

## Special Participation A: ChatGPT on HW1

- Introduction

At first, I simply gave ChatGPT the command “Solve this” along with a screenshot of the problem. However, the responses were not very good, so I consulted with ChatGPT to refine and create a more appropriate initial command. Also, since it tended to misread small images, I made sure to take larger screenshots.

Note:

ChatGPT was set to Thinking mode.

In the following, the gray sections represent the commands, the uncolored sections are ChatGPT’s responses, and the red text contains my commentary on ChatGPT’s answers.

Link to ChatGPT:

[https://chatgpt.com/g/g-p-69002ba693e08191b423a557a254a250-special-participation/shared/c/690035f9-63f8-8332-9361-cfed492e732b?owner\\_user\\_id=user-9DA7E48mSHEKUmgRD3SiXzUm](https://chatgpt.com/g/g-p-69002ba693e08191b423a557a254a250-special-participation/shared/c/690035f9-63f8-8332-9361-cfed492e732b?owner_user_id=user-9DA7E48mSHEKUmgRD3SiXzUm)

- Interactions

You are a large language model being evaluated for your ability to reason through graduate-level machine learning theory questions without coding. I will give you one question at a time as an image (each image corresponds to one subproblem of a homework assignment).

Your goal is to:

Read and interpret the problem correctly.

Produce the analytical or conceptual solution, showing intermediate reasoning.

Avoid hallucinating or making up definitions — if information is missing, explicitly say so.

Answer in a way suitable for a graduate student homework (clear math notation and concise justification).

After each answer, I will interactively challenge or clarify parts of your reasoning.

Rules:

You may use LaTeX-style math notation ( $\Delta\theta$ ,  $\mathbb{I}\cdot\mathbb{I}_2$ , etc.).

Do not skip derivation steps.

Think step-by-step but write concisely.

If the problem contains multiple subparts (a), (b), (c), answer each clearly.

Once ready, I will start by giving you the first question image. Please confirm you understand the task.

Got it—I understand the task and the rules.

When you send each image, I will:

1. restate the problem in my own words to confirm interpretation,
2. solve it step-by-step with clear intermediate derivations,
3. flag any missing information or ambiguous assumptions,
4. present final results in clean math notation, with each subpart (a), (b), (c) labeled.

Please share the first question image when you're ready.

## 1. Why Learning Rates Cannot be Too Big

This is input. 思考時間 is thinking time because my language setting is Japanese.

or problem where we have two depending on  $\eta, \sigma_\ell, \sigma_s$ , which of  $\sigma_s$ , i.e.  $\sigma_\ell \gg \sigma_s$  and we have th slower?

all will be dominated by the slowe convergence to the solution? n if we had more parallel problems the choice of possible learning r

single learning rate  $\eta$  to solve t

SVD, how is the simple scalar an will we converge? Which of t eral least-squares problems of it

$$\begin{bmatrix} \sigma_\ell & 0 \\ 0 & \sigma_s \end{bmatrix} \begin{bmatrix} w[1] \\ w[2] \end{bmatrix} = \begin{bmatrix} y[1] \\ y[2] \end{bmatrix}$$

思考時間: 1m 42s

Output is below.

I'll restate the setup and then answer (a)–(g).

We minimize  $L(w) = (y - \sigma w)^2$  with GD step size  $\eta > 0$ .

The update is

$$w_{t+1} = w_t - \eta \nabla L(w_t) = w_t - \eta \cdot 2\sigma(\sigma w_t - y) = (1 - 2\eta\sigma^2)w_t + 2\eta\sigma y. \quad (*)$$

Let  $w^* = \frac{y}{\sigma}$  (the exact solution). Define the error  $e_t := w_t - w^*$ . Subtracting  $w^*$  from (\*) gives

$$e_{t+1} = (1 - 2\eta\sigma^2) e_t \Rightarrow e_t = (1 - 2\eta\sigma^2)^t e_0. \quad (**)$$

#### (a) Stability range for $\eta$

Stability (convergence for any  $w_0$ ) requires  $|1 - 2\eta\sigma^2| < 1$ .

Thus

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

#### (b) Iteration complexity below the bound

From (\*\*), with  $w_0 = 0 \Rightarrow e_0 = -w^*$ ,

$$|e_t| = |1 - 2\eta\sigma^2|^t |w^*|.$$

To ensure relative error  $\leq \varepsilon$ , i.e.  $|w_t - w^*| \leq \varepsilon |w^*|$ ,

$$|1 - 2\eta\sigma^2|^t \leq \varepsilon \implies t \geq \frac{\ln(\varepsilon)}{\ln|1 - 2\eta\sigma^2|}.$$

(Here  $|1 - 2\eta\sigma^2| < 1$ , so the denominator is negative and the RHS is positive.)

A common smooth bound: if  $0 < \eta \leq \frac{1}{2\sigma^2}$  then

$$|1 - 2\eta\sigma^2| \leq e^{-2\eta\sigma^2}, \text{ giving}$$

$$t \geq \frac{1}{2\eta\sigma^2} \ln(1/\varepsilon).$$

Now the 2-D diagonal problem with  $\text{diag}(\sigma_\ell, \sigma_s)$  where  $\sigma_\ell \gg \sigma_s$ , single step size  $\eta$ , start at  $\mathbf{w}_0 = \mathbf{0}$ .

For coordinate  $i \in \{\ell, s\}$ , the error recurrence is

$$e_{i,t+1} = (1 - 2\eta\sigma_i^2) e_{i,t}, \quad r_i(\eta) := |1 - 2\eta\sigma_i^2|.$$

#### (c) Convergence range and limiting $\sigma_i$

We need both coordinates stable:

$$|1 - 2\eta\sigma_\ell^2| < 1, \quad |1 - 2\eta\sigma_s^2| < 1 \iff 0 < \eta < \frac{1}{\max(\sigma_\ell^2, \sigma_s^2)} = \frac{1}{\sigma_\ell^2}.$$

Hence  $\sigma_\ell$  (the larger one) limits the learning rate.

**(d) Which coordinate converges faster?**

"Faster" means smaller contraction factor  $r_i(\eta)$ .

- For small  $\eta$  (in particular  $\eta \leq \frac{1}{2\sigma_\ell^2}$ ),  
 $r_i(\eta) = 1 - 2\eta\sigma_i^2$  decreases with  $\sigma_i$ ; thus the  $\ell$ -coordinate (larger  $\sigma$ ) is faster.
- As  $\eta$  grows, each coordinate has its own best  $\eta_i^* = \frac{1}{2\sigma_i^2}$  (where  $r_i = 0$ ).

Since  $\eta_\ell^* < \eta_s^*$ , once  $\eta$  passes  $\eta_\ell^*$  the  $\ell$ -coordinate begins to overshoot (its factor becomes negative, with magnitude increasing), while the  $s$ -coordinate continues improving; near  $\eta \approx \eta_s^*$  (if feasible) the  $s$ -coordinate is faster.

In our feasible range  $\eta < 1/\sigma_\ell^2$ , typically  $\eta_s^* = 1/(2\sigma_s^2) > 1/\sigma_\ell^2$ , so  $r_s(\eta)$  simply decreases with  $\eta$  over the whole admissible range.

**(e) Step size for fastest overall convergence**

Overall rate is governed by the worst coordinate:

$$\rho(\eta) = \max\{r_\ell(\eta), r_s(\eta)\}.$$

The optimal constant  $\eta$  equalizes the two magnitudes with one on each side of zero:

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

(which indeed lies in  $[\frac{1}{2\sigma_\ell^2}, \frac{1}{2\sigma_s^2}]$  and below  $1/\sigma_\ell^2$ ).

The resulting optimal contraction (spectral radius) is

$$\boxed{\rho^* = \frac{\sigma_\ell^2 - \sigma_s^2}{\sigma_\ell^2 + \sigma_s^2} = \frac{\kappa - 1}{\kappa + 1}, \quad \kappa := \frac{\sigma_\ell^2}{\sigma_s^2}}$$

so the gap (condition number)  $\kappa$  controls the speed.  $\downarrow$

**(f) Many parallel coordinates with  $\sigma_s \leq \sigma_i \leq \sigma_\ell$**

The admissible range is still  $0 < \eta < 1/\sigma_\ell^2$ .

To minimize the worst-case contraction  $\max_i |1 - 2\eta\sigma_i^2|$ , the same equalization of extremes applies, giving

$$\eta^* = \frac{1}{\sigma_\ell^2 + \sigma_s^2}, \quad \max_i |1 - 2\eta^*\sigma_i^2| = \frac{\sigma_\ell^2 - \sigma_s^2}{\sigma_\ell^2 + \sigma_s^2}.$$

Intermediate  $\sigma_i$ 's only matter through how far they sit between the extremes; the extremes ( $\sigma_{\max}, \sigma_{\min}$ ) determine both the allowable range and the optimal global step.

**(g) Relevance to general least squares  $Xw \approx y$**

For  $f(w) = \|y - Xw\|_2^2$ , GD uses

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

Let  $X = U\Sigma V^\top$  be the SVD, with singular values  $\sigma_i$ . In the rotated basis  $z = V^\top w$ , the dynamics decouple into independent scalar recurrences

$$z_{i,t+1} - z_i^* = (1 - \eta\lambda_i) (z_{i,t} - z_i^*), \quad \lambda_i = 2\sigma_i^2,$$

exactly the scalar form above. Therefore:

- Convergence iff  $0 < \eta < 2/\lambda_{\max} = 1/\sigma_{\max}^2$ .
- The global rate is  $\max_i |1 - \eta\lambda_i|$ ; the optimal constant step is

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

with optimal contraction

$$\rho^* = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\sigma_{\max}^2 - \sigma_{\min}^2}{\sigma_{\max}^2 + \sigma_{\min}^2} = \frac{\kappa - 1}{\kappa + 1},$$

where  $\kappa = \lambda_{\max}/\lambda_{\min} = (\sigma_{\max}/\sigma_{\min})^2$  is the (squared) condition number.

Thus the scalar/diagonal analysis exactly characterizes GD on least squares when viewed in the SVD coordinates.

All of ChatGPT's answers were correct. A notable characteristic of its responses was a tendency to present the solutions in a highly mathematical manner. In particular, for problem (g), the explanation was more mathematical than necessary. This is likely because the initial command included the phrase “showing intermediate reasoning.”



### (b) SVD coordinates and the initial condition in those coordinates

Take a compact SVD  $X = U\Sigma V^\top$  with

$\Sigma = [\text{diag}(\sigma_1, \dots, \sigma_n) \ 0_{n \times (d-n)}]$ , and all

$\sigma_i > 0$  (full row rank).

Define the orthonormal change of variables  $w'' := V^\top w'$  (equivalently  $w'' = V w'$  under the notation in the problem; we'll use  $V^\top$  to emphasize an orthogonal rotation).

Then

$$X w' = 0 \iff U \Sigma V^\top w' = 0 \iff \Sigma w'' = 0 \iff [\tilde{X} \ 0_{n \times (d-n)}] w'' = 0,$$

with  $\tilde{X} = \text{diag}(\sigma_1, \dots, \sigma_n)$ . \tag{7}

For the initial condition:  $w'_0 = -w^*$  is the min-norm interpolant and hence lies in the row-space of  $X$ , which is  $\text{span}\{V_1, \dots, V_n\}$ . Therefore in  $w''_0 = V^\top w'_0$  the last  $d - n$  coordinates are **all zero**.

### (c) Reducing to a square system and relation to the original rows

Write  $w'' = (\tilde{w}, \bar{w})$  with  $\tilde{w} \in \mathbb{R}^n$  the first  $n$  coordinates and  $\bar{w} \in \mathbb{R}^{d-n}$  the last part. From (7),

$$\tilde{X} \tilde{w} = 0, \quad \tilde{X} = \text{diag}(\sigma_1, \dots, \sigma_n). \tag{8}$$

Left-multiplying the original equations by the orthonormal  $U^\top$  simply forms orthonormal linear combinations of the original rows; thus each of the  $n$  scalar equations in (8) is the image, under this orthonormal row transformation and the  $V$ -rotation of variables, of the correspondingly indexed equation in  $X w' = 0$  (or equivalently, of (5)). So we can focus on the **square** system (8).

### (d) SGD update in original and rotated coordinates

With minibatch size 1 and  $I_t \sim \text{Unif}\{1, \dots, n\}$ ,

$$w_{t+1} = w_t - \eta \nabla \mathcal{L}_{I_t}(w_t), \quad \mathcal{L}_i(w) = (y_i - x_i^\top w)^2.$$

Since in the shifted, homogeneous system  $y$  disappears,  $\nabla \mathcal{L}_i(w') = -2(-x_i^\top w')x_i = 2(x_i^\top w')x_i$ .

Carrying the same coordinate changes as in (b) (so  $x_i$  becomes its rotated version  $\tilde{x}_i$ , the  $i$ -th row of  $\tilde{X}$ ), the SGD step on (8) is

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

(10)

which is exactly the rotated form of the original SGD step.

**(e) From a one-step contraction in expectation to a high-probability guarantee**

Assume there exists  $\rho \in (0, 1)$  such that for a Lyapunov function  $\mathcal{L}$ ,

$$\mathbb{E}[\mathcal{L}(\tilde{w}_{t+1}) \mid \tilde{w}_t] \leq (1 - \rho) \mathcal{L}(\tilde{w}_t). \quad (11)$$

Iterating and taking expectations gives

$$\mathbb{E}[\mathcal{L}(\tilde{w}_T)] \leq (1 - \rho)^T \mathcal{L}(\tilde{w}_0).$$

By Markov's inequality,

$$\Pr(\mathcal{L}(\tilde{w}_T) \geq \varepsilon) \leq \frac{\mathbb{E}[\mathcal{L}(\tilde{w}_T)]}{\varepsilon} \leq \frac{(1 - \rho)^T \mathcal{L}(\tilde{w}_0)}{\varepsilon}.$$

So for any  $\varepsilon > 0, \delta > 0$ , choosing

$$T \geq \frac{\ln(\mathcal{L}(\tilde{w}_0)/(\varepsilon\delta))}{\ln(\frac{1}{1-\rho})}$$

ensures  $\Pr(\mathcal{L}(\tilde{w}_T) < \varepsilon) \geq 1 - \delta$ . \tag{12}

**(f) A natural Lyapunov function**

Let

$$\mathcal{L}(\tilde{w}) = \tilde{w}^\top \tilde{X}^\top \tilde{X} \tilde{w} = \|\tilde{X} \tilde{w}\|_2^2. \quad (13)$$

This is non-negative for all  $\tilde{w}$ . In the square, full-rank case,  $\tilde{X}$  is invertible, so

$$\mathcal{L}(\tilde{w}) = 0 \iff \tilde{X} \tilde{w} = 0 \iff \tilde{w} = 0.$$

Thus (13) is a valid Lyapunov function.

**(g) One-step decomposition**

Let  $\Delta_t := \tilde{w}_{t+1} - \tilde{w}_t = -2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t$ .

Then

$$\begin{aligned} \mathcal{L}(\tilde{w}_{t+1}) &= \|\tilde{X}(\tilde{w}_t + \Delta_t)\|^2 = \|\tilde{X} \tilde{w}_t\|^2 + 2\langle \tilde{X} \tilde{w}_t, \tilde{X} \Delta_t \rangle + \|\tilde{X} \Delta_t\|^2 \\ &= \mathcal{L}(\tilde{w}_t) + A + B, \end{aligned} \quad (14)$$

with the explicit forms

$$A = 2 \tilde{w}_t^\top \tilde{X}^\top \tilde{X} \Delta_t = -4\eta \tilde{w}_t^\top \tilde{X}^\top \tilde{X} \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t,$$

$$B = \|\tilde{X} \Delta_t\|^2 = 4\eta^2 \|\tilde{X} \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t\|^2.$$



**(h) The linear term contracts in expectation**

Let  $G_i := \tilde{x}_i$ . Averaging over  $I_t \sim \text{Unif}\{1, \dots, n\}$ ,

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

Let  $\lambda_{\min} > 0$  be the smallest eigenvalue of  $\tilde{X}^\top \tilde{X}$ . Then

$$\tilde{w}_t^\top (\tilde{X}^\top \tilde{X})^2 \tilde{w}_t \geq \lambda_{\min} \tilde{w}_t^\top (\tilde{X}^\top \tilde{X}) \tilde{w}_t = \lambda_{\min} \mathcal{L}(\tilde{w}_t).$$

Hence

$$\mathbb{E}[A \mid \tilde{w}_t] \leq -c_1 \eta \mathcal{L}(\tilde{w}_t), \quad c_1 := \frac{4}{n} \lambda_{\min}(\tilde{X}^\top \tilde{X}) > 0. \quad (15)$$

**(i) The quadratic term is small**

Let  $\sigma_{\max}$  be the largest singular value of  $\tilde{X}$ , and let

$\beta := \max_i \|G_i\|_2$  (max row norm). Then

$$\|\tilde{X} G_i G_i^\top \tilde{w}_t\| \leq \|\tilde{X}\| \|G_i\| |G_i^\top \tilde{w}_t| \leq \sigma_{\max} \beta |G_i^\top \tilde{w}_t|.$$

Therefore

$$B \leq 4\eta^2 \sigma_{\max}^2 \beta^2 (G_t^\top \tilde{w}_t)^2.$$

Taking expectation and using

$$\frac{1}{n} \sum_{i=1}^n (G_i^\top \tilde{w}_t)^2 = \|\tilde{X} \tilde{w}_t\|^2 / n = \mathcal{L}(\tilde{w}_t) / n,$$

$$\mathbb{E}[B \mid \tilde{w}_t] \leq c_2 \eta^2 \mathcal{L}(\tilde{w}_t), \quad c_2 := \frac{4}{n} \sigma_{\max}^2(\tilde{X}) \beta^2 > 0. \quad (16)$$

**(j) Putting it together**

From (14)–(16),

$$\mathbb{E}[\mathcal{L}(\tilde{w}_{t+1}) \mid \tilde{w}_t] \leq (1 - c_1 \eta + c_2 \eta^2) \mathcal{L}(\tilde{w}_t). \quad (17)$$

Because  $c_1 > 0, c_2 > 0$ , the quadratic  $q(\eta) = 1 - c_1 \eta + c_2 \eta^2$  satisfies  $q(0) = 1$  and  $q'(0) = -c_1 < 0$ .

Hence there exists a sufficiently small  $\eta \in (0, c_1/c_2)$  with  $q(\eta) < 1$ . For such  $\eta$ , (11) holds with  $\rho = 1 - q(\eta) > 0$ , yielding exponential decay in expectation and the high-probability guarantee from (e).

In problem (b), ChatGPT mistakenly eliminated (ignored) the  $U$  matrix when simplifying  $Xw' = U \Sigma V^T w' = 0$  to  $\Sigma w'' = 0$ . As a result, all subsequent problems except (e) were answered incorrectly, since they relied on this mistaken step. However, given the incorrect assumption in (b), ChatGPT's reasoning and solutions were internally consistent and correct within that framework. This shows that ChatGPT solved the questions sequentially, building upon its previous answers.

So I gave ChatGPT a hit below.

Please re-solve the entire problem. I believe there was a mistake in the SVD coordinate transformation in part (b). Hint: When you analyzed  $Xw' = 0$  using the SVD  $X = U\Sigma V^T$  and  $w'' = V^T w'$ , you simplified the system to  $\Sigma w'' = 0$ . However, the  $U$  matrix should not be dropped. The correct simplification is  $U\Sigma w'' = 0$ . This implies the new system matrix  $\tilde{X}$  (from equation 15) should be  $\tilde{X} = U\Sigma$ , not just the diagonal matrix  $\Sigma$ . Could you please re-derive the solutions for all parts (especially c, d, h, i) starting from this corrected transformation?

Output is as follows.

思考時間: 1m 55s

Great catch—you're right that we shouldn't silently "drop"  $U$  when we move to SVD coordinates if we want to keep a faithful per-row SGD view. Below I re-derive **all parts (a)–(j)** with the corrected transformation.

We start with the over-parameterized system

$$Xw = y, \quad X \in \mathbb{R}^{n \times d}, \quad d > n, \quad \text{rank}(X) = n,$$

and initialize  $w_0 = 0$ .

Let the **compact SVD** be  $X = U\Sigma V^T$ , where

$$U \in \mathbb{R}^{n \times n}, \quad V \in \mathbb{R}^{d \times d} \text{ orthonormal}, \quad \Sigma = [\text{diag}(\sigma_1, \dots, \sigma_n) \quad 0_{n \times (d-n)}], \quad \sigma_i > 0.$$

**(a) Shift to the min-norm solution**

The min-norm interpolating solution is

$$w^\star = X^\dagger y = X^\top (X X^\top)^{-1} y.$$

Let  $w' := w - w^\star$ . Then  $X w' = X w - X w^\star = y - y = 0$ ; i.e.,

$$X w' = 0.$$

Initial condition:  $w'_0 = w_0 - w^\star = -w^\star$ .

---

**(b) SVD-based coordinates (keeping  $U$ )**

Define the orthonormal change of variables

$$w'' := V^\top w' \iff w' = V w''.$$

Plug into  $X w' = 0$ :

$$X w' = U \Sigma V^\top (V w'') = U \Sigma w'' = 0.$$

Thus the **correct** reduced system is

$$\boxed{[\tilde{X} \ 0_{n \times (d-n)}] w'' = 0, \quad \tilde{X} := U \operatorname{diag}(\sigma_1, \dots, \sigma_n).} \quad (7)$$

(Left multiplying by  $U^\top$  would indeed give  $\Sigma w'' = 0$ , but keeping  $U$  is important for respecting row-wise sampling in SGD.)

Initial condition in these coordinates:

since  $w^\star$  (and hence  $w'_0 = -w^\star$ ) lies in the row-space of  $X = \operatorname{span}(V_{1:n})$ , we have

$$w''_0 = V^\top w'_0 = \begin{bmatrix} \tilde{w}_0 \\ \bar{w}_0 \end{bmatrix}, \quad \bar{w}_0 = 0 \in \mathbb{R}^{d-n}.$$

So the **last  $d - n$  entries are zero**.

---

**(c) Reduction to a square system and row correspondence**

Write  $w'' = (\tilde{w}, \bar{w})$  with  $\tilde{w} \in \mathbb{R}^n$ ,  $\bar{w} \in \mathbb{R}^{d-n}$ .

From (7) and  $\bar{w}_0 = 0$ , the dynamics stay in the first  $n$  coordinates and we get the **square full-rank system**

$$\boxed{\tilde{X} \tilde{w} = 0, \quad \tilde{X} = U \operatorname{diag}(\sigma_1, \dots, \sigma_n) \text{ invertible.}} \quad (8)$$

Moreover, each **row** of (8) is the corresponding orthonormal linear combination (via the same row of  $U$ ) of the equations in  $X w' = 0$ . Hence a per-row SGD step on (5) maps, under the coordinate changes, to a per-row SGD step on (8) indexed by the **same row index**.

**(d) SGD update in original variables matches SGD on (8)**

With minibatch size 1, sampling with replacement, let  $I_t \sim \operatorname{Unif}\{1, \dots, n\}$ . For the loss  $\mathcal{L}_i(w) = (y_i - x_i^\top w)^2$ ,

the shifted problem is homogeneous, and the per-row gradient is

$$\nabla \mathcal{L}_i(w') = 2(x_i^\top w') x_i.$$

Therefore

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

Transforming with  $w'_t = V w''_t$  and using  $x_i$  rotated to  $\tilde{x}_i$  (the  $i$ -th row of  $\tilde{X}$ ), one obtains for the first  $n$  coordinates:

$$\boxed{\tilde{w}_{t+1} = \tilde{w}_t - 2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t,} \quad (10)$$

where now  $\tilde{x}_i^\top$  is the  $i$ -th **row of**  $\tilde{X} = U \operatorname{diag}(\sigma)$  (not a standard basis vector times  $\sigma_i$ ).

This is the exact image of the original SGD step.

**(e) One-step contraction in expectation implies a high-probability guarantee**

If there exists  $\rho \in (0, 1)$  and a Lyapunov function  $\mathcal{L}$  such that

$$\mathbb{E}[\mathcal{L}(\tilde{w}_{t+1}) \mid \tilde{w}_t] \leq (1 - \rho) \mathcal{L}(\tilde{w}_t), \quad (11)$$

then iterating and using Markov's inequality yields: for any  $\varepsilon, \delta > 0$  there exists

$$T \geq \frac{\ln(\mathcal{L}(\tilde{w}_0)/(\varepsilon\delta))}{\ln(1/(1-\rho))}$$

such that

$$\Pr(\mathcal{L}(\tilde{w}_T) < \varepsilon) \geq 1 - \delta. \quad (12)$$

**(f) Lyapunov function**

Take

$$\mathcal{L}(\tilde{w}) := \tilde{w}^\top \tilde{X}^\top \tilde{X} \tilde{w} = \|\tilde{X} \tilde{w}\|_2^2. \quad (13)$$

Since  $\tilde{X}$  is  $n \times n$  and invertible (its singular values are  $\{\sigma_i\}_{i=1}^n$ ),

$\mathcal{L}(\tilde{w}) \geq 0$  with equality iff  $\tilde{w} = 0$ .

**(g) One-step decomposition (explicit  $A, B$ )**

Let  $\Delta_t := \tilde{w}_{t+1} - \tilde{w}_t = -2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t$ .

Then

$$\begin{aligned} \mathcal{L}(\tilde{w}_{t+1}) &= \|\tilde{X}(\tilde{w}_t + \Delta_t)\|^2 = \|\tilde{X} \tilde{w}_t\|^2 + 2\langle \tilde{X} \tilde{w}_t, \tilde{X} \Delta_t \rangle + \|\tilde{X} \Delta_t\|^2 \\ &= \mathcal{L}(\tilde{w}_t) + A + B, \end{aligned} \quad (14)$$

with

$$A = 2 \tilde{w}_t^\top \tilde{X}^\top \tilde{X} \Delta_t = -4\eta \tilde{w}_t^\top \tilde{X}^\top \tilde{X} \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t,$$

$$B = \|\tilde{X} \Delta_t\|^2 = 4\eta^2 \|\tilde{X} \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t\|^2.$$

**(h) The linear term gives contraction in expectation**

Average over  $I_t$ :

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

Let  $\lambda_{\min}(\tilde{X}^\top \tilde{X}) = \sigma_{\min}(\tilde{X})^2 = \sigma_n^2 > 0$ .

Then

$$\tilde{w}_t^\top (\tilde{X}^\top \tilde{X})^2 \tilde{w}_t \geq \lambda_{\min} \tilde{w}_t^\top (\tilde{X}^\top \tilde{X}) \tilde{w}_t = \lambda_{\min} \mathcal{L}(\tilde{w}_t),$$

so

$$\boxed{\mathbb{E}[A \mid \tilde{w}_t] \leq -c_1 \eta \mathcal{L}(\tilde{w}_t), \quad c_1 := \frac{4}{n} \lambda_{\min}(\tilde{X}^\top \tilde{X}) = \frac{4}{n} \sigma_n^2 > 0.} \quad (15)$$

**(i) The quadratic term is small in expectation**

Let  $\sigma_{\max} := \sigma_{\max}(\tilde{X})$  and

$\beta := \max_i \|\tilde{x}_i\|_2$  (maximum row norm of  $\tilde{X}$ ).

Then

$$\|\tilde{X} \tilde{x}_i \tilde{x}_i^\top \tilde{w}_t\| \leq \|\tilde{X}\| \|\tilde{x}_i\| |\tilde{x}_i^\top \tilde{w}_t| \leq \sigma_{\max} \beta |\tilde{x}_i^\top \tilde{w}_t|.$$

Therefore

$$B \leq 4\eta^2 \sigma_{\max}^2 \beta^2 (\tilde{x}_t^\top \tilde{w}_t)^2.$$

Taking expectation and using

$$\frac{1}{n} \sum_{i=1}^n (\tilde{x}_i^\top \tilde{w}_t)^2 = \frac{1}{n} \|\tilde{X} \tilde{w}_t\|^2 = \mathcal{L}(\tilde{w}_t)/n,$$

we obtain

$$\boxed{\mathbb{E}[B \mid \tilde{w}_t] \leq c_2 \eta^2 \mathcal{L}(\tilde{w}_t), \quad c_2 := \frac{4}{n} \sigma_{\max}^2(\tilde{X}) \beta^2 > 0.} \quad (16)$$

(Note how both constants now depend on  $\tilde{X} = U \text{diag}(\sigma)$ ; the singular values are unchanged by  $U$ , but  $\beta$  (max row norm) does depend on  $U$ .)

(j) **Contraction for small enough  $\eta$**

Combine (14)–(16):

$$\mathbb{E}[\mathcal{L}(\tilde{w}_{t+1}) \mid \tilde{w}_t] \leq (1 - c_1\eta + c_2\eta^2) \mathcal{L}(\tilde{w}_t). \quad (17)$$

Since  $c_1 > 0, c_2 > 0$ , the quadratic  $q(\eta) = 1 - c_1\eta + c_2\eta^2$  satisfies  $q(0) = 1, q'(0) = -c_1 < 0$ , and  $\lim_{\eta \rightarrow 0^+} q(\eta) < 1$ . Hence there exists

$$0 < \eta < \frac{c_1}{c_2} \text{ such that } 1 - c_1\eta + c_2\eta^2 < 1,$$

which gives the desired one-step contraction (11) and thus the high-probability convergence in (12).

By providing a hint, ChatGPT was able to reach the correct answer.

The nature of this failure is quite interesting. ChatGPT did not make an arithmetic mistake; rather, it performed a “conceptual oversimplification”—it incorrectly simplified  $U\Sigma w''=0$  to  $\Sigma w''=0$ . This shows that the LLM ignored the broader context (in this case, the row-wise analysis of SGD in parts (c) and beyond) and instead made a locally valid transformation (since  $U$  is invertible) that did not align with the overall goal of the problem.

Moreover, the fact that ChatGPT arrived at the correct answer once a hint was provided suggests that it is particularly strong at conditional reasoning when given scaffolding—that is, when guided with structured hints or intermediate support.

### 3. Accelerating Gradient Descent with Momentum

This is input.

$$\mathcal{L}(w) = ||y - X$$

are time than others? Now, apply what you  
 in  $\mathcal{H}$  with the exponential variables  $x_1$  and  $x_2$ .  $x_1$  and  $x_2$  are  
 and  $\sigma$  is negligible.  
 of  $x_1$  and  $x_2$  is small, then we have  $\log(x_1) \approx \log(x_2) \approx \log(\sigma)$ .  
 $\log(x_1) \approx \log(x_2) \approx \log(\sigma)$ .

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = N \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

These are the next questions.

思考時間: 5m 39s



Output is as follows.

I'll restate the setup and then solve (a)–(i) cleanly.

We minimize the least-squares loss

$$\mathcal{L}(w) = \|y - Xw\|_2^2,$$

using **gradient descent with momentum** (EMA of gradients):

$$\boxed{\begin{aligned} w_{t+1} &= w_t - \eta z_{t+1}, \\ z_{t+1} &= (1 - \beta) z_t + \beta g_t, \quad g_t = \nabla \mathcal{L}(w_t) = 2X^\top (Xw_t - y), \end{aligned}} \quad (19)$$

where  $\eta > 0$  and  $\beta \in (0, 1]$  (when  $\beta = 1$  this becomes plain GD).

Let  $w^* = (X^\top X)^{-1} X^\top y$  and define the shifted variable  $e_t := w_t - w^*$ .

Using the compact SVD  $X = U\Sigma V^\top$  with singular values

$\sigma_1 \geq \dots \geq \sigma_r > 0$  (here  $r = \text{rank } X$ ), we work in the right-singular basis.

Define the reparameterized coordinates

$$\boxed{x_t := V^\top e_t, \quad a_t := V^\top z_t.} \quad (26)$$

#### (a) Rewrite (19) in $(x_t, a_t)$ coordinates (per singular direction)

Because  $g_t = 2X^\top X e_t = 2V\Sigma^\top \Sigma V^\top e_t = 2V \text{diag}(\sigma_i^2) V^\top e_t$ ,

we have  $V^\top g_t = 2 \text{diag}(\sigma_i^2) x_t$ .

Left-multiplying (19) by  $V^\top$  gives the scalar recurrences (for each  $i$ ):

$$\boxed{\begin{aligned} a_{t+1}[i] &= (1 - \beta) a_t[i] + 2\beta \sigma_i^2 x_t[i], \\ x_{t+1}[i] &= x_t[i] - \eta a_{t+1}[i]. \end{aligned}} \quad \text{\tag{25}}$$

#### (b) The $2 \times 2$ system matrix $R_i$

From (25'),

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

hence

$$\boxed{\begin{bmatrix} a_{t+1}[i] \\ x_{t+1}[i] \end{bmatrix} = R_i \begin{bmatrix} a_t[i] \\ x_t[i] \end{bmatrix}, \quad R_i = \begin{bmatrix} 1 - \beta & 2\beta\sigma_i^2 \\ -\eta(1 - \beta) & 1 - 2\eta\beta\sigma_i^2 \end{bmatrix}.} \quad (27)$$

Useful invariants:

$$\text{tr}(R_i) = \tau_i = 2 - \beta - 2\eta\beta\sigma_i^2, \quad \det(R_i) = \delta = 1 - \beta. \quad (*)$$

(c) Eigenvalues of  $R_i$  and when they are real / repeated / complex

The characteristic polynomial is  $\lambda^2 - \tau_i \lambda + \delta = 0$ .

Thus

$$\lambda_{i,\pm} = \frac{\tau_i \pm \sqrt{\tau_i^2 - 4\delta}}{2}.$$

Let the discriminant be

$$D_i := \tau_i^2 - 4\delta = (2 - \beta - 2\eta\beta\sigma_i^2)^2 - 4(1 - \beta).$$

- Purely real:  $D_i \geq 0$ .
- Repeated and real:  $D_i = 0$ .
- Complex-conjugate:  $D_i < 0$ .

(d) Repeated eigenvalues: stability and learning-rate at which repetition occurs

Repeated eigenvalues  $\iff D_i = 0$ :

$$|2 - \beta - 2\eta\beta\sigma_i^2| = 2\sqrt{1 - \beta}.$$

Hence the two critical learning rates are

$$\eta_{i,\pm}^{\text{rep}} = \frac{2 - \beta \pm 2\sqrt{1 - \beta}}{2\beta \sigma_i^2} \quad (\eta_{i,-}^{\text{rep}} < \eta_{i,+}^{\text{rep}}).$$

At repetition the single eigenvalue equals

$|\lambda| = \sqrt{\det(R_i)} = \sqrt{1 - \beta} < 1$  for  $\beta > 0$ , so it is **stable**.

The **largest**  $\eta$  producing repeated eigenvalues is  $\eta_{i,+}^{\text{rep}}$ .

**(e) Real (non-repeated) eigenvalues: stability region and  $\eta$ -range**

For a real 2nd-order polynomial  $\lambda^2 - \tau\lambda + \delta$ , the Jury conditions for both roots to lie strictly inside the unit circle are

$$1 - \delta > 0, \quad 1 - \tau + \delta > 0, \quad 1 + \tau + \delta > 0.$$

With  $\delta = 1 - \beta$  and  $\tau = \tau_i$  from (\*), these give

$$\beta > 0, \quad 2\eta\beta\sigma_i^2 > 0, \quad 4 - 2\beta - 2\eta\beta\sigma_i^2 > 0.$$

Therefore (using  $\eta > 0$ ),

$$0 < \eta < \frac{2 - \beta}{\beta \sigma_i^2}. \quad (\text{stable-upper-bound})$$

Combining with "real" ( $D_i \geq 0$ ),  
the **stable real-eigs** ranges are

$$\eta \in (0, \eta_{i,-}^{\text{rep}}] \cup [\eta_{i,+}^{\text{rep}}, \frac{2-\beta}{\beta \sigma_i^2}).$$

**(f) Complex eigenvalues: stability and highest  $\eta$  giving complex roots**

Complex roots occur for

$$\eta \in (\eta_{i,-}^{\text{rep}}, \eta_{i,+}^{\text{rep}}).$$

For complex conjugates of a real  $2 \times 2$  system,

$$|\lambda_{i,\pm}| = \sqrt{\det(R_i)} = \sqrt{1 - \beta};$$

hence **whenever the eigenvalues are complex they are automatically stable** (since  $\beta > 0$ ),

provided  $\eta$  also satisfies the upper-bound in

(stable-upper-bound), which it does because

$$\eta_{i,+}^{\text{rep}} < \frac{2-\beta}{\beta \sigma_i^2}.$$

The **largest  $\eta$**  with complex roots is thus  $\eta_{i,+}^{\text{rep}}$ .

**(g) Numerical design for fastest overall convergence (two modes)**

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

Compute the repetition thresholds (use  $\sqrt{1-\beta} = \sqrt{0.9} \approx 0.948683$ ):

$$2 - \beta = 1.9, \quad 2\sqrt{1-\beta} \approx 1.897366.$$

Thus

$$N_- := 2 - \beta - 2\sqrt{1-\beta} \approx 0.002634, \quad N_+ := 2 - \beta + 2\sqrt{1-\beta} \approx 3.797366.$$

For  $\sigma^2 = 5$  (large mode): denominator  $2\beta\sigma^2 = 1$

$$\eta_{+, \max}^{\text{rep}} \approx 3.797366, \quad \eta_{-, \max}^{\text{rep}} \approx 0.002634.$$

For  $\sigma^2 = 0.05$  (small mode): denominator  $2\beta\sigma^2 = 0.01$

$$\eta_{+, \min}^{\text{rep}} \approx 379.7366, \quad \eta_{-, \min}^{\text{rep}} \approx 0.2634.$$

To make **both** modes have complex eigenvalues (and hence both contract at the same modulus  $\sqrt{1-\beta}$ ), choose any

$$\eta \in (\max\{0.002634, 0.2634\}, \min\{3.797366, 379.7366\}) = (0.2634, 3.7974).$$

In this entire intersection, the asymptotic spectral radius of each mode is

$$\rho = \sqrt{1-\beta} = \sqrt{0.9} \approx 0.948683.$$

**Iterations to reach 99.9% of the optimum** (i.e., reduce the error by a factor  $10^{-3}$ ):

$$t_{\text{mom}} \geq \frac{\ln(10^{-3})}{\ln \rho} = \frac{6.9078}{-\ln(0.948683)} \approx \frac{6.9078}{0.05268} \approx 131. \text{ (about 132)}$$

For **ordinary GD**, with condition number

$$\kappa = \sigma_{\max}/\sigma_{\min} = 10, \text{ the optimal rate (with } \eta^* = 1/(\sigma_{\min}^2 + \sigma_{\max}^2))$$

is

$$\rho_{\text{GD}} = \frac{\kappa^2 - 1}{\kappa^2 + 1} = \frac{100 - 1}{100 + 1} \approx 0.980198.$$

Hence

$$t_{\text{GD}} \geq \frac{\ln(10^{-3})}{\ln \rho_{\text{GD}}} \approx \frac{6.9078}{0.019999} \approx 345.$$

**Momentum is ~2.6× fewer iterations** here.

A concrete choice:  $\eta = 1$  (it lies in  $(0.2634, 3.7974)$ )

yields the fastest asymptotic rate  $\rho = \sqrt{0.9}$ .

**(h) How do  $\sigma_i$  affect gradients and parameter updates?**

In the SVD basis,

$$V^\top g_t = 2 \operatorname{diag}(\sigma_i^2) x_t \Rightarrow g_t[i] \propto \sigma_i^2 x_t[i].$$

Thus directions with **larger**  $\sigma_i$  produce **larger gradients** for the same error coordinate  $x_t[i]$  and hence update more aggressively; directions with **small**  $\sigma_i$  move slowly. Momentum averages these gradients:

$$a_{t+1}[i] = (1 - \beta)a_t[i] + 2\beta\sigma_i^2 x_t[i],$$

thereby smoothing noise and, crucially, allowing a larger  $\eta$  without instability.

---

**(i) Which converges faster here, GD or GD+momentum, and why?**

With the momentum choice from (g), each mode's eigenvalues are complex with modulus  $|\lambda| = \sqrt{1 - \beta}$  (independent of  $\sigma_i$ ), so the **worst-case rate** is  $\sqrt{1 - \beta} \approx 0.949$ .

For GD the best possible rate is  $(\kappa^2 - 1)/(\kappa^2 + 1) \approx 0.980$ .

Therefore **momentum converges faster** because it decouples the worst-case rate from the condition number and caps it at  $\sqrt{1 - \beta}$ , while GD's rate degrades with  $\kappa$ .

For this problem as well, ChatGPT spent more than five minutes thinking. All of its answers were correct.

Similar to Problem 1, ChatGPT showed a tendency to present its solutions in a highly mathematical manner. For the reasoning and explanation questions (h) and (i), it again provided more mathematical detail than was necessary.

## 4. Optimizers

This is input.

ptimiz

**Algorithm 2 Adam Optimizer** establishes the relationship between vector-valued weights  $\theta$  refers to adding a scalar bias.

```

1: Given  $\eta = 0.001, \beta_1 = 0.9$ 
2: Initialize time step  $t \leftarrow 0$ 
3:  $m \leftarrow 0, v \leftarrow 0$ 
4: Repeat
5:    $t \leftarrow t + 1$ 
6:    $g_t \leftarrow \nabla f(\theta_{t-1})$ 
7:    $m \leftarrow \beta_1 m + (1 - \beta_1) g_t$ 
8:    $v \leftarrow \beta_2 v + (1 - \beta_2) (g_t \odot g_t)$ 
9:    $\theta_t \leftarrow \theta_{t-1} - \eta m / \sqrt{v}$ 
10: Until the stopping condition

```

in the pseudocode of Adam.

These are the next questions.

This is output.

Here are clean, step-by-step answers.

### (a) Fill in Adam (no bias correction)

With step size  $\eta$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and gradient  $g_t = \nabla f_t(\theta_{t-1})$ :

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

where  $\odot$  denotes elementwise square/multiplication.

Then update  $\theta_t \leftarrow \theta_{t-1} - \eta m_t / \sqrt{v_t}$  (optionally with an  $\epsilon$  in the denominator; the prompt says "without bias correction," so no  $\hat{m}_t, \hat{v}_t$ ).

(b) Weight decay  $\leftrightarrow L_2$  regularization; find  $\gamma$

- $L_2$ -regularized loss (per step  $t$ ):

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

- Plain SGD on the regularized loss:

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

- SGD with weight decay  $\gamma$  on the unregularized loss:

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

Comparing the two updates, they are **identical** if

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

So, SGD with weight decay  $\gamma$  using  $f_t$  is equivalent to SGD (no decay) on the  $L_2$ -regularized loss with  $\lambda = \gamma/\eta$ . (Remark: for adaptive methods like Adam, this equivalence breaks unless the decay is “decoupled” as in AdamW.)

For this problem, ChatGPT solved it almost immediately. This is probably because part (a) simply asked about the structure of Adam, so there was little need for deeper reasoning. All of ChatGPT’s answers were correct, though there were some minor differences in notation in its responses.

## 5. Regularization and Instance Noise

This is input.

Instructions: You are given a set of  $N$  data points  $\{x_i, y_i\}_{i=1}^N$ , where each  $x_i \in \mathbb{R}^d$  and each  $y_i \in \mathbb{R}$ . The data is noisy, meaning that the observed  $y_i$  is not the true value but rather  $y_i = y_i^* + \epsilon_i$ , where  $y_i^*$  is the true value and  $\epsilon_i$  is the noise. The noise is assumed to be independent and identically distributed (i.i.d.) with mean 0 and variance  $\sigma^2$ . The goal is to find a linear model  $\hat{y} = w^T x$  that minimizes the expected squared error  $\mathbb{E}[\|y - \hat{y}\|^2]$  over the noise  $\epsilon_i$ . The expected squared error can be written as  $\mathbb{E}[\|y - \hat{y}\|^2] = \mathbb{E}[\|y^* + \epsilon - \hat{y}\|^2] = \mathbb{E}[\|y^* - \hat{y}\|^2] + \mathbb{E}[\|\epsilon\|^2] + 2\mathbb{E}[\epsilon^T (y^* - \hat{y})]$ . Since  $\epsilon$  is independent of  $y^*$  and  $\hat{y}$ , the cross term vanishes, and we have  $\mathbb{E}[\|y - \hat{y}\|^2] = \mathbb{E}[\|y^* - \hat{y}\|^2] + \mathbb{E}[\|\epsilon\|^2]$ . The first term is the squared bias, and the second term is the variance of the noise. The expected squared error is minimized when the bias is zero, which occurs when  $\hat{y} = y^*$ . In this case, the expected squared error is simply the variance of the noise,  $\mathbb{E}[\|\epsilon\|^2] = N\sigma^2$ . Therefore, the expected squared error is minimized when the model is unbiased, and the minimum value is  $N\sigma^2$ .

we expect the expected squared error to be  $N\sigma^2$  for which gradient-descent does this differ from data? check your work for part 1

These are the next questions.

思考時間: 2m 8s

This is output.

**(a) Expected noisy LS equals a regularized LS and identify  $\lambda$**

Noisy inputs:  $\tilde{x}_i = x_i + N_i$ ,  $N_i \sim \mathcal{N}(0, \sigma^2 I_n)$  i.i.d.

$$\begin{aligned} \mathbb{E} \|\tilde{X}w - y\|^2 &= \sum_{i=1}^m \mathbb{E}[(\tilde{x}_i^\top w - y_i)^2] \\ &= \sum_{i=1}^m \left[ (x_i^\top w - y_i)^2 + 2(x_i^\top w - y_i) \underbrace{\mathbb{E}[N_i^\top w]}_0 + \mathbb{E}[(N_i^\top w)^2] \right] \\ &= \sum_{i=1}^m (x_i^\top w - y_i)^2 + \sum_{i=1}^m w^\top \mathbb{E}[N_i N_i^\top] w \\ &= \|Xw - y\|^2 + m \sigma^2 \|w\|^2. \end{aligned}$$

Multiplying by  $1/m$  (as in (29)) gives

$$\min_w \frac{1}{m} \|Xw - y\|^2 + \underbrace{\sigma^2}_{\lambda} \|w\|^2.$$

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

We now switch to the scalar toy model (one datapoint  $x \in \mathbb{R}$ , label  $y \in \mathbb{R}$ ), with fresh noise  $\tilde{X}_t = x + N_t$ ,  $N_t \sim \mathcal{N}(0, \sigma^2)$  i.i.d. each step, and loss  $L(w) = \frac{1}{2}(\tilde{X}_t w - y)^2$ .

Initialize  $w_0 = 0$ . Gradient step:

$$w_{t+1} = w_t - \eta (\tilde{X}_t w_t - y) \tilde{X}_t.$$

**(b) Recurrence for  $\mathbb{E}[w_{t+1}]$**

Using independence of  $N_t$  and the past (hence and  $w_t$ ) and  $\mathbb{E}[N_t] = 0$ ,

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

So

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



### (c) Learning-rate range for convergence of $\mathbb{E}[w_t]$

This is an affine scalar recurrence with contraction factor  $r = 1 - \eta(x^2 + \sigma^2)$ .

Convergence requires  $|r| < 1$ , i.e.

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

### (d) Limit of $\mathbb{E}[w_t]$ and comparison to noiseless optimum

The fixed point  $w_\infty$  solves

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

$$w_\infty = \lim_{t \rightarrow \infty} \mathbb{E}[w_t] = \frac{x y}{x^2 + \sigma^2}.$$

Without augmentation ( $\sigma^2 = 0$ ) the least-squares minimizer is  $w_{\text{clean}}^* = y/x$  (for  $x \neq 0$ ).

Noise therefore **shrinks** toward zero:

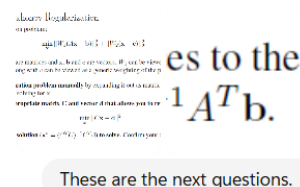
$$w_\infty = \frac{x^2}{x^2 + \sigma^2} \frac{y}{x} = \frac{1}{1 + \sigma^2/x^2} w_{\text{clean}}^*.$$

This exactly matches part (a): instance noise acts like  $L_2$  regularization with  $\lambda = \sigma^2$ , biasing the estimator toward 0.

Although it took about two minutes, all of ChatGPT's answers were correct, and it was able to solve the problem by following the given hint.

## 6. General Case Tikhonov Regularization

This is input.



This is output.

I'll solve 6(a)–(c) step-by-step.

Let  $A \in \mathbb{R}^{m \times n}$ ,  $W_1 \in \mathbb{R}^{m \times m}$ ,  $W_2 \in \mathbb{R}^{n \times n}$  (typically PSD),  $b, c \in \mathbb{R}^m, \mathbb{R}^n$ . The problem is

$$\min_x \|W_1(Ax - b)\|_2^2 + \|W_2(x - c)\|_2^2. \quad (\text{P})$$

---

### (a) Solve by setting the gradient to zero

Expand and differentiate:

$$\begin{aligned} \|W_1(Ax - b)\|_2^2 &= (Ax - b)^\top W_1^\top W_1 (Ax - b), \\ \|W_2(x - c)\|_2^2 &= (x - c)^\top W_2^\top W_2 (x - c). \end{aligned}$$

Hence

$$\nabla_x = 2A^\top W_1^\top W_1 (Ax - b) + 2W_2^\top W_2 (x - c).$$

Set  $\nabla_x = 0$ :

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

Assuming the coefficient matrix is invertible (e.g.,  $W_2^\top W_2 \succcurlyeq 0$  makes it so), the solution is

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

### (b) Write as OLS $\min_x \|Cx - d\|^2$ and solve

Stack the two weighted terms:

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

Then

$$\|Cx - d\|^2 = \|W_1(Ax - b)\|^2 + \|W_2(x - c)\|^2,$$

so (P) is  $\min_x \|Cx - d\|^2$ . The normal equations give

$$C^\top C = A^\top W_1^\top W_1 A + W_2^\top W_2, \quad C^\top d = A^\top W_1^\top W_1 b + W_2^\top W_2 c,$$

hence the OLS solution

$$x^* = (C^\top C)^{-1} C^\top d$$

matches the expression from part (a).

(c) Recover ridge regression  $x^* = (A^\top A + \lambda I)^{-1} A^\top b$

Choose

$$W_1 = I_m, \quad W_2 = \sqrt{\lambda} I_n, \quad c = \mathbf{0}.$$

Then (P) becomes  $\min_x \|Ax - b\|^2 + \lambda \|x\|^2$ , whose solution is

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

as required.

All of ChatGPT's answers were correct, and they followed the same logic as the official solutions.

## 7. An Alternate MAP Interpretation of Ridge Regression

This is input.

ridge regression is Ridge Regression / given an observation  $Y = y$  (We s  
 $p(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$   
 Ridge regression is a method for fitting a linear model to data. It is a type of regression analysis that uses a penalty term to shrink the coefficients of the model. This helps to prevent overfitting and improves the model's performance on new data. The penalty term is called the "ridge" and is proportional to the square of the magnitude of the coefficients. The resulting model is called the "ridge regression" model.  
 Ridge regression is a method for fitting a linear model to data. It is a type of regression analysis that uses a penalty term to shrink the coefficients of the model. This helps to prevent overfitting and improves the model's performance on new data. The penalty term is called the "ridge" and is proportional to the square of the magnitude of the coefficients. The resulting model is called the "ridge regression" model.  
 Ridge regression is a method for fitting a linear model to data. It is a type of regression analysis that uses a penalty term to shrink the coefficients of the model. This helps to prevent overfitting and improves the model's performance on new data. The penalty term is called the "ridge" and is proportional to the square of the magnitude of the coefficients. The resulting model is called the "ridge regression" model.  
 $E[W|Y = y] = \Sigma_{WY} \Sigma_Y^{-1} y$   
 Ridge regression is a method for fitting a linear model to data. It is a type of regression analysis that uses a penalty term to shrink the coefficients of the model. This helps to prevent overfitting and improves the model's performance on new data. The penalty term is called the "ridge" and is proportional to the square of the magnitude of the coefficients. The resulting model is called the "ridge regression" model.

These are the next questions.

思考時間: 1m 8s

This is output.

I'll show that the two ridge forms

$$\hat{w} = (X^\top X + \lambda I)^{-1} X^\top y \quad \text{and} \quad \hat{w} = X^\top (X X^\top + \lambda I)^{-1} y \quad (30-31)$$

are identical by deriving (31) from a MAP/conditioning view and then proving the algebraic identity relating the two.

---

### Generative/MAP view

Assume the standard ridge generative model

- prior on parameters:  $W \sim \mathcal{N}(0, I_d)$ ;
- observation noise:  $N \sim \mathcal{N}(0, I_n)$  independent of  $W$ ;
- labels:  $Y = XW + \sqrt{\lambda} N$ .

Then  $(W, Y)$  is jointly Gaussian (zero mean). Compute the needed covariances:

$$\begin{aligned} \Sigma_{YY} &= \mathbb{E}[YY^\top] = \mathbb{E}[(XW + \sqrt{\lambda}N)(XW + \sqrt{\lambda}N)^\top] = X\mathbb{E}[WW^\top]X^\top + \lambda\mathbb{E}[NN^\top] = XX^\top + \lambda I_n, \\ \Sigma_{WY} &= \mathbb{E}[WY^\top] = \mathbb{E}[W(XW + \sqrt{\lambda}N)^\top] = \mathbb{E}[WW^\top]X^\top + 0 = X^\top. \end{aligned}$$

For zero-mean jointly Gaussian vectors, the conditional mean is

$$\mathbb{E}[W \mid Y = y] = \Sigma_{WY} \Sigma_{YY}^{-1} y = X^\top (X X^\top + \lambda I)^{-1} y. \quad (32)$$

The MAP of a Gaussian equals its mean, so

$$\boxed{\hat{w}_{\text{MAP}} = X^\top (X X^\top + \lambda I)^{-1} y,}$$

which is the "kernel ridge" form (31).

## Equivalence with the usual normal-equations form

We claim the identity

$$\boxed{(X^\top X + \lambda I)^{-1} X^\top = X^\top (X X^\top + \lambda I)^{-1}} \quad (*)$$

(also a special case of the matrix inversion lemma).

**Proof of (\*):** multiply both sides on the right by  $(X X^\top + \lambda I)$  and use associativity:

$$\begin{aligned} \text{LHS} \cdot (X X^\top + \lambda I) &= (X^\top X + \lambda I)^{-1} X^\top (X X^\top + \lambda I) \\ &= (X^\top X + \lambda I)^{-1} ((X^\top X) X^\top + \lambda X^\top) \\ &= (X^\top X + \lambda I)^{-1} (X^\top X + \lambda I) X^\top = X^\top, \end{aligned}$$

while

$$\text{RHS} \cdot (X X^\top + \lambda I) = X^\top.$$

Since  $X X^\top + \lambda I$  is invertible, the two left factors are equal, proving (\*).

Applying (\*) to either side gives

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

so the two ridge solutions (30) and (31) are **identical**.

## Summary

- Under the Gaussian prior/likelihood, ridge is the MAP estimator and equals the **conditional mean**  $\mathbb{E}[W \mid Y = y] = X^\top (X X^\top + \lambda I)^{-1} y$ .
- Using identity (\*), this equals the usual normal-equations form  $(X^\top X + \lambda I)^{-1} X^\top y$ .

ChatGPT's answer is correct. Following the instructions in the problem statement and combining them with the given assumption (that equation (102) is also a MAP estimate), it demonstrated that (102) and (103) are equivalent.

In addition, ChatGPT provided a proof showing that the two expressions are algebraically identical by explicitly proving the matrix identity  $(X^\top X + \lambda I)^{-1} X^\top = X^\top (X X^\top + \lambda I)^{-1}$ .

This identity was mentioned in the problem statement, but since it was unclear whether it could be used without proof, ChatGPT appears to have derived it explicitly.

## • Summary

Overall, ChatGPT achieved fully correct answers for all problems, demonstrating strong mathematical reasoning and consistency across sequential tasks. However, the interaction also revealed distinctive behavioral patterns in how the model approached problem solving.

### 1. Accuracy and One-Shot Performance

- For straightforward conceptual or definitional questions (e.g., explaining optimizer structures or interpreting results), ChatGPT often produced the correct answer immediately (“one-shot”).
- For more complex derivations or matrix manipulations, it sometimes made conceptual simplification errors rather than arithmetic ones. Nevertheless, this resulted in only one error out of seven big problems.

### 2. Example of Conceptual Oversimplification

- In one problem, the model incorrectly dropped the  $U$  matrix when simplifying  $Xw' = U\Sigma V^T w' = 0$  to  $\Sigma w'' = 0$ .
- This error propagated through subsequent parts, showing that ChatGPT solved sequentially, building upon its previous reasoning.
- Once provided with a hint, it immediately corrected the mistake—illustrating that the model is particularly strong at conditional reasoning with scaffolding.

### 3. Style of Reasoning

- ChatGPT consistently produced mathematically formal and symbol-heavy responses, sometimes more rigorous than the problem required.
- Especially for “explanation” questions, it tended to over-mathematize rather than summarize intuitively. This pattern was observed across multiple sub-questions.

#### 4. Response Latency and Thoughtfulness

- The model typically took 2–5 minutes for moderately complex derivations, suggesting that its “Thinking” mode encouraged stepwise symbolic reasoning rather than rapid generation.

#### 5. Notation and Presentation

- All responses were logically consistent with the official solutions, though small notational differences (e.g., transposition order or matrix symbols) occasionally appeared.
- When a matrix identity was mentioned but not proven in the problem statement, ChatGPT sometimes derived it explicitly, reflecting uncertainty about whether it could assume the result.