Smoothing

• introduction

• smoothing via conjugate

• examples
First-order convex optimization methods

complexity of finding $\epsilon$-suboptimal point of $f$

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

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

- fast proximal gradient methods (lecture 7)
  \[ O \left( \frac{\sqrt{L}}{\epsilon} \right) \text{ iterations} \]
Nondifferentiable optimization by smoothing

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

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

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

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

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

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

(\(\mu\) is the strong convexity constant of \(f\); see page 1-17)

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

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

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

\[
< f(x)
\]
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 $f(x) - y^T x$)

• from page 4-18, $\nabla f^*(y) = x_y$

• from strong convexity and page 6 (with $x_u = \nabla f^*(u)$, $x_v = \nabla f^*(v)$)

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

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

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

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

where $h$ is closed and convex with bounded domain

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

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

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

$f_\mu$ 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 \( d(y) = y^2/2 \) gives Huber penalty

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

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 $|x|$

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

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

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

proximity function for $C$

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

smooth approximation

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

$$ = \mu \log \left( \frac{e^{x/\mu} + e^{-x/\mu}}{2} \right) $$
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

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

• \( \nabla f_\mu \) 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

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

- Upper bound follows from

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

\[ f(x) = \sup_{\|y\|_\infty \leq 1} (Ax - b)^T y \]

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 $X \in S^n$,

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

proximity function: negative matrix entropy

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

smooth approximation

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

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

nuclear norm \( f(X) = \|X\|_* \) 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

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

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

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

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

this is equivalent to regularizing the primal problem

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