Subgradient Methods

• subgradient method and stepsize rules

• convergence results and proof

• projected subgradient method

• projected subgradient method for dual

• optimal network flow

Prof. S. Boyd, EE392o, Stanford University
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
Step size rules

step sizes are fixed ahead of time

- **constant step size**: \( \alpha_k = h \) (constant)
- **constant step length**: \( \alpha_k = h/\|g^{(k)}\|_2 \) (so \( \|x^{(k+1)} - x^{(k)}\|_2 = h \))
- **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
  \[
  \lim_{k \to \infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty
  \]
Convergence results

- we assume $\|g\|_2 \leq G$ for all $g \in \partial f$ (equivalent to Lipschitz condition on $f$), $p^* = \inf_x f(x) > -\infty$

- define $f_{\text{best}}^{(k)} = \min_{i=0,...,k} f(x^{(i)})$, best value found in $k$ iterations, and $\bar{f} = \lim_{k \to \infty} f_{\text{best}}^{(k)}$

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

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

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

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

let \( x^* \) be any minimizer of \( f \)

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

using \( p^* = f(x^*) \geq f(x^{(k)}) + g^{(k)T}(x^* - x^{(k)}) \)
applying recursively, and using \( \| g^{(i)} \|_2 \leq G \), we get

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

now we use

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

to get

\[
f_{\text{best}}^{(k)} - p^* \leq \frac{\| x^{(1)} - x^* \|_2^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}.
\]
**constant step size:** for $\alpha_k = h$ we get

$$f_{\text{best}}^{(k)} - p^* \leq \frac{\|x^{(1)} - x^*\|_2^2 + G^2kh^2}{2kh}$$

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

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

then

\[
f^{(k)}_{\text{best}} - p^* \leq \frac{\|x^{(1)} - x^*\|_2^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}
\]

as \( k \to \infty \), numerator converges to a finite number, denominator converges to \( \infty \), so \( f^{(k)}_{\text{best}} \to p^* \)
Example: Piecewise linear minimization

\[
\text{minimize } \quad \begin{array}{c}
   f(x) = \max_{i=1, \ldots, m} (a_i^T x + b_i)
\end{array}
\]

to find a subgradient of \( f \): find index \( j \) for which

\[
\begin{array}{c}
   a_j^T x + b_j = \max_{i=1, \ldots, m} (a_i^T x + b_i)
\end{array}
\]

and take \( g = a_j \)

subgradient method: \( x^{(k+1)} = x^{(k)} - \alpha_k a_j \)
problem instance with $n = 10$ variables, $m = 100$ terms

constant step length, $h = 0.05, 0.02, 0.005$
$f_{\text{best}}^{(k)} - p^*$ and upper bound, constant step length $h = 0.02$
constant step size $h = 0.05, 0.02, 0.005$
diminishing step rule $\alpha = 0.1/\sqrt{k}$ and square summable step size rule $\alpha = 0.1/k$. 
$f_{\text{best}}^{(k)} - p^*$ and upper bound, diminishing step size rule $\alpha = 0.1/\sqrt{k}$
constant step length $h = 0.02$, diminishing step size rule $\alpha = 0.1/\sqrt{k}$, and square summable step rule $\alpha = 0.1/k$
Projected subgradient method

solves constrained optimization problem

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in \mathcal{C},
\end{align*}
\]

where \( f : \mathbb{R}^n \to \mathbb{R}, \mathcal{C} \subseteq \mathbb{R}^n \) are convex

projected subgradient method is given by

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

\( P \) is (Euclidean) projection on \( \mathcal{C} \), and \( g^{(k)} \in \partial f(x^{(k)}) \)
same convergence results:

- for constep step size, converges to neighborhood of optimal (for $f$ differentiable and $h$ small enough, converges)
- for diminishing nonsummable step sizes, converges

key idea: projection does not increase distance to $x^*$

approximate projected subgradient: $P$ only needs to satisfy

$$P(u) \in \mathcal{C}, \quad \|P(u) - z\|_2 \leq \|u - z\|_2 \text{ for any } z \in \mathcal{C}$$
Projected subgradient for dual problem

(convex) primal:

\[
\begin{align*}
\text{minimize} \quad & f_0(x) \\
\text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m
\end{align*}
\]

solve dual problem

\[
\begin{align*}
\text{maximize} \quad & g(\lambda) \\
\text{subject to} \quad & \lambda \succeq 0
\end{align*}
\]

via projected subgradient method:

\[
\lambda^{(k+1)} = \left( \lambda^{(k)} - \alpha_k h \right)_+ , \quad h \in \partial(-g)(\lambda^{(k)})
\]
Subgradient of negative dual function

assume $f_0$ is strictly convex, and denote, for $\lambda \geq 0$,

$$x^*(\lambda) = \arg\min_z (f_0(z) + \lambda_1 f_1(z) + \cdots + \lambda_m f_m(z))$$

so $g(\lambda) = f_0(x^*(\lambda)) + \lambda_1 f_1(x^*(\lambda)) + \cdots + \lambda_m f_m(x^*(\lambda))$

a subgradient of $-g$ at $\lambda$ is given by $h_i = -f_i(x^*(\lambda))$

projected subgradient method for dual:

$$x^{(k)} = x^*(\lambda^{(k)}), \quad \lambda_i^{(k+1)} = \left(\lambda_i^{(k)} + \alpha_k f_i(x^{(k)})\right)_+$$
note:

- primal iterates $x^{(k)}$ are not feasible, but become feasible in limit
- subgradient of $-g$ directly gives the violation of the primal constraints
- dual function values $g(\lambda^{(k)})$ converge to $p^*$
Example: Optimal network flow

- connected directed graph with $n$ links, $p$ nodes
- variable $x_j$ denotes the flow or traffic on arc $j$ (can be $< 0$)
- given external source (or sink) flow $s_i$ at node $i$, $1^T s = 0$
- flow conservation: $Ax = s$, where $A \in \mathbb{R}^{p \times n}$ is the node incidence matrix
- $\phi_j : \mathbb{R} \to \mathbb{R}$ convex flow cost function for link $j$

optimal (single commodity) network flow problem:

\[
\begin{align*}
\text{minimize} & \quad \sum_{j=1}^{n} \phi_j(x_j) \\
\text{subject to} & \quad Ax = s
\end{align*}
\]
Dual network flow problem

Langrangian is

\[ L(x, \nu) = \sum_{j=1}^{n} \phi_j(x_j) + \nu^T (s - Ax) \]

\[ = \sum_{j=1}^{n} (\phi_j(x_j) - \Delta \nu_j x_j) + \nu^T s \]

- we interpret \( \nu_i \) as potential at node \( i \)
- \( \Delta \nu_j \) denotes potential difference across link \( j \)
dual function is

\[ q(\nu) = \inf_x L(x, \nu) = \sum_{j=1}^{n} \inf_{x_j} (\phi_j(x_j) - \Delta \nu_j x_j) + \nu^T s \]

\[ = - \sum_{j=1}^{n} \phi_j^*(\Delta \nu_j) + \nu^T s \]

(\( \phi_j^* \) is conjugate function of \( \phi_j \))

dual network flow problem:

\[ \text{maximize} \quad q(\nu) = - \sum_{j=1}^{n} \phi_j^*(\Delta \nu_j) + \nu^T s \]
Optimal network flow via dual

assume $\phi_i$ strictly convex, and denote

$$x_j^*(\Delta \nu_j) = \arg\min_{x_j} (\phi_j(x_j) - \Delta \nu_j x_j)$$

if $\nu^*$ is optimal solution of the dual network flow problem,

$$x_j^* = x_j^*(\Delta \nu_j^*)$$

is optimal flow
Electrical network analogy

- electrical network with node incidence matrix $A$, nonlinear resistors in branches
- variable $x_j$ is the current flow in branch $j$
- source $s_i$ is external current injected at node $i$ (must sum to zero)
- flow conservation equation $Ax = s$ is Kirchhoff Current Law (KCL)
- dual variables are node potentials; $\Delta \nu_j$ is $j$th branch voltage
- branch current-voltage characteristic is $x_j = x_j^*(\Delta \nu_j)$

then, current and potentials in circuit are optimal flows and dual variables
Subgradient of negative dual function

A subgradient of the negative dual function $-q$ at $\nu$ is

$$g = Ax^*(\Delta \nu) - s$$

$i$th component is $g_i = a_i^T x^*(\Delta \nu) - s_i$, which is flow excess at node $i$.
Subgradient method for dual

subgradient method applied to dual can be expressed as:

\[ x_j := x_j^*(\Delta \nu_j) \]
\[ g_i := \alpha_i^T x - s_i \]
\[ \nu_i := \nu_i - \alpha g_i \]

interpretation:

- optimize each flow, given potential difference, without regard for flow conservation
- evaluate flow excess
- update potentials to correct flow excesses
Example: Minimum queueing delay

Flow cost function

\[
\phi_j(x_j) = \frac{|x_j|}{c_j - |x_j|}, \quad \text{dom } \phi_j = (-c_j, c_j)
\]

where \(c_j > 0\) are given link capacities
(\(\phi_j(x_j)\) gives expected waiting time in queue with exponential arrivals at rate \(x_j\), exponential service at rate \(c_j\))

Conjugate is

\[
\phi_j^*(y) = \begin{cases} 
0 & |y| \leq 1/c_j \\
\left(\sqrt{|c_jy|} - 1\right)^2 & |y| > 1/c_j
\end{cases}
\]
cost function $\phi(x)$ (left) and its conjugate $\phi^*(y)$ (right), $c = 1$

(note conjugate is differentiable)
$x_j^*(\Delta \nu_j)$, for $c_j = 1$

gives flow as function of potential difference across link
A specific example

network with 5 nodes, 7 links, capacities $c_j = 1$
Optimal flow

optimal flows shown as width of arrows; optimal dual variables shown in nodes; potential differences shown on links
Convergence of dual function

constant stepsize rule, $\alpha = 0.1, 1, 2, 3$

for $\alpha = 1, 2$, converges to $p^* = 2.48$ in around about 40 iterations
Convergence of primal residual

The diagram shows the convergence of the primal residual \( \|Ax(k) - s\| \) as a function of iteration \( k \). The curves represent different values of \( \alpha \) as indicated in the legend:

- \( \alpha = 0 \) (black dashed line)
- \( \alpha = 1 \) (blue dashed line)
- \( \alpha = 2 \) (red dashed line)
- \( \alpha = 3 \) (green line)

As \( k \) increases, the residual \( \|Ax(k) - s\| \) decreases, indicating convergence to a solution.
convergence of dual function, nonsummable diminishing stepsize rules
convergence of primal residual, nonsummable diminishing stepsize rules
Convergence of dual variables

$\nu^{(k)}$ versus iteration number $k$, constant stepsize rule $\alpha = 2$

($\nu_5$ is fixed as zero)