REGULARIZATION

1. Multiple linear regression

Convention: We view $\mathbb{R}^k$ as a subset of $\mathbb{R}^{k+1}$ via the following identification

$$v \in \mathbb{R}^k \quad \longleftrightarrow \quad (1, v) \in \mathbb{R}^{k+1}. \tag{1}$$

$p-1$ predictor variables:

$$(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^{1\times(p-1)} \times \mathbb{R}$$

Viewing x_i as an element of $\mathbb{R}^{1\times p}$ via (1), define:

$$x := \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n\times p}, \quad y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n\times 1}$$

For $\beta \in \mathbb{R}^{p\times 1}$, consider the equation:

$$x\beta = y$$

Equivalently:

$$\beta_0 + \beta_1 x_{i,1} + \cdots \beta_{p-1} x_{i,p-1} = y_i, \quad i = 1, \dots, n.$$

Recall: The *column space of* x is the subspace $C(x)$ of $\mathbb{R}^{n\times 1}$ characterized by any of the following equivalent conditions:

- $C(x)$ is the set of all linear combinations of the columns of x
- $C(x) = \{x\beta : \beta \in \mathbb{R}^{p\times 1}\}$
- $C(x) = \{y \in \mathbb{R}^{n\times 1} : x\beta = y \text{ has a solution}\}$

$C(x)$ is also called the *image of* x.

Let $\widehat{y} \in \mathbb{R}^{n\times 1}$ be the vector characterized by any of the equivalent conditions:

- $\widehat{y} = \underset{z\in C(x)}{\operatorname{argmin}} \|z - y\|$
- $\widehat{y}$ is the vector in the column space of x closest to y.
- $\widehat{y}$ is the orthogonal projection of y onto the column space of x.

In partricular, $x\beta = \widehat{y}$ has a solution.

2. The case $\operatorname{rank}(x) = p$

Suppose $\operatorname{rank}(x) = p$. Then $\beta \mapsto x\beta$ maps $\mathbb{R}^{p\times 1}$ bijectively onto $C(x)$ and, therefore, there is a unique vector $\widehat{\beta} \in \mathbb{R}^{p\times 1}$ — the *least squares solution of* $x\beta = y$ — such that

$$x\widehat{\beta} = \widehat{y}.$$

The vector $\widehat{\beta}$ is characterized by the fact that it minimizes the sum of squared errors in approximating y by a vector of the form $x\beta$:

$$\widehat{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^{p\times 1}} \|x\beta - y\|^2$$

Since $\operatorname{rank}(x) = p$, the matrix $x^T x \in \mathbb{R}^{p\times p}$ is invertible and the system

$$x^T x\beta = x^T y$$

has unique solution; this solution is just $\widehat{\beta}$:

$$\widehat{\beta} = (x^T x)^{-1} x^T y$$

Thus,

$$\widehat{y} = x\widehat{\beta} = Py,$$

where

$$P := x(x^T x)^{-1} x^T.$$

The matrix P is called the *projection matrix* because it describes orthogonal projection from $\mathbb{R}^{n\times 1}$ onto $C(x)$.

If we view the y_i as realizations of random variable Y_i and let

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n, \end{bmatrix}$$

then we may view

$$\widehat{\beta} = \widehat{\beta}(Y_1, \ldots, Y_n) = (x^T x)^{-1} x^T Y$$

as an estimator.

Theorem 1. *Suppose* $\operatorname{rank}(x) = p$ *and*

$$Y \sim N(x\beta, \Sigma).$$

Then $\widehat{\beta}$ is an unbiased estimator of β.

Proof. Use the linearity of expectation:

$$\begin{aligned}\mathbb{E}\,\widehat{\beta} = \mathbb{E}\left[(x^T x)^{-1} x^T Y\right] &= (x^T x)^{-1} x^T\, \mathbb{E}\, Y \\ &= (x^T x)^{-1} x^T (x\beta) = (x^T x)^{-1}(x^T x)\beta = I\beta = \beta\end{aligned}$$ $\square$

$$\begin{aligned}
\operatorname{Var}\widehat{\beta} &= \operatorname{Var}(x^T x)^{-1}x^T Y \\
&= (x^T x)^{-1}x^T(\operatorname{Var}Y)((x^T x)^{-1}x^T)^T \\
&= (x^T x)^{-1}x^T(\sigma^2 I)x(x^T x)^{-1} \\
&= \sigma^2(x^T x)^{-1}(x^T x)(x^T x)^{-1} \\
&= \sigma^2(x^T x)^{-1}
\end{aligned}$$

3. The case $\operatorname{rank}(x) \le p$

We consider a *regularized* version of multiple linear regression. Let $\lambda > 0$ and consider the problem of minimizing

$$\mathrm{SSE}_\lambda(\beta) := \|x\beta - y\|^2 + \lambda^2\|\beta\|^2$$

Let

$$\xi \; := \begin{bmatrix} x \\ \lambda I^{p\times p} \end{bmatrix} \in \mathbb{R}^{(n+p)\times p}, \quad \eta := \begin{bmatrix} y \\ 0^{p\times 1} \end{bmatrix} \in \mathbb{R}^{(n+p)\times 1}$$

and consider the equation

$$\xi\beta = \eta.$$

The columns of ξ are linearly independent (why?), so $\operatorname{rank}(\xi) = p$. Therefore, by the discussion of the previous section, $\xi^T\xi$ is invertible and

$$\widehat{\beta}_\lambda := (\xi^T\xi)^{-1}\xi^T\eta$$

minimizes

$$\|\xi\beta - \eta\|^2 = \left\| \begin{bmatrix} x\beta - y \\ \lambda\beta \end{bmatrix} \right\|^2 = \|x\beta - y\|^2 + \lambda^2\|\beta\|^2 = \mathrm{SSE}_\lambda(\beta).$$

We have:

$$\begin{aligned}
\xi^T\xi &= \begin{bmatrix} x^T & \lambda I \end{bmatrix} \begin{bmatrix} x \\ \lambda I \end{bmatrix} \\
&= x^T x + \lambda^2 I, \\
\xi^T\eta &= \begin{bmatrix} x^T & \lambda I \end{bmatrix} \begin{bmatrix} y \\ 0^{p\times 1} \end{bmatrix} \\
&= x^T y
\end{aligned}$$

Therefore,

$$\widehat{\beta}_\lambda = (x^T x + \lambda^2 I)^{-1}x^T y.$$

Let

$$W_\lambda = (x^T x + \lambda^2 I)^{-1}x^T x.$$

Theorem 2. *Suppose* $\operatorname{rank}(x) = p$, *so that* $\widehat{\beta}$ *is defined. Then*

$$\beta_\lambda = W_\lambda\widehat{\beta}.$$

Proof. Just compute.

$$\begin{aligned}
W_\lambda \widehat{\beta} &= (x^T x + \lambda^2 I)^{-1} x^T x \widehat{\beta} \\
&= (x^T x + \lambda^2 I)^{-1} x^T x (x^T x)^{-1} x^T y \\
&= (x^T x + \lambda^2 I)^{-1} x^T y \\
&= \widehat{\beta}_\lambda.
\end{aligned}$$

□

$$E = (a + bx - y)^2 + \lambda(a^2 + b^2)$$

$$\begin{aligned}
E_x &= 0 & (1+\lambda)a + xb &= y \\
E_y &= 0 & xa + (x^2 + \lambda)b &= xy
\end{aligned}$$

Solution:

$$\widehat{a} = \frac{y}{1 + \lambda + x^2}, \qquad \widehat{b} = \frac{xy}{1 + \lambda + x^2}$$

$$\lambda = 0 : \qquad \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \frac{y}{1 + x^2} \begin{bmatrix} 1 \\ x \end{bmatrix} \in C\left(\begin{bmatrix} 1 & x \end{bmatrix}^T \right)$$

$$\begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \frac{y}{1 + x^2} \begin{bmatrix} 1 \\ x \end{bmatrix} \quad \text{is the least squares solution of} \quad \begin{bmatrix} 1 & x \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y \end{bmatrix}$$

4. The singular value decomposition

Theorem 3 (Singular Value Decomposotion, subspace formulation)**.** *Let $A : \mathbb{R}^n \to \mathbb{R}^m$ have rank r. Then there are orthonormal bases*

$$u_1, \dots, u_r \quad \text{of} \quad C(A^T) \subseteq \mathbb{R}^n$$

and

$$v_1, \dots, v_r \quad \text{of} \quad C(A) \subseteq \mathbb{R}^m$$

and numbers

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

such that

$$Au_i = \sigma_i v_i \quad \text{for} \quad i = 1, \dots, r.$$

The σ_i are uniquely determined by A.

Corollary 4 (Singular value decomposotion, basis formulation)**.** *Let $A \in \mathbb{R}^{m \times n}$ have rank r. Then there are orthogonal matrices*

$$U \in \mathbb{R}^{n \times n} \quad \text{and} \quad V \in \mathbb{R}^{m \times m}$$

and numbers

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

such that

$$A = V \Sigma U^T$$

where $\Sigma \in \mathbb{R}^{m \times n}$ *is defined by*

$$\Sigma_{ij} = \begin{cases} \sigma_i & \textit{if } i = j < r, \\ 0 & \textit{otherwise.} \end{cases} \tag{2}$$

The σ_i *are uniquely determined by* A.

Proof. By Theorem 3, there are orthonormal bases

$$u_1, \ldots, u_r \quad \text{of} \quad C(A^T) \subseteq \mathbb{R}^n$$

and

$$v_1, \ldots, v_r \quad \text{of} \quad C(A) \subseteq \mathbb{R}^m$$

and numbers

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

such that

$$Au_i = \sigma_i v_i \quad \text{for} \quad i = 1, \ldots, r. \tag{3}$$

Let

$$u_{r+1}, \ldots, u_n \quad \text{be an orthonormal basis of} \quad C(A^T)^\perp = N(A)$$

and let

$$v_{r+1}, \ldots, v_m \quad \text{be an orthonormal basis of} \quad C(A)^\perp = N(A^T).$$

Then $u_1, \ldots, u_r, u_{r+1}, \ldots, u_n$ and $v_1, \ldots, v_r, v_{r+1}, \ldots, v_m$ are orthonormal bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Equivalently,

$$U := \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} \in \mathbb{R}^{n \times n} \quad \text{and} \quad V := \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix} \in \mathbb{R}^{m \times m},$$

are orthogonal matrices. Define $\Sigma \in \mathbb{R}^{m \times n}$ by (2). Since U is orthogonal,

$$U^T u_i = U^{-1} u_i = e_i.$$

Therefore,

$$V \Sigma U^T u_i = V \Sigma e_i = V(\sigma_i e_i) = \sigma_i V e_i = \sigma_i v_i \overset{(3)}{=} Au_i,$$

for $i = 1, \ldots, r$ and

$$V \Sigma U^T u_i = V \Sigma e_i = V 0 = 0 = Au_i,$$

for $i = r+1, \ldots, n$. (We have $Au_i = 0$ for $i = r+1, \ldots, n$ because $u_i \in C(A^T)^\perp = N(A)$.) Since $u_1, \ldots, u_n$ is a basis of $\mathbb{R}^n$ and $Au_i = V \Sigma U^T u_i$ for all i, we must have $A = V \Sigma U^T$. $\square$

$$A = U \Sigma V^T \iff AV = U\sigma \iff Av_j = \sigma_j u_j$$

Definition 5. *The vector* v_j *is called the* j-th principal component *of* A.

Let $X_1, \ldots, X_p$ be n-dimensional random variables such that

$$\mathbb{E}[X_i] = 0$$

for all i. Let

$$X = \begin{bmatrix} X_1 & \cdots & X_p \end{bmatrix}$$

Then

$$K := \mathbb{E}\left[X^T X\right]$$

is the covariance matrix of the random *row* vector X.

$$\underset{a}{\operatorname{argmax}} \frac{\operatorname{var}(Xa)}{\|a\|^2} = \underset{a}{\operatorname{argmax}} \frac{a^T K a}{\|a\|^2} = \underset{\|a\|=1}{\operatorname{argmax}}\, a^T K a$$

5. Regularization

Let $x_1 > 0$. Suppose we want to fit a line of the form $y = \widehat{b}x$ to the one point data set, $\{(x_1, Y_1)\}$, where

$$Y_1 \sim N(bx_1, \sigma^2).$$

The parameters b and σ^2 are unknown. In this case, a natural estimator of b is

$$\widehat{b} = \widehat{b}(Y_1) = \frac{1}{x_1} Y_1.$$

Let $x_0 > 0$ and let $Y_0 \sim N(bx_0, \sigma^2)$ be independent of Y_1. We want to compute the expected prediction error

$$\text{EPE} := \mathbb{E}\left[\left(\widehat{b}(Y_1)x_0 - Y_0\right)^2\right]$$

$$\mathbb{E}\left[\left(\widehat{b}(Y_1)x_0 - Y_0\right)^2\right] = x_0^2\, \mathbb{E}\left[\widehat{b}(Y_1)^2\right] - 2x_0\, \mathbb{E}\left[\widehat{b}(Y_1) Y_0\right] + \mathbb{E}\left[Y_0^2\right]$$

$$\begin{aligned}
\mathbb{E}\left[\widehat{b}(Y_1)^2\right] &= \frac{1}{x_1^2}\, \mathbb{E}\left[Y_1^2\right] \\
&= \frac{1}{x_1^2}\left(\mathbb{E}\left[Y_1\right]^2 + \operatorname{Var} Y_1\right) \\
&= \frac{1}{x_1^2}\left(b^2 x_1^2 + \sigma^2\right) \\
&= b^2 + \frac{\sigma^2}{x_1^2}
\end{aligned}$$

Since Y_0 and Y_1 are independent,

$$\mathbb{E}\left[\widehat{b}(Y_1) Y_0\right] = \mathbb{E}\left[\widehat{b}(Y_1)\right] \mathbb{E}\left[Y_0\right] = b(bx_0) = b^2 x_0$$

$$\mathbb{E}\left[Y_0^2\right] = b^2 x_0^2 + \sigma^2$$

Therefore,

$$\begin{aligned}
\text{EPE} &= x_0^2\left(b^2 + \frac{\sigma^2}{x_1^2}\right) - 2b^2 x_0^2 + (b^2 x_0^2 + \sigma^2) \\
&= \sigma^2 + \frac{x_0^2 \sigma^2}{x_1^2}
\end{aligned}$$

$$
\begin{aligned}
\operatorname{Bias}\left(\widehat{b}(Y_1)x_0, bx_0\right) &= \mathbb{E}\left[\widehat{b}(Y_1)x_0\right] - bx_0 \\
&= bx_0 - bx_0 \\
&= 0
\end{aligned}
$$

and

$$
\begin{aligned}
\operatorname{Var}\left(\widehat{b}(Y_1)x_0\right) &= x_0^2 \operatorname{Var}\left(\frac{1}{x_1}Y_1\right) \\
&= \frac{x_0^2\sigma^2}{x_1^2}
\end{aligned}
$$

Thus, we recover the bias-variance decomposition:

$$
\mathrm{EPE} = \sigma^2 + \operatorname{Var}\left(\widehat{b}(Y_1)x_0\right) + \operatorname{Bias}\left(\widehat{b}(Y_1)x_0, bx_0\right)^2
$$

(Is $\widehat{b}$ an UMVU?)

Can we find an estimator, $\widetilde{b}(Y_1)$, of b, possibly biased, such that

$$
\mathbb{E}\left[(\widetilde{b}(Y_1)x_0 - Y_0)^2\right] < \mathbb{E}\left[(\widehat{b}(Y_1)x_0 - Y_0)^2\right]?
$$

For $\lambda \geq 0$, define

$$
\widehat{b}_\lambda := \frac{1}{x_1 + \lambda}Y_1
$$

$$
\mathbb{E}\left[(\widehat{b}_\lambda(Y_1)x_0 - Y_0)^2\right] = x_0^2\mathbb{E}\left[\widehat{b}_\lambda(Y_1)^2\right] - 2x_0\mathbb{E}\left[\widehat{b}_\lambda(Y_1)Y_0\right] + \mathbb{E}\left[Y_0^2\right]
$$

$$
\begin{aligned}
\mathbb{E}\left[\widehat{b}_\lambda(Y_1)^2\right] &= \frac{1}{(x_1+\lambda)^2}\mathbb{E}\left[Y_1^2\right] \\
&= \frac{1}{(x_1+\lambda)^2}\left(\mathbb{E}\left[Y_1\right]^2 + \operatorname{Var} Y_1\right) \\
&= \frac{1}{(x_1+\lambda)^2}\left(b^2x_1^2 + \sigma^2\right)
\end{aligned}
$$

Since Y_0 and Y_1 are independent,

$$
\begin{aligned}
\mathbb{E}\left[\widehat{b}_\lambda(Y_1)Y_0\right] &= \mathbb{E}\left[\widehat{b}_\lambda(Y_1)\right]\mathbb{E}\left[Y_0\right] \\
&= \frac{bx_1}{x_1+\lambda}bx_0 \\
&= \frac{b^2x_0x_1}{x_1+\lambda}
\end{aligned}
$$

$$
\mathbb{E}\left[Y_0^2\right] = b^2x_0^2 + \sigma^2
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\left[(\widehat{b}_\lambda(Y_1)x_0 - Y_0)^2\right] &= \frac{b^2x_0^2x_1^2 + x_0^2\sigma^2}{(x_1+\lambda)^2} - \frac{2b^2x_0^2x_1}{x_1+\lambda} + b^2x_0^2 + \sigma^2 \\
&= \sigma^2 + \frac{x_0^2}{(x_1+\lambda)^2}\big(b^2x_1^2 + \sigma^2 \\
&\qquad - 2b^2x_1^2 - 2b^2x_1\lambda + b^2x_1^2 + 2b^2x_1\lambda + b^2\lambda^2\big) \\
&= \sigma^2 + \frac{x_0^2}{(x_1+\lambda)^2}\left(\sigma^2 + b^2\lambda^2\right) \\
&= \sigma^2 + \frac{x_0^2\sigma^2}{(x_1+\lambda)^2} + \frac{b^2\lambda^2x_0^2}{(x_1+\lambda)^2}
\end{aligned}
$$

Double check:

$$
\begin{aligned}
\mathrm{Bias}\left(\widehat{b}_\lambda(Y_1)x_0, bx_0\right) &= \mathbb{E}\left[\widehat{b}_\lambda(Y_1)x_0\right] - bx_0 \\
&= \frac{bx_0x_1}{x_1+\lambda} - bx_0 \\
&= -\frac{b\lambda x_0}{x_1+\lambda}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{Var}\left(\widehat{b}_\lambda(Y_1)x_0\right) &= x_0^2\,\mathrm{Var}\left(\frac{1}{x_1+\lambda}Y_1\right) \\
&= \frac{x_0^2\sigma^2}{(x_1+\lambda)^2}
\end{aligned}
$$

So:

$$
\mathbb{E}\left[(\widehat{b}_\lambda x_0 - Y_0)^2\right] = \sigma^2 + \mathrm{Var}\left(\widehat{b}_\lambda x_0\right) + \mathrm{Bias}\left(\widehat{b}_\lambda x_0, bx_0\right)^2
$$

$$
\mathbb{E}\left[(\widehat{b}_\lambda x_0 - Y_0)^2\right] = \sigma^2 + x_0^2\sigma^2\frac{1+(\frac{b}{\sigma})^2\lambda^2}{(x_1+\lambda)^2}, \qquad \mathbb{E}\left[(\widehat{b}x_0 - Y_0)^2\right] = \sigma^2 + x_0^2\frac{\sigma^2}{x_1^2}
$$

Let $c = b/\sigma$. Then

$$
\begin{aligned}
\frac{1}{\sigma^2x_0^2}\left(\mathbb{E}\left[(\widehat{b}_\lambda x_0 - Y_0)^2\right] - \mathbb{E}\left[(\widehat{b}x_0 - Y_0)^2\right]\right) &= \frac{1+c^2\lambda^2}{(x_1+\lambda)^2} - \frac{1}{x_1^2} \\
&= \frac{(1+c^2\lambda^2)x_1^2 - (x_1+\lambda)^2}{x_1^2(x_1+\lambda)^2} \\
&= \lambda(c^2\lambda x_1^2 - 2x_1 - \lambda) \\
&= (c^2x_1^2 - 1)\lambda\left(\lambda - \frac{2x_1}{(c^2x_1^2-1)}\right)
\end{aligned}
$$

Let

$$
F(\lambda) = \mathbb{E}\left[(\widehat{b}_\lambda x_0 - Y_0)^2\right].
$$

$$
\begin{aligned}
F'(\lambda) &= x_0 \frac{d}{d\lambda} \frac{\sigma^2 + b^2\lambda^2}{(x_1+\lambda)^2} \\
&= x_0 \frac{2b^2\lambda(x_1+\lambda)^2 - (\sigma^2 + b^2\lambda^2)2(x_1+\lambda)}{(x_1+\lambda)^4} \\
&= 2x_0 \frac{b^2\lambda(x_1+\lambda) - (\sigma^2 + b^2\lambda^2)}{(x_1+\lambda)^3} \\
&= 2x_0 \frac{b^2 x_1 \lambda - \sigma^2}{(x_1+\lambda)^3}
\end{aligned}
$$

Since $x_0, x_1, b, \sigma^2 > 0$,

$$
\lambda^* := \operatorname*{argmin}_{\lambda \geq 0} \mathbb{E}\left[(\widehat{b}_\lambda x_0 - Y_0)^2\right] = \frac{\sigma^2}{b^2 x_1} > 0
$$

Moreover, since

$$
F(0) = \mathbb{E}\left[(\widehat{b} x_0 - Y_0)^2\right],
$$

we deduce that

$$
\mathrm{EPE}_\lambda < \mathrm{EPE}\,.
$$

6. k-fold cross validation

$D = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$, $x_i \in \mathbb{R}^p$, $Y_i \sim N(x_i\beta, \sigma^2)$ independent

Let $I_1, \ldots, I_k$ be a partition of $\{1, \ldots, n\}$ into k parts of about equal size.

For $i = 1, \ldots, n$ let

$$
D_i = \{(x_i, Y_i) : i \in I_i\}
$$

and let

$$
D_{-i} = D - D_i
$$

(x_i, Y_i), $i \in I_i$, from D:

$$
D_i = D \setminus \{(x_i, Y_i) : i \in I_i\}
$$

For $j = 1, \ldots, n$, let $\widehat{\theta}_{-i}$ be an estimator of β computed using the dataset D_{-i}. Let

$$
\mathrm{MSE}_i = \mathbb{E}\left[\frac{1}{n}\sum_{i \in I_i} (x_i \widehat{\theta}_{-i} - Y_i)^2\right]
$$

be the mean square error computed using the dataset D_i and set

$$
\mathrm{CV}_k = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i\,.
$$

We use CV_k as a proxy for prediction error. We tune any tuneable (hyper)parameters to minimize CV_k.

When $k = 1$, this is called *leave one out cross validation (LOOCV)*. In practice, k is often 5 or 10.

7. More ridge regression

Let $t > 0$. Suppose that x^* is a solution of

$$\text{minimize } \|x - b\|^2 \text{ subject to } \|x\|^2 = t^2.$$

and that x^{**} is a solution of

$$\text{minimize } \|x - b\|^2 \text{ subject to } \|x\|^2 \leq t^2.$$

For $r > 0$, let

$$S_r = \{x \in \mathbb{R}^n : \|x\| = r\}, \quad B_r = \{x \in \mathbb{R}^n : \|x\| < r\}.$$

Theorem 6. *Let $r > 0$ and define*

$$x^* = \operatorname*{argmin}_{x \in S_r} f(x), \quad x^{**} = \operatorname*{argmin}_{x \in B_r \cup S_r} f(x).$$

Let λ^* be the Lagrange multiplier such that

$$\nabla_{x,\lambda} f(x^*) + \lambda^* \nabla_{x,\lambda} g(x^*) = 0,$$

where $g(x) := \|x\|^2 - r^2$. Assume $\lambda^* \neq 0$. Then $x^* = x^{**}$ if and only if $\lambda^* > 0$.

Let

$$f(x) = \|Ax - b\|^2, \quad g(x) = \|x\|^2, \quad \text{and} \quad L(x, \lambda) = f(x) + \lambda g(x).$$

Then there is a Lagrange multiplier λ^* such that

$$\nabla_{x,\lambda} L(x^*, \lambda^*) = 0.$$

Thus,

$$\nabla_x f(x^*) = -\lambda \nabla_x g(x^*). \tag{4}$$

Now $\nabla_x f(x^*)$ points in the direction of steepest ascent of $f(x)$ at $x = x^*$ while $\nabla_x g(x^*)$ is an outward-pointing normal vector to the solution set of $g(x) = 0$, i.e., to the sphere of radius t centered at 0.

Suppose $\lambda^* > 0$. Since $\nabla_x g(x^*) \neq 0$ (why?), $\nabla_x f(x^*) \neq 0$. Moreover, $\nabla_x f(x^*)$ and $\nabla_x g(x^*)$ point in the opposite directions. Since the latter is an outward-pointing normal to the $g(x) = 0$, the former points inward. Thus, $f(x)$ increases as x moves into $B(0, t)$ from x^*. It follows that x^*

Claim: Suppose $x^* \neq x^{**}$. Then $\lambda^* < 0$.

(Equivalently, if $\lambda^* \geq 0$, then $x^* = x^{**}$.)

By the strict convexity of $f(x)$,

$$\begin{aligned}
f(x^* + \delta v) &= f(x^* + \delta(x^{**} - x^*)) \\
&= f((1 - \delta)x^* + \delta x^{**}) \\
&< (1 - \delta) f(x^*) + \delta f(x^{**}) \\
&\leq (1 - \delta) f(x^*) + \delta f(x^*) \\
&= f(x^*)
\end{aligned}$$

As $\nabla f \cdot v$ coincides with the directional derivative in the v-direction,

$$\nabla f(x^*) \cdot v = \lim_{\delta \downarrow 0} \frac{f(x^* + \delta v) - f(x^*)}{\delta} < 0.$$

It follows that $\nabla f(x^*)$ points outward from S_r. But so does $\nabla g(x^*)$. Therefore, by (4), we must have $\lambda^* < 0$.

Therefore, Let v be the vector pointing from x^* to x^{**} Since $f(x^{**}) \leq f(x^*)$. Since $f(x)$ is convex,

$$0 \geq f((t-1)x^*) - f(tx^{**}) \geq f((t-1))$$

Theorem 7. *Suppose*

$$x^* = \operatorname*{argmin}_x \left(\|Ax - b\| + \lambda \|x\|^2\right), \tag{5}$$

where $\lambda > 0$. *Then*

$$x^* = \operatorname*{argmin}_{\|x\| \leq \|x^*\|} \|Ax - b\|.$$

Conversely, suppose

$$\operatorname*{argmin}_{\|x\| \leq r} \|Ax - b\| = \operatorname*{argmin}_{\|x\| = r} \|Ax - b\| =: x^*;$$

Then there is a $\lambda^* > 0$ *such that*

$$x^* = \operatorname*{argmin}_x \left(\|Ax - b\| + \lambda^* \|x\|^2\right).$$

Proof. Suppose (5) holds for some $\lambda > 0$. Set

$$f(x) = \|Ax - b\|, \quad g(x) = \|x\|^2 - \|x^*\|^2, \quad \text{and} \quad h(x) = f(x) + \lambda g(x).$$

Then

$$\nabla_x h(x) = \nabla_x \left(\|Ax - b\| + \lambda \|x\|^2\right) = 0$$

□