Quiz #1: Kernel Methods for Machine Learning

Problem 1

Given data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, ridge regression solves: for some $\lambda \geq 0$,

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \lambda \|\boldsymbol{\beta}\|^2 .$$

(1) Why is λ important?

(2) What happens if λ is too small or too large?

(3) In practice, how would you choose the value of λ?

Solutions:

(1) The regularization parameter λ controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage on the coefficients toward zero. When there exist many correlated variables in a linear regression model (which typically happen when $n < p$), their coefficients can become poorly determined and exhibit high variance without any regularization. Usually this may lead to poor generalization on test data, a phenomenon typically known as over-fitting. By imposing a size constraint such as ridge on the coefficients during training, this problem can be alleviated.

(2) If λ is too small (and returns least-squares estimator when $\lambda = 0$ at an extreme), estimated coefficients can exhibit high variance due to overfitting, leading to poor generalization on test data. If λ is too large (and returns a zero estimator when $\lambda = +\infty$ at the other extreme), estimated coefficients can exhibit high bias, also leading to poor generalization on test data.

(3) Choosing λ during training is typically known as parameter tuning or model selection, as an attempt to improve generalization on unseen data by trading off bias and variance of the estimated coefficients. A practical approach is cross-validation: for a predetermined list of λ's, pick the one that gives best cross-validated prediction error.

Problem 2

Given data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, the ridge regression with an intercept solves: for some $\lambda \geq 0$,

$$\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \lambda \|\boldsymbol{\beta}\|^2 .$$

(1) Find the optimal solutions $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) \in \mathbb{R}^{p+1}$ that solve this problem.

(2) How could you solve this problem, suppose you already have a solver for ridge regression without intercept?

<u>**Solutions**</u>:

(1) Let us denote by $\ell(\beta_0, \boldsymbol{\beta})$ the objective function. By taking partial derivatives over the variables and setting them to zero, we have:

$$\frac{\partial \ell}{\partial \beta_0} = -\frac{2}{n} \sum_{i=1}^n (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i) \;=\; 0 , \tag{1}$$

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\frac{2}{n} \sum_{i=1}^n (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)\mathbf{x}_i + 2\lambda \boldsymbol{\beta} \;=\; 0 . \tag{2}$$

Denote in matrix form by $\mathbf{X} := (\mathbf{x}_1 | \ldots | \mathbf{x}_n)^\top \in \mathbb{R}^{n \times p}$ the design matrix, $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ the response vector, $\mathbf{1} := (1, \ldots, 1)^\top \in \mathbb{R}^n$ the vector of 1's of length n. (1) and (2) are equivalent to:

$$\beta_0 \;=\; \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{1} , \tag{3}$$

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} + \lambda n \boldsymbol{\beta} \;=\; \mathbf{X}^\top \mathbf{y} - \beta_0 \mathbf{X}^\top \mathbf{1} . \tag{4}$$

Plug (3) into (4) and we get:

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} + \lambda n \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \left(\frac{1}{n} \mathbf{1}\mathbf{1}^\top \right) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) .$$

Denote by $\mathbf{I}$ the n-dimensional identity matrix, and $\mathbf{J} := \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{n \times n}$, we have:

$$(\mathbf{X}^\top \mathbf{J} \mathbf{X} + \lambda n \mathbf{I}) \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{J} \mathbf{y} ,$$

which gives the solution to

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{J} \mathbf{X} + \lambda n \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{J} \mathbf{y} . \tag{5}$$

Plugging $\hat{\boldsymbol{\beta}}$ into (3), we get:

$$\hat{\beta}_0 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{1} . \tag{6}$$

Optional: Suppose you would like to find $\hat{\beta}_0$ directly. Starting with (4), we have:

$$\boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^\top(\mathbf{y} - \beta_0\mathbf{1})\,.$$

Plug it into (3) and some mathematical deductions give the solution to:

$$\begin{aligned}\hat{\beta}_0 &= \frac{\mathbf{1}^\top\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^\top\right)\mathbf{y}}{\mathbf{1}^\top\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^\top\right)\mathbf{1}} \\ &= \frac{\mathbf{1}^\top(\mathbf{X}\mathbf{X}^\top + \lambda n\mathbf{I})^{-1}\mathbf{y}}{\mathbf{1}^\top(\mathbf{X}\mathbf{X}^\top + \lambda n\mathbf{I})^{-1}\mathbf{1}}\,.\end{aligned} \tag{7}$$

Note that we have also used equality $\mathbf{1}^\top\mathbf{1} = n$ in the deduction, and the second equality is due to the matrix inversion lemma.

(2) It is easy to verify that $\mathbf{J}^\top = \mathbf{J}$ and $\mathbf{J}^2 = \mathbf{J}$. Therefore, if we further define centered data:

$$\tilde{\mathbf{X}} := \mathbf{J}\mathbf{X}, \quad \tilde{\mathbf{y}} := \mathbf{J}\mathbf{y}\,,$$

we have that the estimated coefficients (5) can be written as:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \lambda n\mathbf{I})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}}\,.$$

This identifies the form of the estimated coefficients given by ridge regression without an intercept.

In words, the estimated coefficients of data $(\mathbf{X}, \mathbf{y})$ using ridge regression with an intercept is the same as the estimated coefficients of centered data $(\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$ using ridge regression without an intercept.

Optional: Alternatively, we could estimate the intercept $\hat{\beta}_0$ by (7). Define $z_i = y_i - \hat{\beta}_0$, thus we get an equivalent ridge problem without intercept:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n (z_i - \boldsymbol{\beta}^\top\mathbf{x}_i)^2 + \lambda\|\boldsymbol{\beta}\|^2\,.$$

Problem 2*

(1) The Gaussian density in $\mathbb{R}$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+$ is:

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,.$$

Given a set of data points $x_1, \ldots, x_n \in \mathbb{R}$, compute the log-likelihood

$$\ell(\mu, \sigma^2) = \sum_{i=1}^n \log p_{\mu,\sigma^2}(x_i)\,,$$

and find the maximum likelihood estimates of the parameters by solving:

$$(\hat{\mu}, \widehat{\sigma^2}) := \operatorname*{arg\,max}_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+} \ell(\mu, \sigma^2)\,.$$

(2) The Gaussian density in $\mathbb{R}^p$ with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and a symmetric positive-definite matrix $\boldsymbol{\Omega} \in \mathbb{R}_+^{p\times p}$, known as the precision matrix, is:

$$p_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{x}) = \sqrt{\frac{\det(\boldsymbol{\Omega})}{(2\pi)^p}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}(\mathbf{x}-\boldsymbol{\mu})\right\} .$$

Given a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, compute the log-likelihood

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Omega}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{x}_i) ,$$

and find the maximum likelihood estimates of the parameters by solving:

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}}) := \underset{\boldsymbol{\mu}\in\mathbb{R}^p, \boldsymbol{\Omega}\in\mathbb{R}_+^{p\times p}}{\arg\max} \ \ell(\boldsymbol{\mu}, \boldsymbol{\Omega}) . \tag{8}$$

(*Hint: for any vector* $\mathbf{u} \in \mathbb{R}^p$ *and any matrix* $\mathbf{C} \in \mathbb{R}^{p\times p}$*, you may try to find a matrix* $\mathbf{V} \in \mathbb{R}^{p\times p}$ *such that* $\mathbf{u}^\top \mathbf{C}\mathbf{u} = \operatorname{tr}(\mathbf{C}\mathbf{V})$.)

(3) Have you noticed a problem when solving (8) if $n < p$? How could you fix it?

Solutions:

(1) By definition,

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \log p_{\mu,\sigma^2}(x_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 .$$

In order to find the maximum likelihood estimates (MLE), let us take the partial derivatives over both parameters and setting them to zero:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \quad = \quad 0 , \tag{9}$$

$$\frac{\partial \ell}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \quad = \quad 0 . \tag{10}$$

From (9), we have the MLE:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Plugging $\hat{\mu}$ into (10), we have the MLE:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 .$$

(2) By definition,

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\Omega}) &= \sum_{i=1}^{n} \log p_{\boldsymbol{\mu},\boldsymbol{\Omega}}(\mathbf{x}_i) \\ &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det(\boldsymbol{\Omega}) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Omega}(\mathbf{x}_i - \boldsymbol{\mu})\,.\end{aligned}$$

Let us first take the derivative over $\boldsymbol{\mu}$ and set it to zero:

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n} \boldsymbol{\Omega}(\mathbf{x}_i - \boldsymbol{\mu}) = 0\,.$$

Since $\Omega \in \mathbb{R}_+^{p\times p}$ is always invertible, we get the MLE:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\,.$$

Plugging this into the data log-likelihood ℓ, and denote the sample covariance matrix by

$$\mathbf{S} := \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top\,, \tag{11}$$

we get

$$\begin{aligned}\ell(\hat{\boldsymbol{\mu}}, \boldsymbol{\Omega}) &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det(\boldsymbol{\Omega}) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \boldsymbol{\Omega}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) \\ &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det(\boldsymbol{\Omega}) - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left((\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \boldsymbol{\Omega}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})\right) \\ &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det(\boldsymbol{\Omega}) - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left(\boldsymbol{\Omega}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top\right) \\ &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det(\boldsymbol{\Omega}) - \frac{n}{2}\operatorname{tr}\left(\boldsymbol{\Omega}\mathbf{S}\right)\,,\end{aligned}$$

where the first equality is due to $a = \operatorname{tr}(a)$ for any scalar a. Taking the derivative over $\boldsymbol{\Omega}$ and setting it to zero, we get:

$$\left.\frac{\partial \ell}{\partial \boldsymbol{\Omega}}\right|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}} = \frac{n}{2}\boldsymbol{\Omega}^{-1} - \frac{n}{2}\mathbf{S} = 0\,. \tag{12}$$

In order to find a solution to this equation, $\mathbf{S}$ needs to be invertible, and thus we have MLE:

$$\hat{\boldsymbol{\Omega}} = \mathbf{S}^{-1}\,,$$

where $\mathbf{S}$ is defined in (11).

(3) If $n < p$, the sample covariance matrix $\mathbf{S}$ is not invertible. In order to fix this problem, we could resort to regularize the maximum likelihood problem. For example, we could add a trace-norm regularization to $\boldsymbol{\Omega}$ when solving (8): for some $\lambda > 0$,

$$\max_{\boldsymbol{\mu}\in\mathbb{R}^p, \boldsymbol{\Omega}\in\mathbb{R}_+^{p\times p}} \tilde{\ell}(\boldsymbol{\mu}, \boldsymbol{\Omega}) := \ell(\boldsymbol{\mu}, \boldsymbol{\Omega}) - \lambda \operatorname{tr}(\boldsymbol{\Omega}).$$

Following similar deduction, (12) now becomes

$$\left.\frac{\partial \tilde{\ell}}{\partial \boldsymbol{\Omega}}\right|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}} = \frac{n}{2}\boldsymbol{\Omega}^{-1} - \frac{n}{2}\mathbf{S} - \lambda\mathbf{I} = 0,$$

which always has a solution:

$$\hat{\boldsymbol{\Omega}} = \left(\mathbf{S} + \frac{2\lambda}{n}\mathbf{I}\right)^{-1}.$$

Problem 3

Definition. *Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, the Fenchel dual of f is the function $f^* : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$f^*(\mathbf{z}) = \max_{\mathbf{x}\in\mathbb{R}^n} \left[\mathbf{z}^\top \mathbf{x} - f(\mathbf{x})\right].$$

Given a $n \times p$ matrix $\mathbf{X}$, a convex function $R : \mathbb{R}^n \to \mathbb{R}$, $\lambda \geq 0$, let us consider an L_2-regularized optimization problem of the form:

$$\min_{\mathbf{w}\in\mathbb{R}^p} R(\mathbf{X}\mathbf{w}) + \lambda\|\mathbf{w}\|^2. \tag{13}$$

Our goal is to derive a dual problem of (13). To this end, let us rewrite (13) equivalently as:

$$\begin{array}{rl} \min_{\mathbf{w}\in\mathbb{R}^p, \mathbf{u}\in\mathbb{R}^n} & R(\mathbf{u}) + \lambda\|\mathbf{w}\|^2 \\ \text{s.t.} & \mathbf{X}\mathbf{w} = \mathbf{u}. \end{array} \tag{14}$$

(1) Show that the Lagrangian of (14) is:

$$\mathcal{L}(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}) = R(\mathbf{u}) + \lambda\|\mathbf{w}\|^2 + \boldsymbol{\alpha}^\top(\mathbf{X}\mathbf{w} - \mathbf{u}),$$

where $\boldsymbol{\alpha} \in \mathbb{R}^n$ is a vector of Lagrange multipliers.

(2) Find an expression of the Lagrange dual function $q(\boldsymbol{\alpha}) = \min_{\mathbf{w},\mathbf{u}} \mathcal{L}(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha})$ using the Fenchel dual R^*.

(3) If $R(\mathbf{u}) = \sum_{i=1}^n \ell_i(u_i)$, show that $R^*(\boldsymbol{\alpha}) = \sum_{i=1}^n \ell_i^*(\alpha_i)$.

(4) Application to ridge regression and ridge logistic regression. Derive a dual problem for the ridge regression: given $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$,

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \lambda \|\boldsymbol{\beta}\|^2 \,,$$

and for ridge logistic regression: given $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \{-1, +1\}$,

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \log(1 + e^{-y_i \boldsymbol{\beta}^\top \mathbf{x}_i}) + \lambda \|\boldsymbol{\beta}\|^2 \,.$$

<u>Solutions</u>:

(1) There are n equality constraints in the optimization problem (14), and each of them has a Lagrange multiplier in the Lagrangian, denoted by $\alpha_i, i = 1, \ldots, n$. Collecting them into a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$, the Lagrangian of (14) is indeed $\mathcal{L}$ by definition.

(2) By definition, we have

$$\begin{aligned} q(\boldsymbol{\alpha}) &= \min_{\mathbf{u}} \left[R(\mathbf{u}) - \boldsymbol{\alpha}^\top \mathbf{u}\right] + \min_{\mathbf{w}} \left[\lambda \|\mathbf{w}\|^2 + \boldsymbol{\alpha}^\top \mathbf{X}\mathbf{w}\right] \\ &= -R^*(\boldsymbol{\alpha}) - \frac{1}{4\lambda} \boldsymbol{\alpha}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{\alpha} \,. \end{aligned} \tag{15}$$

Note that, for the first minimization over $\mathbf{u}$ we have used the property that $\min g = -\max[-g]$ and the definition of the Fenchel dual of R, and the minimum of the second minimization over $\mathbf{w}$ is attained at $\hat{\mathbf{w}} = -\frac{1}{2\lambda} \mathbf{X}^\top \boldsymbol{\alpha}$.

(3) By definition of the Fenchel dual and the special form of $R(\mathbf{u}) = \sum_{i=1}^n \ell_i(u_i)$, we have

$$\begin{aligned} R^*(\boldsymbol{\alpha}) &= \max_{\mathbf{u} \in \mathbb{R}^n} \left[\boldsymbol{\alpha}^\top \mathbf{u} - R(\mathbf{u})\right] \\ &= \max_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^n \left[\alpha_i u_i - \ell_i(u_i)\right] \\ &= \sum_{i=1}^n \max_{u_i \in \mathbb{R}} \left[\alpha_i u_i - \ell_i(u_i)\right] \\ &= \sum_{i=1}^n \ell_i^*(\alpha_i) \,. \end{aligned} \tag{16}$$

(4) **Application to ridge regression.** Now $R(\mathbf{u}) = \sum_{i=1}^{n} \frac{1}{n}(y_i - u_i)^2$. Let us first derive the Fenchel dual of $\ell_i(u_i) = \frac{1}{n}(u_i - y_i)^2$. By definition we have

$$\ell_i^*(\alpha_i) = \max_{u_i \in \mathbb{R}} \left[\alpha_i u_i - \frac{1}{n}(u_i - y_i)^2 \right] = \frac{n}{4}\alpha_i^2 + \alpha_i y_i \,,$$

where the maximum is attained at $\hat{u}_i = \frac{n}{2}\alpha_i + y_i$. By (16), we have

$$R^*(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \ell_i^*(\alpha_i) = \sum_{i=1}^{n} \left(\frac{n}{4}\alpha_i^2 + \alpha_i y_i \right) = \frac{n}{4}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \mathbf{y}^\top \boldsymbol{\alpha} \,.$$

By (15), we have the Lagrange dual function to ridge regression:

$$q(\boldsymbol{\alpha}) = -\frac{n}{4}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \mathbf{y}^\top \boldsymbol{\alpha} - \frac{1}{4\lambda}\boldsymbol{\alpha}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{\alpha} \,.$$

Therefore, a dual problem to ridge regression is:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} q(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \left[-\frac{1}{4\lambda}\boldsymbol{\alpha}^\top (\mathbf{X}\mathbf{X}^\top + \lambda n \mathbf{I})\boldsymbol{\alpha} - \mathbf{y}^\top \boldsymbol{\alpha} \right] \,.$$

Application to ridge logistic regression. Now $R(\mathbf{u}) = \sum_{i=1}^{n} \frac{1}{n} \log(1+e^{-y_i u_i})$, and $\ell_i(u_i) = \frac{1}{n}\log(1 + e^{-y_i u_i})$. We have

$$\begin{aligned}
\ell_i^*(\alpha_i) &= \max_{u_i \in \mathbb{R}} \left[\alpha_i u_i - \frac{1}{n}\log(1 + e^{-y_i u_i}) \right] \\
&= \frac{1}{n} \max_{u_i \in \mathbb{R}} \left[\log \frac{e^{n\alpha_i u_i}}{1 + e^{-y_i u_i}} \right] \\
&= \frac{1}{n} \max_{u_i \in \mathbb{R}} \left[\log \frac{1}{e^{-n\alpha_i u_i} + e^{-(y_i + n\alpha_i) u_i}} \right] \\
&= -\frac{1}{n} \min_{u_i \in \mathbb{R}} \log \left[e^{-n\alpha_i u_i} + e^{-(y_i + n\alpha_i) u_i} \right] \\
&= -\frac{1}{n} \log \left[\min_{u_i \in \mathbb{R}} \left[e^{-n\alpha_i u_i} + e^{-(y_i + n\alpha_i) u_i} \right] \right] \\
&= -\frac{1}{n} \log \left[\min_{u_i \in \mathbb{R}} \left[e^{(y_i u_i)\cdot(-n y_i \alpha_i)} + e^{(y_i u_i)\cdot(-n y_i \alpha_i - 1)} \right] \right] \\
&= -\frac{1}{n} \log \left[\min_{t_i \in \mathbb{R}_+} \left[t_i^{p_i} + t_i^{p_i - 1} \right] \right] \,,
\end{aligned}$$

where we have changed the optimization variable from $u_i \in \mathbb{R}$ to $t_i = e^{y_i u_i} \in \mathbb{R}_+$, and denoted

$$p_i := -n y_i \alpha_i \,. \tag{17}$$

Note that in the deduction, we have frequently used the fact that $y_i^2 = 1$ since $y_i \in \{-1, +1\}$, and $\max(-g) = -\min g$, and that $\exp(\cdot)$ and $\log(\cdot)$ are monotonically increasing functions.

<u>Claim:</u> for any a that is not a function of x, we have

$$\min_{x>0} \left[x^a + x^{a-1} \right] = \begin{cases} \frac{1}{a^a (1-a)^{1-a}} & \text{if } 0 < a < 1\,, \\ 0 & \text{otherwise.} \end{cases}$$

The claim can be easily verified. Using the claim,

$$\ell_i^*(\alpha_i) = \begin{cases} \frac{1}{n} \left(p_i \log p_i + (1-p_i) \log(1-p_i) \right) & \text{if } 0 < p_i < 1\,, \\ +\infty & \text{otherwise.} \end{cases}$$

By (16), we have

$$R^*(\boldsymbol{\alpha}) = \begin{cases} \frac{1}{n} \sum_{i=1}^n \left(p_i \log p_i + (1-p_i) \log(1-p_i) \right) & \text{if } 0 < p_i < 1\,,\ i = 1, \ldots, n\,, \\ +\infty & \text{otherwise.} \end{cases}$$

By (15) and plugging (17) back in, we have a dual problem to ridge logistic regression:

$$\begin{array}{rl} \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} & -\frac{1}{4\lambda} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j \\ & \quad -\frac{1}{n} \sum_{i=1}^n \left(-n y_i \alpha_i \log(-n y_i \alpha_i) + (1 + n y_i \alpha_i) \log(1 + n y_i \alpha_i) \right) \\ \text{s.t.} & -\frac{1}{n} < y_i \alpha_i < 0\,,\ i = 1, \ldots, n, \end{array}$$

where $\mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$, are the row vectors of $\mathbf{X} \in \mathbb{R}^{n \times p}$.