L4. SUBGRADIENT METHOD
Unconstrained Minimization

\[ \min_{x \in \mathbb{R}^n} f(x) \]

Setting:
- \( f \) is (continuously) differentiable \( \rightarrow \) gradient descent
- \( f \) is not differentiable \( \rightarrow \) subgradient “method”
- \( f \) is convex
Subgradient

For a function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), a subgradient at \( x \) is any vector \( g \in \mathbb{R}^n \) s.t.

\[
f(y) \geq f(x) + g^T (y - x), \quad \forall y \in \mathbb{R}^n
\]

Fact:

- Subgradient always exists when \( f \) is convex
- When \( f \) is nonconvex, subgradient may not exist
- When \( f \) is convex and differentiable, \( g = \nabla f(x) \)
Examples

1-norm in 1-D \( f(x) = |x| \)

Hinge-Loss \( f(x) = \max(1 - x, 0) \)
Gradient vs. Subgradient

Neg. gradient of differentiable $f$ is a descent direction
Neg. subgradient is NOT always descent

Ex. $f(x_1, x_2) = |x_1| + 2|x_2|$
Subdifferential

\( \partial f(x) \): the set of all subgradients of \( f \) at \( x \)

Fact (when \( f \) is convex)

- Nonempty set
- If \( f \) is differentiable at \( x \), then \( \partial f(x) = \{ \nabla f(x) \} \)
- If \( \partial f(x) = \{ g \} \), then \( f \) is differentiable and \( g = \nabla f(x) \)
- Closed and convex set
  - Hint: an intersection of half-spaces
Ex.

1-norm in n-dim: \( f(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i| \)

\[ \partial f(x) = \begin{bmatrix} g \in \mathbb{R}^n : g_i = \begin{cases} 1 & \text{if } x_i > 0 \\ -1 & \text{if } x_i < 0 \\ [-1, 1] & \text{if } x_i = 0 \end{cases} \end{bmatrix} \]
Subgradient Method

\[ \min_{x \in \mathbb{R}^n} f(x) \]

Changes to GD:

1. Update using subgradients

\[ x_{k+1} = x_k - \alpha_k g_k \]

\( g_k \) : any subgradient of \( f(x) \) at \( x_k \), \( g_k \in \partial f(x_k) \)
Subgradient Method

2. Keeping x_best:

- Subgradient is not necessarily a descent direction
- That is, no guarantee of monotonic convergence, i.e.

\[ f(x_{k+1}) \leq f(x_k) \]

- Therefore we need to keep \( X_{\text{best}} \) with the smallest function value over iterates
3. Stepsize

• Linesearch is not possible since search direction is not always descent

• Constant steps: \( \alpha_k = c \)

• Diminishing steps: \( \alpha_k \to 0 \)
Subgradient Method

Assumptions:

• Objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex
• $f$ is proper: $f(x) > -\infty \ \forall x$, $f(x) < \infty$ for some $x$
• Subgradient norm is uniformly bounded, i.e.,
  \[ \|g\|_2 \leq M, \quad g \in \partial f(x), \ \forall x \in \mathbb{R}^n \]
• Starting point is reasonable,
  \[ \|x_0 - x^*\|_2 \leq D \]
Convergence (1): constant stepsize

\[ f(x_{k,\text{best}}) - f(x^*) \leq \frac{D^2 + M^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i} \]

- Constant stepsize \( \alpha_i = C \):
  - If \( k \) is fixed, then optimal constant stepsize?

\[ c = \frac{D}{M \sqrt{k}} \]

\[ f(x_{k,\text{best}}) - f(x^*) \leq \frac{DM}{2\sqrt{k}} \]

Sublinear convergence \( \mathcal{O}(1/\sqrt{k}) \)

(GD: sublinear but with \( \mathcal{O}(1/k) \))
Convergence (2): diminishing, $1/\sqrt{k}$

\[ f(x_{k,\text{best}}) - f(x^*) \leq \frac{D^2 + M^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i} \]

- Diminishing stepsize (not square summable)
  - E.g.
  \[ \alpha_i = \frac{D}{M \sqrt{i}} \]

\[ f(x_{k,\text{best}}) - f(x^*) \leq O \left( \frac{DM \ln k}{2 \sqrt{k}} \right) \]
Convergence (3): diminishing, $1/k$

$$f(x_k, \text{best}) - f(x^*) \leq \frac{D^2 + M^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}$$

- Square summable but non-summable diminishing stepsizes, i.e.,
  $$\sum_{i=1}^{\infty} \alpha_k^2 < \infty, \sum_{i=1}^{\infty} \alpha_k = \infty, \alpha_k \to 0$$

  - E.g. $$\alpha_k = \frac{1}{k}, \sum_{i=1}^{\infty} \frac{1}{k^2} = \frac{\pi}{6} \text{ (Euler)}, \sum_{i=1}^{k} \frac{1}{i} < 1 + \ln k$$

$$f(x_k, \text{best}) - f(x^*) \to 0$$

Can be a better choice for strongly convex case (we’ll see more in SGD)
Subgradient Method: Summary

\[ \min_{x \in \mathbb{R}^n} f(x) \]

For \( k = 1, 2, \ldots \)

- Choose a subgradient \( g_k \in \partial f(x_k) \)
- Choose a stepsize \( \alpha_k \)
- Take a step \( x_{k+1} = x_k - \alpha_k g_k \)

Negative subgradient may not be a descent direction

Cannot do linesearch

No guarantee that \( f(x_{k+1}) \leq f(x_k) \)
Convergence: Summary

For difference choices of stepsizes:

- Constant (optimal, fixed k)
  \[ \alpha_i = \frac{D}{M \sqrt{k}} \]
  \[ f(x_{k, \text{best}}) - f(x^*) \leq \frac{DM}{2 \sqrt{k}} \]

- Diminishing, not square summable
  \[ \alpha_k = \frac{D}{M \sqrt{k}} \]
  \[ f(x_{k, \text{best}}) - f(x^*) \leq O \left( \frac{DM \ln k}{2 \sqrt{k}} \right) \]

- Diminishing, square summable
  \[ \alpha_k = O \left( \frac{1}{k} \right) \]
  \[ f(x_{k, \text{best}}) - f(x^*) \rightarrow 0 \]
Ex. LASSO (\(\ell_1\)-Penalized Regression)

\[
X \in \mathbb{R}^{m \times n}, \ y \in \mathbb{R}^m, \ \lambda > 0 \\
\min_{w \in \mathbb{R}^n} f(w) = \frac{1}{2m} \|y - Xw\|_2^2 + \lambda \|w\|_1
\]

Check:

- \(f\) convex?
- \(f\) differentiable (smooth)?
- \(f\) strongly convex?
LASSO

\[
\min_{w \in \mathbb{R}^n} f(w) = \frac{1}{2m} \|y - Xw\|_2^2 + \lambda \|w\|_1
\]

Subgradient? \[ g(w) = -X^T(y - Xw)/m + \lambda h(w) \]

\[ [h(w)]_i = \begin{cases} 
+1 & \text{if } w_i > 0 \\
-1 & \text{if } w_i < 0 \\
[-1, 1] & \text{if } w_i = 0 
\end{cases} \]

• An easy choice: \[ [h(w)]_i = \text{sign}(w_i) \]
Non-Monotonic Progress

TU Dortmund, Dr. Sangkyun Lee
Convergence: $f_{\text{best}} (m > n)$
Convergence: $f_{\text{best}}$ ($m < n$)