Subgradient Methods

• subgradient method and stepsize rules
• convergence results and proof
• optimal step size and alternating projections
• speeding up subgradient methods
Subgradient method

**Subgradient method** is simple algorithm to minimize nondifferentiable convex function $f$

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- $x^{(k)}$ is the $k$th iterate
- $g^{(k)}$ is any subgradient of $f$ at $x^{(k)}$
- $\alpha_k > 0$ is the $k$th step size

Not a descent method, so we keep track of best point so far

$$f^{(k)}_{\text{best}} = \min_{i=1,\ldots,k} f(x^{(i)})$$
Step size rules

Step sizes are fixed ahead of time

*constant step size*: \( \alpha_k = \alpha \) (constant)

*constant step length*: \( \alpha_k = \gamma / \| g^{(k)} \|_2 \) (so \( \| x^{(k+1)} - x^{(k)} \|_2 = \gamma \))

*square summable but not summable*: step sizes satisfy

\[
\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k = \infty
\]

*nonsummable diminishing*: step sizes satisfy

\[
\lim_{k \to \infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty
\]
Assumptions

• $f^* = \inf_x f(x) > -\infty$, with $f(x^*) = f^*$

• $\|g\|_2 \leq G$ for all $g \in \partial f$ (equivalent to Lipschitz condition on $f$)

• $\|x^{(1)} - x^*\|_2 \leq R$

these assumptions are stronger than needed, just to simplify proofs
Convergence results

define $\bar{f} = \lim_{k \to \infty} f_{\text{best}}^{(k)}$

- **constant step size:** $\bar{f} - f^* \leq G^2 \alpha / 2$, i.e., **converges to** $G^2 \alpha / 2$-suboptimal
  (converges to $f^*$ if $f$ differentiable, $\alpha$ small enough)

- **constant step length:** $\bar{f} - f^* \leq G \gamma / 2$, i.e., **converges to** $G \gamma / 2$-suboptimal

- **diminishing step size rule:** $\bar{f} = f^*$, i.e., **converges**
Convergence proof

key quantity: Euclidean distance to the optimal set, not the function value

let $x^\star$ be any minimizer of $f$

\[
\|x^{(k+1)} - x^\star\|_2^2 = \|x^{(k)} - \alpha_k g^{(k)} - x^\star\|_2^2 \\
= \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k g^{(k)}T(x^{(k)} - x^\star) + \alpha_k^2 \|g^{(k)}\|_2^2 \\
\leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^\star) + \alpha_k^2 \|g^{(k)}\|_2^2
\]

using $f^\star = f(x^\star) \geq f(x^{(k)}) + g^{(k)}T(x^\star - x^{(k)})$
apply recursively to get

\[
\| x^{(k+1)} - x^* \|_2^2 \leq \| x^{(1)} - x^* \|_2^2 - 2 \sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^*) + \sum_{i=1}^{k} \alpha_i^2 \| g^{(i)} \|_2^2
\]

\[
\leq R^2 - 2 \sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^*) + G^2 \sum_{i=1}^{k} \alpha_i^2
\]

now we use

\[
\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^*) \geq (f_{\text{best}}^{(k)} - f^*) \left( \sum_{i=1}^{k} \alpha_i \right)
\]

to get

\[
f_{\text{best}}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}.
\]
**constant step size:** for $\alpha_k = \alpha$ we get

$$f_{\text{best}}^{(k)} - f^* \leq \frac{R^2 + G^2 k \alpha^2}{2k\alpha}$$

righthand side converges to $G^2 \alpha / 2$ as $k \to \infty$

**constant step length:** for $\alpha_k = \gamma / \| g^{(k)} \|_2$ we get

$$f_{\text{best}}^{(k)} - f^* \leq \frac{R^2 + \sum_{i=1}^{k} \alpha_i^2 \| g^{(i)} \|_2^2}{2 \sum_{i=1}^{k} \alpha_i} \leq \frac{R^2 + \gamma^2 k}{2\gamma k/G},$$

righthand side converges to $G\gamma / 2$ as $k \to \infty$
square summable but not summable step sizes:
suppose step sizes satisfy

\[
\sum_{i=1}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k = \infty
\]

then

\[
f^{(k)}_{\text{best}} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}
\]

as \( k \to \infty \), numerator converges to a finite number, denominator converges to \( \infty \), so \( f^{(k)}_{\text{best}} \to f^* \)
Stopping criterion

• terminating when \( \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i} \leq \epsilon \) is really, really, slow

• optimal choice of \( \alpha_i \) to achieve \( \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i} \leq \epsilon \) for smallest \( k \):

\[
\alpha_i = \left( \frac{R}{G} \right) / \sqrt{k}, \quad i = 1, \ldots, k
\]

number of steps required: \( k = \left( \frac{RG}{\epsilon} \right)^2 \)

• the truth: there really isn’t a good stopping criterion for the subgradient method . . .
Example: Piecewise linear minimization

\[
\begin{align*}
\text{minimize} \quad f(x) &= \max_{i=1,\ldots,m} (a_i^T x + b_i) \\
\text{to find a subgradient of } f: \text{ find index } j \text{ for which} \\
&\quad a_j^T x + b_j = \max_{i=1,\ldots,m} (a_i^T x + b_i) \\
\text{and take } g = a_j
\end{align*}
\]

\[
\begin{align*}
\text{subgradient method: } x^{(k+1)} &= x^{(k)} - \alpha_k a_j
\end{align*}
\]
problem instance with $n = 20$ variables, $m = 100$ terms, $f^* \approx 1.1$

$f^{(k)}_{\text{best}} - f^*$, constant step length $\gamma = 0.05, 0.01, 0.005$

![Graph showing the decrease of $f^{(k)}_{\text{best}} - f^*$ with different step lengths.](image)
diminishing step rules $\alpha_k = 0.1/\sqrt{k}$ and $\alpha_k = 1/\sqrt{k}$, square summable step size rules $\alpha_k = 1/k$ and $\alpha_k = 10/k$
Optimal step size when $f^*$ is known

- choice due to Polyak:
  \[
  \alpha_k = \frac{f(x^{(k)}) - f^*}{\|g^{(k)}\|_2^2}
  \]
  (can also use when optimal value is estimated)

- motivation: start with basic inequality
  \[\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) + \alpha_k^2 \|g^{(k)}\|_2^2\]
  and choose $\alpha_k$ to minimize righthand side
• yields
\[
\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(k)} - x^*\|_2^2 - \frac{(f(x^{(k)}) - f^*)^2}{\|g^{(k)}\|_2^2}
\]
(in particular, \(\|x^{(k)} - x^*\|_2\) decreases each step)

• applying recursively,
\[
\sum_{i=1}^{k} \frac{(f(x^{(i)}) - f^*)^2}{\|g^{(i)}\|_2^2} \leq R^2
\]
and so
\[
\sum_{i=1}^{k} (f(x^{(i)}) - f^*)^2 \leq R^2 G^2
\]
which proves \(f(x^{(k)}) \rightarrow f^*\)
PWL example with Polyak’s step size, $\alpha_k = 0.1/\sqrt{k}$, $\alpha_k = 1/k$
Finding a point in the intersection of convex sets

$C = C_1 \cap \cdots \cap C_m$ is nonempty, $C_1, \ldots, C_m \subseteq \mathbb{R}^n$ closed and convex

find a point in $C$ by minimizing

$$f(x) = \max\{\text{dist}(x, C_1), \ldots, \text{dist}(x, C_m)\}$$

with $\text{dist}(x, C_j) = f(x)$, a subgradient of $f$ is

$$g = \nabla \text{dist}(x, C_j) = \frac{x - P_{C_j}(x)}{\|x - P_{C_j}(x)\|_2}$$
subgradient update with optimal step size:

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

\[ = x^{(k)} - f(x^{(k)}) \frac{x - P_{C_j}(x)}{\|x - P_{C_j}(x)\|_2} \]

\[ = P_{C_j}(x^{(k)}) \]

- a version of the famous *alternating projections* algorithm

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

- for \( m = 2 \) sets, projections alternate onto one set, then the other

- convergence: \( \text{dist}(x^{(k)}, C) \to 0 \) as \( k \to \infty \)
Alternating projections

first few iterations:

\[ x^{(1)} \]
\[ x^{(2)} \]
\[ x^{(3)} \]
\[ x^{(4)} \]
\[ \cdots \]

eventually converges to a point \( x^* \in C_1 \cap C_2 \)

\[ C_1 \]
\[ C_2 \]
Example: Positive semidefinite matrix completion

- some entries of matrix in $S^n$ fixed; find values for others so completed matrix is PSD

- $C_1 = S^n_+, C_2$ is (affine) set in $S^n$ with specified fixed entries

- projection onto $C_1$ by eigenvalue decomposition, truncation: for $X = \sum_{i=1}^{n} \lambda_i q_i q_i^T$,

$$P_{C_1}(X) = \sum_{i=1}^{n} \max\{0, \lambda_i\} q_i q_i^T$$

- projection of $X$ onto $C_2$ by re-setting specified entries to fixed values
specific example: $50 \times 50$ matrix missing about half of its entries

- initialize $X^{(1)}$ with unknown entries set to 0
convergence is linear:

\[ \| X^{(k+1)} - X^{(k)} \|_F \]
Polyak step size when $f^*$ isn’t known

- use step size

$$\alpha_k = \frac{f(x^{(k)}) - f^{(k)}_{\text{best}} + \gamma_k}{\|g^{(k)}\|_2}$$

with $\sum_{k=1}^{\infty} \gamma_k = \infty$, $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$

- $f^{(k)}_{\text{best}} - \gamma_k$ serves as estimate of $f^*$

- $\gamma_k$ is in scale of objective value

- can show $f^{(k)}_{\text{best}} \to f^*$

Prof. J. Duchi, EE364b, Stanford University
PWL example with Polyak’s step size, using $f^*$, and estimated with $\gamma_k = 10/(10 + k)$
Speeding up subgradient methods

• subgradient methods are very slow

• often convergence can be improved by keeping memory of past steps

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

(heavy ball method)

other ideas: localization methods, conjugate directions, . . .
A couple of speedup algorithms

\[ x^{(k+1)} = x^{(k)} - \alpha_k s^{(k)}, \quad \alpha_k = \frac{f(x^{(k)}) - f^*}{\|s^{(k)}\|^2_2} \]

(we assume \( f^* \) is known or can be estimated)

- ‘filtered’ subgradient, \( s^{(k)} = (1 - \beta)g^{(k)} + \beta s^{(k-1)} \), where \( \beta \in [0, 1) \)

- Camerini, Fratta, and Maffioli (1975)

\[ s^{(k)} = g^{(k)} + \beta_k s^{(k-1)}, \quad \beta_k = \max\{0, -\gamma_k(s^{(k-1)})^T g^{(k)}/\|s^{(k-1)}\|_2^2\} \]

where \( \gamma_k \in [0, 2) \) (\( \gamma_k = 1.5 \) ‘recommended’)

Prof. J. Duchi, EE364b, Stanford University
PWL example, Polyak’s step, filtered subgradient, CFM step

\[ f(k) - f^* \]

- Polyak
- filtered \( \beta = 0.25 \)
- CFM

\[ f_{\text{best}}(k) \]

Prof. J. Duchi, EE364b, Stanford University