

I'd like to ask you some questions about optimization. I will give you one problem at a time as images (note that a question may contain multiple subproblems). I want you to (1) interpret the problem, (2) produce the solution, and (3) show all step-by-step reasoning/ derivations...

◆ Show thinking ▾

I understand the task. I am ready to help you with your optimization problems.

Please upload the first question. I will:

1. **Interpret the problem** from the image. 
2. **Produce the solution** for each part.
3. **Show all step-by-step reasoning and derivations** clearly and concisely, explicitly stating which subproblem I am addressing.

Go ahead and share the first image!



This is the first problem

Show thinking ▾

Here are the step-by-step solutions for each part of your problem.

Part (a): Stability of the Recurrence

Goal: Determine the values of the learning rate $\eta > 0$ for which the recurrence relation $w_{t+1} = (1 - 2\eta\sigma^2)w_t + 2\eta\sigma y$ is stable.

Reasoning:

1. This is a linear difference equation of the form $w_{t+1} = Aw_t + B$, where $A = (1 - 2\eta\sigma^2)$ and $B = 2\eta\sigma y$.
2. For a linear recurrence $w_{t+1} = Aw_t + B$ to converge (be stable), the magnitude of the multiplicative factor A must be strictly less than 1. This relates to the hint about the "unit circle"—if $|A| > 1$, the value will grow exponentially; if $|A| = 1$, it will not converge to a single value (it might oscillate or stay constant).
3. Therefore, the stability condition is:

$$|1 - 2\eta\sigma^2| < 1$$

4. This inequality splits into two parts:

$$-1 < 1 - 2\eta\sigma^2 < 1$$

Derivation:

- Right inequality:

$$1 - 2\eta\sigma^2 < 1$$

$$-2\eta\sigma^2 < 0$$

Since $\sigma > 0$ and we are given $\eta > 0$, this is always true for positive learning rates.

- Left inequality:

$$-1 < 1 - 2\eta\sigma^2$$

$$2\eta\sigma^2 < 2$$

$$\eta\sigma^2 < 1$$

$$\eta < \frac{1}{\sigma^2}$$

Solution (a): For the recurrence to be stable, the learning rate must satisfy:

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

Part (b): Convergence Speed

Goal: Determine how many iterations t are needed to get within a factor $(1 - \epsilon)$ of the optimal solution $w^* = \frac{y}{\sigma}$.

Reasoning:

1. Let's rewrite the recurrence in terms of the error. Let error $e_t = w_t - w^*$.
2. We know the optimal solution satisfies $\sigma w^* = y$. Substituting $y = \sigma w^*$ into the update rule:

$$w_{t+1} = (1 - 2\eta\sigma^2)w_t + 2\eta\sigma(\sigma w^*)$$

$$w_{t+1} = (1 - 2\eta\sigma^2)w_t + 2\eta\sigma^2 w^*$$

3. Subtract w^* from both sides:

$$w_{t+1} - w^* = (1 - 2\eta\sigma^2)w_t + 2\eta\sigma^2 w^* - w^*$$

$$e_{t+1} = (1 - 2\eta\sigma^2)w_t - (1 - 2\eta\sigma^2)w^*$$

$$e_{t+1} = (1 - 2\eta\sigma^2)(w_t - w^*)$$

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

4. This means the error at step t is:

$$e_t = (1 - 2\eta\sigma^2)^t e_0$$

5. We want the relative error to be within $(1 - \epsilon)$. In optimization contexts, "within a factor $(1 - \epsilon)$ " usually implies we want the *ratio* of current error to initial error to be small, specifically ϵ . However, the phrasing "within a factor $(1 - \epsilon)$ of w^* " is slightly ambiguous. The standard interpretation for "getting within ϵ " is that the error ratio is ϵ . Let's solve for the ratio $\frac{|e_t|}{|e_0|} \leq \epsilon$.

Derivation: We want $|1 - 2\eta\sigma^2|^t \leq \epsilon$. Take the natural log of both sides:

$$t \ln |1 - 2\eta\sigma^2| \leq \ln \epsilon$$

$$t \geq \frac{\ln \epsilon}{\ln |1 - 2\eta\sigma^2|}$$

Note: If the prompt implies reaching $(1 - \epsilon)w^$, we are looking for the error to drop to ϵ relative to the start (assuming start is 0).**

Solution (b): The number of iterations t required is roughly:

$$t \approx \frac{\ln \epsilon}{\ln |1 - 2\eta\sigma^2|}$$

(Note: Since $\ln(\text{fraction})$ is negative, dividing by it flips the inequality \leq to \geq .) 

Part (c): Vector Convergence Stability

Goal: For the vector problem with large σ_ℓ and tiny σ_s ($\sigma_\ell \gg \sigma_s$), find the range of η for convergence and identify the limiting dimension.

Reasoning:



1. The system is diagonal (decoupled). The update for $w[1]$ depends only on σ_ℓ , and $w[2]$ depends only on σ_s .

$$w[1]_{t+1} = (1 - 2\eta\sigma_\ell^2)w[1]_t + 2\eta\sigma_\ell y[1]$$

$$w[2]_{t+1} = (1 - 2\eta\sigma_s^2)w[2]_t + 2\eta\sigma_s y[2]$$

2. For the whole system to converge, **both** individual components must be stable.
3. From part (a), the stability conditions are:
 - For dimension 1: $\eta < \frac{1}{\sigma_\ell^2}$
 - For dimension 2: $\eta < \frac{1}{\sigma_s^2}$
4. Since $\sigma_\ell \gg \sigma_s$, it follows that $\sigma_\ell^2 \gg \sigma_s^2$, and therefore $\frac{1}{\sigma_\ell^2} \ll \frac{1}{\sigma_s^2}$.
5. We must satisfy the stricter constraint (the smaller upper bound).

Solution (c): The range for convergence is:

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

The dimension with the **large eigenvalue** (σ_ℓ) limits our learning rate. We cannot use a large step size because the "steep" direction (σ_ℓ) would become unstable.



Part (d): Fast vs. Slow Dimensions

Goal: Identify which dimension converges faster/slower given the η constraint.

Reasoning:

1. Assume we pick a valid learning rate, e.g., $\eta \approx \frac{1}{\sigma_\ell^2}$ (close to the max stable limit).
2. The convergence rate is determined by the factor $\rho = |1 - 2\eta\sigma^2|$. Closer to 0 is fast; closer to 1 is slow.
3. **For the large σ_ℓ :** If η is close to $1/\sigma_\ell^2$, the factor is $|1 - 2(1/\sigma_\ell^2)\sigma_\ell^2| = |-1| = 1$ (oscillation). If we choose optimal $\eta = 1/(2\sigma_\ell^2)$, the factor is 0 (instant convergence). Generally, this direction changes *rapidly*.
4. **For the small σ_s :** With η constrained to be very small (to satisfy σ_ℓ), the term $2\eta\sigma_s^2$ is extremely small because both η and σ_s are small. Factor: $|1 - 2\eta\sigma_s^2| \approx 1$. This means the error decays very slowly (e.g., 0.9999^t).

Solution (d):

- **Faster Convergence:** The dimension corresponding to σ_ℓ (**large sigma**).
- **Slower Convergence:** The dimension corresponding to σ_s (**small sigma**).

Part (e): Fastest Overall Convergence

Goal: Find the optimal η that maximizes the speed of the slowest component (since the overall time is dominated by the slowest part).

Reasoning:

1. We have two convergence rates: $\rho_\ell = |1 - 2\eta\sigma_\ell^2|$ and $\rho_s = |1 - 2\eta\sigma_s^2|$.
2. We want to minimize $\max(\rho_\ell, \rho_s)$.
3. ρ_s decreases as η increases (approaching 0 from 1). 
4. ρ_ℓ initially decreases as η increases, hits 0, and then increases (approaching |-1|).
5. The optimal trade-off occurs when the rates are balanced on opposite sides of zero (one is positive, one is negative, but magnitudes are equal).

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

$$1 - 2\eta\sigma_s^2 = -1 + 2\eta\sigma_\ell^2$$

$$2 = 2\eta(\sigma_\ell^2 + \sigma_s^2)$$

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

Solution (e): The value of η for best overall convergence is:

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

Part (f): Influence of Intermediate σ_i

Goal: Determine if having more parallel problems with σ_i between σ_ℓ and σ_s changes the choice of η .

Reasoning:

1. Convergence is limited by the "edges" of the spectrum:
 - **Stability** is limited by the largest value (σ_ℓ).
 - **Speed** is limited by the smallest value (σ_s) because it requires the most steps to traverse.
2. Any σ_i strictly between σ_s and σ_ℓ will have a convergence factor $|1 - 2\eta\sigma_i^2|$ that is strictly smaller than the maximum of the edge cases.
3. Therefore, the intermediate values do not bottleneck the process; the "worst-case" constraints are already defined by the min and max eigenvalues.

Solution (f): No, they would **not influence** the choice of the optimal learning rate. The learning rate is constrained by the largest curvature (σ_ℓ) for stability and the smallest curvature (σ_s) for convergence speed. The "condition number" (ratio of largest to smallest singular value) dictates the difficulty, regardless of how many intermediate values exist.

Part (g): Relevance to SVD and General Least Squares

Goal: Connect this scalar analysis to general Least Squares $Xw \approx y$.

Reasoning:

1. The general Least Squares objective is $\|Xw - y\|^2$. The Hessian (curvature) of this objective is $2X^T X$.
2. Using Singular Value Decomposition (SVD), we can write $X = U\Sigma V^T$.
3. We can rotate the coordinate system into the basis of the singular vectors V . In this rotated basis, the variables become decoupled (diagonalized), exactly like the "parallel problems" in part (c) and (f).
4. The values σ_i and σ_s correspond to the largest and smallest singular values of the data matrix X .

Solution (g): This analysis is directly relevant because performing SVD on X allows us to diagonalize the general least squares problem into independent scalar problems along the principal components (eigenvectors of $X^T X$).

- The convergence speed of Gradient Descent on the general problem is determined by the condition number $\kappa = \frac{\sigma_{\max}}{\sigma_{\min}}$ of the matrix X .
- Gradient descent will struggle to converge quickly if the matrix X is "ill-conditioned" (large difference between largest and smallest singular values), exactly as shown in the parallel scalar analysis.

Would you like me to tackle another problem or clarify any of these steps?



I want you to tackle the next problem.

◆ Show thinking ▾

This is a comprehensive problem about the convergence of Stochastic Gradient Descent (SGD) in the interpolating regime (where a solution achieving zero loss exists). This problem walks through coordinate transformations, the behavior of SGD updates, and proving exponential convergence using a Lyapunov function.

I will break this down into manageable parts as presented in the problem statement.

Part (a): Coordinate Transformation to Min-Norm Solution

Goal:

1. Express the min-norm solution w^* explicitly with respect to X and y .
2. Change coordinates to $w' = w - w^*$ to rewrite the equation $Xw = y$ as $Xw' = 0$.
3. Determine the new initial condition w'_0 .

1. Min-Norm Solution w^* : The problem states X is a wide matrix ($d > n$) with full row rank. The equation is $Xw = y$. The minimum norm solution for an underdetermined system (wide matrix) is given by the pseudo-inverse formula:

$$w^* = X^T(XX^T)^{-1}y$$

(Reasoning: w^* must lie in the row space of X . Any component in the null space would increase the norm without affecting Xw . So $w^* = X^T\alpha$. Then $X(X^T\alpha) = y \Rightarrow \alpha = (XX^T)^{-1}y$.)

2. Change Coordinates: We define $w' = w - w^*$. Substitute $w = w' + w^*$ into the original equation $Xw = y$:

$$X(w' + w^*) = y$$

$$Xw' + Xw^* = y$$

Since $Xw^* = y$, this simplifies to:

$$Xw' + y = y \implies Xw' = 0$$

This matches equation (6) in the prompt.

3. Initial Condition w'_0 : We are given $w_0 = 0$. Substitute this into the definition of w' :

$$w'_0 = w_0 - w^* = 0 - w^* = -w^*$$

So, the initial condition is:

$$w'_0 = -X^T(XX^T)^{-1}y$$

Part (b): SVD Coordinate Transformation

Goal: Show that there exists an orthonormal transformation V such that the system simplifies to $[\tilde{X} \quad 0_{n \times (d-n)}]w'' = 0$, and show w'' has zeros in the last $(d-n)$ positions.

1. SVD Setup: Let the Full SVD of X be $X = U\Sigma V^T$.

- U is $n \times n$ orthogonal.
- Σ is $n \times d$ diagonal-like matrix $[\tilde{\Sigma} \quad 0]$ where $\tilde{\Sigma}$ is $n \times n$ diagonal with positive entries (full row rank).
- V is $d \times d$ orthogonal.

2. Transformation: We have $Xw' = 0$. Substitute the SVD:

$$U\Sigma V^T w' = 0$$

Since U is invertible (orthogonal), we can multiply by U^T on the left to remove it:



$$\Sigma(V^T w') = 0$$

Let's define the new variable $w'' = V^T w'$ (which means $w' = Vw''$ since V is orthogonal). Now the equation is:

$$\Sigma w'' = 0$$

Writing out the structure of Σ :

$$[\tilde{\Sigma} \quad 0_{n \times (d-n)}]w'' = 0$$

We can define $\tilde{X} = \tilde{\Sigma}$ (a diagonal, invertible $n \times n$ matrix). So the equation becomes:

$$[\tilde{X} \quad 0]w'' = 0$$

3. Initial Condition w''_0 : Recall $w'_0 = -w^* = -X^T(XX^T)^{-1}y$. Using SVD: $X^T = V\Sigma^T U^T$ and $XX^T = U\Sigma V^T V\Sigma^T U^T = U(\Sigma\Sigma^T)U^T$.

$$w'_0 = -V\Sigma^T U^T (U(\Sigma\Sigma^T)U^T)^{-1}y$$

$$w'_0 = -V\Sigma^T U^T (U(\Sigma\Sigma^T)^{-1}U^T)y$$

$$w'_0 = -V\Sigma^T (\Sigma\Sigma^T)^{-1}U^T y$$

Now, compute $w'_0 = V^T w'_0$:

$$w''_0 = V^T (-V\Sigma^T (\Sigma\Sigma^T)^{-1}U^T y)$$

$$w''_0 = -(V^T V)\Sigma^T (\Sigma\Sigma^T)^{-1}U^T y$$

$$w''_0 = -\Sigma^T (\Sigma\Sigma^T)^{-1}(U^T y)$$

Notice the structure of $\Sigma^T = \begin{bmatrix} \bar{\Sigma} \\ 0 \end{bmatrix}$. The term $\Sigma \Sigma^T = \bar{\Sigma}^2$ is an $n \times n$ diagonal matrix. The product $\Sigma^T (\Sigma \Sigma^T)^{-1}$ becomes:

$$\begin{bmatrix} \bar{\Sigma} \\ 0 \end{bmatrix} \bar{\Sigma}^{-2} = \begin{bmatrix} \bar{\Sigma}^{-1} \\ 0 \end{bmatrix}$$

Because of that block of zeros at the bottom of Σ^T , the resulting vector w''_0 will have non-zero values only in the first n entries. The last $d - n$ entries are strictly **zero**.

Part (c): Reducing to a Square System

Goal: Argue why we can focus on the square system $\tilde{X}\tilde{w} = 0$.

Reasoning: From part (b), our system is $[\tilde{X} \quad 0]w'' = 0$. Let's split w'' into two parts: the top n components (\tilde{w}) and the bottom $d - n$ components (w_{null}).

$$[\tilde{X} \quad 0] \begin{bmatrix} \tilde{w} \\ w_{\text{null}} \end{bmatrix} = \tilde{X}\tilde{w} + 0 \cdot w_{\text{null}} = 0$$

This implies $\tilde{X}\tilde{w} = 0$. Crucially, the update dynamics of SGD only depend on the gradients. Since the last $d - n$ columns of the rotated matrix are **zero**, the gradient with respect to w_{null} is always zero. Furthermore, the initial condition for the bottom part is zero (from part b).

- Start at 0.
- Gradient is always 0.
- Therefore, w_{null} stays 0 forever.



Thus, we can completely ignore the bottom part and focus solely on the top n dimensions, governed by the square, invertible matrix \tilde{X} (which is effectively $\bar{\Sigma}$). Equation (8) $\tilde{X}\tilde{w} = 0$ captures all the non-trivial dynamics.

Part (d): SGD Update Equivalence

Goal: Show the SGD update matches $\tilde{w}_{t+1} = \tilde{w}_t - 2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^T \tilde{w}_t$.

Reasoning: The original SGD update (eq 9) is:

$$w_{t+1} = w_t - \eta \nabla L_{I_t}(w_t)$$

The loss for index i is $L_i(w) = (y_i - x_i^T w)^2$. Gradient: $\nabla L_i(w) = -2(y_i - x_i^T w)x_i$. So,

$$w_{t+1} = w_t + 2\eta(y_{I_t} - x_{I_t}^T w_t)x_{I_t}$$

Now, transform to \tilde{w} coordinates. Recall $y_i - x_i^T w_t$ is the residual. In the new coordinates, the full residual is $0 - \tilde{x}_{I_t}^T \tilde{w}_t$ (since target is 0 and w_{null} is 0). The vector x_{I_t} in the original space corresponds to the row I_t of X . In the rotated space (multiplied by V^T), the row x_{I_t} becomes the row \tilde{x}_{I_t} (padded with zeros). The update in the rotated space (\tilde{w}) is just the projection of the original update onto the first n components. Substituting $y_{I_t} = x_{I_t}^T w^*$ (since solution exists) and using $w' = w - w^*$:

$$w'_{t+1} = w'_t - 2\eta(x_{I_t}^T w'_t)x_{I_t}$$

Rotate by V^T :

$$w''_{t+1} = w''_t - 2\eta(x_{I_t}^T VV^T w'_t)V^T x_{I_t}$$

(Using identity $I = VV^T$ inside the scalar product). Note that $V^T x_{I_t}$ is the I_t -th row of the transformed matrix $[\tilde{X} \quad 0]$. Let's call the top part \tilde{x}_{I_t} . So the update for the top n components \tilde{w} is:

$$\tilde{w}_{t+1} = \tilde{w}_t - 2\eta(\tilde{x}_{I_t}^T \tilde{w}_t)\tilde{x}_{I_t}$$

Rearranging the scalar product:

$$\tilde{w}_{t+1} = \tilde{w}_t - 2\eta \tilde{x}_{I_t} (\tilde{x}_{I_t}^T \tilde{w}_t)$$

$$\tilde{w}_{t+1} = \tilde{w}_t - 2\eta (\tilde{x}_{I_t} \tilde{x}_{I_t}^T) \tilde{w}_t$$

$$\tilde{w}_{t+1} = (I - 2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^T) \tilde{w}_t$$

This matches equation (10).

Part (e): Convergence Probability Bound

Goal: Show that exponential convergence in expectation ($E[\mathcal{L}_{t+1}] \leq (1 - \rho)\mathcal{L}_t$) implies high probability convergence $P(\mathcal{L}(\tilde{w}_T) < \epsilon) \geq 1 - \delta$.

Reasoning: We have $E[\mathcal{L}(\tilde{w}_T)] \leq (1 - \rho)^T \mathcal{L}(\tilde{w}_0)$. Let $C = \mathcal{L}(\tilde{w}_0)$.

$$E[\mathcal{L}(\tilde{w}_T)] \leq C(1 - \rho)^T$$

We want to bound the probability that $\mathcal{L}(\tilde{w}_T) \geq \epsilon$. Using **Markov's Inequality** ($P(X \geq a) \leq \frac{E[X]}{a}$ for non-negative X):



$$P(\mathcal{L}(\tilde{w}_T) \geq \epsilon) \leq \frac{E[\mathcal{L}(\tilde{w}_T)]}{\epsilon} \leq \frac{C(1 - \rho)^T}{\epsilon}$$

We want this probability to be less than δ :

$$\frac{C(1 - \rho)^T}{\epsilon} \leq \delta$$

$$(1 - \rho)^T \leq \frac{\epsilon \delta}{C}$$

Taking logs:

$$T \ln(1 - \rho) \leq \ln\left(\frac{\epsilon \delta}{C}\right)$$

Since $\ln(1 - \rho)$ is negative, dividing by it flips the inequality:

$$T \geq \frac{\ln\left(\frac{\epsilon \delta}{C}\right)}{\ln(1 - \rho)}$$

Since a valid T exists for any ϵ, δ , the implication holds.

Part (f): Candidate Lyapunov Function

Goal: Argue why $\mathcal{L}(\tilde{w}) = \tilde{w}^T \tilde{X}^T \tilde{X} \tilde{w}$ is a valid Lyapunov function.

Reasoning:



1. **Non-negativity:** $\mathcal{L}(\tilde{w}) = \|\tilde{X} \tilde{w}\|_2^2$. The squared Euclidean norm is always non-negative.
2. **Zero only at solution:** We want it to be zero *only* at $\tilde{w} = 0$. \tilde{X} is an $n \times n$ matrix. From part (b), $\tilde{X} = \tilde{\Sigma}$ contains the singular values of the original full-rank matrix X . Since X has full row rank (n), all n singular values are strictly positive. Therefore, \tilde{X} is invertible (full rank). $\|\tilde{X} \tilde{w}\| = 0 \iff \tilde{X} \tilde{w} = 0 \iff \tilde{w} = 0$. So this is a valid Lyapunov function measuring the distance to the solution in the "prediction space".

Part (g) & (h): Decomposing the Update and Bounding Term A

Goal: Expand $\mathcal{L}(\tilde{w}_{t+1})$ and bound the linear term A .

$$\text{Expansion: } \tilde{w}_{t+1} = \tilde{w}_t + \Delta_t, \text{ where } \Delta_t = -2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^T \tilde{w}_t. \mathcal{L}(\tilde{w}_{t+1}) = \|\tilde{X}(\tilde{w}_t + \Delta_t)\|^2 =$$

$$\underbrace{(\tilde{w}_t + \Delta_t)^T \tilde{X}^T \tilde{X}(\tilde{w}_t + \Delta_t)}_A = \tilde{w}_t^T \tilde{X}^T \tilde{X} \tilde{w}_t + 2\tilde{w}_t^T \tilde{X}^T \tilde{X} \Delta_t + \Delta_t^T \tilde{X}^T \tilde{X} \Delta_t = \mathcal{L}(\tilde{w}_t) +$$

$$\underbrace{2\tilde{w}_t^T \tilde{X}^T \tilde{X} \Delta_t}_B + \underbrace{\Delta_t^T \tilde{X}^T \tilde{X} \Delta_t}_C$$

Bounding Expected A (Part h):

$$E[A|\tilde{w}_t] = E[2\tilde{w}_t^T \tilde{X}^T \tilde{X}(-2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^T \tilde{w}_t)]$$

The expectation is over the choice of index I_t (uniform $1 \dots n$). $E[\tilde{x}_{I_t} \tilde{x}_{I_t}^T] = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^T$.
Recall that the rows \tilde{x}_i form the matrix \tilde{X} . So $\sum \tilde{x}_i \tilde{x}_i^T = \tilde{X}^T \tilde{X}$. Therefore $E[\tilde{x}_{I_t} \tilde{x}_{I_t}^T] = \frac{1}{n} \tilde{X}^T \tilde{X}$

Substitute this back:

$$E[A] = -4\eta \tilde{w}_t^T \tilde{X}^T \tilde{X} \left(\frac{1}{n} \tilde{X}^T \tilde{X} \right) \tilde{w}_t$$

$$E[A] = -\frac{4\eta}{n} \tilde{w}_t^T (\tilde{X}^T \tilde{X})^2 \tilde{w}_t$$

We need to bound this by $-c_1 \eta \mathcal{L}(\tilde{w}_t) = -c_1 \eta \tilde{w}_t^T (\tilde{X}^T \tilde{X}) \tilde{w}_t$. Let $M = \tilde{X}^T \tilde{X}$. We are comparing $-w^T M^2 w$ vs $-w^T M w$. Since \tilde{X} is diagonal (from SVD), M is diagonal with entries σ_i^2 . M^2 has entries σ_i^4 . We know $\sigma_i^2 \geq \sigma_{\min}^2 > 0$ (smallest singular value). So $M^2 \geq \sigma_{\min}^2 M$ (in the PSD sense).

$$-w^T M^2 w \leq -\sigma_{\min}^2 w^T M w$$

$$E[A] \leq -\frac{4\eta}{n} \sigma_{\min}^2 \mathcal{L}(\tilde{w}_t)$$

So $c_1 = \frac{4\sigma_{\min}^2}{n}$.

Part (i): Bounding Term B

Goal: Show $E[B|\tilde{w}_t] \leq c_2 \eta^2 \mathcal{L}(\tilde{w}_t)$.

Expansion of B: $B = \Delta_t^T \tilde{X}^T \tilde{X} \Delta_t = -2\eta \tilde{x}_{I_t} (\tilde{x}_{I_t}^T \tilde{w}_t)$. Let scalar $r_t = \tilde{x}_{I_t}^T \tilde{w}_t$. $B = 4\eta^2 r_t^2 \tilde{x}_{I_t}^T (\tilde{X}^T \tilde{X}) \tilde{x}_{I_t}$. Term $r_t^2 = (\tilde{x}_{I_t}^T \tilde{w}_t)^2$. We need to bound $E[r_t^2 \|\tilde{X} \tilde{x}_{I_t}\|^2]$.

Let $\beta = \max_i \|\tilde{x}_i\|^2$ (max squared norm of rows). Also $\|\tilde{X} \tilde{x}_{I_t}\|^2 \leq \sigma_{\max}^2 \|\tilde{x}_{I_t}\|^2 \leq \sigma_{\max}^2 \beta$. So $B \leq 4\eta^2 (\tilde{x}_{I_t}^T \tilde{w}_t)^2 (\sigma_{\max}^2 \beta)$. Taking expectation: $E[(\tilde{x}_{I_t}^T \tilde{w}_t)^2] = \tilde{w}_t^T E[\tilde{x}_{I_t} \tilde{x}_{I_t}^T] \tilde{w}_t = \tilde{w}_t^T (\frac{1}{n} \tilde{X}^T \tilde{X}) \tilde{w}_t = \frac{1}{n} \mathcal{L}(\tilde{w}_t)$. Combine:

$$E[B] \leq 4\eta^2 (\sigma_{\max}^2 \beta) \frac{1}{n} \mathcal{L}(\tilde{w}_t)$$

$$E[B] \leq \frac{4\beta \sigma_{\max}^2}{n} \eta^2 \mathcal{L}(\tilde{w}_t)$$

So $c_2 = \frac{4\beta \sigma_{\max}^2}{n}$.

Part (j): Combining for Exponential Convergence

Goal: Show $E[\mathcal{L}_{t+1}] \leq (1 - c_1\eta + c_2\eta^2)\mathcal{L}_t$.

Reasoning: From (g), (h), (i): $E[\mathcal{L}_{t+1}] = \mathcal{L}_t + E[A] + E[B]$ $E[\mathcal{L}_{t+1}] \leq \mathcal{L}_t - c_1\eta\mathcal{L}_t + c_2\eta^2\mathcal{L}_t$
 $E[\mathcal{L}_{t+1}] \leq (1 - c_1\eta + c_2\eta^2)\mathcal{L}_t$

For convergence, we need the factor $(1 - c_1\eta + c_2\eta^2) < 1$.

$$-c_1\eta + c_2\eta^2 < 0$$

$$c_2\eta^2 < c_1\eta$$

$$\eta < \frac{c_1}{c_2}$$

Substituting our constants:

$$\eta < \frac{4\sigma_{\min}^2/n}{4\beta\sigma_{\max}^2/n} = \frac{\sigma_{\min}^2}{\beta\sigma_{\max}^2}$$

As long as η is small enough (positive and below this bound), the term is < 1 , proving exponential convergence in expectation.

Part (k): Ridge Regression Interpretation

Goal: Explain why feature augmentation allows SGD to converge exponentially for Ridge Regression, unlike standard optimization.

Reasoning: Ridge regression minimizes $\|Xw - y\|^2 + \lambda\|w\|^2$.

1. **Standard Optimization View:** The objective is strongly convex, but the gradients

$\nabla(\lambda\|w\|^2) = 2\lambda w$ are deterministic. However, the stochastic gradients from the data part might not vanish at the optimum if we use standard mini-batching, preventing standard SGD with constant step size from converging to the exact solution (it bounces around in a "noise ball").

2. **Feature Augmentation View:** Ridge regression is mathematically equivalent to Ordinary Least Squares on an augmented dataset:

$$X' = \begin{bmatrix} X \\ \sqrt{\lambda}I \end{bmatrix}, \quad y' = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

We are now solving $X'w \approx y'$. Crucially, because of the $\sqrt{\lambda}I$ block, the matrix X' has full column rank (trivial null space). More importantly, as discussed in this problem, this fits the **interpolating regime** framework if we consider the "noise" in SGD as sampling rows from this augmented matrix. By treating the regularization term as just another data point (row), the "stochastic" gradients for the regularizer are compatible with the data gradients in a way that allows the system to be an interpolating system (or close to it in the augmented space logic). Specifically, the problem establishes that for *interpolating* least squares (where loss can go to 0), SGD with constant step size converges exponentially. By augmenting features, we effectively transform the ridge problem into a form where we can leverage this interpolating property (or structural equivalent), allowing **constant step size SGD** to converge exponentially fast, rather than the slow $1/t$ decay required when gradients don't vanish at the optimum.

Summary of Observation: Implementation matters. Augmenting the data matrix makes the problem look like a zero-residual least squares problem (in the augmented space), enabling the fast exponential convergence proved in parts (a)-(j).

For part (c), you did not answer why each of the n constituent equations (corresponding to rows of $\tilde{X}\tilde{w} = 0$) can be obtained by means of coordinate changes from the same indexed equation in $Xw = y$. From now and onwards, do not skip any parts...



◆ Show thinking ▾

You are absolutely right. I apologize for the oversight in part (c). I solved the system globally using the full SVD (multiplying by U^T) in the previous step, which mixed the rows and obscured the direct correspondence between the individual equations.

Here is the corrected and specific derivation for **Part (c)**, ensuring we establish the row-wise coordinate change.

Part (c): Row-wise Coordinate Change

Goal: Show that each of the n constituent equations (rows) of the square system $\tilde{X}\tilde{w} = 0$ can be obtained by means of coordinate changes from the **same indexed equation** (row i) in the original system $Xw' = 0$.

Derivation:

1. **The Original i -th Equation:** Recall our system is $Xw' = 0$. This system consists of n equations. The i -th equation corresponds to the i -th row of X (let's denote the row vector as x_i^T).

$$\text{Equation } i : \quad x_i^T w' = 0$$

2. **The Coordinate Transformation:** In part (b), we introduced the change of variables $w' = Vw''$, where V is the $d \times d$ orthogonal matrix from the SVD of X ($X = U\Sigma V^T$).

Substitute this into the i -th equation:

$$x_i^T (Vw'') = 0 \implies (x_i^T V)w'' = 0$$

3. **Analyzing the Transformed Row Vector ($x_i^T V$):** The vector $x_i^T V$ represents the i -th row of the matrix product XV . Let's break down V into two blocks:

- V_{range} : The first n columns (corresponding to the row space).
- V_{null} : The last $d - n$ columns (corresponding to the null space).

$$V = [V_{\text{range}} \quad V_{\text{null}}]$$

Now, look at the product $x_i^T V$:

$$x_i^T V = [x_i^T V_{\text{range}} \quad x_i^T V_{\text{null}}]$$

- **Term 1:** Define $\tilde{x}_i^T = x_i^T V_{\text{range}}$. This is a vector of size $1 \times n$.
- **Term 2:** Since x_i is a row of X , it lies entirely in the row space of X . It is orthogonal to the null space basis V_{null} . Therefore, $x_i^T V_{\text{null}} = 0$.

So, the transformed row vector is:

$$x_i^T V = [\tilde{x}_i^T \quad 0_{1 \times (d-n)}]$$

4. **The Resulting Equation:** Substitute this back into the transformed equation:

$$[\tilde{x}_i^T \quad 0] \begin{bmatrix} \tilde{w} \\ w_{\text{null}} \end{bmatrix} = 0$$

$$\tilde{x}_i^T \tilde{w} + 0 = 0$$

$$\tilde{x}_i^T \tilde{w} = 0$$

Conclusion: We have shown that simply applying the coordinate change $w' = Vw''$ to the i -th row of the original system directly yields the i -th row of the new square system $\tilde{X}\tilde{w} = 0$.

- The matrix \tilde{X} in equation (8) is composed of these rows: $\tilde{X} = X V_{\text{range}}$.
- The index i is preserved perfectly; we did not mix the equations (which would have happened if we multiplied the whole system by U^T).

(Note: This definition of $\tilde{X} = X V_{\text{range}}$ still satisfies $\tilde{X}^T \tilde{X} = V_{\text{range}}^T X^T X V_{\text{range}} = \Sigma^2$, so all the convergence analysis using eigenvalues in the later parts remains exactly the same.)



Please proceed with the next question (image) when you are ready.



This is the next problem.



Show thinking ▾

This problem explores how "Momentum" changes the convergence behavior of gradient descent. It involves analyzing a 2×2 recurrence relation that governs the evolution of the parameters w_t and the momentum term z_t .

I will break this down step-by-step.

Part (a): Rewrite the System with Reparameterized Variables

Goal: Rewrite the momentum update equations in terms of the transformed variables $x_t = V^T(w_t - w^*)$ and $a_t = V^T z_t$.

Reasoning:

1. **Original System:**

- $w_{t+1} = w_t - \eta z_{t+1}$
- $z_{t+1} = (1 - \beta)z_t + \beta(2X^T X w_t - 2X^T y)$ (using the gradient $g_t = \nabla \mathcal{L}(w_t)$)

2. **Gradient Simplification:** Recall that $w^* = (X^T X)^{-1} X^T y$, so $X^T X w^* = X^T y$. We can rewrite the gradient term:

$$g_t = 2X^T X w_t - 2X^T y = 2X^T X(w_t - w^*)$$

3. **Variable Transformation:** We multiply everything by V^T to move into the singular vector basis (where $X^T X$ becomes diagonal Σ^2).

- Let $e_t = w_t - w^*$. The gradient is $2X^T X e_t$.
- Multiply w -update by V^T :

$$V^T(w_{t+1} - w^*) = V^T(w_t - w^*) - \eta V^T z_{t+1}$$

$$x_{t+1} = x_t - \eta a_{t+1}$$

- Multiply z -update by V^T :

$$V^T z_{t+1} = (1 - \beta)V^T z_t + \beta V^T(2X^T X e_t)$$

$$a_{t+1} = (1 - \beta)a_t + 2\beta(V^T X^T X V V^T e_t)$$

(Insert identity $I = VV^T$). Note that $V^T X^T X V = \Sigma^2$.

$$a_{t+1} = (1 - \beta)a_t + 2\beta\Sigma^2 x_t$$

Component-wise Form: Since Σ^2 is diagonal with entries σ_i^2 , the system decouples for each component i . For the i -th component (dropping index $[i]$ for brevity):

1. $a_{t+1} = (1 - \beta)a_t + 2\beta\sigma_i^2 x_t$
2. $x_{t+1} = x_t - \eta a_{t+1}$ Substitute the expression for a_{t+1} into the x -update:

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

$$x_{t+1} = (1 - 2\eta\beta\sigma_i^2)x_t - \eta(1 - \beta)a_t$$

Part (b): Derive the State Matrix R_i

Goal: Find the matrix R_i such that $\begin{bmatrix} a_{t+1} \\ x_{t+1} \end{bmatrix} = R_i \begin{bmatrix} a_t \\ x_t \end{bmatrix}$.

Derivation: From part (a), we have the equations:

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

Arranging this into matrix-vector multiplication form:

$$\begin{bmatrix} a_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} 1 - \beta & 2\beta\sigma_i^2 \\ -\eta(1 - \beta) & 1 - 2\eta\beta\sigma_i^2 \end{bmatrix} \begin{bmatrix} a_t \\ x_t \end{bmatrix}$$

Solution (b):

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

Part (c): Analyze Eigenvalues of R_i

Goal: Find eigenvalues symbolically and determine conditions for them being Real, Repeated, or Complex.

Characteristic Equation: We need $\det(R_i - \lambda I) = 0$. Let λ_1, λ_2 be the eigenvalues. $\text{Trace}(R_i) = (1 - \beta) + (1 - 2\eta\beta\sigma_i^2) = 2 - \beta - 2\eta\beta\sigma_i^2$. $\text{Determinant}(R_i) = (1 - \beta)(1 - 2\eta\beta\sigma_i^2) - (2\beta\sigma_i^2)(-\eta(1 - \beta)) = (1 - \beta)[1 - 2\eta\beta\sigma_i^2 + 2\eta\beta\sigma_i^2] = 1 - \beta$.

The characteristic polynomial is $\lambda^2 - \text{Tr}(R_i)\lambda + \det(R_i) = 0$.

$$\lambda^2 - (2 - \beta - 2\eta\beta\sigma_i^2)\lambda + (1 - \beta) = 0$$

Solve using quadratic formula:

$$\lambda = \frac{\text{Tr} \pm \sqrt{\text{Tr}^2 - 4 \det}}{2}$$

Discriminant $\Delta = \text{Tr}^2 - 4(1 - \beta)$. Let $k = \eta\beta\sigma_i^2$. Then $\text{Tr} = 2 - \beta - 2k$.

$$\Delta = (2 - \beta - 2k)^2 - 4(1 - \beta)$$

Conditions:

1. **Purely Real:** Discriminant $\Delta > 0$.
2. **Repeated Real:** Discriminant $\Delta = 0$.
3. **Complex:** Discriminant $\Delta < 0$.

Let's simplify the repeated condition ($\Delta = 0$) to find the boundary:

$$(2 - \beta - 2k)^2 = 4(1 - \beta)$$

Take square root: $2 - \beta - 2k = \pm 2\sqrt{1 - \beta}$.

$$2k = (2 - \beta) \mp 2\sqrt{1 - \beta}$$

Since $k = \eta\beta\sigma_i^2$, we can solve for η . Note that $(1 - \sqrt{1 - \beta})^2 = 1 - 2\sqrt{1 - \beta} + (1 - \beta) = 2 - \beta - 2\sqrt{1 - \beta}$. So, $2k = (1 \pm \sqrt{1 - \beta})^2$. The relevant root for stability transitions is usually the smaller one related to learning rates.

Summary for (c):

- **Repeated:** When $\eta\beta\sigma_i^2 = \frac{1}{2}(1 - \sqrt{1 - \beta})^2$ (approx).
- **Complex:** When η is larger (inside the parabola of the discriminant), leading to oscillation.
- **Real:** When η is very small (overdamped) or very large (unstable).

Part (d): Repeated Eigenvalues Case

Goal: Stability condition and highest learning rate for this case.

Stability: For repeated eigenvalues, we are at the critical damping point. The eigenvalue is $\lambda = \frac{\text{Tr}}{2}$. From the condition $\Delta = 0$, we know $\text{Tr}^2 = 4(1 - \beta)$, so $\text{Tr} = 2\sqrt{1 - \beta}$. Thus $\lambda = \sqrt{1 - \beta}$. For stability (inside unit circle), we need $|\lambda| < 1$. $\sqrt{1 - \beta} < 1 \implies 1 - \beta < 1 \implies \beta > 0$. Also we need $\beta < 1$ for the term under square root to be real. So stability holds if $0 < \beta < 1$.

Highest Learning Rate η : Using the condition derived in (c) for $\Delta = 0$:

$$2\eta\beta\sigma_i^2 = (1 - \sqrt{1 - \beta})^2$$

(We pick the minus sign because we are looking for the transition from small η to complex. The plus sign would correspond to a much larger η). Actually, let's check. Standard momentum analysis suggests optimal convergence happens exactly at this transition.

$$\eta = \frac{(1 - \sqrt{1 - \beta})^2}{2\beta\sigma_i^2}$$

Part (e): Real Eigenvalues Case

Goal: Stability condition and range of η .

Regime: This is the "overdamped" regime where η is small. Condition: $\Delta > 0$ and η is small.

Condition for stability: Both eigenvalues must be < 1 . Since $\lambda_1 \lambda_2 = \det(R_i) = 1 - \beta$, if $0 < \beta < 1$, the product is < 1 . We just need the larger eigenvalue to be < 1 . This typically holds for $0 < \eta < \frac{(1-\sqrt{1-\beta})^2}{2\beta\sigma_i^2}$.

Part (f): Complex Eigenvalues Case

Goal: Stability condition and highest η .

Regime: This is the "underdamped" regime where momentum causes oscillation. Eigenvalues are complex conjugates: $\lambda, \bar{\lambda}$. The magnitude squared is $|\lambda|^2 = \lambda\bar{\lambda} = \det(R_i) = 1 - \beta$.

Observation: The magnitude of the eigenvalues depends *only* on β , not on η (as long as they are complex). Stability condition: $|1 - \beta| < 1$. Given $\beta > 0$, this is always true if $\beta < 1$.

Upper Bound on η : However, we must remain in the complex regime. Also, we must ensure the real part doesn't blow up? No, actually, if the determinant is fixed at $1 - \beta$, the eigenvalues stay on the circle of radius $\sqrt{1 - \beta}$ regardless of η , until they collide on the real axis again at the other end of the discriminant parabola (instability side). The condition for complex eigenvalues is $\Delta < 0$:

$$(2 - \beta - 2\eta\beta\sigma_i^2)^2 < 4(1 - \beta)$$

$$-2\sqrt{1 - \beta} < 2 - \beta - 2\eta\beta\sigma_i^2 < 2\sqrt{1 - \beta}$$

Right inequality: $2 - \beta - 2\sqrt{1 - \beta} < 2\eta\beta\sigma_i^2$ ($1 - \sqrt{1 - \beta})^2 < 2\eta\beta\sigma_i^2$) Left inequality (Upper bound on η): $2\eta\beta\sigma_i^2 < 2 - \beta + 2\sqrt{1 - \beta} = (1 + \sqrt{1 - \beta})^2$

Solution (f): The learning rate range for stable, complex eigenvalues is:

$$\frac{(1 - \sqrt{1 - \beta})^2}{2\beta\sigma_i^2} < \eta < \frac{(1 + \sqrt{1 - \beta})^2}{2\beta\sigma_i^2}$$

The highest η is the upper bound.

Part (g): Numerical Comparison (The "Meat" of the problem)

Parameters: $\beta = 0.1$, $\sigma_{\max}^2 = 5$, $\sigma_{\min}^2 = 0.05$.

1. Momentum Strategy: We want the "fastest convergence". The convergence rate is determined by the spectral radius $\rho(R_i) = \max |\lambda_i|$.

- **Complex regime:** Magnitude is $\sqrt{1 - \beta} = \sqrt{0.9} \approx 0.948$. This is constant for all eigenmodes that are complex.
- **Real regime:** One eigenvalue is larger than $\sqrt{1 - \beta}$, meaning slower convergence.
- **Optimal Strategy:** We want all modes (from σ_{\min} to σ_{\max}) to be complex or at the boundary (repeated). The bottleneck is usually σ_{\min} trying to be real (too slow) or σ_{\max} becoming unstable. Ideally, we set η such that the "slowest" mode (σ_{\min}) is pushed as close to the complex transition as possible (critical damping), or ideally into the complex regime where all magnitudes are $\sqrt{1 - \beta}$.

However, with a fixed $\beta = 0.1$, the magnitude of convergence is fixed at $\sqrt{0.9}$ for complex modes. Is this better than GD? GD rate $\approx \frac{\kappa-1}{\kappa+1} \cdot \kappa = 5/0.05 = 100$. GD rate $\approx 99/101 \approx 0.98$. Momentum rate ≈ 0.948 . Momentum is much faster per step.

Choosing η : To get complex eigenvalues for all $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, we need η to satisfy the lower bound condition for the smallest sigma: $2\eta\beta\sigma_{\min}^2 > (1 - \sqrt{1 - \beta})^2 \eta >$

$$\frac{(1-\sqrt{0.9})^2}{2(0.1)(0.05)} \approx \frac{(1-0.9486)^2}{0.01} = \frac{0.0026}{0.01} = 0.26$$

Check upper bound for σ_{\max} : $\eta < \frac{(1+\sqrt{0.9})^2}{2(0.1)(5)} \approx \frac{(1.948)^2}{1} \approx 3.8$. So $\eta \approx 0.26$ (or slightly higher) works. Let's pick η to put σ_{\min} at the critical point. $\eta \approx 0.26$.

2. Iteration Calculation: Target: Error < 0.001 (99.9%). Condition: Rate^t = 0.001. $t =$

$$\frac{\ln(0.001)}{\ln(\text{Rate})} = \frac{-6.9}{\ln(\text{Rate})}.$$

- **Momentum:** Rate $\approx \sqrt{0.9} \approx 0.9486$. $\ln(0.9486) \approx -0.0527$. $t \approx 6.9/0.0527 \approx \underline{131}$ iterations.
- **Ordinary GD:** Optimal rate $\frac{\kappa-1}{\kappa+1} = \frac{99}{101} \approx 0.9802$. $\ln(0.9802) \approx -0.02$. $t \approx 6.9/0.02 \approx \underline{345}$ iterations.

Comparison: Momentum is significantly faster ($\sim 2.6x$ fewer steps). 

Part (h) & (i): Notebook Conceptual Questions

Q (h): How does σ_i influence updates?

- High σ_i (high curvature) directions oscillate or converge fast.
- Low σ_i (flat) directions converge slowly.
- Momentum helps accelerate the low σ_i directions by accumulating gradients (z_t grows), effectively increasing the step size for those dimensions.

Q (i): Which converges faster?

- **Momentum** converges faster. 
- **Why?** Ordinary GD is limited by the condition number κ . The rate is $1 - 2/\kappa$. Momentum (with optimal tuning) can achieve a rate closer to $1 - 2/\sqrt{\kappa}$. Because $\sqrt{100} = 10 \ll 100$, momentum handles ill-conditioned problems (like this one with $\kappa = 100$) much better.

Please move onto the next problem: 

◆ Show thinking ▾

This is the next problem, titled "4. Optimizers". It focuses on completing the pseudocode for the Adam Optimizer and establishing the mathematical equivalence between L2 regularization and Weight Decay.

Here are the step-by-step solutions for each part.

Part (a): Complete the Pseudocode for Adam

Goal: Fill in the missing lines (A) and (B) in Algorithm 2 (Adam Optimizer).

Reasoning: The Adam optimizer maintains exponential moving averages of the gradients (m_t) and the squared gradients (v_t).

- m_t (**First Moment**): This is the mean of the gradients. The update rule is a linear interpolation between the previous moment m_{t-1} and the current gradient g_t . The decay rate is β_1 .

$$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

- v_t (**Second Moment**): This is the uncentered variance (mean of squared gradients). The update rule is similar, interpolating between the previous second moment v_{t-1} and the element-wise square of the current gradient g_t^2 . The decay rate is β_2 .

$$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

(Note: The operation g_t^2 is element-wise).

Solution (a):

- (A): $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$
- (B): $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)(g_t \odot g_t)$ (where \odot denotes element-wise multiplication, or simply g_t^2 if vector notation implies element-wise powers).

Part (b): Equivalence of L2 Regularization and Weight Decay for SGD

Goal: Show that SGD with weight decay on the original loss $f_t(\theta)$ is equivalent to regular SGD on the L2-regularized loss $f_t^{reg}(\theta)$, and find the relationship between γ (weight decay factor), λ (regularization strength), and η (learning rate).

Derivation:

1. **Method 1: SGD on L2-Regularized Loss (f_t^{reg})** The regularized objective function is:

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

The gradient of this regularized loss with respect to θ is:

$$\nabla f_t^{reg}(\theta) = \nabla f_t(\theta) + \nabla \left(\frac{\lambda}{2} \|\theta\|_2^2 \right)$$

$$\nabla f_t^{reg}(\theta) = \nabla f_t(\theta) + \lambda \theta$$

The standard SGD update rule is $\theta_{t+1} = \theta_t - \eta \nabla f_t^{reg}(\theta_t)$. Substitute the gradient derived above:

$$\theta_{t+1} = \theta_t - \eta (\nabla f_t(\theta_t) + \lambda \theta_t)$$

$$\theta_{t+1} = \theta_t - \eta \nabla f_t(\theta_t) - \eta \lambda \theta_t$$

Grouping the terms involving θ_t :

$$\theta_{t+1} = (1 - \eta \lambda) \theta_t - \eta \nabla f_t(\theta_t)$$

(Equation 1)

2. **Method 2: SGD with Explicit Weight Decay** The problem statement gives the explicit weight decay update rule:

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

(Equation 2)

3. **Matching the Terms:** Compare **Equation 1** and **Equation 2**:

- Equation 1: $\theta_{t+1} = (1 - \eta\lambda)\theta_t - \eta \nabla f_t(\theta_t)$
- Equation 2: $\theta_{t+1} = (1 - \gamma)\theta_t - \eta \nabla f_t(\theta_t)$

For these two updates to be mathematically identical, the coefficients of θ_t must match:

$$1 - \eta\lambda = 1 - \gamma$$

$$\eta\lambda = \gamma$$

Solution (b): The two methods are equivalent when the weight decay parameter γ is chosen as:

$$\gamma = \eta\lambda$$

Interpretation: This confirms that for standard SGD, "L2 Regularization" (adding a penalty to the loss) and "Weight Decay" (shrinking the weights directly during the update) are mathematically identical. (Note: This equivalence holds for SGD, but **not** for adaptive methods like Adam, where L2 regularization and Weight Decay behave differently!)

Would you like to move on to the next problem or discuss the distinction for Adam mentioned in the note?

Please move onto the next problem

◆ Show thinking ▾



This is the final problem, titled "**5. Regularization and Instance Noise**". It connects the concept of training with noisy data (data augmentation) to L2 regularization.

Here are the step-by-step solutions for each part.

Part (a): Equivalence to Regularized Least Squares

Goal: Show that minimizing the expected loss with noisy data $\check{X}_i = x_i + N_i$ is equivalent to minimizing a regularized least squares objective, and find λ .

Derivation:

1. **The Objective:** We want to minimize $J(w) = \mathbb{E}[\|\check{X}w - y\|^2]$. Here \check{X} is the matrix where row i is $\check{x}_i^T = (x_i + N_i)^T$. So, $\|\check{X}w - y\|^2 = \sum_{i=1}^m (\check{x}_i^T w - y_i)^2$.
2. **Expanding the Expectation:** Let's look at a single term $\mathbb{E}[(x_i + N_i)^T w - y_i]^2$. Let residual $r_i = x_i^T w - y_i$. Then the term inside the square is $(r_i + N_i^T w)^2$.

$$\mathbb{E}[(r_i + N_i^T w)^2] = \mathbb{E}[r_i^2 + 2r_i N_i^T w + (N_i^T w)^2]$$

3. Analyzing Terms:

- r_i^2 is deterministic (no noise), so $\mathbb{E}[r_i^2] = (x_i^T w - y_i)^2$.
- $\mathbb{E}[2r_i N_i^T w] = 2r_i \mathbb{E}[N_i]^T w$. Since $N_i \sim \mathcal{N}(0, \sigma^2 I)$, $\mathbb{E}[N_i] = 0$. So this term vanishes.
- $\mathbb{E}[(N_i^T w)^2] = \mathbb{E}[w^T N_i N_i^T w] = w^T \mathbb{E}[N_i N_i^T] w$. Since the covariance of N_i is $\sigma^2 I$, we have $\mathbb{E}[N_i N_i^T] = \sigma^2 I$. Therefore, $\mathbb{E}[(N_i^T w)^2] = w^T (\sigma^2 I) w = \sigma^2 \|w\|^2$.

4. Summing Over Data Points:

$$J(w) = \sum_{i=1}^m [(x_i^T w - y_i)^2 + \sigma^2 \|w\|^2]$$

$$J(w) = \|\mathbf{X}w - \mathbf{y}\|^2 + m\sigma^2 \|w\|^2$$

5. **Matching the Form:** The problem asks to match the form: $\arg \min \frac{1}{m} \|\mathbf{X}w - \mathbf{y}\|^2 + \lambda \|w\|^2$. Dividing our derived $J(w)$ by m :

$$\frac{1}{m} J(w) = \frac{1}{m} \|\mathbf{X}w - \mathbf{y}\|^2 + \sigma^2 \|w\|^2$$

Solution (a): Comparing the coefficients of $\|w\|^2$, we find:

$$\lambda = \sigma^2$$

Minimizing the squared error on noise-corrupted data is mathematically equivalent to Ridge Regression with regularization strength equal to the noise variance.

Part (b): Gradient Descent Recurrence for Expected Weight

Goal: Write the recurrence relation for $\mathbb{E}[w_{t+1}]$ in terms of x, σ^2, y, η , and $\mathbb{E}[w_t]$ for a single data point scalar case ($x \in \mathbb{R}$).

Setup:

- Objective at step t : $L_t(w) = \frac{1}{2}(\check{X}_t w - y)^2$ where $\check{X}_t = x + N_t$.
- Gradient: $\nabla L_t(w) = (\check{X}_t w - y)\check{X}_t = ((x + N_t)w - y)(x + N_t)$.
- Update Rule: $w_{t+1} = w_t - \eta \nabla L_t(w_t)$.

Derivation:

1. Take the expectation of the update rule with respect to the noise N_t . Note that w_t depends on previous noise but is independent of N_t .

$$\mathbb{E}[w_{t+1}|w_t] = w_t - \eta \mathbb{E}[((x + N_t)w_t - y)(x + N_t)]$$

2. Expand the term inside expectation:

$$\begin{aligned} ((x + N_t)w_t - y)(x + N_t) &= (xw_t + N_t w_t - y)(x + N_t) \\ &= (xw_t - y)x + (xw_t - y)N_t + N_t w_t x + N_t^2 w_t \end{aligned}$$

3. Expectation of terms (knowing $\mathbb{E}[N_t] = 0$ and $\mathbb{E}[N_t^2] = \sigma^2$):

- $\mathbb{E}[(xw_t - y)x] = x(xw_t - y)$
- $\mathbb{E}[(xw_t - y)N_t] = 0$
- $\mathbb{E}[N_t w_t x] = 0$
- $\mathbb{E}[N_t^2 w_t] = \sigma^2 w_t$

4. Combine:

$$\mathbb{E}[\nabla L_t] = x(xw_t - y) + \sigma^2 w_t = (x^2 + \sigma^2)w_t - xy$$

5. Now take full expectation $\mathbb{E}[\cdot]$ over all previous steps:

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

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

Solution (b): The recurrence is:

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

Part (c): Convergence Condition

Goal: Find values of η for which the expectation converges.

Reasoning: The recurrence is of the form $u_{t+1} = \rho u_t + C$, where $\rho = 1 - \eta(x^2 + \sigma^2)$. For this linear dynamic system to be stable (converge), the magnitude of the multiplicative factor must be less than 1.

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

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

- Right side: $-\eta(x^2 + \sigma^2) < 0 \implies \eta > 0$ (since $x^2 + \sigma^2 > 0$).
- Left side: $-1 < 1 - \eta(x^2 + \sigma^2) \implies \eta(x^2 + \sigma^2) < 2 \implies \eta < \frac{2}{x^2 + \sigma^2}$.

Solution (c): The condition for convergence is:

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

Part (d): Expected Limit at Infinity

Goal: Find $\lim_{t \rightarrow \infty} \mathbb{E}[w_t]$ and compare it to the noiseless case.

1. Finding the Limit: At convergence, $\mathbb{E}[w_{t+1}] = \mathbb{E}[w_t] = w_\infty$. Substitute into the recurrence from (b):

$$w_\infty = (1 - \eta(x^2 + \sigma^2))w_\infty + \eta xy$$

$$w_\infty - (1 - \eta(x^2 + \sigma^2))w_\infty = \eta xy$$

$$w_\infty[1 - (1 - \eta x^2 - \eta \sigma^2)] = \eta xy$$

$$w_\infty[\eta(x^2 + \sigma^2)] = \eta xy$$

Divide by η (assuming $\eta > 0$):

$$w_\infty(x^2 + \sigma^2) = xy$$

$$w_\infty = \frac{xy}{x^2 + \sigma^2}$$

2. Comparison to Noiseless Case ($\sigma^2 = 0$): If there were no noise, we would minimize $\frac{1}{2}(xw - y)^2$. The optimal solution is simply $w^* = y/x$ (assuming $x \neq 0$). Looking at our result, if $\sigma^2 = 0$, $w_\infty = \frac{xy}{x^2} = \frac{y}{x}$.

Difference: The term σ^2 in the denominator acts exactly like a regularization term.

$$w_\infty = \frac{xy}{x^2 + \sigma^2}$$



This "shrinks" the weight toward 0 compared to the unregularized solution y/x . This confirms the result from part (a): noise acts as L2 regularization (Ridge Regression), biasing the solution towards zero to reduce variance.

Solution (d):

- **Limit:** $\mathbb{E}[w_t] \rightarrow \frac{xy}{x^2 + \sigma^2}$
- **Comparison:** This is smaller (in magnitude) than the optimal unregularized value $w^* = y/x$. The noise effectively adds a term σ^2 to the "covariance" x^2 , regularizing the inversion.



That concludes the problem set! Do you have any further questions or need clarification on any of the derivations?

Why do you assume that this question is the last question?

Please move onto the next question:



Show thinking ▾

My apologies! I missed that there was a Problem 6 in the uploaded screenshots.

Here are the step-by-step solutions for **Problem 6: General Case Tikhonov Regularization**.

Part (a): Solve the Optimization Problem Manually



Goal: Solve $\min_{\mathbf{x}} \|W_1(A\mathbf{x} - \mathbf{b})\|_2^2 + \|W_2(\mathbf{x} - \mathbf{c})\|_2^2$.

Derivation:

1. **Expand the Objective Function:** Let $J(\mathbf{x}) = \|W_1(A\mathbf{x} - \mathbf{b})\|_2^2 + \|W_2(\mathbf{x} - \mathbf{c})\|_2^2$. Recall that $\|\mathbf{v}\|_2^2 = \mathbf{v}^T \mathbf{v}$.

$$J(\mathbf{x}) = (W_1(A\mathbf{x} - \mathbf{b}))^T (W_1(A\mathbf{x} - \mathbf{b})) + (W_2(\mathbf{x} - \mathbf{c}))^T (W_2(\mathbf{x} - \mathbf{c}))$$

$$J(\mathbf{x}) = (A\mathbf{x} - \mathbf{b})^T W_1^T W_1 (A\mathbf{x} - \mathbf{b}) + (\mathbf{x} - \mathbf{c})^T W_2^T W_2 (\mathbf{x} - \mathbf{c})$$

2. **Take the Gradient with respect to \mathbf{x} :** We use the identity $\nabla_{\mathbf{x}}(\mathbf{x}^T M \mathbf{x}) = 2M\mathbf{x}$ (for symmetric M) and $\nabla_{\mathbf{x}}(\mathbf{x}^T \mathbf{v}) = \mathbf{v}$.

- **Term 1:** $(A\mathbf{x} - \mathbf{b})^T W_1^T W_1 (A\mathbf{x} - \mathbf{b})$ Gradient: $2A^T W_1^T W_1 (A\mathbf{x} - \mathbf{b})$
- **Term 2:** $(\mathbf{x} - \mathbf{c})^T W_2^T W_2 (\mathbf{x} - \mathbf{c})$ Gradient: $2W_2^T W_2 (\mathbf{x} - \mathbf{c})$

Set the total gradient to zero:

$$2A^T W_1^T W_1 (A\mathbf{x} - \mathbf{b}) + 2W_2^T W_2 (\mathbf{x} - \mathbf{c}) = \mathbf{0}$$

3. **Solve for \mathbf{x} :** Divide by 2 and expand:

$$A^T W_1^T W_1 A \mathbf{x} - A^T W_1^T W_1 \mathbf{b} + W_2^T W_2 \mathbf{x} - W_2^T W_2 \mathbf{c} = \mathbf{0}$$

Group the \mathbf{x} terms:

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

Invert the matrix on the left side to find \mathbf{x}^* :

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

Part (b): Construct Matrix C and Vector d for OLS Form

Goal: Rewrite the problem as $\min_{\mathbf{x}} \|C\mathbf{x} - \mathbf{d}\|_2^2$ and confirm the solution matches.

Reasoning: The sum of two squared norms can be written as a single squared norm by stacking vectors/matrices vertically.

$$\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = \left\| \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \right\|_2^2$$

Here, we want:

$$\|W_1(A\mathbf{x} - \mathbf{b})\|_2^2 + \|W_2(\mathbf{x} - \mathbf{c})\|_2^2 = \underbrace{\|W_1 A \mathbf{x} - W_1 \mathbf{b}\|_2^2}_{\text{top part}} + \underbrace{\|W_2 \mathbf{x} - W_2 \mathbf{c}\|_2^2}_{\text{bottom part}}$$

This looks like $\|C\mathbf{x} - \mathbf{d}\|^2$ where:

$$C\mathbf{x} - \mathbf{d} = \begin{bmatrix} W_1 A \mathbf{x} - W_1 \mathbf{b} \\ W_2 \mathbf{x} - W_2 \mathbf{c} \end{bmatrix} = \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix} \mathbf{x} - \begin{bmatrix} W_1 \mathbf{b} \\ W_2 \mathbf{c} \end{bmatrix}$$

Construction:

- **Matrix C:** Stack $W_1 A$ on top of W_2 .

$$C = \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix}$$

- **Vector d:** Stack $W_1 \mathbf{b}$ on top of $W_2 \mathbf{c}$.

$$\mathbf{d} = \begin{bmatrix} W_1 \mathbf{b} \\ W_2 \mathbf{c} \end{bmatrix}$$

Verification using OLS Solution $\mathbf{x}^ = (C^T C)^{-1} C^T \mathbf{d}$:*

1. Compute $C^T C$:

$$C^T C = \begin{bmatrix} (W_1 A)^T & W_2^T \end{bmatrix} \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix} = A^T W_1^T W_1 A + W_2^T W_2$$

2. Compute $C^T \mathbf{d}$:

$$C^T \mathbf{d} = \begin{bmatrix} (W_1 A)^T & W_2^T \end{bmatrix} \begin{bmatrix} W_1 \mathbf{b} \\ W_2 \mathbf{c} \end{bmatrix} = A^T W_1^T W_1 \mathbf{b} + W_2^T W_2 \mathbf{c}$$

3. Result:

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

This matches exactly with the result from **Part (a)**.

Part (c): Reduce to Simple Ridge Regression

Goal: Choose W_1, W_2, \mathbf{c} such that the solution becomes $\mathbf{x}^* = (A^T A + \lambda I)^{-1} A^T \mathbf{b}$.

Reasoning: The standard Ridge Regression objective is:

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_2^2$$

Comparing this to our general form:

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

1. **Match the First Term:** We need $\|W_1(A\mathbf{x} - \mathbf{b})\|_2^2$ to look like $\|A\mathbf{x} - \mathbf{b}\|_2^2$. This happens if W_1 is the identity matrix (I). (Technically, any unitary matrix would preserve the norm, but Identity is the simplest choice).
2. **Match the Second Term:** We need $\|W_2(\mathbf{x} - \mathbf{c})\|_2^2$ to look like $\lambda \|\mathbf{x}\|_2^2 = \|\sqrt{\lambda} \mathbf{x}\|_2^2$.
 - Set $\mathbf{c} = \mathbf{0}$ (to remove the shift).
 - Set $W_2 = \sqrt{\lambda} I$ (a diagonal matrix with $\sqrt{\lambda}$ on the diagonal). Then $\|W_2 \mathbf{x}\|_2^2 = \|\sqrt{\lambda} I \mathbf{x}\|_2^2 = (\sqrt{\lambda})^2 \|\mathbf{x}\|_2^2 = \lambda \|\mathbf{x}\|_2^2$.

Solution (c): To recover standard Ridge Regression, choose:

- $W_1 = I$ (Identity matrix)
- $W_2 = \sqrt{\lambda} I$
- $\mathbf{c} = \mathbf{0}$ (Zero vector)

Verification: Substitute these into the formula from (a):

$$\mathbf{x}^* = (A^T I^T I A + (\sqrt{\lambda} I)^T (\sqrt{\lambda} I))^{-1} (A^T I^T I \mathbf{b} + (\sqrt{\lambda} I)^T (\sqrt{\lambda} I) \mathbf{0})$$

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



Show thinking ▾

Here is the step-by-step solution for **Problem 7: An Alternate MAP Interpretation of Ridge Regression.**

Goal:

Show that the two forms of the Ridge Regression solution are identical using properties of joint Gaussian distributions:

1. **Primal Form:** $\hat{\mathbf{w}} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$
2. **Dual (Kernel) Form:** $\hat{\mathbf{w}} = X^T (X X^T + \lambda I)^{-1} \mathbf{y}$

We are given that $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, I)$ and $\mathbf{Y} = X\mathbf{W} + \sqrt{\lambda}\mathbf{N}$, where $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, I)$. We need to find the MAP estimate, which corresponds to the conditional expectation $E[\mathbf{W}|\mathbf{Y} = \mathbf{y}]$.

Step 1: Determine the Joint Distribution of \mathbf{W} and \mathbf{Y}

The vectors \mathbf{W} and \mathbf{Y} are jointly Gaussian because \mathbf{Y} is a linear transformation of Gaussian vector \mathbf{W} plus independent Gaussian noise. The joint distribution can be characterized by their means and the covariance matrix:

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{WW} & \Sigma_{WY} \\ \Sigma_{YW} & \Sigma_{YY} \end{bmatrix} \right)$$

Compute the Covariance Components:

1. $\Sigma_{WW} = E[\mathbf{W}\mathbf{W}^T]$: Since $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, I)$, we have:

$$\Sigma_{WW} = I$$

2. $\Sigma_{WY} = E[\mathbf{W}\mathbf{Y}^T]$: Substitute $\mathbf{Y} = X\mathbf{W} + \sqrt{\lambda}\mathbf{N}$:

$$E[\mathbf{W}(X\mathbf{W} + \sqrt{\lambda}\mathbf{N})^T] = E[\mathbf{W}\mathbf{W}^T X^T] + E[\mathbf{W}\mathbf{N}^T \sqrt{\lambda}]$$

Since \mathbf{W} and \mathbf{N} are independent, $E[\mathbf{W}\mathbf{N}^T] = 0$. Also $E[\mathbf{W}\mathbf{W}^T] = I$.

$$\Sigma_{WY} = IX^T = X^T$$

(Consequently, $\Sigma_{YW} = \Sigma_{WY}^T = X$).

3. $\Sigma_{YY} = E[\mathbf{Y}\mathbf{Y}^T]$:

$$E[(X\mathbf{W} + \sqrt{\lambda}\mathbf{N})(X\mathbf{W} + \sqrt{\lambda}\mathbf{N})^T]$$

Expand terms:

$$= E[X\mathbf{W}\mathbf{W}^T X^T] + E[\sqrt{\lambda}\mathbf{N}\mathbf{N}^T \sqrt{\lambda}] + \text{cross terms (0)}$$

$$= XE[\mathbf{W}\mathbf{W}^T]X^T + \lambda E[\mathbf{N}\mathbf{N}^T]$$

$$= XIX^T + \lambda I = XX^T + \lambda I$$

Step 2: Calculate the Conditional Expectation

For jointly normal variables with mean zero, the conditional expectation $E[\mathbf{W}|\mathbf{Y} = \mathbf{y}]$ is given by the formula (Equation 32 in the prompt):

$$E[\mathbf{W}|\mathbf{Y} = \mathbf{y}] = \Sigma_{WY}\Sigma_{YY}^{-1}\mathbf{y}$$

Substitute the covariance matrices we computed in Step 1:

- $\Sigma_{WY} = X^T$
- $\Sigma_{YY} = XX^T + \lambda I$

Thus:

$$\hat{\mathbf{w}} = X^T(XX^T + \lambda I)^{-1}\mathbf{y}$$

This confirms the **Dual (Kernel) Form** of Ridge Regression directly from the Gaussian conditioning formula.

Step 3: Prove Equivalence to the Primal Form

We now need to show that:

$$X^T(XX^T + \lambda I)^{-1}\mathbf{y} = (X^TX + \lambda I)^{-1}X^T\mathbf{y}$$

This is a classic matrix identity proof using the **Woodbury Matrix Identity** or simply algebraic manipulation.

Algebraic Proof: Start with the identity:

$$X^T(XX^T + \lambda I) = (X^TX + \lambda I)X^T$$

(Let's verify this by expanding both sides):

- **LHS:** $X^TXX^T + X^T\lambda I = X^TXX^T + \lambda X^T$
- **RHS:** $X^TXX^T + \lambda IX^T = X^TXX^T + \lambda X^T$ Both sides are identical.

Now, perform the following operations on this equality:

1. Multiply by $(X^TX + \lambda I)^{-1}$ on the **left**:



$$(X^TX + \lambda I)^{-1}X^T(XX^T + \lambda I) = X^T$$

2. Multiply by $(XX^T + \lambda I)^{-1}$ on the **right**:

$$(X^TX + \lambda I)^{-1}X^T = X^T(XX^T + \lambda I)^{-1}$$

Finally, multiply both sides by \mathbf{y} from the right:

$$(X^TX + \lambda I)^{-1}X^T\mathbf{y} = X^T(XX^T + \lambda I)^{-1}\mathbf{y}$$

Conclusion: We have shown that the MAP estimate derived from the Gaussian prior interpretation yields the dual form $\hat{\mathbf{w}} = X^T(XX^T + \lambda I)^{-1}\mathbf{y}$, and this form is algebraically identical to the standard primal form $\hat{\mathbf{w}} = (X^TX + \lambda I)^{-1}X^T\mathbf{y}$.

Is there anything else you would like help with regarding these homework problems?