Subgradient methods

Yuxin Chen
Princeton University, Spring 2018
Outline

- Steepest descent
- Subgradients
- Projected subgradient descent
  - Convex and Lipschitz problems
  - Strongly convex and Lipschitz problems
Nondifferentiable problems

Differentiability of objective function $f$ is essential for validity of gradient methods

However, there is no shortage of interesting cases (e.g. $\ell_1$ minimization, nuclear norm minimization) where non-differentiability is present at some points
Generalizing steepest descent?

minimize \( x \) \( f(x) \) subject to \( x \in C \)

- find search direction \( d^t \) that minimizes directional derivative

\[
d^t \in \arg \min_{d : \|d\|_2 \leq 1} \ f'(x^t, d)
\]

where \( f'(x, d) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha} \)

- updates

\[
x^{t+1} = x^t + \eta_t d^t
\]
Issues

- Finding steepest descent direction (or even finding descent direction) may involve *expensive* computation

- Stepsize rule is tricky to choose: for certain popular stepsize rule (like exact line search), steepest descent might converge to non-optimal points
Wolfe’s example

\[
f(x_1, x_2) = \begin{cases} 
5(9x_1^2 + 16x_2^2)^{\frac{1}{2}} & \text{if } x_1 > |x_2| \\
9x_1 + 16|x_2| & \text{if } x_1 \leq |x_2|
\end{cases}
\]

- (0,0) is non-differentiable point
- if one starts from \(x^0 = \left(\frac{16}{9}, 1\right)\) and uses exact line search, then
  - \(\{x^t\}\) are all differentiable points
  - \(x^t \to (0, 0)\) as \(t \to \infty\)
Wolfe’s example

\[ f(x_1, x_2) = \begin{cases} 
5(9x_1^2 + 16x_2^2)^{\frac{1}{2}} & \text{if } x_1 > |x_2| \\
9x_1 + 16|x_2| & \text{if } x_1 \leq |x_2| 
\end{cases} \]

- even though it never hits non-differentiable points, steepest descent with exact line search gets stuck around non-optimal point (i.e. \((0,0)\))

- **problem**: steepest descent direction may undergo large/discontinuous change close to convergence limit
Practically, one often resorts to subgradient-based approach

\[ x^{t+1} = \mathcal{P}_C(x^t - \eta_t g^t) \]  

(4.1)

where \( g^t \) is any subgradient of \( f \) at \( x^t \)

- as we will see, this update rule does not necessarily yield cost reduction
Subgradients
We say \( g \) is subgradient of \( f \) at point \( x \) if
\[
f(z) \geq f(x) + g^\top(z - x) , \quad \forall z
\]
\[(4.2)\]

- set of all subgradients of \( f \) at \( x \) is called subdifferential of \( f \) at \( x \), denoted by \( \partial f(x) \)
Example: $f(x) = |x|$
Example: subgradient of norms at 0

Let \( f(x) = \|x\| \) for any norm \( \| \cdot \| \), then for any \( g \) obeying \( \|g\|* \leq 1 \),

\[
g \in \partial f(0)
\]

where \( \| \cdot \|_* \) denotes dual norm of \( \| \cdot \| \) (i.e. \( \|x\|_* := \sup_{z: \|z\| \leq 1} \langle z, x \rangle \))

Proof: To see this, it suffices to prove that

\[
f(z) \geq f(0) + \langle g, z - 0 \rangle, \quad \forall z
\]

\[
\iff \langle g, z \rangle \leq \|z\|, \quad \forall z
\]

This follows from generalized Cauchy-Schwarz, i.e.

\[
\langle g, z \rangle \leq \|g\|_* \|z\| \leq \|z\|
\]
Example: \( \max\{f_1(x), f_2(x)\} \)

\[
f(x) = \max\{f_1(x), f_2(x)\}
\]

\[
\partial f(x) = \begin{cases} 
\{f'_1(x)\}, & \text{if } f_1(x) > f_2(x) \\
[f'_1(x), f'_2(x)], & \text{if } f_1(x) = f_2(x) \\
\{f'_2(x)\}, & \text{if } f_1(x) < f_2(x) 
\end{cases}
\]
Basic rules

- **scaling:** $\partial (\alpha f) = \alpha \partial f$

- **summation:** $\partial (f_1 + f_2) = \partial f_1 + \partial f_2$
Example: $\ell_1$ norm

$$f(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i| = \sum_{i: x_i \neq 0} \text{sgn}(x_i) e_i$$

since

$$\partial f_i(x) = \begin{cases} 
\text{sgn}(x_i) e_i, & \text{if } x_i \neq 0 \\
[-1, 1] \cdot e_i, & \text{if } x_i = 0 
\end{cases}$$

we have

$$\sum_{i: x_i \neq 0} \text{sgn}(x_i) e_i \in \partial f(x)$$
Basic rules (cont.)

• affine transformation: if \( h(x) = f(Ax + b) \), then

\[
\partial h(x) = A^\top \partial f(Ax + b)
\]
**Example:** \( \|Ax + b\|_1 \)

\[
h(x) = \|Ax + b\|_1
\]

Letting \( f(x) = \|x\|_1 \) and \( A = [a_1, \cdots, a_m]^\top \), we have

\[
g = \sum_{i : a_i^\top x + b_i \neq 0} \text{sgn}(a_i^\top x + b_i)e_i \in \partial f(Ax + b).
\]

\[
\implies A^\top g = \sum_{i : a_i^\top x + b_i \neq 0} \text{sgn}(a_i^\top x + b_i)a_i \in \partial h(x)
\]

**Subgradient methods**
Basic rules (cont.)

- **chain rule:** suppose $f$ is convex, and $g$ is differentiable, nondecreasing, and convex. Let $h = g \circ f$, then
  
  $$\partial h(x) = g'(f(x)) \partial f(x)$$

- **composition:** suppose $f(x) = h(f_1(x), \cdots, f_n(x))$, where $f_i$’s are convex, and $h$ is differentiable, nondecreasing, and convex. Let $q = \nabla h(y) \big|_{y = [f_1(x), \cdots, f_n(x)]}$, and $g_i \in \partial f_i(x)$. Then
  
  $$q_1 g_1 + \cdots + q_n g_n \in \partial f(x)$$
Basic rules (cont.)

- **pointwise maximum**: if \( f(x) = \max_{1 \leq i \leq k} f_i(x) \), then
  \[
  \partial f(x) = \text{conv} \left\{ \bigcup \{ \partial f_i(x) \mid f_i(x) = f(x) \} \right\}
  \]
  convex hull of subdifferentials of all active functions

- **pointwise supremum**: if \( f(x) = \sup_{\alpha \in \mathcal{F}} f_\alpha(x) \), then
  \[
  \partial f(x) = \text{closure} \left( \text{conv} \left\{ \bigcup \{ \partial f_\alpha(x) \mid f_\alpha(x) = f(x) \} \right\} \right)
  \]
Example: piece-wise linear function

\[ f(x) = \max_{1 \leq i \leq m} \{ a_i^\top x + b_i \} \]

pick any \( a_j \) s.t. \( a_j^\top x + b_j = \max_i \{ a_i^\top x + b_i \} \), then

\[ a_j \in \partial f(x) \]
Example: $\ell_\infty$ norm

\[
f(x) = \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|
\]

if $x \neq 0$, then pick any $x_j$ obeying $|x_j| = \max_i |x_i|$ to get

\[
\text{sgn}(x_j) e_j \in \partial f(x)
\]
Example: maximum eigenvalue

\[ f(\mathbf{x}) = \lambda_{\text{max}} (x_1 \mathbf{A}_1 + \cdots + x_n \mathbf{A}_n) \]

where \( \mathbf{A}_1, \cdots, \mathbf{A}_n \) are real symmetric matrices

Rewrite

\[ f(\mathbf{x}) = \sup_{\mathbf{y} \colon \|\mathbf{y}\|_2 = 1} \mathbf{y}^\top (x_1 \mathbf{A}_1 + \cdots + x_n \mathbf{A}_n) \mathbf{y} \]

as supremum of affine functions of \( \mathbf{x} \). Therefore, taking \( \mathbf{y} \) as leading eigenvector of \( x_1 \mathbf{A}_1 + \cdots + x_n \mathbf{A}_n \), we have

\[ [\mathbf{y}^\top \mathbf{A}_1 \mathbf{y}, \cdots, \mathbf{y}^\top \mathbf{A}_n \mathbf{y}]^\top \in \partial f(\mathbf{x}) \]
Example: nuclear norm

Let \( X \in \mathbb{R}^{m \times n} \) with SVD \( X = U \Sigma V^\top \) and

\[
 f(X) = \min\{n,m\} \sum_{i=1}^{\min\{n,m\}} \sigma_i(X)
\]

where \( \sigma_i(x) \) is \( i \)th largest singular value of \( X \)

Rewrite

\[
 f(X) = \sup_{\text{orthonormal } A, B} \langle AB^\top, X \rangle := \sup_{\text{orthonormal } A, B} f_{A,B}(X)
\]

Recognizing that \( f_{A,B}(X) \) is maximized by \( A = U \) and \( B = V \) and that \( \nabla f_{A,B}(X) = AB^\top \), we have

\[
 UV^\top \in \partial f(X)
\]
Negative subgradient is not necessarily descent direction

Example: \( f(x) = |x_1| + 3|x_2| \)

at \( x = (1, 0) \):

- \( g_1 = (1, 0) \in \partial f(x) \), and \(-g_1\) is descent direction
- \( g_2 = (1, 3) \in \partial f(x) \), but \(-g_2\) is not a descent direction

Reason: lack of continuity — one can change direction significantly without violating validity of subgradient
Negative subgradient is not necessarily descent direction

Since $f(x^t)$ is not necessarily monotone, we will keep track of best point

$$f^{\text{best}, t} := \min_{1 \leq i \leq t} f(x^i)$$

We also denote by $f^{\text{opt}} := \min_x f(x)$ optimal objective value
Convex and Lipschitz problems

Clearly, we cannot analyze all nonsmooth functions. A nice (and widely encountered) class to start with is Lipschitz functions, i.e. set of all $f$ obeying

$$|f(x) - f(z)| \leq L_f \|x - z\|_2 \quad \forall x \text{ and } z$$
Fundamental inequality for projected subgradient methods

We’d like to optimize $\|x^{t+1} - x^*\|_2^2$, but don’t have access to $x^*$

Key idea (majorization-minimization): find another function that majorizes $\|x^{t+1} - x^*\|_2^2$, and optimize majorizing function

**Lemma 4.1**

Projected subgradient update rule (4.1) obeys

$$\|x^{t+1} - x^*\|_2^2 \leq \|x^t - x^*\|_2^2 - 2\eta_t (f(x^t) - f^{opt}) + \eta_t^2 \|g^t\|_2^2 \quad (4.3)$$

fixed

majorizing function
Proof of Lemma 4.1

\[ \|x^{t+1} - x^*\|^2 = \|P_C(x^t - \eta_t g^t) - P_C(x^*)\|^2 \]
\[ \leq \|x^t - \eta_t g^t - x^*\|^2 \quad \text{(nonexpansiveness of projection)} \]
\[ = \|x^t - x^*\|^2 - 2\eta_t \langle x^t - x^*, g^t \rangle + \eta_t^2 \|g^t\|^2 \]
\[ \leq \|x^t - x^*\|^2 - 2\eta_t (f(x^t) - f(x^*)) + \eta_t^2 \|g^t\|^2 \]

where last line uses subgradient inequality

\[ f(x^*) - f(x^t) \geq \langle x^* - x^t, g^t \rangle \]
Polyak’s stepsize rule

Majorizing function given in (4.3) suggests stepsize (Polyak ’87)

\[ \eta_t = \frac{f(x^t) - f^{\text{opt}}}{\|g_t\|^2} \]  

(4.4)

which leads to error reduction

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

(4.5)

- useful if \( f^{\text{opt}} \) is known
- estimation error is monotonically decreasing with Polyak’s stepsize
Let $C_1$, $C_2$ be closed convex sets and suppose $C_1 \cap C_2 \neq \emptyset$

\[
\text{find } \ x \in C_1 \cap C_2
\]

\[
\iff
\]

\[
\text{minimize}_x \quad \max \{ \text{dist}_{C_1}(x), \text{dist}_{C_2}(x) \}
\]

where $\text{dist}_{C}(x) := \min_{z \in C} \| x - z \|_2$
Example: projection onto intersection of convex sets

Let $C_1, C_2$ be closed convex sets and suppose $C_1 \cap C_2 = \emptyset$

Find $x \in C_1 \cap C_2$ that minimizes $\max\{\text{dist} C_1(x), \text{dist} C_2(x)\}$

where $\text{dist} C(x) = \min z \in C \land x \neq z \in C_2$

For this problem, subgradient method with Polyak’s stepsize rule is equivalent to alternating projection

\[
x^{t+1} = P_{C_1}(x^t), \quad x^{t+2} = P_{C_2}(x^{t+1})
\]
Example: projection onto intersection of convex sets

**Proof:** Use subgradient rule for pointwise max function to get

\[ g^t \in \partial \text{dist}_{C_i}(x^t) \]

where \( i = \arg \max_{j=1,2} \text{dist}_{C_j}(x^t) \)

If \( \text{dist}_{C_i}(x^t) \neq 0 \), then one has

\[ g^t = \nabla \text{dist}_{C_i}(x^t) = \frac{x^t - P_{C_i}(x^t)}{\text{dist}_{C_i}(x^t)} \]

which follows since \( \nabla \left( \frac{1}{2} \text{dist}_{C_i}^2(x^t) \right) = x^t - P_{C_i}(x^t) \) (homework) and

\[ \nabla \left( \frac{1}{2} \text{dist}_{C_i}^2(x^t) \right) = \text{dist}_{C_i}(x^t) \cdot \nabla \text{dist}_{C_i}(x^t) \]
Example: projection onto intersection of convex sets

Proof (cont.): Adopting Polya’s stepsize rule and recognizing that $\|g^t\|_2 = 1$ we reach

$$x^{t+1} = x^t - \eta_t g^t = x^t - \frac{\text{dist}_{C_i}(x^t)}{\|g^t\|_2^2} \left( x^t - \mathcal{P}_{C_i}(x^t) \right)$$

$$= \mathcal{P}_{C_i}(x^t)$$

where $i = \arg \max_{j=1,2} \text{dist}_{C_j}(x^t)$
Theorem 4.2 (Convergence of projected subgradient method with Polyak’s stepsize)

Suppose \( f \) is convex and \( L_f \)-Lipschitz continuous. Then projected subgradient (4.1) with Polyak’s stepsize rule obeys

\[
f^{\text{best},t} - f^{\text{opt}} \leq \frac{L_f \| \mathbf{x}^0 - \mathbf{x}^* \|_2}{\sqrt{t + 1}}
\]

- sublinear convergence rate \( O(1/\sqrt{t}) \)
Proof of Theorem 4.2

We have seen from (4.5) that

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

\[
\leq \left\{ \|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2 \right\} L_f^2
\]

Applying recursively for all iterations (from 0th to tth) and summing them up yield

\[
\sum_{k=0}^{t} (f(x^k) - f(x^*))^2 \leq \left\{ \|x^0 - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2 \right\} L_f^2
\]

\[
\Rightarrow \quad (t + 1)(f^{\text{best},t} - f^{\text{opt}})^2 \leq \|x^0 - x^*\|_2^2 L_f^2
\]

which concludes proof
Unfortunately, Polyak’s stepsize rule requires knowledge of $f^{opt}$, which is often unknown \textit{a priori}.

This calls for simpler rules for setting stepsizes.
Theorem 4.3 (Subgradient methods for convex and Lipschitz functions)

Suppose \( f \) is convex and \( L_f \)-Lipschitz continuous. Then projected subgradient update rule (4.1) obeys

\[
 f^{\text{best},t} - f^{\text{opt}} \leq \frac{\|x^0 - x^*\|^2}{2} + L_f^2 \sum_{i=0}^{t} \eta_i^2 \cdot \frac{2 \sum_{i=0}^{t} \eta_i}{2 \sum_{i=0}^{t} \eta_i}
\]
Implications: stepsize rules

- **Constant step size** $\eta_t \equiv \eta$:

  \[
  \lim_{t \to \infty} f^{\text{best},t} \leq \frac{L_f^2 \eta}{2}
  \]

  i.e. may converge to non-optimal point

- **Diminishing step size obeying** $\sum_t \eta_t^2 < \infty$ and $\sum_t \eta_t \to \infty$:

  \[
  \lim_{t \to \infty} f^{\text{best},t} = 0
  \]

  i.e. converges to optimal point

Subgradient methods
Implications: stepsize rule

- Optimal choice? $\eta_t = \frac{1}{\sqrt{t}}$:

$$f^{\text{best},t} - f^{\text{opt}} \lesssim \frac{\|x^0 - x^*\|_2^2 + L_f^2 \log t}{\sqrt{t}}$$

i.e. attains $\varepsilon$-accuracy within about $O(1/\varepsilon^2)$ iterations (ignoring logarithmic factor)
Proof of Theorem 4.3

Applying Lemma 4.1 recursively gives

\[ \|x^{t+1} - x^*\|^2_2 \leq \|x^0 - x^*\|^2_2 - 2 \sum_{i=0}^{t} \eta_i (f(x^i) - f^{opt}) + \sum_{i=0}^{t} \eta_i^2 \|g^i\|^2_2 \]

Rearranging terms, we are left with

\[ 2 \sum_{i=0}^{t} \eta_i (f(x^i) - f^{opt}) \leq \|x^0 - x^*\|^2_2 - \|x^{t+1} - x^*\|^2_2 + \sum_{i=0}^{t} \eta_i^2 \|g^i\|^2_2 \]

\[ \leq \|x^0 - x^*\|^2_2 + L_f^2 \sum_{i=0}^{t} \eta_i^2 \]

\[ \Rightarrow f_{\text{best},t} - f^{opt} \leq \frac{\sum_{i=0}^{t} \eta_i (f(x^i) - f^{opt})}{\sum_{i=0}^{t} \eta_i} \leq \frac{\|x^0 - x^*\|^2_2 + L_f^2 \sum_{i=0}^{t} \eta_i^2}{2 \sum_{i=0}^{t} \eta_i} \]

Subgradient methods 4-39
If $f$ is strongly convex, then convergence rate can be improved to $O(1/t)$, as long as stepsize diminishes at $O(1/t)$

**Theorem 4.4 (Subgradient methods for strongly convex and Lipschitz functions)**

Let $f$ be $\mu$-strongly convex and $L_f$-Lipschitz continuous over $C$. If $\eta_t \equiv \eta = \frac{2}{\mu(t+1)}$, then

$$f_{\text{best},t} - f^{\text{opt}} \leq \frac{2L_f^2}{\mu} \cdot \frac{1}{t+1}$$
Proof of Theorem 4.4

When $f$ is $\mu$-strongly convex, we can improve Lemma 4.1 to (exercise)

$$\|x^{t+1} - x^*\|^2_2 \leq (1 - \mu \eta_t)\|x^t - x^*\|^2_2 - 2\eta_t \left(f(x^t) - f^{\text{opt}}\right) + \eta_t^2 \|g^t\|^2_2$$

$$\implies f(x^t) - f^{\text{opt}} \leq \frac{1 - \mu \eta_t}{2\eta_t} \|x^t - x^*\|^2_2 - \frac{1}{2\eta_t} \|x^{t+1} - x^*\|^2_2 + \frac{\eta_t}{2} \|g^t\|^2_2$$

Since $\eta_t = 2/(\mu(t + 1))$, we have

$$f(x^t) - f^{\text{opt}} \leq \frac{\mu(t - 1)}{4} \|x^t - x^*\|^2_2 - \frac{\mu(t + 1)}{4} \|x^{t+1} - x^*\|^2_2 + \frac{1}{\mu(t + 1)} \|g^t\|^2_2$$

and hence

$$t \left(f(x^t) - f^{\text{opt}}\right) \leq \frac{\mu t(t - 1)}{4} \|x^t - x^*\|^2_2 - \frac{\mu t(t + 1)}{4} \|x^{t+1} - x^*\|^2_2 + \frac{1}{\mu} \|g^t\|^2_2$$
Proof of Theorem 4.4 (cont.)

Summing over all iterations before $t$, we get

$$\sum_{k=0}^{t} k \left( f(x^k) - f^{\text{opt}} \right) \leq 0 - \frac{\mu t(t+1)}{4} \|x^{t+1} - x^*\|_2^2 + \frac{1}{\mu} \sum_{k=0}^{t} \|g^k\|_2^2$$

$$\leq \frac{t}{\mu} L_f^2$$

$$\implies f^{\text{best},k} - f^{\text{opt}} \leq \frac{L_f^2}{\mu} \frac{t}{\sum_{k=0}^{t} k} \leq \frac{2L_f^2}{\mu} \frac{1}{t + 1}$$
Numerical example
## Summary: subgradient methods

<table>
<thead>
<tr>
<th></th>
<th>stepsize rule</th>
<th>convergence rate</th>
<th>iteration complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>convex &amp; Lipschitz</td>
<td>$\eta_t \approx \frac{1}{\sqrt{t}}$</td>
<td>$O\left(\frac{1}{\sqrt{t}}\right)$</td>
<td>$O\left(\frac{1}{\varepsilon^2}\right)$</td>
</tr>
<tr>
<td>problems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strongly convex &amp;</td>
<td>$\eta_t \approx \frac{1}{t}$</td>
<td>$O\left(\frac{1}{t}\right)$</td>
<td>$O\left(\frac{1}{\varepsilon}\right)$</td>
</tr>
<tr>
<td>Lipschitz problems</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reference


