^k - \gamma_k \nabla R(f) = f^k - \gamma_k \mathbb{E}_{(\mathbf{a},b)}[\nabla L(f^k(\mathbf{a}), b)].$$

This can not be implemented in practice as the distribution of $(\mathbf{a}, b)$ is unknown.

Unconstrained convex minimization

Problem (**Mathematical formulation**)

Consider the following convex minimization problem:

$$f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \mathbb{E}[h(\mathbf{x},\theta)] \right\}$$

- θ *is a random vector whose probability distribution is supported on set* Θ.
- $f(\mathbf{x}) := \mathbb{E}[h(\mathbf{x},\theta)]$ *is proper, closed, and convex.*
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in \mathit{dom}(f) : f(\mathbf{x}^\star) = f^\star\}$ *is nonempty.*

Unconstrained convex minimization

Problem (**Mathematical formulation**)

Consider the following convex minimization problem:

$$f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \mathbb{E}[h(\mathbf{x},\theta)] \right\}$$

- θ *is a random vector whose probability distribution is supported on set* Θ.
- $f(\mathbf{x}) := \mathbb{E}[h(\mathbf{x},\theta)]$ *is proper, closed, and convex.*
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in$ *dom*$(f) : f(\mathbf{x}^\star) = f^\star\}$ *is nonempty.*

Example: Convex optimization with finite sums

The problem

$$\arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\},$$

can be rewritten as

$$\arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \mathbb{E}_i[f_i(\mathbf{x})] \right\}, \qquad i \text{ is uniformly distributed over} \{1, 2, \cdots, n\}.$$

Stochastic gradient method (SG)

Stochastic gradient method (SG)
1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in \,]0, +\infty[^{\mathbb{N}}$. **2.** For $k = 0, 1, \ldots$ perform: $$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k).$$

- $G(\mathbf{x}^k, \theta_k)$ is an unbiased estimate of the full gradient, i.e., it satisfies

$$\mathbb{E}[G(\mathbf{x}^k, \theta_k)] = \nabla f(\mathbf{x}^k).$$

Stochastic gradient method (SG)

Stochastic gradient method (SG)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k).$$

- $G(\mathbf{x}^k, \theta_k)$ is an unbiased estimate of the full gradient, i.e., it satisfies

$$\mathbb{E}[G(\mathbf{x}^k, \theta_k)] = \nabla f(\mathbf{x}^k).$$

Remark

- The cost of computing $G(\mathbf{x}^k, \theta_k)$ is typically much cheaper than that of $\nabla f(\mathbf{x}^k)$.
- As $G(\mathbf{x}^k, \theta_k)$ is an unbiased estimate of the full gradient, we expect that SG would also perform well.
- We assume that $\{\theta_k\}$ are jointly independent.
- SG is not a monotonic descent method.

Example: Convex optimization with finite sums

Consider the problem:

$$\min_{\mathbf{x}\in\mathbb{R}^p}\left\{f(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x})\right\}$$

Example: Convex optimization with finite sums

Consider the problem:

$$\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ f(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} f_j(\mathbf{x}) \right\}$$

Stochastic gradient methods (SG)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f_i(\mathbf{x}^k) \qquad i \text{ is uniformly distributed over} \{1, ..., n\}$$

- Note: $\mathbb{E}_i[\nabla f_i(\mathbf{x}^k)] = \sum_{j=1}^n \nabla f_j(\mathbf{x}^k)/n = \nabla f(\mathbf{x}^k)$.
- The computational cost of SG per iteration is independent of n.

Theoretical analysis

Recall: convergence of gradient descent method

- strong convexity and smoothness assumptions imply linear convergence, i.e.,

$$f(\mathbf{x}^k) - f^\star \leq O(\rho^k), \qquad \rho < 1.$$

- smoothness assumption implies

$$f(\mathbf{x}^k) - f^\star \leq O(1/k).$$

For SG methods, we will show that

Convergence of SG

- strong convexity implies

$$\mathbb{E} f(\mathbf{x}^k) - f^\star \leq O(1/k).$$

- without strong convexity,

$$\mathbb{E} f(\mathbf{x}^k) - f^\star \leq O(1/\sqrt{k}).$$

Convergence of SG I: strongly convex case

Theorem (Convergence in expectation [2])

Suppose that:

1. f *is* μ*-strongly convex,*
2. $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2] \leq M^2$,
3. $\gamma_k = \gamma_0/(k+1)$ *with* $\gamma_0 > \frac{1}{2\mu}$.

Then,

$$\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|^2] \leq \frac{C}{k}, \qquad C = \max\left\{\frac{\gamma_0^2 M^2}{2\gamma_0\mu - 1}, \|\mathbf{x}^0 - \mathbf{x}^\star\|^2, \gamma_0^2 M^2\right\}.$$

If, in addition, ∇f *is* L*-Lipschitz continuous, then,*

$$\mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{x}^\star)] \leq \frac{CL}{2k}.$$

• $\mathcal{O}(1/k)$ rate is optimal for SG under strong convexity .
• As will be given in Lecture 9, Assumption 2 can be replaced by a less strict condition, $\mathbb{E}[\|G(\mathbf{x}, \theta) - \nabla f(\mathbf{x})\|^2] \leq M^2$.

Convergence of SG II: non-strongly convex case

Theorem (Convergence in expectation [8])

Suppose that:

1. $\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|^2] \leq D^2$ *for all* k,
2. $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2] \leq M^2$,
3. $\gamma_k = \gamma_0/\sqrt{k}$.

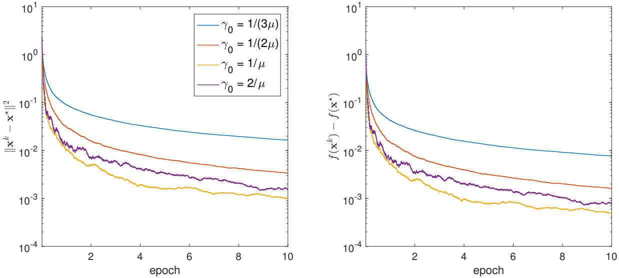
Then,

$$\mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{x}^\star)] \leq \left(\frac{D^2}{\gamma_0} + \gamma_0 M^2\right) \frac{2 + \log k}{\sqrt{k}}.$$

- Proof of this theorem can be found in [8].

Example: SG method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$

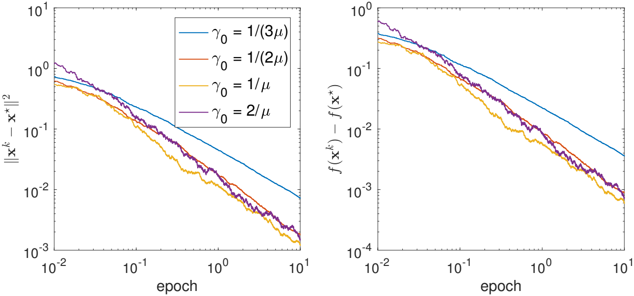


Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 50 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise with variance 1.
- $\gamma_k = \gamma_0/(k + k_0)$.

Example: SG method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$



Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 50 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise with variance 1.
- $\gamma_k = \gamma_0/(k + k_0)$.

$\gamma_0 = 1/\mu$ *is the best choice.*

Convergence for SG-A I: strongly convex case

Stochastic gradient method with averaging (SG-A)
1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$. **2a.** For $k = 0, 1, \ldots$ perform: $$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k).$$ **2b.** $\bar{\mathbf{x}}^k = \frac{1}{k}\sum_{j=1}^k \mathbf{x}^j$.

Theorem (Convergence of SG-A [9])

Assume

1. f *is* μ*-strongly convex,*
2. $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2] \leq M^2$,
3. $\gamma_k = \gamma_0/k$ *for some* $\gamma_0 \geq 1/\mu$.

Then,

$$\mathbb{E}[f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star)] \leq \frac{\gamma_0 M^2(1 + \log k)}{2k}.$$

Convergence for SG-A II: non-strongly convex case

Stochastic gradient method with averaging (SG-A)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2a. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k).$$

2b. $\bar{\mathbf{x}}^k = (\sum_{j=0}^k \gamma_j)^{-1} \sum_{j=0}^k \gamma_j \mathbf{x}^j$.

Theorem (Convergence of SG-A [2])

Denote $D = \|\mathbf{x}^0 - \mathbf{x}^\star\|$ *and assume* $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2] \le M^2$.
Then,

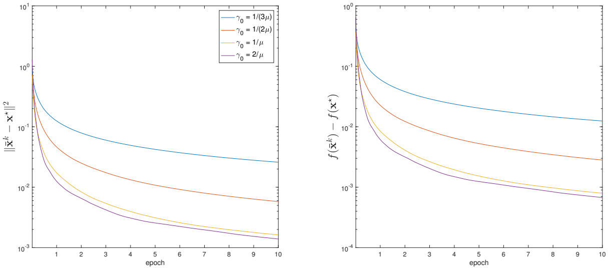
$$\mathbb{E}[f(\bar{\mathbf{x}}^{k+1}) - f(\mathbf{x}^\star)] \le \frac{D^2 + M^2 \sum_{j=0}^k \gamma_j^2}{2\sum_{j=0}^k \gamma_j}.$$

In addition, choosing $\gamma_k = D/(M\sqrt{k+1})$, *we get,*

$$\mathbb{E}[f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star)] \le \frac{MD(2 + \log k)}{\sqrt{k}}.$$

Example: SG-A method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$

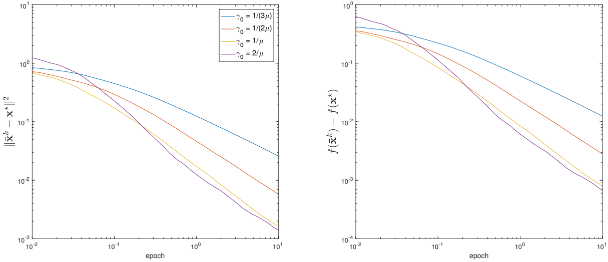


Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 50 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise with variance 1.
- $\bar{\mathbf{x}}^k = \frac{1}{k} \sum_{i=0}^{k-1} \mathbf{x}^i$, and $\gamma_k = \gamma_0/(k + k_0)$.

Example: SG-A method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$



Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 50 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise with variance 1.
- $\bar{\mathbf{x}}^k = \frac{1}{k}\sum_{i=0}^{k-1} \mathbf{x}^i$, and $\gamma_k = \gamma_0/(k + k_0)$.

SG-A is more stable than SG. $\gamma_0 = 2/\mu$ is the best choice.

*Adaptive stochastic gradient methods (Adagrad)

AdaGrad (diagonal form) [11]

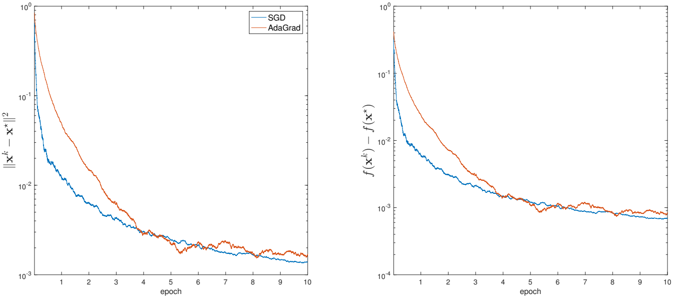
1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and δ.
2. For $k = 0, 1, \ldots$ perform:

$$\begin{cases} H_k = \delta I + \mathrm{diag}\left(\sum_{i=1}^{k} G(\mathbf{x}^i, \theta_i) G(\mathbf{x}^i, \theta_i)^T\right) \\ \mathbf{x}^{k+1} = \mathbf{x}^k - \gamma H_k^{-1/2} G(\mathbf{x}^k, \theta_k). \end{cases}$$

- The step-size for each coordinate is different.
- The algorithm is a stochastic version of the adaptive GD from Lecture 4.

*Example: AdaGrad vs SG

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$



Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 50 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise with variance 1.
- $\gamma_k = 1/(\mu(k + k_0))$ for SG. $\delta = 10^{-2}$ for AdaGrad.

Important remark!

All the results we have shown so far can be generalized for the non-smooth objectives, simply by replacing the gradient with a subgradient.

We will talk about the subgradient methods in the next lecture.

Recall: Accelerated gradient descent algorithm

- In what follows, we will assume that ∇f is L-Lipschitz continuous.

Accelerated Gradient algorithm for $\mathcal{F}_L^{1,1}$ (smoothness)

1. Set $\mathbf{x}^0 = \mathbf{y}^0 \in \mathrm{dom}(f)$ and $t_0 := 1$.

2. For $k = 0, 1, \ldots,$ iterate

$$\begin{cases} \mathbf{y}^{k+1} &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \\ t_{k+1} &= (1 + \sqrt{4t_k^2 + 1})/2 \\ \mathbf{x}^{k+1} &= \mathbf{y}^{k+1} + \frac{(t_k - 1)}{t_{k+1}}(\mathbf{y}^{k+1} - \mathbf{y}^k) \end{cases}$$

Recall: Accelerated gradient descent algorithm

- In what follows, we will assume that ∇f is L-Lipschitz continuous.

Accelerated Gradient algorithm for $\mathcal{F}_L^{1,1}$ (smoothness)
1. Set $\mathbf{x}^0 = \mathbf{y}^0 \in \mathrm{dom}(f)$ and $t_0 := 1$. **2.** For $k = 0, 1, \ldots,$ iterate $\begin{cases} \mathbf{y}^{k+1} &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \\ t_{k+1} &= (1 + \sqrt{4t_k^2 + 1})/2 \\ \mathbf{x}^{k+1} &= \mathbf{y}^{k+1} + \frac{(t_k - 1)}{t_{k+1}}(\mathbf{y}^{k+1} - \mathbf{y}^k) \end{cases}$

Accelerated Gradient algorithm for $\mathcal{F}_{L,\mu}^{1,1}$ (smoothness + stronglyConvex)
1. Choose $\mathbf{x}^0 = \mathbf{y}^0 \in \mathrm{dom}(f)$ **2.** For $k = 0, 1, \ldots,$ iterate $\begin{cases} \mathbf{y}^{k+1} &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \\ \mathbf{x}^{k+1} &= \mathbf{y}^{k+1} + \gamma(\mathbf{y}^{k+1} - \mathbf{y}^k) \end{cases}$ where $\gamma = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$.

Recall: Accelerated gradient descent algorithm

- In what follows, we will assume that ∇f is L-Lipschitz continuous.

Accelerated Gradient algorithm for $\mathcal{F}_L^{1,1}$ (smoothness)
1. Set $\mathbf{x}^0 = \mathbf{y}^0 \in \text{dom}(f)$ and $t_0 := 1$. **2.** For $k = 0, 1, \ldots,$ iterate $\begin{cases} \mathbf{y}^{k+1} &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \\ t_{k+1} &= (1 + \sqrt{4t_k^2 + 1})/2 \\ \mathbf{x}^{k+1} &= \mathbf{y}^{k+1} + \frac{(t_k - 1)}{t_{k+1}}(\mathbf{y}^{k+1} - \mathbf{y}^k) \end{cases}$

Accelerated Gradient algorithm for $\mathcal{F}_{L,\mu}^{1,1}$ (smoothness + stronglyConvex)
1. Choose $\mathbf{x}^0 = \mathbf{y}^0 \in \text{dom}(f)$ **2.** For $k = 0, 1, \ldots,$ iterate $\begin{cases} \mathbf{y}^{k+1} &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \\ \mathbf{x}^{k+1} &= \mathbf{y}^{k+1} + \gamma(\mathbf{y}^{k+1} - \mathbf{y}^k) \end{cases}$ where $\gamma = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$.

Can we use similar accelerated techniques for stochastic gradient methods?

Accelerated stochastic gradient method I

Accelerated stochastic gradient method (AccSG)
0. $0 \leq \mu$-strong convexity of F. **1.** Choose $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{0}$, $(\gamma_k)_{k\in\mathbb{N}}$, $(\alpha_k)_{k\in\mathbb{N}} \in \,]0, +\infty[^{\mathbb{N}}$, $\alpha_0 = 1, \gamma_0 = L + \mu$. **2.** For $k = 0, 1, \ldots$ perform: **2a.** $\mathbf{x}^{k+1} = (1-\alpha_k)\mathbf{y}^k + \alpha_k \mathbf{z}^k$. **2b.** $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} - \frac{1}{\gamma_k} G(\mathbf{x}^{k+1}, \theta_k)$. **2c.** $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{1}{\gamma_k \alpha_k + \mu}\left(\gamma_k(\mathbf{x}^{k+1} - \mathbf{y}^{k+1}) + \mu(\mathbf{z}^k - \mathbf{x}^{k+1})\right)$.

Accelerated stochastic gradient method I

Theorem (Convergence of AccSG with strong convexity [3])

Define $\lambda_k = \prod_{j=1}^{k}(1-\alpha_j)$ *and* $\lambda_0 = 1$. *Let*

1. f *is* μ*-strongly convex,*
2. $\mathbb{E}[\|\mathbf{z}^k - \mathbf{x}^\star\|^2] \leq D^2$,
3. $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k)\|^2] \leq M^2$.
4. $\gamma_k = L + \frac{\mu}{\lambda_{k-1}}$ *and* $\alpha_k = \sqrt{\lambda_{k-1} + \frac{\lambda_{k-1}^2}{4}} - \frac{\lambda_{k-1}}{2}$.

Then,

$$\mathbb{E}[f(\mathbf{y}^{k+1}) - f(\mathbf{x}^\star)] \leq \frac{2(L+\mu)D^2}{k^2} + \frac{6M^2}{\mu k}.$$

The accelerated technique can be used to reduce the error term related to $\mathbb{E}[\|\mathbf{z}^k - \mathbf{x}^\star\|^2]$.

Accelerated stochastic gradient method II

Accelerated stochastic gradient method (AccSG)

1. Choose $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{0}$, $(\gamma_k)_{k\in\mathbb{N}}$, $(\alpha_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$, $\alpha_0 = 1, \gamma_0 = L$.
2. For $k = 0, 1, \dots$ perform:

2a. $\mathbf{x}^{k+1} = (1-\alpha_k)\mathbf{y}^k + \alpha_k \mathbf{z}^k$.
2b. $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} - \frac{1}{\gamma_k} G(\mathbf{x}^{k+1}, \theta_k)$.
2c. $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{1}{\alpha_k}(\mathbf{x}^{k+1} - \mathbf{y}^{k+1})$.

Theorem (Convergence of AccSG without strong convexity [3])

Let:

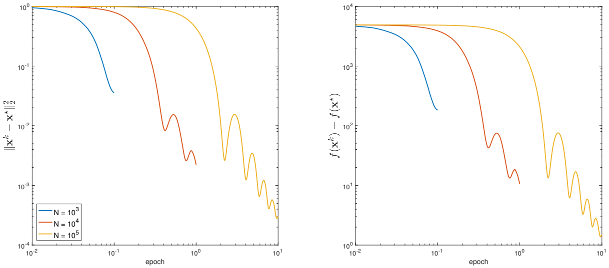
1. $\mathbb{E}[\|\mathbf{z}^k - \mathbf{x}^\star\|^2] \leq D^2$,
2. $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k) - \nabla f(\mathbf{x}^k)\|^2] \leq M^2$,
3. $\gamma_k = c(k+1)^{3/2} + L$ *for a fixed* $c > 0$, *and* $\alpha_k = 2/(k+2)$.

Then,

$$\mathbb{E}[f(\mathbf{y}^{k+1}) - f(\mathbf{x}^\star)] \leq \frac{3D^2 L}{k^2} + \left(3D^2 c + \frac{5M^2}{3c}\right)\frac{1}{\sqrt{k}}.$$

Example: AccSG

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$



Synthetic problem setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 10 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise. SNR is 30dB.
- $\gamma_k = c_0(N+1)^{3/2} + L$, where N is the number of total iterations.

Convex optimization with finite sums

Problem (Convex optimization with finite sums)

We consider the following simple example in the next few slides:

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\}$$

- f_j *is proper, closed, and convex.*
- ∇f_j *is* L_j*-Lipschitz continuous for* $j = 1, \ldots, n$.
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in \text{dom}(f) : f(\mathbf{x}^\star) = f^\star\}$ *is nonempty.*

- One prevalent choice is given by

$$G(\mathbf{x}^k, i_k) = \nabla f_{i_k}(\mathbf{x}^k), \qquad i_k \text{ is uniformly distributed over } \{1, 2, \cdots, n\}$$

An observation of SG

Lemma B

Assume f is Lipschitz smooth with constant L and $\{\mathbf{x}^k\}$ is generated by SG. Then,

$$\mathbb{E}[f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)] \leq (\gamma_k^2 L - \gamma_k)\mathbb{E}[\|\nabla f(\mathbf{x}^k)\|^2] + L\gamma_k^2 \mathbb{E}[\|G(\mathbf{x}^k, i_k) - \nabla f(\mathbf{x}^k)\|^2]$$

An observation of SG

Lemma B

Assume f is Lipschitz smooth with constant L and $\{\mathbf{x}^k\}$ is generated by SG. Then,

$$\mathbb{E}[f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)] \leq (\gamma_k^2 L - \gamma_k)\mathbb{E}[\|\nabla f(\mathbf{x}^k)\|^2] + L\gamma_k^2\mathbb{E}[\|G(\mathbf{x}^k, i_k) - \nabla f(\mathbf{x}^k)\|^2]$$

***Proof.** From the smoothness of f,

$$\langle \nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k\rangle \leq L\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2,$$

and by the convexity of f,

$$\mathbb{E}[f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)] \leq \mathbb{E}[\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^k\rangle].$$

Thus,

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)] &\leq \mathbb{E}[\langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k\rangle] + L\mathbb{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2] \\
&= -\gamma_k\mathbb{E}[\|\nabla f(\mathbf{x}^k)\|^2] + L\gamma_k^2\mathbb{E}[\|G(\mathbf{x}^k, i_k)\|^2] \\
&= (\gamma_k^2 L - \gamma_k)\mathbb{E}[\|\nabla f(\mathbf{x}^k)\|^2] + L\gamma_k^2\mathbb{E}[\|G(\mathbf{x}^k, i_k) - \nabla f(\mathbf{x}^k)\|^2].
\end{aligned}$$

An observation of SG

Lemma B

Assume f is Lipschitz smooth with constant L and $\{\mathbf{x}^k\}$ is generated by SG. Then,

$$\mathbb{E}[f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)] \leq (\gamma_k^2 L - \gamma_k)\mathbb{E}[\|\nabla f(\mathbf{x}^k)\|^2] + L\gamma_k^2 \mathbb{E}[\|G(\mathbf{x}^k, i_k) - \nabla f(\mathbf{x}^k)\|^2]$$

- Here, $G(\mathbf{x}^k, i_k) = \nabla f_{i_k}$.
- The first term dominates at the beginning, and the variance in gradient will dominate later (as if $\nabla f(\mathbf{x}^k) \to 0$).
- To ensure convergence, $\gamma_k \to 0$. $\Longrightarrow$ Slow convergence!

Can we decrease the variance while using a constant step-size?

- Choose a stochastic gradient, s.t. $\mathbb{E}\left[\|G(\mathbf{x}^k; i_k)\|^2\right] \to 0$.

Variance reduction techniques: SVRG

- Select the stochastic gradient ∇f_{i_k}, and computes a gradient estimate

$$\mathbf{r}_k = \nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\tilde{\mathbf{x}}) + \nabla f(\tilde{\mathbf{x}}),$$

where $\tilde{\mathbf{x}}$ is a good approximation of $\mathbf{x}^\star$. As $\tilde{\mathbf{x}} \to \mathbf{x}^\star$ and $\mathbf{x}^k \to \mathbf{x}^\star$,

$$\nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\tilde{\mathbf{x}}) + \nabla f(\tilde{\mathbf{x}}) \to 0.$$

Therefore,

$$\mathbb{E}\left[\|\nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\tilde{\mathbf{x}}) + \nabla f(\tilde{\mathbf{x}})\|^2\right] \to 0.$$

Stochastic gradient algorithm with variance reduction

Stochastic gradient algorithm with variance reduction (SVRG) [10, 5]

1. Choose $\widetilde{\mathbf{x}}^0 \in \mathbb{R}^p$ as a starting point and $\gamma > 0$ and $q \in \mathbb{N}_+$.
2. For $s = 0, 1, 2 \cdots$, perform:
   - **2a**. $\widetilde{\mathbf{x}} = \widetilde{\mathbf{x}}^s, \quad \widetilde{\mathbf{v}} = \nabla f(\widetilde{\mathbf{x}}), \quad \mathbf{x}^0 = \widetilde{\mathbf{x}}$.
   - **2b**. For $k = 0, 1, \cdots q-1$, perform:

$$\begin{cases} \text{Pick } i_k \in \{1, \ldots, n\} \text{ uniformly at random} \\ \mathbf{r}_k = \nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}) + \widetilde{\mathbf{v}} \\ \mathbf{x}^{k+1} := \mathbf{x}^k - \gamma \mathbf{r}_k, \end{cases} \tag{1}$$

   - **2c**. Update $\widetilde{\mathbf{x}}^{s+1} = \frac{1}{m} \sum_{j=0}^{q-1} \mathbf{x}^j$.

Common features

- The SVRG method uses a multistage scheme to reduce the variance of the stochastic gradient $\mathbf{r}_k$ where $\mathbf{x}^k$ and $\widetilde{\mathbf{x}}^s$ tend to $\mathbf{x}_\star$.
- Learning rate γ is not necessarily tend to 0.
- Each stage, SVRG uses $n + 2q$ component gradient evaluations: n for the full gradient at the beginning of each stage, and $2q$ for each of the q stochastic gradient steps.

Convergence analysis

Assumption A5.

(i) f is μ-strongly convex

(ii) The learning rate $0 < \gamma < 1/(4L_{\max})$, where $L_{\max} = \max_{1 \leq j \leq n} L_j$.

(iii) q is large enough such that

$$\kappa = \frac{1}{\mu\gamma(1 - 4\gamma L_{\max})q} + \frac{4\gamma L_{\max}(q+1)}{(1 - 4\gamma L_{\max})q} < 1.$$

Theorem

Assumptions:

- *The sequence $\{\widetilde{\mathbf{x}^s}\}_{k \geq 0}$ is generated by SVRG.*
- *Assumption A5 is satisfied.*

Conclusion: *Linear convergence is obtained:*

$$\mathbb{E} f(\widetilde{\mathbf{x}^s}) - f(\mathbf{x}^\star) \leq \kappa^s (f(\widetilde{\mathbf{x}^0}) - f(\mathbf{x}^\star)).$$

Choice of γ and q, and complexity

Chose γ and q such that $\kappa \in (0,1)$:

For example
$$\gamma = 0.1/L_{\max}, q = 100(L_{\max}/\mu) \Longrightarrow \kappa \approx 5/6.$$

Complexity

$$\mathbb{E}f(\widetilde{\mathbf{x}^s}) - f(\mathbf{x}^\star) \leq \varepsilon, \quad \text{when } s \geq \log((f(\widetilde{\mathbf{x}^0}) - f(\mathbf{x}^\star))/\epsilon)/\log(\kappa^{-1})$$

Since at each stage needs $n + 2q$ component gradient evaluations, with $q = \mathcal{O}(L_{\max}/\mu)$, we get the overall complexity is

$$\mathcal{O}\Big((n + L_{\max}/\mu)\log(1/\epsilon)\Big).$$

Taxonomy of algorithms

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\}.$$

- $f(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x})$: μ-strongly convex with L-Lipschitz continuous gradient.

Gradient descent	Acc. MB SVRG	SVRG/SAGA/SARAH	SGM
Linear	Linear	Linear	Sublinear

Table: Rate of convergence.

- $\kappa = L/\mu$ and $s_0 = 8\sqrt{\kappa}n(\sqrt{2}\alpha(n-1) + 8\sqrt{\kappa})^{-1}$ for $0 < \alpha \leq 1/8$.

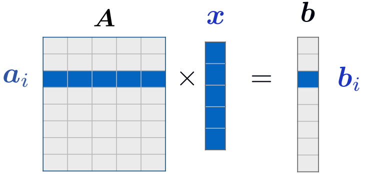
SVRG/SAGA/SARAH	Acc. MB SVRG $s < \lceil s_0 \rceil$	AccGrad
$\mathcal{O}((n+\kappa)\log(1/\varepsilon))$	$\mathcal{O}((n+\kappa\frac{n-s}{n-1})\log(1/\varepsilon))$	$\mathcal{O}((n\kappa)\log(1/\varepsilon))$

Table: Complexity to obtain ε-solution.

SAGA/SARAH/AccMBSVRG can be found in the next few slides.

*Another way of parsing data

Example (Least squares): $\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$



Using a subset of rows

We have mainly focused on using a subset of rows instead of the full data at each iteration.
This way, we compute an unbiased estimate $G(\mathbf{x}^k, i_k)$ of the gradient using

- a subset of data points: $(\mathbf{a}_{i_k}, b_{i_k})$,
- and the whole decision variable: $\mathbf{x}^k$:

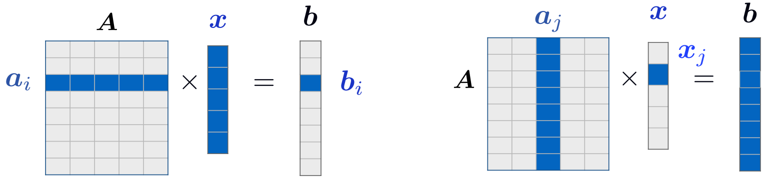
$$G(\mathbf{x}^k, i_k) = \mathbf{a}_{i_k}^T(\langle \mathbf{a}_{i_k}^T, \mathbf{x}\rangle - \mathbf{b}_{i_k}).$$

Estimate $G(\mathbf{x}^k, i_k)$ is dense, so we update the whole decision variable.

Next: Using a subset of columns.

⋆Another way of parsing data

Example (Least squares): $\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$



Using a subset of columns

Denote the standard basis vectors by $\mathbf{e}_i$, and the corresponding directional derivatives by ∇_i. Let $\mathbf{a}_i$ represent the ith column of matrix $\mathbf{A}$. Consider the following unbiased estimate:

$$G(\mathbf{x}^k, i_k) = p\nabla_{i_k} f(\mathbf{x}^k)\mathbf{e}_{i_k} = p\langle \mathbf{a}_{i_k}, \mathbf{a}_{i_k}\mathbf{x}^k_{i_k} - \mathbf{b}\rangle \mathbf{e}_{i_k}.$$

This way, we compute an unbiased estimate $G(\mathbf{x}^k, i_k)$ of the gradient using

- a subset of columns $(\mathbf{a}_{i_k})$ and the whole measurement vector $\mathbf{b}$,
- and only the chosen coordinates of decision variable: $\mathbf{x}^k_{i_k}$.

Estimate $G(\mathbf{x}^k, i_k)$ is sparse, only coordinates chosen by i_k are nonzero. Hence, we update these coordinates only.

*Variance reduction techniques: SAGA

Stochastic Average Gradient (SAGA) [6]
1a. Choose $\tilde{\mathbf{x}}_i^0 = \mathbf{x}^0 \in \mathbb{R}^p, \forall i,\ q \in \mathbb{N}_+$ and stepsize $\gamma > 0$. **1b.** Store $\nabla f_i(\tilde{\mathbf{x}}_i^0)$ in a table data-structure with length n. **2.** For $k = 0, 1 \ldots$ perform: **2a.** pick $i_k \in \{1, \ldots, n\}$ uniformly at random **2b.** Take $\tilde{\mathbf{x}}_{i_k}^{k+1} = \mathbf{x}^k$, store $\nabla f_{i_k}(\tilde{\mathbf{x}}_{i_k}^{k+1})$ in the table and leave other entries the same. **2c.** $\mathbf{r}_k = \nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\tilde{\mathbf{x}}_{i_k}^k) + \frac{1}{n}\sum_{j=1}^n \nabla f_j(\tilde{\mathbf{x}}_j^k)$ **3.** $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma \mathbf{r}_k$

Recipe:

In each iteration:

- Store last gradient evaluated at each datapoint.
- Previous gradient for datapoint j is $\nabla f_j(\tilde{\mathbf{x}}_j^k)$.
- Perform SG-iterations with the following stochastic gradient

$$\mathbf{r}_k = \nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\tilde{\mathbf{x}}_{i_k}^k) + \frac{1}{n}\sum_{j=1}^n \nabla f_j(\tilde{\mathbf{x}}_j^k).$$

*Variance reduction techniques: SAGA

- Select the stochastic gradient $\mathbf{r}_k$ as

$$\mathbf{r}_k = \nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\tilde{\mathbf{x}}_{i_k}^k) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\tilde{\mathbf{x}}_j^k),$$

where, at each iteration, $\tilde{\mathbf{x}}$ is updated as $\tilde{\mathbf{x}}_{i_k}^k = \mathbf{x}^k$ and $\tilde{\mathbf{x}}_j^k$ stays the same for $j \neq i_k$.
As $\tilde{\mathbf{x}}_j^k \to \mathbf{x}^\star$ and $\mathbf{x}^k \to \mathbf{x}^\star$,

$$\nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\tilde{\mathbf{x}}_{i_k}^k) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\tilde{\mathbf{x}}_j^k) \to 0.$$

Therefore,

$$\mathbb{E}\left[\|\nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\tilde{\mathbf{x}}_{i_k}^k) + \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\tilde{\mathbf{x}}_j^k)\|^2\right] \to 0.$$

*Convergence of SAGA

$$\boxed{f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\}.}$$

Theorem (Convergence of SAGA [6])

Suppose that f *is* μ*-strongly convex and that the stepsize is* $\gamma = \frac{1}{2(\mu n + L)}$ *with*

$$\rho = 1 - \frac{\mu}{2(\mu n + L)} < 1,$$

$$C = \|\mathbf{x}^0 - \mathbf{x}^\star\|^2 + \frac{n}{\mu n + L}[f(\mathbf{x}^0) - \langle \nabla f(\mathbf{x}^\star), \mathbf{x}^0 - \mathbf{x}^\star \rangle - f(\mathbf{x}^\star)]$$

Then

$$\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|^2] \leq \rho^k C.$$

- Allows the constant step-size.
- Obtains linear rate convergence.

⋆Variance reduction techniques: SARAH

- Select the stochastic gradient $\mathbf{r}_k$

$$\mathbf{r}_k = \nabla f_{i_k}(\mathbf{x}^k) - \nabla f_{i_k}(\mathbf{x}^{k-1}) + \mathbf{r}_{k-1},$$

The variance reduction in SARAH can be characterized as

$$\mathbb{E}[\|\mathbf{r}_k\|^2] \le \left[1 - \left(\frac{2}{\gamma L} - 1\right)\mu^2\gamma^2\right]^k \mathbb{E}[\|\nabla f(\mathbf{x}^0)\|^2].$$

*Variance reduction techniques: SARAH

Stochastic Recursive Gradient Algorithm (SARAH) [7]

1. Choose $\overline{\mathbf{x}}^0 \in \mathbb{R}^p$, $q \in \mathbb{N}_+$ and stepsize $\gamma > 0$.

2. For $k = 0, 1 \ldots$ perform:

2. $\mathbf{x}^0 = \overline{\mathbf{x}}^k$, $\mathbf{r}_0 = \frac{1}{n}\sum_{j=1}^n f_j(\overline{\mathbf{x}}^0)$

2a. $\mathbf{x}^1 = \mathbf{x}^0 - \gamma \mathbf{r}_0$

2b. For $l = 1 \ldots, q-1$, perform:

$$\begin{cases} \text{pick } i_l \in \{1, \ldots, n\} \text{ uniformly at random,} \\ \mathbf{r}_l = \nabla f_{i_l}(\mathbf{x}^l) - \nabla f_{i_l}(\mathbf{x}^{l-1}) + \mathbf{r}_{l-1}, \\ \mathbf{x}^{l+1} = \mathbf{x}^l - \gamma \mathbf{r}_l. \end{cases}$$

3 Update $\overline{\mathbf{x}}^{k+1} = \mathbf{x}^l$ where l is chosen uniformly at random from $\{0, \ldots, q\}$.

Recipe:

In a cycle of q inner iterations:

- Compute stochastic step direction by recursively adding and subtracting component gradients to and from the previous direction.

$$\mathbf{r}_l = \nabla f_{i_l}(\mathbf{x}^l) - \nabla f_{i_l}(\mathbf{x}^{l-1}) + \mathbf{r}_{l-1}.$$

- Perform q SG-iterations with $\mathbf{r}_l$.
- Update next iteration by picking uniformly at random from q previous iterations.

*Convergence of SARAH

$$\boxed{f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\}.}$$

Theorem (Convergence of SARAH [7])

Suppose that f is μ-strongly convex and that the stepsize γ and number of inner iterations q satisfies

$$\rho_q = \frac{1}{\mu\gamma(1+q)} + \frac{L_{max}\gamma}{2 - L_{max}\gamma} < 1.$$

Then

$$\mathbb{E}[\|\nabla f(\overline{\mathbf{x}}^k)\|^2] \leq \rho_q^k \|\nabla f(\overline{\mathbf{x}}^0)\|^2.$$

*Variance reduction techniques: Mini-batch variance reduction

Accelerated mini-batch SVRG (Acc. MB SVRG)
1. Choose $q \in \mathbb{N}_+$, initialization $\overline{\mathbf{x}}^0 \in \mathbb{R}^p$, stepsize $\gamma > 0$, accelerated stepsize $\beta = (1 - \sqrt{\mu\gamma})/(1 + \sqrt{\mu\gamma})$.
2. For $k = 0, 1, \ldots$ perform:
2a. $\bar{\mathbf{x}} = \bar{\mathbf{x}}^k$, $\mathbf{x}^0 = \mathbf{y}^1 = \overline{\mathbf{x}}$; $\nabla f(\overline{\mathbf{x}}) = \frac{1}{n}\sum_{j=1}^n \nabla f_j(\overline{\mathbf{x}})$.
2b. For $l = 0, 1, \ldots, q-1$, perform:
$\begin{cases} \text{pick } I_l \subset \{1, \ldots, n\}\text{: mini-batch of size} s, \\ \mathbf{r}_l = \nabla f_{I_l}(\mathbf{y}_l) - \nabla f_{I_l}(\overline{\mathbf{x}}) + \nabla f(\overline{\mathbf{x}}), \\ \mathbf{x}^{l+1} = \mathbf{y}^l - \gamma \mathbf{r}_l \\ \mathbf{y}^{l+1} = \mathbf{x}^{l+1} + \beta(\mathbf{x}^{l+1} - \mathbf{x}^l). \end{cases}$
3. Update $\overline{\mathbf{x}}^{k+1} = \mathbf{x}^q$.

- A mini-batch of size s is indexed by $I = \{i_1, \ldots, i_s\}$, where each $i_j \in \{1, \ldots, n\}$ is chosen uniformly at random, and

$$f_I = \frac{1}{s}\sum_{j=1}^{s} f_{i_j}.$$

- s components are chosen instead of one + an accelerated step.

⋆Convergence of Acc. MB SVRG

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\}$$

Theorem (Convergence of Acc. MB SVRG [4])

Suppose that:

1. $0 < \gamma \leq \gamma_{\max} = \min\left\{ \frac{(\alpha q)^2 (n-1)^2 \mu}{64(n-s)^2 L_{max}^2}, \frac{1}{2L_{max}} \right\}$ *for some* $0 < \alpha < 1/8$.
2. $q \geq \frac{1}{(1-\alpha)\sqrt{\mu\gamma}} \log \frac{1-\alpha}{\alpha}$.

Then,

$$\mathbb{E}[f(\overline{\mathbf{x}}^k) - f^\star] \leq \rho^k (f(\overline{\mathbf{x}}^0) - f^\star),$$

where $\rho = 2\alpha(2+\alpha)/(1-\alpha) < 1$.

References I

[1] V. N. Vapnik.
An overview of statistical learning theory.
IEEE Trans. Inf. Theory, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[2] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro.
Robust stochastic approximation approach to stochastic programming.
SIAM J. Optim., vol. 19, pp. 1574–1609, 2008.

[3] J. T. Kwok, C. Hu and W. Pan.
Accelerated gradient methods for stochastic optimization and online learning.
Advances in Neural Information Processing Systems, vol. 22, pp. 781–789, 2009.

[4] A. Nitanda.
Stochastic proximal gradient descent with acceleration techniques.
Advances in Neural Information Processing Systems, pp. 1574–1582, 2014.

[5] L. Xiao, and T. Zhang.
A proximal stochastic gradient method with progressive variance reduction.
SIAM Journal on Optimization 2057-2075, 2014.

[6] A. Defazio, F. Bach, and S. Lacoste-Julien.
SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives.
Advances in Neural Information Processing Systems, pp. 1646–1654, 2014.

References II

[7] L. Nguyen, J. Liu, K. Scheinberg, and M. Takac.
SARAH: A novel method for machine learning problems using stochastic recursive gradient.
International Conference on Machine Learning, 2017.

[8] S., Ohad, and T. Zhang.
Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.
International Conference on Machine Learning, 2013.

[9] S.-S., Shai, et al.
Pegasos: Primal estimated sub-gradient solver for svm.
Mathematical Programming 127.1 (2011): 3-30.

[10] R. Johnson, and T. Zhang.
Accelerating stochastic gradient descent using predictive variance reduction.
Advances in neural information processing systems 315–323, 2013

[11] J. Duchi, E. Hazan, and Y. Singer.
Adaptive subgradient methods for online learning and stochastic optimization.
Journal of Machine Learning Research 12, 2121-2159.

*Proof: A Basic Lemma for SG

Lemma A

Let f be μ-strongly convex ($\mu \geq 0$) and $\mathbb{E}\|G(\mathbf{x}^k, \theta_k)\|^2 \leq M^2$. For all *fixed* $\mathbf{x} \in \mathbb{R}^p$,

$$\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 \leq (1 - \gamma_k \mu)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}\|^2 - 2\gamma_k \mathbb{E}\left(f(\mathbf{x}^k) - f(\mathbf{x})\right) + \gamma_k^2 M^2.$$

*Proof: A Basic Lemma for SG

Lemma A

Let f be μ-strongly convex $(\mu \geq 0)$ and $\mathbb{E}\|G(\mathbf{x}^k, \theta_k)\|^2 \leq M^2$. For all *fixed* $\mathbf{x} \in \mathbb{R}^p$,

$$\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 \leq (1 - \gamma_k \mu)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}\|^2 - 2\gamma_k \mathbb{E}\left(f(\mathbf{x}^k) - f(\mathbf{x})\right) + \gamma_k^2 M^2.$$

- $\mu = 0$ corresponds to the non-strongly convex case.
- This lemma will be used several times in this lecture.

*Proof: A Basic Lemma for SG

Lemma A

Let f be μ-strongly convex ($\mu \geq 0$) and $\mathbb{E}\|G(\mathbf{x}^k, \theta_k)\|^2 \leq M^2$. For all *fixed* $\mathbf{x} \in \mathbb{R}^p$,

$$\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 \leq (1 - \gamma_k \mu)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}\|^2 - 2\gamma_k \mathbb{E}\left(f(\mathbf{x}^k) - f(\mathbf{x})\right) + \gamma_k^2 M^2.$$

Proof of Lemma A. According to the iterative relationship, and expanding the inner product,

$$\begin{aligned}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 &= \|\mathbf{x}^k - \gamma_k G(\mathbf{x}^k, \theta_k) - \mathbf{x}\|^2 \\ &= \|\mathbf{x}^k - \mathbf{x}\|^2 - 2\gamma_k \langle G(\mathbf{x}^k, \theta_k), \mathbf{x}^k - \mathbf{x}\rangle + \gamma_k^2 \|G(\mathbf{x}^k, \theta_k)\|^2.\end{aligned}$$

Noting that $\mathbf{x}_k$ is independent from θ_k, thus $\mathbb{E}_{\theta_k}[G(\mathbf{x}^k, \theta_k)] = \nabla f(\mathbf{x}^k)$. Taking the expectation with respect to the random variable θ_k on both sides,

$$\mathbb{E}_{\theta_k}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 = \|\mathbf{x}^k - \mathbf{x}\|^2 - 2\gamma_k \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}\rangle + \gamma_k^2 \mathbb{E}_{\theta_k}\|G(\mathbf{x}^k, \theta_k)\|^2.$$

Using the strong convexity of f which implies

$$\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}\rangle \geq f(\mathbf{x}^k) - f(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}\|^2,$$

$$\mathbb{E}_{\theta_k}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 \leq (1 - \gamma_k \mu)\|\mathbf{x}^k - \mathbf{x}\|^2 - 2\gamma_k \left(f(\mathbf{x}^k) - f(\mathbf{x})\right) + \gamma_k^2 \mathbb{E}_{\theta_k}\|G(\mathbf{x}^k, \theta_k)\|^2.$$

*Proof for Slide 14

Since f is μ-strongly convex,

$$f(\mathbf{x}^k) - f(\mathbf{x}^\star) + \langle \nabla f(\mathbf{x}^\star), \mathbf{x}^\star - \mathbf{x}^k \rangle \geq \frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}^\star\|^2.$$

Moreover, as $\mathbf{x}^\star$ is a minimizer, we have $\nabla f(\mathbf{x}^\star) = 0$ and as a result,

$$f(\mathbf{x}^k) - f(\mathbf{x}^\star) \geq \frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}^\star\|^2.$$

Applying the above and Lemma A, one can easily show that

$$\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2 \leq (1 - 2\gamma_k\mu)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^\star\|^2 + \gamma_k^2 M^2.$$

Introducing with $\gamma_k = \frac{\gamma_0}{k+1}$,

$$\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^\star\|^2 \leq (1 - 2\gamma_0\mu k^{-1})\mathbb{E}\|\mathbf{x}^{k-1} - \mathbf{x}^\star\|^2 + \gamma_0^2 M^2 k^{-2}.$$

By an inductive argument, one can prove the theorem.

⋆Proof for Slide 17

By Lemma A,

$$2\gamma_k \mathbb{E}\left(f(\mathbf{x}^k) - f(\mathbf{x})\right) \leq (1-\gamma_k\mu)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}\|^2 - \mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 + \gamma_k^2 M^2.$$

Dividing both sides by γ_k, and then introducing with $\gamma_k = \gamma_0/k$ ($\gamma_0 \geq 1/\mu$),

$$\begin{aligned} 2\mathbb{E}\left(f(\mathbf{x}^k) - f(\mathbf{x})\right) &\leq \left(\frac{k}{\gamma_0} - \mu\right)\mathbb{E}\|\mathbf{x}^k - \mathbf{x}\|^2 - \frac{k}{\gamma_0}\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 + \frac{\gamma_0}{k}M^2 \\ &\leq \frac{k-1}{\gamma_0}\mathbb{E}\|\mathbf{x}^k - \mathbf{x}\|^2 - \frac{k}{\gamma_0}\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2 + \frac{\gamma_0}{k}M^2. \end{aligned}$$

Summing up over $k = 1, 2, \cdots, t$,

$$2\mathbb{E}\sum_{k=1}^{t}\left(f(\mathbf{x}^k) - f(\mathbf{x})\right) \leq \sum_{k=1}^{t}\frac{\gamma_0}{k}M^2 \leq \gamma_0 M^2(1+\log t).$$

Dividing both sides by $2t$, and using the convexity of f which implies

$$\frac{1}{t}\sum_{k=1}^{t} f(\mathbf{x}^k) \geq f\left(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}^k\right) = f(\bar{\mathbf{x}}^t),$$

one can get the desired results.

*Proof for Slide 18

Applying Lemma A, we have

$$\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2 \leq \mathbb{E}\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - 2\gamma_k \mathbb{E}\left(f(\mathbf{x}^k) - f(\mathbf{x}^\star)\right) + \gamma_k^2 M^2.$$

Rearranging terms,

$$2\gamma_k \mathbb{E}\left(f(\mathbf{x}^k) - f(\mathbf{x}^\star)\right) \leq \mathbb{E}\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2 + \gamma_k^2 M^2.$$

Summing up over $k = 1, \cdots, t$,

$$2\mathbb{E}\sum_{k=1}^{t} \gamma_k \left(f(\mathbf{x}^k) - f(\mathbf{x}^\star)\right) \leq \|\mathbf{x}^1 - \mathbf{x}^\star\|^2 - \mathbb{E}\|\mathbf{x}^{t+1} - \mathbf{x}^\star\|^2 + \sum_{k=1}^{t} \gamma_k^2 M^2.$$

Dividing both sides by $2\sum_{k=1}^t \gamma_k$, and noting that the convexity of f implies

$$\frac{\sum_{k=1}^t \gamma_k f(\mathbf{x}^k)}{\sum_{k=1}^t \gamma_k} \geq f\left(\frac{\sum_{k=1}^t \gamma_k \mathbf{x}^k}{\sum_{j=1}^t \gamma_j}\right) = f(\bar{\mathbf{x}}),$$

we get

$$\mathbb{E}\left(f(\mathbf{x}^k) - f(\mathbf{x}^\star)\right) \leq \frac{\|\mathbf{x}^1 - \mathbf{x}^\star\|^2}{2\sum_{k=1}^t \gamma_k} + \frac{\sum_{k=1}^t \gamma_k^2 M^2}{2\sum_{k=1}^t \gamma_k}.$$