Notes on LASSO

Jungbin Jun

1 Lagrange Duality

1.1 Primal Optimization Problem

Consider the following primal optimization problem:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0, \quad i \in [k], \\ & h_j(x) = 0, \quad j \in [r]. \end{aligned}$$

Define the (general) Lagrangian

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_{i \in [k]} \alpha_i \, g_i(x) + \sum_{j \in [r]} \beta_j \, h_j(x),$$

where $\alpha_i \ge 0$ for all $i \in [k]$ and $\beta_j \in \mathbb{R}$ for all $j \in [r]$.
Next, define the function

$$\Theta_P(x) = \max_{\alpha \ge 0, \, \beta} \mathcal{L}(x, \alpha, \beta).$$

It is straightforward to verify that

$$\Theta(x) = \begin{cases} f(x), & \text{if } x \text{ satisfies all the primal constraints,} \\ \infty, & \text{otherwise.} \end{cases}$$

Therefore, the primal optimization problem can be equivalently written as

$$\min_x \Theta_P(x) = \min_x \max_{\alpha \ge 0, \, \beta} \mathcal{L}(x, \alpha, \beta) = \min_{x \in \mathcal{F}} f(x),$$

where $\mathcal{F}$ is a feasible set for primal optimization problem.

1.2 Dual Optimization Problem

Define the dual function

$$\Theta_D(\alpha, \beta) = \inf_x \mathcal{L}(x, \alpha, \beta).$$

Then, the dual optimization problem is given by

$$\max_{\alpha \ge 0, \, \beta} \Theta_D(\alpha, \beta) = \max_{\alpha \ge 0, \, \beta} \inf_x \mathcal{L}(x, \alpha, \beta).$$

By well-known inequality ($\sup_y \inf_x F(x, y) \le \inf_y \sup_x F(x, y)$ for any F), the following holds.

$$\max_{\alpha \ge 0, \, \beta} \inf_x \mathcal{L}(x, \alpha, \beta) \;\le\; \inf_x \max_{\alpha \ge 0, \, \beta} \mathcal{L}(x, \alpha, \beta),$$

which is referred to as *weak duality*.
Under certain conditions (under KKT conditions!), however, equality holds, i.e.,

$$\max_{\alpha \geq 0,\, \beta} \inf_{x} \mathcal{L}(x, \alpha, \beta) = \inf_{x} \max_{\alpha \geq 0,\, \beta} \mathcal{L}(x, \alpha, \beta).$$

This phenomenon is known as *strong duality*.
Moreover, the dual problem is often easier to solve than the primal problem, both analytically and computationally.

1.3 Karush-Kuhn-Tucker(KKT) Condition

When the optimization problem satisfies the Karush–Kuhn–Tucker (KKT) conditions, the optimal solution of the dual problem coincides with that of the primal problem. In this case, strong duality holds.
The KKT conditions are given as follows:

- Stationarity : $\nabla_x \mathcal{L}(x^*, \alpha^*, \beta^*) = 0$
- Complementary Slackness : $\alpha_i^* g_i(x^*) = 0, \quad \forall i \in [k]$
- Primal Feasibility : $g_i(x^*) \leq 0,\ h_j(x^*) = 0 \quad \forall i \in [k],\ \forall j \in [r]$
- Dual Feasibility : $\alpha_i \geq 0 \quad \forall i \in [k]$

If $\alpha_i > 0$, then $g_i(x) = 0$. (Active Constraint).
If $\alpha_i = 0$, then $g_i(x) \leq 0$. (Inactive Constraint)

2 Subgradient

Definition 1 (Subgradient)**.** A subgradient of a convex function f at x is any $g \in \mathbb{R}^d$ such that

$$f(y) \geq f(x) + g^\top (y - x) \quad \forall y \in \mathrm{dom}(f).$$

Definition 2 (Subdifferential)**.** A set of all subgradient is called as subdifferential of f at x which is

$$\partial f(x) = \{g \in \mathbb{R}^d : g \text{ is a subgradient of } f \text{ at } x\}.$$

Properties of Subdifferential

- $\partial f(x)$ is closed and convex. (even for non-convex f)
- f is differentiable at $x \iff \partial f(x) = \{\nabla_x f(x)\}$.
- $\partial(a_1 f_1 + a_2 f_2) = a_1 \partial f_1 + a_2 \partial f_2$.
- $g(x) = f(Ax + b) \Rightarrow \partial g(x) = A^\top \partial f(Ax + b)$.

Theorem 1. *For any f,*

$$f(x^*) = \min_x f(x) \iff 0 \in \partial f(x^*)$$

which implies that x^* is a minimizer of f if and only if 0 is a subgradient of f at x^*.

Proof. If $0 \in \partial f(x^*)$, then

$$f(y) \geq f(x^*) + 0^\top (y - x) = f(x^*)$$

which implies that x^* is a minimizer for f. □

3 LASSO

$$\hat{\beta}^{\text{LASSO}} = \underset{b \in \mathbb{R}^{p+1}}{\arg\min} \left[\frac{1}{2n} \|Y - Xb\|^2 + \lambda \|b\|_1 \right]$$

3.1 Subgradient of LASSO

By Theorem 1, the estimator $\hat{\beta}$ satisfies

$$\begin{aligned}
\hat{\beta} &= \arg\min_{b \in \mathbb{R}^{p+1}} \left\{ \frac{1}{2n} \|Y - Xb\|_2^2 + \lambda \|b\|_1 \right\} \\
&\iff 0 \in \partial_b \left(\frac{1}{2n} \|Y - Xb\|_2^2 + \lambda \|b\|_1 \right) \\
&\iff 0 \in \frac{1}{n} X^\top (Xb - Y) + \lambda \, \partial \|b\|_1 \\
&\iff \frac{1}{n} X^\top (Y - Xb) = \lambda \, u, \qquad \exists \, u \in \partial \|b\|_1.
\end{aligned}$$

Note that $\|b\|_1 = \sum_{j=1}^{p} |b_j|$ and by the property of subdifferential,

$$\partial \|b\|_1 = \partial \left(\sum_{j=1}^{p} |b_j| \right) = \sum_{j=1}^{p} \partial |b_j|.$$

Thus, we have

$$u_i \in \begin{cases} \{1\} & b_j > 0, \\ \{-1\} & b_j < 0, \\ [-1, 1] & b_j = 0. \end{cases}$$

Finally, we have

$$\begin{cases} x_i^\top (y - X\hat{\beta}) = \lambda \, \text{sgn}(\hat{\beta}_j) & \hat{\beta}_j \neq 0 \\ |x_i^\top (y - X\hat{\beta})| \leq \lambda & \hat{\beta}_j = 0. \end{cases}$$