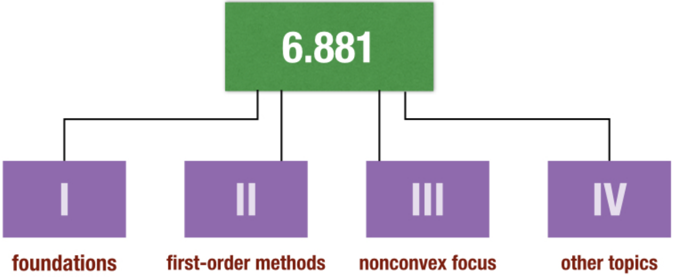
Optimization for Machine Learning

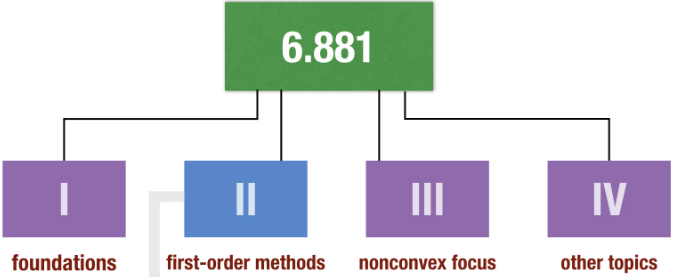
Lecture 7: First-order methods

6.881: MIT

Suvrit Sra
Massachusetts Institute of Technology

11 Mar, 2021


6.881
I
II
III
IV
foundations
first-order methods
nonconvex focus
other topics



1. GradDesc, SGD, Subgradient method
2. Accelerated gradient, mirror descent
3. Faster stochastic methods, variance reduction
4. Franke-Wolfe (projection free) methods
5. Operator splitting methods
6. Coordinate descent, BCD, Altmin
7. CCCP, EM method, and related ideas

First-order methods

$$x \leftarrow x - \eta g(x)$$

First-order methods

$$x \leftarrow x - \eta \boxed{g(x)}$$

First-order methods

$$x \leftarrow x - \eta g(x)$$

$\nabla f(x)$

GD

First-order methods

$$x \leftarrow x - \eta \boxed{g(x)}$$

$\nabla f(x)$

GD

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

SGD

First-order methods

$$x \leftarrow x - \eta \boxed{g(x)}$$

$\nabla f(x)$

GD

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

SGD

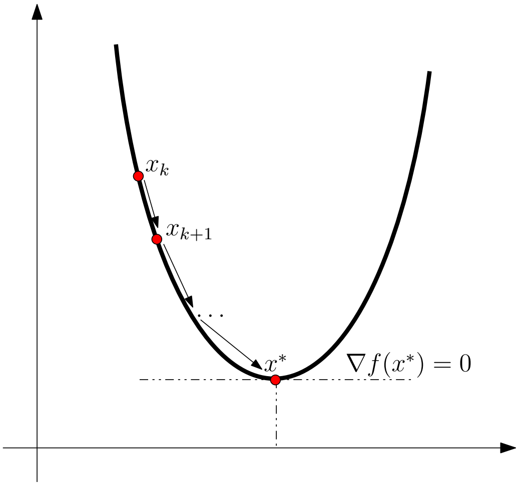
$g(x) \in \partial f(x)$

Subgrad

Gradient Descent

$$x \leftarrow x - \eta \nabla f(x)$$

Descent methods



Descent methods

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$
- Consider updating x using

$$x(\eta) = x + \eta d,$$

where **direction** $d \in \mathbb{R}^n$ obtuse to $\nabla f(x)$, i.e.,

$$\langle \nabla f(x), d \rangle < 0.$$

Descent methods

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$
- Consider updating x using

$$x(\eta) = x + \eta d,$$

where **direction** $d \in \mathbb{R}^n$ obtuse to $\nabla f(x)$, i.e.,

$$\langle \nabla f(x), d \rangle < 0.$$

- Again, we have the Taylor expansion

$$f(x(\eta)) = f(x) + \eta \langle \nabla f(x), d \rangle + o(\eta),$$

where $\langle \nabla f(x), d \rangle$ dominates $o(\eta)$ for small η

Descent methods

- Suppose we have a vector $x \in \mathbb{R}^n$ for which $\nabla f(x) \neq 0$
- Consider updating x using

$$x(\eta) = x + \eta d,$$

where **direction** $d \in \mathbb{R}^n$ obtuse to $\nabla f(x)$, i.e.,

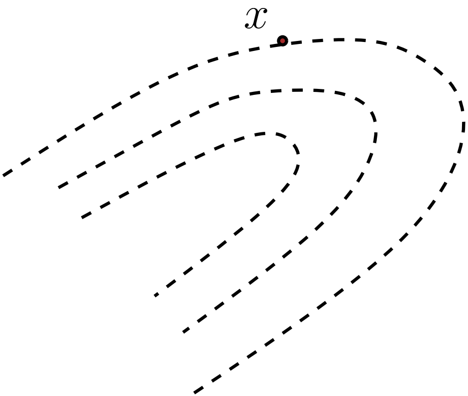
$$\langle \nabla f(x), d \rangle < 0.$$

- Again, we have the Taylor expansion

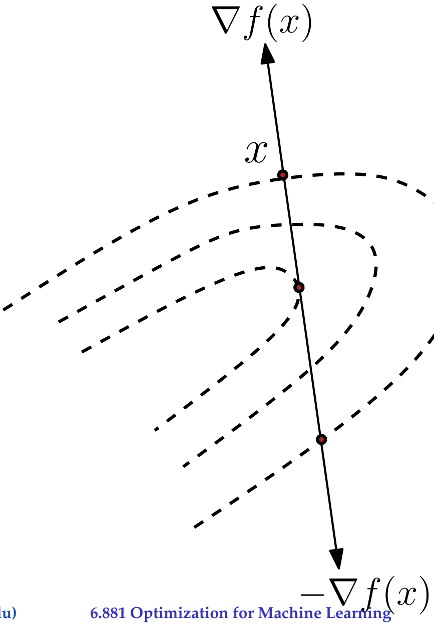
$$f(x(\eta)) = f(x) + \eta \langle \nabla f(x), d \rangle + o(\eta),$$

where $\langle \nabla f(x), d \rangle$ dominates $o(\eta)$ for small η
- Since d is obtuse to $\nabla f(x)$, this implies $f(x(\eta)) < f(x)$

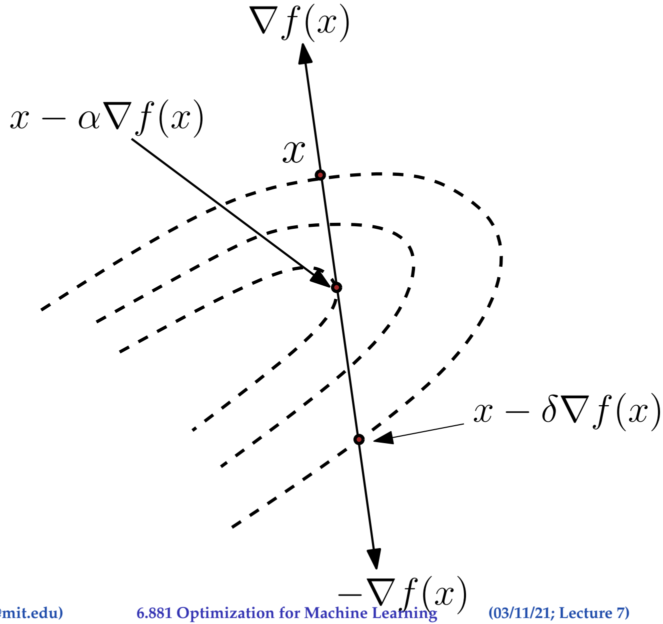
Descent methods



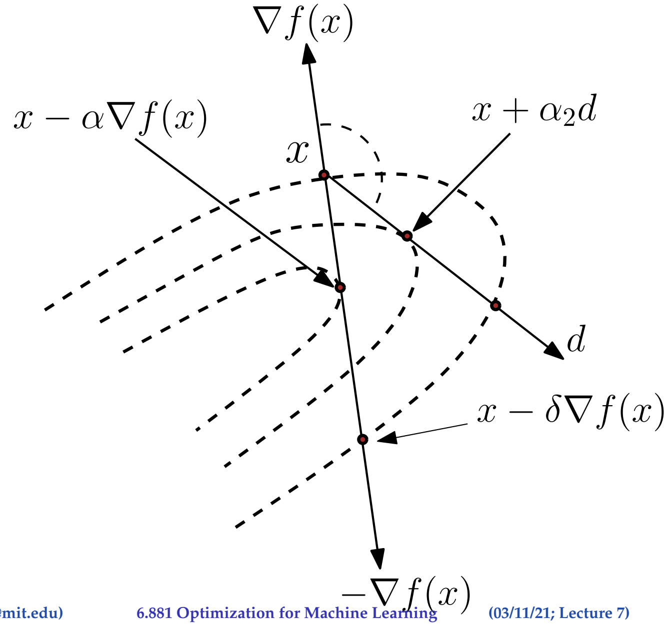
Descent methods



Descent methods



Descent methods



Gradient-based methods

1. Start with some guess x^0;
2. For each $k = 0, 1, \ldots$
   - $x^{k+1} \leftarrow x^k + \eta_k d^k$
   - Stop somehow (e.g., if $\|\nabla f(x^{k+1})\| \leq \varepsilon$)

Gradient based methods

$$\boxed{x^{k+1} = x^k + \eta_k d^k, \quad k = 0, 1, \ldots}$$

Gradient based methods

$$\boxed{x^{k+1} = x^k + \eta_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\eta_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$

Gradient based methods

$$\boxed{x^{k+1} = x^k + \eta_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\eta_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k),\, d^k \rangle < 0$$

Gradient based methods

$$\boxed{x^{k+1} = x^k + \eta_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\eta_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k),\, d^k \rangle < 0$$

Numerous ways to select η_k and d^k

Gradient based methods

$$\boxed{x^{k+1} = x^k + \eta_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\eta_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k),\, d^k \rangle < 0$$

$$\boxed{\text{Numerous ways to select } \eta_k \text{ and } d^k}$$

Many methods **seek monotonic descent**

$$f(x^{k+1}) < f(x^k)$$

Gradient methods – direction

$$x^{k+1} = x^k + \eta_k d^k, \quad k = 0, 1, \ldots$$

- Different choices of direction d^k
  - **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
  - **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
  - **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
  - **Steepest descent:** $D^k = \boldsymbol{I}$
  - **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}$
  - **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.

Gradient methods – direction

$$x^{k+1} = x^k + \eta_k d^k, \quad k = 0, 1, \ldots$$

▶ Different choices of direction d^k

- **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
- **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
- **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
- **Steepest descent:** $D^k = \boldsymbol{I}$
- **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}$
- **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.
- $\ldots$

Exercise: Verify that $\langle \nabla f(x^k), d^k \rangle < 0$ for above choices

Stepsize selection

- **Constant:** $\eta_k = 1/L$ (for suitable value of L)

Stepsize selection

- **Constant:** $\eta_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\eta_k \to 0$ but $\sum_k \eta_k = \infty$.

Stepsize selection

- **Constant:** $\eta_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\eta_k \to 0$ but $\sum_k \eta_k = \infty$.
  Exercise: Prove that the latter condition ensures that $\{x^k\}$ does not converge to nonstationary points.

Stepsize selection

- **Constant:** $\eta_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\eta_k \to 0$ but $\sum_k \eta_k = \infty$.
  Exercise: Prove that the latter condition ensures that $\{x^k\}$ does not converge to nonstationary points.
  Sketch: Say, $x^k \to \bar{x}$; then for sufficiently large m and n, $(m > n)$

$$x^m \approx x^n \approx \bar{x}, x^m \approx x^n - \left(\sum_{k=n}^{m-1} \eta_k\right) \nabla f(\bar{x}).$$

  The sum can be made arbitrarily large, contradicting nonstationarity of $\bar{x}$

Stepsize selection*

- **Exact:** $\eta_k := \underset{\eta \geq 0}{\operatorname{argmin}} f(x^k + \eta d^k)$
- **Limited min:** $\eta_k = \underset{0 \leq \eta \leq s}{\operatorname{argmin}} f(x^k + \eta d^k)$
- **Armijo-rule**. Given **fixed** scalars, s, β, σ with $0 < \beta < 1$ and $0 < \sigma < 1$ (chosen experimentally). Set

$$\eta_k = \beta^{m_k} s,$$

where we **try** $\beta^m s$ for $m = 0, 1, \ldots$ until **sufficient descent**

$$f(x^k) - f(x + \beta^m s d^k) \geq -\sigma \beta^m s \langle \nabla f(x^k), d^k \rangle$$

If $\langle \nabla f(x^k), d^k \rangle < 0$, stepsize guaranteed to exist
Usually, σ small $\in [10^{-5}, 0.1]$, while β from $1/2$ to $1/10$ depending on how confident we are about initial stepsize s.

Barzilai-Borwein step-size*

- Stepsize computation can be expensive
- Convergence analysis depends on monotonic descent

Barzilai-Borwein step-size⋆

- Stepsize computation can be expensive
- Convergence analysis depends on monotonic descent
- Give up search for stepsizes
- Use constants or closed-form formulae for stepsizes
- Don't insist on monotonic descent?
- (e.g., diminishing stepsizes non-monotonic descent)

Barzilai-Borwein step-size*

- Stepsize computation can be expensive
- Convergence analysis depends on monotonic descent
- Give up search for stepsizes
- Use constants or closed-form formulae for stepsizes
- Don't insist on monotonic descent?
- (e.g., diminishing stepsizes non-monotonic descent)

Barzilai & Borwein stepsizes

$$x^{k+1} = x^k - \eta^k \nabla f(x^k), \quad k = 0, 1, \ldots$$

Barzilai-Borwein step-size*

- Stepsize computation can be expensive
- Convergence analysis depends on monotonic descent
- Give up search for stepsizes
- Use constants or closed-form formulae for stepsizes
- Don't insist on monotonic descent?
- (e.g., diminishing stepsizes non-monotonic descent)

Barzilai & Borwein stepsizes

$$x^{k+1} = x^k - \eta^k \nabla f(x^k), \quad k = 0, 1, \ldots$$

$$\eta_k = \frac{\langle u^k,\, v^k \rangle}{\|v^k\|^2}, \qquad \eta_k = \frac{\|u^k\|^2}{\langle u^k,\, v^k \rangle}$$

$$u^k = x^k - x^{k-1}, \quad v^k = \nabla f(x^k) - \nabla f(x^{k-1})$$

Challenge. Analyze convergence of GD using BB stepsizes.

Exercise

- Let D be the $(n-1) \times n$ *differencing* matrix

$$D = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & & \ddots & \\ & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times n},$$

- $f(x) = \frac{1}{2}\|D^T x - b\|_2^2 = \frac{1}{2}(\|D^T x\|_2^2 + \|b\|_2^2 - 2\langle D^T x, b\rangle)$
- Notice that $\nabla f(x) = D(D^T x - b)$
- Try different choices of b, and different initial vectors x_0
- **Exercise:** Experiment to see how large n must be before gradient method starts outperforming CVX
- **Exercise:** Minimize $f(x)$ for large n; e.g., $n = 10^6$, $n = 10^7$
- **Exercise:** Repeat same exercise with constraints: $x_i \in [-1, 1]$.

Convergence

(remarks)

Gradient descent – convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

Gradient descent – convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

- Gradient vectors of closeby points are close to each other
- Objective function has "bounded curvature"
- Speed at which gradient varies is bounded
- **Exercise:** If $f \in C_L^1$ is twice diff. then $\|\nabla^2 f(x)\|_2 \leq L$.

Gradient descent – convergence

Convergence of gradient norm

Theorem. Let $f \in C_L^1$. $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Gradient descent – convergence

Convergence of gradient norm

Theorem. Let $f \in C_L^1$. $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Theorem. Let $f \in C_L^1$. $\min_{1 \le k \le T} \|\nabla f(x^k)\|_2 = O(1/T)$

Gradient descent – convergence

Convergence of gradient norm

Theorem. Let $f \in C_L^1$. $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Theorem. Let $f \in C_L^1$. $\min_{1 \leq k \leq T} \|\nabla f(x^k)\|_2 = O(1/T)$

Convergence rate: function suboptimality

Theorem. Let $f \in C_L^1$ be convex; let $\{x^k\}$ be generated as above, with $\eta_k = 1/L$. Then, $f(x^{T+1}) - f(x^*) = O(1/T)$.

Gradient descent – convergence

Convergence of gradient norm

Theorem. Let $f \in C_L^1$. $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Theorem. Let $f \in C_L^1$. $\min_{1 \le k \le T} \|\nabla f(x^k)\|_2 = O(1/T)$

Convergence rate: function suboptimality

Theorem. Let $f \in C_L^1$ be convex; let $\{x^k\}$ be generated as above, with $\eta_k = 1/L$. Then, $f(x^{T+1}) - f(x^*) = O(1/T)$.

Theorem. If $f \in S_{L,\mu}^1$, $\eta = \frac{2}{L+\mu}$, and $\{x^k\}$ generated by GD. Then, $f(x^T) - f^* \le \frac{L}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^{2T} \|x^0 - x^*\|_2^2$, where $\kappa := L/\mu$ is the *condition number*.

Proof

(sketches)

Key result: The Descent Lemma

Lemma (Descent lemma). Let $f \in C^1_L$. Then,

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|_2^2$$

Key result: The Descent Lemma

Lemma (Descent lemma). Let $f \in C^1_L$. Then,

$$f(x) \le f(y) + \langle \nabla f(y),\, x - y\rangle + \tfrac{L}{2}\|x-y\|_2^2$$

Proof. By Taylor's theorem, for $z_t = y + t(x-y)$ we have

$$f(x) = f(y) + \textstyle\int_0^1 \langle \nabla f(z_t),\, x-y\rangle dt.$$

Adding and subtracting $\langle \nabla f(y),\, x-y\rangle$ we obtain

$$\begin{aligned}
|f(x) - f(y) - \langle \nabla f(y),\, x-y\rangle| &= \left|\textstyle\int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x-y\rangle dt\right| \\
&\le \textstyle\int_0^1 |\langle \nabla f(z_t) - \nabla f(y),\, x-y\rangle| dt \\
&\le \textstyle\int_0^1 \|\nabla f(z_t) - \nabla f(y)\|_2 \|x-y\|_2 dt \\
&\le L \textstyle\int_0^1 t\|x-y\|_2^2 dt \\
&= \tfrac{L}{2}\|x-y\|_2^2.
\end{aligned}$$

Bounds $f(x)$ above and below with quadratic functions

Descent lemma – corollary

Coroll. 1 If $f \in C^1_L$, and $0 < \eta_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

Descent lemma – corollary

Coroll. 1 If $f \in C^1_L$, and $0 < \eta_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

Proof.

$$\begin{aligned} f(x^{k+1}) &\le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2 \\ &= f(x^k) - \eta_k \|\nabla f(x^k)\|_2^2 + \tfrac{\eta_k^2 L}{2}\|\nabla f(x^k)\|_2^2 \\ &= f(x^k) - \eta_k(1 - \tfrac{\eta_k}{2}L)\|\nabla f(x^k)\|_2^2 \end{aligned}$$

If $\eta_k < 2/L$ we have descent. min over η_k to get best bound, giving $\eta_k = 1/L$.

Descent lemma – corollary

Coroll. 1 If $f \in C^1_L$, and $0 < \eta_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

Proof.

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2 \\ &= f(x^k) - \eta_k \|\nabla f(x^k)\|_2^2 + \tfrac{\eta_k^2 L}{2}\|\nabla f(x^k)\|_2^2 \\ &= f(x^k) - \eta_k(1 - \tfrac{\eta_k}{2}L)\|\nabla f(x^k)\|_2^2 \end{aligned}$$

If $\eta_k < 2/L$ we have descent. min over η_k to get best bound, giving $\eta_k = 1/L$.

Alternative bigger picture

Minimize global upper bound:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{L}{2}\|x - y\|^2$$

$f(x) \leq F(x, y)$, where $F(x, x) = f(x)$

Explore: Other global upper bounds and corresponding algorithms.

Convergence of gradient norm

- We showed that
$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2L}\|\nabla f(x^k)\|_2^2,$$

Convergence of gradient norm

- We showed that

$$f(x^k) - f(x^{k+1}) \geq \tfrac{1}{2L}\|\nabla f(x^k)\|_2^2,$$

- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain

$$\frac{1}{2L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1})$$

Convergence of gradient norm

- We showed that
$$f(x^k) - f(x^{k+1}) \geq \tfrac{1}{2L}\|\nabla f(x^k)\|_2^2,$$
- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain
$$\frac{1}{2L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1}) \leq f(x^0) - f^*.$$

Convergence of gradient norm

- We showed that
$$f(x^k) - f(x^{k+1}) \geq \tfrac{1}{2L}\|\nabla f(x^k)\|_2^2,$$
- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain
$$\frac{1}{2L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1}) \leq f(x^0) - f^*.$$
- We assume $f^* > -\infty$, so rhs is some fixed positive constant

Convergence of gradient norm

- We showed that
$$f(x^k) - f(x^{k+1}) \geq \tfrac{1}{2L}\|\nabla f(x^k)\|_2^2,$$
- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain
$$\frac{1}{2L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1}) \leq f(x^0) - f^*.$$
- We assume $f^* > -\infty$, so rhs is some fixed positive constant
- Thus, as $k \to \infty$, lhs must converge; thus
$$\|\nabla f(x^k)\|_2 \to 0 \quad \text{as} \quad k \to \infty.$$

Convergence of gradient norm

- We showed that
$$f(x^k) - f(x^{k+1}) \geq \tfrac{1}{2L}\|\nabla f(x^k)\|_2^2,$$
- Sum up above inequalities for $k = 0, 1, \ldots, T$ to obtain
$$\frac{1}{2L}\sum_{k=0}^{T}\|\nabla f(x^k)\|_2^2 \quad \leq \quad f(x^0) - f(x^{T+1}) \leq f(x^0) - f^*.$$
- We assume $f^* > -\infty$, so rhs is some fixed positive constant
- Thus, as $k \to \infty$, lhs must converge; thus
$$\|\nabla f(x^k)\|_2 \to 0 \quad \text{as} \quad k \to \infty.$$
- $\min_{0\leq k\leq T}\|\nabla f(x^k)\|^2 \leq \frac{1}{T+1}\sum_{k=0}^{T}\|\nabla f(x^k)\|^2$, so $O(\frac{1}{\varepsilon})$ for $\|\nabla f\|^2 \leq \varepsilon$
- Notice, we **did not require** f to be convex ...

SGD

Convergence rate – strongly convex

Theorem. If $f \in S^1_{L,\mu}$, $0 < \eta < 2/(L+\mu)$, then the gradient method generates a sequence $\{x^k\}$ that satisfies

$$\|x^k - x^*\|_2^2 \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^k \|x^0 - x^*\|_2.$$

Moreover, if $\eta = 2/(L+\mu)$ then

$$f(x^k) - f^* \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - x^*\|_2^2,$$

where $\kappa := L/\mu$ is the *condition number*.

Convergence – strongly convex case

Assumption: *Strong convexity*; denote $f \in S^1_{L,\mu}$

$$f(x) \geq f(y) + \langle \nabla f(y),\, x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2$$

Convergence – strongly convex case

Assumption: *Strong convexity*; denote $f \in S^1_{L,\mu}$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2$$

Descent lemma convex corollary

Corollary 2. If f is a **convex** function $\in C^1_L$, then

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle,$$

Exercise: Prove Cor. 2. (*Hint:* Consider $\phi(y) = f(y) - \langle \nabla f(x_0), y \rangle$).

Convergence – strongly convex case

Assumption: *Strong convexity*; denote $f \in S^1_{L,\mu}$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2$$

Descent lemma convex corollary

Corollary 2. If f is a **convex** function $\in C^1_L$, then

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle,$$

Exercise: Prove Cor. 2. (*Hint:* Consider $\phi(y) = f(y) - \langle \nabla f(x_0), y \rangle$).

Valuable refinement for the strongly convex case. . .

Convergence – strongly convex case

Corollary 2. If f is a **convex** function $\in C_L^1$, then

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y),\, x - y\rangle,$$

Thm 2. Suppose $f \in S_{L,\mu}^1$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Convergence – strongly convex case

Corollary 2. If f is a **convex** function $\in C^1_L$, then

$$\tfrac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y),\, x - y\rangle,$$

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

- Consider the **convex** function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$

Convergence – strongly convex case

Corollary 2. If f is a **convex** function $\in C^1_L$, then

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y),\, x - y\rangle,$$

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

- Consider the **convex** function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- If $\mu = L$, then immediate from strong convexity and Cor. 2

Convergence – strongly convex case

Corollary 2. If f is a **convex** function $\in C^1_L$, then

$$\tfrac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y),\, x - y\rangle,$$

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

- Consider the **convex** function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- If $\mu = L$, then immediate from strong convexity and Cor. 2
- If $\mu < L$, then $\phi \in C^1_{L-\mu}$; now invoke Cor. 2

$$\langle \nabla \phi(x) - \nabla \phi(y),\, x - y\rangle \geq \frac{1}{L - \mu}\|\nabla \phi(x) - \nabla \phi(y)\|_2^2$$

Convergence – strongly convex case

Corollary 2. If f is a **convex** function $\in C^1_L$, then

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y),\, x - y \rangle,$$

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

- Consider the **convex** function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- If $\mu = L$, then immediate from strong convexity and Cor. 2
- If $\mu < L$, then $\phi \in C^1_{L-\mu}$; now invoke Cor. 2

$$\langle \nabla \phi(x) - \nabla \phi(y),\, x - y \rangle \geq \frac{1}{L - \mu}\|\nabla \phi(x) - \nabla \phi(y)\|_2^2$$

Let's put this to use now....

Convergence rate – strongly convex

Theorem. If $f \in S^1_{L,\mu}$, $0 < \eta < 2/(L+\mu)$, then the gradient method generates a sequence $\{x^k\}$ that satisfies

$$\|x^k - x^*\|_2^2 \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^k \|x^0 - x^*\|_2.$$

Moreover, if $\eta = 2/(L+\mu)$ then

$$f(x^k) - f^* \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - x^*\|_2^2,$$

where $\kappa := L/\mu$ is the *condition number*.

Strongly convex – rate

- **Key idea:** Analyze $r_k = \|x^k - x^*\|_2$ recursively; consider thus,

Strongly convex – rate

- **Key idea:** Analyze $r_k = \|x^k - x^*\|_2$ recursively; consider thus,

$$r_{k+1}^2 = \|x^k - x^* - \eta \nabla f(x^k)\|_2^2$$

Strongly convex – rate

- **Key idea:** Analyze $r_k = \|x^k - x^*\|_2$ recursively; consider thus,

$$
\begin{aligned}
r_{k+1}^2 &= \|x^k - x^* - \eta \nabla f(x^k)\|_2^2 \\
&= r_k^2 - 2\eta\langle \nabla f(x^k),\, x^k - x^* \rangle + \eta^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Strongly convex – rate

- **Key idea:** Analyze $r_k = \|x^k - x^*\|_2$ recursively; consider thus,

$$\begin{aligned} r_{k+1}^2 &= \|x^k - x^* - \eta\nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2\eta\langle\nabla f(x^k), x^k - x^*\rangle + \eta^2\|\nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2\eta\langle\nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle + \eta^2\|\nabla f(x^k)\|_2^2 \end{aligned}$$

Strongly convex – rate

- **Key idea:** Analyze $r_k = \|x^k - x^*\|_2$ recursively; consider thus,

$$\begin{aligned}
r_{k+1}^2 &= \|x^k - x^* - \eta\nabla f(x^k)\|_2^2 \\
&= r_k^2 - 2\eta\langle\nabla f(x^k),\, x^k - x^*\rangle + \eta^2\|\nabla f(x^k)\|_2^2 \\
&= r_k^2 - 2\eta\langle\nabla f(x^k) - \nabla f(x^*),\, x^k - x^*\rangle + \eta^2\|\nabla f(x^k)\|_2^2 \\
&\le \left(1 - \frac{2\eta\mu L}{\mu + L}\right) r_k^2 + \eta\left(\eta - \frac{2}{\mu + L}\right)\|\nabla f(x^k)\|_2^2
\end{aligned}$$

Strongly convex – rate

- **Key idea:** Analyze $r_k = \|x^k - x^*\|_2$ recursively; consider thus,

$$
\begin{aligned}
r_{k+1}^2 &= \|x^k - x^* - \eta\nabla f(x^k)\|_2^2 \\
&= r_k^2 - 2\eta\langle \nabla f(x^k),\, x^k - x^*\rangle + \eta^2\|\nabla f(x^k)\|_2^2 \\
&= r_k^2 - 2\eta\langle \nabla f(x^k) - \nabla f(x^*),\, x^k - x^*\rangle + \eta^2\|\nabla f(x^k)\|_2^2 \\
&\leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right) r_k^2 + \eta\left(\eta - \frac{2}{\mu + L}\right)\|\nabla f(x^k)\|_2^2
\end{aligned}
$$

where we used Thm. 2 with $\nabla f(x^*) = 0$ for last inequality.

Exercise: Complete the proof of the theorem now.

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \eta_k g^k\|_2^2$.

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \eta_k g^k\|_2^2$.

$$r_{k+1}^2 \quad = \quad r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k,\, x^k - x^*\rangle$$

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \eta_k g^k\|_2^2$.

$$\begin{aligned} r_{k+1}^2 &= r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k,\, x^k - x^*\rangle \\ &= r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k - g^*,\, x^k - x^*\rangle \quad \text{as } g^* = 0 \end{aligned}$$

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \eta_k g^k\|_2^2$.

$$\begin{aligned}
r_{k+1}^2 &= r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k,\, x^k - x^*\rangle \\
&= r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k - g^*,\, x^k - x^*\rangle \quad \text{as } g^* = 0 \\
&\leq r_k^2 + \eta_k^2\|g^k\|_2^2 - \tfrac{2\eta_k}{L}\|g^k - g^*\|_2^2 \qquad \text{(Coroll. 2, pg. 24)}
\end{aligned}$$

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \eta_k g^k\|_2^2$.

$$\begin{aligned}
r_{k+1}^2 &= r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k, x^k - x^*\rangle \\
&= r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k - g^*, x^k - x^*\rangle \quad \text{as } g^* = 0 \\
&\leq r_k^2 + \eta_k^2\|g^k\|_2^2 - \tfrac{2\eta_k}{L}\|g^k - g^*\|_2^2 \qquad \text{(Coroll. 2, pg. 24)} \\
&= r_k^2 - \eta_k(\tfrac{2}{L} - \eta_k)\|g^k\|_2^2.
\end{aligned}$$

Convergence rate – (weakly) convex

- ⋆ Want to prove the well-known $O(1/T)$ rate
- ⋆ Let $\eta_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Lemma Distance to min shrinks monotonically; $r_{k+1} \leq r_k$

Proof. Descent lemma implies that: $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$

Consider, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \eta_k g^k\|_2^2$.

$$
\begin{aligned}
r_{k+1}^2 &= r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k,\, x^k - x^*\rangle \\
&= r_k^2 + \eta_k^2\|g^k\|_2^2 - 2\eta_k\langle g^k - g^*,\, x^k - x^*\rangle \quad \text{as } g^* = 0 \\
&\leq r_k^2 + \eta_k^2\|g^k\|_2^2 - \tfrac{2\eta_k}{L}\|g^k - g^*\|_2^2 \qquad \text{(Coroll. 2, pg. 24)} \\
&= r_k^2 - \eta_k\left(\tfrac{2}{L} - \eta_k\right)\|g^k\|_2^2.
\end{aligned}
$$

Since $\eta_k < 2/L$, it follows that $r_{k+1} \leq r_k$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1 - \beta_k)$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1 - \beta_k)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\leq} \langle g^k, x^k - x^* \rangle$$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \le \Delta_k(1 - \beta_k)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\le} \langle g^k, x^k - x^* \rangle \overset{\text{CS}}{\le} \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1 - \beta_k)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\leq} \langle g^k, x^k - x^* \rangle \overset{\text{CS}}{\leq} \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

That is, $\|g^k\|_2 \geq \Delta_k / r_k$.

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \le \Delta_k(1-\beta_k)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\le} \langle g^k, x^k - x^* \rangle \overset{\text{CS}}{\le} \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

That is, $\|g^k\|_2 \ge \Delta_k / r_k$. In particular, since $r_k \le r_0$, we have

$$\|g^k\|_2 \ge \frac{\Delta_k}{r_0}.$$

Convergence rate

Lemma Let $\Delta_k := f(x^k) - f(x^*)$. Then, $\Delta_{k+1} \leq \Delta_k(1 - \beta_k)$

$$f(x^k) - f(x^*) = \Delta_k \overset{\text{cvx } f}{\leq} \langle g^k, x^k - x^* \rangle \overset{\text{CS}}{\leq} \|g^k\|_2 \underbrace{\|x^k - x^*\|_2}_{r_k}.$$

That is, $\|g^k\|_2 \geq \Delta_k / r_k$. In particular, since $r_k \leq r_0$, we have

$$\|g^k\|_2 \geq \frac{\Delta_k}{r_0}.$$

Now we have a bound on the gradient norm...

Convergence rate

Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\left(1 - \tfrac{\Delta_k}{2Lr_0^2}\right)$$

Convergence rate

Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\big(1 - \tfrac{\Delta_k}{2Lr_0^2}\big) = \Delta_k(1 - \beta_k).$$

Convergence rate

Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\left(1 - \tfrac{\Delta_k}{2Lr_0^2}\right) = \Delta_k(1 - \beta_k).$$

$$\boxed{\text{But we want to bound: } f(x^{T+1}) - f(x^*)}$$

Convergence rate

Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\left(1 - \tfrac{\Delta_k}{2Lr_0^2}\right) = \Delta_k(1 - \beta_k).$$

$$\boxed{\text{But we want to bound: } f(x^{T+1}) - f(x^*)}$$

$$\Delta_{k+1} \leq \Delta_k(1 - \beta_k) \implies \quad \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k}(1 + \beta_k) = \frac{1}{\Delta_k} + \frac{1}{2Lr_0^2}.$$

Convergence rate

Recall $f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|g^k\|_2^2$; subtracting f^* from both sides

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2Lr_0^2} = \Delta_k\left(1 - \tfrac{\Delta_k}{2Lr_0^2}\right) = \Delta_k(1-\beta_k).$$

$$\boxed{\text{But we want to bound: } f(x^{T+1}) - f(x^*)}$$

$$\Delta_{k+1} \leq \Delta_k(1-\beta_k) \Longrightarrow \quad \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k}(1+\beta_k) = \frac{1}{\Delta_k} + \frac{1}{2Lr_0^2}.$$

▶ Sum both sides over $k = 0, \ldots, T$ (telescoping) to obtain

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

Convergence rate

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

Convergence rate

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

- Rearrange to conclude

$$f(x^T) - f^* \leq \frac{2L\Delta_0 r_0^2}{2Lr_0^2 + T\Delta_0}$$

Convergence rate

$$\frac{1}{\Delta_{T+1}} \geq \frac{1}{\Delta_0} + \frac{T+1}{2Lr_0^2}$$

- Rearrange to conclude

$$f(x^T) - f^* \leq \frac{2L\Delta_0 r_0^2}{2Lr_0^2 + T\Delta_0}$$

- Use descent lemma to bound $\Delta_0 \leq (L/2)\|x^0 - x^*\|_2^2$; simplify

$$f(x^T) - f(x^*) \leq \frac{2L\Delta_0\|x^0 - x^*\|_2^2}{T+4} = O(1/T).$$

SGD

$$x \leftarrow x - \eta g$$

Why SGD?

Regularized Empirical Risk Minimization

$$\min_x \quad \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, x^T a_i) + \lambda r(x).$$

(e.g., logistic regression, deep learning, SVMs, etc.)

Why SGD?

Regularized Empirical Risk Minimization

$$\min_x \quad \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, x^T a_i) + \lambda r(x).$$

(e.g., logistic regression, deep learning, SVMs, etc.)

- training data: $(a_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$ (i.i.d.)

Why SGD?

Regularized Empirical Risk Minimization

$$\min_x \quad \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, x^T a_i) + \lambda r(x).$$

(e.g., logistic regression, deep learning, SVMs, etc.)

- training data: $(a_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$ (i.i.d.)
- large-scale ML: Both d and n are large:
  - d: dimension of each input sample
  - n: number of training data points / samples
- Assume training data "sparse"; so total datasize $\ll dn$.
- Running time $O(\#\text{nnz})$

Finite-sum problems

$$\min_{x \in \mathbb{R}^d} \quad f(x) =$$

Finite-sum problems

$$\min_{x \in \mathbb{R}^d} \quad f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

Finite-sum problems

$$\min_{x \in \mathbb{R}^d} \quad f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

Gradient / subgradient methods

$$\begin{aligned} x_{k+1} &= x_k - \eta_k \nabla f(x_k) \\ x_{k+1} &= x_k - \eta_k g(x_k), \quad g \in \partial f(x_k) \end{aligned}$$

If n is large, each iteration above is expensive

Stochastic gradients

At iteration k, we randomly pick an integer
$i(k) \in \{1, 2, \ldots, n\}$

$$x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$$

- The update requires only gradient for $f_{i(k)}$
- Uses unbiased estimate $\mathbb{E}[\nabla f_{i(k)}] = \nabla f$
- One iteration now n times faster using $\nabla f(x)$
- Can such a method work? If so, how fast? Why?

Intuition – (Bertsekas)

- Assume all variables involved are **scalars**.

$$\min \quad f(x) = \tfrac{1}{2} \sum\nolimits_{i=1}^{n} (a_i x - b_i)^2$$

Intuition – (Bertsekas)

- Assume all variables involved are **scalars**.

$$\min \quad f(x) = \tfrac{1}{2}\sum_{i=1}^{n}(a_i x - b_i)^2$$

- Solving $f'(x) = 0$ we obtain

$$x^* = \frac{\sum_i a_i b_i}{\sum_i a_i^2}$$

Intuition – (Bertsekas)

- Assume all variables involved are **scalars**.

$$\min \quad f(x) = \tfrac{1}{2}\sum\nolimits_{i=1}^{n}(a_i x - b_i)^2$$

- Solving $f'(x) = 0$ we obtain

$$x^* = \frac{\sum_i a_i b_i}{\sum_i a_i^2}$$

- Minimum of a single $f_i(x) = \frac{1}{2}(a_i x - b_i)^2$ is $x_i^* = b_i/a_i$

Intuition – (Bertsekas)

- Assume all variables involved are **scalars**.

$$\min \quad f(x) = \frac{1}{2}\sum_{i=1}^{n}(a_i x - b_i)^2$$

- Solving $f'(x) = 0$ we obtain

$$x^* = \frac{\sum_i a_i b_i}{\sum_i a_i^2}$$

- Minimum of a single $f_i(x) = \frac{1}{2}(a_i x - b_i)^2$ is $x_i^* = b_i/a_i$
- Notice now that

$$x^* \in [\min_i x_i^*, \max_i x_i^*] =: R$$

(Use: $\sum_i a_i b_i = \sum_i a_i^2 (b_i/a_i)$)

Intuition – (Bertsekas)

- Assume all variables involved are **scalars**.

$$\min \quad f(x) = \tfrac{1}{2} \sum\nolimits_{i=1}^{n} (a_i x - b_i)^2$$

- Notice: $x^* \in \left[\min_i x_i^*, \max_i x_i^*\right] =: R$

Intuition – (Bertsekas)

- Assume all variables involved are **scalars**.

$$\min \quad f(x) = \tfrac{1}{2}\sum_{i=1}^{n}(a_i x - b_i)^2$$

- Notice: $x^* \in \left[\min_i x_i^*, \max_i x_i^*\right] =: R$
- If we have a scalar x that lies outside R?
- We see that

$$\nabla f_i(x) = a_i(a_i x - b_i)$$
$$\nabla f(x) = \sum_i a_i(a_i x - b_i)$$

Intuition – (Bertsekas)

- Assume all variables involved are **scalars**.

$$\min \quad f(x) = \tfrac{1}{2} \sum_{i=1}^{n} (a_i x - b_i)^2$$

- Notice: $x^* \in \left[\min_i x_i^*, \max_i x_i^*\right] =: R$
- If we have a scalar x that lies outside R?
- We see that

$$\nabla f_i(x) = a_i(a_i x - b_i)$$

$$\nabla f(x) = \sum_i a_i(a_i x - b_i)$$

- $\nabla f_i(x)$ has **same sign** as $\nabla f(x)$. So using $\nabla f_i(x)$ **instead** of $\nabla f(x)$ also ensures progress.

Intuition – (Bertsekas)

- Assume all variables involved are **scalars**.

$$\min \quad f(x) = \frac{1}{2}\sum_{i=1}^{n}(a_i x - b_i)^2$$

- Notice: $x^* \in \left[\min_i x_i^*, \max_i x_i^*\right] =: R$
- If we have a scalar x that lies outside R?
- We see that

$$\nabla f_i(x) = a_i(a_i x - b_i)$$
$$\nabla f(x) = \sum_i a_i(a_i x - b_i)$$

- $\nabla f_i(x)$ has **same sign** as $\nabla f(x)$. So using $\nabla f_i(x)$ **instead** of $\nabla f(x)$ also ensures progress.
- But once inside region R, **no guarantee** that SGD will make progress towards optimum.

SGD: two variants

$$\boxed{\min \tfrac{1}{n} \textstyle\sum_i f_i(x)}$$

SGD: two variants

$$\min \frac{1}{n} \sum_i f_i(x)$$

- Start with feasible x_0
- For $k = 0, 1, \ldots,$
  - Option 1: Randomly pick an index i (with replacement)

SGD: two variants

$$\min \frac{1}{n} \sum_i f_i(x)$$

- Start with feasible x_0
- For $k = 0, 1, \ldots,$
  - Option 1: Randomly pick an index i (with replacement)
  - Option 2: Pick index i without replacement
  - Use $g_k = \nabla f_i(x)$ as the "stochastic gradient"
  - Update $x_{k+1} = x_k - \eta_k g_k$

SGD: two variants

$$\min \frac{1}{n} \sum_i f_i(x)$$

- Start with feasible x_0
- For $k = 0, 1, \ldots,$
  - Option 1: Randomly pick an index i (with replacement)
  - Option 2: Pick index i without replacement
  - Use $g_k = \nabla f_i(x)$ as the "stochastic gradient"
  - Update $x_{k+1} = x_k - \eta_k g_k$

Explore. Which version would you use? Why?

SGD: mini-batches

$$\min_x \quad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Idea: Use a mini-batch of stochastic gradients

$$x_{k+1} = x_k - \frac{\eta_k}{|I_k|}\sum_{j\in I_k} \nabla f_j(x_k)$$

SGD: mini-batches

$$\min_x \quad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Idea: Use a mini-batch of stochastic gradients

$$x_{k+1} = x_k - \frac{\eta_k}{|I_k|}\sum_{j\in I_k} \nabla f_j(x_k)$$

- Iteration k samples set I_k, and uses $|I_k|$ stochastic gradients
- Increases parallelism, reduces communication

Explore: Large mini-batches not that "favorable" for DNNs. 
(also known as: "large-batch training")

SGD: some theoretical challenges

$$x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$$

- Proving that it "works"
- Theoretical results lagging behind practice (without replacement SGD widely used, most published theory studies with replacement)

SGD: some theoretical challenges

$$x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$$

- Proving that it "works"
- Theoretical results lagging behind practice (without replacement SGD widely used, most published theory studies with replacement)

Explore: Why does SGD work so well for neural networks? (i.e., why does it deliver such low training losses despite non-convexiity, and how does it influence generalization behavior of neural networks?)

SGD for empirical risk / finite sums

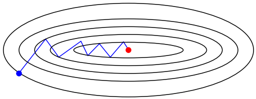
$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

– **Iteration**: $x^{k+1} = x^k - \eta_k f'_{i(t)}(x^k)$
  - Sampling with replacement: $i(k) \sim \mathrm{Unif}(\{1, \ldots, n\})$
  - Polyak-Ruppert averaging: $\bar{x}_k = \frac{1}{k+1} \sum_{j=0}^{k} x^j$

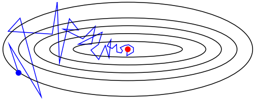
– **Convergence rate** if each f_i convex L-smooth, and f is μ-strongly-convex:

$$\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leqslant \begin{cases} O(1/\sqrt{k}) & \text{if } \eta_k = 1/(L\sqrt{k}) \\ O(L/(\mu k)) = O(\kappa/k) & \text{if } \eta_k = 1/(\mu k) \end{cases}$$

SGD vs GD (strongly convex case)

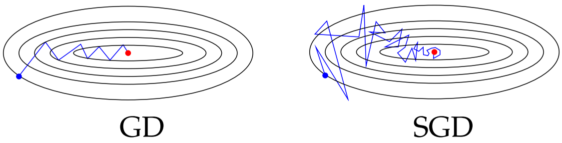


GD



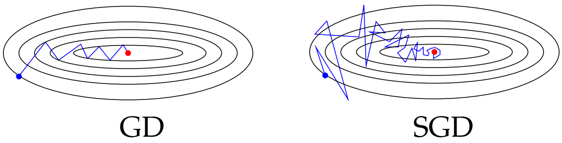
SGD

SGD vs GD (strongly convex case)



- Batch GD:
  – Linear (e.g., exponential) convergence rate in $O(e^{-k/\kappa})$
  – Iteration complexity is linear in n ($O(n \log 1/\varepsilon)$)

SGD vs GD (strongly convex case)



- Batch GD:
  - Linear (e.g., exponential) convergence rate in $O(e^{-k/\kappa})$
  - Iteration complexity is linear in n ($O(n \log 1/\varepsilon)$)
- SGD:
  - Sampling with replacement: $i(k)$ random element of $\{1, \ldots, n\}$
  - Convergence rate $O(\kappa/k)$
  - Iteration complexity independent of n ($O(1/\varepsilon^2)$)

Convergence

(some theory)

SGD: nonconvex (smooth) case

$f(x) = \frac{1}{n}\sum_i f_i(x)$ and $x_{k+1} = x_k - \eta_k \nabla f_{i_k}(x_k)$

SGD: nonconvex (smooth) case

$f(x) = \frac{1}{n}\sum_i f_i(x)$ and $x_{k+1} = x_k - \eta_k \nabla f_{i_k}(x_k)$

- Assumption 1: L-smooth components $f_i \in C^1_L$
- Assumption 2: Unbiased gradients $\mathbb{E}[\nabla f_{i_t}(x) - \nabla f(x)] = 0$
- Assumption 3: Bounded noise: $\mathbb{E}[\|\nabla f_{i_k}(x) - \nabla f(x)\|^2] = \sigma^2$
- Assumption 4: Bounded gradient: $\|\nabla f_i(x)\| \leq G$

SGD: nonconvex (smooth) case

$f(x) = \frac{1}{n}\sum_i f_i(x)$ and $x_{k+1} = x_k - \eta_k \nabla f_{i_k}(x_k)$

- Assumption 1: L-smooth components $f_i \in C^1_L$
- Assumption 2: Unbiased gradients $\mathbb{E}[\nabla f_{i_t}(x) - \nabla f(x)] = 0$
- Assumption 3: Bounded noise: $\mathbb{E}[\|\nabla f_{i_k}(x) - \nabla f(x)\|^2] = \sigma^2$
- Assumption 4: Bounded gradient: $\|\nabla f_i(x)\| \leq G$

Theorem. Under above assumptions, for suitable stepsize SGD satisfies

$$\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}[\|\nabla f(x_k)\|^2] \leq \frac{1}{\sqrt{T}}\left(\frac{f(x_1) - f(x^*)}{c} + \frac{Lc}{2}G^2\right),$$

for some constant c; hence $\min_k \mathbb{E}[\|\nabla f(x_k)\|^2] = O(1/\sqrt{T})$.

SGD: nonconvex (smooth) case

Proof: Using L-smoothness of f_i and taking expectations we obtain

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] + \mathbb{E}\left[\langle \nabla f(x_k), -\eta_k \nabla f_{i_k}(x_k)\rangle + \tfrac{L}{2}\|\eta_k \nabla f_{i_k}(x_k)\|^2\right]$$

SGD: nonconvex (smooth) case

Proof: Using L-smoothness of f_i and taking expectations we obtain

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})] &\leq \mathbb{E}[f(x_k)] + \mathbb{E}\left[\langle \nabla f(x_k), -\eta_k \nabla f_{i_k}(x_k)\rangle + \tfrac{L}{2}\|\eta_k \nabla f_{i_k}(x_k)\|^2\right] \\
&\leq \mathbb{E}[f(x_k)] - \eta_k \mathbb{E}\left[\|\nabla f(x_k)\|^2\right] + \tfrac{L\eta_k^2}{2} G^2.
\end{aligned}$$

SGD: nonconvex (smooth) case

Proof: Using L-smoothness of f_i and taking expectations we obtain

$$
\begin{aligned}
\mathbb{E}[f(x_{k+1})] &\leq \mathbb{E}[f(x_k)] + \mathbb{E}\left[\langle \nabla f(x_k),\, -\eta_k \nabla f_{i_k}(x_k)\rangle + \tfrac{L}{2}\|\eta_k \nabla f_{i_k}(x_k)\|^2\right] \\
&\leq \mathbb{E}[f(x_k)] - \eta_k \mathbb{E}\left[\|\nabla f(x_k)\|^2\right] + \tfrac{L\eta_k^2}{2} G^2.
\end{aligned}
$$

Rearranging the terms above we obtain

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

SGD: nonconvex (smooth) case

Proof: Using L-smoothness of f_i and taking expectations we obtain

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] &\le \mathbb{E}[f(x_k)] + \mathbb{E}\left[\langle \nabla f(x_k), -\eta_k \nabla f_{i_k}(x_k)\rangle + \tfrac{L}{2}\|\eta_k \nabla f_{i_k}(x_k)\|^2\right] \\ &\le \mathbb{E}[f(x_k)] - \eta_k \mathbb{E}\left[\|\nabla f(x_k)\|^2\right] + \tfrac{L\eta_k^2}{2} G^2.\end{aligned}$$

Rearranging the terms above we obtain

$$\mathbb{E}\left[\|\nabla f(x_k)\|^2\right] \le \frac{1}{\eta_k}\mathbb{E}\left[f(x_k) - f(x_{k+1})\right] + \frac{L\eta_k}{2} G^2.$$

Choose $\eta_k = \frac{c}{\sqrt{T}}$ for some constant c and sum over $k = 0$ to $T-1$ to obtain

$$\begin{aligned}\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}\left[\|\nabla f(x_k)\|^2\right] &\le \frac{1}{\sqrt{T}c}\mathbb{E}[f(x_1) - f(x_{T+1})] + \frac{Lc}{2\sqrt{T}} G^2 \\ &\le \frac{1}{\sqrt{T}}\left(\frac{f(x_1) - f(x^*)}{c} + \frac{Lc}{2} G^2\right).\end{aligned}$$

SGD: convex case

- $\min_{x \in \mathcal{X}} f(x) := \mathbb{E}[F(x, \xi)]$

SGD: convex case

- $\min_{x \in \mathcal{X}} f(x) := \mathbb{E}[F(x, \xi)]$
- Let ξ_k denote the randomness at step k
- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- SGD analysis hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$
- SGD iteration: $x_{k+1} \leftarrow P_{\mathcal{X}}(x_k - \eta_k g_k)$ ($P_{\mathcal{X}}$: projection)

SGD: convex case

- $\min_{x \in \mathcal{X}} f(x) := \mathbb{E}[F(x, \xi)]$
- Let ξ_k denote the randomness at step k
- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- SGD analysis hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$
- SGD iteration: $x_{k+1} \leftarrow P_{\mathcal{X}}(x_k - \eta_k g_k)$ ($P_{\mathcal{X}}$: projection)

Denote: $R_k := \|x_k - x^*\|^2$ and $r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x_k - x^*\|^2]$

SGD: convex case

- $\min_{x\in\mathcal{X}} f(x) := \mathbb{E}[F(x,\xi)]$
- Let ξ_k denote the randomness at step k
- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- SGD analysis hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$
- SGD iteration: $x_{k+1} \leftarrow P_{\mathcal{X}}(x_k - \eta_k g_k)$ ($P_{\mathcal{X}}$: projection)

Denote: $R_k := \|x_k - x^*\|^2$ and $r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x_k - x^*\|^2]$

Bounding R_{k+1}

$$\begin{aligned} R_{k+1} &= \|x_{k+1} - x^*\|_2^2 = \|P_{\mathcal{X}}(x_k - \eta_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\ &\le \|x_k - x^* - \eta_k g_k\|_2^2 \end{aligned}$$

SGD: convex case

- $\min_{x \in \mathcal{X}} f(x) := \mathbb{E}[F(x,\xi)]$
- Let ξ_k denote the randomness at step k
- x_k depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, x_k does not depend on ξ_k
- SGD analysis hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$
- SGD iteration: $x_{k+1} \leftarrow P_{\mathcal{X}}(x_k - \eta_k g_k)$ ($P_{\mathcal{X}}$: projection)

Denote: $R_k := \|x_k - x^*\|^2$ and $r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x_k - x^*\|^2]$

Bounding R_{k+1}

$$\begin{aligned} R_{k+1} &= \|x_{k+1} - x^*\|_2^2 = \|P_{\mathcal{X}}(x_k - \eta_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\ &\leq \|x_k - x^* - \eta_k g_k\|_2^2 \\ &= R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle. \end{aligned}$$

SGD – analysis for strongly cvx

$$\boxed{R_{k+1} \le R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

SGD – analysis for strongly cvx

$$\boxed{R_{k+1} \leq R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \leq G$, and take expectation:

$$r_{k+1} \leq r_k + \eta_k^2 G^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

Unbiasedness $\mathbb{E}[g_k] = \nabla f(x_k)$ and μ-strong convexity give

$$r_{k+1} \leq r_k + \eta_k^2 G^2 - 2\eta_k \mathbb{E}[f(x_k) - f(x^*) + \tfrac{\mu}{2}\|x_k - x^*\|^2].$$

SGD – analysis for strongly cvx

$$\boxed{R_{k+1} \le R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \le G$, and take expectation:

$$r_{k+1} \le r_k + \eta_k^2 G^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

Unbiasedness $\mathbb{E}[g_k] = \nabla f(x_k)$ and μ-strong convexity give

$$r_{k+1} \le r_k + \eta_k^2 G^2 - 2\eta_k \mathbb{E}[f(x_k) - f(x^*) + \tfrac{\mu}{2}\|x_k - x^*\|^2].$$

Rearranging and dividing by $2\eta_k$ we get

$$\mathbb{E}[f(x_k) - f(x^*)] \le \frac{\eta_k G^2}{2} + \frac{\eta_k^{-1} - \mu}{2} r_k - \frac{1}{2\eta_k} r_{k+1}.$$

SGD – analysis for strongly cvx

$$\boxed{R_{k+1} \le R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \le G$, and take expectation:

$$r_{k+1} \le r_k + \eta_k^2 G^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

Unbiasedness $\mathbb{E}[g_k] = \nabla f(x_k)$ and μ-strong convexity give

$$r_{k+1} \le r_k + \eta_k^2 G^2 - 2\eta_k \mathbb{E}[f(x_k) - f(x^*) + \tfrac{\mu}{2}\|x_k - x^*\|^2].$$

Rearranging and dividing by $2\eta_k$ we get

$$\mathbb{E}[f(x_k) - f(x^*)] \le \frac{\eta_k G^2}{2} + \frac{\eta_k^{-1} - \mu}{2} r_k - \frac{1}{2\eta_k} r_{k+1}.$$

Put $\eta_k = 1/\mu k$, and telescope (and one more trick...)

SGD – analysis for strongly cvx

$$\mathbb{E}[f(x_k) - f(x^*)] \leq \frac{G^2}{2\mu k} + \frac{\mu(k-1)}{2} r_k - \frac{\mu k}{2} r_{k+1}. \qquad (**)$$

Using convexity, observe that

$$\mathbb{E}f\left(\tfrac{1}{T}\sum\nolimits_{k=1}^{T} x_k\right) - f(x^*) \leq \tfrac{1}{T}\sum\nolimits_{k=1}^{T} \mathbb{E}[f(x_k) - f(x^*)]$$

Using $(**)$, after telescoping, clearing junk **Verify!** we get

$$\frac{1}{T}\sum\nolimits_{k=1}^{T} \mathbb{E}[f(x_k) - f(x^*)] \leq \frac{G^2}{2\mu T}\sum_{k=1}^{T}\frac{1}{k} \leq \frac{G^2}{2\mu T}(1 + \log T).$$

We've obtained the rate $O\left(\frac{G^2 \log T}{2\mu T}\right)$

SGD – exercise

Exercise: Suppose f_i is convex and $f(x)$ is μ-strongly convex. Let $\bar{x}_k := \sum_{i=0}^{k} \theta_i x_i$, where $\theta_i = \frac{2(i+1)}{(k+1)(k+2)}$, we obtain

$$\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \frac{2G^2}{\mu(k+1)}.$$

Question: What if we want to **not** use averaged iterates?

SGD: weakly convex case

$$\boxed{R_{k+1} \leq R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

SGD: weakly convex case

$$\boxed{R_{k+1} \le R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \le G$ on compact set $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \le r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

SGD: weakly convex case

$$\boxed{R_{k+1} \le R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \le G$ on compact set $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \le r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

- We need to now get a handle on the last term

SGD: weakly convex case

$$\boxed{R_{k+1} \leq R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \leq G$ on compact set $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \leq r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

- We need to now get a handle on the last term
- Since x_k is independent of ξ_k, we have

$$\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle] \quad =$$

SGD: weakly convex case

$$\boxed{R_{k+1} \le R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \le G$ on compact set $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \le r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

- We need to now get a handle on the last term
- Since x_k is independent of ξ_k, we have

$$\begin{aligned} \mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k)\rangle] &= \mathbb{E}\left\{\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k)\rangle \mid \xi_{[1..(k-1)]}]\right\} \\ &= \end{aligned}$$

SGD: weakly convex case

$$\boxed{R_{k+1} \leq R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \leq G$ on compact set $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \leq r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

- We need to now get a handle on the last term
- Since x_k is independent of ξ_k, we have

$$\begin{aligned}
\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle] &= \mathbb{E}\left\{\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle \mid \xi_{[1..(k-1)]}]\right\} \\
&= \mathbb{E}\left\{\langle x_k - x^*,\, \mathbb{E}[g(x_k, \xi_k) \mid \xi_{[1..(k-1)]}] \rangle\right\} \\
&=
\end{aligned}$$

SGD: weakly convex case

$$\boxed{R_{k+1} \le R_k + \eta_k^2 \|g_k\|_2^2 - 2\eta_k \langle g_k,\, x_k - x^* \rangle}$$

- **Assume:** $\|g_k\|_2 \le G$ on compact set $\mathcal{X}$
- Taking expectation:

$$r_{k+1} \le r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle g_k,\, x_k - x^* \rangle].$$

- We need to now get a handle on the last term
- Since x_k is independent of ξ_k, we have

$$\begin{aligned}
\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle] &= \mathbb{E}\left\{\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k)\rangle \mid \xi_{[1..(k-1)]}]\right\} \\
&= \mathbb{E}\left\{\langle x_k - x^*,\, \mathbb{E}[g(x_k, \xi_k) \mid \xi_{[1..(k-1)]}]\rangle\right\} \\
&= \mathbb{E}[\langle x_k - x^*,\, G_k\rangle], \quad G_k \in \partial F(x_k).
\end{aligned}$$

SGD: weakly convex case

$$\boxed{\text{It remains to bound: } \mathbb{E}[\langle x_k - x^*,\, G_k\rangle]}$$

SGD: weakly convex case

$$\boxed{\text{It remains to bound: } \mathbb{E}[\langle x_k - x^*,\ G_k\rangle]}$$

- Since F is cvx, $F(x) \geq F(x_k) + \langle G_k,\ x - x_k\rangle$ for any $x \in \mathcal{X}$.

SGD: weakly convex case

It remains to bound: $\mathbb{E}[\langle x_k - x^*, G_k \rangle]$

- Since F is cvx, $F(x) \geq F(x_k) + \langle G_k, x - x_k \rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular

$$2\eta_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\eta_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]$$

SGD: weakly convex case

$$\boxed{\text{It remains to bound: } \mathbb{E}[\langle x_k - x^*,\, G_k\rangle]}$$

- Since F is cvx, $F(x) \geq F(x_k) + \langle G_k,\, x - x_k\rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular

$$2\eta_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\eta_k \mathbb{E}[\langle G_k,\, x^* - x_k\rangle]$$

Plug this bound back into the r_{k+1} inequality:

$$r_{k+1} \quad \leq \quad r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle G_k,\, x_k - x^*\rangle]$$

SGD: weakly convex case

$$\boxed{\text{It remains to bound: } \mathbb{E}[\langle x_k - x^*,\, G_k\rangle]}$$

- Since F is cvx, $F(x) \geq F(x_k) + \langle G_k,\, x - x_k\rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular

$$2\eta_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\eta_k \mathbb{E}[\langle G_k,\, x^* - x_k\rangle]$$

Plug this bound back into the r_{k+1} inequality:

$$\begin{aligned} r_{k+1} &\leq r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle G_k,\, x_k - x^*\rangle] \\ 2\eta_k \mathbb{E}[\langle G_k,\, x_k - x^*\rangle] &\leq r_k - r_{k+1} + \eta_k M^2 \end{aligned}$$

SGD: weakly convex case

It remains to bound: $\mathbb{E}[\langle x_k - x^*,\, G_k\rangle]$

- Since F is cvx, $F(x) \geq F(x_k) + \langle G_k,\, x - x_k\rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular

$$2\eta_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\eta_k \mathbb{E}[\langle G_k,\, x^* - x_k\rangle]$$

Plug this bound back into the r_{k+1} inequality:

$$\begin{aligned} r_{k+1} &\leq r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle G_k,\, x_k - x^*\rangle] \\ 2\eta_k \mathbb{E}[\langle G_k,\, x_k - x^*\rangle] &\leq r_k - r_{k+1} + \eta_k M^2 \\ 2\eta_k \mathbb{E}[F(x_k) - F(x^*)] &\leq r_k - r_{k+1} + \eta_k M^2. \end{aligned}$$

SGD: weakly convex case

$$\boxed{\text{It remains to bound: } \mathbb{E}[\langle x_k - x^*, G_k \rangle]}$$

- Since F is cvx, $F(x) \geq F(x_k) + \langle G_k, x - x_k \rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular

$$2\eta_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\eta_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]$$

Plug this bound back into the r_{k+1} inequality:

$$\begin{aligned} r_{k+1} &\leq r_k + \eta_k^2 M^2 - 2\eta_k \mathbb{E}[\langle G_k, x_k - x^* \rangle] \\ 2\eta_k \mathbb{E}[\langle G_k, x_k - x^* \rangle] &\leq r_k - r_{k+1} + \eta_k M^2 \\ 2\eta_k \mathbb{E}[F(x_k) - F(x^*)] &\leq r_k - r_{k+1} + \eta_k M^2. \end{aligned}$$

$$\boxed{\text{We've bounded the expected progress; What now?}}$$

SGD: weakly convex case

$$2\eta_k \mathbb{E}[F(x_k) - F(x^*)] \le r_k - r_{k+1} + \eta_k M^2.$$

SGD: weakly convex case

$$2\eta_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \eta_k M^2.$$

Sum up over $i = 1, \ldots, k$, to obtain

$$\sum\nolimits_{i=1}^{k} (2\eta_i \mathbb{E}[F(x_i) - f(x^*)]) \quad \leq \quad r_1 - r_{k+1} + M^2 \sum\nolimits_i \eta_i^2$$

SGD: weakly convex case

$$2\eta_k \mathbb{E}[F(x_k) - F(x^*)] \le r_k - r_{k+1} + \eta_k M^2.$$

Sum up over $i = 1, \ldots, k$, to obtain

$$\begin{aligned}\sum\nolimits_{i=1}^{k} (2\eta_i \mathbb{E}[F(x_i) - f(x^*)]) &\le r_1 - r_{k+1} + M^2 \sum\nolimits_i \eta_i^2 \\ &\le r_1 + M^2 \sum\nolimits_i \eta_i^2.\end{aligned}$$

SGD: weakly convex case

$$2\eta_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \eta_k M^2.$$

Sum up over $i = 1, \ldots, k$, to obtain

$$\begin{aligned} \sum\nolimits_{i=1}^{k} (2\eta_i \mathbb{E}[F(x_i) - f(x^*)]) &\leq r_1 - r_{k+1} + M^2 \sum\nolimits_i \eta_i^2 \\ &\leq r_1 + M^2 \sum\nolimits_i \eta_i^2. \end{aligned}$$

Divide both sides by $\sum_i \eta_i$, so

SGD: weakly convex case

$$2\eta_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \eta_k M^2.$$

Sum up over $i = 1, \ldots, k$, to obtain

$$\begin{aligned} \sum\nolimits_{i=1}^{k} (2\eta_i \mathbb{E}[F(x_i) - f(x^*)]) &\leq r_1 - r_{k+1} + M^2 \sum\nolimits_i \eta_i^2 \\ &\leq r_1 + M^2 \sum\nolimits_i \eta_i^2. \end{aligned}$$

Divide both sides by $\sum_i \eta_i$, so

- Set $\gamma_i = \frac{\eta_i}{\sum_i^k \eta_i}$.
- Thus, $\gamma_i \geq 0$ and $\sum_i \gamma_i = 1$

SGD: weakly convex case

$$2\eta_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \eta_k M^2.$$

Sum up over $i = 1, \ldots, k$, to obtain

$$\begin{aligned} \sum\nolimits_{i=1}^{k} (2\eta_i \mathbb{E}[F(x_i) - f(x^*)]) &\leq r_1 - r_{k+1} + M^2 \sum\nolimits_i \eta_i^2 \\ &\leq r_1 + M^2 \sum\nolimits_i \eta_i^2. \end{aligned}$$

Divide both sides by $\sum_i \eta_i$, so

- Set $\gamma_i = \frac{\eta_i}{\sum_i^k \eta_i}$.
- Thus, $\gamma_i \geq 0$ and $\sum_i \gamma_i = 1$

$$\mathbb{E}\left[\sum\nolimits_i \gamma_i (F(x_i) - F(x^*))\right] \leq \frac{r_1 + M^2 \sum_i \eta_i^2}{2 \sum_i \eta_i}$$

SGD: weakly convex case

- Bound looks similar to bound in subgradient method

SGD: weakly convex case

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k

SGD: weakly convex case

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$

SGD: weakly convex case

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$
- Easier to talk about **averaged**

$$\bar{x}_k := \sum\nolimits_i^k \gamma_i x_i.$$

SGD: weakly convex case

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$
- Easier to talk about **averaged**

$$\bar{x}_k := \sum_i^k \gamma_i x_i.$$

- $f(\bar{x}_k) \leq \sum_i \gamma_i F(x_i)$ due to convexity

SGD: weakly convex case

- Bound looks similar to bound in subgradient method
- But we wish to say something about x_k
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$
- Easier to talk about **averaged**

$$\bar{x}_k := \sum_i^k \gamma_i x_i.$$

- $f(\bar{x}_k) \leq \sum_i \gamma_i F(x_i)$ due to convexity

So we finally obtain the inequality

$$\boxed{\mathbb{E}\left[F(\bar{x}_k) - F(x^*)\right] \leq \frac{r_1 + M^2 \sum_i \eta_i^2}{2 \sum_i \eta_i}.}$$

SGD: weakly convex case

- Let $D_{\mathcal{X}} := \max_{x\in\mathcal{X}} \|x - x^*\|_2$ (act. only need $\|x_1 - x^*\| \le D_{\mathcal{X}}$)
- Assume $\eta_i = \eta$ is a constant. Observe that

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \le \frac{D_{\mathcal{X}}^2 + M^2 k\eta^2}{2k\eta}$$

- Minimize the rhs over $\eta > 0$ to obtain $\mathbb{E}[F(\bar{x}_k) - F(x^*)] \le \frac{D_{\mathcal{X}} M}{\sqrt{k}}$
- If k is not fixed in advance, then choose

$$\eta_i = \frac{\theta D_{\mathcal{X}}}{M\sqrt{i}}, \quad i = 1, 2, \ldots$$

- Analyze $\mathbb{E}[F(\bar{x}_k) - F(x^*)]$ with this choice of stepsize

SGD: weakly convex case

- Let $D_{\mathcal{X}} := \max_{x \in \mathcal{X}} \|x - x^*\|_2$ (act. only need $\|x_1 - x^*\| \leq D_{\mathcal{X}}$)
- Assume $\eta_i = \eta$ is a constant. Observe that

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D_{\mathcal{X}}^2 + M^2 k \eta^2}{2k\eta}$$

- Minimize the rhs over $\eta > 0$ to obtain $\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D_{\mathcal{X}} M}{\sqrt{k}}$
- If k is not fixed in advance, then choose

$$\eta_i = \frac{\theta D_{\mathcal{X}}}{M\sqrt{i}}, \quad i = 1, 2, \ldots$$

- Analyze $\mathbb{E}[F(\bar{x}_k) - F(x^*)]$ with this choice of stepsize

$$\boxed{\text{We showed } O(1/\sqrt{k}) \text{ rate}}$$

SGD: weakly convex and smooth

Exercise: Assuming the cost (and component functions) are L-smooth and convex, study the convergence rate of SGD.
Hint: Use bounded noise: $\mathbb{E}[\|\nabla f_{i_k}(x) - \nabla f(x)\|^2] = \sigma^2$.