# EE263 Final Exam Solutions, December 2025

- This is a 75-minute in-class exam.
- Write your answers in the provided blue book. Make sure to clearly label each page with the problem number you are working on.
- If we ask you to "show" something is true, we are expecting a concise, intuitive derivation *not a formal proof*.
- You may use any books or paper notes.
- You may not use any electronic resources. For example, you cannot use a laptop, tablet or phone, run Julia code, search online, or consult a large language model.
- You may not discuss the exam or course material with others until everyone has taken the exam.
- If you have a question, please ask the staff. We have done our best to make the exam unambiguous. So unless there is a mistake, we are unlikely to say much.
- Please box or otherwise highlight your solution on the page.
- The problems on this exam are split into parts. Each part has a solution under 1/2 page. Thus, if you find yourself producing a complicated and time-consuming solution, stop and reconsider your assumptions.
- Good luck!

**1. Multiple choice (10 points).** For each of the problems below, select exactly one correct answer. Throughout this question $A \in \mathbb{R}^{m\times n}$ and no assumption on the relationship between $m$ and $n$ is made unless explicitly stated.

a) Suppose $A \in \mathbb{R}^{m\times n}$ with $m > n$ and $\operatorname{rank}(A) = n$. Which statement about the system $Ax = y$ is correct?

(A) For every $y \in \mathbb{R}^m$, there is at least one solution and generally infinitely many.

(B) For every $y \in \mathbb{R}^m$, there is a unique solution.

(C) There exists some $y$ for which $Ax = y$ has no solution, but whenever a solution exists, it is unique.

(D) There exist $y_1, y_2$ such that $Ax = y_1$ has no solutions and $Ax = y_2$ has infinitely many solutions.

b) Suppose $A \in \mathbb{R}^{m\times n}$ with $m > n$ and full column rank, and let

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

be a full QR factorization, where $Q_1 \in \mathbb{R}^{m\times n}$, $Q_2 \in \mathbb{R}^{m\times(m-n)}$ have orthonormal columns and $R_1 \in \mathbb{R}^{n\times n}$ is upper triangular and invertible. For a given $y \in \mathbb{R}^m$, the orthogonal projection of $y$ onto $\operatorname{range}(A)$ is

(A) $A^\mathsf{T}(AA^\mathsf{T})^{-1}y$.

(B) $(A^\mathsf{T}A)^{-1}A^\mathsf{T}y$.

(C) $Q_1Q_1^\mathsf{T}y$.

(D) $Q_2Q_2^\mathsf{T}y$.

c) Let $A \in \mathbb{R}^{m\times n}$ have singular value decomposition

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\mathsf{T},$$

where $r = \operatorname{rank}(A)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Which expression equals the Frobenius norm $\|A\|_F$?

(A) $\displaystyle\sum_{i=1}^{r} \sigma_i$.

(B) $\displaystyle\sum_{i=1}^{r} |\sigma_i|$.

(C) $\displaystyle\sqrt{\sum_{i=1}^{r} \sigma_i^2}$.

(D) $\displaystyle\sqrt{\max_{1\leq i\leq r} \sigma_i^2}$.

d) Let $A \in \mathbb{R}^{m\times n}$ be arbitrary and let $A^\dagger$ denote its pseudo-inverse. For a given $y \in \mathbb{R}^m$, define $x^\star = A^\dagger y$. Which statement is always true?

(A) $x^\star$ is the unique solution of $Ax = y$ for every $y \in \mathbb{R}^m$.

(B) $x^\star$ minimizes $\|x\|_2$ subject to $Ax = y$ for every $y \in \mathbb{R}^m$.

(C) $x^\star$ minimizes $\|Ax - y\|_2$ over all $x$, and among all minimizers of $\|Ax - y\|_2$ has the smallest norm.

(D) $x^\star$ is in the range of $A$.

e) Let $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$. Consider the statement

$$z^\mathsf{T} y = 0 \quad \text{for every } z \in \text{null}(A^\mathsf{T}). \qquad (*)$$

For which $y$ does $(*)$ hold?

(A) $(*)$ holds if and only if $y \in \text{range}(A)$.

(B) $(*)$ holds if and only if $y \in \text{null}(A)$.

(C) $(*)$ holds if and only if $y \in \text{range}(A^\mathsf{T})$.

(D) $(*)$ holds for every $y \in \mathbb{R}^m$.

f) Let $x \sim \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^n$ with $\Sigma > 0$. For $\alpha > 0$ define the ellipsoid

$$\mathcal{E}_\alpha = \left\{ x \in \mathbb{R}^n \;\middle|\; (x - \mu)^\mathsf{T} \Sigma^{-1} (x - \mu) \le \alpha \right\}.$$

Which statement about $\mathcal{E}_\alpha$ is correct?

(A) The principal axes of $\mathcal{E}_\alpha$ are the eigenvectors of $\Sigma$, and the $i$th semi-axis length is $\sqrt{\alpha\, \lambda_i(\Sigma)}$.

(B) The principal axes of $\mathcal{E}_\alpha$ are the eigenvectors of $\Sigma^{-1}$, and the $i$th semi-axis length is $\sqrt{\alpha / \lambda_i(\Sigma)}$.

(C) $\mathcal{E}_\alpha$ is a ball of radius $\sqrt{\alpha}$ centered at $\mu$ for every positive definite $\Sigma$.

(D) Doubling $\alpha$ doubles every semi-axis length.

g) Let $A \in \mathbb{R}^{m \times n}$ have singular value decomposition $A = U \Sigma V^\mathsf{T}$ and pseudo-inverse $A^\dagger = V \Sigma^\dagger U^\mathsf{T}$. Which statement is always true?

(A) $AA^\dagger$ is the orthogonal projector onto range$(A)$.

(B) $A^\dagger A$ is the orthogonal projector onto null$(A)$.

(C) $AA^\dagger = I_m$ for every $A$.

(D) $A^\dagger A = I_n$ for every $A$.

h) Let $L$ be the Laplacian of an undirected weighted graph with $n$ nodes. Suppose $L$ has exactly 3 linearly independent eigenvectors associated with eigenvalue 0. Which of the following statements is true?

(A) The graph is connected and has 3 self-loops.

(B) The graph has 3 connected components.

(C) The graph has 3 cycles.

(D) The graph has 3 nodes of maximum degree.

**Solution.**

a) (C)

b) (C)

c) (C)

d) (C)

e) (A)

f) (A)

g) (A)

h) (B)

**2. Warehouse location (10 points).** We want to locate two warehouses in $\mathbb{R}^2$ to serve four cities. The city locations are described by two dimensional vectors

$$c_1, c_2, c_3, c_4 \in \mathbb{R}^2,$$

and warehouse $k$ is placed at $x^{(k)} \in \mathbb{R}^2$, $k = 1, 2$. Cities 1 and 2 are served by warehouse 1; cities 3 and 4 are served by warehouse 2.

We choose a quadratic shipping cost

$$J_{\text{ship}}(x) = \|x^{(1)} - c_1\|_2^2 + \|x^{(1)} - c_2\|_2^2 + \|x^{(2)} - c_3\|_2^2 + \|x^{(2)} - c_4\|_2^2,$$

and we also penalize the warehouses being far apart

$$J_{\text{pen}}(x) = \|x^{(1)} - x^{(2)}\|_2^2.$$

For a given penalty parameter $\mu \geq 0$ we define the total cost

$$J_\mu(x) = J_{\text{ship}}(x) + \mu J_{\text{pen}}(x),$$

where

$$x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix} \in \mathbb{R}^4.$$

a) Find a matrix $A_{\text{ship}} \in \mathbb{R}^{8\times 4}$ and a vector $y \in \mathbb{R}^8$ such that

$$J_{\text{ship}}(x) = \|A_{\text{ship}}x - y\|_2^2.$$

b) Find a matrix $A_{\text{pen}} \in \mathbb{R}^{2\times 4}$ such that

$$J_{\text{pen}}(x) = \|A_{\text{pen}}x\|_2^2.$$

c) Show that for any $\mu \geq 0$ we can write

$$J_\mu(x) = \|\hat{A}x - \hat{y}\|_2^2$$

for some matrix $\hat{A}$ and vector $\hat{y}$ that you should write explicitly in terms of $A_{\text{ship}}$, $A_{\text{pen}}$, $y$, and $\mu$.

d) Does the problem

$$\min_x J_\mu(x)$$

always have a unique minimizer for every $\mu \geq 0$? Justify your answer.

e) Give a short geometric interpretation of the optimal warehouse locations for $\mu = 0$. Consider the specific city locations

$$c_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad c_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \qquad c_3 = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \qquad c_4 = \begin{bmatrix} 2 \\ 4 \end{bmatrix},$$

and denote the optimal solution for a given $\mu$ by $x^\star(\mu)$. What is $x^\star(0)$?

f) Give a short geometric interpretation of the optimal warehouse locations for $\mu \to \infty$. What is $x^\star(\mu)$ for $\mu \to \infty$ for the specific city locations above?

g) For this problem we can express the solution as

$$x^\star(\mu) = \frac{1}{2} P \begin{bmatrix} I_2 & 0 \\ 0 & \frac{1}{1+\mu} I_2 \end{bmatrix} P^{\mathsf{T}} A_{\text{ship}}^{\mathsf{T}} y,$$

where $P$ is some orthogonal matrix that is not a function of $\mu$. Sketch the solutions $x^\star(\mu)$ for $\mu = 0$ and $\mu \to \infty$ and on the same diagram show where the solutions $x^\star(\mu)$ for $\mu \in (0, \infty)$ lie. Justify your answer using the provided expression for $x^\star(\mu)$.

**Solution.**

a) We want

$$J_{\text{ship}}(x) = \left\| \begin{bmatrix} x^{(1)} - c_1 \\ x^{(1)} - c_2 \\ x^{(2)} - c_3 \\ x^{(2)} - c_4 \end{bmatrix} \right\|_2^2 = \left\| A_{\text{ship}} \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix} - y \right\|_2^2.$$

This is achieved by taking

$$A_{\text{ship}} = \begin{bmatrix} I_2 & 0 \\ I_2 & 0 \\ 0 & I_2 \\ 0 & I_2 \end{bmatrix} \in \mathbb{R}^{8\times 4}, \qquad y = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} \in \mathbb{R}^8.$$

b) We want

$$J_{\text{pen}}(x) = \left\| x^{(1)} - x^{(2)} \right\|_2^2 = \left\| A_{\text{pen}} \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix} \right\|_2^2.$$

This is achieved by taking

$$A_{\text{pen}} = \begin{bmatrix} I_2 & -I_2 \end{bmatrix} \in \mathbb{R}^{2\times 4}.$$

c) We have

$$J_\mu(x) = \|A_{\text{ship}}x - y\|_2^2 + \mu\|A_{\text{pen}}x\|_2^2 = \left\| \begin{bmatrix} A_{\text{ship}}x - y \\ \sqrt{\mu}\, A_{\text{pen}}x \end{bmatrix} \right\|_2^2.$$

We can factor this into

$$J_\mu(x) = \left\| \begin{bmatrix} A_{\text{ship}} \\ \sqrt{\mu}\, A_{\text{pen}} \end{bmatrix} x - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|_2^2.$$

Thus

$$\hat{A} = \begin{bmatrix} A_{\text{ship}} \\ \sqrt{\mu}\, A_{\text{pen}} \end{bmatrix}, \qquad \hat{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

d) Yes, the minimizer is always unique for every $\mu \geq 0$. Observe that $A_{\text{ship}}$ has full column rank and addition of extra rows cannot destroy column independence.

e) For $\mu = 0$ we ignore the penalty and just minimize the shipping cost. Geometrically, the shipping cost is minimized by placing each warehouse at the midpoint of the segment joining the two cities it serves. This can be derived algerbracally by noting that the problem separates into optimization over each warehouse location independently:

$$\min_{x^{(1)}} \left(\|x^{(1)} - c_1\|_2^2 + \|x^{(1)} - c_2\|_2^2\right),$$

$$\min_{x^{(2)}} \left(\|x^{(2)} - c_3\|_2^2 + \|x^{(2)} - c_4\|_2^2\right).$$

The solutions are simply:

$$x^{(1)\star}(0) = \frac{c_1 + c_2}{2} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$x^{(2)\star}(0) = \frac{c_3 + c_4}{2} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}.$$

So

$$x^\star(0) = \begin{bmatrix} x^{(1)\star}(0) \\ x^{(2)\star}(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 4 \end{bmatrix}.$$

f) The penalty term discourages keeping the warehouses apart. In the limit $\mu \to \infty$ we effectively constrain the warehouses to be at the same location.

$$x^{(1)} = x^{(2)} = z \in \mathbb{R}^2,$$

and then minimize

$$\sum_{i=1}^{4} \|z - c_i\|_2^2.$$

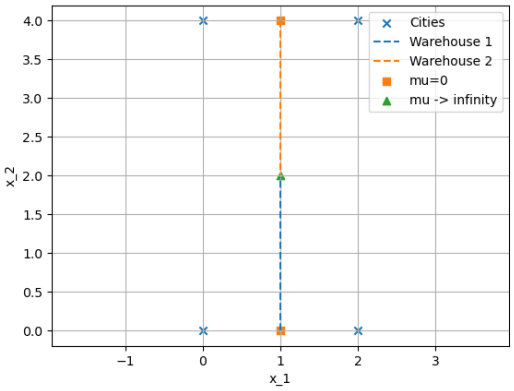
This is now similar to the problem in the previous part and this is minimized by the centroid of all four cities. For the specific locations we get:

$$\frac{1}{4} \sum_{i=1}^{4} c_i = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

So as

$$\mu \to \infty, \quad x^\star(\mu) \to \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}.$$

g) From the expression given we see that as $\mu$ changes and everything else is fixed the warehouses move monotonically along fixed directions defined by the columns of $P$. The solutions for $\mu \in (0, \infty)$ lie on a straight line between the solutions for $\mu = 0$ and $\mu \to \infty$.



**3. Two variations on least-squares regression (10 points).** In least-squares regression, we seek to predict a *target* $y$ as a linear function of features $x_1, \ldots, x_p$: that is, $y \approx \beta_1 x_1 + \cdots + \beta_p x_p$. Given a dataset of $N > p$ samples $y^{(i)} \in \mathbb{R}$ and $x^{(i)} \in \mathbb{R}^p$, we define

$$y \in \mathbb{R}^N = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}, \; X \in \mathbb{R}^{N \times p} = \begin{bmatrix} (x^{(1)})^\top \\ \vdots \\ (x^{(N)})^\top \end{bmatrix}.$$

a) Suppose $X$ is not full rank. (Statisticians call this multicollinearity, but you do not need to know any statistics to solve this problem.) What can you say about the singular values of $X$? What about the least-squares solution $\beta = (X^\top X)^{-1} X^\top y$?

b) Explain how to find $\beta$ for this problem such that $\hat{y} = X\beta$ minimizes the objective

$$J(\hat{y}) = \|\hat{y} - y\|^2.$$

The pseudoinverse $X^\dagger$ may be useful. Clearly explain why your method works when $X$ is not full rank.

c) Now we consider a different situation. Given $X \in \mathbb{R}^{N \times p}, N > p$, you compute $\beta$ using least-squares. Your friend (who has not taken ee263) proposes a shortcut. Instead of fitting the entire $\beta$ vector using your method from part (b), he fits each $\beta_j$ to solve the single-variable regression

$$\text{minimize} \sum_{i=1}^{N} (y_i - \beta_j X_{ij})^2$$

Then he arranges the $\beta_j$'s into a new vector $\tilde{\beta}$. Which of the following is true?

i. Your friend's approach yields a LOWER $J$ value than yours.
ii. Your friend's approach yields a HIGHER $J$ value than yours.
iii. You and your friend have the SAME $J$ value.

Briefly justify your answer.

d) You explain the SVD to your friend, and he decides to modify his approach by computing the SVD $X = U\Sigma V^\top$. Then, for each column $u_j$ in $U$, he solves the single-variable regression problem from part (c) to obtain a coefficient $\beta_j$. Which of the following is true?

i. Your friend's new approach yields a LOWER $J$ value than yours.
ii. Your friend's new approach yields a HIGHER $J$ value than yours.
iii. You and your friend now have the SAME $J$ value.

Briefly justify your answer.

Hint: It may help to think of the objective in terms of the columns of $X$: that is,

$$J(\hat{y}) = \|y - \hat{y}\|^2 = \|y - (\beta_1 x_1 + \cdots + \beta_p x_p)\|^2$$

where $x_i$ is the $i$th column of $X$.

e) In part (d), which of the following is true?

i. Your friend's $\tilde{\beta}$ vector is DIFFERENT from your $\beta$.
ii. Your friend's $\tilde{\beta}$ vector is the SAME as your $\beta$.

Briefly justify your answer.

**Solution.**

a) If $X$ is not full rank, at least one singular value is zero. Then $(X^\top X)^{-1}$ won't exist and we cannot compute $\beta$. We also gave full credit to answers that said the global optimum may not be unique, or that one can use the SVD to compute $X^\dagger$ when $X$ is not full rank.

b) The pseudoinverse $X^\dagger$ can be used to find the least-squares solution $\hat{y} = XX^\dagger y$.

The pseudoinverse is guaranteed to exist because it is defined in terms of the SVD. When $X \in \mathbb{N} \times \text{ı}$ has rank $\ell < p$, we can write the SVD as

$$X = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix}$$

where $U_1 \in \mathbb{R}^{N \times \ell}$, $\Sigma_1 \in \mathbb{R}^{\ell \times \ell}$, and $V_1 \in \mathbb{R}^{\ell \times p}$.

Since $X = U_1 \Sigma_1 V_1^\top$, $X^\dagger = V_1 \Sigma_1^{-1} U_1^\top$.

c) Your friend's approach yields a higher $J$ value. One way to explain this is that if features $x_i$ and $x_j$ are correlated, meaning $x_i^\top x_j \neq 0$, when your friend performs the single-variable linear regression, he "accounts for" the projection of $x_i$ on $x_j$ twice in $\beta_j$ and $\beta_i$. So his $J$ cannot be the same as yours. You are computing the global optimal solution, so your friend cannot outperform you. So his $J$ must be higher than yours.

Some students claimed your friend's $J$ must be lower than yours. This can never be true, because multivariate least squares has the BLUE (best linear unbiased estimator) property. Another way to see this is that the least-squares objective is convex, thus the least-squares solution yields a globally optimal $\beta$. You will learn a lot more about convexity and optimality if you take EE 364A.

d) The problem your friend ran into was that columns $x_i$ and $x_j$ are correlated, meaning $x_i^\top x_j \neq 0$. So if he can construct a matrix $\tilde{X}$ with the same span as $X$ and orthogonal feature vectors, he should achieve the same $J$ as you did. The orthogonal matrix $U$ has this property, so your friend achieves the same $J$ you did with the multivariate regression.

Many students correctly intuited the answer but provided an incorrect explanation. We accepted any explanation that referred to orthogonality, including mathematically vague but conceptually correct answers such as "columns $x_i$ and $x_j$ interact with each other".

e) Your friend's feature matrix $U$ has a different scale than the original $X$, so he obtains a different $\beta$ vector.

Many students thought that if your friend's approach yields the same $J$ value as yours, $\tilde{\beta}$ must equal $\beta$. This is not true, because your friend's approach in part (d) uses the orthogonal matrix $U$, which you can think of as a scaled/rotated version of the original feature columns in $X$.

Here is some code you can run if you want to experiment with understanding these concepts:

```
using LinearAlgebra
X = rand(10,3);
# y = Ax + noise
y = X*[0.1; -0.2; -0.15] .+ 0.05*randn(10);

# your method
beta_0 = pinv(X) * y
yhat_0 = X*beta_0
println("multivariate least squares (yours) ", round.(beta_0, digits=3))
```

```
# non-orthogonal features one at a time (doesn't reach global optimum)
beta_1 = [pinv(X[:,1:1])*y; pinv(X[:,2:2])*y; pinv(X[:,3:3])*y];
yhat_1 = X*beta_1

println("non-orthogonal (part c)            ", round.(beta_1, digits=3))

# orthogonal features with multivariate regression
# (we didn't ask about this, but it's the same as the one-at-a-time method in part (d))
F = svd(X);
X_tilde = F.U;
beta_2 = pinv(X_tilde)*y;
yhat_2 = X_tilde*beta_2;
println("multivariate least squares on U    ", round.(beta_2, digits=3))

# orthogonal features one at a time
beta_3 = [pinv(X_tilde[:,1:1])*y; pinv(X_tilde[:,2:2])*y; pinv(X_tilde[:,3:3])*y];
yhat_3 = X_tilde*beta_3
println("your friend's method (part d)      ", round.(beta_3, digits=3))

println("multivariate least squares (yours) ", round(norm(yhat_0 - y), digits=3))
println("non-orthogonal (part c)            ", round(norm(yhat_1 - y), digits=3))
println("multivariate least squares on U    ", round(norm(yhat_2 - y), digits=3))
println("your friend's method (part d)      ", round(norm(yhat_3 - y), digits=3))
```

And here is an example of what the code produces:

```
$ julia demo_code.jl
multivariate least squares (yours) [0.128, -0.197, -0.189]
non-orthogonal (part c)            [-0.335, -0.281, -0.288]
multivariate least squares on U    [0.612, 0.092, -0.1]
your friend's method (part d)      [0.612, 0.092, -0.1]
multivariate least squares (yours) 0.169
non-orthogonal (part c)            1.118
multivariate least squares on U    0.169
your friend's method (part d)      0.169
```

**4. Regularized least squares and MMSE (10 points).** Let $x \in \mathbb{R}^n$ be a random vector with prior distribution $x \sim \mathcal{N}(0, \tau^2 I)$. We observe $y \in \mathbb{R}^m$ according to the linear Gaussian model

$$y = Ax + w,$$

where $A \in \mathbb{R}^{m \times n}$ is known and $w \sim \mathcal{N}(0, \sigma^2 I)$ is Gaussian noise independent of $x$. We measure $y = y_{\text{meas}}$ and wish to estimate $x$.

a) Is the random vector $\begin{bmatrix} x \\ y \end{bmatrix}$ jointly Gaussian? Justify your answer using the properties of Gaussian vectors. Write down the covariance matrix for $\begin{bmatrix} x \\ y \end{bmatrix}$.

b) Write down the MMSE estimate of $x$ given $y = y_{\text{meas}}$ in terms of $A$, $\tau$, $\sigma$, and $y_{\text{meas}}$.

c) Now consider the regularized least-squares problem

$$\min_x \quad \|y_{\text{meas}} - Ax\|_2^2 + \lambda \|x\|_2^2$$

with $\lambda > 0$. State or derive the formula for the minimizer of this problem, $x_{\text{reg}}$, in terms of $A$, $\lambda$, and $y_{\text{meas}}$.

d) Show that $x_{\text{reg}}$ is equal to the MMSE for appropriate choice of $\lambda$ and express that $\lambda$ in terms of $\tau$ and $\sigma$. You might find the following identity useful:

$$B(I + CB)^{-1} = (I + BC)^{-1} B.$$

e) Assume that $m \geq n$ and that $A$ has full column rank with SVD

$$A = U \Sigma V^{\mathsf{T}}, \qquad \Sigma = \text{diag}(s_1, \ldots, s_n), \quad s_1 \geq \cdots \geq s_n > 0.$$

Express $x_{\text{mmse}}$ in the right singular-vector basis as

$$x_{\text{mmse}} = \sum_{i=1}^{n} \beta_i v_i,$$

and write $\beta_i$ as functions of $s_i$, $\sigma$, $\tau$, and $u_i^{\mathsf{T}} y_{\text{meas}}$.

f) Let $x_{\text{ls}} = A^{\dagger} y_{\text{meas}}$ be the ordinary least-squares estimate. Show that

$$v_i^{\mathsf{T}} x_{\text{mmse}} = \gamma_i \, v_i^{\mathsf{T}} x_{\text{ls}}$$

for suitable factors $\gamma_i$ that you should give explicitly. What values can $\gamma_i$ take? Briefly interpret how $\gamma_i$ depends on $s_i$, $\sigma$, and $\tau$.

g) Consider the scalar case $n = m = 1$. Express the conditional mean-square error $\mathbb{E}\left((x - x_{\text{mmse}})^2 \mid y = y_{\text{meas}}\right)$ in terms of $s$, $\sigma$, $\tau$, and $y_{\text{meas}}$. Give the limits of this error as $\sigma \to 0$ and as $\sigma \to \infty$, and briefly interpret the results.

**Solution.**

a) We have $x$ and $w$ independent Gaussian vectors, so the vector $\begin{bmatrix} x \\ w \end{bmatrix}$ is jointly Gaussian. The vector $\begin{bmatrix} x \\ y \end{bmatrix}$ is an affine transformation of $\begin{bmatrix} x \\ w \end{bmatrix}$, so it is also jointly Gaussian. The covariance matrix is:

$$\begin{bmatrix} I & 0 \\ A & I \end{bmatrix} \begin{bmatrix} \tau^2 I & 0 \\ 0 & \sigma^2 I \end{bmatrix} \begin{bmatrix} I & A^{\mathsf{T}} \\ 0 & I \end{bmatrix} = \begin{bmatrix} \tau^2 I & \tau^2 A^{\mathsf{T}} \\ \tau^2 A & \tau^2 A A^{\mathsf{T}} + \sigma^2 I \end{bmatrix}.$$

b) For jointly Gaussian $\begin{bmatrix} x \\ y \end{bmatrix}$, the MMSE estimator (conditional mean) is

$$x_{\text{mmse}}(y) = \mathbb{E}[x \mid y] = \Sigma_{xy}\Sigma_{yy}^{-1}y.$$

Using the matrices above,

$$x_{\text{mmse}} = \tau^2 A^\mathsf{T}(\tau^2 AA^\mathsf{T} + \sigma^2 I)^{-1} y_{\text{meas}}.$$

c) We need to solve

$$\min_x \ \|y_{\text{meas}} - Ax\|_2^2 + \lambda\|x\|_2^2.$$

Differentiating the objective and setting to zero gives

$$\nabla_x \left(\|y_{\text{meas}} - Ax\|_2^2 + \lambda\|x\|_2^2\right) = -2A^\mathsf{T}(y_{\text{meas}} - Ax) + 2\lambda x = 0,$$

so

$$(A^\mathsf{T}A + \lambda I)x_{\text{reg}} = A^\mathsf{T} y_{\text{meas}}.$$

Thus

$$x_{\text{reg}} = (A^\mathsf{T}A + \lambda I)^{-1} A^\mathsf{T} y_{\text{meas}}.$$

d) Let's first rewrite the two expressions to bring them closer to the identity form. For the MMSE estimator, we can factor out $\sigma^2$ to get

$$x_{\text{mmse}} = \frac{\tau^2}{\sigma^2} A^\mathsf{T} \left(\frac{\tau^2}{\sigma^2} AA^\mathsf{T} + I\right)^{-1} y_{\text{meas}}.$$

For the regularized least-squares estimator, we can factor out $\lambda$ to get

$$x_{\text{reg}} = \left(\frac{1}{\lambda} A^\mathsf{T}A + I\right)^{-1} \frac{1}{\lambda} A^\mathsf{T} y_{\text{meas}}.$$

We can now apply the identity to the $x_{\text{reg}}$ expression with $B = \frac{1}{\lambda}A^\mathsf{T}$ and $C = A$. This gives

$$x_{\text{reg}} = \frac{1}{\lambda} A^\mathsf{T} \left(\frac{1}{\lambda} AA^\mathsf{T} + I\right)^{-1} y_{\text{meas}}.$$

From this we see that $x_{\text{reg}} = x_{\text{mmse}}$ if

$$\lambda = \frac{\sigma^2}{\tau^2}.$$

e) Using the SVD of $A = U\Sigma V^\mathsf{T}$ we can write

$$A^\mathsf{T}A = V\Sigma^2 V^\mathsf{T}.$$

Substituing this into the expression for $x_{\text{mmse}}$ and simplifying we get

$$x_{\text{mmse}} = V\left[\tau^2\Sigma(\tau^2\Sigma^2 + \sigma^2 I)^{-1}\right] U^\mathsf{T} y_{\text{meas}}.$$

The matrix in the brackets is diagonal so the expression can be written as spectral decomposition

$$x_{\text{mmse}} = \sum_{i=1}^{n} \frac{\tau^2 s_i}{\tau^2 s_i^2 + \sigma^2} \left(u_i^\mathsf{T} y_{\text{meas}}\right) v_i.$$

Comparing with the expression in the question we see that

$$\beta_i = \frac{\tau^2 s_i}{\tau^2 s_i^2 + \sigma^2} \left(u_i^\mathsf{T} y_{\text{meas}}\right).$$

f) The least-squares solution is

$$x_{\text{ls}} = V\Sigma^{-1}U^{\mathsf{T}}y_{\text{meas}} = \sum_{i=1}^{n} \frac{1}{s_i}(u_i^{\mathsf{T}}y_{\text{meas}})\, v_i.$$

Hence

$$v_i^{\mathsf{T}}x_{\text{ls}} = \frac{1}{s_i}u_i^{\mathsf{T}}y_{\text{meas}}.$$

From the previous part we have

$$v_i^{\mathsf{T}}x_{\text{mmse}} = \frac{\tau^2 s_i}{\tau^2 s_i^2 + \sigma^2}\, u_i^{\mathsf{T}}y_{\text{meas}}.$$

Therefore

$$\gamma_i = \frac{\tau^2 s_i^2}{\tau^2 s_i^2 + \sigma^2} = \frac{1}{1 + \dfrac{\sigma^2}{\tau^2 s_i^2}}.$$

Since $\tau^2 s_i^2 > 0$ and $\sigma^2 > 0$, we have

$$0 < \gamma_i < 1.$$

Interpretation:

- If $s_i^2$ is large (direction well-excited by $A$) or $\sigma^2$ is small (low noise) or $\tau^2$ is large (more prior variance allowed), then $\tau^2 s_i^2 \gg \sigma^2$, so $\gamma_i \approx 1$ and the MMSE estimate is close to the LS estimate in that singular direction.
- If $s_i^2$ is small (ill-conditioned direction), or $\sigma^2$ is large (high noise), or $\tau^2$ is very small (strong prior that $x$ is small), then $\tau^2 s_i^2 \ll \sigma^2$, so $\gamma_i \approx 0$ and the MMSE estimator heavily shrinks the LS estimate toward zero in that direction.

g) The conditional mean-square error for the scalar case is given by

$$\mathbb{E}\left((x - x_{\text{mmse}})^2 \mid y = y_{\text{meas}}\right) = \tau^2 - \frac{s^2\tau^4}{s^2\tau^2 + \sigma^2} = \frac{\tau^2\sigma^2}{\sigma^2 + s^2\tau^2}.$$

The limit of the error as $\sigma \to 0$ is

$$\lim_{\sigma\to 0} \frac{\tau^2\sigma^2}{\sigma^2 + s^2\tau^2} = 0.$$

The observation noise vanishes and $y$ essentially reveals $x$ exactly.

The limit of the error as $\sigma \to \infty$ is

$$\lim_{\sigma\to \infty} \frac{\tau^2\sigma^2}{\sigma^2 + s^2\tau^2} = \tau^2.$$

The measurement becomes pure noise and we learn nothing beyond the prior, so the best estimator is 0 and the MSE equals the prior variance $\tau^2$.