8. Smoothing

- introduction
- smoothing via conjugate
- examples

First-order convex optimization methods

iteration complexity of finding ϵ-suboptimal solution

- subgradient method: f nondifferentiable with Lipschitz constant G

$$O\big((G/\epsilon)^2\big)$$

- proximal gradient method: minimize $f + \Psi$
  - f differentiable with L-Lipschtiz continuous gradient
  - Ψ nondifferentiable but "simple" ($\mathbf{prox}_\Psi$ easy to compute)

$$O(L/\epsilon)$$

- accelerated proximal gradient methods: same problem class as in proximal gradient method

$$O(\sqrt{L/\epsilon})$$

Nondifferentiable optimization by smoothing

for nondifferentiable f that cannot be handled by proximal gradient methods

- replace f with differentiable approximation f_μ (parametrized by μ)
- minimize f_μ by accelerated gradient method

complexity: #iterations for accelerated gradient method depends on L_μ/ϵ_μ

- L_μ is Lipschitz constant of ∇f_μ
- ϵ_μ is accuracy with which the smoothed problem is solved

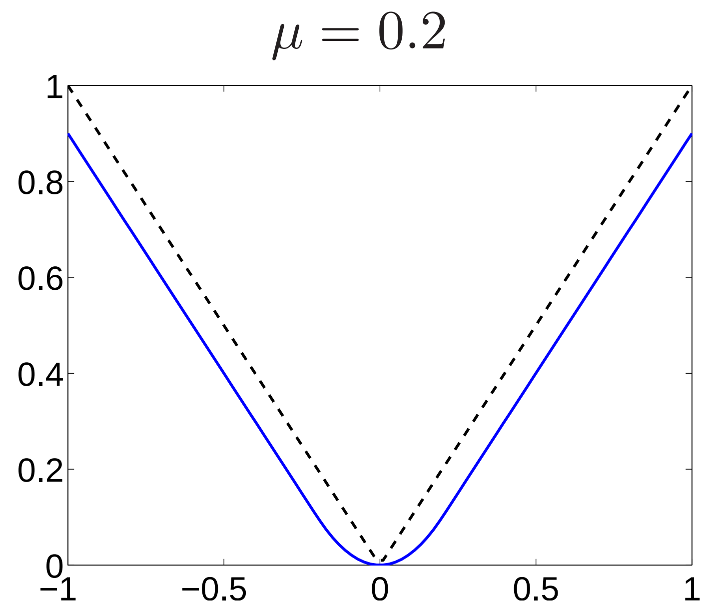
trade-off in amount of smoothing (choice of μ)

- large L_μ (less smoothing) gives more accurate approximation
- small L_μ (more smoothing) gives faster convergence

Example: Huber penalty as smoothed absolute value

$$\phi(z) = |z|$$

$$\phi_\mu(z) = \begin{cases} z^2/(2\mu) & |z| \le \mu \\ |z| - \mu/2 & |z| \ge \mu \end{cases}$$



μ controls accuracy and smoothness

- accuracy

$$|z| - \frac{\mu}{2} \le \phi_\mu(z) \le |z|$$

- smoothness

$$\phi''_\mu(z) \le \frac{1}{\mu}$$

Huber penalty approximation of 1-norm minimization

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

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

$$f(x) - f^\star \leq f_\mu(x) - f_\mu^\star + \frac{m\mu}{2}$$

  to achieve $f(x) - f^\star \leq \epsilon$: need $f_\mu(x) - f_\mu^\star \leq \epsilon_\mu$ with $\epsilon_\mu = \epsilon - m\mu/2$

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

complexity: (more general version later) 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., $O(\sqrt{L_\mu/\epsilon_\mu}\,) = O(1/\epsilon)$ complexity using accelerated 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

$$f(y) \;\geq\; f(x) + \frac{\mu}{2}\|y - x\|_2^2, \qquad \forall\, y \in \mathbf{dom}\, f$$

(μ is strong convexity constant of f, see p. 2-4)

proof: if some y does not satisfy the inequality, then for small positive θ

$$\begin{aligned} f((1-\theta)x + \theta y) &\leq (1-\theta)f(x) + \theta f(y) - \frac{\mu}{2}\theta(1-\theta)\|y - x\|_2^2 \\ &= f(x) + \theta\left(f(y) - f(x) - \frac{\mu}{2}\|y - x\|_2^2\right) + \frac{\mu}{2}\theta^2\|y - x\|_2^2 \\ &< f(x) \end{aligned}$$

contradicts x being minimizer

Conjugate of strongly convex function

suppose f is closed and strongly convex with constant μ

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} \big(y^T x - f(x)\big)$$

- f^* is defined and differentiable at all y, with gradient

$$\nabla f^*(y) = \underset{x}{\operatorname{argmax}} \big(y^T x - f(x)\big)$$

- ∇f^* is Lipschitz continuous with constant $1/\mu$, i.e.,

$$\big\| \nabla f^*(u) - \nabla f^*(v) \big\|_2 \le \frac{1}{\mu} \|u - v\|_2$$

outline of proof

- $y^T x - f(x)$ has a unique maximizer x_y for every y (follows from closedness and strong convexity of $f(x) - y^T x$)

- since $f^*(y)$ is supremum of a family of affine functions, $\nabla f^*(y) = x_y$

- from strong convexity and ($x_u = \nabla f^*(u) \iff u = \nabla f(x_u)$)

$$f(x_u) \geq f(x_v) + v^T(x_u - x_v) + \frac{\mu}{2}\|x_u - x_v\|_2^2$$
$$f(x_v) \geq f(x_u) + u^T(x_v - x_u) + \frac{\mu}{2}\|x_u - x_v\|_2^2$$

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 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
- $C \subseteq \mathbf{dom}\, d$

$d(x)$ measures "distance" of x to the **center** $\quad x_d = \operatorname{argmin}_{x \in C} d(x) \quad$ of C

normalization

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

$$d(x) \geq \frac{1}{2}\|x - x_d\|_2^2, \qquad \forall\, x \in C$$

Common proximity functions

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

Smoothing via conjugate

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

$$\begin{aligned} f(x) &= \sup_{y \in \mathbf{dom}\, h} \left((Ax+b)^T y - h(y) \right) \\ &= h^*(Ax+b) \end{aligned}$$

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

smooth approximation: choose proximity function d for $C = \mathbf{cl}(\mathbf{dom}\, h)$

$$\begin{aligned} f_\mu(x) &= \sup_{y \in \mathbf{dom}\, h} \left((Ax+b)^T y - h(y) - \mu d(y) \right) \\ &= (h+\mu d)^*(Ax+b) \end{aligned}$$

then f_μ is differentiable because $h + \mu d$ is strongly convex

Example: absolute value

conjugate representation

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

proximity function: choosing $d(y) = y^2/2$ gives Huber penalty

$$f_\mu(x) = \sup_{-1 \le y \le 1} \left(xy - \mu y^2/2\right) = \begin{cases} x^2/(2\mu) & |x| \le \mu \\ |x| - \mu/2 & |x| > \mu \end{cases}$$

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

$$f_\mu(x) = \sup_{-1 \le y \le 1} \left(xy - \mu + \mu\sqrt{1 - y^2}\right) = \sqrt{x^2 + \mu^2} - \mu$$

another conjugate representation of $|x|$

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

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

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

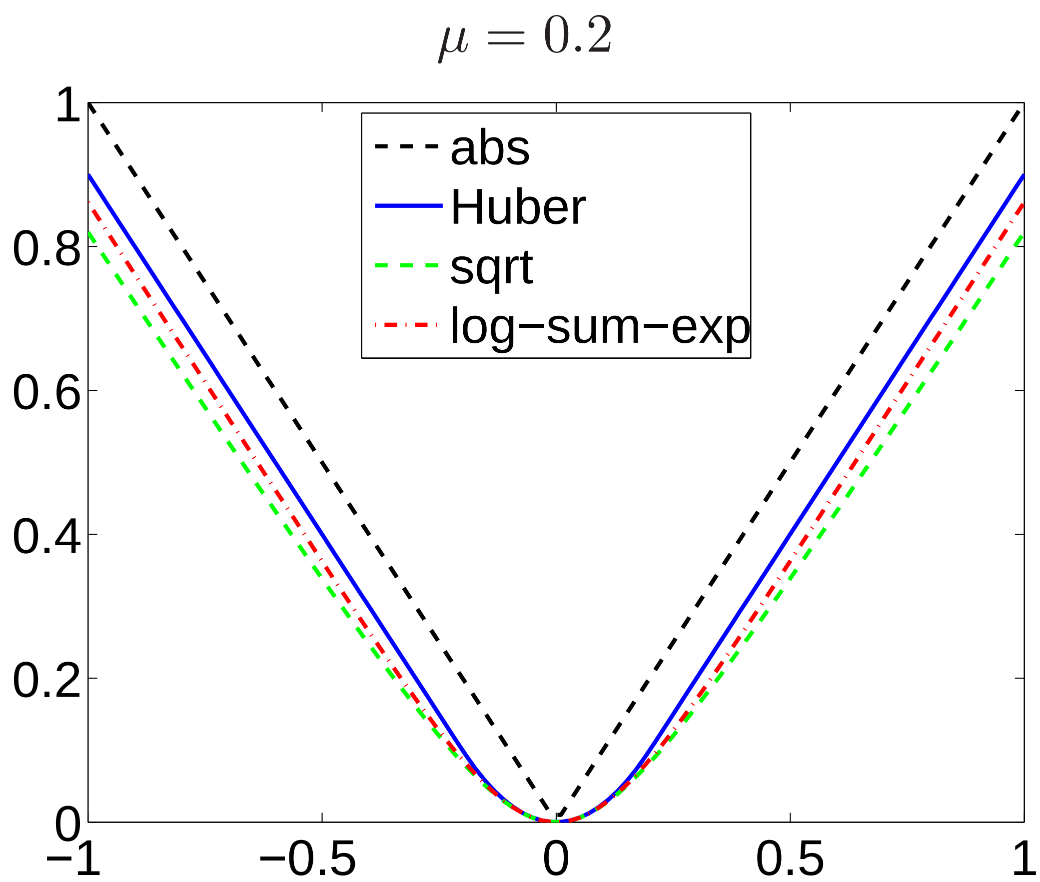
proximity function for C

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

smooth approximation (soft-max approximation for $|x| = \max\{-x, x\}$)

$$\begin{aligned} f_\mu(x) &= \sup_{y_1+y_2=1} \left(xy_1 - xy_2 + \mu(y_1 \log y_1 + y_2 \log y_2 + \log 2)\right) \\ &= \mu \log\left(\frac{e^{x/\mu} + e^{-x/\mu}}{2}\right) \end{aligned}$$

comparison: three smooth approximations of absolute value



Gradient of smooth approximation

$$
\begin{aligned}
f_\mu(x) &= (h + \mu d)^*(Ax + b) \\
&= \sup_{y \in \mathbf{dom}\, h} \big((Ax+b)^T y - h(y) - \mu d(y)\big)
\end{aligned}
$$

from properties of the conjugate strongly convex function (page 8–7)

- f_μ is differentiable, with gradient

$$
\nabla f_\mu(x) = A^T \operatorname*{argmax}_{y \in \mathbf{dom}\, h} \big((Ax+b)^T y - h(y) - \mu d(y)\big)
$$

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

$$
L_\mu = \frac{\|A\|_2^2}{\mu}
$$

Accuracy of smooth approximation

$$f(x) - \mu D \le f_\mu(x) \le f(x), \qquad D = \sup_{y \in \mathbf{dom}\, h} d(y)$$

note $D \le +\infty$ because $\mathbf{dom}\, h$ is bounded and $\mathbf{dom}\, h \subseteq \mathbf{dom}\, d$

- lower bound follows from

$$\begin{aligned} f_\mu(x) &= \sup_{y \in \mathbf{dom}\, h} \left((Ax+b)^T y - h(y) - \mu d(y)\right) \\ &\ge \sup_{y \in \mathbf{dom}\, h} \left((Ax+b)^T y - h(y) - \mu D\right) \\ &= f(x) - \mu D \end{aligned}$$

- upper bound follows from

$$f_\mu(x) \le \sup_{y \in \mathbf{dom}\, h} \left((Ax+b)^T y - h(y)\right) = f(x)$$

Complexity

minimize nondifferentiable function f with accuracy $f(x) - f^\star \leq \epsilon$

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

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

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

iteration 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 $O(\sqrt{L_\mu/\epsilon_\mu}) = O(1/\epsilon)$ iteration bound for fast gradient method
- efficiency in practice can be improved by decreasing μ gradually (homotopy continuation)

Outline

- introduction
- smoothing via conjugate
- **examples**

Piecewise-linear approximation

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

conjugate representation

$$f(x) = \sup_{y \succeq 0,\, \mathbf{1}^T y = 1} (Ax + b)^T y$$

proximity function

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

smooth approximation

$$f_\mu(x) = \mu \log \left(\sum_{i=1}^{m} e^{(a_i^T x + b_i)/\mu} \right) - \mu \log m$$

1-norm approximation

$$f(x) = \|Ax - b\|_1$$

conjugate representation

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

proximity function

$$d(y) = \sum_{i=1}^{m} w_i y_i^2 \qquad \text{(with } w_i \geq 1\text{)}$$

smooth approximation: Huber approximation

$$f_\mu(x) = \sum_{i=1}^{m} \phi_{\mu w_i}(a_i^T x - b_i)$$

Maximum eigenvalue

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

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

proximity function: negative matrix entropy

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

smooth approximation

$$\begin{aligned} f_\mu(X) &= \sup_{Y \succeq 0, \mathbf{tr}\, Y = 1} \big(\mathbf{tr}(XY) - \mu d(Y)\big) \\ &= \mu \log\left(\sum_{i=1}^{n} e^{\lambda_i(X)/\mu}\right) - \mu \log n \end{aligned}$$

Nuclear norm

nuclear norm $f(X) = \|X\|_*$ is sum of singular values of $X \in \mathbf{R}^{m \times n}$

conjugate representation (dual norm)

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

proximity function

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

smooth approximation

$$f_\mu(X) \;=\; \sum_{\|Y\|_2 \leq 1} \big(\mathbf{tr}(X^T Y) - \mu d(Y)\big) \;=\; \sum_i \phi_\mu(\sigma_i(X))$$

the sum of Huber penalties applied to the singular values of X

Lagrange dual function

$$
\begin{array}{ll}
\text{minimize} & f_0(x) \\
\text{subject to} & f_i(x) \le 0, \quad i = 1, \ldots, m \\
& x \in C
\end{array}
$$

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 regularize the primal problem

$$
\begin{array}{ll}
\text{minimize} & f_0(x) + \mu d(x) \\
\text{subject to} & f_i(x) \le 0, \quad i = 1, \ldots, m \\
& x \in C
\end{array}
$$

Smoothing for minimizing strongly convex function

conjugate representation: suppose f can be expressed as

$$f(x) \;=\; \hat{f}(x) + \sup_{y \in \mathbf{dom}\, h} \big((Ax)^T y - h(y)\big) \;=\; \hat{f}(x) + h^*(Ax+b)$$

- $\hat{f}$ is strongly convex with **known** constant $\hat{\mu} > 0$
- $\nabla \hat{f}$ is Lipschitz continuous with constant $\hat{L}$

smooth approximation:

$$f_\mu(x) \;=\; \hat{f}(x) + \sup_{y \in \mathbf{dom}\, h} \big((Ax)^T y - h(y) - \mu d(y)\big) \;=\; \hat{f}(x) + (h+\mu d)^*(Ax+b)$$

- f_μ is strongly convex with constant $\hat{\mu} > 0$
- ∇f_μ is Lipschitz continuous with constant $L_\mu = \hat{L} + \frac{\|A\|_2^2}{\mu}$

Complexity

minimize $f(x) = \hat{f}(x) + h^*(Ax+b)$ with accuracy $f(x) - f^\star \leq \epsilon$

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

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

- f_μ is $\hat{\mu}$-strongly convex and ∇f_μ has Lipschitz constant $L_\mu = L + \frac{\|A\|_2^2}{\mu}$

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

- accelerated gradient method gives iteration complexity (need to know $\hat{\mu}$)

$$O\left(\sqrt{\frac{L_\mu}{\hat{\mu}}}\log\frac{1}{\epsilon_\mu}\right) = O\left(\sqrt{\frac{\hat{L}\epsilon + 2D\|A\|_2^2}{\hat{\mu}\epsilon}}\log\frac{1}{\epsilon}\right) = O\left(\frac{1}{\sqrt{\epsilon}}\log\frac{1}{\epsilon}\right)$$

Sources and References

- this lecture is a modified version of lecture on smoothing from: L. Vandenberghe, *Lecture notes for EE236C - Optimization Methods for Large-Scale Systems* (Spring 2011), UCLA.
- Yu. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming (2005).
- Yu. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM Journal on Optimization (2005)