Smoothing

- introduction
- smoothing via conjugate
- examples

First-order convex optimization methods

complexity of finding ϵ-suboptimal point of f

- subgradient method: f nondifferentiable with Lipschitz constant G

$$O\left((G/\epsilon)^2\right) \text{ iterations}$$

- proximal gradient method: $f = g + h$, h a 'simple' nondifferentiable function, g differentiable with L-Lipschitz continuous gradient

$$O\left(L/\epsilon\right) \text{ iterations}$$

- fast proximal gradient methods (lecture 7)

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

Nondifferentiable optimization by smoothing

for nondifferentiable f that cannot be handled by proximal gradient method

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

complexity: #iterations for (fast) gradient method depends on L_μ/ϵ_μ

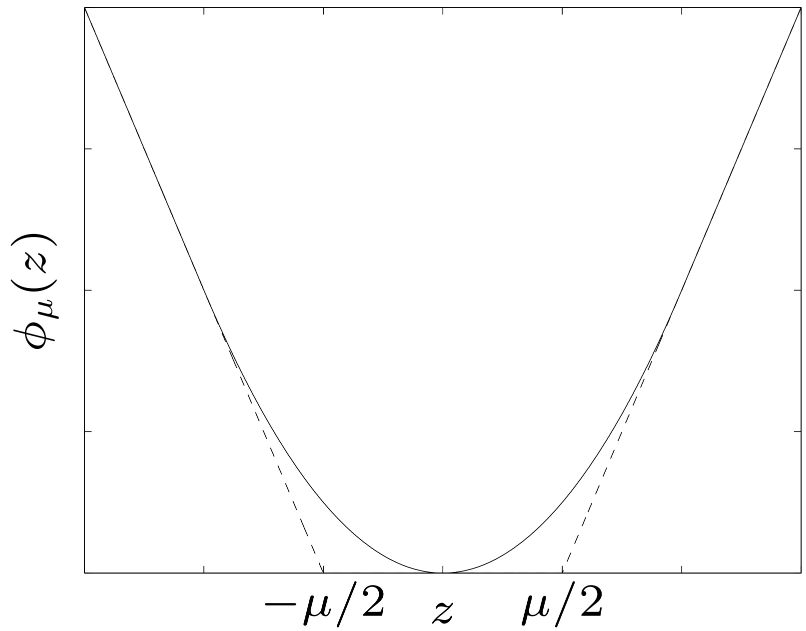
- L_μ is Lipschitz constant of ∇f_μ
- ϵ_μ is accuracy with which the smooth problem is solved

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

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



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

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

complexity: for $\mu = \epsilon/m$

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

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

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

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

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

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

Conjugate of strongly convex function

suppose f is closed and strongly convex with constant μ and conjugate

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

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

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

- ∇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 $x_u = \nabla f^*(u)$, $x_v = \nabla f^*(v)$)

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

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

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

normalization

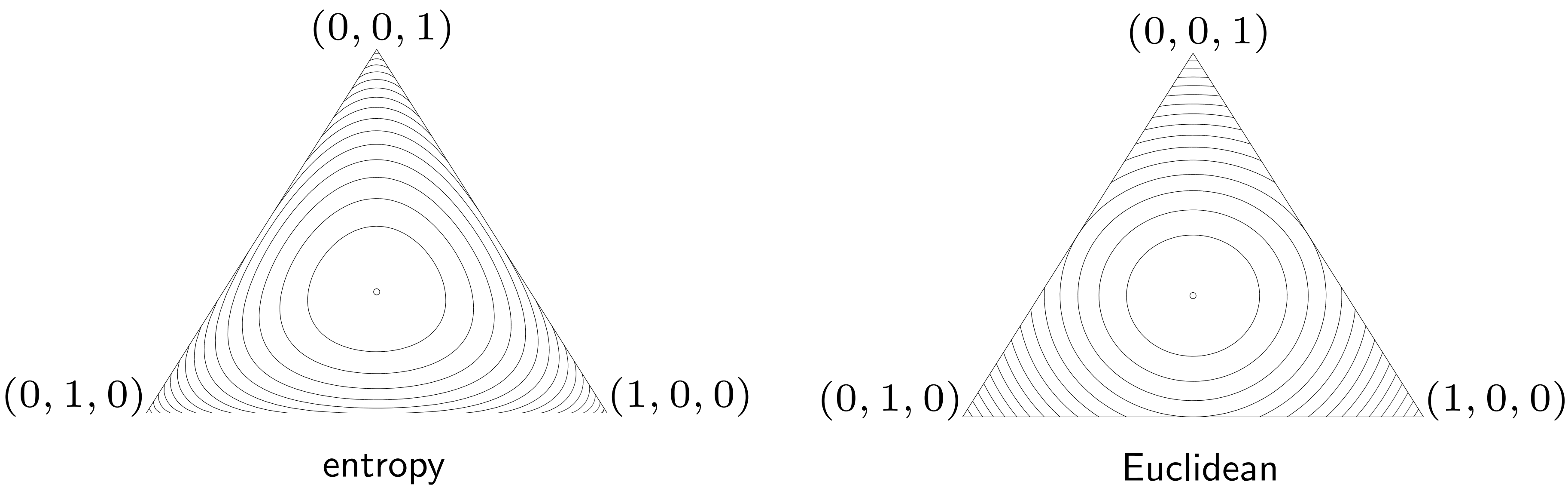
- we will assume the strong convexity constant is 1 and $\inf_{x \in C} d(x) = 0$
- for a normalized proximity function

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

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 $C = \{x \succeq 0 \mid \mathbf{1}^T x = 1\}$, $x_d = (1/n)\mathbf{1}$

example (probability simplex): entropy and $d(x) = (1/2)\|x - (1/n)\mathbf{1}\|_2^2$



Smoothing via conjugate

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

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

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

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

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

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

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

another conjugate representation of $|x|$

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

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

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

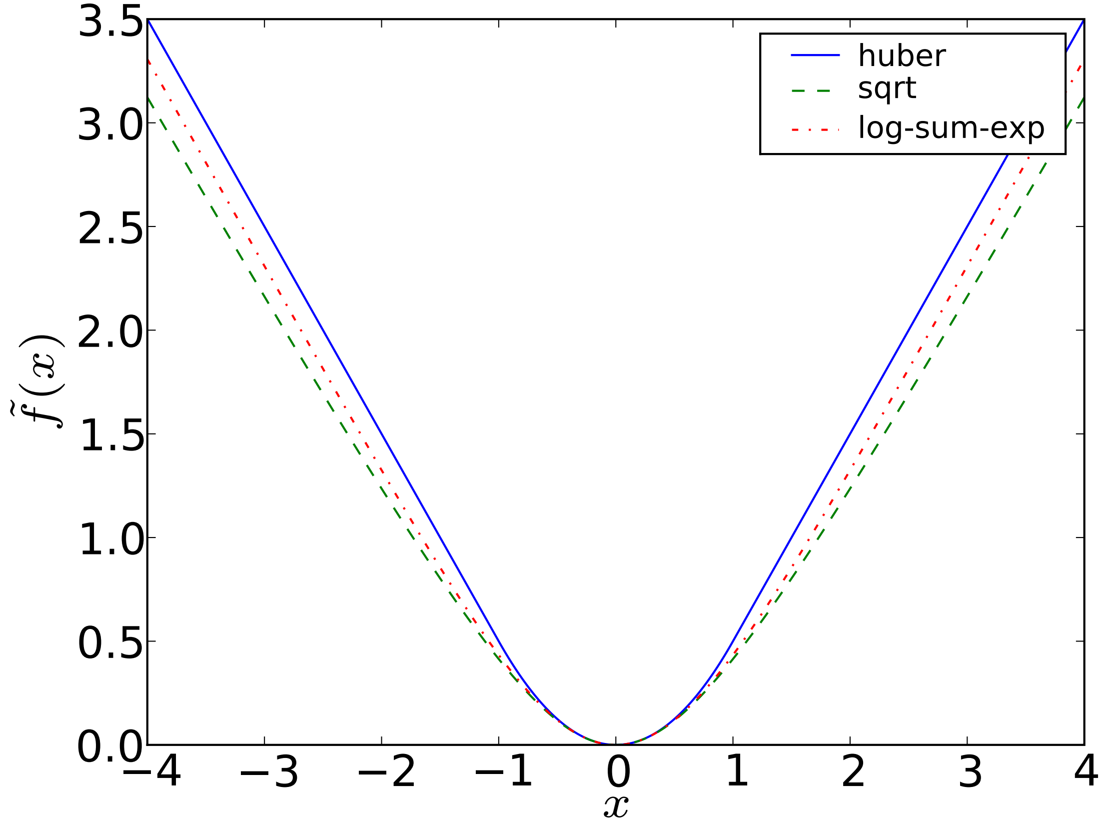
proximity function for C

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

smooth approximation

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

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 of strongly convex function (page 7)

- f_μ is differentiable, with gradient

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

- ∇f_μ is Lipschitz continuous with constant

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

Accuracy of smooth approximation

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

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

- lower bound follows from

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

- upper bound follows from

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

Complexity

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

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

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

- Lipschitz constant of f_μ 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 μ 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

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

proximity function

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

smooth approximation

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

1-Norm approximation

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

conjugate representation

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

proximity function

$$d(y) = \frac{1}{2} \sum_i w_i y_i^2 \qquad \text{(with } w_i > 1\text{)}$$

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 $X \in \mathbf{S}^n$,

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

proximity function: negative matrix entropy

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

smooth approximation

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

Nuclear norm

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

conjugate representation

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

proximity function

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

smooth approximation

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

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

Lagrange dual function

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

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 regularizing the primal problem

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

References

- D. Bertsekas, *Nonlinear Programming* (1995), §5.4.5
- Yu. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming (2005)