Agenda

Nesterov’s method for the minimization of nonsmooth functions

1. Smoothing
2. Conjugate functions
3. Properties of conjugate functions
4. Smoothing by conjugation
5. Nesterov 2005 algorithm
6. Examples
Nonsmooth optimization

minimize \( f(x) \)
subject to \( x \in C \) (can be \( \mathbb{R}^n \))

- \( f \) cvx, \( \text{dom}(f) = \mathbb{R}^n \)
- \( f \) not differentiable
Examples

minimize $\|x\|_1$
subject to $Ax = b$

minimize $\|x\|_{TV}$
subject to $\|x - b\|_2 \leq \delta$

minimize $\|Ax - b\|_1$

Objectives are not smooth
Motivation

- $f$ convex with Lipschitz gradient,
  
  $O(\sqrt{L/\epsilon})$

  iterations for $\epsilon$ approximation

- $f$ convex and Lipschitz with constant $G$,
  
  $O(G^2/\epsilon^2)$

  iterations via method of subgradients for $\epsilon$ approximation

This lecture: speed up convergence for nonsmooth functions

- Build smooth approximation $f_\mu$ to objective functional
- Minimize smooth approximation
Example: smoothed $\ell_1$ norm

Huber function

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

$$\nabla h_\mu(t) = \begin{cases} 
\frac{t}{\mu} & |t| \leq \mu \\
\text{sgn}(t) & |t| \geq \mu
\end{cases}$$

- $h_\mu$ approximates $|t|$
  $$h_\mu(t) \leq |t| \leq h_\mu(t) + \frac{\mu}{2}$$

- $h_\mu'$ is Lipschitz with constant $1/\mu$

Smooth approximation of $\ell_1$ norm: $f_\mu(x) = \sum_i h_\mu(x_i)$

$$f_\mu(x) \leq \|x\|_{\ell_1} \leq f_\mu(x) + \frac{\mu n}{2}$$
Basic idea

Get an $\epsilon/2$ approximation of

$$\begin{align*}
\text{minimize} & \quad f_\mu(x) \\
\text{subject to} & \quad x \in C
\end{align*}$$

with

$$O\left(\sqrt{\frac{1}{\mu \epsilon}}\right)$$

iterations. Then

$$f(x_k) - f^* \leq f_\mu(x_k) + \frac{\mu n}{2} - f_\mu(x^*) \leq \epsilon/2 + \frac{\mu n}{2}$$

Set $\mu n = \epsilon$ and cost is

$$O\left(\frac{\sqrt{n}}{\epsilon}\right)$$

at most
References


- L. Vandenberghe. Lecture Notes for EE 236C, UCLA

Strongly convex functions

$f$ is strongly convex with parameter $\mu$ if for all $x, y$ and all $\lambda \in [0, 1]$

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

- For differentiable functions, equivalent to

\[
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \quad \forall x, y
\]

- For $C^2$ functions, equivalent to

\[
\nabla^2 f(x) \succeq \mu I \quad \forall x
\]

A strongly convex function has a unique minimizer
Conjugate of strongly convex functions

**Definition**

Conjugate of cvx $f$

$$f^*(x) = \sup_{u \in \text{dom}(f)} \langle u, x \rangle - f(u)$$

**Lemma (Properties of conjugate of strongly convex functions)**

Assume $\text{dom}(f)$ cvx and closed

- $f^*$ is well defined and differentiable:

$$\nabla f^*(x) = u^* = \arg \max \langle u, x \rangle - f(u)$$

- $\nabla f^*$ is Lipshitz and obeys

$$\| \nabla f^*(x) - \nabla f^*(y) \|_2 \leq \mu^{-1} \| x - y \|_2$$
Proof. First,

$$f^*(x + h) = \sup_u \{ \langle u, x + h \rangle - f(u) \} \geq f^*(x) + \langle u^*, h \rangle$$

Second, we claim that because of strong convexity, $h(u) = f(u) - \langle u, x \rangle$ obeys

$$h(u) \geq h(u^*) + \frac{\mu}{2} \| u - u^* \|^2$$

(Easy to check if $f$ is differentiable.) Therefore,

$$f^*(x + h) = \sup_u \{ \langle u, x + h \rangle - f(u) \} \leq f^*(x) + \sup_u \{ \langle u, h \rangle - \frac{\mu}{2} \| u - u^* \|^2 \}$$

$$= f^*(x) + \langle u^*, h \rangle + \sup_v \{ \langle v, h \rangle - \frac{\mu}{2} \| v \|^2 \}$$

$$= f^*(x) + \langle u^*, h \rangle + \frac{1}{2\mu} \| h \|^2$$

Conclusion:

- $f^*$ is differentiable and $\nabla f^*(x) = u^*$
- Lipschitz constant is at most $1/\mu$
For second claim, assume $f$ differentiable. Optimality conditions

\[ \nabla f(u) = x, \quad \nabla f(v) = y \]

Then

\[
\begin{align*}
    f(v) &\geq f(u) + \langle x, v - u \rangle + \frac{\mu}{2} \|u - v\|^2 \\
    f(u) &\geq f(v) + \langle y, u - v \rangle + \frac{\mu}{2} \|u - v\|^2
\end{align*}
\]

This gives

\[
\mu \|u - v\|^2 \leq \langle x - y, u - v \rangle \leq \|x - y\| \|u - v\|
\]

which is the claim.

One can use a density argument in the case $f$ not differentiable.
Proximity function

Definition

\( d \) is a prox function for closed cvx set \( C \) if

- \( d \) is continuous on \( C \)
- \( d \) is strongly cvx on \( C \)

Assume \( d \) normalized, \( \mu = 1 \) and \( \inf_C d(x) = 0 \), so that

\[
\begin{align*}
d(x) &\geq \frac{1}{2} \| x - x_c \|^2
\end{align*}
\]

\( x_c \) is prox-center
Examples of prox functions

- \( d(x) = \frac{1}{2} \| x - x_c \|_2^2, \ x_c \in C \)
- \( d(x) = \frac{1}{2} \sum_i w_i (x_i - x_{c,i})^2, \ w_i \geq 1 \text{ and } x_c \in C \)
- \( d(x) = \sum_i x_i \log x_i + \log n, \ C = \{ x : x \geq 0 \text{ and } \sum_i x_i = 1 \} \)
- \( d(X) = \frac{1}{2} \| X - X_c \|_F^2, \ X_c \in C \)
Smoothing by conjugation

Suppose that $f$ can be written as

$$f(x) = \sup_{u \in \text{dom}(g)} \langle u, x \rangle - g(u) = g^*(x)$$

Smooth approximation

$$f_\mu(x) = \sup_{u \in \text{dom}(g)} \langle u, x \rangle - (g(u) + \mu d(u)) = (g + \mu d)^*(x)$$

- $\nabla f_\mu$ is Lipschitz with constant at most $\mu^{-1}$
- if $\text{dom}(g)$ bdd and $D = \sup_{u \in \text{dom}(g)} d(u)$

$$f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu D$$
Example: \( f(x) = |x| \)

- Express \( f(x) = |x| \) as (conjugate representation)

\[
|x| = \sup_{|u| \leq 1} ux
\]

- Smooth via

\[
f_\mu(x) = \sup_{|u| \leq 1} \{ux - \frac{\mu}{2}u^2\} = h_\mu(x) = \begin{cases} 
\frac{x^2}{2\mu} & |x| \leq \mu \\
|x| - \frac{\mu}{2} & |x| \geq \mu 
\end{cases}
\]

This is Huber’s function
- Many other choices of proximity functions
Express $f(x) = |x|$ as (conjugate representation)

$$|x| = \sup(u_1 - u_2)x \quad u_1, u_2 \geq 0, \quad u_1 + u_2 = 1$$

Prox

$$d(u) = u_1 \log u_1 + u_2 \log u_2 + \log 2$$

Smooth approximation

$$h_\mu(x) = \mu \log[\cosh(x/\mu)]$$
Smoothing norms

Pair of primal/dual norms \( f(x) = \|x\| = \sup_{\|u\|_\star \leq 1} \langle u, x \rangle \)

\[
f_\mu(x) = \sup_{\|u\|_{\star \leq 1}} \langle u, x \rangle - \mu d(u)
\]

With \( \ell_1 \) norm and \( d(u) = \frac{1}{2} \|u\|^2_2 \)

\[
f_\mu(x) = \sup_{\|u\|_\infty \leq 1} \langle u, x \rangle - \mu d(u) = \sum_i h_\mu(x_i)
\]

Suppose \( f(x) = \sum_i |x_i - x_{i-1}| = \|Dx\|_1 \), then

\[
f(x) = \sup_{\|u\|_\infty \leq 1} \langle D^* u, x \rangle
\]

and smooth approximation

\[
f_\mu(x) = \sup_{\|u\|_\infty \leq 1} \langle D^* u, x \rangle - \mu d(u)
\]
Complexity analysis

Want to find solution to nonsmooth problem with accuracy $\epsilon$

1. Construct smooth approximation such that

$$f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu D$$

which gives

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

($\nabla f_\mu$ is Lipschitz constant at most $\mu^{-1}$)

2. Choose $\mu$ s.t. $\mu D = \epsilon/2$

3. Minimize $f_\mu$ with accuracy $\epsilon/2$

Solution is $\epsilon$ accurate and number of iterations is

$$O\left(\sqrt{\frac{1}{\mu\epsilon}}\right) = O\left(\frac{\sqrt{D}}{\epsilon}\right)$$

This is much better than $O(1/\epsilon^2)$
Example 1: Chebyshev approximation

\begin{align*}
\text{minimize} \quad & \|Ax - b\|_\infty \\
A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^m
\end{align*}

- Conjugate representation

\[ f(x) = \sup_{(u,v) \in Q} \langle u - v, Ax - b \rangle \]

\[ Q = \{(u, v) : u \geq 0, v \geq 0, 1^T u + 1^T v = 1\} \]

- Prox

\[ d(u, v) = \sum_i u_i \log u_i + \sum_i v_i \log v_i + \log 2m \]

- Smooth approximation: \( f_\mu(x) = \sup_{(u,v) \in Q} \{\langle u - v, Ax - b \rangle - \mu d(u, v)\} \)

\[ f_\mu(x) = \mu \sum_{i=1}^m \log \left[ \cosh \left( \frac{a_i^T x - b_i}{\mu} \right) \right] \]

- Accuracy

\[ f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu \log 2m \]

Efficient Chebyshev approximation
Example 2: Robust regression

minimize $\|Ax - b\|_1$

$A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$

- Conjugate representation

$$f(x) = \sup_{\|u\|_\infty \leq 1} \langle u, Ax - b \rangle$$

- Prox

$$d(u) = \frac{1}{2} \|u\|_2^2$$

- Smooth approximation

$$f_\mu(x) = \sup_{\|u\|_\infty \leq 1} \{ \langle u, Ax - b \rangle - \mu \|u\|_2^2 \} = \sum_{i=1}^{m} h_\mu(a_i^T x - b_i)$$

$h_\mu$ is the Huber penalty function
Example 3: nuclear-norm minimization

\[
\begin{align*}
\text{minimize} & \quad \|X\|_* \\
\text{subject to} & \quad A(X) = b
\end{align*}
\]

- Conjugate representation
  \[
  \|X\|_* = \sum_i \sigma_i(X) = \sup_{\|U\| \leq 1} \langle U, X \rangle
  \]

- Prox
  \[
  d(U) = \frac{1}{2} \|U\|_F^2
  \]

- Smooth approximation
  \[
  f_\mu(x) = \sup_{\|U\| \leq 1} \left\{ \langle U, X \rangle - \frac{\mu}{2} \|U\|_F^2 \right\} = \sum_{i=1}^{m} h_\mu(\sigma_i(X))
  \]

\(h_\mu\) is the Huber penalty function
Nesterov’s 2005 algorithm

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in C
\end{align*}
\]

\(f\) cvx and \(\nabla f\) Lipshitz with constant \(L\)

- Choose \(x_0\) and prox function for \(C (\mu = 1)\)
- For \(k = 0, 1, 2, \ldots\) (sequence of interest is \(y_k\))
  1. \(y_k = \arg \min_{x \in C} \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2\)
  2. \(z_k = \arg \min_{x \in C} \sum_{i=0}^{k} \frac{i+1}{2} \langle \nabla f(x_i), x - x_i \rangle + \frac{L}{2} d(x)\)
  3. \(x_{k+1} = \theta_k z_k + (1 - \theta_k)y_k, \ \theta_k = 2/(k + 3)\)
Theorem (Nesterov 2005)

\[ f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k + 1)(k + 2)} \]

- Same convergence as before
- If \( f \) not smooth, apply algorithm to \( f_\mu \)
Case study: total-variation denoising

Model

\[ b = I + z \]

- \( I \) is an \( n \times n \) image
- \( z \) noise
- \( b \) is the observed noisy image

Recovery via TV minimization (\( \delta \) is a bound on the noise level)

\[
\begin{align*}
\text{minimize} & \quad \|x\|_{TV} \\
\text{subject to} & \quad \|x - b\|_2 \leq \delta
\end{align*}
\]
Conjugate representation

\[ f(x) = \sup \{ \langle u, Dx \rangle : \|u_{ij}\|_2 \leq 1 \} \]

Prox \( d(u) = \frac{1}{2} \|u\|_2^2 \)

Smooth approximation

\[ f_\mu(x) = \sup \{ \langle u, Dx \rangle - \frac{\mu}{2} \|u\|_2^2 : \|u_{ij}\|_2 \leq 1 \} \]

\nabla f_\mu \text{ Lipshitz with constant at most } \mu^{-1}\|D\|^2 \approx 8\mu^{-1}

Approximation

\[ f_\mu(x) \leq f(x) \leq f_\mu(x) + \frac{\mu n^2}{2} \]
Nesterov’s method

\[
\begin{align*}
& \text{minimize} \quad \|x\|_{\text{TV}} \\
& \text{subject to} \quad \|x - b\|_2 \leq \delta
\end{align*}
\]

Choose \(x_0\) and for \(k = 0, 1, 2, \ldots\) (sequence of interest is \(y_k\))

1. Compute \(\nabla f_\mu(x_k) = D^T u_k\)

\[
\begin{align*}
\quad & u_k = \arg \max \left\{ u^T D x - \frac{\mu}{2} \|u\|_2^2 : \|u_{ij}\|_2 \leq 1 \right\} \\
\end{align*}
\]

2. \(y_k = \arg \min \left\{ \langle \nabla f_\mu(x_k), x - x_k \rangle + \frac{L\mu}{2} \|x - x_k\|^2 : \|x - b\|_2 \leq \delta \right\} \)

3. \(z_k = \arg \min \left\{ \sum_{i=0}^{k} \frac{i+1}{2} \nabla f_\mu(x_i), x - x_i \rangle + \frac{L\mu}{2} \|x - b\|^2 : \|x - b\|_2 \leq \delta \right\} \)

4. update \(x_{k+1} = \theta_k z_k + (1 - \theta_k) y_k, \theta_k = 2 / (k + 3)\)

After change of variables, each step is of the form

\[
\begin{align*}
& \text{minimize} \quad \frac{1}{2} \|x\|_2^2 - c^T x \\
& \text{subject to} \quad \|x\|_2 \leq t
\end{align*}
\]

with solution given by

\[
\min (1, t / \|c\|_2) c
\]
Possible stopping criterion

Stop when PD gap less than a tol.

Dual problem

minimize \[ -\delta \| D^T u \|_2 + \langle b, D^T u \rangle \]
subject to \[ \| u_{ij} \|_2 \leq 1 \]

Duality gap

\[ \| x \|_{TV} + \delta \| D^T u \|_2 - \langle b, D^T u \rangle \]