Recap

\[ x \leftarrow x - \eta \nabla f(x) \]
Gradient methods – upper bounds

**Theorem.** (Upper bound I). Let $f \in C_L^1$. Then,

$$\min_k \| \nabla f(x^k) \| \leq \varepsilon \text{ in } O(1/\varepsilon^2) \text{ iterations.}$$
**Theorem.** (Upper bound I). Let $f \in C^1_L$. Then,

$$\min_k \| \nabla f(x^k) \| \leq \varepsilon \text{ in } O(1/\varepsilon^2) \text{ iterations.}$$

**Theorem.** (Upper bound II). Let $f \in S^1_{L, \mu}$. Then,

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

**Theorem.** (Upper bound I). Let $f \in C^1_L$. Then,
\[ \min_k \| \nabla f(x^k) \| \leq \varepsilon \text{ in } O(1/\varepsilon^2) \text{ iterations.} \]

**Theorem.** (Upper bound II). Let $f \in S^1_{L, \mu}$. Then,
\[ f(x^k) - f(x^*) \leq \frac{L}{2} \left( \frac{\kappa - 1}{\kappa + 1} \right)^{2k} \| x^0 - x^* \|_2^2 \]

**Theorem.** (Upper bound III). Let $f \in C^1_L$ be convex. Then,
\[ f(x^k) - f(x^*) \leq \frac{2L(f(x^0) - f(x^*))}{k + 4} \| x^0 - x^* \|_2^2. \]
**Theorem.** (Carmon-Duchi-Hinder-Sidford 2017). There exists an $f \in C^1_L$, such that $\|\nabla f(x)\| \leq \varepsilon$ requires $\Omega(\varepsilon^{-2})$ gradient evaluations.
Theorem. (Carmon-Duchi-Hinder-Sidford 2017). There exists an $f \in C^1_L$, such that $\|\nabla f(x)\| \leq \varepsilon$ requires $\Omega(\varepsilon^{-2})$ gradient evaluations.

Theorem. (Nesterov). There exists $f \in S^\infty_{L,\mu}$ ($\mu > 0$, $\kappa > 1$) s.t.

$$f(x^k) - f(x^*) \geq \frac{\mu}{2} \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} \|x^0 - x^*\|_2^2,$$
Theorem. (Carmon-Duchi-Hinder-Sidford 2017). There exists an $f \in C^1_L$, such that $\|\nabla f(x)\| \leq \varepsilon$ requires $\Omega(\varepsilon^{-2})$ gradient evaluations.

Theorem. (Nesterov). There exists $f \in S_{L,\mu}^\infty (\mu > 0, \kappa > 1)$ s.t.

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

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

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

$$\|x^k - x^0\|^2 \geq \frac{1}{8}\|x^0 - x^*\|^2.$$
Fast gradient methods
We saw upper bounds: $O(1/k)$, and linear rate involving $\kappa$

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

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

Can we close the gap?
Accelerated gradient methods

We saw upper bounds: $O(1/k)$, and linear rate involving $\kappa$

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

Can we close the gap?

Nesterov (1983) closed the gap!
Accelerated gradient methods

We saw upper bounds: $O(1/k)$, and linear rate involving $\kappa$

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

Can we close the gap?

Nesterov (1983) closed the gap!

Key idea 1: Don’t insist on $f(x_{k+1}) \leq f(x_k)$

Key idea 2: “multi-step” method
If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. Nesterov’s method is:

\[
\begin{align*}
    x_{k+1} &= y_k - \frac{1}{L} \nabla f(y_k) \\
    y_{k+1} &= x_k + 1 + \sqrt{L} - \sqrt{\mu} \sqrt{L} + \sqrt{\mu} (x_{k+1} - x_k)
\end{align*}
\]
If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. Nesterov’s method is:

1. Choose $y_0 = x_0 \in \mathbb{R}^n$
2. $k$-th iteration ($k \geq 0$):
If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. Nesterov’s method is:

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

Why does it work? Why does multi-step help?
Accelerated gradient – strongly convex

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu / L}$. Nesterov’s method is:

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

Why does it work? Why does multi-step help?

► Acceleration a big topic; several competing ways on deriving and analyzing accelerated methods. Still worthy of further study.
Nesterov Accelerated gradient method

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

Let $y_0 \leftarrow x_0$, $q = \mu/L$

$k$-th iteration ($k \geq 0$):

- 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 ($k \geq 0$):
   - Let $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$
   - Obtain $\alpha_{k+1}$ from: $\alpha_{k+1}^2 = (1 - \alpha_{k+1}) \alpha_k^2 + q \alpha_{k+1}$
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 ($k \geq 0$):
   - Let $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$
   - Obtain $\alpha_{k+1}$ from: $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$
   - Let $\beta_k = \alpha_k(1 - \alpha_k)/(\alpha_k^2 + \alpha_{k+1})$, and set
     $y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$

Exercise:
Choose $\alpha_k$ and $\beta_k$ to obtain method on the previous slide

Toy exercise:
Can we go even faster with more than 1 “multi-step”?
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 ($k \geq 0$):
   - Let $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$
   - Obtain $\alpha_{k+1}$ from: $\alpha_{k+1}^2 = (1 - \alpha_{k+1}) \alpha_k^2 + q \alpha_{k+1}$
   - 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

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

where constants $c_1, c_2$ depend on $\alpha_0, L, \mu$.

**Exercise:** Choose $\alpha_k$ and $\beta_k$ to obtain method on the previous slide.
Choose $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$

Let $y_0 \leftarrow x_0$, $q = \mu / L$

$k$-th iteration ($k \geq 0$):

- Let $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$
- Obtain $\alpha_{k+1}$ from: $\alpha_{k+1}^2 = (1 - \alpha_{k+1}) \alpha_k^2 + q \alpha_{k+1}$
- 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

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

where constants $c_1, c_2$ depend on $\alpha_0, L, \mu$.

**Exercise:** Choose $\alpha_k$ and $\beta_k$ to obtain method on the previous slide

**Toy exercise:** Can we go even faster with more than 1 “multi-step”?
Subgradient methods
Unconstrained convex problem

$$\min_x f(x)$$

How to optimize?

1. Start with some guess $$x_0$$; set $$k = 0$$
2. If $$0 \in \partial f(x_k)$$, stop; output $$x_k$$
3. Otherwise, generate next guess $$x_{k+1}$$
4. Repeat above procedure

▶ In reality: we stop in finite time
▶ Only solve problem approximately
▶ $$f(x_k) \leq f(x^*) + \varepsilon$$

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (3/02/20; Lecture 8)
Unconstrained convex problem

\[
\min_x f(x)
\]

How to optimize?

1. Start with some guess \( x^0 \); set \( k = 0 \)
Unconstrained convex problem

\[ \min_{x} f(x) \]

How to optimize?

1. Start with some guess \( x^0 \); set \( k = 0 \)
2. If \( 0 \in \partial f(x^k) \), stop; output \( x^k \)
Unconstrained convex problem

\[ \min_x f(x) \]

How to optimize?

1. Start with some guess \( x^0 \); set \( k = 0 \)
2. If \( 0 \in \partial f(x^k) \), stop; output \( x^k \)
3. Otherwise, generate next guess \( x^{k+1} \)

In reality: we stop in finite time
Only solve problem approximately

\[ f(x^k) \leq f(x^\ast) + \epsilon \]

shorthand
\[ f_k \leq f^\ast + \epsilon \]

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (3/02/20; Lecture 8)
Unconstrained convex problem

\[
\min_x f(x)
\]

How to optimize?

1. Start with some guess \(x^0\); set \(k = 0\)
2. If \(0 \in \partial f(x^k)\), stop; output \(x^k\)
3. Otherwise, generate next guess \(x^{k+1}\)
4. Repeat above procedure
Unconstrained convex problem

\[
\min_x f(x)
\]

How to optimize?

1. Start with some guess \(x^0\); set \(k = 0\)
2. If \(0 \in \partial f(x^k)\), **stop**; output \(x^k\)
3. Otherwise, generate next guess \(x^{k+1}\)
4. Repeat above procedure
   - In reality: we stop in finite time
   - Only solve problem approximately
   - \(f(x^k) \leq f(x^*) + \varepsilon\)
   - **shorthand** \(f^k \leq f^* + \varepsilon\)
Subgradient method

\[ x^{k+1} = x^k - \alpha_k g^k \]

where \( g^k \in \partial f(x^k) \) is any subgradient
Subgradient method

\[ x^{k+1} = x^k - \alpha_k g^k \]

where \( g^k \in \partial f(x^k) \) is any subgradient

**Stepsize** \( \alpha_k > 0 \) must be chosen
Subgradient method

\[ x^{k+1} = x^k - \alpha_k g^k \]

where \( g^k \in \partial f(x^k) \) is any subgradient

**Stepsize** \( \alpha_k > 0 \) must be chosen

- Method generates sequence \( \{x^k\}_{k \geq 0} \)
- Does this sequence converge to an optimal solution \( x^* \)?
- If yes, then how fast?
- What if have constraints: \( x \in C \)?
Example

\[ \min \quad \frac{1}{2} \|Ax - b\|^2_2 + \lambda \|x\|_1 \]

\[ x^{k+1} = x^k - \alpha_k (A^T(Ax^k - b) + \lambda \text{sgn}(x^k)) \]
Example

\[
\min \quad \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1 \\
x^{k+1} = x^k - \alpha_k (A^T(Ax^k - b) + \lambda \text{sgn}(x^k))
\]

(More careful implementation)
Subgradient method – stepsizes

- **Constant** Set \( \alpha_k = \alpha > 0, \text{ for } k \geq 0 \)
- **Scaled constant** \( \alpha_k = \alpha / \|g^k\|_2 \) (\( \|x^{k+1} - x^k\|_2 = \alpha \))
Subgradient method – stepsizes

- **Constant** Set $\alpha_k = \alpha > 0$, for $k \geq 0$

- **Scaled constant** $\alpha_k = \alpha / \| g^k \|_2$ ($\| x^{k+1} - x^k \|_2 = \alpha$)

- **Square summable but not summable**
  \[ \sum_k \alpha_k^2 < \infty, \quad \sum_k \alpha_k = \infty \]

- **Diminishing scalar**
  \[ \lim_{k} \alpha_k = 0, \quad \sum_k \alpha_k = \infty \]

- **Adaptive stepsizes** (not covered)

  Not a descent method!
  Work with best $f^k$ so far: $f^k_{\min} := \min_{0 \leq i \leq k} f^i$
Assumptions

- Min is attained: \( f^\ast := \inf_x f(x) > -\infty \), with \( f(x^\ast) = f^\ast \)
Convergence analysis

Assumptions

► Min is attained: \( f^* := \inf_x f(x) > -\infty \), with \( f(x^*) = f^* \)

► Bounded subgradients: \( \|g\|_2 \leq G \) for all \( g \in \partial f \)
  \( (f(x) - f(y) = \langle g_\xi, x - y \rangle; \text{use Cauchy-Schwarz or Hölder}) \)
Assumptions

- Min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$
- Bounded subgradients: $\|g\|_2 \leq G$ for all $g \in \partial f$
  
  \[ f(x) - f(y) = \langle g_\xi, x - y \rangle; \text{ use Cauchy-Schwarz or Hölder} \]

- Bounded domain: $\|x^0 - x^*\|_2 \leq R$
Convergence analysis

Assumptions

- Min is attained: \( f^* := \inf_x f(x) > -\infty \), with \( f(x^*) = f^* \)
- Bounded subgradients: \( \|g\|_2 \leq G \) for all \( g \in \partial f \)

\[ f(x) - f(y) = \langle g_\xi, x - y \rangle; \text{ use Cauchy-Schwarz or Hölder} \]
- Bounded domain: \( \|x^0 - x^*\|_2 \leq R \)

Convergence results for: \( f_{\text{min}}^k := \min_{0 \leq i \leq k} f^i \)
Lyapunov function: Distance to $x^*$, not function values

\[
\|x_{k+1} - x^*\|^2_2 \\
\leq \|x_k - x^*\|^2_2 \\
+ \alpha_k^2 \|g_k\|^2_2 \\
- 2\alpha_k (f(x_k) - f^*) \\
\]

since $f^* = f(x^*) \geq f(x_k) + \langle g_k, x^* - x_k \rangle$

Apply same argument to $\|x_k - x^*\|^2_2$ recursively

\[
\|x_{k+1} - x^*\|^2_2 \leq \|x_0 - x^*\|^2_2 \\
+ \sum_{t=1}^{\infty} \alpha_t^2 \|g_t\|^2_2 \\
- 2\sum_{t=1}^{\infty} \alpha_t (f_t - f^*) \\
\]

Now use our convenient assumptions!
Subgradient method – convergence

Lyapunov function: Distance to $x^*$, not function values

$$\|x^{k+1} - x^*\|^2_2 = \|x^k - \alpha_k g^k - x^*\|^2_2$$
Subgradient method – convergence

Lyapunov function: Distance to $x^*$, not function values

\[
\|x^{k+1} - x^*\|_2^2 = \|x^k - \alpha_k g^k - x^*\|_2^2 \\
= \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\langle \alpha_k g^k, x^k - x^* \rangle
\]
Subgradient method – convergence

Lyapunov function: Distance to $x^*$, not function values

\[ \|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 + \alpha_k^2 \|g^k\|^2 - 2\alpha_k (f(x^k) - f^*) , \]

since $f^* = f(x^*) \geq f(x^k) + \langle g^k, x^* - x^k \rangle$
Lyapunov function: Distance to $x^*$, not function values

\[
\|x^{k+1} - x^*\|_2^2 = \|x^k - \alpha_k g^k - x^*\|_2^2 \\
= \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\langle \alpha_k g^k, x^k - x^* \rangle \\
\leq \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\alpha_k (f(x^k) - f^*) ,
\]

since $f^* = f(x^*) \geq f(x^k) + \langle g^k, x^* - x^k \rangle$

Apply same argument to $\|x^k - x^*\|_2^2$ recursively
Subgradient method – convergence

Lyapunov function: Distance to $x^*$, not function values

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

since $f^* = f(x^*) \geq f(x^k) + \langle g^k, x^* - x^k \rangle$

Apply same argument to $\|x^k - x^*\|_2^2$ recursively

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^0 - x^*\|_2^2 + \sum_{t=1}^{k} \alpha_t^2 \|g^t\|_2^2 - 2 \sum_{t=1}^{k} \alpha_t (f^t - f^*)$$
Subgradient method – convergence

Lyapunov function: Distance to $x^*$, not function values

$$\|x^{k+1} - x^*\|^2_2 = \|x^k - \alpha_k g^k - x^*\|^2_2$$

$$\|x^k - x^*\|^2_2 + \alpha_k^2 \|g^k\|^2_2 - 2 \langle \alpha_k g^k, x^k - x^* \rangle$$

$$\leq \|x^k - x^*\|^2_2 + \alpha_k^2 \|g^k\|^2_2 - 2 \alpha_k (f(x^k) - f^*)$$

since $f^* = f(x^*) \geq f(x^k) + \langle g^k, x^* - x^k \rangle$

Apply same argument to $\|x^k - x^*\|^2_2$ recursively

$$\|x^{k+1} - x^*\|^2_2 \leq \|x^0 - x^*\|^2_2 + \sum_{t=1}^{k} \alpha_t^2 \|g^k\|^2_2 - 2 \sum_{t=1}^{k} \alpha_t (f^t - f^*)$$

Now use our convenient assumptions!
Subgradient method – convergence

\[ \|x^{k+1} - x^*\|^2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2 \sum_{t=1}^{k} \alpha_t (f^t - f^*). \]

To get a bound on the last term, simply notice (for \( t \leq k \))

\[ f^t \geq f^t_{\min} \geq f_{\min} \]

since \( f_{\min} := \min_{0 \leq i \leq t} f(x^i) \)
\[ \|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2 \sum_{t=1}^{k} \alpha_t (f^t - f^*) . \]

To get a bound on the last term, simply notice (for \( t \leq k \))

\[ f^t \geq f^t_{\min} \geq f^k_{\min} \quad \text{since} \quad f^t_{\min} := \min_{0 \leq i \leq t} f(x^i) \]

\[ \sum_{t=1}^{k} \alpha_t (f^t - f^*) \geq 2(f^k_{\min} - f^*) \sum_{t=1}^{k} \alpha_t . \]
Subgradient method – convergence

\[ \|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2 \sum_{t=1}^{k} \alpha_t (f^t - f^*). \]

- To get a bound on the last term, simply notice (for \( t \leq k \))

\[ f^t \geq f^t_{\min} \geq f^k_{\min} \quad \text{since} \quad f^t_{\min} := \min_{0 \leq i \leq t} f(x^i) \]

- Plugging this in yields the bound

\[ 2 \sum_{t=1}^{k} \alpha_t (f^t - f^*) \geq 2(f^k_{\min} - f^*) \sum_{t=1}^{k} \alpha_t. \]

- So that we finally have

\[ 0 \leq \|x^{k+1} - x^*\|_2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2(f^k_{\min} - f^*) \sum_{t=1}^{k} \alpha_t \]
Subgradient method – convergence

\[ \|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2 \sum_{t=1}^{k} \alpha_t (f^t - f^*). \]

To get a bound on the last term, simply notice (for \( t \leq k \))

\[ f^t \geq f^t_{\min} \geq f^k_{\min} \quad \text{since} \quad f^t_{\min} := \min_{0 \leq i \leq t} f(x^i) \]

Plugging this in yields the bound

\[ 2 \sum_{t=1}^{k} \alpha_t (f^t - f^*) \geq 2(f^k_{\min} - f^*) \sum_{t=1}^{k} \alpha_t. \]

So that we finally have

\[ 0 \leq \|x^{k+1} - x^*\|_2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2(f^k_{\min} - f^*) \sum_{t=1}^{k} \alpha_t \]

\[ f^k_{\min} - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \]
Subgradient method – convergence

\[
\begin{align*}
f_k^{\min} - f^* & \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t}
\end{align*}
\]

**Exercise:** Analyze \( \lim_{k \to \infty} f_k^{\min} - f^* \) for the different choices of stepsize that we mentioned.
Subgradient method – convergence

\[ f^k_{\text{min}} - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \]

**Exercise:** Analyze \( \lim_{k \to \infty} f^k_{\text{min}} - f^* \) for the different choices of stepsize that we mentioned.

**Constant step:** \( \alpha_k = \alpha \); We obtain

\[ f^k_{\text{min}} - f^* \leq \frac{R^2 + G^2 k\alpha^2}{2k\alpha} \]
Subgradient method – convergence

\[ f^k_{\text{min}} - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \]

Exercise: Analyze \( \lim_{k \to \infty} f^k_{\text{min}} - f^* \) for the different choices of stepsize that we mentioned.

**Constant step:** \( \alpha_k = \alpha \); We obtain

\[ f^k_{\text{min}} - f^* \leq \frac{R^2 + G^2 k \alpha^2}{2k \alpha} \to \frac{G^2 \alpha}{2} \text{ as } k \to \infty. \]
Subgradient method – convergence

\[ f_k^{\min} - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \]

**Exercise:** Analyze \( \lim_{k \to \infty} f_k^{\min} - f^* \) for the different choices of stepsize that we mentioned.

**Constant step:** \( \alpha_k = \alpha \); We obtain

\[ f_k^{\min} - f^* \leq \frac{R^2 + G^2 k \alpha^2}{2k \alpha} \rightarrow \frac{G^2 \alpha}{2} \quad \text{as } k \rightarrow \infty. \]

**Square summable, not summable:** \( \sum_k \alpha_k^2 < \infty, \sum_k \alpha_k = \infty \)
Subgradient method – convergence

\[ f^k_{\text{min}} - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \]

**Exercise:** Analyze \( \lim_{k \to \infty} f^k_{\text{min}} - f^* \) for the different choices of stepsize that we mentioned.

**Constant step:** \( \alpha_k = \alpha \); We obtain

\[ f^k_{\text{min}} - f^* \leq \frac{R^2 + G^2 k \alpha^2}{2k\alpha} \rightarrow \frac{G^2 \alpha}{2} \quad \text{as } k \to \infty. \]

**Square summable, not summable:** \( \sum_k \alpha_k^2 < \infty, \sum_k \alpha_k = \infty \)

As \( k \to \infty \), numerator \(< \infty\) but denominator \( \to \infty \); so \( f^k_{\text{min}} \to f^* \)
Subgradient method – convergence

\[
f^k_{\min} - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t}
\]

**Exercise:** Analyze \( \lim_{k \to \infty} f^k_{\min} - f^* \) for the different choices of stepsize that we mentioned.

**Constant step:** \( \alpha_k = \alpha \); We obtain

\[
f^k_{\min} - f^* \leq \frac{R^2 + G^2 k \alpha^2}{2k \alpha} \to \frac{G^2 \alpha}{2} \text{ as } k \to \infty.
\]

**Square summable, not summable:** \( \sum_k \alpha_k^2 < \infty, \sum_k \alpha_k = \infty \)

As \( k \to \infty \), numerator \( < \infty \) but denominator \( \to \infty \); so \( f^k_{\min} \to f^* \)

In practice, fair bit of stepsize tuning needed, e.g. \( \alpha_t = a/(b + t) \)
Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should $k$ be?
Suppose we want $f_{\text{min}}^k - f^* \leq \varepsilon$, how big should $k$ be?

Optimize the bound for $\alpha_t$: want

$$f_{\text{min}}^k - f^* \leq \varepsilon$$
Subgradient method – convergence

- Suppose we want $f_{\text{min}}^k - f^* \leq \varepsilon$, how big should $k$ be?
- Optimize the bound for $\alpha_t$: want

  $$f_{\text{min}}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \leq \varepsilon$$

- For fixed $k$: best possible stepsize is constant

- Then, after $k$ steps $f_{\text{min}}^k - f^* \leq \frac{RG}{\sqrt{k}}$.

- For accuracy $\varepsilon$, we need at least $\left(\frac{RG}{\varepsilon}\right)^2 = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ steps (quite slow)
Subgradient method – convergence

► Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should $k$ be?

► Optimize the bound for $\alpha_t$: want

$$f_{\min}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2 \sum_{t=1}^k \alpha_t} \leq \varepsilon$$

► For fixed $k$: best possible stepsize is constant $\alpha$

$$\frac{R^2 + G^2 k \alpha^2}{2k \alpha} \leq \varepsilon \quad \Rightarrow \quad \alpha = \frac{R}{G \sqrt{k}}$$
Subgradient method – convergence

- Suppose we want \( f_{\min}^k - f^* \leq \varepsilon \), how big should \( k \) be?
- Optimize the bound for \( \alpha_t \): want

\[
\frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \leq \varepsilon
\]

- For fixed \( k \): best possible stepsize is constant \( \alpha \)

\[
\frac{R^2 + G^2 k \alpha^2}{2k \alpha} \leq \varepsilon \quad \Rightarrow \quad \alpha = \frac{R}{G \sqrt{k}}
\]

- Then, after \( k \) steps \( f_{\min}^k - f^* \leq RG/\sqrt{k} \).
- For accuracy \( \varepsilon \), we need at least \((RG/\varepsilon)^2 = O(1/\varepsilon^2)\) steps
Suppose we want $f_{\text{min}}^k - f^* \leq \varepsilon$, how big should $k$ be?

Optimize the bound for $\alpha_t$: want

$$f_{\text{min}}^k - f^* \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2 \sum_{t=1}^{k} \alpha_t} \leq \varepsilon$$

For fixed $k$: best possible stepsize is constant $\alpha$

$$\frac{R^2 + G^2 k \alpha^2}{2 k \alpha} \leq \varepsilon \quad \Rightarrow \quad \alpha = \frac{R}{G \sqrt{k}}$$

Then, after $k$ steps $f_{\text{min}}^k - f^* \leq RG/\sqrt{k}$.

For accuracy $\varepsilon$, we need at least $(RG/\varepsilon)^2 = O(1/\varepsilon^2)$ steps

(quite slow)
Exercise

Support vector machines

Let $\mathcal{D} := \{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \{\pm 1\}\}$

We wish to find $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$\min_{w, b} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{m} \max[0, 1 - y_i(w^T x_i + b)]$$

Derive and implement a subgradient method

Plot evolution of objective function

Experiment with different values of $C > 0$

Plot and keep track of $f_{\min}^k := \min_{0 \leq t \leq k} f(x^t)$
Polyak’s stepsize

Assume $f^*$ is known (or can be estimated). Then use

$$\alpha_t = \frac{f^t - f^*}{\|g^t\|_2^2}$$
Polyak’s stepsize

▶ Assume $f^*$ is known (or can be estimated). Then use

$$
\alpha_t = \frac{f^t - f^*}{\|g^t\|^2/2}
$$

▶ Motivation: recall bound

$$
\|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - 2\alpha_t(f^t - f^*) + \alpha_t^2\|g^t\|^2
$$

and minimize RHS.
Polyak’s stepsize

- Assume $f^*$ is known (or can be estimated). Then use

$$
\alpha_t = \frac{f^t - f^*}{\|g^t\|_2^2}
$$

- Motivation: recall bound

$$
\|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - 2\alpha_t (f^t - f^*) + \alpha_t^2 \|g^t\|^2
$$

and minimize RHS.

- Let’s plug in $\alpha_t$:

$$
\|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - \frac{(f^t - f^*)^2}{\|g_t\|^2}
$$
Polyak’s stepsize

\[ \| x^{t+1} - x^* \|^2 \leq \| x^t - x^* \|^2 - \frac{(f^t - f^*)^2}{\| g_t \|^2} \]
Polyak’s stepsize

\[ ||x^{t+1} - x^*||^2 \leq ||x^t - x^*||^2 - \frac{(f^t - f^*)^2}{||g_t||^2} \]

- **Observation 1**  \( ||x^t - x^*|| \) decreases
Polyak’s stepsize

\[ \|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - \frac{(f^t - f^*)^2}{\|g_t\|^2} \]

- **Observation 1** \( \|x^t - x^*\| \) decreases
- **Recursion:**

\[
\sum_{t=1}^{k} \frac{(f^t - f^*)^2}{\|g_t\|^2} \leq \|x^1 - x^*\|^2 \leq R^2
\]
Polyak’s stepsize

$$\|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - \frac{(f^t - f^*)^2}{\|g_t\|^2}$$

- **Observation 1** $\|x^t - x^*\|$ decreases
- **Recursion:**

$$\sum_{t=1}^{k} \frac{(f^t - f^*)^2}{\|g_t\|^2} \leq \|x^1 - x^*\|^2 \leq R^2$$

- **Now use** $\|g^t\| \leq G$

$$\sum_{t=1}^{k} (f^t - f^*)^2 \leq R^2G^2$$
Polyak’s stepsize

\[ \|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - \frac{(f^t - f^*)^2}{\|g_t\|^2} \]

- **Observation 1**  \[\|x^t - x^*\|\] decreases
- **Recursion:**
  \[
  \sum_{t=1}^{k} \frac{(f^t - f^*)^2}{\|g_t\|^2} \leq \|x^1 - x^*\|^2 \leq R^2
  \]

- **Now use** \[\|g^t\| \leq G\]
  \[
  \sum_{t=1}^{k} (f^t - f^*)^2 \leq R^2 G^2
  \]

- **Observation 2**  \[f^t \rightarrow f^*\]
Polyak’s stepsize

\[ \|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - \frac{(f^t - f^*)^2}{\|g_t\|^2} \]

- **Observation 1** \( \|x^t - x^*\| \) decreases
- **Recursion:**

\[
\sum_{t=1}^{k} \frac{(f^t - f^*)^2}{\|g_t\|^2} \leq \|x^1 - x^*\|^2 \leq R^2
\]

- **Now use** \( \|g^t\| \leq G \)

\[
\sum_{t=1}^{k} (f^t - f^*)^2 \leq R^2 G^2
\]

- **Observation 2** \( f^t \to f^* \)
- **for accuracy** \( \varepsilon \), need \( k = (RG/\varepsilon)^2 \)
Constrained optimization

\[ \min f(x) \quad \text{s.t.} \quad x \in C \]
Constrained optimization

\[
\min f(x) \quad \text{s.t.} \quad x \in \mathcal{C}
\]

- Previously:
  \[
  x^{t+1} = x^t - \alpha_t g^t
  \]

- This could be infeasible!
  Solution: projection
Projected subgradient method

\[ x^{k+1} = P_C(x^k - \alpha_k g^k) \]

where \( g^k \in \partial f(x^k) \) is any subgradient

Projection closest feasible point

\[ P_C(x) = \arg \min_{y \in C} \|x - y\|_2 \]
(Assume \( C \) is closed and convex, then projection is unique)

Great as long as projection is “easy”

Same questions as before: Does it converge? For which stepsizes? How fast?
Projected subgradient method

\[ x^{k+1} = P_C(x^k - \alpha_k g^k) \]

where \( g^k \in \partial f(x^k) \) is any subgradient

- **Projection** closest feasible point
  \[ P_C(x) = \arg \min_{y \in C} \|x - y\|^2 \]

  (Assume \( C \) is closed and convex, then projection is unique)
Projected subgradient method

\[ x^{k+1} = P_C(x^k - \alpha_k g^k) \]
where \( g^k \in \partial f(x^k) \) is any subgradient

- **Projection** closest feasible point

\[ P_C(x) = \arg \min_{y \in C} \|x - y\|^2 \]

(Assume \( C \) is closed and convex, then projection is unique)

- Great as long as projection is “easy”
- Same questions as before:
  - Does it converge?
  - For which stepsizes?
  - How fast?
Convergence

Assumptions

► Min is attained: \( f^* := \inf_x f(x) > -\infty \), with \( f(x^*) = f^* \)
► Bounded subgradients: \( \|g\|_2 \leq G \) for all \( g \in \partial f \)
► Bounded domain: \( \|x^0 - x^*\|_2 \leq R \)
Convergence

Assumptions

- Min is attained: \( f^* := \inf_x f(x) > -\infty \), with \( f(x^*) = f^* \)
- Bounded subgradients: \( \|g\|_2 \leq G \) for all \( g \in \partial f \)
- Bounded domain: \( \|x^0 - x^*\|_2 \leq R \)

Analysis

- Let \( z^{t+1} = x^t - \alpha_t g^t \).
- Then \( x^{t+1} = P_C(z^{t+1}) \).
Convergence

Assumptions

- Min is attained: $f^* := \inf_x f(x) > -\infty$, with $f(x^*) = f^*$
- Bounded subgradients: $\|g\|_2 \leq G$ for all $g \in \partial f$
- Bounded domain: $\|x^0 - x^*\|_2 \leq R$

Analysis

- Let $z^{t+1} = x^t - \alpha_t g^t$.
- Then $x^{t+1} = P_C(z^{t+1})$.
- Recall analysis of unconstrained method:

\[
\|z^{t+1} - x^*\|_2^2 = \|x^t - \alpha_t g^t - x^*\|_2^2 \\
\leq \|x^t - x^*\|_2^2 + \alpha_t^2 \|g^t\|_2^2 - 2\alpha_t (f(x^t) - f^*) \\
\ldots
\]

- Need to relate to $\|x^{t+1} - x^*\|_2^2$, the rest of the proof is the same as above.
Projection Theorem

Let $\mathcal{C}$ be nonempty, closed and convex.

- Optimality conditions: $y^* = P_C(z)$ iff
  \[ \langle z - y^*, y - y^* \rangle \leq 0 \text{ for all } y \in \mathcal{C} \]

- The projection is nonexpansive:
  \[ \|P_C(x) - P_C(z)\| \leq \|x - z\|^2 \text{ for all } x, z \in \mathbb{R}^n. \]
Use nonexpansiveness of projection:

\[
\|x_t^* - \alpha_t g^t - x^*\|_2^2 \\
\leq \|x_t - x^*\|_2^2 + \alpha_t^2 \|g^t\|_2^2 - 2\alpha_t (f(x_t) - f^*)
\]

\ldots
Convergence

Use nonexpansiveness of projection:

\[ \|x^{t+1} - x^*\|^2_2 = \|P_C(x^t - \alpha_t g^t) - x^*\|^2_2 \]
\[ \leq \|x^t - \alpha_t g^t - x^*\|^2_2 \]
\[ \leq \|x^t - x^*\|^2_2 + \alpha_t^2 \|g^t\|^2_2 - 2\alpha_t (f(x^t) - f^*) \]

\[ \ldots \]

Same convergence results as in unconstrained case:

- within neighborhood of optimal for constant step size
- converges for diminishing non-summable
Examples

\[
\begin{align*}
\min & \quad \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1 \\
\text{s.t.} & \quad x \in C
\end{align*}
\]

- **Nonnegativity** \( x \geq 0 \)

\[
P_C(z) = [z]_+
\]

Update step: \( x^{k+1} = [x^k - \alpha_k (A^T(Ax^k - b) + \lambda \text{sgn}(x^k))]_+ \)
Examples

\[
\min \quad \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1 \\
\text{s.t. } x \in C
\]

► **Nonnegativity** \( x \geq 0 \)

\[
P_C(z) = [z]_+ \\
\text{Update step: } x^{k+1} = [x^k - \alpha_k(A^T(Ax^k - b) + \lambda \text{sgn}(x^k))]_+
\]

► **\( \ell_\infty \)-ball** \( \|x\|_\infty \leq 1 \)

\[
\text{Projection: } \min \|x - z\|^2 \text{ s.t. } x \leq 1 \text{ and } x \geq -1
\]
Examples

\[
\begin{align*}
\text{min} & \quad \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1 \\
\text{s.t.} & \quad x \in C
\end{align*}
\]

- **Nonnegativity** \( x \geq 0 \)

\[
P_C(z) = [z]_+ \\
\text{Update step: } x^{k+1} = [x^k - \alpha_k (A^T (Ax^k - b) + \lambda \text{sgn}(x^k))]_+
\]

- **\( \ell_\infty \)-ball** \( \|x\|_\infty \leq 1 \)

\[
\text{Projection: } \min \|x - z\|^2 \text{ s.t. } x \leq 1 \text{ and } x \geq -1 \\
\text{this is separable, so do it coordinate-wise:} \\
P_C(z) = y \text{ where } y_i = \text{sgn}(z_i) \min\{|z_i|, 1\}
\]
Examples

\[
\min \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1
\]

s.t. \(x \in C\)

- **Nonnegativity** \(x \geq 0\)
  \[
P_C(z) = [z]_+
\]
  Update step: \(x^{k+1} = [x^k - \alpha_k (A^T(Ax^k - b) + \lambda \text{sgn}(x^k))]_+\)

- **\(\ell_\infty\)-ball** \(\|x\|_\infty \leq 1\)
  Projection: \(\min \|x - z\|_2\) s.t. \(x \leq 1\) and \(x \geq -1\)
  this is separable, so do it coordinate-wise:
  \[
P_C(z) = y \text{ where } y_i = \text{sgn}(z_i) \min\{|z_i|, 1\}
\]
  Update step:
  \[
z^{k+1} = x^k - \alpha_k (A^T(Ax^k - b) + \lambda \text{sgn}(x^k))
  \]
  \[
x^{k+1}_i = \text{sgn}(z^{k+1}_i) \min\{|z^{k+1}_i|, 1\}
\]
Examples

1. Linear equality constraints $Ax = b$ ($A \in \mathbb{R}^{n \times m}$ has rank $n$)

$$P_C(x) = z - A^\top (AA^\top)^{-1} (Az - b)$$

$$= (I - A^\top (A^\top A)^{-1} A)z + A^\top (AA^\top)^{-1} b$$
Examples

- **Linear equality constraints** $Ax = b$ ($A \in \mathbb{R}^{n \times m}$ has rank $n$)

$$P_C(x) = z - A^\top (AA^\top)^{-1} (Az - b)$$

$$= (I - A^\top (A^\top A)^{-1} A)z + A^\top (AA^\top)^{-1} b$$

Update step, using $Ax^t = b$:

$$x^{t+1} = P_C(x^t - \alpha_t g^t)$$

$$= x^t - \alpha_t (I - A^\top (AA^\top)^{-1} A)g^t$$
Examples

- **Linear equality constraints** \(Ax = b\) (\(A \in \mathbb{R}^{n \times m}\) has rank \(n\))

  \[
P_C(x) = z - A^\top (A A^\top)^{-1} (Az - b) \\
  = (I - A^\top (A^\top A)^{-1} A)z + A^\top (A A^\top)^{-1} b
  \]

  Update step, using \(Ax^t = b\):

  \[
x^{t+1} = P_C(x^t - \alpha_t g^t) \\
  = x^t - \alpha_t (I - A^\top (A A^\top)^{-1} A) g^t
  \]

- **Simplex** \(x^\top 1 = 1\) and \(x \geq 0\)

  more complex but doable, similarly \(\ell_1\)-norm ball
Some remarks

► Why care?
  - simple
  - low-memory
  - stochastic version possible

Mirror Descent

► Improvements using more information (heavy-ball, filtered subgradient, . . . )

► Don’t forget the dual! may be more amenable to optimization
duality gap
Some remarks

- Why care?
  - simple
  - low-memory
  - stochastic version possible

- Another perspective

\[ x^{k+1} = \min_{x \in C} \langle x, g^k \rangle + \frac{1}{2\alpha_k} \| x - x_k \|^2 \]

Mirror Descent
Some remarks

► Why care?
  ■ simple
  ■ low-memory
  ■ stochastic version possible

► Another perspective

\[ x^{k+1} = \min_{x \in C} \langle x, g^k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2 \]

Mirror Descent

► Improvements using more information (heavy-ball, filtered subgradient, …)
Some remarks

► Why care?
  ■ simple
  ■ low-memory
  ■ stochastic version possible

► Another perspective

\[ x^{k+1} = \min_{x \in C} \langle x, g^k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2 \]

Mirror Descent

► Improvements using more information (heavy-ball, filtered subgradient, …)

► Don’t forget the dual!
  ■ may be more amenable to optimization
  ■ duality gap
What we did not cover

♠ Adaptive stepsize tricks
♠ Space dilation methods, quasi-Newton style subgrads
♠ Barrier subgradient method
♠ Sparse subgradient method
♠ Ellipsoid method, center of gravity, etc. as subgradient methods
♠ And many more