MATH 680 Computation Intensive Statistics

November 27, 2019

Dual Norm and Conjugate Function

1 Dual Norm

- Let $\|x\|$ be a norm, e.g.,
  - ℓ_p norm: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, for $p \geq 1$.
  - Trace norm: $\|X\|_{tr} = \sum_{i=1}^r \sigma_i(X)$
- **Dual norm:** for a vector x, we define its dual norm $\|x\|_*$ as

$$\|x\|_* = \max_{\|z\| \leq 1} z^\top x,$$

  where $\|\cdot\|$ is the original norm.

  We have the inequality (Cauchy-schwarz like)

$$|z^\top x| \leq \|z\| \|x\|_*.$$

  This is because $\|x^*\| = \max_{\|z\| \leq 1} z^\top x \geq \left(\frac{z}{\|z\|}\right)^\top x$

- The dual norm of the ℓ_1 norm is the ℓ_∞ norm. Let $\|z\| = \sum_{i=1}^p |z_i| = \|z\|_1$ (ℓ_1 norm).

$$\max_{\sum_i |z_i| \leq 1} \sum_i z_i y_i$$

$$= \max_i |y_i| = \|y\|_\infty.$$

- The dual norm of the ℓ_2 norm is the ℓ_2 norm. Since $\|z\|_2 \le 1$,

$$\max_{\|z\|_2 \le 1} z^\top y \le \|z\|_2 \|y\|_2 \le \|y\|_2,$$

where the "=" is taken when

$$z = \begin{cases} \|y\|_2^{-1} \cdot y, & y \neq 0 \\ 0, & y = 0. \end{cases}$$

- The dual norm of the ℓ_p norm ($p > 1$) is the ℓ_q norm ($q > 1$) and $\frac{1}{p} + \frac{1}{q} = 1$. Since

$$|a_1 b_1 + \cdots + a_k b_k| \le (a_1^p + \cdots + a_k^p)^{1/p} (b_1^q + \cdots + b_k^q)^{1/q}$$

for $\frac{1}{p} + \frac{1}{q} = 1$, $p > 1$, and $q > 1$.

- Trace norm dual: $(\|X\|_{tr})_* = \|X\|_{op} = \sigma_1(X)$.

- Dual norm of dual norm: can show that $\|x\|_{**} = \|x\|$

2 Conjugate Function

- **Conjugate function:** given $f : \mathbb{R}^n \to \mathbb{R}$, the function

$$f^*(y) = \max_x y^\top x - f(x)$$

is called its conjugate.

Proposition 1. *f^* is always convex.*

Proof. For any y_1, y_2 and $0 \leq \alpha \leq 1$, let $y_\alpha = \alpha y_1 + (1-\alpha) y_2$. Then,

$$f^*(y_\alpha) = \max_x y_\alpha^\top x - f(x).$$

Note that

$$y_\alpha^\top x - f(x) = \alpha \left(y_1^\top x - f(x) \right) + (1-\alpha) \left(y_2^\top x - f(x) \right)$$

which implies that

$$f^*(y_\alpha) \leq \alpha f^*(y_1) + (1-\alpha) f^*(y_2).$$

□

- **Fenchel's inequality:** $f(x) + f^*(y) \geq y^\top x$.
- Conjugate of conjugate f^{**} satisfies $f^{**} \leq f$

Proof. We have

$$\begin{aligned} & f^*(y) \geq y^\top x - f(x) \\ \Longrightarrow\ & f(x) \geq y^\top x - f^*(y) \\ \Longrightarrow\ & f(x) \geq \max_y y^\top x - f^*(y) = f^{**}, \end{aligned}$$

so $f \geq f^{**}$. □

If f is closed (continuous) and convex, then $f^{**} = f$. Also for any x, y.

$$
\begin{aligned}
& y \in \partial f(x) \Longleftrightarrow x \in \partial f^*(y) \\
\Longleftrightarrow & x \in \arg\min_z f(z) - y^\top z \Longleftrightarrow x \in \arg\max_z y^\top z - f(z) \\
\Longleftrightarrow & f(x) + f^*(y) = y^\top x
\end{aligned}
$$

If f is strictly convex, then $\nabla f^*(y) = \arg\min_z f(z) - y^\top z$.

Proof. We can easily see that

$$
\begin{aligned}
& y \in \partial f(x) \\
\Longleftrightarrow & 0 \in \partial(f(x) - y^\top x) \\
\Longleftrightarrow & x \in \arg\min_z f(z) - y^\top z \\
\Longleftrightarrow & x \in \arg\max_z y^\top z - f(z) \\
\Longleftrightarrow & y^\top x - f(x) = \max_z (y^\top z - f(z)) = f^*(y)
\end{aligned}
$$

Now we just need to prove $y^\top x - f^*(y) = f(x) \Longleftrightarrow x \in \partial f^*(y)$. Since

$$
\begin{aligned}
& y^\top x - f^*(y) = f(x) \\
\Longleftrightarrow & y^\top x - f^*(y) = \max_z z^\top x - f^*(z) \qquad (f = f^{**}) \\
\Longleftrightarrow & y \in \arg\max_z z^\top x - f^*(z) \\
\Longleftrightarrow & y \in \arg\min_z f^*(z) - z^\top x \\
\Longleftrightarrow & 0 \in \partial(f^*(y) - y^\top x) \\
\Longleftrightarrow & x \in \partial f^*(y)
\end{aligned}
$$

$\square$

- If $f(u, v) = f_1(u) + f_2(v)$, then $(u \in \mathbb{R}^n, v \in \mathbb{R}^m)$,

$$
f^*(w, z) = f_1^*(w) + f_2^*(z)
$$

- Example: $f(x) = \frac{1}{2}x^\intercal Qx$, $Q \succ 0$.

$$
\begin{aligned}
f^*(y) &= \max_x y^\top x - \frac{1}{2}x^\top Qx \\
&= -\min_x \frac{1}{2}x^\top Qx - y^\top x \qquad\qquad (taking\ x = Q^{-1}y) \\
&= -\min_x \frac{1}{2}(Q^{-1}y)^\top Q(Q^{-1}y) - y^\top Q^{-1}y \\
&= \frac{1}{2}y^\top Q^{-1}y.
\end{aligned}
$$

- Fenchel's inequality gives

$$
f(x) + f^*(y) \geq x^\top y \Longrightarrow \frac{1}{2}x^\top Qx + \frac{1}{2}y^\top Q^{-1}y \geq x^\top y
$$

- **Conjugate of indicator function:** if $f(x) = I_c(x)$, then its conjugate is

$$
f^*(y) = I_C^*(y) = \max_{x \in C} y^\top x
$$

  Since $f^*(y) = \max_x y^\top x - I_C(x) = \max_{x \in C} y^\top x$.

- **Conjugate of norm:** If $f(x) = \|x\|$ (any norm), then its conjugate is

$$
f^*(y) = \begin{cases} 0 & \|y\|_* \leq 1 \\ +\infty & \|y\|_* > 1 \end{cases}
$$

  or can be written as

$$
f^*(y) = I_{\{z : \|z\|_* \leq 1\}}(y)
$$

  Proof. recall the definition of dual norm

$$
\|y\|_* = \max_{\|x\| \leq 1} x^\top y
$$

  to evaluate

$$
f^*(y) = \max_x y^\top x - \|x\|
$$

we distinguish two cases $\square$

- If $\|y\|_* \leq 1$, then by definition of dual norm

$$y^\top x \leq \|x\|\|y\|_* \leq \|x\| \qquad \forall x$$

and equality holds if $x = 0$; Therefore $f^*(y) = \max_x y^\top x - \|x\| = 0$.

- If $\|y\|_* > 1$, by the definition of dual norm $\|y\|_* = \max_{\|x\| \leq 1} x^\top y > 1$, there exists an x with $\|x\| \leq 1$, $x^\top y > 1$, then

$$f^*(y) \geq y^\top (tx) - \|tx\| = t(y^\top x - \|x\|) \xrightarrow{t \to \infty} \infty$$

3 Conjugates and Dual Problem

Conjugates appear frequently in derivation of dual problems, via

$$\begin{aligned} f^*(u) &= \max_x u^\top x - f(x) \\ &= -\min_x f(x) - u^\top x \end{aligned}$$

Therefore

$$-f^*(u) = \min_x f(x) - u^\top x$$

in minimization of the Lagrangian.

E.g. consider

$$\begin{aligned} &\min_x f(x) + g(x) \\ \Longleftrightarrow &\min_x f(x) + g(z) \qquad \text{subject to } x = z \end{aligned}$$

Lagrange dual function is

$$g(u) = \min f(x) + g(z) + u^\top (z - x) = -f^*(u) - g^*(-u)$$

Hence the dual problem is

$$\max_u -f^*(u) - g^*(-u)$$

4 Lasso Dual (through duality)

The Lasso primal is

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Introduce $z = X\beta$, and the dual variable u

$$\begin{aligned} &\min \frac{1}{2}\|y - z\|^2 + \lambda\|\beta\|_1 \\ &\text{s.t. } X\beta - z = 0. \end{aligned}$$

Then we have Lagrangian

$$L(\beta, z, u) = \frac{1}{2}\|y - z\|^2 + \lambda\|\beta\|_1 + u^\top(X\beta - z)$$

and Lagrange dual function

$$\begin{aligned} g(u) &= \min_{\beta,z} L(\beta, z, u) \\ &= \min_\beta \left\{\lambda\|\beta\|_1 - (X^\top u)^\top\beta\right\} + \min_z \left\{\frac{1}{2}\|y - z\|_2^2 + u^\top z\right\} \\ &= -\lambda \max_\beta \left(\frac{(X^\top u)^\top}{\lambda}\beta - \|\beta\|_1\right) + \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 \\ &= -\lambda I_{\{z:\|z\|_\infty \le 1\}}\left(\frac{X^\top u}{\lambda}\right) + \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 \end{aligned}$$

Thus the dual problem is

$$\max_u -\frac{1}{2}\|y - u\|_2^2 - \lambda I_{\{z:\|z\|_\infty \le 1\}}\left(\frac{X^\top u}{\lambda}\right)$$

which is equivalent to

$$\begin{aligned} \max_u -\frac{1}{2}\|y - u\|_2^2 \\ \text{subject to} \|\frac{X^\top u}{\lambda}\|_\infty \le 1 \end{aligned}$$

which is equivalent to

$$\min_{u} \|y - u\|_2^2$$
$$\text{subject to} \|X^\top u\|_\infty \leq \lambda$$

Note that the problem now becomes solving $u \in \mathbb{R}^n$ instead of solving $\beta \in \mathbb{R}^p$. Suppose now we have solve the dual problem and the solution is $u^\star$. Then $\beta^\star, z^\star$ must minimize $L(\beta, z, u^\star)$

$$\nabla_z L(\beta, z, u^\star) = 0 \iff z^\star = y - u^\star \iff X\beta^\star = y - u^\star$$

Therefore

$$\|X^\top u^*\|_\infty \leq \lambda \Longrightarrow \|X^\top (y - X\beta^*)\|_\infty \leq \lambda \text{ (fitted residual)}$$

and

$$\beta^* = \arg\max_{\beta} \left\{ \frac{(X^\top u^\star)^\top}{\lambda} \beta - \|\beta\|_1 \right\}.$$

5 Lasso Dual (through KKT)

Recall the definition of polyhedron. A set $C \subseteq \mathbb{R}^n$ is called a convex Polyhedron if C is the intersection of many half-spaces:

$$C = \cap_{i=1}^k \{x \in \mathbb{R}^n : a_i^\top x \leq b_i\}$$
$$\text{Where} \quad a_1, \ldots, a_k \in \mathbb{R}^n \quad \text{and} \quad b_1, \ldots, b_k \in \mathbb{R}$$

The aim of this section is to show that the LASSO problem can be formulated as a projection onto a polyhedron. Before we delve into the details of the derivation, we state a couple of preliminary results that will be used in later proofs:

- A polyhedron is a closed and convex set.
- For any closed and convex set $C \subseteq \mathbb{R}^n$ and point $x \in \mathbb{R}^n$, there is a unique point $u \in C$ minimizing

$\|x-u\|_2$. This point is a projection of x onto set C, which we denote by $\Pi_C(x)$.

In the linear regression setting, with response variable $y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n\times p}$, if we regress Y on X using the LASSO, the optimal model parameters $X\hat{\beta}$ can be written as:

$$X\hat{\beta} = y - \Pi_C(y) = (I - \Pi_C)(y),$$

where C is a polyhedron

Proof. Given $y \in \mathbb{R}^n$.

$$\theta = \Pi_C(y)$$

onto a closed convex set $C \subseteq \mathbb{R}^n$ can be characterised as the unique point satisfying

$$\langle y-\theta, \theta-u\rangle \geq 0, \quad \forall u \in C. \tag{1}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. Based on this, if we define

$$\theta = y - X\hat{\beta}(y),$$

or equivalently:

$$X\hat{\beta} = y - \theta$$

can be regarded as a function of y, We want to show that the inequality 1 holds for all $u \in C$, where C is defined as

$$C := \bigcap_{j=1}^{p} \left(\left\{u \in \mathbb{R}^n \colon X_j^\top u \leq \lambda\right\} \cap \left\{u \in \mathbb{R}^n \colon X_j^\top u \geq -\lambda\right\}\right),$$

which is equivalent to

$$\left\{u \in \mathbb{R}^n \colon \left\|X^\top u\right\|_\infty \leq \lambda\right\}.$$

To show this, we can see that

$$\begin{aligned}\langle y-\theta, \theta-u\rangle &= \langle X\hat{\beta}, y - X\hat{\beta} - u\rangle \\ &= \langle X\hat{\beta}, y - X\hat{\beta}\rangle - \langle X^\top u, \hat{\beta}\rangle.\end{aligned}$$

From the KKT conditions for LASSO, we know that the optimization problem

$$\min_{\beta} g(\beta) + h(\beta) = \min_{\beta} \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{g(\beta)} + \underbrace{\lambda\|\beta\|_1}_{h(\beta)}$$

satisfies the **stationarity** condition, which can be stated as:

$$\begin{aligned} 0 &\in \partial\left(g(\hat{\beta}) + h(\hat{\beta})\right) \\ &= \nabla g(\hat{\beta}) + \partial h(\hat{\beta}) \\ &= -X^{\top}\left(y - X\hat{\beta}\right) + \lambda \partial \left\|\hat{\beta}\right\|_1 . \end{aligned}$$

Thus

$$X^{\top}\left(y - X\hat{\beta}\right) = \lambda\gamma, \tag{2}$$

where

$$\gamma_j = \left(\partial\left\|\hat{\beta}\right\|_1\right)_j = \begin{cases} \text{sgn}\left(\hat{\beta}_j\right) & \text{if } \hat{\beta}_j \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_j = 0 \end{cases} .$$

$\square$

Taking the inner product with $\hat{\beta}$ on both sides of 2, we always have

$$\langle X\hat{\beta}, y - X\hat{\beta}\rangle = \lambda\left\|\hat{\beta}\right\|_1 = \max_{\|w\|_\infty \leq \lambda} w^{\top}\hat{\beta}.$$

The RHS holds since $\beta_j\gamma_j = \begin{cases} 0 & \hat{\beta}_j = 0 \\ |\hat{\beta}_j| & \text{otherwise} \end{cases}$. Therefore,

$$\langle y - \theta, \theta - u\rangle = \max_{\|X^{\top}u\|_\infty \leq \lambda} \langle X^{\top}u, \hat{\beta}\rangle - \langle X^{\top}u, \hat{\beta}\rangle \geq 0 \ \ \forall u \in C,$$

which implies that θ is indeed a projection of y onto C,

$$\theta = y - X\hat{\beta}(y) = \Pi_C(y).$$

6 Screening

Let f be differentiable and strictly convex, let $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$. Consider

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

The KKT conditions are

$$\begin{cases} X_j^T(y - X\beta^*) = \lambda \cdot \text{sgn}(\beta_j^*) & \beta_j^* \neq 0 \\ |X_j^T(y - X\beta^*)| \leq \lambda & \beta_j^* = 0 \end{cases}$$

which implies that

$$|X_j^T(y - X\beta^*)| < \lambda \quad \Longrightarrow \quad \beta_j^* = 0 \tag{3}$$

Suppose that the dual solution is $u^\star$. Then $\beta^\star, z^\star$ must minimize $L(\beta, z, u^\star)$

$$\begin{aligned} &\nabla_z L(\beta, z, u^\star) = 0 \\ \Longleftrightarrow &\nabla_z \left\{\frac{1}{2}\|y - z\|_2^2 + u^\top z\right\} = 0 \\ \Longleftrightarrow &-(y - z^\star) + u^\star = 0 \\ \Longleftrightarrow &y - X\beta^\star = u^\star \end{aligned}$$

Replace $y - X\beta^\star$ using u^* in (3) we get

$$|X_j^T u^*| < \lambda \quad \Longrightarrow \quad \beta_j^* = 0.$$

This is when u^* is in the interior of the slab defined by the feature X_j, therefore