Math 273a: Optimization
Convex Conjugacy

Instructor: Wotao Yin
Department of Mathematics, UCLA
Fall 2015

online discussions on piazza.com
Convex conjugate (the Legendre transform)

Let $f$ be a closed proper convex function.

The convex conjugate of $f$ is

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

- $f^*$ is convex (in $y$).

Reason: for each fixed $x$, $(y^T x - f(x))$ is linear in $y$. Hence, $f^*$ is point-wise maximum of linear functions, that is, point $y$ and over $x$.

- As long as $f$ is proper, $f^*$ is proper closed convex.
Geometry

- For fixed $y$ and $z$, consider linear function: $g(x) = y^T x - z$
  - the corresponding hyperplane is
    $$\mathcal{H} = \{(x, g(x)) : x \in \mathbb{R}^n\} \subset \mathbb{R}^{n+1}$$
  - $[y; -1]$ is the *downward normal direction* of $\mathcal{H}$
  - $\mathcal{H}$ crosses the $(n+1)$th axis at $g(0) = -z$
  - $y$ (tilt) and $z$ (height) uniquely define the hyperplane $\mathcal{H}$ and function $g$

**Our task**: for each $y$, determine $z$ using the function $f$, thus determining $g$
• If we set **two rules**

1. \( g(x) = f(x) \) at some point \( x \), i.e., \( \mathcal{H} \) intersects \( \text{epi}(f) \),
2. \( \mathcal{H} \) is *as low as possible*, i.e., \(-z \) is *as small as possible*,

then \( \mathcal{H} \) will be the **supporting hyperplane** of \( f \)

By rule #1, \( \exists \) point \( x \in \text{dom} f \), \( \exists \ y^T x - z = f(x) \) or \( -z = f(x) - y^T x \)

By rule #2, \(-z = \inf_{\text{dom} f} \{ f(x) - y^T x \} \implies z = \sup_{x \in \text{dom} f} \{ y^T x - f(x) \} \)

Therefore,

- \( z = f^*(x) \)
- \( g(y) = y^T x - f^*(x) \)
- \( \mathcal{H} \) is the supporting hyperplane
Geometry

1. $\mathcal{H}$ intersects $\text{epi}(f)$

2. $\mathcal{H}$ is as low as possible

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

$f^*$ almost completely characterizes $f$. “Almost” is because covers only up to closure. This is a result of the Hahn-Banach Separation Theorem.
Relation to Lagrange duality

Consider convex problem

$$\min_x f(x) \quad \text{subject to } Ax = b.$$  

Lagrangian:

$$\mathcal{L}(x; y) = f(x) - y^T(Ax - b).$$

Lagrange dual function:

$$d(y) = - \inf_{x \in \text{dom } f} \mathcal{L}(x; y) = \sup_{x \in \text{dom } f} \{y^T(Ax - b) - f(x)\} = f^*(A^T y) - b^T y.$$  

Lagrange dual problem (given in terms of convex conjugate $f^*$):

$$\min_y f^*(A^T y) - b^T y \quad \text{or} \quad \max_y b^T y - f^*(A^T y).$$
Exercise

Derive a Lagrange dual problem for

\[
\min_{x \in \mathbb{R}^n} \quad f(x) + g(Ax)
\]
Primal and dual subdifferentials

Suppose that \([y; -1]\) is the *downward normal* of the hyperplane touching \(\text{epi}(f)\) at \(x\); therefore,

\[ y = \nabla f(x^*) \]

In general, for proper closed convex \(f\),

\[ y \in \partial f(x^*) \]

Therefore,

\[ \text{dom} f^* = \{ \partial f(x) : x \in \text{dom} f \} \]

**Theorem (biconjugation)**

*Let \(x \in \text{dom} f\) and \(y \in \text{dom} f^*\). Then,*

\[ y \in \partial f(x) \iff x \in \partial f^*(y) \]

*If the relation holds, then*

\[ f(x) + f^*(y) = y^T x. \]

The result is very useful in deriving optimality conditions.
Fenchel’s inequality

**Theorem (Fenchel’s inequality)**

For arbitrary \( x \in \text{dom} f \) and \( y \in \text{dom} f^* \), we have

\[
f(x) + f^*(y) \geq y^T x.
\]

**Proof.**

Since \( x \) is not necessarily the maximizing point for \( f(y) = \sup_x \{\cdots\} \), we have

\[
f^*(y) \geq y^T x - f(x).
\]
Theorem

If \( f \) is proper, closed, convex, then \((f^*)^* = f\), i.e.,

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

Proof.

Consider linear function \( g_{y,z} \), defined as \( g_{y,z}(x) = y^T x - z \).

Step 1.

\[
g_{y,z} \leq f \iff y^T x - z \leq f(x), \ \forall x
\]

\[
\iff y^T x - f(x) \leq z, \ \forall x
\]

\[
\iff \sup_{x} \{y^T x - f(x)\} \leq z
\]

\[
\iff f^*(y) \leq z
\]

\[
\iff (y, z) \in \text{epi}(f^*)
\]

(cont.)
Proof.

**Step 2.** From the Hahn-Banach Separation Theorem,

\[ f(x) = \sup_{y,z} \{ g_{y,z}(x) : g_{y,z} \leq f \}, \quad \forall x \in \text{dom} f. \]

**Step 3.**

\[
\sup_{y,z} \{ g_{y,z}(x) : g_{y,z} \leq f \} = \sup_{y,z} \{ g_{y,z}(x) : f^*(y) \leq z \} \text{ by Step 1}
\]

\[
= \sup_{y,z} \{ y^T x - z : f^*(y) \leq z \}
\]

\[
= \sup_y \{ y^T x - f^*(y) \}
\]

\[
= (f^*)^*(x)
\]

Combining Steps 2 and 3, we get \( f = (f^*)^*. \)
Examples

- \( f(x) = \nu_C(x) \), *indicator function* of nonempty closed convex set \( C \); then

\[ \sigma_C^* := f^*(y) = \sup_{x \in C} y^T x \]

is the *support function* of \( C \)

- Applying the theorem, we get \((\nu_C^*)^* = \nu_C\) and \((\sigma_C^*)^* = \sigma_C\).
Examples

- $f(x) = \iota_{\{-1 \leq x \leq 1\}}$, indicator function of the unit hypercube; then
  \[
  f^*(y) = \sup_{-1 \leq x \leq 1} y^T x = \|y\|_1
  \]

- $f(x) = \iota_{\{\|x\|_2 \leq 1\}}$, then
  \[
  f^*(y) = \|y\|_2
  \]

- $f(x) = \frac{1}{p} \|x\|_p^p$, $1 < p < \infty$, then
  \[
  f^*(y) = \frac{1}{q} \|x\|_q^q
  \]

  \[
  \frac{1}{p} + \frac{1}{q} = 1
  \]

- lots of smooth examples ......
Previously, we can represent $f$ by $f^*$ via

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

We can introduce a more general representation:

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

so that $h^*$ might be simpler than $f^*$ (or $h$ is simpler than $f$) in form.
Example $f(x) = \|x\|_1$

- Let $\mathcal{C} = \{y = [y_1; y_2]: y_1 + y_2 = 1, \ y_1, y_2 \geq 0\}$ and

\[
A = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.
\]

- We have

\[
\|x\|_1 = \sup_y \{(y_1 - y_2)^T x - \iota_{\mathcal{C}}(y)\} = \sup_y \{y^T Ax - \iota_{\mathcal{C}}(y)\}.
\]

- Since $\iota_{\mathcal{C}}^*([x_1; x_2]) = 1^T(\max\{x_1, x_2\})$, where max is taken entry-wise, and

\[
\sup_{y \in \text{dom} h^*} \{(Ax - b)^T y - h^*(y)\} = h(Ax - b)
\]

we have

\[
\|x\|_1 = \iota_{\mathcal{C}}^*(Ax) = 1^T(\max\{x, -x\}).
\]
**Application: dual smoothing**

**Idea:** “strongly convexify” $h^* \implies f$ becomes Lipschitz-differentiable

- Suppose that $f$ is represented in terms of $h^*$ as

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

- Let us strongly convexify $h^*$ by adding strongly convex function $d$:

  $$\hat{h}^*(y) = h^*(y) + \mu d(y)$$

  (a simple choice is $d(y) = \frac{1}{2} \| y \|_2^2$)

- Obtain a Lipschitz-differentiable approximation:

  $$f_\mu(x) = \sup_{y \in \text{dom} h^*} \{ y^T (A x - b) - \hat{h}^*(y) \}$$

- $f_\mu(x)$ is differentiable since $h^*(y) + \mu d(y)$ is strongly convex.
Example: augmented $\ell_1$

- **primal problem:** $\min \{ \|x\|_1 : Ax = b \}$
- **dual problem:** $\max \{ b^T y + \iota_{[-1,1]^n}(A^T y) \}$
- $f(y) = \iota_{[-1,1]^n}(y)$ is non-differentiable

**plan:** strongly convexify the primal so that dual becomes smooth and can be quickly solved

- let $f^*(x) = \|x\|_1$ and $f(y) = \iota_{[-1,1]^n}(y) = \sup_x \{ y^T x - f^*(x) \}$,
- add $\frac{\mu}{2} \|x\|_2$ to $f^*(x)$ and obtain

$$f_\mu(y) = \sup_x \{ y^T x - (\|x\|_1 + \frac{\mu}{2} \|x\|_2^2) \} = \frac{1}{2\mu} \|y - \text{Proj}_{[-1,1]^n}(y)\|_2^2$$

- $f_\mu(y)$ is differentiable; $\nabla f_\mu(y) = \frac{1}{\mu} \text{shrink}(y)$.
- On the other hand, we can also directly smooth $f^*(x) = \|x\|_1$ and obtain differentiable $f_\mu^*(x)$ by adding $d(y)$ to $f(y)$. (see the next slide ...)
Example: smoothed absolute value

Let $x \in \mathbb{R}$. Recall

$$f(x) = |x| = \sup_y \{yx - \iota_{[-1,1]}(y)\}$$

- add $d(y) = y^2/2$ to $\iota_{[-1,1]}(y)$ and obtain:

$$f_\mu = \sup_y \{yx - (\iota_{[-1,1]}(y) + \mu y^2/2)\} = \begin{cases} x^2/(2\mu), & |x| \leq \mu, \\ |x| - \mu/2, & |x| > \mu, \end{cases}$$

which is the Huber function.

- If $|x| \leq \mu$, the maximizing $y$ stays within $[-1, 1]$ even if $\iota_{[-1,1]}(y)$ vanishes, so it leaves with $\mu y^2/2$ only.
- If $|x| > \mu$, the maximizing $y$ occurs at $\pm 1$, so $-\mu y^2/2 = -\mu/2$.
- The Huber function is used in robust least squares.
add \( d(y) = 1 - \sqrt{1 - y^2} \), which is well defined and strongly convex in \([-1, 1]\):

\[
f^*_\mu = \sup_y \{yx - (\iota_{[-1,1]}(y) - \mu \sqrt{1 - y^2})\} - \mu = \sqrt{x^2 + \mu^2} - \mu,
\]

which is used in reweighted least-squares methods.

Recall

\[
|x| = \sup_y \{(y_1 - y_2)x - \iota_C(y)\}
\]

for \( C = \{y : y_1 + y_2 = 1, \ y_1, y_2 \geq 0\} \).

Add negative entropy \( d(y) = y_1 \log y_1 + y_2 \log y_2 + \log 2 \):

\[
f^*_\mu(x) = \sup_y \{(y_1 - y_2)x - (\iota_C(y) + \mu d(y))\} = \mu \log \frac{e^{x/\mu} + e^{-x/\mu}}{2}.
\]
$x \log(x)$ is strongly convex between $[0, C]$ for any finite $C > 0$
Compare three smoothed functions

 Courtesy of L. Vandenberghe
Example: smoothed maximum eigenvalue

- Let $X \in S^n$. Consider
  \[ f(X) = \lambda_{\text{max}}(X), \]
  which is the "$\ell_\infty$ norm" on the matrix spectrum.

- Recall the dual of $\ell_\infty$ is $\ell_1$, and we have
  \[ \|x\|_\infty = \sup_{y} \{ x^T y - \iota_{\{1^T y = 1, y \geq 0\}}(y) \} \]

- Let $C = \{ Y \in S^n : \text{tr}Y = 1, Y \succeq 0 \}$. We have
  \[ f(X) = \lambda_{\text{max}}(X) = \sup_{Y} \{ Y \bullet X - \iota_C(Y) \} \]

- Next, strongly convexify $\iota_C(Y)$:
Negative entropy of \( \{\lambda_i(Y)\} \):

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

(Courtesy of L. Vandenberghe)

Smoothed function

\[
f_{\mu}(X) = \sup_{Y} \{ Y \cdot X - (\nu C(Y) + \mu d(Y)) \} = \mu \log \left( \frac{1}{n} \sum_{i=1}^{n} e^{\lambda_i(X)/\mu} \right)
\]
Application: smoothed minimization

Instead of solving
\[
\min_x f(x),
\]
solve
\[
\min_{x} f_\mu(x) = \sup_{y \in \text{dom} h^*} \left\{ y^T (Ax + b) - [h^*(y) + \mu d(y)] \right\}
\]
by gradient descent, with acceleration, line search, etc......
Since $h^*(y) + \mu d(y)$ is strongly convex, $\nabla f_\mu(x)$ is given by:

$$\nabla f_\mu(x) = A^T \bar{y}, \quad \text{where } \bar{y} = \arg\max_{y \in \text{dom } h^*} \{ y^T (Ax + b) - [h^*(y) + \mu d(y)] \}.$$ 

**Theorem**

*If $d(y)$ is strongly convex with modulus $\nu > 0$, then*

- $h^*(y) + \mu d(y)$ is strongly convex with modulus at least $\mu \nu$
- $\nabla f_\mu(x)$ is Lipschitz continuous with constant no more than $\|A\|^2 / \mu \nu$. 
Nonsmooth optimization

Examples:

\[
\begin{align*}
\min & \|Ax - b\|_1 \\
\min & \text{TV}(x) \quad \text{s.t. } \|Ax - b\|_2 \leq \sigma
\end{align*}
\]

Worst-case complexity for \( \epsilon \)-approximation:

- If \( f \) is convex and \( \nabla f \) is \( L \)-Lipschitz, accelerated gradient method takes
  \[
  O\left(\sqrt{L}/\epsilon\right)
  \]
  iterations.

- If \( f \) is convex and nonsmooth, \( f \) is \( G \)-Lipschitz, subgradient method takes
  \[
  O\left(G^2/\epsilon^2\right)
  \]
  iterations.

Smooth optimization has much better complexities.
Nesterov’s complexity analysis

1. Construct smooth approximate satisfying

\[ f_\mu \leq f \leq f_\mu + \mu D \]

and consequently

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

2. Choose \( \mu \) such that \( \mu D \leq \epsilon/2 \) \( \Rightarrow \) \( \frac{1}{\mu} \geq \frac{2D}{\epsilon} \)

3. Minimize \( f_\mu \) such that \( f_\mu(x) - f^*_\mu \leq \epsilon/2 \)

Step 3 has complexity

\[ O\left(\sqrt{\frac{1}{\mu \epsilon}}\right) = O\left(\frac{\sqrt{D}}{\epsilon}\right), \]

which can be much better than the subgradient method’s

\[ O\left(G^2/\epsilon^2\right). \]
Theorem

Consider

\[ h(x) = \sup_{y \in \text{dom} h^*} y^T x - h^*(y), \]

\[ h_\mu(x) = \sup_{y \in \text{dom} h^*} y^T x - (h^*(y) + \mu d(y)), \]

where \( d(y) \geq 0 \) is strongly convex with modulus \( \nu > 0 \). Then,

1. \( \nabla h_\mu \) is \( (\mu \nu)^{-1} \)-Lipschitz;

2. if \( d(y) \leq D \) for \( y \in \text{dom} h^* \), then

\[ h_\mu(x) \leq h(x) \leq h_\mu(x) + \mu D, \quad x \in \text{dom} h. \]
Example: Huber function

Recall

\[ h_\mu(x) = \sup_y \{yx - (\iota_{[-1,1]}(y) + \mu y^2/2)\} = \begin{cases} 
    x^2/(2\mu), & |x| \leq \mu, \\
    |x| - \mu/2, & |x| > \mu,
\end{cases} \]

We have

\[ h_\mu(x) \leq |x| \leq h_\mu(x) + \mu/2 \]

and

\[ h'_\mu(x) = \begin{cases} 
    x/\mu, & |x| \leq \mu, \\
    \text{sign}(x), & |x| > \mu,
\end{cases} \]

which is Lipschitz with constant \( \mu^{-1} \).

Apply to \( \ell_1 \)-norm:

\[ \sum_i h_\mu(x_i) \leq \|x\|_1 \leq \sum_i h_\mu(x_i) + n\mu/2. \]
Robust least squares

Consider

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

- Representation

$$\|x\|_1 = \sup_y \{y^T(Ax - b) - \iota_C(y)\}$$

where $C = \{y : \|y\|_\infty \leq 1\}$.

- Add $\mu d(y) = \frac{\mu}{2} \|y\|_2^2$ to $\iota_C(y)$ and obtain

$$\min_x f_\mu(x) = \sum_{i=1}^m h_\mu(a_i^T x - b_i).$$
Other examples and questions

- Total variation / analysis $\ell_1$ minimization examples
- Nuclear norm examples
- More other one nonsmooth terms
- ...
- Stopping criteria