Kimi K2 - Thinking Solution to Homework 1

Hanyang Gu

November 30, 2025

Question 1

Consider the scalar equation $\sigma w = y$ with $\sigma > 0$. Minimising the squared loss $L(w) = (y - \sigma w)^2$ by gradient descent with step size η yields the recurrence

$$w_{t+1} = w_t + 2\eta\sigma\,(y - \sigma w_t) = (1 - 2\eta\sigma^2)w_t + 2\eta\sigma y. \tag{1}$$

The unique optimum is $w^* = y/\sigma$.

(a) Stability of the recurrence

Let $\alpha = 1 - 2\eta\sigma^2$. Then (1) becomes

$$w_{t+1} = \alpha\, w_t + 2\eta\sigma y.$$

The homogeneous part $w_{t+1}^{(h)} = \alpha w_t^{(h)}$ decays to zero iff $|\alpha| < 1$. Since $\eta > 0$,

$$-1 < 1 - 2\eta\sigma^2 < 1 \iff 0 < \eta < \frac{1}{\sigma^2}.$$

Hence the recurrence is stable precisely for

$$\boxed{0 < \eta < \frac{1}{\sigma^2}}.$$

Evaluation: Correct. The derivation matches the instructor's solution. The condition for stability is correctly identified as keeping the geometric ratio $|1 - 2\eta\sigma^2| < 1$, which leads to the bound $0 < \eta < 1/\sigma^2$.

(b) Convergence speed

Define the error $e_t = w_t - w^*$. Using $w^* = y/\sigma$ we obtain from (1)

$$e_{t+1} = (1 - 2\eta\sigma^2)\, e_t = \alpha\, e_t.$$

Thus $e_t = \alpha^t e_0$. Because $w_0 = 0$, $e_0 = -w^*$ and

$$|e_t| = |w^*|\, |\alpha|^t.$$

To be within a factor $(1 - \varepsilon)$ of the optimum, i.e. $|e_t| \le \varepsilon |w^*|$, we need $|\alpha|^t \le \varepsilon$. Solving for t,

$$t \ge \frac{\ln \varepsilon}{\ln |\alpha|} = \frac{\ln(1/\varepsilon)}{-\ln |1 - 2\eta\sigma^2|}\,. \tag{1}$$

Hence the number of iterations required is

$$\boxed{t \;\ge\; \frac{\ln(1/\varepsilon)}{-\ln |1 - 2\eta\sigma^2|}}.$$

Evaluation: Correct. The analysis of convergence speed is accurate. The derivation $T > \frac{\log \epsilon}{\log |1-2\eta\sigma^2|}$ correctly links the number of iterations to the contraction factor and the desired error tolerance ϵ.

(c) Two-parameter (vector) problem

Now consider the diagonal system

$$\begin{bmatrix} \sigma_\ell & 0 \\ 0 & \sigma_s \end{bmatrix} \begin{bmatrix} w[1] \\ w[2] \end{bmatrix} = \begin{bmatrix} y[1] \\ y[2] \end{bmatrix}, \qquad \sigma_\ell \gg \sigma_s > 0.$$

Gradient descent with a single learning rate η updates each coordinate independently:

$$w_{t+1}[i] = (1 - 2\eta\sigma_i^2)\, w_t[i] + 2\eta\sigma_i\, y[i], \qquad i \in \{\ell, s\}.$$

From part (a) each coordinate converges iff $0 < \eta < 1/\sigma_i^2$. Because $1/\sigma_\ell^2 < 1/\sigma_s^2$, the overall system converges precisely when

$$\boxed{0 < \eta < \frac{1}{\sigma_\ell^2}}.$$

Thus the *large* singular value σ_ℓ limits the admissible learning rate.

Evaluation: Correct. The stability condition for the vector case is correctly identified as being limited by the direction with the largest curvature (σ_ℓ). The result $\eta < 1/\sigma_\ell^2$ matches the solution.

(d) Which dimension converges faster?

For coordinate i the error contracts by the factor $|\alpha_i| = |1 - 2\eta\sigma_i^2|$. Since $\sigma_\ell > \sigma_s$ and η satisfies the bound of part (c), we have

$$|\alpha_\ell| = |1 - 2\eta\sigma_\ell^2| \leq |1 - 2\eta\sigma_s^2| = |\alpha_s|.$$

Consequently the direction associated with the *small* singular value σ_s decays more slowly. In short,

$$\text{large } \sigma_\ell \text{ direction converges faster,} \qquad \text{small } \sigma_s \text{ direction converges slower.}$$

Evaluation: Correct. The reasoning is sound: for a stable learning rate, the contraction factor $|1 - 2\eta\sigma^2|$ is smaller (better) for the large curvature direction σ_ℓ and closer to 1 (slower) for the small curvature direction σ_s. Thus, the σ_ℓ dimension converges faster.

(e) Optimal learning rate for fastest overall convergence

Overall convergence is dictated by the slowest (largest) contraction factor. To make the two factors as small as possible we equalise their magnitudes:

$$|1 - 2\eta\sigma_\ell^2| = |1 - 2\eta\sigma_s^2|.$$

For the admissible range of η, the left-hand side becomes negative while the right-hand side stays positive, so the equality reduces to

$$-(1 - 2\eta\sigma_\ell^2) = 1 - 2\eta\sigma_s^2.$$

Solving for η gives

$$\boxed{\eta^* = \frac{1}{\sigma_\ell^2 + \sigma_s^2}}.$$

At this step size both coordinates contract with the same factor

$$|\alpha_\ell| = |\alpha_s| = \frac{\sigma_\ell^2 - \sigma_s^2}{\sigma_\ell^2 + \sigma_s^2},$$

which is the smallest possible worst-case factor, yielding the fastest overall convergence.
Evaluation: Perfect match. The optimal rate is correctly found by balancing the convergence speeds of the two extremes, setting $|1 - 2\eta\sigma_\ell^2| = |1 - 2\eta\sigma_s^2|$. The resulting formula $\eta = \frac{1}{\sigma_\ell^2 + \sigma_s^2}$ is correct.

(f) Many parallel problems with intermediate singular values

Suppose we have additional dimensions with singular values σ_i satisfying

$$\sigma_s \leq \sigma_i \leq \sigma_\ell \quad \text{for } i = 3, \dots, k.$$

The stability condition remains $\eta < 1/\sigma_\ell^2$; the intermediate σ_i do not tighten this bound. For any fixed η, the contraction factor $|\alpha_i| = |1 - 2\eta\sigma_i^2|$ is monotone in σ_i^2, so the worst-case factor is still attained at the extremes σ_ℓ and σ_s. Consequently the optimal step size is the same as in part (e):

$$\boxed{\eta^* = \frac{1}{\sigma_\ell^2 + \sigma_s^2}}.$$

Intermediate singular values affect the *individual* convergence rates but not the feasible range of η nor the rate-optimal choice.

Evaluation: Correct. The answer correctly notes that intermediate singular values do not constrain the convergence rate or stability beyond the bounds already set by $\sigma_{\min}$ and $\sigma_{\max}$.

(g) Relevance to general least-squares via the SVD

Consider the general least-squares problem

$$\min_w \ \|Xw - y\|_2^2,$$

where $X \in \mathbb{R}^{m \times d}$ has singular value decomposition $X = U\Sigma V^\top$ with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Gradient descent with step size η reads

$$w_{t+1} = w_t - 2\eta X^\top (Xw_t - y).$$

Introduce the rotated variables $\tilde{w}_t = V^\top w_t$ and $\tilde{y} = U^\top y$. Using $X^\top = V\Sigma^\top U^\top$ we obtain

$$\tilde{w}_{t+1} = \tilde{w}_t - 2\eta\Sigma^\top(\Sigma\tilde{w}_t - \tilde{y}).$$

Because Σ is diagonal, each component evolves independently as

$$\tilde{w}_{t+1}[i] = (1 - 2\eta\sigma_i^2)\,\tilde{w}_t[i] + 2\eta\sigma_i\,\tilde{y}[i].$$

Thus the scalar analysis of parts (a)–(f) applies *coordinate-wise* in the orthonormal basis furnished by the SVD. In particular:

- Stability requires $0 < \eta < \frac{1}{\sigma_{\max}^2}$, where $\sigma_{\max} = \sigma_1$.
- The convergence rate along direction i is governed by $|1 - 2\eta\sigma_i^2|$.

- The overall convergence is bottlenecked by the smallest singular value $\sigma_{\min}$ (the direction of smallest curvature).

- The optimal constant step size for the whole problem is

$$\boxed{\eta^* = \frac{1}{\sigma_{\max}^2 + \sigma_{\min}^2}},$$

  which balances the contraction of the slowest and fastest directions, exactly as derived in part (e).

Hence the simple scalar and parallel-scalar analyses provide a complete picture of gradient-descent dynamics for arbitrary least-squares problems after the change of basis induced by the SVD.

Evaluation: Correct. The connection to general least squares is properly explained using the SVD to diagonalize the problem, showing that the analysis of scalar/parallel problems applies directly to the rotated coordinates.

Question 2.

We consider the wide linear system $X\mathbf{w} = \mathbf{y}$ with $X \in \mathbb{R}^{n \times d}$ having full row rank ($d > n$). There are infinitely many exact solutions. We analyze SGD with batch size 1 and step-size $\eta > 0$.

(a) Min-norm solution and change of coordinates

Since X has full row rank, $XX^\top$ is invertible and the minimum-norm solution to $X\mathbf{w} = \mathbf{y}$ is obtained by solving the normal equations with a Lagrange multiplier:

$$\mathbf{w}^* = X^\top (XX^\top)^{-1}\mathbf{y}.$$

Define the centered variable $\mathbf{w}' = \mathbf{w} - \mathbf{w}^*$. Then

$$X\mathbf{w}' = X\mathbf{w} - X\mathbf{w}^* = \mathbf{y} - \mathbf{y} = \mathbf{0},$$

so the new system is exactly $X\mathbf{w}' = \mathbf{0}$. With the initial condition $\mathbf{w}_0 = \mathbf{0}$, we have

$$\boxed{\mathbf{w}'_0 = \mathbf{w}_0 - \mathbf{w}^* = -X^\top (XX^\top)^{-1}\mathbf{y}}.$$

Evaluation: Correct. The derivation of the minimum-norm solution $\mathbf{w}^*$ and the centered initial condition $\mathbf{w}'_0 = -\mathbf{w}^*$ is algebraically correct and matches the solution key.

(b) SVD coordinates

Write the thin SVD of X as $X = U\Sigma V_1^{\mathsf{T}}$ where $U \in \mathbb{R}^{n\times n}$ is orthogonal, $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n) \in \mathbb{R}^{n\times n}$ with $\sigma_i > 0$, and $V_1 \in \mathbb{R}^{d\times n}$ has orthonormal columns. Complete V_1 to a full orthogonal matrix $V = [V_1\ V_2] \in \mathbb{R}^{d\times d}$.

Set $\mathbf{w}'' = V^{\mathsf{T}}\mathbf{w}'$ and partition $\mathbf{w}'' = [\widetilde{\mathbf{w}}; \mathbf{w}_\perp]$ with $\widetilde{\mathbf{w}} \in \mathbb{R}^n$ and $\mathbf{w}_\perp \in \mathbb{R}^{d-n}$. Since $X\mathbf{w}' = U\Sigma V_1^{\mathsf{T}}\mathbf{w}' = U\Sigma\widetilde{\mathbf{w}}$, the constraint $X\mathbf{w}' = \mathbf{0}$ is equivalent to $\widetilde{\mathbf{w}} = \mathbf{0}$. Thus the transformed system reads

$$[\,\widetilde{X} \quad \mathbf{0}_{n\times(d-n)}\,]\mathbf{w}'' = \mathbf{0}, \qquad \text{where } \widetilde{X} = U\Sigma.$$

From part (a), $\mathbf{w}_0' = -X^{\mathsf{T}}(XX^{\mathsf{T}})^{-1}\mathbf{y} = -V_1\Sigma^{-1}U^{\mathsf{T}}\mathbf{y}$. Therefore

$$\mathbf{w}_0'' = V^{\mathsf{T}}\mathbf{w}_0' = [V_1^{\mathsf{T}}; V_2^{\mathsf{T}}](-V_1\Sigma^{-1}U^{\mathsf{T}}\mathbf{y}) = [-\Sigma^{-1}U^{\mathsf{T}}\mathbf{y};\ \mathbf{0}].$$

Hence the last $(d-n)$ components of $\mathbf{w}_0''$ are zero:

$$\boxed{\mathbf{w}_{\perp,0} = \mathbf{0}}.$$

Evaluation: Correct. The transformation to SVD coordinates is handled well. The conclusion that the null-space component $\mathbf{w}_{\perp,0}$ initializes to zero is correctly derived from the property that $\mathbf{w}^*$ lies entirely in the row space of X.

(c) Reduction to a square system

Because $\mathbf{w}_\perp$ never appears in the constraint $[\widetilde{X}\ \mathbf{0}]\mathbf{w}'' = \mathbf{0}$ and its initial value is zero, SGD never updates $\mathbf{w}_\perp$ (the gradient in those directions is zero). Consequently we may restrict attention to the first n components $\widetilde{\mathbf{w}}$ which evolve independently of $\mathbf{w}_\perp$.

Define the square system

$$\widetilde{X}\,\widetilde{\mathbf{w}} = \mathbf{0}, \tag{8}$$

where $\widetilde{X} = U\Sigma \in \mathbb{R}^{n\times n}$ is invertible. Each row of (8) is obtained from the corresponding row of the original system $X\mathbf{w} = \mathbf{y}$ by:

i) centering: replace $\mathbf{y}$ with $\mathbf{0}$ and $\mathbf{w}$ with $\mathbf{w}'$;

ii) the orthonormal change of variables $\mathbf{w}'' = V^{\mathsf{T}}\mathbf{w}'$;

iii) restriction to the first n coordinates, which eliminates the trivial null-space directions.

Thus analyzing SGD on (8) is equivalent to analyzing SGD on the original problem.

Evaluation: Correct logic. The reduction to the square system $\widetilde{X}\widetilde{\mathbf{w}} = \mathbf{0}$ is valid because the orthogonal components start at zero and receive zero gradient updates. The explanation of the coordinate changes (centering, rotation) is accurate.

(d) SGD step equivalence

SGD on the original objective draws index $I_t \sim \mathrm{Unif}\{1, \dots, n\}$ and updates

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}_{I_t}(\mathbf{w}_t) = \mathbf{w}_t + 2\eta\big(y[I_t] - \mathbf{x}_{I_t}^{\mathsf{T}} \mathbf{w}_t\big)\mathbf{x}_{I_t}.$$

Using $\mathbf{w}' = \mathbf{w} - \mathbf{w}^*$ and the fact that $\mathbf{x}_{I_t}^{\mathsf{T}} \mathbf{w}^* = y[I_t]$, this becomes

$$\mathbf{w}'_{t+1} = \mathbf{w}'_t - 2\eta\big(\mathbf{x}_{I_t}^{\mathsf{T}} \mathbf{w}'_t\big)\mathbf{x}_{I_t}.$$

Now apply the orthonormal transformation $\mathbf{w}'' = V^{\mathsf{T}} \mathbf{w}'$:

$$\mathbf{w}''_{t+1} = \mathbf{w}''_t - 2\eta\big(\mathbf{x}_{I_t}^{\mathsf{T}} V \mathbf{w}''_t\big) V^{\mathsf{T}} \mathbf{x}_{I_t}.$$

Recall from (b) that $V^{\mathsf{T}} \mathbf{x}_i = \begin{bmatrix} \widetilde{\mathbf{x}}_i \\ \mathbf{0} \end{bmatrix}$ where $\widetilde{\mathbf{x}}_i = \Sigma U^{\mathsf{T}} \mathbf{e}_i \in \mathbb{R}^n$ is the ith row of $\widetilde{X}$ (written as a column). Consequently the update decouples:

$$\begin{bmatrix} \widetilde{\mathbf{w}}_{t+1} \\ \mathbf{w}_{\perp,t+1} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{w}}_t \\ \mathbf{w}_{\perp,t} \end{bmatrix} - 2\eta \begin{bmatrix} (\widetilde{\mathbf{x}}_{I_t}^{\mathsf{T}} \widetilde{\mathbf{w}}_t)\, \widetilde{\mathbf{x}}_{I_t} \\ \mathbf{0} \end{bmatrix}.$$

Since $\mathbf{w}_{\perp,0} = \mathbf{0}$ and it is never updated, we have $\mathbf{w}_{\perp,t} = \mathbf{0}$ for all t. The n-dimensional dynamics are exactly

$$\boxed{\widetilde{\mathbf{w}}_{t+1} = \widetilde{\mathbf{w}}_t - 2\eta\, \widetilde{\mathbf{x}}_{I_t} \widetilde{\mathbf{x}}_{I_t}^{\mathsf{T}} \widetilde{\mathbf{w}}_t}, \tag{10}$$

which is precisely SGD applied to the square system (8).
Evaluation: Correct derivation. The proof that the SGD step on the original system is equivalent to the step on the reduced square system is rigorous. The use of the SVD relation $V^{\top} \mathbf{x}_{I_t}$ to decouple the update is correct.

(e) From conditional expectation to high-probability bound

Assume we have a non-negative stochastic Lyapunov function $\mathcal{L}(\widetilde{\mathbf{w}})$ satisfying

$$\mathbb{E}\big[\mathcal{L}(\widetilde{\mathbf{w}}_{t+1}) \mid \widetilde{\mathbf{w}}_t\big] \le (1-\rho)\, \mathcal{L}(\widetilde{\mathbf{w}}_t), \qquad 1 > \rho > 0, \tag{11}$$

for all $\widetilde{\mathbf{w}}_t$. Taking expectations and iterating,

$$\mathbb{E}[\mathcal{L}(\widetilde{\mathbf{w}}_t)] \le (1-\rho)^t\, \mathcal{L}(\widetilde{\mathbf{w}}_0).$$

Let $\epsilon, \delta > 0$ be arbitrary. By Markov's inequality,

$$\Pr\big(\mathcal{L}(\widetilde{\mathbf{w}}_t) \ge \epsilon\big) \le \frac{\mathbb{E}[\mathcal{L}(\widetilde{\mathbf{w}}_t)]}{\epsilon} \le \frac{(1-\rho)^t\, \mathcal{L}(\widetilde{\mathbf{w}}_0)}{\epsilon}.$$

Choose T large enough that $(1-\rho)^T\,\mathcal{L}(\widetilde{\mathbf{w}}_0)/\epsilon \le \delta$, i.e.

$$T \ge \frac{\ln(\delta\epsilon/\mathcal{L}(\widetilde{\mathbf{w}}_0))}{\ln(1-\rho)}.$$

Since $\ln(1-\rho) < 0$, this is equivalent to

$$T \ge \frac{\ln\bigl(\mathcal{L}(\widetilde{\mathbf{w}}_0)/(\delta\epsilon)\bigr)}{-\ln(1-\rho)}.$$

Such a T always exists, proving that for all $\epsilon, \delta > 0$,

$$\boxed{\Pr\bigl(\mathcal{L}(\widetilde{\mathbf{w}}_T) < \epsilon\bigr) \ge 1-\delta}.$$

Evaluation: Correct approach. The use of Markov's inequality to translate the expectation bound on the Lyapunov function into a high-probability bound is the standard and correct technique for this proof.

(f) Candidate Lyapunov function

Define

$$\mathcal{L}(\widetilde{\mathbf{w}}) = \widetilde{\mathbf{w}}^{\mathsf{T}}\widetilde{X}^{\mathsf{T}}\widetilde{X}\,\widetilde{\mathbf{w}} = \bigl\|\widetilde{X}\,\widetilde{\mathbf{w}}\bigr\|^2. \tag{13}$$

Since it is a squared Euclidean norm, $\mathcal{L}(\widetilde{\mathbf{w}}) \ge 0$ for all $\widetilde{\mathbf{w}}$. Moreover, $\mathcal{L}(\widetilde{\mathbf{w}}) = 0$ iff $\widetilde{X}\,\widetilde{\mathbf{w}} = \mathbf{0}$. Because $\widetilde{X}$ is invertible (its singular values $\sigma_1, \dots, \sigma_n$ are strictly positive), $\widetilde{X}\,\widetilde{\mathbf{w}} = \mathbf{0}$ implies $\widetilde{\mathbf{w}} = \mathbf{0}$. Thus $\mathcal{L}(\widetilde{\mathbf{w}}) \ge 0$ with equality *only* at $\widetilde{\mathbf{w}} = \mathbf{0}$.
Evaluation: Correct. The explanation that $\mathcal{L}(\widetilde{\mathbf{w}})$ is a valid Lyapunov function because $\widetilde{X}$ is invertible (full rank), making the quadratic form positive definite, is correct.

(g) Decomposition of the Lyapunov update

Let $\Delta\widetilde{\mathbf{w}}_t = \widetilde{\mathbf{w}}_{t+1} - \widetilde{\mathbf{w}}_t$. Expanding

$$\mathcal{L}(\widetilde{\mathbf{w}}_{t+1}) = (\widetilde{\mathbf{w}}_t + \Delta\widetilde{\mathbf{w}}_t)^{\mathsf{T}}\widetilde{X}^{\mathsf{T}}\widetilde{X}(\widetilde{\mathbf{w}}_t + \Delta\widetilde{\mathbf{w}}_t)$$

gives the decomposition

$$\mathcal{L}(\widetilde{\mathbf{w}}_{t+1}) = \mathcal{L}(\widetilde{\mathbf{w}}_t) + A + B, \tag{14}$$

where

$$A = 2\,\widetilde{\mathbf{w}}_t^{\mathsf{T}}\widetilde{X}^{\mathsf{T}}\widetilde{X}\,\Delta\widetilde{\mathbf{w}}_t,$$

$$B = \Delta\widetilde{\mathbf{w}}_t^{\mathsf{T}}\widetilde{X}^{\mathsf{T}}\widetilde{X}\,\Delta\widetilde{\mathbf{w}}_t.$$

Using the SGD update (10), $\Delta \widetilde{\mathbf{w}}_t = -2\eta\, \widetilde{\mathbf{x}}_{I_t} \widetilde{\mathbf{x}}_{I_t}^\top \widetilde{\mathbf{w}}_t$, where $\widetilde{\mathbf{x}}_{I_t}$ is the I_t-th column of $\widetilde{X}^\top$. Explicitly,

$$A = -4\eta\, \widetilde{\mathbf{w}}_t^\top \widetilde{X}^\top \widetilde{X}\, \widetilde{\mathbf{x}}_{I_t} \widetilde{\mathbf{x}}_{I_t}^\top \widetilde{\mathbf{w}}_t,$$

$$B = 4\eta^2\, \widetilde{\mathbf{w}}_t^\top \widetilde{\mathbf{x}}_{I_t} \widetilde{\mathbf{x}}_{I_t}^\top \widetilde{X}^\top \widetilde{X}\, \widetilde{\mathbf{x}}_{I_t} \widetilde{\mathbf{x}}_{I_t}^\top \widetilde{\mathbf{w}}_t.$$

Evaluation: Correct expansion. The linear (A) and quadratic (B) terms of the Lyapunov update are correctly identified by expanding the quadratic form.

(h) Bounding the linear term A

Since I_t is uniform on $\{1, \dots, n\}$,

$$\mathbb{E}\left[\widetilde{\mathbf{x}}_{I_t} \widetilde{\mathbf{x}}_{I_t}^\top \mid \widetilde{\mathbf{w}}_t\right] = \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\top = \frac{1}{n} \widetilde{X}^\top \widetilde{X}.$$

Therefore

$$\mathbb{E}[A \mid \widetilde{\mathbf{w}}_t] = -4\eta\, \widetilde{\mathbf{w}}_t^\top \widetilde{X}^\top \widetilde{X} \left(\frac{1}{n} \widetilde{X}^\top \widetilde{X}\right) \widetilde{\mathbf{w}}_t = -\frac{4\eta}{n}\, \widetilde{\mathbf{w}}_t^\top \left(\widetilde{X}^\top \widetilde{X}\right)^2 \widetilde{\mathbf{w}}_t.$$

Let $M = \widetilde{X}^\top \widetilde{X}$ and denote its eigenvalues by $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2 > 0$. Because M is symmetric positive definite,

$$\widetilde{\mathbf{w}}_t^\top M^2 \widetilde{\mathbf{w}}_t \geq \sigma_{\min}^2\, \widetilde{\mathbf{w}}_t^\top M \widetilde{\mathbf{w}}_t = \sigma_{\min}^2\, \mathcal{L}(\widetilde{\mathbf{w}}_t),$$

where $\sigma_{\min} = \sigma_n$ is the smallest singular value of $\widetilde{X}$. Consequently

$$\boxed{\mathbb{E}[A \mid \widetilde{\mathbf{w}}_t] \leq -\frac{4\sigma_{\min}^2}{n}\, \eta\, \mathcal{L}(\widetilde{\mathbf{w}}_t)} = -c_1 \eta\, \mathcal{L}(\widetilde{\mathbf{w}}_t),$$

with $c_1 = 4\sigma_{\min}^2 / n > 0$.
Evaluation: Correct. The expectation of the linear term is correctly bounded using the minimum singular value $\sigma_{\min}^2$ of the active subspace, matching the solution's drift analysis.

(i) Bounding the quadratic term B

From (g) and the uniform sampling of I_t,

$$\mathbb{E}[B \mid \widetilde{\mathbf{w}}_t] = 4\eta^2\, \widetilde{\mathbf{w}}_t^\top \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\top \widetilde{X}^\top \widetilde{X}\, \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\top\right) \widetilde{\mathbf{w}}_t.$$

For any vector $\mathbf{v}$,

$$\mathbf{v}^\top \widetilde{X}^\top \widetilde{X}\, \mathbf{v} \leq \|\widetilde{X}^\top \widetilde{X}\|\, \|\mathbf{v}\|^2 = \sigma_{\max}^2\, \|\mathbf{v}\|^2,$$

where $\sigma_{\max} = \sigma_1$ is the largest singular value of $\widetilde{X}$. Thus the scalar factor in each term satisfies

$$\widetilde{\mathbf{x}}_i^\mathsf{T} \widetilde{X}^\mathsf{T} \widetilde{X}\, \widetilde{\mathbf{x}}_i \le \sigma_{\max}^2 \, \|\widetilde{\mathbf{x}}_i\|^2.$$

Let $\beta = \max_i \|\widetilde{\mathbf{x}}_i\|$ (the largest row norm of $\widetilde{X}$). Then

$$\widetilde{\mathbf{x}}_i^\mathsf{T} \widetilde{X}^\mathsf{T} \widetilde{X}\, \widetilde{\mathbf{x}}_i \le \sigma_{\max}^2 \beta^2.$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\mathsf{T} \widetilde{X}^\mathsf{T} \widetilde{X}\, \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\mathsf{T} \preceq \frac{1}{n}\sum_{i=1}^{n} \sigma_{\max}^2 \beta^2 \, \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\mathsf{T} = \sigma_{\max}^2 \beta^2 \, M.$$

Therefore

$$\boxed{\mathbb{E}[B \mid \widetilde{\mathbf{w}}_t] \le 4\eta^2 \sigma_{\max}^2 \beta^2 \, \widetilde{\mathbf{w}}_t^\mathsf{T} M \widetilde{\mathbf{w}}_t} = c_2 \eta^2 \, \mathcal{L}(\widetilde{\mathbf{w}}_t),$$

with $c_2 = 4\sigma_{\max}^2 \beta^2 > 0$.

Evaluation: Correct. The quadratic noise term is correctly bounded using the maximum singular value $\sigma_{\max}^2$ and the row norm bound, consistent with the solution key.

(j) Exponential decay of the Lyapunov function

Combining the bounds from (h) and (i),

$$\begin{aligned} \mathbb{E}\big[\mathcal{L}(\widetilde{\mathbf{w}}_{t+1}) \mid \widetilde{\mathbf{w}}_t\big] &\le \mathcal{L}(\widetilde{\mathbf{w}}_t) - c_1 \eta \, \mathcal{L}(\widetilde{\mathbf{w}}_t) + c_2 \eta^2 \, \mathcal{L}(\widetilde{\mathbf{w}}_t) \\ &= \big(1 - c_1 \eta + c_2 \eta^2\big) \, \mathcal{L}(\widetilde{\mathbf{w}}_t), \end{aligned}$$

with $c_1, c_2 > 0$ as above. To guarantee contraction we need $1 - c_1\eta + c_2\eta^2 < 1$, which is equivalent to $\eta < c_1/c_2$. Since the right-hand side is strictly positive, such an η always exists. For example, choosing

$$\eta^* = \frac{c_1}{2c_2} = \frac{\sigma_{\min}^2}{2n\sigma_{\max}^2 \beta^2}$$

yields

$$1 - c_1\eta^* + c_2(\eta^*)^2 = 1 - \frac{c_1^2}{4c_2} < 1.$$

Thus there exists a step size $\eta > 0$ for which the Lyapunov function decays exponentially in expectation.

Evaluation: Correct. The final assembly of the bounds to show exponential decay $1 - c_1\eta + c_2\eta^2 < 1$ for small η is correct.

(k) Ridge regression and feature augmentation

The ordinary ridge regression problem

$$\min_{\mathbf{w}} \; \|X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2$$

does *not* satisfy the interpolation condition: the regularizer prevents the loss from reaching zero. Consequently, SGD with a constant step size converges only sublinearly.

The "feature-augmentation" trick rewrites the problem as

$$\min_{\mathbf{w}} \left\| \begin{bmatrix} X \\ \sqrt{\lambda}\, I_d \end{bmatrix} \mathbf{w} - \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \right\|^2,$$

which *is* an interpolation problem: the augmented matrix is tall and, for appropriate λ, has full column rank, so there exists a $\mathbf{w}$ that drives the augmented residual to zero. Applying SGD to this formulation (with a constant step size satisfying the analysis above) yields exponential (linear) convergence.

Typical observations from the Jupyter notebook would be:

- **Direct SGD on ridge regression**: convergence rate is $O(1/t)$ (sublinear), and the step size must be decayed to guarantee convergence.
- **SGD on the augmented formulation**: error decays geometrically as $\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq C \cdot (1-\rho)^t$ for some $\rho \in (0,1)$, i.e. exponential convergence.
- The practical convergence rate ρ depends on the condition number of the augmented matrix $\begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix}$; smaller λ generally leads to faster convergence but may increase the variance of the stochastic gradients.

The dramatic difference between these two implementations illustrates that mathematically equivalent formulations need not be equivalent from an algorithmic standpoint. Understanding these nuances is crucial for efficient large-scale machine learning.

Evaluation: Good summary. The response correctly captures the insight that feature augmentation allows SGD to converge exponentially (like in the interpolating regime) by providing "memory" for each data point, effectively resolving the conflicts that usually slow down SGD.

Solution to Homework 1, Question 3

We consider the least-squares objective $\mathcal{L}(w) = \|y - Xw\|_2^2$ and gradient descent with momentum:

$$\begin{aligned} w_{t+1} &= w_t - \eta z_{t+1}, \\ z_{t+1} &= (1-\beta) z_t + \beta \nabla \mathcal{L}(w_t), \end{aligned} \tag{1}$$

where $g_t = \nabla \mathcal{L}(w_t) = 2X^T(Xw_t - y)$. The minimizer is $w^* = (X^T X)^{-1} X^T y$. Let the SVD of X be $X = U \Sigma V^T$ with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

(a) Reparameterized dynamics

Let $e_t = w_t - w^*$. Since $Xw^* = y$, we have $Xw_t - y = Xe_t$. Substituting into (1),

$$\begin{aligned} e_{t+1} &= e_t - \eta z_{t+1}, \\ z_{t+1} &= (1-\beta)z_t + 2\beta X^T X e_t. \end{aligned}$$

Apply the orthonormal change of variables $x_t = V^T e_t$ and $a_t = V^T z_t$. Using $X^T X = V\Sigma^T \Sigma V^T$,

$$\begin{aligned} x_{t+1} &= x_t - \eta a_{t+1}, \\ a_{t+1} &= (1-\beta)a_t + 2\beta\Sigma^T\Sigma x_t. \end{aligned}$$

Because $\Sigma^T\Sigma$ is diagonal with entries σ_i^2, the system decouples coordinate-wise:

$$\boxed{\begin{aligned} x_{t+1}[i] &= x_t[i] - \eta a_{t+1}[i], \\ a_{t+1}[i] &= (1-\beta)a_t[i] + 2\beta\sigma_i^2 x_t[i]. \end{aligned}} \tag{2}$$

Evaluation: Correct derivation. The reparameterization into eigen-coordinates x_t and a_t matches the solution steps. The update rules for the transformed variables are derived correctly from the momentum equations.

(b) The 2×2 system matrix R_i

Eliminate $a_{t+1}[i]$ from the first equation of (2):

$$x_{t+1}[i] = x_t[i] - \eta\big((1-\beta)a_t[i] + 2\beta\sigma_i^2 x_t[i]\big) = (1 - 2\eta\beta\sigma_i^2)x_t[i] - \eta(1-\beta)a_t[i].$$

Together with the second equation of (2) this yields

$$\begin{bmatrix} a_{t+1}[i] \\ x_{t+1}[i] \end{bmatrix} = \underbrace{\begin{bmatrix} 1-\beta & 2\beta\sigma_i^2 \\ -\eta(1-\beta) & 1-2\eta\beta\sigma_i^2 \end{bmatrix}}_{R_i} \begin{bmatrix} a_t[i] \\ x_t[i] \end{bmatrix}.$$

Hence the system matrix for coordinate i is

$$\boxed{R_i = \begin{bmatrix} 1-\beta & 2\beta\sigma_i^2 \\ -\eta(1-\beta) & 1-2\eta\beta\sigma_i^2 \end{bmatrix}}. \tag{3}$$

Evaluation: Correct. The system matrix R_i is correctly identified. The entries $[(1-\beta), 2\beta\sigma_i^2]$ and $[-\eta(1-\beta), 1-2\eta\beta\sigma_i^2]$ (or their transpose equivalent depending on vector definition) match the dynamics described in the solution.

(c) Eigenvalues of R_i

The characteristic polynomial of R_i is

$$\det(R_i - \lambda I) = \lambda^2 - T_i\lambda + (1-\beta) = 0,$$

where $T_i = \mathrm{tr}(R_i) = 2 - \beta - 2\eta\beta\sigma_i^2$. The eigenvalues are

$$\lambda_{i,\pm} = \frac{T_i \pm \sqrt{T_i^2 - 4(1-\beta)}}{2}. \tag{4}$$

The discriminant is $D_i = T_i^2 - 4(1-\beta)$.

- **Purely real:** $D_i > 0 \iff |T_i| > 2\sqrt{1-\beta}$.
- **Repeated and purely real:** $D_i = 0 \iff |T_i| = 2\sqrt{1-\beta}$.
- **Complex:** $D_i < 0 \iff |T_i| < 2\sqrt{1-\beta}$.

Evaluation: Correct. The characteristic equation and the resulting eigenvalues λ are correctly derived. The conditions for real vs. complex eigenvalues based on the discriminant are also correct.

(d) Repeated eigenvalues: stability and maximal η

When $D_i = 0$, the eigenvalue is $\lambda_i = T_i/2$. The condition $D_i = 0$ gives

$$2 - \beta - 2\eta\beta\sigma_i^2 = \pm 2\sqrt{1-\beta} \quad \Longrightarrow \quad \boxed{\eta = \frac{2 - \beta \mp 2\sqrt{1-\beta}}{2\beta\sigma_i^2}}. \tag{5}$$

For stability we require $|\lambda_i| < 1$, i.e. $|T_i| < 2$. Since $|T_i| = 2\sqrt{1-\beta}$, this is equivalent to $\sqrt{1-\beta} < 1$, which holds for all $\beta \in (0,1)$. Hence repeated eigenvalues are strictly stable inside the unit circle for any $\beta > 0$.

The larger learning rate (corresponding to the minus sign in (5)) is

$$\boxed{\eta_{\text{repeat}}(\sigma_i) = \frac{2 - \beta - 2\sqrt{1-\beta}}{2\beta\sigma_i^2}}. \tag{6}$$

Evaluation: Correct. The analysis that repeated/complex eigenvalues lie on a circle of radius $\sqrt{1-\beta}$ (hence always stable for $\beta \in (0,1)$) matches the solution. The condition for the maximal η for repeated roots is also correct.

(e) Real eigenvalues: stability and range of η

For real eigenvalues we need $D_i > 0 \iff |T_i| > 2\sqrt{1-\beta}$. Stability requires both eigenvalues to satisfy $|\lambda_{i,\pm}| < 1$. For a quadratic $\lambda^2 - T_i\lambda + (1-\beta) = 0$ with $0 < \beta < 1$, the Jury stability criteria reduce to

$$|T_i| < 1 + (1-\beta) = 2 - \beta.$$

Since $T_i = 2 - \beta - 2\eta\beta\sigma_i^2$, this yields

$$|2 - \beta - 2\eta\beta\sigma_i^2| < 2 - \beta \quad \Longrightarrow \quad \boxed{0 < \eta < \frac{2-\beta}{\beta\sigma_i^2}}. \tag{7}$$

Thus for fixed β, the admissible learning rate is inversely proportional to σ_i^2.

The condition $D_i > 0$ gives two disjoint η-intervals:

$$0 < \eta < \frac{2-\beta-2\sqrt{1-\beta}}{2\beta\sigma_i^2} \quad \text{or} \quad \eta > \frac{2-\beta+2\sqrt{1-\beta}}{2\beta\sigma_i^2}.$$

Only the first interval satisfies (7), so the stable real regime is

$$\boxed{0 < \eta < \frac{2-\beta-2\sqrt{1-\beta}}{2\beta\sigma_i^2}}. \tag{8}$$

Evaluation: Correct. The stability analysis for real eigenvalues (requiring them to stay within $(-1,1)$) leads to the correct bounds on η. The derivations align with the Jury stability conditions or direct root analysis.

(f) Complex eigenvalues: stability and maximal η

Complex eigenvalues occur when $D_i < 0 \iff |T_i| < 2\sqrt{1-\beta}$, which translates to

$$\boxed{\frac{2-\beta-2\sqrt{1-\beta}}{2\beta\sigma_i^2} < \eta < \frac{2-\beta+2\sqrt{1-\beta}}{2\beta\sigma_i^2}}. \tag{9}$$

For complex conjugate pairs, $|\lambda_{i,\pm}|^2 = 1 - \beta$. Hence the spectral radius is $\sqrt{1-\beta}$, independent of η and σ_i. Stability requires $\sqrt{1-\beta} < 1$, which holds for all $\beta \in (0,1)$. The upper bound in (9) is therefore the maximal learning rate that keeps the eigenvalues complex:

$$\boxed{\eta_{\text{complex,max}}(\sigma_i) = \frac{2-\beta+2\sqrt{1-\beta}}{2\beta\sigma_i^2}}. \tag{10}$$

Evaluation: Correct. The observation that complex eigenvalues have constant magnitude $\sqrt{1-\beta}$ and are thus stable is correct. The range for η yielding complex roots is correctly identified.

(g) Optimal learning rate for a concrete problem

Let $\beta = 0.1$, $\sigma_{\max}^2 = 5$ and $\sigma_{\min}^2 = 0.05$ (condition number $\kappa = 100$).

Momentum GD: The fastest overall convergence is achieved by making the worst-case direction converge as quickly as possible. This occurs when the eigenvalues for the *largest* curvature ($\sigma_{\max}$) are just at the boundary of the complex region, because the complex region gives a constant spectral radius $\sqrt{1-\beta}$ that is smaller than any real eigenvalue magnitude in the stable regime. Using (10) with $\sigma_{\max}$,

$$\eta_{\text{mom}}^* = \frac{2 - 0.1 + 2\sqrt{1-0.1}}{2 \cdot 0.1 \cdot 5} = \frac{1.9 + 2\sqrt{0.9}}{1} \approx \frac{1.9 + 1.8974}{1} = \boxed{3.7974}.$$

With this choice, the $i = \max$ direction has eigenvalues $\lambda_{\max} = -\sqrt{1-\beta} \approx -0.9487$ (repeated), while the $i = \min$ direction is safely inside the complex region with $|\lambda_{\min}| = \sqrt{1-\beta} \approx 0.9487$. Hence the overall spectral radius is $\rho_{\text{mom}} = \sqrt{1-\beta} = \sqrt{0.9} \approx 0.9487$.

The number of iterations to reach $\varepsilon = 0.001$ (i.e., 99.9% of optimum) satisfies $\rho_{\text{mom}}^t \le \varepsilon$, so

$$t_{\text{mom}} \ge \frac{\ln \varepsilon}{\ln \rho_{\text{mom}}} = \frac{\ln(10^{-3})}{\ln(\sqrt{0.9})} \approx \frac{-6.9078}{-0.05268} \approx \boxed{131 \text{ iterations}}.$$

Ordinary GD: The optimal step size and rate are

$$\eta_{\text{gd}}^* = \frac{1}{\sigma_{\max}^2 + \sigma_{\min}^2} = \frac{1}{5 + 0.05} \approx 0.198, \qquad \rho_{\text{gd}} = \frac{\kappa - 1}{\kappa + 1} = \frac{100 - 1}{100 + 1} \approx 0.9802.$$

Iterations needed:

$$t_{\text{gd}} \ge \frac{\ln(10^{-3})}{\ln(0.9802)} \approx \frac{-6.9078}{-0.0200} \approx \boxed{345 \text{ iterations}}.$$

Conclusion: Momentum GD converges roughly $345/131 \approx 2.6$ times faster than ordinary GD on this problem.

Evaluation: Excellent. The calculation of the optimal η by ensuring the eigenvalues for both $\sigma_{\min}$ and $\sigma_{\max}$ are complex (or repeated) matches the advanced analysis in the solution. The estimated iteration count is consistent with the theoretical acceleration provided by momentum.

Question 4: Optimizers

(a) Completing the Adam pseudocode

The Adam optimizer maintains two exponentially weighted averages:

- m_t: first moment (mean) of the gradient
- v_t: second moment (uncentered variance) of the gradient

The missing update rules are:

$$\begin{aligned} &\text{(A)} \quad m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t, \\ &\text{(B)} \quad v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \end{aligned}$$

where g_t^2 denotes elementwise squaring.
Evaluation: Correct. The update rules for the first moment m_t (Equation A) and second moment v_t (Equation B) are correctly identified as exponential moving averages, which is the standard definition of Adam.

(b) Equivalence of weight decay and L2 regularization

Weight decay update:

$$\theta_{t+1} = (1-\gamma)\theta_t - \eta \nabla f(\theta_t). \tag{1}$$

L2-regularized loss:

$$f_t^{\text{reg}}(\theta) = f_t(\theta) + \frac{\lambda}{2}\|\theta\|_2^2.$$

SGD on this loss gives:

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \nabla f_t^{\text{reg}}(\theta_t) \\ &= \theta_t - \eta\big(\nabla f(\theta_t) + \lambda \theta_t\big) \\ &= (1-\eta\lambda)\theta_t - \eta \nabla f(\theta_t). \end{aligned} \tag{2}$$

Comparing (1) and (2), we see they are identical when:

$$\boxed{\gamma = \eta\lambda}.$$

Evaluation: Correct derivation. The mathematical equivalence between the weight decay update and the L2-regularized SGD update is correctly shown. The relation $\gamma = \eta\lambda$ is accurately derived.

Question 5: Regularization and Instance Noise

(a) Equivalence to ridge regression

Noisy data: $\check{\mathbf{X}}_i = \mathbf{x}_i + \mathbf{N}_i$, $\mathbf{N}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. The objective is:

$$\arg\min_{\mathbf{w}} \mathbb{E}\big[\|\check{X}\mathbf{w} - \mathbf{y}\|^2\big],$$

where $\check{X}$ has rows $\check{\mathbf{X}}_i^\top$.

Expanding the squared norm:

$$\begin{aligned}\mathbb{E}\big[\|\check{X}\mathbf{w}-\mathbf{y}\|^2\big] &= \mathbb{E}\big[\mathbf{w}^\top\check{X}^\top\check{X}\mathbf{w}-2\mathbf{y}^\top\check{X}\mathbf{w}+\|\mathbf{y}\|^2\big] \\ &= \mathbf{w}^\top\mathbb{E}[\check{X}^\top\check{X}]\mathbf{w}-2\mathbf{y}^\top\mathbb{E}[\check{X}]\mathbf{w}+\|\mathbf{y}\|^2.\end{aligned}$$

Since $\mathbb{E}[\check{\mathbf{X}}_i]=\mathbf{x}_i$, we have $\mathbb{E}[\check{X}]=X$. For the second moment:

$$\begin{aligned}\check{X}^\top\check{X} &= \sum_{i=1}^m \check{\mathbf{X}}_i\check{\mathbf{X}}_i^\top = \sum_{i=1}^m(\mathbf{x}_i+\mathbf{N}_i)(\mathbf{x}_i+\mathbf{N}_i)^\top, \\ \mathbb{E}[\check{X}^\top\check{X}] &= \sum_{i=1}^m\big(\mathbf{x}_i\mathbf{x}_i^\top+\mathbb{E}[\mathbf{N}_i\mathbf{N}_i^\top]\big) = X^\top X+m\sigma^2 I_n.\end{aligned}$$

Substituting back:

$$\begin{aligned}\mathbb{E}\big[\|\check{X}\mathbf{w}-\mathbf{y}\|^2\big] &= \mathbf{w}^\top(X^\top X+m\sigma^2 I_n)\mathbf{w}-2\mathbf{y}^\top X\mathbf{w}+\|\mathbf{y}\|^2 \\ &= \|X\mathbf{w}-\mathbf{y}\|^2+m\sigma^2\|\mathbf{w}\|^2 \\ &= m\left(\frac{1}{m}\|X\mathbf{w}-\mathbf{y}\|^2+\sigma^2\|\mathbf{w}\|^2\right).\end{aligned}$$

Ignoring the constant factor m, this is ridge regression with:

$$\boxed{\lambda=\sigma^2}.$$

Comments:
Evaluation: Correct. The derivation shows that minimizing the expected loss with additive noise $\mathbf{N}_i$ introduces a term $\mathbb{E}[\|\mathbf{N}_i\|^2]$ which acts exactly as an L2 penalty on the weights. The identification $\lambda=\sigma^2$ is correct.

(b) Gradient descent recurrence for scalar case

For a single scalar datapoint x with noise $N_t\sim\mathcal{N}(0,\sigma^2)$, the noisy input is $\check{X}_t=x+N_t$. The loss is $\mathcal{L}(w)=\frac{1}{2}(\check{X}_t w-y)^2$. The gradient descent update is:

$$\begin{aligned}w_{t+1} &= w_t-\eta\nabla\mathcal{L}(w_t) \\ &= w_t-\eta(\check{X}_t w_t-y)\check{X}_t \\ &= (1-\eta\check{X}_t^2)w_t+\eta y\check{X}_t.\end{aligned}$$

Taking expectation conditioned on w_t (but not on the current noise N_t):

$$\begin{aligned}\mathbb{E}[w_{t+1}] &= \mathbb{E}\big[(1-\eta(x+N_t)^2)w_t\big]+\eta y\mathbb{E}[x+N_t] \\ &= (1-\eta\mathbb{E}[(x+N_t)^2])\mathbb{E}[w_t]+\eta xy \\ &= \big(1-\eta(x^2+\sigma^2)\big)\mathbb{E}[w_t]+\eta xy.\end{aligned}$$

Thus the recurrence is:

$$\boxed{\mathbb{E}[w_{t+1}] = \big(1 - \eta(x^2 + \sigma^2)\big)\mathbb{E}[w_t] + \eta xy}.$$

Comments: Evaluation: Correct. The recurrence for the expected weight $\mathbb{E}[w_{t+1}]$ is correctly derived. The analysis properly accounts for the noise variance in the second moment of the input $E[(x+N)^2] = x^2 + \sigma^2$.

(c) Convergence condition for learning rate

The recurrence from part (b) is stable iff the contraction factor satisfies:

$$|\, 1 - \eta(x^2 + \sigma^2)\, | < 1.$$

This gives:

$$\boxed{0 < \eta < \frac{2}{x^2 + \sigma^2}}.$$

Evaluation: Correct. The stability condition is correctly updated to reflect the increased "effective" curvature due to noise: $\eta < \frac{2}{x^2+\sigma^2}$.

(d) Limiting value and comparison to noise-free optimum

The fixed point w^* satisfies:

$$\begin{aligned} w^* &= \big(1 - \eta(x^2 + \sigma^2)\big)w^* + \eta xy \\ \eta(x^2 + \sigma^2)w^* &= \eta xy \\ w^* &= \frac{xy}{x^2 + \sigma^2} = \frac{x^2}{x^2 + \sigma^2} \cdot \frac{y}{x}. \end{aligned}$$

The noise-free optimum is $w_{\text{opt}} = y/x$. Therefore:

$$\boxed{\mathbb{E}[w_\infty] = \frac{x^2}{x^2 + \sigma^2}\, w_{\text{opt}}}.$$

The expectation converges to a *shrunken* version of the true optimum, with shrinkage factor $\frac{x^2}{x^2+\sigma^2} < 1$. The noise acts as a regularizer, biasing the solution toward zero.
Evaluation: Correct. The fixed point analysis shows the "shrinkage" effect where the solution converges to a value smaller than the true optimum y/x by a factor of $\frac{x^2}{x^2+\sigma^2}$, which matches the behavior of Ridge Regression.

Question 6: General Case Tikhonov Regularization

(a) Manual solution

The objective is:

$$f(\mathbf{x}) = \|W_1(A\mathbf{x} - \mathbf{b})\|_2^2 + \|W_2(\mathbf{x} - \mathbf{c})\|_2^2.$$

Expanding both terms:

$$\begin{aligned} f(\mathbf{x}) &= (A\mathbf{x} - \mathbf{b})^\top W_1^\top W_1 (A\mathbf{x} - \mathbf{b}) + (\mathbf{x} - \mathbf{c})^\top W_2^\top W_2 (\mathbf{x} - \mathbf{c}) \\ &= \mathbf{x}^\top A^\top W_1^\top W_1 A\mathbf{x} - 2\mathbf{b}^\top W_1^\top W_1 A\mathbf{x} + \mathbf{b}^\top W_1^\top W_1 \mathbf{b} \\ &\quad + \mathbf{x}^\top W_2^\top W_2 \mathbf{x} - 2\mathbf{c}^\top W_2^\top W_2 \mathbf{x} + \mathbf{c}^\top W_2^\top W_2 \mathbf{c}. \end{aligned}$$

The gradient is:

$$\nabla f(\mathbf{x}) = 2A^\top W_1^\top W_1 A\mathbf{x} - 2A^\top W_1^\top W_1 \mathbf{b} + 2W_2^\top W_2 \mathbf{x} - 2W_2^\top W_2 \mathbf{c}.$$

Setting $\nabla f(\mathbf{x}) = \mathbf{0}$ and solving:

$$(A^\top W_1^\top W_1 A + W_2^\top W_2)\mathbf{x} = A^\top W_1^\top W_1 \mathbf{b} + W_2^\top W_2 \mathbf{c}.$$

The solution is:

$$\boxed{\mathbf{x}^* = (A^\top W_1^\top W_1 A + W_2^\top W_2)^{-1}(A^\top W_1^\top W_1 \mathbf{b} + W_2^\top W_2 \mathbf{c})}.$$

Evaluation: Correct. The manual derivation of the gradient and the closed-form solution $\mathbf{x}^*$ involves correct matrix calculus and aligns with the generalized normal equations.

(b) OLS formulation

Construct:

$$C = \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix}, \qquad \mathbf{d} = \begin{bmatrix} W_1 \mathbf{b} \\ W_2 \mathbf{c} \end{bmatrix}.$$

Then:

$$\begin{aligned} \|C\mathbf{x} - \mathbf{d}\|^2 &= \left\| \begin{bmatrix} W_1 A\mathbf{x} - W_1 \mathbf{b} \\ W_2 \mathbf{x} - W_2 \mathbf{c} \end{bmatrix} \right\|^2 \\ &= \|W_1(A\mathbf{x} - \mathbf{b})\|^2 + \|W_2(\mathbf{x} - \mathbf{c})\|^2. \end{aligned}$$

The OLS solution is:

$$\mathbf{x}^* = (C^\top C)^{-1} C^\top \mathbf{d} = (A^\top W_1^\top W_1 A + W_2^\top W_2)^{-1}(A^\top W_1^\top W_1 \mathbf{b} + W_2^\top W_2 \mathbf{c}),$$

which matches part (a).

Evaluation: Correct. The construction of the stacked matrix C and vector $\mathbf{d}$ effectively maps the generalized problem to a standard Ordinary Least Squares problem, yielding the same solution.

(c) Reduction to standard ridge regression

Standard ridge regression: $\mathbf{x}^* = (A^\top A + \lambda I)^{-1} A^\top \mathbf{b}$.

Choose:

$$W_1 = I, \qquad W_2 = \sqrt{\lambda} I, \qquad \mathbf{c} = \mathbf{0}.$$

Then:

$$\begin{aligned} A^\top W_1^\top W_1 A + W_2^\top W_2 &= A^\top A + \lambda I, \\ A^\top W_1^\top W_1 \mathbf{b} + W_2^\top W_2 \mathbf{c} &= A^\top \mathbf{b}. \end{aligned}$$

Substituting into the solution from (a):

$$\boxed{\mathbf{x}^* = (A^\top A + \lambda I)^{-1} A^\top \mathbf{b}}.$$

Evaluation: Correct. The choice of matrices $W_1 = I$, $W_2 = \sqrt{\lambda} I$ and $\mathbf{c} = \mathbf{0}$ correctly reduces the general Tikhonov form to the specific standard Ridge Regression objective.

Question 7: Alternate MAP Interpretation of Ridge Regression

We show the equivalence of:

$$\hat{\mathbf{w}}_1 = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y} \quad \text{and} \quad \hat{\mathbf{w}}_2 = X^\top (X X^\top + \lambda I)^{-1} \mathbf{y}.$$

Matrix algebra proof

Multiply both sides of the identity

$$(X^\top X + \lambda I) X^\top = X^\top (X X^\top + \lambda I)$$

by $(X^\top X + \lambda I)^{-1}$ on the left and $(X X^\top + \lambda I)^{-1}$ on the right:

$$X^\top (X X^\top + \lambda I)^{-1} = (X^\top X + \lambda I)^{-1} X^\top.$$

Multiplying by $\mathbf{y}$ yields $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$.

Probabilistic (MAP) proof

Model specification:

$$\begin{aligned} \mathbf{W} &\sim \mathcal{N}(\mathbf{0}, I_d), \\ \mathbf{Y} \mid \mathbf{W} &= X\mathbf{W} + \sqrt{\lambda} \mathbf{N}, \quad \mathbf{N} \sim \mathcal{N}(\mathbf{0}, I_n). \end{aligned}$$

The joint distribution is Gaussian:

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{WW} & \Sigma_{WY} \\ \Sigma_{YW} & \Sigma_{YY} \end{bmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} I_d & X^\top \\ X & XX^\top + \lambda I_n \end{bmatrix}\right).$$

The MAP estimate given $\mathbf{Y} = \mathbf{y}$ equals the conditional mean:

$$\hat{\mathbf{w}} = \mathbb{E}[\mathbf{W} \mid \mathbf{Y} = \mathbf{y}] = \Sigma_{WY}\Sigma_{YY}^{-1}\mathbf{y} = X^\top(XX^\top + \lambda I_n)^{-1}\mathbf{y}.$$

This is $\hat{\mathbf{w}}_2$.

The standard ridge solution $\hat{\mathbf{w}}_1$ is obtained by maximizing the log-posterior:

$$\hat{\mathbf{w}}_1 = \arg\min_{\mathbf{w}}\left\{\frac{1}{\lambda}\|X\mathbf{w} - \mathbf{y}\|^2 + \|\mathbf{w}\|^2\right\} = \arg\min_{\mathbf{w}}\left\{\|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2\right\}.$$

Both formulations yield the same optimal $\mathbf{w}$ because they describe the same underlying probabilistic model, just conditioned on different variables. The matrix identity above confirms their algebraic equivalence.

Evaluation: Correct and rigorous. The response provides two valid proofs: an algebraic one using matrix identities (Woodbury/inversion lemma variants) and a probabilistic one using the properties of Joint Gaussians and conditional expectations. Both correctly conclude that $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$.