Stochastic Gradient Descent

Pontus Giselsson

Outline

- **Stochastic gradient method**
- Nonconvex setting
- Convex setting
- Step-sizes and rates
- Refined step-size and rate analysis
- Rate comparison to proximal gradient method
- Stochastic gradient descent variations

Proximal gradient method

- Proximal gradient method is applied problems of the form

$$\underset{x}{\text{minimize}}\, f(x) + g(x)$$

  where, for instance:
  - $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth (not necessarily convex)
  - $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed convex
- For large problems, gradient can be expensive to compute
  $\Rightarrow$ replace by unbiased stochastic approximation of gradient

Unbiased stochastic gradient approximation

- Stochastic gradient *estimator*:
  - notation: $\widehat{\nabla} f(x)$
  - outputs random vector in $\mathbb{R}^n$ for each $x \in \mathbb{R}^n$
- Stochastic gradient *realization*:
  - notation: $\widetilde{\nabla} f(x) : \mathbb{R}^n \to \mathbb{R}^n$
  - outputs, $\forall x \in \mathbb{R}^n$, vector in $\mathbb{R}^n$ drawn from distribution of $\widehat{\nabla} f(x)$
- An unbiased stochastic gradient estimator $\widehat{\nabla} f$ satisfies $\forall x \in \mathbb{R}^n$:

$$\mathbb{E}\widehat{\nabla} f(x) = \nabla f(x)$$

- If x is random vector in $\mathbb{R}^n$, unbiased estimator satisfies

$$\mathbb{E}[\widehat{\nabla} f(x)|x] = \nabla f(x)$$

(both are random vectors in $\mathbb{R}^n$)

Stochastic gradient descent (SGD)

- The following iteration generates $(x_k)_{k\in\mathbb{N}}$ of *random* variables:

$$x_{k+1} = \mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \widehat{\nabla} f(x_k))$$

  since $\widehat{\nabla} f$ outputs random vectors in $\mathbb{R}^n$
- Stochastic gradient descent finds a *realization* of this sequence:

$$x_{k+1} = \mathrm{prox}_{\gamma_k g}(x_k - \gamma_k \widetilde{\nabla} f(x_k))$$

  where $(x_k)_{k\in\mathbb{N}}$ here is a realization with values in $\mathbb{R}^n$
- Sloppy in notation for when x_k is *random variable* vs *realization*
- Can be efficient if evaluating $\widetilde{\nabla} f$ much cheaper than ∇f

Stochastic gradients – Finite sum problems

- Consider *finite sum problems* of the form

$$\underset{x}{\text{minimize}}\, \underbrace{\frac{1}{N}\left(\sum_{i=1}^{N} f_i(x)\right)}_{f(x)} + g(x)$$

  where $\frac{1}{N}$ is for convenience
- Training problems of this form, where sum over training data
- Stochastic gradient: select f_i at random and take gradient step

Single function stochastic gradient

- Let I be a $\{1, \ldots, N\}$-valued random variable
- Let, as before, $\widehat{\nabla} f$ denote the stochastic gradient estimator
- Realization: let i be drawn from probability distribution of I

$$\widetilde{\nabla} f(x) = \nabla f_i(x)$$

where we will use uniform probability distribution

$$p_i = p(I = i) = \tfrac{1}{N}$$

- Stochastic gradient is unbiased:

$$\mathbb{E}[\widehat{\nabla} f(x)] = \sum_{i=1}^{N} p_i \nabla f_i(x) = \tfrac{1}{N} \sum_{i=1}^{N} \nabla f_i(x) = \nabla f(x)$$

Mini-batch stochastic gradient

- Let $\mathcal{B}$ be set of K-sample mini-batches to choose from:
  - Example: 2-sample mini-batches and $N = 4$:

$$\mathcal{B} = \{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}\}$$

  - Number of mini batches $\binom{N}{K}$, each item in $\binom{N-1}{K-1}$ batches
- Let $\mathbb{B}$ be $\mathcal{B}$-valued random variable
- Let, as before, $\widehat{\nabla} f$ denote stochastic gradient estimator
- Realization: let B be drawn from probability distribution of $\mathbb{B}$

$$\widetilde{\nabla} f(x) = \frac{1}{K} \sum_{i \in B} \nabla f_i(x)$$

where we will use uniform probability distribution

$$p_B = p(\mathbb{B} = B) = \frac{1}{\binom{N}{K}}$$

- Stochastic gradient is unbiased:

$$\mathbb{E}\widehat{\nabla} f(x) = \frac{1}{\binom{N}{K}} \sum_{B \in \mathcal{B}} \frac{1}{K} \sum_{i \in B} \nabla f_i(x) = \frac{\binom{N-1}{K-1}}{\binom{N}{K} K} \sum_{i=1}^{N} \nabla f_i(x) = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x) = \nabla f(x)$$

Stochastic gradient descent for finite sum problems

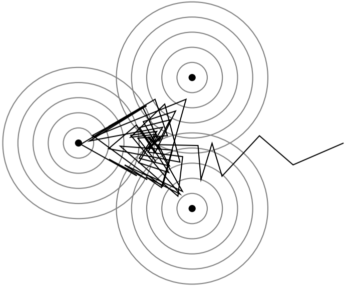
- The algorithm, choose $x_0 \in \mathbb{R}^n$ and iterate:
  1. Sample a mini-batch $B_k \in \mathcal{B}$ of K indices uniformly
  2. Update

$$x_{k+1} = \mathrm{prox}_{\gamma_k g}(x_k - \tfrac{\gamma_k}{K} \sum_{j \in B_k} \nabla f_j(x_k))$$

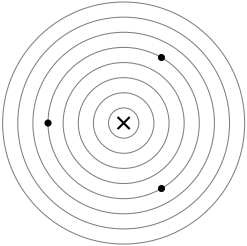
- Can have $\mathcal{B} = \{\{1\}, \ldots, \{N\}\}$ and sample only one function
- Gives realization of underlying stochastic process
- How about convergence?

SGD – Example

- Let $c_1 + c_2 + c_3 = 0$
- Solve $\text{minimize}_x(\frac{1}{2}(\|x - c_1\|_2^2 + \|x - c_2\|_2^2 + \|x - c_3\|_2^2) = \frac{3}{2}\|x\|_2^2 + c$
- Stochastic gradient method with $\gamma_k = 1/3$



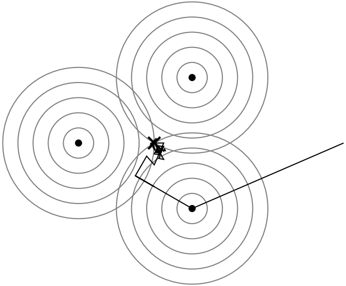
Levelsets of summands



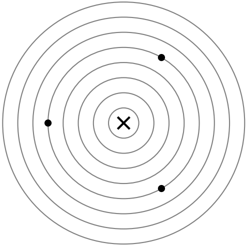
Levelset of sum

SGD – Example

- Let $c_1 + c_2 + c_3 = 0$
- Solve $\text{minimize}_x(\frac{1}{2}(\|x - c_1\|_2^2 + \|x - c_2\|_2^2 + \|x - c_3\|_2^2) = \frac{3}{2}\|x\|_2^2 + c$
- Stochastic gradient method with $\gamma_k = 1/k$



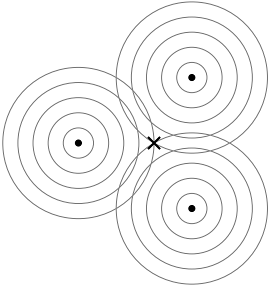
Levelsets of summands



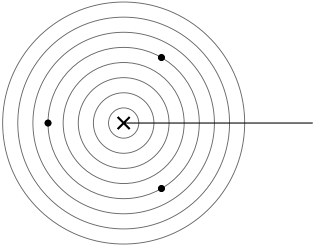
Levelset of sum

SGD – Example

- Let $c_1 + c_2 + c_3 = 0$
- Solve $\text{minimize}_x(\frac{1}{2}(\|x - c_1\|_2^2 + \|x - c_2\|_2^2 + \|x - c_3\|_2^2) = \frac{3}{2}\|x\|_2^2 + c$
- Gradient method with $\gamma_k = 1/3$



Levelsets of summands



Levelset of sum

- SGD will not converge for constant steps (unlike gradient method)

Fixed step-size SGD does not converge to solution

- We can at most hope for finding point $\bar{x}$ such that

$$0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$$

  i.e., the proximal gradient fixed-point characterization
- Consider setting $g = 0$ and assume x_k such that $0 = \nabla f(x_k)$
  - That $0 = \nabla f(x_k)$ does *not* imply $0 = \nabla f_i(x_k)$ for all f_i, hence

$$x_{k+1} = x_k - \gamma_k \nabla f_i(x_k) \neq x_k$$

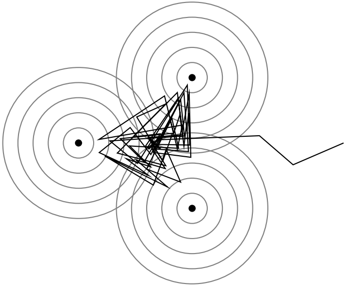
    i.e., will move away from prox-grad fixed-point for fixed $\gamma_k > 0$
  - Need diminishing step-size rule to hope for convergence

Last iterate vs best and average

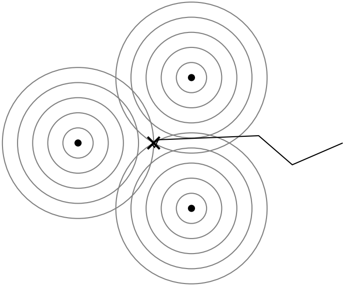
- Last iterate moves away from fixed-point
- Behavior can better for:
  - Best iterate (smallest function value)
  - Average iterate (Polyak-Ruppert averaging)

Best iterate sequence

- Output best (in function value) iterate instead of last iterate
- Example: SGD with constant steps and best iterate



SGD with constant step-size

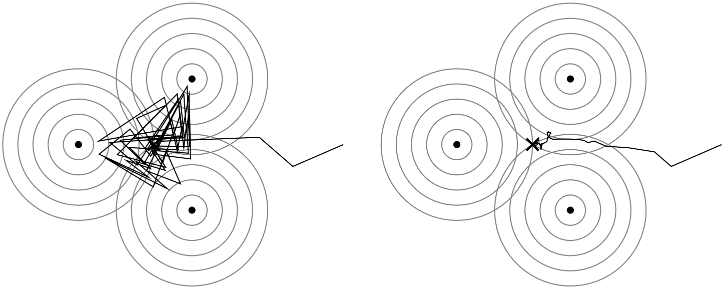


Best iterate in sequence

- Not usful in practice: Function value comparison too expensive

Polyak-Ruppert averaging

- Polyak-Ruppert averaging:
  - Output average of iterations instead of last iteration
- Example: SGD with constant steps and its average sequence



SGD with constant step-size   Average of SGD sequence

Rate outlook

- Sublinear convergence in:
  - Nonconvex and convex settings
  - Strongly convex setting (unlike proximal gradient method)
- Convergence rate dependent on step-size choice

Outline

- Stochastic gradient method
- **Nonconvex setting**
- Convex setting
- Step-sizes and rates
- Refined step-size and rate analysis
- Rate comparison to proximal gradient method
- Stochastic gradient descent variations

Stochastic gradient descent

- We consider problems of the form

$$\text{minimize}\, f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is not necessarily convex

- We will analyze stochastic gradient descent

$$x_{k+1} = x_k - \gamma_k \widehat{\nabla} f(x_k)$$

where $\widehat{\nabla} f(x_k)$ is an unbiased estimate of $\nabla f(x_k)$ for all x_k

- Will show sublinear convergence rates that depend on step-sizes

Nonconvex setting – Assumptions

(i)	$f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth, for all $x, y \in \mathbb{R}^n$: $$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \tfrac{\beta}{2}\|y - x\|_2^2$$
(ii)	Stochastic gradient of f is unbiased: $\mathbb{E}[\widehat{\nabla} f(x) \vert x] = \nabla f(x)$
(iii)	Variance is bounded: $\mathbb{E}[\|\widehat{\nabla} f(x)\|_2^2 \vert x] \leq \|\nabla f(x)\|_2^2 + M^2$
(iv)	No nonsmooth term, i.e., $g = 0$
(v)	A minimizer $x^\star$ exists and $p^\star = f(x^\star)$ is optimal value
(vi)	Step-sizes $\gamma_k > 0$ satisfy $\sum_{k=0}^{\infty} \gamma_k = \infty$ and $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$

- (iii): variance is bounded by M^2 since

$$\mathbb{E}[\|\widehat{\nabla} f(x)\|_2^2 | x] \geq \mathrm{Var}[\|\widehat{\nabla} f(x)\|_2 | x] + \|\nabla f(x)\|_2^2$$

- (iii): analysis is slightly simpler if assuming $\mathbb{E}[\|\widehat{\nabla} f(x)\|_2^2 | x] \leq G$

Nonconvex setting – Analysis

- Upper bound on f in Assumption (i) gives

$$
\begin{aligned}
&\mathbb{E}[f(x_{k+1})|x_k] \\
&\qquad \leq \mathbb{E}[f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \tfrac{\beta}{2}\|x_{k+1} - x_k\|_2^2 | x_k] \\
&\qquad = f(x_k) - \gamma_k \nabla f(x_k)^T \mathbb{E}[\widehat{\nabla} f(x_k)|x_k] + \tfrac{\beta\gamma_k^2}{2}\mathbb{E}[\|\widehat{\nabla} f(x_k)\|_2^2 | x_k] \\
&\qquad \leq f(x_k) - \gamma_k \nabla f(x_k)^T \nabla f(x_k) + \tfrac{\beta\gamma_k^2}{2}(\|\nabla f(x_k)\|_2^2 + M^2) \\
&\qquad = f(x_k) - \gamma_k(1 - \tfrac{\beta\gamma_k}{2})\|\nabla f(x_k)\|_2^2 + \tfrac{\beta\gamma_k^2}{2} M^2
\end{aligned}
$$

- Let $\gamma_k \leq \frac{1}{\beta}$ (true for large enough k since $(\gamma_k^2)_{k\in\mathbb{N}}$ summable):

$$
\mathbb{E}[f(x_{k+1})|x_k] \leq f(x_k) - \tfrac{\gamma_k}{2}\|\nabla f(x_k)\|_2^2 + \tfrac{\beta\gamma_k^2}{2} M^2
$$

- Subtracting $p^\star$ from both sides gives

$$
\mathbb{E}[f(x_{k+1})|x_k] - p^\star \leq f(x_k) - p^\star - \tfrac{\gamma_k}{2}\|\nabla f(x_k)\|_2^2 + \tfrac{\beta\gamma_k^2}{2} M^2
$$

Lyapunov inequality

- Take expected value and use law of total expectation to get:

$$\underbrace{\mathbb{E}[f(x_{k+1})] - p^\star}_{V_{k+1}} \leq \underbrace{\mathbb{E}[f(x_k)] - p^\star}_{V_k} - \tfrac{\gamma_k}{2} \underbrace{\mathbb{E}[\|\nabla f(x_k)\|_2^2]}_{R_k} + \underbrace{\tfrac{\beta\gamma_k^2}{2} M^2}_{W_k}$$

- Consequences:
  - $V_k = \mathbb{E}[f(x_k)] - p^\star$ converges (not necessarily to 0)
  - $\sum_{l=0}^k \frac{\gamma_l}{2} R_l \leq V_0 + \sum_{l=0}^k W_k$, which, when multiplied by 2 gives

$$\sum_{l=0}^{k} \gamma_l \mathbb{E}[\|\nabla f(x_l)\|_2^2] \leq 2(f(x_0) - p^\star) + \sum_{l=0}^{k} \gamma_l^2 \beta M^2$$

Minimum expected gradient norm bound

- Lyapunov inequality consequence restated:

$$\sum_{l=0}^{k} \gamma_l \mathbb{E}[\|\nabla f(x_l)\|_2^2] \le 2(f(x_0) - p^\star) + \sum_{l=0}^{k} \gamma_l^2 \beta M^2$$

- Using that

$$\min_{l=0,\ldots,k} \mathbb{E}[\|\nabla f(x_l)\|_2^2] \sum_{l=0}^{k} \gamma_l \le \sum_{l=0}^{k} \gamma_l \mathbb{E}[\|\nabla f(x_l)\|_2^2]$$

$$\mathbb{E}[\min_{l=0,\ldots,k} \|\nabla f(x_l)\|_2^2] \le \min_{l=0,\ldots,k} \mathbb{E}[\|\nabla f(x_l)\|_2^2]$$

where second is Jensen's inequality on concave $\min_l$, we get

$$\mathbb{E}[\min_{l=0,\ldots,k} \|\nabla f(x_l)\|_2^2] \le \frac{2(f(x_0) - p^\star) + \sum_{l=0}^{k} \gamma_l^2 \beta M^2}{\sum_{l=0}^{k} \gamma_l}$$

where terms in the numerator:

- $2(f(x_0) - p^\star)$ is due to initial suboptimality
- $\sum_{l=0}^{k} \gamma_l^2 \beta M^2$ is due to noise in gradient estimates
(if $M = 0$, use $\gamma_k = \frac{1}{\beta}$ to recover (proximal) gradient bound)

Minimum expected gradient norm convergence

- What conclusions can we draw from

$$\mathbb{E}[\min_{l=0,\ldots,k} \|\nabla f(x_l)\|_2^2] \leq \frac{2(f(x_0) - p^\star) + \sum_{l=0}^k \gamma_l^2 \beta M^2}{\sum_{l=0}^k \gamma_l}$$

- Let $C = \sum_{l=0}^\infty \gamma_l^2 < \infty$ (finite since $(\gamma_k^2)_{k\in\mathbb{N}}$ summable) then

$$\mathbb{E}[\min_{l=0,\ldots,k} \|\nabla f(x_l)\|_2^2] \leq \frac{2(f(x_0) - p^\star) + C\beta M^2}{\sum_{l=0}^k \gamma_l} \to 0$$

  as $k \to \infty$ since $(\gamma_k)_{k\in\mathbb{N}}$ is not summable
- Consequences:
  - Expected value of smallest gradient norm converges to 0
  - Minimum gradient converges to 0 in probability
  - We don't know what happens with latest expected value

Outline

- Stochastic gradient method
- Nonconvex setting
- **Convex setting**
- Step-sizes and rates
- Refined step-size and rate analysis
- Rate comparison to proximal gradient method
- Stochastic gradient descent variations

Stochastic gradient descent

- We consider problems of the form

$$\text{minimize}\, f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex

- We will analyze stochastic gradient descent

$$x_{k+1} = x_k - \gamma_k \widehat{\nabla} f(x_k)$$

where $\widehat{\nabla} f(x_k)$ is an unbiased estimate of $\nabla f(x_k)$ for all x_k

- Will show sublinear convergence rates that depend on step-sizes

Convex setting – Assumptions

(i)	$f : \mathbb{R}^n \to \mathbb{R}$ is convex but not necessarily differentiable
(ii)	Stochastic subgradient of f is unbiased: $\mathbb{E}[\widehat{\nabla} f(x) \vert x] \in \partial f(x)$
(iii)	Second moment is bounded: $\mathbb{E}[\Vert\widehat{\nabla} f(x)\Vert_2^2 \vert x] \leq G^2$
(iv)	A minimizer $x^\star$ exists and $p^\star = f(x^\star)$ is optimal value
(v)	Step-sizes $\gamma_k > 0$ satisfy $\sum_{k=0}^\infty \gamma_k = \infty$ and $\sum_{k=0}^\infty \gamma_k^2 < \infty$

- Do not assume smoothness or differentiability of f
- (iii): assumption is stronger than variance bound:

$$\mathbb{E}[\|\widehat{\nabla} f(x)\|_2^2 | x] \leq \|\nabla f(x)\|_2^2 + M^2$$

but can be relaxed under smoothness assumptions

Convex setting – Analysis

- Let, by (ii), $\mathbb{E}[\widehat{\nabla} f(x_k)|x_k] = g_k \in \partial f(x_k)$, then

$$
\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^\star\|_2^2|x_k] \\
&\quad = \mathbb{E}[\|x_k - \gamma_k \widehat{\nabla} f(x_k) - x^\star\|_2^2|x_k] \\
&\quad = \|x_k - x^\star\|_2^2 - 2\gamma_k \mathbb{E}_k[\widehat{\nabla} f(x_k)|x_k]^T(x_k - x^\star) + \gamma_k^2 \mathbb{E}[\|\widehat{\nabla} f(x_k)\|_2^2|x_k] \\
&\quad \leq \|x_k - x^\star\|_2^2 - 2\gamma_k g_k^T(x_k - x^\star) + \gamma_k^2 G^2
\end{aligned}
$$

- Use subgradient definition $f(x^\star) \geq f(x_k) + g_k^T(x^\star - x_k)$ to get

$$
\mathbb{E}[\|x_{k+1} - x^\star\|_2^2|x_k] \leq \|x_k - x^\star\|_2^2 - 2\gamma_k(f(x_k) - f(x^\star)) + \gamma_k^2 G^2
$$

Lyapunov inequality

- Take expected value and use law of total expectation to get:

$$\underbrace{\mathbb{E}[\|x_{k+1} - x^\star\|_2^2]}_{V_{k+1}} \leq \underbrace{\mathbb{E}[\|x_k - x^\star\|_2^2]}_{V_k} - 2\gamma_k \underbrace{\mathbb{E}[(f(x_k) - f(x^\star))]}_{R_k} + \underbrace{\gamma_k^2 G^2}_{W_k}$$

- Consequences:
  - $V_k = \mathbb{E}[\|x_k - x^\star\|_2^2]$ converges (not necessarily to 0)
  - $\sum_{l=0}^{k} 2\gamma_l R_l \leq V_0 + \sum_{l=0}^{k} W_k$, which gives

$$\sum_{l=0}^{k} 2\gamma_l \mathbb{E}[(f(x_l) - f(x^\star))] \leq \|x_0 - x^\star\|_2^2 + \sum_{l=0}^{k} \gamma_l^2 G^2$$

Minimum expected function value bound

- What are the consequences of:

$$\sum_{l=0}^{k} 2\gamma_l \mathbb{E}[(f(x_l) - f(x^\star))] \leq \|x_0 - x^\star\|_2^2 + \sum_{l=0}^{k} \gamma_l^2 G^2$$

- By using

$$\min_{l=0,\ldots,k} \mathbb{E}[f(x_l) - f(x^\star)] \sum_{l=0}^{k} \gamma_l \leq \sum_{l=0}^{k} \gamma_l \mathbb{E}[f(x_l) - f(x^\star)]$$

$$\mathbb{E}[\min_{l=0,\ldots,k} f(x_l) - f(x^\star)] \leq \min_{l=0,\ldots,k} \mathbb{E}[f(x_l) - f(x^\star)]$$

where second is Jensen's inequality on concave $\min_l$, we get

$$\mathbb{E}[\min_{l=0,\ldots,k} f(x_k) - f(x^\star)] \leq \frac{\|x_0 - x^\star\|_2^2 + \sum_{l=0}^{k} \gamma_l^2 G^2}{2 \sum_{l=0}^{k} \gamma_l}$$

- The last iterate not bounded

Weighted average expected function value bound

- Let us define the weighted average $\bar{x}_k = \displaystyle\sum_{l=0}^{k} \frac{\gamma_l}{\sum_{j=0}^{k} \gamma_j} x_l$
- By Jensen's inequality for convex f, we have

$$f(\bar{x}_k) = f\left(\sum_{l=0}^{k} \frac{\gamma_l}{\sum_{j=0}^{k} \gamma_j} x_l \right) \leq \sum_{l=0}^{k} \frac{\gamma_l}{\sum_{j=0}^{k} \gamma_j} f(x_l)$$

- Subtract $f(x^\star)$, multiply by $\left(\sum_{j=0}^{k} \gamma_j \right)$, and take expectation:

$$\left(\sum_{j=0}^{k} \gamma_j \right) \mathbb{E}[f(\bar{x}_k) - f(x^\star)] \leq \sum_{l=0}^{k} \gamma_l \mathbb{E}[f(x_l) - f(x^\star)]$$

- This gives the following bound for the average:

$$\mathbb{E}[f(\bar{x}_k) - f(x^\star)] \leq \frac{\|x_0 - x^\star\|_2^2 + \sum_{l=0}^{k} \gamma_l^2 G^2}{2 \sum_{l=0}^{k} \gamma_l}$$

Expected function value convergence

- Let $C = \sum_{l=0}^{\infty} \gamma_l^2 < \infty$ (finite since $(\gamma_k^2)_{k\in\mathbb{N}}$ summable) then

$$Q_k \leq \frac{\|x_0 - x^\star\|_2^2 + CG^2}{2\sum_{l=0}^{k} \gamma_l} \to 0$$

as $k \to \infty$ since $(\gamma_k)_{k\in\mathbb{N}}$ is not summable, where

$$Q_k = \mathbb{E}[\min_{l=0,\ldots,k} f(x_k) - f(x^\star)] \qquad \text{or} \qquad Q_k = \mathbb{E}[f(\bar{x}_k) - f(x^\star)]$$

- Expected smallest and average function value converge to $f(x^\star)$
- Function values converge in probability to optimal function $f(x^\star)$
- We have no last iterate convergence bound

Smoothness

- We did not assume smoothness (or differentability) for result
- What happens if we add smoothness?
  - Rate is not improved, but can improve constant
  - We can replace $\mathbb{E}[\|\widehat{\nabla} f(x)\|_2^2 | x] \leq G$ assumption by weaker

    $$\mathbb{E}[\|\widehat{\nabla} f(x)\|_2^2 | x] \leq \|\nabla f(x)\|_2^2 + M^2$$

    that bounds variance (as in nonconvex analysis)
  - If $\gamma_k \leq \frac{1}{\beta}$, it can shown that

    $$\mathbb{E}[\min_{l=0,\ldots,k} f(x_k) - f(x^\star)] \leq \frac{\|x_0 - x^\star\|_2^2 + \sum_{l=0}^k \gamma_l^2 M^2}{2 \sum_{l=0}^k \gamma_l}$$

    where, similar to in the smooth nonconvex setting, the term:
    - $\|x_0 - x^\star\|_2^2$ is due to initial suboptimality
    - $\sum_{l=0}^k \gamma_l^2 M^2$ is due to variance in gradient estimates

Strong convexity

- Assumption: f smooth and strongly convex
- Proximal gradient method achieves linear convergence
- Stochastic gradient descent does not achieve linear convergence

Outline

- Stochastic gradient method
- Nonconvex setting
- Convex setting
- **Step-sizes and rates**
- Refined step-size and rate analysis
- Rate comparison to proximal gradient method
- Stochastic gradient descent variations

Unifying convergence results

- Convergence in nonconvex and convex settings are:

$$Q_k \leq \frac{V_0 + DC}{b \sum_{l=0}^{k} \gamma_l}$$

where $C = \sum_{l=0}^{\infty} \gamma_l^2 < \infty$ by summability of $(\gamma_k)_{k\in\mathbb{N}}$

- Convex setting: $D = G^2$, $b = 2$, $V_0 = \|x_0 - x^\star\|_2^2$

$$Q_k = \mathbb{E}[\min_{i\in\{0,\ldots,k\}} f(x_i) - f(x^\star)] \qquad \text{or} \qquad Q_k = \mathbb{E}[f(\bar{x}_k) - f(x^\star)]$$

- Nonconvex setting: $D = \beta M^2$, $b = 1$, $V_0 = 2(f(x_0) - p^\star)$, and

$$Q_k = \mathbb{E}[\min_{i\in\{0,\ldots,k\}} \|\nabla f(x_i)\|_2^2]$$

Step-size requirements

- Step-size requirement $\sum_{l=0}^{\infty} \gamma_l = \infty$ makes upper bound

$$Q_k \leq \frac{V_0 + DC}{b \sum_{l=0}^{k} \gamma_l} \rightarrow 0$$

  as $k \rightarrow \infty$, with Q_k from previous slide, since $C = \sum_{l=0}^{\infty} \gamma_l^2 < \infty$
- Step-sizes that satisfy $\sum_{l=0}^{\infty} \gamma_l = \infty$ and $\sum_{l=0}^{\infty} \gamma_l^2 < \infty$
  - $\gamma_k = c/k$, with $c > 0$
  - $\gamma_k = c/k^{\alpha}$ for $\alpha \in (0.5, 1)$, with $c > 0$

Estimating rates via integrals

- For convergence need to verify $\sum_{l=0}^{\infty} \gamma_l = \infty$ and $\sum_{l=0}^{\infty} \gamma_l^2 < \infty$
- To estimate rate we need to lower bound $\sum_{l=0}^{k} \gamma_l$
- Assume $\gamma_l = \phi(l)$ with decreasing and nonnegative $\phi : \mathbb{R}_+ \to \mathbb{R}_+$
- We can estimate sums using integral formula:

$$\int_{t=0}^{k} \phi(t)dt + \phi(k) \leq \sum_{l=0}^{k} \phi(l) \leq \int_{t=0}^{k} \phi(t)dt + \phi(0)$$

  (we can remove $\phi(k) \geq 0$ from lower bound to simplify)
- Will use upper bound on $\sum_{l=0}^{k} \gamma_l^2$ and lower bound on $\sum_{l=0}^{k} \gamma_l$

Estimating rates – Example $\gamma_k = \frac{c}{k+1}$

- Let $\gamma_k = \phi(k)$ with $\phi(k) = \frac{c}{k+1}$ and estimate the sum

$$\sum_{l=0}^{k} \gamma_l \geq \int_{t=0}^{k} \frac{c}{t+1} dt = c \log(k+1) \to \infty$$

as $k \to \infty$ and

$$\sum_{l=0}^{k} \gamma_l^2 \leq \int_{t=0}^{k} \frac{c^2}{(t+1)^2} dt + \phi(0)^2 = c^2(1 - \tfrac{1}{k+1}) + c^2 \leq 2c^2 < \infty$$

- We arrive at the following (slow) $O(1/\log(k+1))$ rate:

$$Q_k \leq \frac{V_0 + DC}{b \sum_{l=0}^{k} \gamma_l} \leq \frac{V_0 + 2Dc^2}{bc \log(k+1)} = \frac{V_0/c + 2Dc}{b \log(k+1)}$$

- The constant c trades off the two constant terms V_0 and $2D$

Estimating rates – Example $\gamma_k = \frac{c}{(k+1)^\alpha}$

- Let $\gamma_k = \phi(k)$ with $\phi(k) = \frac{c}{(k+1)^\alpha}$ and $\alpha \in (0.5, 1)$ and estimate

$$\sum_{l=0}^{k} \gamma_l \geq \int_{t=0}^{k} \frac{c}{(t+1)^\alpha} dt = \tfrac{c}{1-\alpha}((k+1)^{1-\alpha} - 1) \to \infty$$

as $k \to \infty$ and, since $\phi(0)^2 = c^2$:

$$\sum_{l=0}^{k} \gamma_l^2 - c^2 \leq \int_{t=0}^{k} \frac{c^2}{(t+1)^{2\alpha}} dt = c^2 [\tfrac{(t+1)^{1-2\alpha}}{1-2\alpha}]_{t=0}^{k} \leq \tfrac{c^2}{2\alpha-1} < \infty$$

- We arrive at the following $O(1/(k+1)^{1-\alpha})$ rate:

$$Q_k \leq \frac{V_0 + DC}{b \sum_{l=0}^{k} \gamma_l} \leq \frac{(1-\alpha)(V_0 + Dc^2 \frac{2\alpha}{2\alpha-1})}{bc((k+1)^{1-\alpha} - 1)}$$

- Comments:
  - Rate improves with smaller α: $\frac{1}{(k+1)^{1-\alpha}} \to \sqrt{k+1}$ as $\alpha \to 0.5$
  - Constant worse with smaller α: $(1-\alpha) \nearrow$, $\frac{2\alpha}{2\alpha-1} \nearrow$ as $\alpha \searrow 0.5$

Outline

- Stochastic gradient method
- Nonconvex setting
- Convex setting
- Step-sizes and rates
- **Refined step-size and rate analysis**
- Rate comparison to proximal gradient method
- Stochastic gradient descent variations

Refining the step-size analysis

- Have not assumed $\sum_{l=0}^{\infty} \gamma_l^2$ finite for general convergence bound

$$Q_k \leq \frac{V_0 + D \sum_{l=0}^{k} \gamma_l^2}{b \sum_{l=0}^{k} \gamma_l}$$

- We can divide the sum into two parts

$$Q_k \leq \frac{V_0}{b \sum_{l=0}^{k} \gamma_l} + \frac{D}{b \frac{\sum_{l=0}^{k} \gamma_l}{\sum_{l=0}^{k} \gamma_l^2}}$$

- So $Q_k \to 0$ if $\sum_{l=0}^{k} \gamma_l \to \infty$ and $\frac{\sum_{l=0}^{k} \gamma_l}{\sum_{l=0}^{k} \gamma_l^2} \to \infty$
(don't need $\sum_{l=0}^{k} \gamma_l^2 < \infty$ for $Q_k \to 0$)

Refined step-size analysis interpretation

- Let $\psi_1(k) \leq \sum_{l=0}^{k} \gamma_l$ and $\psi_2(k) \leq \frac{\sum_{l=0}^{k} \gamma_l}{\sum_{l=0}^{k} \gamma_l^2}$ and restate bound:

$$Q_k \leq \frac{V_0}{b\psi_1(k)} + \frac{D}{b\psi_2(k)}$$

- ψ_1 decides how fast V_0 $(f(x_0) - p^\star$ or $\|x_0 - x^\star\|_2^2)$ is supressed
- ψ_2 decides how fast D, that comes from noise, is supressed
- There is a tradeoff between supressing these quantities
- Actual convergence very much dependent on constants V_0 and D

Estimating rates – Example $\gamma_k = \frac{c}{(k+1)^\alpha}$

- Let now $\alpha \in (0, 0.5)$ and estimate

$$\sum_{l=0}^{k} \gamma_l \geq \tfrac{c}{1-\alpha}((k+1)^{1-\alpha} - 1)$$

  squared sum does not converge, but can be shown to satisfy

$$\sum_{l=0}^{k} \gamma_l^2 \leq \tfrac{c^2}{1-2\alpha}((k+1)^{1-2\alpha} - 2\alpha)$$

- We use these to arrive at the following rate when $\gamma_k = \frac{c}{(k+1)^\alpha}$:

$$Q_k \leq \frac{(1-\alpha)V_0}{2bc((k+1)^{1-\alpha} - 1)} + \frac{(1-\alpha)Dc}{b(1-2\alpha)\frac{k^{1-\alpha}-1}{(k+1)^{1-2\alpha}-2\alpha}}$$

  where rate is worst of these: $O(\frac{(k+1)^{1-2\alpha}}{(k+1)^{1-\alpha}}) = O(\frac{1}{(k+1)^\alpha})$
- Comments:
  - Rate improves with larger α: $\frac{1}{(k+1)^\alpha} \to \sqrt{k+1}$ as $\alpha \to 0.5$
  - Constant worse with larger α: $\frac{1}{1-2\alpha} \nearrow$ as $\alpha \nearrow 0.5$

Estimating rates – Example $\gamma_k = \frac{c}{\sqrt{k+1}}$

- We know from before that

$$\sum_{l=0}^{k} \gamma_l = \sum_{l=0}^{k} \tfrac{c}{\sqrt{l+1}} \geq 2c(\sqrt{k+1} - 1)$$

and that the sum of step-sizes does not converge, but satisfies

$$\sum_{l=0}^{k} \gamma_l^2 = \sum_{l=0}^{k} \tfrac{c^2}{l+1} \leq c^2 \log(k+1)$$

- Since $\sum_{l=0}^{k} \gamma_l \to \infty$ and $\sum_{l=0}^{k} \gamma_l / \sum_{l=0}^{k} \gamma_l^2 \to \infty$ also

$$Q_k \leq \frac{V_0}{2bc\sqrt{k+1}} + \frac{Dc}{2b\frac{\sqrt{k+1}-1}{\log{(k+1)}}} \to 0$$

with rate $O(\frac{\log{(k+1)}}{\sqrt{(k+1)}})$ (since slower than $O(\frac{1}{\sqrt{k+1}})$)

Comparing rates for $\gamma_k = \frac{c}{k+1}$ and $\gamma_k = \frac{c}{\sqrt{k+1}}$

- Rates for $\gamma_k = \frac{c}{k+1}$ and $\gamma_k = \frac{c}{\sqrt{k+1}}$ respectively:

$$Q_k \leq \frac{V_0/c + 2Dc}{b \log{(k+1)}} \quad \text{and} \quad Q_k \leq \frac{V_0}{2bc\sqrt{k+1}} + \frac{Dc}{2b\frac{\sqrt{k+1}-1}{\log{(k+1)}}}$$

- Constants in the two terms similar or same
- Rate better for $\gamma_k = \frac{c}{\sqrt{k+1}}$ $(O(\frac{\log(k+1)}{\sqrt{k+1}})$ vs $O(\frac{1}{\log(k+1)}))$
- This is worst-case analysis, might not reflect actual performance

Outline

- Stochastic gradient method
- Nonconvex setting
- Convex setting
- Step-sizes and rates
- Refined step-size and rate analysis
- **Rate comparison to proximal gradient method**
- Stochastic gradient descent variations

Rate comparison

Setting	Quantity	Gradient	Stochastic $\gamma_k = \frac{1}{k^\alpha}$	
			$\alpha = 1$	$\alpha = 0.5$
Nonconvex	$\min_{l \in \{0,\dots,k\}} \|\nabla f(x_l)\|_2^2$	$O(\frac{1}{k})$	$O(\frac{1}{\log k})$	$O(\frac{\log k}{\sqrt{k}})$
Convex	$\min_{l \in \{0,\dots,k\}} (f(x_l) - f(x^\star))$	$O(\frac{1}{k})$	$O(\frac{1}{\log k})$	$O(\frac{\log k}{\sqrt{k}})$
Strongly convex	-	linear	sublinear	sublinear

- For stochastic, we have expectation around convergence quantity
- For convex gradient method, smallest suboptimality is the latest
- Constants similar except extra term from gradient estimate noise
- Stochastic gradient descent rate slower in all settings
- However, every iteration in stochastic gradient descent cheaper

Finite sum comparison

- We consider

$$\text{minimize} \sum_{i=1}^{N} f_i(x)$$

where N is large and use one f_i for each stochastic gradient
- N iterations of stochastic gradient is at cost of 1 full gradient
- Progress after k epochs (stochastic) vs k iterations (full):

Setting	Quantity	Gradient	Stochastic $\gamma_k = \frac{1}{k^\alpha}$	
			$\alpha = 1$	$\alpha = 0.5$
Nonconvex	$\min_{l \in \{0,\ldots,k\}} \|\nabla f(x_l)\|_2^2$	$O(\frac{1}{k})$	$O(\frac{1}{\log Nk})$	$O(\frac{\log Nk}{\sqrt{Nk}})$
Convex	$\min_{l \in \{0,\ldots,k\}} (f(x_l) - f(x^\star))$	$O(\frac{1}{k})$	$O(\frac{1}{\log Nk})$	$O(\frac{\log Nk}{\sqrt{Nk}})$

Finite sum comparison – Quantification

- Assume that finite sum of N equals 10 million summands
- Assume constant for SGD 10x larger than for GD
- Computational budget is that we run $k = 10$ iterations/epochs
- Replacing upper bounds with numbers:

Setting	**Quantity**	**Gradient**	**Stochastic** $\gamma_k = \frac{1}{k^\alpha}$	
			$\alpha = 1$	$\alpha = 0.5$
Nonconvex	$\min_{l\in\{0,\ldots,k\}} \|\nabla f(x_l)\|_2^2$	0.1	0.54	0.018
Convex	$\min_{l\in\{0,\ldots,k\}} (f(x_l) - f(x^\star))$	0.1	0.54	0.018

- Stochastic gives better worst case guarantees
- Significant difference between stochastic methods
- Actual performance depends a lot on relation between constants

Outline

- Stochastic gradient method
- Nonconvex setting
- Convex setting
- Step-sizes and rates
- Refined step-size and rate analysis
- Rate comparison to proximal gradient method
- **Stochastic gradient descent variations**

Adaptive diagonal scaling

- Diagonal scaling gives one step-size (learning rate) per variable
- Gives SGD with diagonal scaling $H_k = \mathbf{diag}(h_{1,k}, \ldots, h_{N,k})$

$$x_{k+1} = x_k - \gamma H_k^{-1} \widehat{\nabla} f(x_k)$$

  where the inverse is $H_k^{-1} = \mathbf{diag}(\frac{1}{h_{1,k}}, \ldots, \frac{1}{h_{N,k}})$
- A few methods exists that adaptively select individual step sizes
  - Adagrad
  - RMSProp
  - Adam
  - Adamax
  - Adadelta
- Among these, Adagrad was first but Adam most popular
- Sometimes improve convergence compared to SGD
- Will briefly motivate Adagrad and show how Adam differs

Motivation for Adagrad

- Consider SGD with diagonal scaling H_k:

$$x_{k+1} = x_k - \gamma H_k^{-1} \widehat{\nabla} f(x_k)$$

- Update our analysis in the convex setting by
  - expanding the square in the H_k norm and
  - assuming deterministic $H_k \succeq H_{k-1}$ for all k
  - not replacing $\mathbb{E}[\|\widehat{\nabla} f(x_k)\|^2_{H_k^{-1}} | x_k]$ by upper bound G^2
  - using fixed step-size $\gamma_k = \gamma$

  we get bound (that converges if H_k increases fast enough)

$$\mathbb{E}[f(x_k) - f(x^\star)] \leq \frac{\gamma^{-1}\|x_0 - x^\star\|^2_{H_0} + \gamma \sum_{l=0}^{k} \mathbb{E}[\|\widehat{\nabla} f(x_l)\|^2_{H_l^{-1}}]}{2(k+1)}$$

- Adagrad idea: select H_k to optimize constant

Adagrad idea for selecting H_k

- Assume $H_k = H$ has been constant and optimize bound constant

$$\gamma^{-1}\|x_0 - x^\star\|_H^2 + \gamma \sum_{l=0}^{k} \mathbb{E}[\|\widehat{\nabla} f(x_l)\|_{H^{-1}}^2]$$

- Don't know $\|x_0 - x^\star\|_H^2$, approximate with $\mathrm{tr}(H)\|x_0 - x^\star\|_2^2$
- Estimate sum from realization $\mathbb{E}[\|\widehat{\nabla} f(x_l)\|_{H^{-1}}^2] = \|\widetilde{\nabla} f(x_l)\|_{H^{-1}}^2$
- Let $R = \|x_0 - x^\star\|_2^2$ to get optimization problem

$$\gamma^{-1} R\mathrm{tr}(H) + \gamma \sum_{l=0}^{k} \|\widetilde{\nabla} f(x_l)\|_{H^{-1}}^2$$

- Problem is separable in diagonal elements with solution

$$h_{ii} = \tfrac{\gamma}{\sqrt{R}} \|(\widetilde{\nabla} f(x_l))_i^{0:k}\|_2$$

where $(\widetilde{\nabla} f(x_l))_i^{0:k} = (\widetilde{\nabla} f(x_0)_i, \ldots, \widetilde{\nabla} f(x_k)_i)$
- Since we do not know R, we can set $R = \gamma^2$

Adagrad summary

- Adagrad adds ϵ to above estimate for numerical reasons
- The algorithm is
  1. $\widetilde{\nabla} f(x_k)$ is subgradient or stochastic (sub)gradient of f at x_k
  2. Select metric H_k
     - set $s_k = \sum_{l=0}^{k}(\widetilde{\nabla} f(x_k))^2$
     - set $h_k = \epsilon\mathbf{1} + \sqrt{s_k}$
     - set $H_k = \gamma^{-1}\,\mathbf{diag}(h_k)$
  3. $x_{k+1} = x_k - H_k^{-1} g_k = x_k - \gamma g_k./(\epsilon\mathbf{1} + \sqrt{s_k})$
- Sometimes H_k sums up too fast so too short steps are taken
- Possible reason, in smooth settings, we would get rate constant

$$\gamma^{-1}\|x_0 - x^\star\|_H^2 + \gamma \sum_{l=0}^{k} \mathbb{E}[\|\widehat{\nabla} f(x_l) - \nabla f(x_l)\|_{H^{-1}}^2]$$

where second term in this rate constant

- depends on noise, not full gradient as in Adagrad development
- would give smaller H_k and longer steps
- is more difficult to estimate online

Variations – RMSprop and Adam

- In Adagrad, H_k may grow too fast which gives too short steps
- Instead: Don't *sum* gradient square, estimate variance:

$$\hat{v}_k = b_v \hat{v}_{k-1} + (1 - b_v)(\widetilde{\nabla} f(x_k))^2$$

  where $\hat{v}_0 = 0$, $b_v \in (0, 1)$
- H_k is chosen (approximately) as standard deviation:
  - RMSprop: biased estimate $H_k = \mathbf{diag}(\sqrt{\hat{v}_k} + \epsilon)$
  - Adam: unbiased estimate $H_k = \mathbf{diag}(\sqrt{\frac{\hat{v}_k}{1-b_v^k}} + \epsilon)$

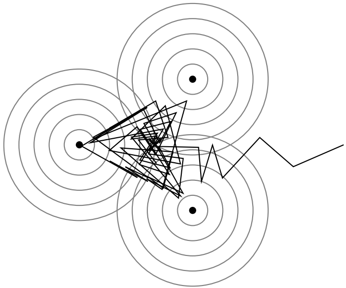
  which is much smaller than in Adagrad $\Rightarrow$ longer steps
- Intuition:
  - Reduce step size for high variance coordinates
  - Increase step size for low variance coordinates
- Adam also filters stochastic gradients for smoother updates

Filtered stochastic gradients

- Let $m_0 = 0$ and $b_m \in (0,1)$, and update

$$\hat{m}_k = b_m \hat{m}_{k-1} + (1 - b_m) g_k$$

- Adam uses unbiased estimate: $\frac{\hat{m}_k}{1-b_m^k}$
- Does not improve convergence properties, but slower changes
- Problem from before, fixed step-size, without filtered gradient



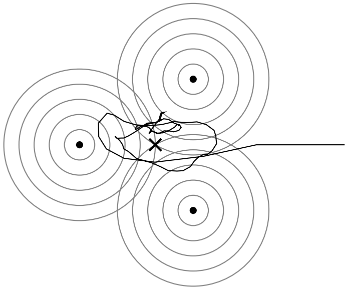
Levelsets of summands

Filtered stochastic gradients

- Let $m_0 = 0$ and $b_m \in (0, 1)$, and update

$$\hat{m}_k = b_m \hat{m}_{k-1} + (1 - b_m) g_k$$

- Adam uses unbiased estimate: $\frac{\hat{m}_k}{1-b_m^k}$
- Does not improve convergence properties, but slower changes
- Problem from before, fixed step-size, with filtered gradient



Levelsets of summands

Adam – Summary

- Initialize $\hat{m}_0 = \hat{v}_0 = 0$, $b_m, b_v \in (0,1)$, and select $\gamma > 0$
    1. $g_k = \widetilde{\nabla} f(x_k)$ (stochastic gradient)
    2. $\hat{m}_k = b_m \hat{m}_{k-1} + (1 - b_m) g_k$
    3. $\hat{v}_k = b_v \hat{v}_{k-1} + (1 - b_v) g_k^2$
    4. $m_k = \hat{m}_k / (1 - b_m^k)$
    5. $v_k = \hat{v}_k / (1 - b_v^k)$
    6. $x_{k+1} = x_k - \gamma m_k./(\sqrt{v_k} + \epsilon \mathbf{1})$
- Suggested choices $b_m = 0.9$ and $b_v = 0.999$
- Similar to Adagrad, but $\sqrt{v_k} \ll \sqrt{s_k} \Rightarrow$ longer steps
- May not work in deterministic setting (unlike Adagrad):
    - If method converges $\nabla f(x_k) \to 0$
    - Then $v_k \to 0$ and steps become very large
    - Needs noise and stochastic gradients to work well