Chapter 5

Subgradient Method
5.1. Subgradient method

to minimize a non-differentiable convex function $f$:

$$\min_x f(x)$$

subgradient method: choose $x^0$ and repeat

$$x^k = x^{k-1} - t_k g^{k-1}, \quad k = 1, 2, \ldots$$

$g^{k-1}$ is any subgradient of $f$ at $x^{k-1}$

step size rules:

- fixed step: $t_k$ constant
- fixed length: $t_k \|g^{k-1}\|_2$ constant (i.e., $\|x^k - x^{k-1}\|_2$ constant)
- diminishing: $t_k \to 0$, $\sum_{k=1}^{\infty} t_k = \infty$
Directional derivative

the subgradient method is not a descent method!

**definition** (general $f$): directional derivative of $f$ at $x$ in the direction of $y$ is

$$f'(x; y) = \lim_{\alpha \to 0} \frac{f(x + \alpha y) - f(x)}{\alpha}$$

(if the limit exists)

- $f'(x; y)$ is the right derivative of $g(\alpha) = f(x + \alpha y)$ at $\alpha = 0$
Descent directions and subgradients

$y$ is a **descent direction** of $f$ at $x$ if $f'(x; y) < 0$

- negative gradient of differentiable $f$ is descent direction (if $\nabla f(x) \neq 0$)
- negative subgradient is **not** always a descent direction

**example:** $f(x_1, x_2) = |x_1| + 2|x_2|$ 

$g = (1, 2) \in \partial f(1, 0)$, but $-g = (-1, -2)$ is not a descent direction at $(1, 0)$
Assumptions

- $f$ has finite optimal value $f^*$, minimizer $x^*$
- $f$ is convex, dom $f = \mathbb{R}^n$
- $f$ is Lipschitz continuous with constant $G > 0$:
  \[ |f(x) - f(y)| \leq G\|x - y\|_2, \quad \forall x, y \]
  this is equivalent to
  \[ \|g\|_2 \leq G, \quad \forall g \in \partial f(x), \forall x \]
  (see next page for proof)
proof:

- assume $\|g\|_2 \leq G$ for all subgradients; choose $g_x \in \partial f(x), g_y \in \partial f(y)$:

$$g_x^\top (x - y) \geq f(x) - f(y) \geq g_y^\top (x - y)$$

by the Cauchy-Schwarz inequality

$$G\|x - y\|_2 \geq f(x) - f(y) \geq -G\|x - y\|_2$$

- assume $\|g\|_2 > G$ for some $g \in \partial f(x)$; take $y = x + g/\|g\|_2$:

$$f(y) \geq f(x) + g^\top (y - x)$$

$$= f(x) + \|g\|_2$$

$$> f(x) + G$$
Analysis of subgradient method

- the subgradient method is not a descent method
- the key quantity in the analysis is the distance to the optimal set

with \( x^+ = x^i \), \( x = x^{i-1} \), \( g = g^{i-1} \), \( t = t_i \):

\[
\|x^+ - x^*\|_2^2 = \|x - tg - x^*\|_2^2 \\
= \|x - x^*\|_2^2 - 2tg^\top(x - x^*) + t^2\|g\|_2^2 \\
\leq \|x - x^*\|_2^2 - 2t(f(x) - f(x^*)) + t^2\|g\|_2^2
\]

combine inequalities for \( i = 1, \ldots, k \), and define \( f_{best}^k = \min_{0 \leq i < k} f(x^i) \):

\[
2(\sum_{i=1}^k t_i)(f_{best}^k - f^*) \leq \|x^0 - x^*\|_2^2 - \|x^k - x^*\|_2^2 + \sum_{i=1}^k t_i^2\|g^{i-1}\|_2^2 \\
\leq \|x^0 - x^*\|_2^2 + \sum_{i=1}^k t_i^2\|g^{i-1}\|_2^2
\]
fixed step size $t_i = t$

$$f_{\text{best}}^k - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2kt} + \frac{G^2t}{2}$$

- does not guarantee convergence of $f_{\text{best}}^k$
- for large $k$, $f_{\text{best}}^k$ is approximately $G^2t/2$-suboptimal

fixed step length $t_i = s/\|g^{-1}\|_2$

$$f_{\text{best}}^k - f^* \leq \frac{G\|x^0 - x^*\|_2^2}{2ks} + \frac{Gs}{2}$$

- does not guarantee convergence of $f_{\text{best}}^k$
- for large $k$, $f_{\text{best}}^k$ is approximately $Gs/2$-suboptimal
**diminishing step size** $t_i \to 0$, $\sum_{i=1}^{\infty} t_i = \infty$

$$f_{best}^k - f^* \leq \frac{\|x^0 - x^*\|_2^2 + G^2 \sum_{i=1}^{k} t_i^2}{2 \sum_{i=1}^{k} t_i}$$

can show that $(\sum_{i=1}^{k} t_i^2) / (\sum_{i=1}^{k} t_i) \to 0$; hence, $f_{best}^k$ converges to $f^*$
5.1.1. Example: 1-norm minimization

\[
\min \| Ax - b \|_1
\]

subgradient is given by \( A^\top \text{sign}(Ax - b) \)

fixed step size \( t_k = s/\|g^{k-1}\|_2, \ s = 0.1, 0.01, 0.001 \)
diminishing step size $t_k = 0.01/\sqrt{k}$, $t_k = 0.01/k$
Optimal step size for fixed number of iterations

if \( s_i = t_i \| g^{i-1} \|_2 \) and \( \| x^0 - x^* \| \leq R \):

\[
\frac{f_{best}^k - f^*}{2} \leq \frac{R^2 + \sum_{i=1}^{k} s_i^2}{2 \sum_{i=1}^{k} s_i/G}
\]

• for given \( k \), bound is minimized by fixed step length \( s_i = s = R/\sqrt{k} \)

• resulting bound after \( k \) steps is

\[
\frac{f_{best}^k - f^*}{\sqrt{k}} \leq \frac{GR}{\sqrt{k}}
\]

• guarantees accuracy \( f_{best}^k - f^* \leq \epsilon \) in \( k = O(1/\epsilon^2) \) iterations
Optimal step size when $f^*$ is known

rhs in first inequality of page 7 is minimized by

$$t_i = \frac{f(x^{i-1}) - f^*}{\|g^{i-1}\|^2_2}$$

optimized bound is

$$\frac{(f(x^{i-1}) - f^*)^2}{\|g^{i-1}\|^2_2} \leq \|x^{i-1} - x^*\|^2_2 - \|x^i - x^*\|^2_2$$

applying recursively (with $\|x^0 - x^*\|^2_2 \leq R$ and $\|g^i\|^2 \leq G$) gives

$$f_{best}^k - f^* \leq \frac{GR}{\sqrt{k}}$$
Exercise: find point in intersection of convex sets

to find a point in the intersection of \( m \) closed convex sets \( C_1, \ldots, C_m, \)

\[
\min f(x) = \max\{d_1(x), \ldots, d_m(x)\}
\]

where \( d_j(x) = \inf_{y \in C_j} \|x - y\|_2 \) is Euclidean distance of \( x \) to \( C_j \)

- \( f^* = 0 \) if the intersection is nonempty
- \( g \in \partial f(\hat{x}) \) if \( g \in \partial d_j(\hat{x}) \) and \( C_j \) is farthest set from \( \hat{x} \)
- subgradient \( g \in \partial d_j(\hat{x}) \) from projection \( P_j(\hat{x}) \) on \( C_j \):

\[
g = 0 \quad (\text{if } \hat{x} \in C_j), \quad g = \frac{1}{d(\hat{x}, C_j)}(\hat{x} - P_j(\hat{x})) \quad (\text{if } \hat{x} \notin C_j)
\]

note that \( \|g\|_2 = 1 \) if \( \hat{x} \notin C_j \)
subgradient method with optimal step size

• optimal step size for \( f^* = 0 \) and \( \|g_{i-1}\|_2 = 1 \) is \( t_i = f(x_{i-1}) \)

• at iteration \( k \), find farthest set \( C_j \) (with \( f(x_{k-1}) = d_j(x_{k-1}) \)); take

\[
x^k = x^{k-1} - \frac{f(x_{k-1})}{d_j(x_{k-1})}(x^{k-1} - P_j(x_{k-1})) \\
= P_j(x_{k-1})
\]

• a version of the alternating projections algorithm

• at each step, project the current point onto the farthest set

• for \( m = 2 \), projections alternate onto one set, then the other
Optimality of the subgradient method

can the bound $f_{\text{best}}^k - f^* \leq GR/\sqrt{k}$ be improved?

problem class:

- $f$ is convex, with a minimizer $x^*$
- we know a starting point $x^0$ with $\|x^0 - x^*\|_2 \leq R$
- we know the Lipschitz constant $G$ of $f$ on $\{x \mid \|x - x^0\|_2 \leq R\}$
- $f$ is defined by an oracle: given $x$, oracle returns $f(x)$ and subgradient

algorithm class: $k$ iterations of any method that chooses $x^i$ in

$$x^0 + \text{span}\{g^0, g^1, \ldots, g^{i-1}\}$$
test problem and oracle

\[ f(x) = \max_{i=1,\ldots,k} x_i + \frac{1}{2} \|x\|_2^2, \quad x^0 = 0 \]

- solution: \( x^* = -\frac{1}{k} (1, \ldots, 1, 0, \ldots, 0) \) and \( f^* = -\frac{1}{2k} \)

- \( R = \|x^0 - x^*\|_2 = 1/\sqrt{k} \) and \( G = 1 + 1/\sqrt{k} \)

- oracle returns subgradient \( e_{\hat{j}} + x \) where \( \hat{j} = \min\{j \mid x_j = \max_{i=1,\ldots,k} x_i\} \)

iteration: for \( i = 0, \ldots, k - 1 \), entries \( x_{i+1}^i, \ldots, x_k^i \) are zero

\[ f_{\text{best}}^k - f^* = \min_{i<k} f(x^i) - f^* \geq -f^* = \frac{GR}{2(1 + \sqrt{k})} \]

conclusion: \( O(1/\sqrt{k}) \) bound cannot be improved
5.1.2. Summary: subgradient method

- handles general nondifferentiable convex problem
- often leads to very simple algorithms
- convergence can be very slow
- no good stopping criterion
- theoretical complexity: $O(1/\epsilon^2)$ iterations to find $\epsilon$-suboptimal point
- an “optimal” 1st-order method: $O(1/\epsilon^2)$ bound cannot be improved