Lecture: Smoothing

http://bicmr.pku.edu.cn/~wenzw/opt-2018-fall.html

Acknowledgement: this slides is based on Prof. Lieven Vandenberghe’s lecture notes
Smoothing

- introduction
- smoothing via conjugate
- examples
First-order convex optimization methods

complexity of finding \( \epsilon \)-suboptimal point of \( f(x) \)

- subgradient method: \( f \) nondifferentiable with Lipschitz constant \( G \)
  \[ O\left(\left(\frac{G}{\epsilon}\right)^2\right) \text{ iterations} \]

- proximal gradient method: \( f = g + h \), where \( h \) is a 'simple' nondifferentiable function, \( g \) is differentiable with \( L \)–Lipschitz continuous gradient
  \[ O\left(\frac{L}{\epsilon}\right) \text{ iterations} \]

- fast proximal gradient methods
  \[ O\left(\sqrt{\frac{L}{\epsilon}}\right) \text{ iterations} \]
for nondifferentiable $f$ that cannot be handled by proximal gradient method

- replace $f$ with differentiable approximation $f_\mu$ (parametrized by $\mu$)
- minimize $f_\mu$ by (fast) gradient method

**Complexity:** #iterations for (fast) gradient method depends on $L_\mu/\epsilon_\mu$

- $L_\mu$ is Lipschitz constant of $\nabla f_\mu$
- $\epsilon_\mu$ is accuracy with which the smooth problem is solved

**trade-off** in amount of smoothing (choice of $\mu$)

- Large $L_\mu$ (less smoothing) gives more accurate approximation
- Small $L_\mu$ (more smoothing) gives faster convergence
Example: Huber penalty as smoothed absolute value

\[
\phi_{\mu}(z) = \begin{cases} 
\frac{z^2}{2\mu} & |z| \leq \mu \\
|z| - \frac{\mu}{2} & |z| \geq \mu 
\end{cases}
\]

\(\mu\) controls accuracy and smoothness

- accuracy

\[|z| - \frac{\mu}{2} \leq \phi_{\mu}(z) \leq |z|\]

- smoothness

\[\phi_{\mu}''(z) \leq \frac{1}{\mu}\]
Huber penalty approximation of 1-norm minimization

\[ f(x) = \|Ax - b\|_1, \quad f_\mu(x) = \sum_{i=1}^{m} \phi_\mu(a_i^T x - b_i) \]

- accuracy: from \( f(x) - m\mu/2 \leq f_\mu(x) \leq f(x) \),

\[ f(x) - f^* \leq f_\mu(x) - f^*_\mu + m\mu/2 \]

- to achieve \( f(x) - f^* \leq \epsilon \), we need \( f_\mu(x) - f^*_\mu \leq \epsilon_\mu \) with \( \epsilon_\mu = \epsilon - m\mu/2 \)

- Lipschitz constant of \( f_\mu \) is \( L_\mu = \|A\|_2^2/\mu \)

**complexity:** for \( \mu = \epsilon/m \)

\[ \frac{L_\mu}{\epsilon_\mu} - \frac{\|A\|_2^2}{\mu(\epsilon - m\mu/2)} = \frac{2m\|A\|_2^2}{\epsilon^2} \]

i.e., \( O(\sqrt{L_\mu/\epsilon_\mu} = O(1/\epsilon) \) iteration complexity for fast gradient method
Outline

- introduction
- smoothing via conjugate
- examples
Minimum of strongly convex function

if \( x \) is a minimizer of a strongly convex function \( f \), then it is unique and

\[
f(y) \geq f(x) + \frac{\mu}{2} \|y - x\|_2^2 \quad \forall y \in \text{dom} f
\]

(\( \mu \) is the strong convexity constant of \( f \))

proof: if some \( y \) does not satisfy the inequality, then for some small \( \theta > 0 \):

\[
f((1 - \theta)x + \theta y) \leq (1 - \theta)f(x) + \theta f(y) - \mu \frac{\theta(1 - \theta)}{2} ||y - x||_2^2
\]

\[
= f(x) + \theta(f(y) - f(x)) - \frac{\mu}{2} ||y - x||_2^2 + \mu \frac{\theta^2}{2} ||x - y||_2^2
\]

\[
< f(x)
\]
Conjugate of strongly convex function

suppose \( f \) is closed and strongly convex with constant \( \mu \) and conjugate

\[
f^*(y) = \sup_{x \in \text{dom} f} (y^T x - f(x))
\]

\( f^* \) is defined and differentiable at all \( y \), with gradient

\[
\nabla f^*(y) = \arg\max_{x} (y^T x - f(x))
\]

\( \nabla f^* \) is Lipschitz continuous with constant \( 1/\mu \)

\[
\|\nabla f^*(u) - \nabla f^*(v)\|_2 \leq \frac{1}{\mu} \|u - v\|_2
\]
Outline of proof

- $y^T x - f(x)$ has a unique maximizer $x_y$ for every $y$ (follows from closedness and strong convexity of $f(x) - y^T x$)
- $\nabla f^*(y) = x_y$
- From strong convexity (with $x_\mu = \nabla f^*(u)$, $x_v = \nabla f^*(v)$)

\[
\begin{align*}
f(x_u) - v^T x_u & \geq f(x_v) - v^T x_v + \frac{\mu}{2} \|x_u - x_v\|^2_2 \\
f(x_v) - u^T x_v & \geq f(x_u) - u^T x_u + \frac{\mu}{2} \|x_u - x_v\|^2_2
\end{align*}
\]

Adding the left- and right-hand sides of the inequalities gives

\[
\mu \|x_u - x_v\|^2_2 \leq (x_u - x_v)^T (u - v)
\]

By the Cauchy-Schwarz inequality, $\mu \|x_u - x_v\|_2 \leq \|u - v\|_2$
Proximity function

$d$ is a **proximity function** for a closed convex set $C$ if

- $d$ is continuous and strongly convex
- $C \subseteq \text{dom}d$

$d(x)$ measures 'distance' of $x$ to the center $x_d = \arg\min_{x \in C} d(x)$ of $C$

**Normalization**

- we will assume the strong convexity constant is 1 and $\inf_{x \in C} d(x) = 0$
- for a normalized proximity function

\[
d(x) \geq \frac{1}{2} \|x - x_d\|_2^2 \quad \forall x \in C
\]
common proximity functions

- $d(x) = \|x - u\|_2^2 / 2$ with $x_d = u \in C$
- $d(x) = \sum_{i=1}^{n} w_i(x_i - u_i)^2 / 2$ with $w_i \geq 1$ and $x_d = u \in C$
- $d(x) = \sum_{i=1}^{n} x_i \log x_i + \log n$ for $C = \{x \geq 0 \mid 1^T x = 1\}$, $x_d = (1/n)1$

example (probability simplex): entropy and $d(x) = (1/2)\|x - (1/n)1\|_2^2$
Smoothing via conjugate

**conjugate (dual) representation:** suppose $f$ can be expressed as

$$f(x) = \sup_{y \in \text{dom} h} \left( (Ax + b)^T y - h(y) \right)$$

$$= h^*(Ax + b)$$

where $h$ is closed and convex with **bounded** domain

**smooth approximation:** choose proximity function $d$ for $C = \text{cldom} h$

$$f_\mu(x) = \sup_{y \in \text{dom} h} \left( (Ax + b)^T y - h(y) - \mu d(y) \right)$$

$$= (h + \mu d)^*(Ax + b)$$

$f_\mu$ is differentiable because $h + \mu d$ is strongly convex
Example: absolute value

conjugate representation

\[ |x| = \sup_{-1 \leq y \leq 1} xy = h^*(x), \quad h(y) = I_{[-1,1]}(y) \]

proximity function: choosing \( d(y) = \frac{y^2}{2} \) gives Huber penalty

\[ f_{\mu}(x) = \sup_{-1 \leq y \leq 1} (xy - \mu y^2/2) = \begin{cases} 
  \frac{x^2}{2\mu} & |x| \leq \mu \\
  |x| - \frac{\mu}{2} & |x| > \mu 
\end{cases} \]

proximity function: choosing \( d(y) = 1 - \sqrt{1 - y^2} \) gives

\[ f_{\mu}(x) = \sup_{-1 \leq y \leq 1} (xy + \mu \sqrt{1 - y^2} - \mu) = \sqrt{x^2 + \mu^2} - \mu \]
another conjugate representation of $x$

$$|x| = \sup_{y_1+y_2=1 \atop y \succeq 0} x(y_1 - y_2)$$

i.e., $|x| = h^*(ax)$ for $h = I_C$,

$$C = \{y \succeq 0 | y_1 + y_2 = 1\}, \quad A = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

proximity function for $C$

$$d(y) = y_1 \log y_1 + y_2 \log y_2 + \log 2$$

smooth approximation

$$f_\mu(x) = \sup_{y_1+y_2=1} (xy_1 - xy_2 + \mu (y_1 \log y_1 + y_2 \log y_2 + \log 2))$$

$$= \mu \log \left( \frac{e^{x/\mu} + e^{-x/\mu}}{2} \right)$$
comparison: three smooth approximations of absolute value
Gradient of smooth approximation

\[ f_\mu(x) = (h + \mu d)^*(Ax + b) \]

\[ = \sup_{y \in \text{dom} h} ((Ax + b)^T y - h(y) - \mu d(y)) \]

from properties of the conjugate of strongly convex function (page 7)

\( f_\mu \) is differentiable, with gradient

\[ \nabla f_\mu(x) = A^T \arg\max_{y \in \text{dom} h} ((Ax + b)^T y - h(y) - \mu d(y)) \]

\( \nabla f_\mu(x) \) is Lipschitz continuous with constant

\[ L_\mu = \frac{\|A\|^2}{\mu} \]
Accuracy of smooth approximation

\[
f(x) - \mu D \leq f_\mu(x) \leq f(x), \quad D = \sup_{y \in \text{dom} h} d(y)
\]

note $D < +\infty$ because \( \text{dom} h \) is bounded and \( \text{dom} h \subseteq \text{dom} d \)

- lower bound follows from

\[
f_\mu(x) = \sup_{y \in \text{dom} h} ((Ax + b)^T y - h(y) - \mu d(y)) \\
\geq \sup_{y \in \text{dom} h} ((Ax + b)^T y - h(y) - \mu D) \\
= f(x) - \mu D
\]

- upper bound follows from

\[
f_\mu(x) \leq \sup_{y \in \text{dom} h} ((Ax + b)^T y - h(y)) = f(x)
\]
Complexity

to find solution of nondifferentiable problem with accuracy $f(x) - f^* \leq \epsilon$

- solve smoothed problem with accuracy $\epsilon_\mu = \epsilon - \mu D$, so that

$$f(x) - f^* \leq f_\mu(x) + \mu D - f^*_\mu \leq \epsilon_\mu + \mu D = \epsilon$$

- Lipschitz constant of $f_\mu$ is $L_\mu = \|A\|_2^2 / \mu$

**complexity:** for $\mu = \epsilon / (2D)$

$$\frac{L_\mu}{\epsilon_\mu} = \frac{\|A\|_2^2}{\mu(\epsilon - \mu D)} = \frac{4D\|A\|_2^2}{\mu \epsilon^2}$$

- gives $O(1/\epsilon)$ iteration bound for fast gradient method

- efficiency in practice can be improved by decreasing $\mu$ gradually
Outline

- introduction
- smoothing via conjugate
- examples
Piecewise-linear approximation

\[ f(x) = \max_{i=1,\ldots,m} (a_i^T x + b_i) \]

conjugate representation

\[ f(x) = \sup_{y \succeq 0, 1^T y = 1} (A x + b)^T y \]

proximity function

\[ d(y) = \sum_{i=1}^{m} y_i \log y_i + \log m \]

smooth approximation

\[ f_\mu(x) = \mu \log \sum_{i=1}^{m} e^{(a_i^T x + b_i)/\mu} - \mu \log m \]
1-Norm approximation

\[ f(x) = \|Ax - b\|_1 \]

conjugate representation

\[ f(x) = \sup \ (Ax - b)^T y \quad \text{subject to} \quad \|y\|_\infty \leq 1 \]

proximity function

\[ d(y) = \frac{1}{2} \sum_i w_i y_i^2 \quad \text{(with } w_i > 1) \]

smooth approximation: Huber approximation

\[ f_\mu(x) = \sum_{i=1}^n \phi_{\mu w_i}(a_i^T x - b_i) \]
Maximum eigenvalue

**conjugate representation:** for $X \in \mathbb{S}^n$,

$$f(X) = \lambda_{\text{max}}(X) = \sup_{Y \succeq 0, \text{tr}Y = 1} \text{tr}(XY)$$

**proximity function:** negative matrix entropy

$$d(Y) = \sum_{i=1}^n \lambda_i(Y) \log \lambda_i(Y) + \log n$$

**smooth approximation**

$$f_\mu(X) = \sup_{Y \succeq 0, \text{tr}Y = 1} \left( \text{tr}(XY) - \mu d(Y) \right)$$

$$= \mu \log \sum_{i=1}^n e^{\lambda_i(X)/\mu} - \mu \log n$$
Nuclear norm

Nuclear norm $f(X) = \|X\|_\star$ is sum of singular values of $X \in \mathbb{R}^{m \times n}$

Conjugate representation

$$f(X) = \sup_{\|Y\|_2 \leq 1} \text{tr}(X^T Y)$$

Proximity function

$$d(Y) = \frac{1}{2} \|Y\|_F^2$$

Smooth approximation

$$f_\mu(X) = \sup_{\|Y\|_2 \leq 1} \text{tr}(X^T Y - \mu d(Y)) = \sum_i \phi_\mu(\sigma_i(X))$$

The sum of the Huber penalties applied to the singular values of $X$
Lagrange dual function

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

\( f_i(x) \) convex, \( C \) closed and bounded

**smooth approximation of dual function:** choose prox. function \( d \) for \( C \)

\[ g_\mu(\lambda) = \inf_{x \in C} (f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \mu d(x)) \]

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