10/36-702 Statistical Machine Learning Homework #2 Solutions

DUE: 3:00 PM February 22, 2019

Problem 1 [10 pts.]

Consider the data $(X_1, Y_1), \dots, (X_n, Y_n)$ where $X_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$. Inspired by the fact that $\mathbb{E}[Y|X = x] = \int yp(x,y)dy/p(x)$, define

$$\widehat{m}(x) = \frac{\int y\widehat{p}(x,y)dy}{\widehat{p}(x)}$$

where

$$\widehat{p}(x) = \frac{1}{n}\sum_i \frac{1}{h}K\left(\frac{X_i - x}{h}\right)$$

and

$$\widehat{p}(x,y) = \frac{1}{n}\sum_i \frac{1}{h^2}K\left(\frac{X_i - x}{h}\right)K\left(\frac{Y_i - y}{h}\right).$$

Assume that $\int K(u)du = 1$ and $\int uK(u)du = 0$. Show that $\widehat{m}(x)$ is exactly the kernel regression estimator that we defined in class.

Solution.

$$\begin{aligned}
\frac{\int y \cdot \widehat{p}(x,y)dy}{\widehat{p}(x)} &= \frac{\frac{1}{nh^2}\int y \sum K(\frac{x-X_i}{h})K(\frac{y-Y_i}{h})dy}{\frac{1}{nh}\sum K(\frac{x-X_i}{h})} \\
&= \frac{\sum K(\frac{x-X_i}{h}) \int y \frac{1}{h}K(\frac{y-Y_i}{h})dy}{\sum K(\frac{x-X_i}{h})} \\
&= \frac{\sum K(\frac{x-X_i}{h})Y_i}{\sum K(\frac{x-X_i}{h})} \\
&= \widehat{m}(x).
\end{aligned}$$

Problem 2 [15 pts.]

Suppose that (X, Y) is bivariate Normal:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \eta \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} \right).$$

(a) (5 pts.) Show that $m(x) = \mathbb{E}[Y|X = x] = \alpha + \beta x$ and find explicit expressions for α and β.

(b) (5 pts.) Find the maximum likelihood estimator $\widehat{m}(x) = \widehat{\alpha} + \widehat{\beta} x$.

(c) (5 pts.) Show that $|\widehat{m}(x) - m(x)|^2 = O_P(n^{-1})$.

Solution.

(a) Some simple calculations show

$$Y|X = x \sim N\left(\eta + \frac{\tau}{\sigma}\rho(x - \mu),\ (1 - \rho^2)\tau^2\right),$$

which gives

$$\alpha = \eta - \frac{\tau\rho\mu}{\sigma} \quad \text{and} \quad \beta = \frac{\tau\rho}{\sigma}.$$

(b) Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, the MLEs for the bivariate normal parameters are

$$\begin{aligned} \widehat{\mu} &= \overline{X} \\ \widehat{\eta} &= \overline{Y} \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ \widehat{\tau}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \\ \widehat{\text{Cov}}(X, Y) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}). \end{aligned}$$

Note $\beta = \frac{\tau\rho}{\sigma} = \frac{\tau\rho\sigma}{\sigma^2}$. Then by the equivariance property of the MLE,

$$\widehat{\beta} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\sigma}^2}$$

and

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X}.$$

Again by equivariance,

$$\widehat{m}(x) = \widehat{\alpha} + \widehat{\beta} x.$$

(c) $\widehat{m}(x)$ is an MLE and satisfies the regularity conditions for asymptotic normality. Therefore,

$$\sqrt{n}(\widehat{m}(x) - m(x)) \rightsquigarrow N(0, I^{-1}(m(x))),$$

which implies

$$\sqrt{n}|\widehat{m}(x) - m(x)| = O_p(1) \implies |\widehat{m}(x) - m(x)|^2 = O_p(n^{-1}).$$

Problem 3 [20 pts.]

Let $m(x) = \mathbb{E}[Y|X = x]$. Let $X \in [0,1]^d$. Divide $[0,1]^d$ into cubes $B_1, \dots, B_N$ whose sides have length h. The data are $(X_1, Y_1), \dots, (X_n, Y_n)$. In this problem, treat the X_i's as fixed. Assume that the number of observations in each bin is positive. Let

$$\widehat{m}(x) = \frac{1}{n(x)} \sum_i Y_i \mathbb{1}(X_i \in B(x))$$

where $B(x)$ is the cube containing x and $n(x) = \sum_i \mathbb{1}(X_i \in B(x))$. Assume that

$$|m(y) - m(x)| \le L\|x - y\|_2$$

for all x, y. You may further assume that $\sup_x \mathrm{Var}(Y|X = x) < \infty$.

(a) (10 pts.) Show that

$$|\mathbb{E}[\widehat{m}(x)] - m(x)| \le C_1 h$$

for some $C_1 > 0$. Also show that

$$\mathrm{Var}(\widehat{m}(x)) \le \frac{C_2}{n(x)}$$

for some $C_2 > 0$.

(b) (10 pts.) Now let X be random and assume that X has a uniform density on $[0,1]^d$. Let $h \equiv h_n = (C \log n/n)^{1/d}$. Show that, for $C > 0$ large enough, $P(\min n_j = 0) \to 0$ as $n \to \infty$ where n_j is the number of observations in cube B_j.

Solution.

(a) We have that X_i are fixed, so that $m(X_i) = Y_i$. Were they not, the below is still applicable by using the law of iterated expectation and the law of total variance.

$$\begin{aligned}
\Big|\mathbb{E}[\widehat{m}(x)] - m(x)\Big| &= \left|\mathbb{E}\left[\frac{1}{n(x)} \sum_i Y_i \mathbb{1}_{\{X_i \in B(x)\}}\right] - m(x)\right| \\
&= \left|\frac{1}{n(x)} \sum_i \Big(\mathbb{E}[Y_i] - m(x)\Big)\mathbb{1}_{\{X_i \in B(x)\}}\right| \\
&= \left|\frac{1}{n(x)} \sum_i \Big(m(X_i) - m(x)\Big)\mathbb{1}_{\{X_i \in B(x)\}}\right| \\
&\le \frac{1}{n(x)} \sum_i \Big|m(X_i) - m(x)\Big|\mathbb{1}_{\{X_i \in B(x)\}} \\
&\le \frac{1}{n(x)} \sum_i L\sqrt{d}h \cdot \mathbb{1}_{\{X_i \in B(x)\}} \\
&= L\sqrt{d}h
\end{aligned}$$

With the first upper bound due to triangular inequality and the second one because, given $x, y \in B_i$:

$$\|x - y\|_2^2 = \sum_{j=1}^{d} (x_j - y_j)^2 \le dh^2 \implies \|x - y\|_2 \le \sqrt{d}h$$

Let $\sup_x \text{Var}(Y|X=x) = M$.

$$\begin{aligned}\text{Var}(\widehat{m}(x)) &= \text{Var}\left(\frac{1}{n(x)}\sum_i Y_i \mathbb{1}_{\{X_i \in B(x)\}}\right)\\ &= \frac{1}{n^2(x)}\sum_i \text{Var}(Y_i)\mathbb{1}_{\{X_i \in B(x)\}}\\ &\le \frac{M}{n(x)}.\end{aligned}$$

(b)

$$\begin{aligned}P(\min_j n_j = 0) &= P\left(\bigcup_{j=1}^{B}\{n_j = 0\}\right)\\ &\le \sum_{j=1}^{B} P(n_j = 0)\\ &= \sum_{j=1}^{B}\prod_{i=1}^{n}\Big(1 - P(X_i \in B_j)\Big)\\ &= \frac{1}{h^d}(1-h^d)^n\\ &= \frac{n}{C\log n}\left(1 - \frac{C\log n}{n}\right)^n\end{aligned}$$

Since $B = \frac{1}{h^d}$.[1] Take $C = 1$. Then

$$\begin{aligned}\frac{n}{C\log n}\left(1 - \frac{C\log n}{n}\right)^n &< \frac{n}{C\log n}e^{-\frac{C\log n}{n}\cdot n}\\ &= \frac{n}{C\log n}n^{-C}\\ &= \frac{1}{C\log n}\\ &\to 0.\end{aligned}$$

[1] if we assume $1/h$ is an integer, otherwise we could use that as an upper bound.

Problem 4 [15 pts.]

Consider the RKHS problem

$$\widehat{f} = \text{argmin}_{f \in \mathcal{H}} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{1}$$

for some Mercer kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. In this problem, you will prove that the above problem is equivalent to the finite dimensional one

$$\widehat{\alpha} = \text{argmin}_{\alpha \in \mathbb{R}^n} \|y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha, \tag{2}$$

where $K \in \mathbb{R}^{n \times n}$ denotes the kernel matrix $K_{ij} = K(x_i, x_j)$.
Recall that the functions $K(\cdot, x_i)$, $i = 1, \ldots, n$ are called the *representers of evaluation*.
Recall that

- $\langle f, K(\cdot, x_i) \rangle_{\mathcal{H}} = f(x_i)$, for any function $f \in \mathcal{H}$
- $\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j)$ for any function $f = \sum_{i=1}^n \alpha_i K(\cdot, x_i)$.

(a) (5 pts.) Let $f = \sum_{i=1}^n \alpha_i K(\cdot, x_i)$, and consider defining a function $\widetilde{f} = f + \rho$, where ρ is any function orthogonal to $K(\cdot, x_i)$, $i = 1, \ldots, n$. Using the properties of the representers, prove that $\widetilde{f}(x_i) = f(x_i)$ for all $i = 1, \ldots, n$, and $\|\widetilde{f}\|_{\mathcal{H}}^2 \geq \|f\|_{\mathcal{H}}^2$.

(b) (10 pts.) Conclude from part (a) that in the infinite-dimensional problem (1), we need only consider functions of the form $f = \sum_{i=1}^n \alpha_i K(\cdot, x_i)$, and that this in turn reduces to (2).

Solution.

(a) Since $f, \widetilde{f} \in \mathcal{H}_K$, for all $i = 1, \ldots, n$

$$\begin{aligned}
\widetilde{f}(x_i) &= \langle \widetilde{f}, K(\cdot, x_i) \rangle_{\mathcal{H}_K} \\
&= \langle f, K(\cdot, x_i) \rangle_{\mathcal{H}_K} + \langle \rho, K(\cdot, x_i) \rangle_{\mathcal{H}_K} \\
&= \langle f, K(\cdot, x_i) \rangle_{\mathcal{H}_K} \\
&= f(x_i).
\end{aligned}$$

Also, because

$$\begin{aligned}
\langle \rho, f \rangle_{\mathcal{H}_K} &= \left\langle \rho, \sum_{i=1}^n \alpha_i K(\cdot, x_i) \right\rangle_{\mathcal{H}_K} \\
&= \sum_{i=1}^n \alpha_i \langle \rho, K(\cdot, x_i) \rangle_{\mathcal{H}_K} \\
&= 0,
\end{aligned}$$

we have,

$$\begin{aligned}
\|\widetilde{f}\|_{\mathcal{H}_K}^2 &= \langle f, f \rangle_{\mathcal{H}_K} + \langle \rho, \rho \rangle_{\mathcal{H}_K} + 2\langle \rho, f \rangle_{\mathcal{H}_K} \\
&= \|f\|_{\mathcal{H}_K}^2 + \|\rho\|_{\mathcal{H}_K}^2 \\
&\geq \|f\|_{\mathcal{H}_K}^2.
\end{aligned}$$

(b) For any $\widetilde{f} \in \mathcal{H}_K$, let $\widetilde{y} = (\widetilde{f}(x_1), \dots, \widetilde{f}(x_n))^T \in \mathbb{R}^n$. Let $f \in \mathcal{H}_K$ be $f = \sum_{i=1}^n \alpha_i K(\cdot, x_i)$, where $\alpha = K^{-1}\widetilde{y}$. Then

$$\begin{aligned}
\langle \widetilde{f} - f, K(\cdot, x_i)\rangle_{\mathcal{H}_K} &= \langle \widetilde{f}, K(\cdot, x_i)\rangle_{\mathcal{H}_K} - \sum_{j=1}^n \alpha_j \langle K(\cdot, x_j), K(\cdot, x_i)\rangle_{\mathcal{H}_K} \\
&= \widetilde{f}(x_i) - \sum_{j=1}^n \alpha_j K(x_i, x_j) \\
&= \widetilde{f}(x_i) - [K(K^{-1}\widetilde{y})]_i \\
&= \widetilde{f}(x_i) - \widetilde{f}(x_i) \\
&= 0.
\end{aligned}$$

Hence, $\widetilde{f} - f \perp K(\cdot, x_i)$ for all $i = 1, \dots, n$, and from (a), this implies $\widetilde{f}(x_i) = f(x_i)$ for all $i = 1, \dots, n$, and $\|\widetilde{f}\|_{\mathcal{H}_K}^2 \geq \|f\|_{\mathcal{H}_K}^2$, where equality holds if and only if $\widetilde{f} = f$. Therefore,

$$\sum_{i=1}^n \left(y_i - f(x_i)\right)^2 + \lambda \|f\|_{\mathcal{H}_K}^2 \leq \sum_{i=1}^n \left(y_i - \widetilde{f}(x_i)\right)^2 + \lambda \|\widetilde{f}\|_{\mathcal{H}_K}^2,$$

where equality holds if and only if $\widetilde{f} = f$. Hence if $\widetilde{f} = \operatorname{argmin}_{f \in \mathcal{H}_K} \sum_{i=1}^n \left(y_i - f(x_i)\right)^2 + \lambda \|f\|_{\mathcal{H}_K}^2$, then $\widetilde{f} = \sum_{i=1}^n \alpha_i K(\cdot, x_i)$ with $\alpha = K^{-1}\widetilde{y}$. So we only need to consider functions of the form $f = \sum_{i=1}^n \alpha_i K(\cdot, x_i)$. By plugging in, we have

$$\begin{aligned}
\sum_{i=1}^n \left(y_i - f(x_i)\right)^2 \lambda \|f\|_{\mathcal{H}_K}^2 &= \sum_{i=1}^n \left(y_i \sum_{j=1}^n \alpha_j K(x_i, x_j)\right)^2 + \lambda \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) \\
&= \|y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha.
\end{aligned}$$

Problem 5 [15 pts.]

Let $X = (X(1), \ldots, X(d)) \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. In the questions below, make any reasonable assumptions that you need but state your assumptions.

(a) (5 pts.) Prove that $\mathbb{E}(Y - m(X))^2$ is minimized by choosing $m(x) = \mathbb{E}(Y|X = x)$.

(b) (5 pts.) Find the function $m(x)$ that minimizes $\mathbb{E}|Y - m(X)|$. (You can assume that the conditional cdf $F(y|X = x)$ is continuous and strictly increasing, for every x.)

(c) (5 pts.) Prove that $\mathbb{E}(Y - \beta^T X)^2$ is minimized by choosing $\beta_* = B^{-1}\alpha$ where $B = \mathbb{E}(XX^T)$ and $\alpha = (\alpha_1, \ldots, \alpha_d)$ and $\alpha_j = \mathbb{E}(YX(j))$.

Solution.

(a) Let $g(x)$ be any function of x. Then

$$\begin{aligned}
\mathbb{E}(Y - g(X))^2 &= \mathbb{E}(Y - m(X) + m(X) - g(X))^2 \\
&= \mathbb{E}(Y - m(X))^2 + \mathbb{E}(m(X) - g(X))^2 + 2\mathbb{E}((Y - m(X))(m(X) - g(X))) \\
&\geq \mathbb{E}(Y - m(X))^2 + 2\mathbb{E}((Y - m(X))(m(X) - g(X))) \\
&= \mathbb{E}(Y - m(X))^2 + 2\mathbb{E}\mathbb{E}\left((Y - m(X))(m(X) - g(X)) \middle| X \right) \\
&= \mathbb{E}(Y - m(X))^2 + 2\mathbb{E}\left((\mathbb{E}(Y|X) - m(X))(m(X) - g(X)) \right) \\
&= \mathbb{E}(Y - m(X))^2 + 2\mathbb{E}\left((m(X) - m(X))(m(X) - g(X)) \right) \\
&= \mathbb{E}(Y - m(X))^2
\end{aligned}$$

(b) Let $g(x)$ be any function of x. Recall that

$$\mathbb{E}[|Y - g(X)|] = \mathbb{E}\{\mathbb{E}[|Y - g(X)| \, |X]\}.$$

The idea is to choose c such that $\mathbb{E}[\big|Y - c\big| \mid X = x]$ is minimized. Now define:

$$r(c) = \mathbb{E}[|Y - c| \mid X = x] = \int |y - c| p_{Y|X=x}(y) dy.$$

The function $h_y(c) = |y - c|$ is differentiable everywhere except when $y = c$. Thus for $c \neq y$

$$h_y'(c) = \begin{cases} 1 & c > y \\ -1 & c < y \end{cases} = \mathbb{1}(c > y) - \mathbb{1}(c < y).$$

Since Y is continuous and has a density function, $P(Y = c) = 0$. So to minimize $r(c)$ we can differentiate under the integral sign and set the derivative equal to 0 to obtain:

$$\begin{aligned}
r'(c) &= \int h_y'(c) p_{Y|X=x}(y) dy = \int_{-\infty}^{c} p_{Y|X=x}(y) dy - \int_{c}^{\infty} p_{Y|X=x}(y) dy \\
&= 2\int_{-\infty}^{c} p_{Y|X=x}(y) dy - 1 = 0 \\
&\iff \int_{-\infty}^{c} p_{Y|X=x}(y) dy = \frac{1}{2},
\end{aligned}$$

so that $c = m(x)$, which is the median of $p_{Y|X=x}(y)$. It is a minimum since $r'(c) < 0$ for $c < m(x)$ and $r'(c) > 0$ for $c > m(x)$. Since m minimizes $\mathbb{E}[|Y - c| \mid X = x]$ at every x for any g we get

$$\mathbb{E}[|Y - g(X)| - |Y - m(X)| \big| X = x] \geq 0$$

which implies

$$R(g) - R(m) = \mathbb{E}[|Y - g(X)| - |Y - m(X)|] = \mathbb{E}\{\mathbb{E}[|Y - g(X)| - Y - m(X)| \big| X]\} \geq 0.$$

(c) By setting the first derivative of the loss function equal to 0 we obtain:

$$\begin{aligned} &\frac{\partial R(\beta)}{\partial \beta} = 0 \\ \implies &\frac{\partial \mathbb{E}(Y - \beta^T X)^2}{\partial \beta} = 0 \\ \implies &\mathbb{E}\Big[-2X(Y - \beta^T X)\Big] = 0 \\ \implies &2B\beta - 2\alpha = 0 \\ \implies &\beta_* = B^{-1}\alpha, \end{aligned}$$

where we can exchange the derivative and expectation by the dominated convergence theorem. The loss function $R(\beta)$ is strictly convex so β_* is its unique minimum.

Problem 6 [25 pts.]

Consider the many Normal means problem where we observe $Y_i \sim N(\theta_i, 1)$ for $i = 1, \cdots, d$. Let $\widehat{\theta}$ minimize the penalized loss

$$\sum_i (Y_i - \theta_i)^2 + \lambda J(\theta).$$

Find an explicit form for $\widehat{\theta}$ in three cases: (i) (10 pts.) $J(\beta) = \|\theta\|_0$, (ii) (10 pts.) $J(\beta) = \|\theta\|_1$ and (iii) (5 pts.) $J(\beta) = \|\theta\|_2^2$.

Solution.

(i) Note that

$$\sum_i (Y_i - \theta_i)^2 + \lambda \|\theta\|_0 = \sum_{j=1}^{d} \Bigg((Y_i - \theta_i)^2 + \lambda \mathbb{1}(\theta_i \neq 0) \Bigg).$$

Then for each term i,

$$\begin{aligned} (Y_i - \theta_i)^2 + \lambda \mathbb{1}(\theta_i \neq 0) &\geq Y_i^2 \mathbb{1}(\theta_i = 0) + \lambda \mathbb{1}(\theta_i \neq 0) \\ &\geq \min \left\{ Y_i^2, \lambda \right\} \end{aligned}$$

and equality holds if and only if

$$\widehat{\theta_i} = \begin{cases} 0 & \text{if } Y_i^2 < \lambda \\ 0 \text{ or } Y_i & \text{if } Y_i^2 = \lambda \\ Y_i & \text{if } Y_i^2 > \lambda. \end{cases}$$

Hence

$$\begin{aligned} \sum_i (Y_i - \theta_i)^2 + \lambda \|\theta\|_0 &= \sum_{j=1}^{d} \Bigg((Y_i - \theta_i)^2 + \lambda \mathbb{1}(\theta_i \neq 0) \Bigg) \\ &\geq \sum_{j=1}^{d} \min \left\{ Y_i^2, \lambda \right\} \end{aligned}$$

and equality holds if and only if

$$\widehat{\theta_i} = \begin{cases} 0 & \text{if } |Y_i| < \sqrt{\lambda} \\ 0 \text{ or } Y_i & \text{if } |Y_i| = \sqrt{\lambda} \\ Y_i & \text{if } |Y_i| > \sqrt{\lambda}. \end{cases}$$

(ii) First write

$$\min_\theta \sum_i (Y_i - \theta_i)^2 + \lambda\|\theta\|_1 = \min_\theta \sum_i \left(-2Y_i\theta_i + \theta_i^2 + \lambda|\theta_i|\right).$$

Now note it is simply equivalent to

$$\begin{aligned} &\min_{\theta_i} -2Y_i\theta_i + \theta_i^2 + \lambda|\theta_i| \\ \Longleftrightarrow &\min_{\theta_i} -2\widehat{\theta}_i^{OLS}\theta_i + \theta_i^2 + \lambda|\theta_i| \end{aligned}$$

for all $i = 1, \dots, d$.

When $\widehat{\theta}_i^{OLS} \geq 0$, then $\widehat{\theta}_i \geq 0$ so

$$-2\widehat{\theta}_i^{OLS}\theta_i + \theta_i^2 + \lambda|\theta_i| = -2\widehat{\theta}_i^{OLS}\theta_i + \theta_i^2 + \lambda\theta_i.$$

Differentiating with respect to θ_i, setting equal to zero, and solving gives

$$\widehat{\theta}_i = \left(\widehat{\theta}_i^{OLS} - \frac{\lambda}{2}\right)\mathbb{1}_{\widehat{\theta}_i^{OLS} \geq \frac{\lambda}{2}\}}.$$

When $\widehat{\theta}_i^{OLS} \leq 0$, the analogous steps give

$$\widehat{\theta}_i = \left(\widehat{\theta}_i^{OLS} + \frac{\lambda}{2}\right)\mathbb{1}_{\{\widehat{\theta}_i^{OLS} \leq -\frac{\lambda}{2}\}},$$

Putting them together gives

$$\widehat{\theta}_i = \begin{cases} \widehat{\theta}_i^{OLS} - \frac{\lambda}{2} & \widehat{\theta}_i^{OLS} \geq \frac{\lambda}{2} \\ 0 & \widehat{\theta}_i^{OLS} \in \left(-\frac{\lambda}{2}, \frac{\lambda}{2}\right) \\ \widehat{\theta}_i^{OLS} + \frac{\lambda}{2} & \widehat{\theta}_i^{OLS} \leq \frac{\lambda}{2} \end{cases}.$$

(iii) Here the objective function is differentiable everywhere. Taking the gradient w.r.t. θ we have

$$\nabla_\theta \left(\sum_i (Y_i - \theta_i)^2 + \lambda\|\theta\|_2^2 \right) = \sum_i \left(-2Y_i\theta_i + 2\lambda\theta_i\right).$$

Setting this equal to 0 and solving for θ gives

$$\widehat{\theta}_i = \frac{Y_i}{1+\lambda}. \tag{3}$$

Since the objective is strictly convex, (3) is the unique solution.