

## 6. Smoothing techniques

- motivation
- smoothing via conjugate
- examples

6-1

### First-order convex optimization methods

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

- $f$  differentiable

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

with fast gradient method ( $L$  is Lipschitz constant for  $\nabla f$ )

- $f = g + h$  with  $g$  differentiable,  $h$  a 'simple' nondifferentiable function

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

with fast gradient method ( $L$  is Lipschitz constant for  $\nabla g$ )

- $f$  general nondifferentiable

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

with subgradient method ( $G$  is Lipschitz constant for  $f$ )

# Nondifferentiable optimization by smoothing

make a differentiable approximation  $\tilde{f}$  and minimize by gradient method

**complexity:**  $O(\sqrt{L/\tilde{\epsilon}})$  iterations

- $L$  is Lipschitz constant of  $\nabla \tilde{f}$  (smaller  $L$  means more smoothing)
- $\tilde{\epsilon}$  is accuracy for smooth problem; needs to be smaller than  $\epsilon$  to account for approximation error

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

- large  $L$  gives more accurate approximation
- small  $L$  gives faster convergence, but requires smaller  $\tilde{\epsilon}$

is the overall complexity better than  $O(1/\epsilon^2)$ ?

## Example: 1-norm approximation

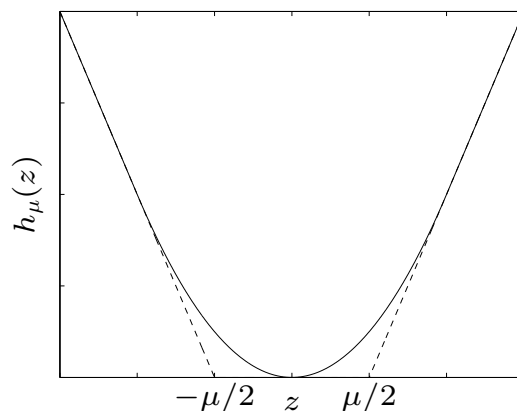
$$\text{minimize } f(x) = \|Ax - b\|_1 = \sum_{i=1}^m |a_i^T x - b_i|$$

**Huber penalty** as smoothed absolute value

$$h_\mu(z) = \begin{cases} z^2/(2\mu) & |z| \leq \mu \\ |z| - \mu/2 & |z| \geq \mu \end{cases}$$

$\mu$  controls accuracy and smoothness

- $h_\mu''(z) \leq 1/\mu$
- $h_\mu(z) \leq |z| \leq h_\mu(z) + \mu/2$



**1-norm approximation by smoothing:** take  $\mu = \epsilon/m$  and solve

$$\text{minimize } \tilde{f}(x) = \sum_{i=1}^m h_{\mu}(a_i^T x - b_i)$$

- $\tilde{f}$  is smooth with  $\nabla^2 \tilde{f}(x) \preceq LI$ ,

$$L = \frac{\lambda_{\max}(A^T A)}{\mu} = \frac{m\|A\|^2}{\epsilon}$$

- if  $\tilde{f}(x) - \tilde{f}^* \leq \epsilon/2$ , then  $f(x) - f^* \leq \tilde{f}(x) + m\mu/2 - \tilde{f}^* \leq \epsilon$

bound on #iterations to reach  $f(x^{(k)}) - f^* \leq \epsilon$  by fast gradient method:

$$O(\sqrt{2L/\epsilon}) = O\left(\frac{\sqrt{2m}\|A\|}{\epsilon}\right)$$

cf. the  $O(1/\epsilon^2)$  bound for subgradient method

## Strongly convex functions

$f$  is strongly convex with parameter  $\mu > 0$  if  $\text{dom } f$  is convex and

$$(1 - \theta)f(x) + \theta f(y) \geq f((1 - \theta)x + \theta y) + \mu \frac{\theta(1 - \theta)}{2} \|x - y\|_2^2$$

for all  $x, y \in \text{dom } f$ , all  $\theta \in [0, 1]$

- for differentiable functions the inequality is equivalent to

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

- for twice differentiable functions the inequality is equivalent to

$$\nabla^2 f(x) \succeq \mu I \quad \forall x \in \text{dom } f$$

## Minimum of strongly convex function

if  $x$  is a minimizer of a strongly convex function  $f$ , then it is unique and

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

**proof** (by contradiction): if  $y \neq x$  satisfies

$$f(y) < f(x) + \frac{\mu}{2} \|y - x\|_2^2,$$

then for small positive  $\theta$ ,

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

## Conjugate of strongly convex function

recall definition of conjugate of  $f$

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

**properties:** for  $f$  strongly convex, continuous, with  $\text{dom } f$  closed

- $f^*$  is defined and differentiable at all  $x$

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

- $\nabla f^*$  is Lipschitz continuous with constant  $1/\mu$

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

## proof of second property

define  $y_u = \nabla f^*(u)$ ,  $y_v = \nabla f^*(v)$

from strong convexity of  $f(y) - x^T y$  and inequality on page 6–7

$$f(y) - u^T y \geq f(y_u) - u^T y_u + \frac{\mu}{2} \|y - y_u\|_2^2$$

$$f(y) - v^T y \geq f(y_v) - v^T y_v + \frac{\mu}{2} \|y - y_v\|_2^2$$

evaluate first inequality at  $y = y_v$  and second at  $y = y_u$ , and add to get

$$(u - v)^T (y_u - y_v) \geq \mu \|y_u - y_v\|_2^2$$

from the Cauchy-Schwarz inequality, this implies

$$\|y_u - y_v\|_2 \leq \frac{1}{\mu} \|u - v\|_2$$

## Proximity functions

### definition

$d$  is a proximity function (prox-function) for closed bounded convex set  $C$  if

- $d$  is continuous and strongly convex (with constant  $\mu > 0$ )
- $C \subseteq \text{dom } d$

we will assume  $d$  is normalized so that  $\mu = 1$  and  $\inf_{x \in C} d(x) = 0$

**prox-center:**  $x_d = \operatorname{argmin}_{x \in C} d(x)$

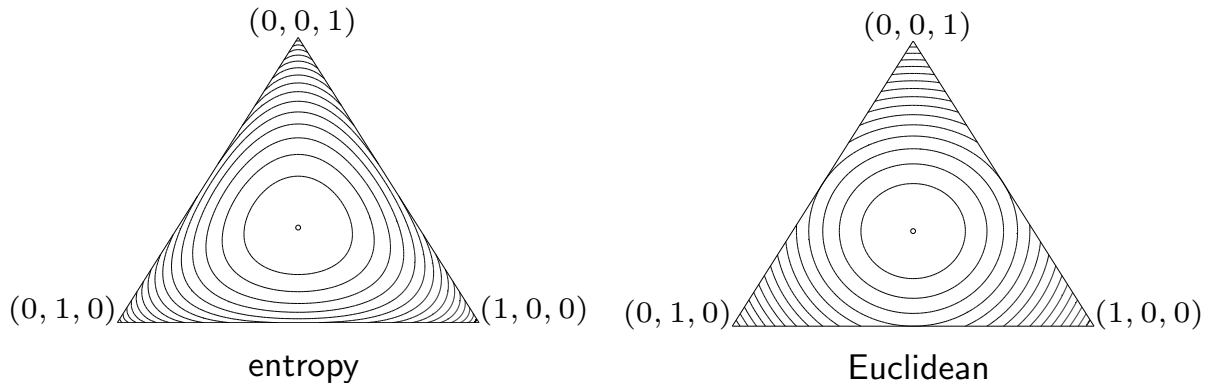
for normalized prox-functions,

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

## common proximity functions

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

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



contour lines of entropy function give a better fit to the set

## Smoothing via conjugate

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

$$f(x) = g^*(Ax) = \sup_{y \in \text{dom } g} (x^T A^T y - g(y))$$

$g$  convex and continuous, with  $\text{dom } g$  closed and bounded

**smooth approximation:**  $\tilde{f} = (g + \mu d)^*(Ax)$

$$\tilde{f}(x) = \sup_{y \in \text{dom } g} (x^T A^T y - g(y) - \mu d(y))$$

where  $d$  is a prox-function for  $\text{dom } g$

## Example: absolute value

conjugate representations of  $f(x) = |x|$ :

$$|x| = \sup_{-1 \leq y \leq 1} xy, \quad |x| = \sup_{\substack{u+v=1 \\ u \geq 0, v \geq 0}} x(u-v)$$

- first representation, prox-function  $d(y) = y^2/2$

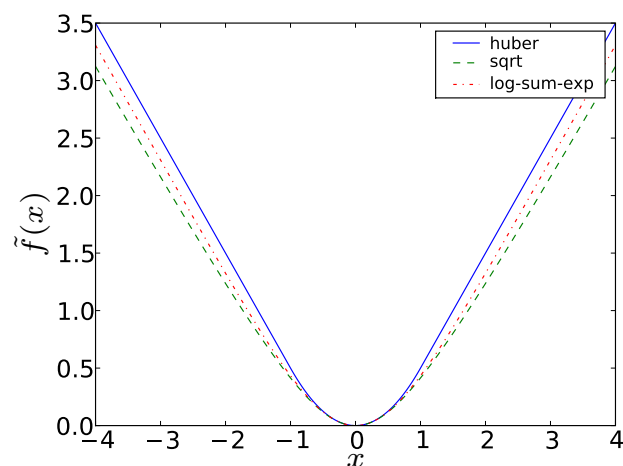
$$\tilde{f}(x) = \sup_{-1 \leq y \leq 1} (xy - \mu y^2/2) = h_\mu(x) = \begin{cases} x^2/(2\mu) & |x| \leq \mu \\ |x| - \mu/2 & |x| > \mu \end{cases}$$

- first representation, prox-function  $d(y) = 1 - \sqrt{1 - y^2}$

$$\tilde{f}(x) = \sup_{-1 \leq y \leq 1} (xy + \mu\sqrt{1 - y^2} - \mu) = \sqrt{x^2 + \mu^2} - \mu$$

- second representation, prox-function  $d(u, v) = u \log u + v \log v + \log 2$

$$\begin{aligned} \tilde{f}(x) &= \sup_{u+v=1} (xu - xv + \mu(u \log u + v \log v + \log 2)) \\ &= \mu \log\left(\frac{e^{x/\mu} + e^{-x/\mu}}{2}\right) \end{aligned}$$



## Properties of smoothed conjugate

- $\tilde{f}$  is differentiable, with gradient

$$\nabla \tilde{f}(x) = A^T \operatorname{argmax}_{y \in \operatorname{dom} g} (x^T A^T y - g(y) - \mu d(y))$$

- $\nabla \tilde{f}$  is Lipschitz continuous with constant  $\|A\|_2^2/\mu$
- since  $\operatorname{dom} g$  is bounded,  $D = \sup_{y \in \operatorname{dom} g} d(y)$  is finite, and

$$\tilde{f}(x) \leq f(x) \leq \tilde{f}(x) + \mu D$$

first two properties follow from the fact that

$$g(y) + \mu d(y) - x^T A^T y$$

is a strongly convex function of  $y$ , with constant  $\mu$

## Complexity

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

- solve smoothed problem with accuracy  $\tilde{\epsilon} = \epsilon - \mu D$ , so that

$$f(x) - f^* \leq \tilde{f}(x) + \mu D - \tilde{f}^* \leq \tilde{\epsilon} + \mu D = \epsilon$$

- number of iterations bounded by

$$O(\sqrt{L/\tilde{\epsilon}}) = O\left(\frac{\|A\|_2}{\sqrt{\mu(\epsilon - \mu D)}}\right)$$

bound is minimal for  $\mu = \epsilon/(2D)$ , and equal to

$$O\left(\frac{2\|A\|_2 D^{1/2}}{\epsilon}\right)$$

## Piecewise-linear approximation

$$\text{minimize } f(x) = \max_{i=1,\dots,m} (a_i^T x - b_i)$$

- conjugate representation

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

- take  $d(y) = \sum_{i=1}^m y_i \log y_i + \log m$  as prox-function

$$\begin{aligned} \tilde{f}(x) &= \sup_{y \succeq 0, \mathbf{1}^T y = 1} ((Ax - b)^T y - \mu d(y)) \\ &= \mu \log \sum_{i=1}^m e^{(a_i^T x - b_i)/\mu} - \mu \log m \end{aligned}$$

- accuracy:  $\tilde{f}(x) \leq f(x) \leq \tilde{f}(x) + \mu \log m$

## 1-norm approximation

$$\text{minimize } f(x) = \|Ax - b\|_1 = \sum_{i=1}^m |a_i^T x - b_i|$$

- conjugate representation

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

- take  $d(y) = (1/2) \sum_i w_i y_i^2$  (with  $w_i > 1$ ) as prox-function

$$\begin{aligned} \tilde{f}(x) &= \sup_{\|y\|_\infty \leq 1} (Ax - b)^T y - \mu d(y) \\ &= \sum_{i=1}^n h_{\mu w_i}(a_i^T x - b_i) \end{aligned}$$

$h_{\mu w_i}(z)$  is the Huber penalty function with parameter  $\mu w_i$

## Maximum eigenvalue

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

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

- choose negative matrix entropy

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

as prox-function

$$\tilde{f}(X) = \sup_{Y \succeq 0, \text{tr} Y = 1} (\text{tr}(XY) - \mu d(Y)) = \mu \log\left(\frac{1}{n} \sum_{i=1}^n e^{\lambda_i(X)/\mu}\right)$$

## Nuclear norm

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

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

- conjugate representation

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

- choose  $d(Y) = (1/2)\|Y\|_F^2$  as prox-function

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

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

## Lagrange dual function

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

$f_i$  convex,  $C$  closed and bounded

### smooth approximation of dual function

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

$d$  is a prox-function for  $C$

this is equivalent to regularizing the primal problem

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

## Reference

Yu. Nesterov, *Smooth minimization of non-smooth functions*,  
Mathematical Programming (2005)