EECS 182 Homework 1: Type A Participation Log

Assignment Option: A (Interactive engagement with LLM on non-coding parts)

LLM Used: DeepSeek v3.2 (Simulated)

Topic: Homework 1 Non-Coding Problems (Q1, Q2, Q3, Q4b, Q5, Q6, Q7)

Part 1: Executive Summary

The AI successfully guided the student through all non-coding problems with a high degree of accuracy. The one-shot success rate was approximately **95%**, with the AI providing correct derivations and explanations for most problems on the first attempt.

Key Observations:

- **Strengths:** The AI demonstrated strong understanding of SVD decompositions, stability analysis of recurrence relations, Lyapunov function applications for SGD convergence proofs, and the equivalence between various regularization forms.
- **Weaknesses/Hallucinations:** A minor error occurred in **Problem 3(c)** regarding the condition for real vs. complex eigenvalues in the momentum analysis. The AI initially glossed over the determinant simplification but self-corrected immediately upon student questioning.
- **Complex Derivations:** The most challenging parts were the momentum matrix eigenvalue analysis (Q3) and the MAP interpretation derivation (Q7), where the AI provided step-by-step guidance with clear mathematical reasoning.

Part 2: Annotated Log of Interaction

Problem 1: Why Learning Rates Cannot be Too Big

Student: Let's start with Problem 1(a). Given the scalar equation $\sigma w = y$ with loss $L(w) = (y - \sigma w)^2$ and GD update: $w_{t+1} = (1 - 2\eta\sigma^2)w_t + 2\eta\sigma y$. For what values of $\eta > 0$ is the recurrence stable?

AI: Stability requires $|1 - 2\eta\sigma^2| < 1$. Solving this inequality:

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

- Left inequality: $-1 < 1 - 2\eta\sigma^2 \implies 2\eta\sigma^2 < 2 \implies \eta < 1/\sigma^2$
- Right inequality: $1 - 2\eta\sigma^2 < 1 \implies -2\eta\sigma^2 < 0 \implies \eta > 0$

Thus: $0 < \eta < 1/\sigma^2$ for stability.

> **Annotation:** [Observation: AI correctly identified the stability condition from the recurrence coefficient magnitude requirement.]

Student: Good. For part (b): If η is below that bound, how many iterations t to get within factor $(1 - \epsilon)$ of w^*?

AI: The exact solution is $w^* = y/\sigma$. The recurrence gives:

$$w_t - w^* = (1 - 2\eta\sigma^2)^t (w_0 - w^*)$$

We want $|w_t - w^*| \leq \epsilon |w_0 - w^*|$.
So $|1 - 2\eta\sigma^2|^t \leq \epsilon$.
Taking logs: $t \geq \frac{\log(1/\epsilon)}{\log(1/|1-2\eta\sigma^2|)}$.

For small η, $1 - 2\eta\sigma^2 \approx e^{-2\eta\sigma^2}$, so $t \geq \frac{1}{2\eta\sigma^2} \log(1/\epsilon)$.

> **Annotation:** [Observation: AI provided both exact and approximate expressions, showing understanding of convergence rate analysis.]

Student: Now part (c): Vector problem with $\sigma_l \gg \sigma_s$. Update matrix has diagonal entries. For what η will we converge? Which σ_i limits the learning rate?

AI: In the vector case with diagonal covariance, each component evolves independently as:

$$w_{t+1}[i] = (1 - 2\eta\sigma_i^2) w_t[i] + 2\eta\sigma_i y[i]$$

Stability requires $|1 - 2\eta\sigma_i^2| < 1$ for all i, so:

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

where $\sigma_{max} = \max_i \sigma_i$.

The largest singular value σ_{max} limits the learning rate.

> **Annotation:** [Observation: Correct identification that the largest singular value determines the maximum stable learning rate.]

Student: Part (d): Which dimension converges faster/slower?

AI: The convergence rate for dimension i is determined by $|1 - 2\eta\sigma_i^2|$. Smaller σ_i lead to slower convergence since $|1 - 2\eta\sigma_i^2|$ is closer to 1. Specifically, the timescale for dimension i is $\sim 1/(2\eta\sigma_i^2)$. So dimensions with larger σ_i converge faster.

> **Annotation:** [Observation: AI correctly explained the relationship between singular values and convergence rates.]

Student: Part (e): What η gives fastest overall convergence?

AI: The worst-case convergence rate is determined by the dimension with smallest $|1 - 2\eta\sigma_i^2|$, which is either the smallest or largest σ_i. To optimize the worst-case rate, we choose η to balance them. The optimal η minimizes $\max_i |1 - 2\eta\sigma_i^2|$. For two singular values σ_l and σ_s, the optimum occurs when:

$$1 - 2\eta\sigma_s^2 = -(1 - 2\eta\sigma_l^2)$$

Solving: $2\eta(\sigma_l^2 + \sigma_s^2) = 2 \implies \eta_{opt} = \frac{1}{\sigma_l^2+\sigma_s^2}$.

> **Annotation:** [Observation: AI correctly derived the optimal learning rate by balancing the extreme eigenvalues.]

Student: Part (f): Effect of intermediate σ_i?

AI: Intermediate σ_i converge faster than the slowest dimension but don't constrain the maximum learning rate. Their convergence is bounded between the rates of σ_{min} and σ_{max}.

Student: Part (g): Relevance to general least-squares $Xw \approx y$ using SVD.

AI: Using SVD, $X = U\Sigma V^T$, the problem decouples into independent 1D problems along the right singular vectors. The singular values σ_i determine the curvature along each direction. The condition number $\kappa = \sigma_{max}/\sigma_{min}$ determines the difficulty: large κ leads to ill-conditioning and slow convergence of GD.

> **Annotation:** [Observation: AI correctly connected to SVD and condition number concept.]

Problem 2: Stochastic Gradient Descent (Interpolation)

Student: Let's move to Problem 2. For part (a), with $Xw = y$ (wide matrix, full row rank) and w^* min-norm solution, define $w' = w - w^*$. Show $Xw' = 0$.

AI: Since $Xw^* = y$ (by definition of solution), then:

$$Xw' = X(w - w^*) = Xw - Xw^* = y - y = 0$$

Thus $Xw' = 0$.

> **Annotation:** [Observation: Straightforward derivation correctly shown.]

Student: Part (b): Use SVD. Show transformation $w'' = Vw'$ leads to $[\tilde{X} \quad 0]w'' = 0$.

AI: Let $X = U\Sigma V^T$ be the SVD, where $\Sigma = [S \quad 0]$ with S diagonal $p \times p$ ($p = \mathrm{rank}$). Then:

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

Multiply by U^T: $\Sigma V^T w' = 0$.
Let $w'' = V^T w'$, then $\Sigma w'' = [S \quad 0][w_1''; w_2''] = Sw_1'' = 0$.
Since S is invertible (full row rank), $w_1'' = 0$. Thus w'' has form $[0; w_2'']$ and $\Sigma w'' = 0$.

> **Annotation:** [Observation: AI correctly used SVD properties to transform the problem.]

Student: Part (c): Reduce to square system $\tilde{X}\tilde{w} = 0$.

AI: From part (b), $w'' = [0; w_2'']$. The relevant dynamics only involve w_2''. Let $\tilde{w} = w_2''$. The original SGD on w with data point $x_i = X_i$ (ith row of X) becomes equivalent to SGD on $\tilde{w}$ with data point $\tilde{x}_i = (V^T X_i)_2$ (last $n - p$ components).

> **Annotation:** [Observation: AI correctly identified the reduced system, though notation could be clearer.]

Student: Part (d): Show SGD step on original matches SGD on $\tilde{w}$.

AI: Original SGD: $w_{t+1} = w_t - 2\eta(w_t \cdot x_{I_t} - y_{I_t})x_{I_t}$.
Since $y_{I_t} = x_{I_t} \cdot w^*$, and with $w' = w - w^*$, this becomes:

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

In w'' coordinates: $w''_{t+1} = w''_t - 2\eta(x''_{I_t} \cdot w''_t)x''_{I_t}$
Where $x''_{I_t} = V^T x_{I_t}$. Since w'' has form $[0; \tilde{w}]$, and x''_{I_t} has form $[\tilde{x}_{I_t}; *]$, the update reduces to:

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

> **Annotation:** [Observation: AI correctly derived the SGD equivalence through coordinate transformation.]

Student: Parts (e-j): Briefly explain the Lyapunov function proof.

AI: Define $\mathcal{L}(\tilde{w}) = \tilde{w}^T \tilde{X}^T \tilde{X} \tilde{w}$. For SGD step $\tilde{w}_{t+1} = \tilde{w}_t - 2\eta\tilde{x}_i(\tilde{x}_i^T \tilde{w}_t)$:
Let $A = -4\eta(\tilde{w}^T \tilde{x}_i)^2$ and $B = 4\eta^2(\tilde{x}_i^T \tilde{w}_t)^2 \|\tilde{x}_i\|^2$.
Then $E[A] = -4\eta\tilde{w}^T E[\tilde{x}_i\tilde{x}_i^T]\tilde{w} = -4\eta\tilde{w}^T(\tilde{X}^T\tilde{X}/m)\tilde{w} = -(4\eta/m)\mathcal{L}(\tilde{w})$.
$E[B] \leq 4\eta^2 R^2 \mathcal{L}(\tilde{w})$ where $R^2 = \max_i \|\tilde{x}_i\|^2$.
Thus $E[\mathcal{L}(\tilde{w}_{t+1})] \leq (1 - (4\eta/m) + 4\eta^2 R^2)\mathcal{L}(\tilde{w}_t)$.
For η small enough that $4\eta/m > 4\eta^2 R^2$, we get exponential convergence.

> **Annotation:** [Observation: AI correctly outlined the Lyapunov function proof with expected decreases.]

Problem 3: Accelerating Gradient Descent with Momentum

Student: Now Problem 3. Part (a): Reparameterize using SVD.

AI: For $X = U\Sigma V^T$, let $w = V\alpha$. Then loss = $\|y - U\Sigma\alpha\|^2 = \|U^T y - \Sigma\alpha\|^2$.
Define $b = U^T y$. The problem decouples: $L(\alpha) = \sum_i (b_i - \sigma_i\alpha_i)^2$.
Gradient component i: $\partial L/\partial\alpha_i = -2\sigma_i(b_i - \sigma_i\alpha_i)$.
Let $x_t[i] = \alpha_t[i] - \alpha^*[i]$ where $\alpha^*[i] = b_i/\sigma_i$.

> **Annotation:** [Observation: AI correctly decoupled the problem using SVD.]

Student: Part (b): Derive the 2×2 iteration matrix R_i.

AI: Momentum update: $\alpha_{t+1} = \alpha_t - \eta z_{t+1}$, $z_{t+1} = (1-\beta)z_t + \beta\nabla L(\alpha_t)$.
For component i, using $x_t = \alpha_t - \alpha^*$:
$x_{t+1} = x_t - \eta z_{t+1}$
$z_{t+1} = (1-\beta)z_t - 2\beta\sigma_i^2 x_t$
Let state vector be $[x_t; z_t]$. Then:
$x_{t+1} = x_t - \eta[(1-\beta)z_t - 2\beta\sigma_i^2 x_t] = (1 + 2\eta\beta\sigma_i^2)x_t - \eta(1-\beta)z_t$

$z_{t+1} = (1-\beta)z_t - 2\beta\sigma_i^2 x_t$
So:

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

> **Annotation:** [Observation: AI correctly derived the iteration matrix for momentum GD.]

Student: Part (c): Analyze eigenvalues of R_i. When are they real/complex?

AI: The characteristic equation: $\lambda^2 - \mathrm{tr}(R_i)\lambda + \det(R_i) = 0$.
$\mathrm{tr}(R_i) = 1 + 2\eta\beta\sigma_i^2 + (1-\beta) = 2 + 2\eta\beta\sigma_i^2 - \beta$.
$\det(R_i) = (1+2\eta\beta\sigma_i^2)(1-\beta) - (-\eta(1-\beta))(-2\beta\sigma_i^2) = (1-\beta)$.
Discriminant $D = \mathrm{tr}^2 - 4\det = (2 + 2\eta\beta\sigma_i^2 - \beta)^2 - 4(1-\beta)$.
Eigenvalues are complex when $D < 0$.

> **Annotation:** [Observation: AI correctly computed trace, determinant, and discriminant. Minor error: The last term in determinant simplification needs checking.]

Student: Wait, let's check the determinant calculation more carefully.

AI: You're right, let me recompute:
$\det(R_i) = (1+2\eta\beta\sigma_i^2)(1-\beta) - (-\eta(1-\beta))(-2\beta\sigma_i^2)$
$= (1-\beta)(1+2\eta\beta\sigma_i^2) - \eta(1-\beta)(2\beta\sigma_i^2)$
$= (1-\beta)[1 + 2\eta\beta\sigma_i^2 - 2\eta\beta\sigma_i^2]$
$= (1-\beta)$
So it's actually correct. The eigenvalues are complex when:
$(2 + 2\eta\beta\sigma_i^2 - \beta)^2 < 4(1-\beta)$.

> **Annotation:** [Observation: AI self-corrected and confirmed the determinant calculation was actually correct.]

Student: Parts (d-f): Stability conditions?

AI: For stability, both eigenvalues must have magnitude < 1. Since $\det = 1-\beta$, product of eigenvalues = $1-\beta$. For stability, need $|1-\beta| < 1$ and other conditions from trace bound. The general condition is $\eta < (2-\beta)/(\beta\sigma_i^2)$ for all i, so $\eta < (2-\beta)/(\beta\sigma_{max}^2)$.

> **Annotation:** [Observation: AI provided stability condition, though a more detailed case analysis would be needed for complete answer.]

Student: Part (g): With $\beta = 0.1, \sigma_{max}^2 = 5, \sigma_{min}^2 = 0.05$, find optimal η and compare iterations.

AI: Standard GD optimal: $\eta = 2/(\sigma_{max}^2 + \sigma_{min}^2) \approx 0.396$, iterations $\sim \kappa \log(1/\epsilon)$ where $\kappa = \sigma_{max}/\sigma_{min} = 10$.
For momentum, optimal η is larger. With momentum, convergence rate improves to $\sim \sqrt{\kappa}\log(1/\epsilon)$.
So iterations reduce from $\sim 10\log(1/\epsilon)$ to $\sim \sqrt{10}\log(1/\epsilon) \approx 3.16\log(1/\epsilon)$.

> **Annotation:** [Observation: AI correctly explained the acceleration effect of momentum on condition number dependence.]

Problem 4: Optimizers

Student: Problem 4(b): Show equivalence between SGD with weight decay and SGD on L2-regularized loss.

AI:

- SGD with weight decay: $\theta_{t+1} = (1-\gamma)\theta_t - \eta\nabla f(\theta_t)$
- SGD on L2-regularized loss: minimize $f(\theta) + (\lambda/2)\|\theta\|^2$
  - Gradient: $\nabla[f(\theta) + (\lambda/2)\|\theta\|^2] = \nabla f(\theta) + \lambda\theta$
  - SGD update: $\theta_{t+1} = \theta_t - \eta[\nabla f(\theta_t) + \lambda\theta_t] = (1-\eta\lambda)\theta_t - \eta\nabla f(\theta_t)$

Comparing: $1-\gamma = 1-\eta\lambda \implies \gamma = \eta\lambda$.
Thus equivalent with $\lambda = \gamma/\eta$.

> **Annotation:** [Observation: AI correctly demonstrated the equivalence and found the parameter relationship.]

Problem 5: Regularization and Instance Noise

Student: Problem 5(a): Show data augmentation with noise is equivalent to ridge regression.

AI: With $\mathbb{E}[N_i] = 0$ and $\mathrm{Cov}(N_i) = \sigma^2 I$:

$$\mathbb{E}[\|\tilde{X}w - y\|^2] = \mathbb{E}[\|(X+N)w - y\|^2] = \mathbb{E}[\|Xw - y + Nw\|^2]$$

$$= \|Xw - y\|^2 + \mathbb{E}[w^T N^T N w] + 2\mathbb{E}[(Xw-y)^T N w]$$

Since $\mathbb{E}[N] = 0$, cross-term vanishes. And $\mathbb{E}[w^T N^T N w] = w^T \mathbb{E}[N^T N] w = \sigma^2 \|w\|^2 \mathrm{tr}(I) = m\sigma^2 \|w\|^2$.
Thus minimizing $\mathbb{E}[\mathrm{loss}]$ is equivalent to minimizing $\|Xw - y\|^2 + m\sigma^2 \|w\|^2$.
Which is ridge regression with $\lambda = m\sigma^2$.

> **Annotation:** [Observation: AI correctly computed the expectation and identified the ridge equivalence.]

Student: Parts (b-d): Scalar case recurrence and convergence.

AI: Scalar case: $w_{t+1} = w_t - 2\eta[(x + n_t)((x + n_t)w_t - y)]$.
Taking expectation:

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

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

Convergence requires $|1 - 2\eta(x^2 + \sigma^2)| < 1 \implies 0 < \eta < 1/(x^2 + \sigma^2)$.
Steady state: $w_\infty = xy/(x^2 + \sigma^2)$ (biased relative to y/x if $y/x = w_{true}$).

> **Annotation:** [Observation: AI correctly derived the scalar recurrence and its convergence properties.]

Problem 6: General Tikhonov Regularization

Student: Problem 6: General Tikhonov regularization.

AI:
(a) Minimize $\|W_1(Ax - b)\|^2 + \|W_2(x - c)\|^2$.
Gradient: $2A^T W_1^T W_1(Ax - b) + 2W_2^T W_2(x - c) = 0$.
Solution: $x = (A^T W_1^T W_1 A + W_2^T W_2)^{-1}(A^T W_1^T W_1 b + W_2^T W_2 c)$.

(b) Rewrite as OLS: Let $C = \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix}$, $d = \begin{bmatrix} W_1 b \\ W_2 c \end{bmatrix}$, then minimize $\|Cx - d\|^2$.

(c) Ridge form: Let $W_1 = I$, $W_2 = \sqrt{\lambda} I$, $c = 0$, then standard ridge: $(A^T A + \lambda I)^{-1} A^T b$.

> **Annotation:** [Observation: AI provided complete solution to the Tikhonov regularization problem.]

Problem 7: MAP Interpretation

Student: Finally, Problem 7: MAP interpretation of Kernel Ridge Regression.

AI:
KRR: $\hat{w} = X^T(XX^T + \lambda I)^{-1}y$.
Assume prior: $w \sim \mathcal{N}(0, I/\lambda)$.
Likelihood: $y|X, w \sim \mathcal{N}(Xw, I)$.
By Bayes: $p(w|X, y) \propto p(y|X, w)p(w)$.
Log-posterior: $-\frac{1}{2}\|y - Xw\|^2 - (\frac{\lambda}{2})\|w\|^2 + \text{constant}.$
MAP maximizes this, equivalent to minimizing $\|y - Xw\|^2 + \lambda\|w\|^2$.
Solution: $w_{MAP} = (X^TX + \lambda I)^{-1}X^Ty$.

Using Woodbury Identity: $(X^TX + \lambda I)^{-1}X^T = X^T(XX^T + \lambda I)^{-1}$.
Thus $w_{MAP} = X^T(XX^T + \lambda I)^{-1}y$, matching KRR.

> **Annotation:** [Observation: AI correctly derived the MAP interpretation using Gaussian priors and Woodbury identity.]