Convex Optimization

(EE227A: UC Berkeley)

Lecture 14
(Gradient methods – II)

07 March, 2013

Suvrit Sra

Organizational

- Take home midterm: will be released on **18th March 2013** on bSpace by 5pm; Solutions (typeset) due **in class, 21st March, 2013** — no exceptions!
- Office hours: 2–4pm, Tuesday, 421 SDH (or by appointment)

- **1 page** project outline due on **3/14**
- Project page link (clickable)

- HW3 out on **3/14**; due on **4/02**
- HW4 out on **4/02**; due on **4/16**
- HW5 out on **4/16**; due on **4/30**

Convergence theory

Gradient descent – convergence

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

Gradient descent – convergence

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

Convergence

Theorem $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Gradient descent – convergence

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

Convergence

Theorem $\|\nabla f(x^k)\|_2 \to 0$ as $k \to \infty$

Convergence rate with constant stepsize

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

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 $f \in C^1_L$

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

- ♣ Gradient vectors of closeby points are close to each other
- ♣ Objective function has "bounded curvature"
- ♣ Speed at which gradient varies is bounded

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

Lemma (Descent). Let $f \in C_L^1$. Then,

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

Descent lemma – corollary

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

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

Descent lemma – corollary

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

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

Descent lemma – corollary

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

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

Descent lemma – corollary

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

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

Thus, if $\alpha_k < 2/L$ we have descent.

Descent lemma – corollary

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

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

Thus, if $\alpha_k < 2/L$ we have descent. Minimize over α_k to get best bound: this yields $\alpha_k = 1/L$—**we'll use this stepsize**

$$
\boxed{f(x^k) - f(x^{k+1}) \ge \alpha_k(1 - \tfrac{\alpha_k}{2}L)\|\nabla f(x^k)\|_2^2}
$$

Convergence

- Let's write the descent corollary as

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

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

Convergence

- Let's write the descent corollary as

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

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

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

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

Convergence

- Let's write the descent corollary as

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

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

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

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

Convergence

- Let's write the descent corollary as

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

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

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

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

- We assume $f^* > -\infty$, so rhs is some fixed positive constant

Convergence

- Let's write the descent corollary as

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

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

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

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

- 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

- Let's write the descent corollary as

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

($c = 1/2$ for $\alpha_k = 1/L$; c has diff. value for other stepsize rules)

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

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

- We assume $f^* > -\infty$, so rhs is some fixed positive constant
- Thus, as $k \to \infty$, lhs must converge; thus

$$\|\nabla f(x^k)\|_2 \to 0 \quad \text{as} \quad k \to \infty.$$

- Notice, we **did not require** f to be convex . . .

Descent lemma – another corollary

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

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

Exercise: Prove this corollary.

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ Shorthand notation $g^k = \nabla f(x^k)$, $g^* = \nabla f(x^*)$
- ⋆ Let $r_k := \|x^k - x^*\|_2$ (distance to optimum)

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ 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 – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ 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 – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ 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, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ 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, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

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

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ 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, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

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

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ 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, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

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

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ 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, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

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

Convergence rate – convex f

- ⋆ Let $\alpha_k = 1/L$
- ⋆ 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, $r_{k+1}^2 = \|x^{k+1} - x^*\|_2^2 = \|x^k - x^* - \alpha_k g^k\|_2^2$.

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

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

Convergence rate

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

Convergence rate

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

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

Convergence rate

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

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

Convergence rate

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

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

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

Convergence rate

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

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

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

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

Convergence rate

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

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

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

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

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

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

Convergence rate

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

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

Convergence rate

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

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

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

Convergence rate

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

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

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

$$\Longrightarrow \quad \frac{1}{\Delta_{k+1}} \ge \frac{1}{\Delta_k}(1+\beta) = \frac{1}{\Delta_k} + \frac{1}{2Lr_0^2}$$

Convergence rate

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

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

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

$$\implies \quad \frac{1}{\Delta_{k+1}} \ge \frac{1}{\Delta_k}(1+\beta) = \frac{1}{\Delta_k} + \frac{1}{2Lr_0^2}$$

- Sum both sides over $k = 0, \ldots, T$ to obtain

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

Convergence rate

- Sum both sides over $k = 0, \ldots, T$ to obtain

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

Convergence rate

- Sum both sides over $k = 0, \ldots, T$ to obtain

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

- Rearrange to conclude

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

Convergence rate

- Sum both sides over $k = 0, \ldots, T$ to obtain

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

- 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

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

Exercise: Prove above simplification.

Rates of convergence

Suppose a sequence $\left\{s^k\right\} \to s$.

Rates of convergence

Suppose a sequence $\{s^k\} \to s$.

- **Linear** If there is a constant $r \in (0,1)$ such that

$$\lim_{k\to\infty} \frac{\|s^{k+1}-s\|_2}{\|s^k - s\|_2} = r.$$

  i.e., distance decreases by **constant factor** at each iteration.

Rates of convergence

Suppose a sequence $\{s^k\} \to s$.

- **Linear** If there is a constant $r \in (0, 1)$ such that

$$\lim_{k\to\infty} \frac{\|s^{k+1} - s\|_2}{\|s^k - s\|_2} = r.$$

  i.e., distance decreases by **constant factor** at each iteration.
- **Sublinear** If $r = 1$ (constant factor decrease not there!)

Rates of convergence

Suppose a sequence $\{s^k\} \to s$.

- **Linear** If there is a constant $r \in (0,1)$ such that

$$\lim_{k\to\infty} \frac{\|s^{k+1} - s\|_2}{\|s^k - s\|_2} = r.$$

  i.e., distance decreases by **constant factor** at each iteration.
- **Sublinear** If $r = 1$ (constant factor decrease not there!)
- **Superlinear** If $r = 0$ (we rarely see this in large-scale opt)

Rates of convergence

Suppose a sequence $\{s^k\} \to s$.

- **Linear** If there is a constant $r \in (0,1)$ such that

$$\lim_{k\to\infty} \frac{\|s^{k+1} - s\|_2}{\|s^k - s\|_2} = r.$$

  i.e., distance decreases by **constant factor** at each iteration.

- **Sublinear** If $r = 1$ (constant factor decrease not there!)
- **Superlinear** If $r = 0$ (we rarely see this in large-scale opt)

Example 1. $\{1/k^c\}$: sublinear as $\lim k^c/(k+1)^c = 1$;
2. $\{sr^k\}$, where $|r| < 1$: linear with rate r

Gradient descent – faster rate

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

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

- Rarely do we have so much convexity!
- The extra convexity makes function "well-conditioned"

Gradient descent – faster rate

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

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

- ♣ Rarely do we have so much convexity!
- ♣ The extra convexity makes function "well-conditioned"
- ♣ **Exercise:** Prove strong convexity $\implies$ strict convexity

Gradient descent – faster rate

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

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

- Rarely do we have so much convexity!
- The extra convexity makes function "well-conditioned"
- **Exercise:** Prove strong convexity $\implies$ strict convexity
- C^1_L was sublinear; strong convexity leads linear rate

Strongly convex case – growth

Thm A. $f \in S^1_{L,\mu}$ is equivalent to

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \mu \|x - y\|_2^2 \qquad \forall\ x, y.$$

Exercise: Prove this claim.

Strongly convex case – growth

Thm A. $f \in S^1_{L,\mu}$ is equivalent to

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \mu \|x - y\|_2^2 \qquad \forall\ x, y.$$

Exercise: Prove this claim.

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

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

Strongly convex case – growth

Thm A. $f \in S^1_{L,\mu}$ is equivalent to

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \mu \|x - y\|_2^2 \qquad \forall\ x, y.$$

Exercise: Prove this claim.

Thm B. 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 $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$

Strongly convex case – growth

Thm A. $f \in S^1_{L,\mu}$ is equivalent to

$$\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geq \mu\|x-y\|_2^2 \qquad \forall\ x, y.$$

Exercise: Prove this claim.

Thm B. 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 $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- $\nabla\phi(x) = \nabla f(x) - \mu x$

Strongly convex case – growth

Thm A. $f \in S^1_{L,\mu}$ is equivalent to

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \mu \|x - y\|_2^2 \qquad \forall\ x, y.$$

Exercise: Prove this claim.

Thm B. 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 $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- $\nabla \phi(x) = \nabla f(x) - \mu x$
- If $\mu = L$, then easily true (due to Thm. A and Coroll. 2)

Strongly convex case – growth

Thm A. $f \in S^1_{L,\mu}$ is equivalent to

$$\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geq \mu \|x - y\|_2^2 \qquad \forall\; x, y.$$

Exercise: Prove this claim.

Thm B. 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 $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$
- $\nabla \phi(x) = \nabla f(x) - \mu x$
- If $\mu = L$, then easily true (due to Thm. A and Coroll. 2)
- If $\mu < L$, then $\phi \in C^1_{L-\mu}$; now invoke Coroll. 2

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

Strongly convex – rate

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

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

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

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

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

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

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

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

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

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

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

Strongly convex – rate

- As before, let $r_k = \|x^k - x^*\|_2$, and consider

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

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

Exercise: Complete the proof using above argument.

Gradient methods – lower bounds

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Gradient methods – lower bounds

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \le k \le \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^k) - f(x^*) \ge \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Theorem Lower bound II (Nesterov). For class of smooth, strongly convex, i.e., $S_{L,\mu}^{\infty}$ ($\mu > 0$, $\kappa > 1$)

$$f(x^k) - f(x^*) \ge \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\|x^0 - x^*\|_2^2.$$

Gradient methods – lower bounds

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \le k \le \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^k) - f(x^*) \ge \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Theorem Lower bound II (Nesterov). For class of smooth, strongly convex, i.e., $S^{\infty}_{L,\mu}$ ($\mu > 0$, $\kappa > 1$)

$$f(x^k) - f(x^*) \ge \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x^0 - x^*\|_2^2.$$

- **Notice gap between lower and upper bounds!**

Gradient methods – lower bounds

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Theorem Lower bound II (Nesterov). For class of smooth, strongly convex, i.e., $S_{L,\mu}^{\infty}$ ($\mu > 0$, $\kappa > 1$)

$$f(x^k) - f(x^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\|x^0 - x^*\|_2^2.$$

- **Notice gap between lower and upper bounds!**
- **We'll come back to these toward end of course**

Exercise

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

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

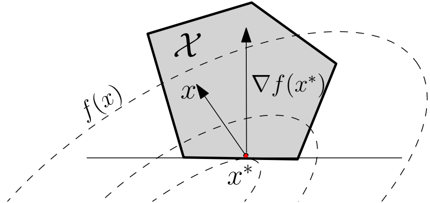
- $f(x) = \frac{1}{2}\|D^T x - b\|_2^2 = \frac{1}{2}(\|D^T x\|_2^2 + \|b\|_2^2 - 2\langle D^T x,\, b\rangle)$
- Try different choices of b, and different initial vectors x_0
- Determine L and μ for above $f(x)$ (nice linalg exercise!)
- **Exercise:** Try $\alpha = 2/(L+\mu)$ and other stepsize choices; report on empirical performance
- **Exercise:** Experiment to see how large n must be before gradient method starts outperforming CVX
- **Exercise:** Minimize $f(x)$ for large n; e.g., $n = 10^6$, $n = 10^7$

Constrained problems

Constrained optimization

$$\min \quad f(x) \quad \text{s.t. } x \in \mathcal{X}$$

$$\langle \nabla f(x^*),\, x - x^* \rangle \geq 0, \qquad \forall x \in \mathcal{X}.$$



Constrained optimization

$$x^{k+1} = x^k + \alpha_k d^k$$

Constrained optimization

$$x^{k+1} = x^k + \alpha_k d^k$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$

Constrained optimization

$$x^{k+1} = x^k + \alpha_k d^k$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

Constrained optimization

$$x^{k+1} = x^k + \alpha_k d^k$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

Since $\mathcal{X}$ is convex, all feasible directions are of the form

$$d^k = \gamma(z - x^k), \quad \gamma > 0,$$

where $z \in \mathcal{X}$ is any feasible vector.

Constrained optimization

$$x^{k+1} = x^k + \alpha_k d^k$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

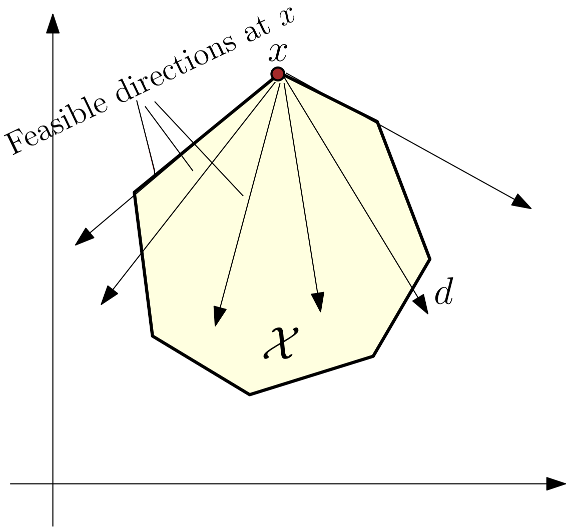
Since $\mathcal{X}$ is convex, all feasible directions are of the form

$$d^k = \gamma(z - x^k), \quad \gamma > 0,$$

where $z \in \mathcal{X}$ is any feasible vector.

$$x^{k+1} = x^k + \alpha_k(z^k - x^k), \quad \alpha_k \in (0, 1]$$

Cone of feasible directions



Conditional gradient method

Optimality: $\langle \nabla f(x^k),\, z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$

Conditional gradient method

Optimality: $\langle \nabla f(x^k),\, z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$
Aim: If not optimal, then generate feasible direction $d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k),\, d^k \rangle < 0$.

Conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$
Aim: If not optimal, then generate feasible direction $d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k), d^k \rangle < 0$.

Frank-Wolfe (Conditional gradient) method

- ▲ Let $z^k \in \operatorname{argmin}_{x \in \mathcal{X}} \langle \nabla f(x^k), x - x^k \rangle$
- ▲ Use different methods to select α_k
- ▲ $x^{k+1} = x^k + \alpha_k (z^k - x^k)$

Conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$
Aim: If not optimal, then generate feasible direction $d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k), d^k \rangle < 0$.

Frank-Wolfe (Conditional gradient) method

▲ Let $z^k \in \operatorname{argmin}_{x \in \mathcal{X}} \langle \nabla f(x^k), x - x^k \rangle$
▲ Use different methods to select α_k
▲ $x^{k+1} = x^k + \alpha_k(z^k - x^k)$

♠ Practical when easy to solve *linear* problem over $\mathcal{X}$.
♠ Currently enjoying huge renewed interest in machine learning.
♠ Several refinements, variants exist. (good for project)

Gradient projection

- FW method can be slow
- If $\mathcal{X}$ not compact, doesn't make sense
- A possible alternative (with other weaknesses though!)

Gradient projection

- FW method can be slow
- If $\mathcal{X}$ not compact, doesn't make sense
- A possible alternative (with other weaknesses though!)

If constraint set $\mathcal{X}$ is simple, i.e., we can easily solve projections

$$\min \quad \tfrac{1}{2}\|x - y\|_2 \quad \text{s.t.} \quad x \in \mathcal{X}.$$

Gradient projection

- FW method can be slow
- If $\mathcal{X}$ not compact, doesn't make sense
- A possible alternative (with other weaknesses though!)

If constraint set $\mathcal{X}$ is simple, i.e., we can easily solve projections

$$\min \quad \tfrac{1}{2}\|x - y\|_2 \quad \text{s.t.} \quad x \in \mathcal{X}.$$

$$x^{k+1} = P_{\mathcal{X}}\big(x^k - \alpha_k \nabla f(x^k)\big), \quad k = 0, 1, \ldots$$

where $P_{\mathcal{X}}$ denotes above orthogonal projection.

Gradient projection – convergence

Depends on the following crucial properties of P

Nonexpansivity: $\|Px - Py\|_2 \leq \|x - y\|_2$

Firm nonxpansivity: $\|Px - Py\|_2^2 \leq \langle Px - Py, x - y \rangle$

Gradient projection – convergence

Depends on the following crucial properties of P

Nonexpansivity: $\|Px - Py\|_2 \leq \|x - y\|_2$
Firm nonxpansivity: $\|Px - Py\|_2^2 \leq \langle Px - Py, x - y \rangle$

♡ Using the above, essentially convergence analysis with $\alpha_k = 1/L$ that we saw for the unconstrained case works.

Gradient projection – convergence

Depends on the following crucial properties of P

Nonexpansivity: $\|Px - Py\|_2 \leq \|x - y\|_2$
Firm nonxpansivity: $\|Px - Py\|_2^2 \leq \langle Px - Py, x - y \rangle$

♡ Using the above, essentially convergence analysis with $\alpha_k = 1/L$ that we saw for the unconstrained case works.

♡ Skipping for now; (see next slides though)

Gradient projection – convergence

Depends on the following crucial properties of P

Nonexpansivity: $\|Px - Py\|_2 \le \|x - y\|_2$
Firm nonxpansivity: $\|Px - Py\|_2^2 \le \langle Px - Py, x - y \rangle$

- ♡ Using the above, essentially convergence analysis with $\alpha_k = 1/L$ that we saw for the unconstrained case works.
- ♡ Skipping for now; (see next slides though)

Exercise: Recall $f(x) = \frac{1}{2}\|D^T x - b\|_2^2$. Write a matlab script to minimize this function over the convex set $\mathcal{X} := \{-1 \le x_i \le 1\}$.

Projection lemma

Theorem Orthogonal projection is firmly nonexpansive

$$\langle Px - Py, x - y \rangle \leq \|x - y\|_2^2$$

Projection lemma

Theorem Orthogonal projection is firmly nonexpansive

$$\langle Px - Py, x - y \rangle \leq \|x - y\|_2^2$$

Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$ (necc and suff)

Projection lemma

Theorem Orthogonal projection is firmly nonexpansive

$$\langle Px - Py, x - y\rangle \le \|x - y\|_2^2$$

Recall: $\langle \nabla f(x^*), x - x^*\rangle \ge 0$ for all $x \in \mathcal{X}$ (necc and suff)

$$\langle Px - Py, y - Py\rangle \quad \le \quad 0$$

Projection lemma

Theorem Orthogonal projection is firmly nonexpansive

$$\langle Px - Py, x - y \rangle \leq \|x - y\|_2^2$$

Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$ (necc and suff)

$$\begin{aligned} \langle Px - Py, y - Py \rangle &\leq 0 \\ \langle Px - Py, Px - x \rangle &\leq 0 \end{aligned}$$

Projection lemma

Theorem Orthogonal projection is firmly nonexpansive

$$\langle Px - Py, x - y \rangle \leq \|x - y\|_2^2$$

Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$ (necc and suff)

$$\begin{aligned} \langle Px - Py, y - Py \rangle &\leq 0 \\ \langle Px - Py, Px - x \rangle &\leq 0 \\ \langle Px - Py, Px - Py \rangle &\leq \langle Px - Py, x - y \rangle \end{aligned}$$

Projection lemma

Theorem Orthogonal projection is firmly nonexpansive

$$\langle Px - Py, x - y\rangle \le \|x - y\|_2^2$$

Recall: $\langle \nabla f(x^*), x - x^*\rangle \ge 0$ for all $x \in \mathcal{X}$ (necc and suff)

$$\begin{aligned} \langle Px - Py, y - Py\rangle &\le 0 \\ \langle Px - Py, Px - x\rangle &\le 0 \\ \langle Px - Py, Px - Py\rangle &\le \langle Px - Py, x - y\rangle \end{aligned}$$

Both nonexpansivity and firm nonexpansivity follow after invoking Cauchy-Schwarz

Gradient projection – convergence hints

$$f(x^{k+1}) \quad \leq \quad f(x^k) + \langle g^k,\, x^{k+1} - x^k \rangle + \tfrac{L}{2} \|x^{k+1} - x^k\|_2^2$$

Gradient projection – convergence hints

$$f(x^{k+1}) \quad \leq \quad f(x^k) + \langle g^k,\, x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2^2$$

Let us look at the latter two terms above:

$$\langle g^k,\, P(x^k - \alpha_k g^k) - P(x^k) \rangle \quad + \quad \tfrac{L}{2}\|P(x^k - \alpha_k g^k) - P(x^k)\|_2^2$$

Gradient projection – convergence hints

$$f(x^{k+1}) \quad \le \quad f(x^k) + \langle g^k, x^{k+1} - x^k\rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2^2$$

Let us look at the latter two terms above:

$$\begin{aligned}\langle g^k, P(x^k - \alpha_k g^k) - P(x^k)\rangle \quad &+ \quad \tfrac{L}{2}\|P(x^k - \alpha_k g^k) - P(x^k)\|_2^2 \\ \langle P(x - \alpha g) - Px, -\alpha g\rangle \quad &\le \quad \|\alpha g\|_2^2\end{aligned}$$

Gradient projection – convergence hints

$$f(x^{k+1}) \quad \leq \quad f(x^k) + \langle g^k, x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2^2$$

Let us look at the latter two terms above:

$$\begin{aligned}
\langle g^k, P(x^k - \alpha_k g^k) - P(x^k) \rangle \quad &+ \quad \tfrac{L}{2}\|P(x^k - \alpha_k g^k) - P(x^k)\|_2^2 \\
\langle P(x - \alpha g) - Px, -\alpha g \rangle \quad &\leq \quad \|\alpha g\|_2^2 \\
\langle P(x - \alpha g) - Px, g \rangle \quad &\geq \quad -\alpha\|g\|_2^2
\end{aligned}$$

Gradient projection – convergence hints

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

Let us look at the latter two terms above:

$$\begin{aligned}
\langle g^k, P(x^k - \alpha_k g^k) - P(x^k) \rangle &+ \tfrac{L}{2}\|P(x^k - \alpha_k g^k) - P(x^k)\|_2^2 \\
\langle P(x - \alpha g) - Px, -\alpha g \rangle &\leq \|\alpha g\|_2^2 \\
\langle P(x - \alpha g) - Px, g \rangle &\geq -\alpha\|g\|_2^2 \\
\tfrac{L}{2}\|P(x - \alpha g) - Px\|_2^2 &\leq \tfrac{L}{2}\alpha^2\|g\|_2^2
\end{aligned}$$

Optimal gradient methods

Optimal gradient methods

We saw *upper bounds:* $O(1/T)$, and linear rate involving κ

We saw *lower bounds:* $O(1/T^2)$, and linear rate involving $\sqrt{\kappa}$

Optimal gradient methods

We saw *upper bounds:* $O(1/T)$, and linear rate involving κ

We saw *lower bounds:* $O(1/T^2)$, and linear rate involving $\sqrt{\kappa}$

Can we close the gap?

Optimal gradient methods

We saw *upper bounds:* $O(1/T)$, and linear rate involving κ

We saw *lower bounds:* $O(1/T^2)$, and linear rate involving $\sqrt{\kappa}$

Can we close the gap?

Nesterov (1983) closed the gap!

Optimal gradient methods

We saw *upper bounds:* $O(1/T)$, and linear rate involving κ

We saw *lower bounds:* $O(1/T^2)$, and linear rate involving $\sqrt{\kappa}$

Can we close the gap?

Nesterov (1983) closed the gap!

Note 1: Don't insist on $f(x_{k+1}) \leq f(x_k)$
Note 2: Use "multi-steps"

Nesterov Accelerated gradient method

1. Choose $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y_0 \leftarrow x_0$, $q = \mu/L$

Nesterov Accelerated gradient method

1. Choose $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y_0 \leftarrow x_0$, $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   - Compute $f(y_k)$ and $\nabla f(y_k)$
     Let $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

Nesterov Accelerated gradient method

1. Choose $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y_0 \leftarrow x_0$, $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   - Compute $f(y_k)$ and $\nabla f(y_k)$
     Let $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
   - Obtain α_{k+1} by solving
     $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$

Nesterov Accelerated gradient method

1. Choose $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y_0 \leftarrow x_0$, $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   - Compute $f(y_k)$ and $\nabla f(y_k)$
     Let $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
   - Obtain α_{k+1} by solving
     $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$
   - Let $\beta_k = \alpha_k(1 - \alpha_k)/(\alpha_k^2 + \alpha_{k+1})$, and set
     $y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$

Nesterov Accelerated gradient method

1. Choose $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$
2. Let $y_0 \leftarrow x_0$, $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   - Compute $f(y_k)$ and $\nabla f(y_k)$
     Let $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
   - Obtain α_{k+1} by solving
     $\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$
   - Let $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$, and set
     $y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$

If $\alpha_0 \geq \sqrt{\mu/L}$, then

$$f(x_T) - f(x^*) \leq c_1 \min\left\{\left(1 - \sqrt{\frac{\mu}{L}}\right)^T, \frac{4L}{(2\sqrt{L} + c_2 T)^2}\right\},$$

where constants c_1, c_2 depend on α_0, L, μ.

Strong-convexity – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. Algo becomes

Strong-convexity – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. Algo becomes

1. Choose $y_0 = x_0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):

Strong-convexity – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. Algo becomes

1. Choose $y_0 = x_0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):
   - $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
   - $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$
     $y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k)$

A simple multi-step method!

References

1. Y. Nesterov. *Introductory lectures on convex optimization*
2. D. Bertsekas. *Nonlinear programming*