CS 189:289A Introduction to Machine Learning
Fall 2021 Jennifer Listgarten and Jitendra Malik — HW3

Due 10/11 at 11:59pm

- We prefer that you typeset your answers using LaTeX or other word processing software. If you haven't yet learned LaTeX, one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted for the written questions.
- In all of the questions, **show your work**, not just the final answer.

Deliverables:

1. Submit a PDF of your homework to the Gradescope assignment entitled "HW3 Write-Up". **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.
   - In your write-up, please state with whom you worked on the homework. This should be on its own page and should be the first page that you submit.
   - In your write-up, please copy the following statement and sign your signature next to it. (Mac Preview and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) We want to make it *extra* clear so that no one inadvertently cheats. *"I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."*

1 Poisson Classification

Recall that the PDF of a Poisson random variable is

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x \in \{0, 1, \dots, \infty\}$$

The PDF is defined for non-negative integral values.

You are given two classes ω_1, ω_2 of Poisson data with parameters λ_1 and λ_2. This means that $x|\omega_1 \sim \text{Poisson}(\lambda_1)$ and $x|\omega_2 \sim \text{Poisson}(\lambda_2)$. Assume that $P(\omega_1) = P(\omega_2) = \frac{1}{2}$.

(a) Find $P(\omega_1|x)$ in terms of λ_1 and λ_2. What type of function is the posterior?

Solution: We use Bayes' Rule

$$\begin{aligned} P(\omega_1|x) &= \frac{P(x|\omega_1)P(\omega_1)}{P(x|\omega_1)P(\omega_1) + P(x|\omega_2)P(\omega_2)} \\ &= \frac{e^{-\lambda_1}\frac{\lambda_1^x}{x!}}{e^{-\lambda_1}\frac{\lambda_1^x}{x!} + e^{-\lambda_2}\frac{\lambda_2^x}{x!}} \\ &= \frac{1}{1 + e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_2}{\lambda_1}\right)^x} \\ &= \frac{1}{1 + e^{-(\lambda_1+\lambda_2)+x(\ln(\lambda_2)-\ln(\lambda_1))}} \end{aligned}$$

The posterior is a logistic function.

(b) Find the optimal rule (decision boundary) for allocating an observation x to a particular class. In the case where $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ and $\lambda_1 = 10$ and $\lambda_2 = 15$, calculate the decision boundary, probability of correct classification for each class, and total error rate.

Solution: The decision boundary is the value for x for which

$$1 < \frac{P(\omega_1|x)}{P(\omega_2|x)} = e^{\lambda_2-\lambda_1}\left(\frac{\lambda_1}{\lambda_2}\right)^x$$

so we should choose class 1 if

$$\boxed{x < \frac{\lambda_1 - \lambda_2}{\ln\lambda_1 - \ln\lambda_2} \approx 12.3}$$

or class 2 otherwise (the inequality holds since $\lambda_1/\lambda_2 < 1$).

Recall that the probability of correctly classifying a vector $\mathbf{x}$ is $1-P(error) = 1-\sum_{-\infty}^{\infty} P(error|x)P(x)$. For the decision boundary $\theta = 12.3$, it would therefore equal:

$$\sum_{i=0}^{12} P(\omega_1|x)P(x) + \sum_{i=13}^{\infty} P(\omega_2|x)P(x) = \sum_{i=0}^{12} P(x|\omega_1)P(w_1) + \sum_{i=13}^{\infty} P(x|\omega_2)P(\omega_2)$$

.

The probability of correctly classifying as class 1 is

$$P(x < 12.3|\omega_1)P(\omega_1) = \sum_{x=0}^{12} e^{-10}10^x/x! * 0.5 \approx \boxed{0.396}$$

and the probability of correctly classifying as class 2 is

$$P(x > 12.3|\omega_2)P(\omega_2) = (1 - \sum_{x=0}^{12} e^{-15}15^x/x!) * 0.5 \approx \boxed{0.366}$$

The total error rate for this decision boundary is $1 - 0.396 - 0.366 \approx 0.238$.

(c) Suppose instead of one, we can obtain two independent measurements x_1 and x_2 for the object to be classified. How does the allocation rule change? In the case where $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ and $\lambda_1 = 10$ and $\lambda_2 = 15$, calculate the new total error.

Solution: If we receive two independent measurements x_1, x_2, then the decision boundary condition for choosing class 1 is

$$1 < \frac{P(\omega_1|x_1,x_2)}{P(\omega_2|x_1,x_2)} = \frac{P(x_1,x_2|\omega_1)P(\omega_1)}{P(x_1,x_2|\omega_2)P(\omega_2)} = e^{2(\lambda_2-\lambda_1)}\left(\frac{\lambda_1}{\lambda_2}\right)^{x_1+x_2}$$

which means that we should choose class 1 if

$$\boxed{\frac{x_1+x_2}{2} < \frac{\lambda_1-\lambda_2}{\ln\lambda_1 - \ln\lambda_2} \approx 12.3}$$

The probability of correctly classifying as class 1 is

$$P(x_1 + x_2 < 24.6|\omega_1)P(\omega_1) = \sum_{x=0}^{24} e^{-20}20^x/x! * 0.5 \approx \boxed{0.4216}$$

(since $x_1 + x_2 \sim \text{Poisson}(2\lambda_1)$ assuming x_1, x_2 are iid from class 1), and the probability of correctly classifying as class 2 is

$$P(x_1 + x_2 > 24.6|\omega_2)P(\omega_2) = 1 - \sum_{x=0}^{24} e^{-30}30^x/x!0.5 \approx \boxed{0.4214}$$

(since $x_1 + x_2 \sim \text{Poisson}(2\lambda_2)$ assuming x_1, x_2 are iid from class 2). The total error rate for this decision boundary is

$$1 - 0.4216 - 0.4214 \approx \boxed{0.157}$$

2 Logistic posterior with exponential class conditionals

We have seen in class that Gaussian class conditionals can lead to a logistic posterior that is linear in X. Now, suppose the class conditionals are exponentially distributed with parameters λ_i, i.e.

$$p(x|Y=i) = \lambda_i \exp(-\lambda_i x), \quad \text{where } i \in \{0,1\}$$
$$Y \sim \text{Bernoulli}(\pi)$$

Show that the posterior distribution of the class label given X is also a logistic function, however with a linear argument in X. What is the decision boundary?

Solution:

We are solving for $P(Y=1|x)$. By Bayes Rule, we have

$$\begin{aligned} P(Y=1|x) &= \frac{P(x|Y=1)P(Y=1)}{P(x|Y=1)P(Y=1)+P(x|Y=0)P(Y=0)} \\ &= \frac{1}{1+\frac{P(Y=0)P(x|Y=0)}{P(Y=1)P(x|Y=1)}} \\ &= \frac{1}{1+\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\exp\left(-\lambda_0 x+\lambda_1 x\right)} \end{aligned}$$

Looking at the bottom right equation, we have

$$\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\exp\left(-\lambda_0 x+\lambda_1 x\right) = \exp\left(-(\lambda_0-\lambda_1)x + \log\left(\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}\right)\right)$$

Now we see that we have a logistic function $\frac{1}{1+\exp\left(-h(x)\right)}$, where $h(x) = ax+b$ is linear (affine) in x. Since we are assuming 0-1 loss, we use the optimal classifier $f^*(x) = 1$ when $P(Y=1|x) > P(Y=0|x)$. Thus, the decision boundary can be found when $P(Y=1|x) = P(Y=0|x) = \frac{1}{2}$. This happens when $h(x) = 0$. Solving for x gives

$$\bar{x} = \frac{\log\frac{\lambda_0}{\lambda_1}\frac{1-\pi}{\pi}}{\lambda_0-\lambda_1}.$$

If we assume $\lambda_0 > \lambda_1$, then the optimal classifier is

$$f^*(x) = \begin{cases} 1 & \text{if } x > \bar{x} \\ 0 & o.w. \end{cases}$$

3 Gaussian Classification

Let $f(x \mid C_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-class, one-dimensional classification problem with classes C_1 and C_2, $P(C_1) = P(C_2) = 1/2$, and $\mu_2 > \mu_1$.

(a) Find the Bayes optimal decision boundary and the corresponding Bayes decision rule.

(b) The Bayes error is the probability of misclassification,

$$P_e = P((\text{misclassified as } C_1) \mid C_2)\, P(C_2) + P((\text{misclassified as } C_2) \mid C_1)\, P(C_1).$$

Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz$$

where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$.

Solution:

(a) The decision boundary occurs at the point, x, where $P(C_1 \mid x) = P(C_2 \mid x)$. Thus we have

$$\begin{aligned}
P(C_1 \mid x) &= P(C_2 \mid x) \\
\implies \frac{f(x \mid C_1)P(C_1)}{f(x \mid C_1)P(C_1) + f(x \mid C_2)P(C_2)} &= \frac{f(x \mid C_2)P(C_2)}{f(x \mid C_1)P(C_1) + f(x \mid C_2)P(C_2)} \quad \text{(by Bayes rule)} \\
\implies f(x \mid C_1)P(C_1) &= f(x \mid C_2)P(C_2) \\
\implies f(x \mid C_1)\frac{1}{2} &= f(x \mid C_2)\frac{1}{2} \\
\implies f(x \mid C_1) &= f(x \mid C_2) \\
\implies (x - \mu_1)^2 &= (x - \mu_2)^2
\end{aligned}$$

This yields the Bayes decision boundary: $x = \frac{\mu_1 + \mu_2}{2}$.

The corresponding decision rule is, given a data point $x \in \mathbb{R}$:

- if $x < \frac{\mu_1 + \mu_2}{2}$, then classify x in class 1 (since $\mu_2 > \mu_1$).
- otherwise, classify x in class 2

(b)

$$\begin{aligned}
P((\text{misclassified as } C_1) \mid C_2) &= P(x < \frac{\mu_1 + \mu_2}{2} \mid C_2) \\
&= \int_{-\infty}^{(\mu_1+\mu_2)/2} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_2)^2/(2\sigma^2)} dx
\end{aligned}$$

$$\text{Let } z = \frac{x - \mu_2}{\sigma}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\frac{\mu_1-\mu_2}{2\sigma}} \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/2} \sigma dz && \left(\text{since } \frac{dx}{dz} = \sigma\right) \\
&= \int_{-\infty}^{\frac{\mu_1-\mu_2}{2\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \int_{-\infty}^{-a} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \int_{a}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz && \text{(by symmetry)} \\
&= \frac{1}{\sqrt{2\pi}} \int_{a}^{+\infty} e^{-z^2/2} dz \\
&= P_e
\end{aligned}
$$

$$
\begin{aligned}
P((\text{misclassified as } C_2) \mid C_1) &= P(x \geq \frac{\mu_1+\mu_2}{2} \mid C_1) \\
&= \int_{(\mu_1+\mu_2)/2}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_1)^2/(2\sigma^2)} dx \\
\text{Let } z = \frac{x-\mu_1}{\sigma} \\
&= \int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \int_{a}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= P_e
\end{aligned}
$$

Therefore:

$$P((\text{misclassified as } C_1) \mid C_2)P(C_2) + P((\text{misclassified as } C_2) \mid C_1)P(C_1) = P_e \cdot \frac{1}{2} + P_e \cdot \frac{1}{2} = P_e$$

4 Bias Variance for Ridge Regression

Recall the statistical model for ridge regression from lecture. We have a design matrix $\mathbf{X}$, where the rows of $\mathbf{X} \in \mathbb{R}^{n\times d}$ are our data points $\mathbf{x}_i \in \mathbb{R}^d$. We assume a linear regression model

$$Y = \mathbf{X}\mathbf{w}^* + \mathbf{z}$$

Where $\mathbf{w}^* \in \mathbb{R}^d$ is the true parameter we are trying to estimate, $\mathbf{z} = [z_1, \ldots, z_n] \sim \mathcal{N}(0, \sigma^2 I_n)$, and $Y = [y_1, \ldots, y_n]$ is the random variable representing our labels.

Throughout this problem, you may assume $\mathbf{X}^\top\mathbf{X}$ is invertible. Given a realization of the labels $Y = \mathbf{y}$, recall these two estimators we have studied:

$$\begin{aligned}\mathbf{w}_{\text{ols}} &= \min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \\ \mathbf{w}_{\text{ridge}} &= \min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2\end{aligned}$$

(a) Write the solution for $\mathbf{w}_{\text{ols}}, \mathbf{w}_{\text{ridge}}$. No need to derive it.

Solution:

$$\begin{aligned}\mathbf{w}_{\text{ols}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ \mathbf{w}_{\text{ridge}} &= (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

(b) Let $\widehat{\mathbf{w}} \in \mathbb{R}^d$ denote any estimator of $\mathbf{w}^*$. In the context of this problem, an estimator $\widehat{\mathbf{w}} = \widehat{\mathbf{w}}(Y)$ is any function which takes the data $\mathbf{X}$ and a realization of Y, and computes a guess of $\mathbf{w}^*$.

Define the MSE (mean squared error) of the estimator $\widehat{\mathbf{w}}$ as

$$\text{MSE}(\widehat{\mathbf{w}}) := \mathbb{E}\left[\left\|\widehat{\mathbf{w}} - \mathbf{w}^*\right\|_2^2\right].$$

Above, the expectation is taken w.r.t. the randomness inherent in z. Note that this is a multivariate generalization of the mean squared error we have seen previously.

Define $\widehat{\boldsymbol{\mu}} := \mathbb{E}[\widehat{\mathbf{w}}]$. Show that the MSE decomposes as such

$$\text{MSE}(\widehat{\mathbf{w}}) = \left\|\widehat{\boldsymbol{\mu}} - \mathbf{w}^*\right\|_2^2 + \text{Tr}(\text{Cov}(\widehat{\mathbf{w}})).$$

Note that this is a multivariate generalization of the bias-variance decomposition we have seen previously.

Hint: The inner product of two vectors is the trace of their outer product. Also, expectation and trace commute, so $\mathbb{E}[\text{Tr}(A)] = \text{Tr}(\mathbb{E}[A])$ for any square matrix A.

Solution:

$$E[\left\|\widehat{w} - w_*\right\|_2^2] = E[\left\|(\widehat{w} - \widehat{\mu}) - (w_* - \widehat{\mu})\right\|_2^2]$$

$$
\begin{aligned}
&= E[\left\|\widehat{w}-\widehat{\mu}\right\|^2 - 2(\widehat{w}-\widehat{\mu})(w_*-\widehat{\mu}) + \left\|w_*-\widehat{\mu}\right\|_2^2] \\
&= E[\left\|\widehat{w}-\widehat{\mu}\right\|^2] - 2E[(\widehat{w}-\widehat{\mu})(w_*-\widehat{\mu})] + E[\left\|w_*-\widehat{\mu}\right\|_2^2] \\
&= E[\left\|\widehat{w}-\widehat{\mu}\right\|^2] - 2E[(\widehat{w}-\widehat{\mu})](w_*-\widehat{\mu}) + \left\|w_*-\widehat{\mu}\right\|_2^2 \\
&= E[\left\|\widehat{w}-\widehat{\mu}\right\|^2] + \left\|w_*-\widehat{\mu}\right\|_2^2 && \text{(since } E[(\widehat{w}-\widehat{\mu})]=0\text{)} \\
&= E[\mathrm{Tr}((\widehat{w}-\widehat{\mu})(\widehat{w}-\widehat{\mu})^\top)] + \left\|w_*-\widehat{\mu}\right\|_2^2 \\
&= \mathrm{Tr}(E[(\widehat{w}-\widehat{\mu})(\widehat{w}-\widehat{\mu})^\top]) + \left\|w_*-\widehat{\mu}\right\|_2^2 \\
&= \mathrm{Tr}(\mathrm{Cov}(\widehat{w})) + \left\|w_*-\widehat{\mu}\right\|_2^2 .
\end{aligned}
$$

(c) Show that

$$\mathbb{E}[\mathbf{w}_{\text{ols}}] = \mathbf{w}^*, \qquad \mathbb{E}[\mathbf{w}_{\text{ridge}}] = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w}^* .$$

That is, $\mathbf{w}_{\text{ols}}$ is an *unbiased* estimator of $\mathbf{w}^*$, whereas $\mathbf{w}_{\text{ridge}}$ is a *biased* estimator of $\mathbf{w}^*$.

Solution: For OLS,

$$
\begin{aligned}
w_{\text{ols}} &= (X^\top X)^{-1}X^\top Y \\
&= (X^\top X)^{-1}X^\top(Xw_* + z) \\
&= w_* + (X^\top X)^{-1}X^\top z .
\end{aligned}
$$

Thus, we have that

$$
\begin{aligned}
E[w_{\text{ols}}] &= E[w_* + (X^\top X)^{-1}X^\top z] \\
&= E[w_*] + E[(X^\top X)^{-1}X^\top z] \\
&= w_* + (X^\top X)^{-1}X^\top E[z] \\
&= w_* && \text{(since } E[z]=0\text{).}
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
w_{\text{ridge}} &= (X^\top X + \lambda I_d)^{-1}X^\top Y \\
&= (X^\top X + \lambda I_d)^{-1}X^\top(Xw_* + z) \\
&= (X^\top X + \lambda I_d)^{-1}X^\top Xw_* + (X^\top X + \lambda I_d)^{-1}X^\top z ,
\end{aligned}
$$

and therefore,

$$
\begin{aligned}
E[w_{\text{ridge}}] &= E[(X^\top X + \lambda I_d)^{-1}X^\top Xw_* + (X^\top X + \lambda I_d)^{-1}X^\top z] \\
&= E[(X^\top X + \lambda I_d)^{-1}X^\top Xw_*] + E[(X^\top X + \lambda I_d)^{-1}X^\top z] \\
&= (X^\top X + \lambda I_d)^{-1}X^\top Xw_* + (X^\top X + \lambda I_d)^{-1}X^\top E[z] \\
&= (X^\top X + \lambda I_d)^{-1}X^\top Xw_* && \text{(since } E[z]=0\text{).}
\end{aligned}
$$

(d) Let $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_d$ denote the d eigenvalues of the matrix $\mathbf{X}^\top \mathbf{X}$ arranged in non-increasing order. Show that

$$\mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ols}})) = \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i}, \qquad \mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ridge}})) = \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i + \lambda)^2} .$$

Finally, use these formulas to conclude that

$$\mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ridge}})) < \mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ols}})) .$$

Hint: Remember the relationship between the trace and the eigenvalues of a matrix. Also, for the ridge variance, consider writing $\mathbf{X}^\top \mathbf{X}$ in terms of its eigen-decomposition $U\Sigma U^\top$.

Solution: For OLS, we have

$$\begin{aligned}
\mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ols}})) &= \mathrm{Tr}(E[(\mathbf{w}_{\mathrm{ols}} - E[\mathbf{w}_{\mathrm{ols}}])(\mathbf{w}_{\mathrm{ols}} - E[\mathbf{w}_{\mathrm{ols}}])^\top]) \\
&= \mathrm{Tr}(E[(w^* + (X^\top X)^{-1} X^\top z - w^*])(w^* + (X^\top X)^{-1} X^\top z - w^*)^\top]) \\
&= \mathrm{Tr}(E[((X^\top X)^{-1} X^\top z])((X^\top X)^{-1} X^\top z)^\top]) \\
&= \mathrm{Tr}(E[(X^\top X)^{-1} X^\top z z^\top X (X^\top X)^{-1}]) \\
&= \mathrm{Tr}((X^\top X)^{-1} X^\top E[z z^\top] X (X^\top X)^{-1}) \\
&= \mathrm{Tr}((X^\top X)^{-1} X^\top (\sigma^2 I_n) X (X^\top X)^{-1}) \\
&= \sigma^2 \mathrm{Tr}((X^\top X)^{-1} X^\top X (X^\top X)^{-1}) \\
&= \sigma^2 \mathrm{Tr}((X^\top X)^{-1}) \\
&= \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i} .
\end{aligned}$$

For Ridge, writing $X^\top X = U\Sigma U^\top$, observe that

$$\begin{aligned}
(X^\top X + \lambda I_d)^{-1} &= U(\Sigma + \lambda I_d)^{-1} U^\top \\
(X^\top X + \lambda I_d)^{-1} X^\top X &= U(\Sigma + \lambda I_d)^{-1} \Sigma U^\top .
\end{aligned}$$

Recall that,

$$\begin{aligned}
w_{\mathrm{ridge}} &= (X^\top X + \lambda I_d)^{-1} X^\top X w_* + (X^\top X + \lambda I_d)^{-1} X^\top z \\
E[w_{\mathrm{ridge}}] &= (X^\top X + \lambda I_d)^{-1} X^\top X w_* .
\end{aligned}$$

Thus we have,

$$\begin{aligned}
\mathrm{Tr}(\mathrm{Cov}(\mathbf{w}_{\mathrm{ols}})) &= \mathrm{Tr}(E[(X^\top X + \lambda I_d)^{-1} X^\top z)(X^\top X + \lambda I_d)^{-1} X^\top z)^\top]) \\
&= \mathrm{Tr}(E[(X^\top X + \lambda I_d)^{-1} X^\top z z^\top X (X^\top X + \lambda I_d)^{-1}]) \\
&= \mathrm{Tr}((X^\top X + \lambda I_d)^{-1} X^\top E[z z^\top] X (X^\top X + \lambda I_d)^{-1}) \\
&= \mathrm{Tr}((X^\top X + \lambda I_d)^{-1} X^\top (\sigma^2 I_n) X (X^\top X + \lambda I_d)^{-1}) \\
&= \sigma^2 \mathrm{Tr}((X^\top X + \lambda I_d)^{-1} X^\top X (X^\top X + \lambda I_d)^{-1})
\end{aligned}$$

$$
\begin{aligned}
&= \sigma^2 \mathrm{Tr}(U(\Sigma + \lambda I_d)^{-1}\Sigma(\Sigma + \lambda I_d)^{-1}U^\top) \\
&= \sigma^2 \mathrm{Tr}((\Sigma + \lambda I_d)^{-1}\Sigma(\Sigma + \lambda I_d)^{-1}U^\top U) && \text{(by cyclic property of the trace)} \\
&= \sigma^2 \mathrm{Tr}((\Sigma + \lambda I_d)^{-1}\Sigma(\Sigma + \lambda I_d)^{-1}) \\
&= \sigma^2 \mathrm{Tr}(\Sigma(\Sigma + \lambda I_d)^{-2}) \\
&= \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i + \lambda)^2} \,.
\end{aligned}
$$

The inequality $\mathrm{Tr}(\mathrm{Cov}(w_{\text{ridge}})) < \mathrm{Tr}(\mathrm{Cov}(w_{\text{ols}}))$ holds because $(\gamma_i + \lambda)^2 > \gamma_i^2$ and $\gamma_i > 0$ for all $1 \le i \le d$. Thus,

$$
\begin{aligned}
\sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i + \lambda)^2} &\le \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{\gamma_i^2} \\
&= \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i} \\
\implies \mathrm{Tr}(\mathrm{Cov}(w_{\text{ridge}})) &< \mathrm{Tr}(\mathrm{Cov}(w_{\text{ols}}))
\end{aligned}
$$