Discussion 13

1. Sphere Enclosure

For $i = 1, \ldots, m$, let B_i be a ball in $\mathbb{R}^n$ with center $\vec{x}_i$, and radius $\rho_i \geq 0$. We wish to find a ball B of minimum radius that contains all the B_i for $i = 1, \ldots, m$. Cast this problem as an SOCP.

Solution: Let $\vec{c} \in \mathbb{R}^n$ and $r \geq 0$ denote the center and radius of the enclosing ball B, respectively. We express the given balls B_i as

$$B_i = \{\vec{x} : \vec{x} = \vec{x}_i + \vec{\delta}_i,\ \|\vec{\delta}_i\|_2 \leq \rho_i\}, \quad i = 1, \ldots, m.$$

We have that $B_i \subseteq B$ if and only if

$$\max_{\vec{x} \in B_i} \|\vec{x} - \vec{c}\|_2 \leq r.$$

Note that

$$\max_{\vec{x} \in B_i} \|\vec{x} - \vec{c}\|_2 = \max_{\|\vec{\delta}_i\|_2 \leq \rho_i} \|\vec{x}_i - \vec{c} + \vec{\delta}_i\|_2 = \|\vec{x}_i - \vec{c}\|_2 + \rho_i.$$

The last step follows by choosing $\vec{\delta}_i$ in the direction of $\vec{x}_i - \vec{c}$.
The problem is then cast as the following SOCP:

$$\min_{\vec{c} \in \mathbb{R}^n, r \in \mathbb{R}} r$$

$$\text{subject to: } \|\vec{x}_i - \vec{c}\|_2 + \rho_i \leq r, i = 1, \ldots, m.$$

2. LASSO vs. Ridge

Consider the data set $\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1,\ldots,n}$ of samples $\vec{x}^{(i)} \in \mathbb{R}^d$ and values $y^{(i)} \in \mathbb{R}$. Define $X = \begin{bmatrix} \vec{x}^{(1)} & \ldots & \vec{x}^{(n)} \end{bmatrix}^\top \in \mathbb{R}^{n\times d}$ and $\vec{y} = \begin{bmatrix} y^{(1)} & \ldots & y^{(n)} \end{bmatrix}^\top \in \mathbb{R}^n$, i.e., X is the $n \times d$ matrix whose i-th row is $(\vec{x}^{(i)})^\top$, for each $i \in \{1, \cdots, n\}$, and $\vec{y}$ is the n-dimensional column vector whose i-th component is y_i, for each $i \in \{1, \cdots, n\}$.

For the sake of simplicity, assume that the data has been centered and whitened so that each feature has mean 0 and variance 1 and the features are uncorrelated, i.e. $X^\top X = nI_{d\times d}$, where $I_{d\times d}$ denotes the $d \times d$ identity matrix. Consider the linear least squares regression with regularization in the ℓ_1-norm, also known as LASSO:

$$\vec{w}^\star = \underset{\vec{w}\in\mathbb{R}^d}{\operatorname{argmin}}\ \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_1 . \tag{1}$$

This problem will compare ℓ_1-regularization with ℓ_2-regularization (ridge regression) to understand their similarities and differences, by looking at the elements of $\vec{w}^\star$ in the solution to each problem.

(a) First, decompose this optimization problem into d univariate optimization problems over each element of $\vec{w}$. *Hint*: Let $\vec{x}_j \in \mathbb{R}^n$ denote the j-th column of X, so that $X = \begin{bmatrix} \vec{x}_1 & \ldots & \vec{x}_d \end{bmatrix}$ and recall that $X^\top X = nI_{d\times d}$.

Solution: Note that

$$\begin{aligned} \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_1 &= \vec{w}^\top X^\top X\vec{w} - 2\vec{y}^\top X\vec{w} + \vec{y}^\top\vec{y} + \lambda\|\vec{w}\|_1 \\ &= \sum_{i=1}^{d} \left[nw_i^2 - 2\vec{y}^\top\vec{x}_i w_i + \lambda |w_i|\right] + \vec{y}^\top\vec{y}. \end{aligned}$$

Hence the original problem becomes

$$\min_{\vec{w}\in\mathbb{R}^d} \sum_{i=1}^{d} \left[nw_i^2 - 2\vec{y}^\top\vec{x}_i w_i + \lambda |w_i|\right], \tag{2}$$

where we have removed $\vec{y}^\top\vec{y}$ from the objective function because we can add it back in after solving the problem. Since the objective is separable in w_i the problem decomposes into the following d univariate optimization problems

$$\min_{w_i\in\mathbb{R}} \left[nw_i^2 - 2\vec{y}^\top\vec{x}_i w_i + \lambda |w_i|\right]. \tag{3}$$

(b) Prove that for any $i \in \{1, \cdots, d\}$, if $\vec{y}^\top\vec{x}_i > \frac{1}{2}\lambda$ then $w_i^\star > 0$. Find $w_i^\star$ in that case.

Solution: For each $i \in \{1, \cdots, d\}$, let $f_i : \mathbb{R} \to \mathbb{R}$ be the objective function of the i-th univariate optimization problem derived above, i.e., for each $w_i \in \mathbb{R}$:

$$\begin{aligned} f_i(w_i) &:= nw_i^2 - 2\vec{y}^\top\vec{x}_i w_i + \lambda|w_i| \\ &= \begin{cases} nw_i^2 + (-2\vec{y}^\top\vec{x}_i + \lambda)w_i, & \text{if } w_i \geq 0, \\ nw_i^2 + (-2\vec{y}^\top\vec{x}_i - \lambda)w_i, & \text{else} \end{cases} \end{aligned}$$

Then the derivative of $f_i(w_i)$ at each $w_i \neq 0$ can be piecewisely defined as

$$\frac{df_i}{dw_i}(w_i) = \begin{cases} 2nw_i + (-2\vec{y}^\top\vec{x}_i + \lambda), & \text{if } w_i > 0, \\ 2nw_i + (-2\vec{y}^\top\vec{x}_i - \lambda), & \text{if } w_i < 0. \end{cases}$$

For convenience, define $g_i, h_i : \mathbb{R} \to \mathbb{R}$ by:

$$g_i(w_i) := nw_i^2 + (-2\vec{y}^\top\vec{x}_i + \lambda)w_i,$$

$$h_i(w_i) := nw_i^2 + (-2\vec{y}^\top \vec{x}_i - \lambda)w_i$$

for each $w_i \in \mathbb{R}$. Notice that g_i and h_i attain (unique) minimizers at $\hat{w}_i := \frac{1}{2n}(2\vec{y}^\top \vec{x}_i - \lambda)$ and $\tilde{w}_i := \frac{1}{2n}(2\vec{y}^\top \vec{x}_i + \lambda)$, respectively.

If $\vec{y}^\top \vec{x}_i > \frac{1}{2}\lambda$, then $\hat{w}_i > 0$, so for each $w_i > 0$ we have

$$f_i(\hat{w}_i) = g_i(\hat{w}_i) \leq g_i(w_i) = f_i(w_i), \tag{4}$$

with equality if and only if $w_i = \hat{w}_i$. Moreover, for any $w_i < 0$:

$$\begin{aligned} \frac{df_i}{dw_i}(w_i) = \frac{dh_i}{dw_i}(w_i) &= 2nw_i + (-2\vec{y}^\top \vec{x}_i - \lambda) \\ &< 0 + (-\lambda - \lambda) \\ &< 0. \end{aligned} \tag{5}$$

This implies that for each $w_i < 0$:

$$f_i(\hat{w}_i) < \lim_{w \to 0^+} f_i(w) = \lim_{w \to 0^-} f_i(w) < f_i(w_i).$$

Above, the first inequality follows from (4), the equality follows from the continuity of f_i at $w = 0$, and the second inequality follows from (5). We thus conclude that $w_i^\star = \hat{w}_i = \frac{1}{2n}(2\vec{y}^\top \vec{x}_i - \lambda) > 0$.

(c) Prove that for any $i \in \{1, \cdots, d\}$, if $\vec{y}^\top \vec{x}_i < -\frac{1}{2}\lambda$ then $w_i^\star < 0$. Find $w_i^\star$ in that case.

Solution: For each $i \in \{1, \cdots, d\}$, let $f_i, g_i, h_i : \mathbb{R} \to \mathbb{R}$ and $\hat{w}_i, \tilde{w}_i \in \mathbb{R}$ be as defined the solution to the above sub-problem. If $\vec{y}^\top \vec{x}_i < -\frac{1}{2}\lambda$, then $\tilde{w}_i < 0$, so for each $w_i < 0$ we have

$$f_i(\tilde{w}_i) = h_i(\tilde{w}_i) \leq h_i(w_i) = f_i(w_i). \tag{6}$$

with equality if and only if $w_i = \tilde{w}_i$. Moreover, for any $w_i > 0$:

$$\begin{aligned} \frac{df_i}{dw_i}(w_i) = \frac{dg_i}{dw_i}(w_i) &= 2nw_i + (-2\vec{y}^\top \vec{x}_i + \lambda) \\ &> 0 + (\lambda + \lambda) \\ &> 0. \end{aligned} \tag{7}$$

This implies that for each $w_i > 0$:

$$f_i(\tilde{w}_i) < \lim_{w \to 0^-} f_i(w) = \lim_{w \to 0^+} f_i(w) < f_i(w_i).$$

Above, the first inequality follows from (6), the equality follows from the continuity of f_i at $w = 0$, and the second inequality follows from (7). We thus conclude that $w_i^\star = \tilde{w}_i = \frac{1}{2n}(2\vec{y}^\top \vec{x}_i + \lambda) < 0$.

(d) Prove that for any $i \in \{1, \cdots, d\}$, if $|\vec{y}^\top \vec{x}_i| < \frac{1}{2}\lambda$ then $w_i^\star = 0$.

Solution: For each $i \in \{1, \cdots, d\}$, let $f_i, g_i, h_i : \mathbb{R} \to \mathbb{R}$ be as defined the solution to the above two sub-problems. If $|\vec{y}^\top \vec{x}_i| < \frac{1}{2}\lambda$, then for any $w_i > 0$:

$$\frac{df_i}{dw_i}(w_i) = \frac{dg_i}{dw_i}(w_i) = 2nw_i + (-2\vec{y}^\top \vec{x}_i + \lambda) > 0,$$

so we have

$$f_i(0) = \lim_{w \to 0^+} f_i(w) < f_i(w_i),$$

since f_i is continuous at $w_i = 0$. Similarly, for any $w_i < 0$

$$\frac{df_i}{dw_i}(w_i) = \frac{dh_i}{dw_i}(w_i) = 2nw_i + (-2\vec{y}^\top \vec{x}_i - \lambda) < 0,$$

so we have

$$f_i(0) = \lim_{w \to 0^-} f_i(w) < f_i(w_i).$$

To summarize, $f_i(0) < f_i(w_i)$ for any $w_i \neq 0$, so $w_i^\star = 0$.

In words, a larger value of λ will force more entries of $\vec{w}$ to be zero — i.e. larger λ will imply higher sparsity.

(e) Now consider the case of ridge regression, which uses the the ℓ_2 regularization $\lambda \|\vec{w}\|_2^2$.

$$\vec{w}^\star = \operatorname*{argmin}_{\vec{w} \in \mathbb{R}^d} \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_2^2. \tag{8}$$

Write down the new condition for $\vec{w}_i^\star$ to be 0. How does this differ from the condition obtained in part (4) and what does this suggest about LASSO?

Solution: In the case of ridge regression the optimal weight vector $\vec{w}$ is given by

$$w_i^\star = \frac{\vec{y}^\top \vec{x}_i}{n + \lambda}, \ i = 1, \dots, d. \tag{9}$$

So $w_i^\star$ is only zero when $\vec{y}^\top \vec{x}_i = 0$, in contrast to LASSO where $w_i^\star$ is zero when $\vec{y}^\top \vec{x}_i \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right]$. This suggest that LASSO forces a lot of coordinates to be zero, i.e. induces sparsity to the optimal weight vector.