Midterm 2 Practice — Solutions

M 375T Mathematical Foundations of Machine Learning — Fall 2026

Part A — PageRank and walks on directed graphs

A1. (a) Arrows $0 \to 1$, $0 \to 2$, $1 \to 2$, $2 \to 0$, $2 \to 3$. $d = (2, 1, 2, 0)$ (out-degrees). Vertex 3 is dangling: it has no outgoing link, $d_3 = 0$, and W_{3j}/d_3 is 0/0. (b) $P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$. The code starts from the uniform matrix and overwrites only the rows with $d > 0$, so the dangling row stays uniform: from page 3 the surfer jumps to a uniformly chosen page. Line 1: `[1. 1. 1. 1.]`. (c) $\mu_1 = (0, \frac{1}{2}, \frac{1}{2}, 0)$; $\mu_2 = \frac{1}{2}(\text{row } 1) + \frac{1}{2}(\text{row } 2) = (\frac{1}{4}, 0, \frac{1}{2}, \frac{1}{4})$. Line 2: `[0.25 0. 0.5 0.25]`. (d) No: $\pi \propto d$ needs W symmetric; here $d_3 = 0$ would give page 3 no mass although page 2 sends mass to it. Vertex 2: $(P^\mathsf{T}\pi)_2 = \frac{1}{2}\pi_0 + \pi_1 + \frac{1}{4}\pi_3 = \frac{1}{17}(2 + 3 + 1) = \frac{6}{17} = \pi_2$. Page 2 receives links from pages 0 and 1, and page 1 gives it all of its mass: rank comes from in-links, weighted by the rank of the pages that link.

A2. (a) With probability α the surfer follows a link (a step of P); with probability $1 - \alpha$ it jumps to a uniformly chosen page. $\pi_j = \alpha \sum_i \pi_i P_{ij} + \frac{1-\alpha}{n}$. (b) $G_{1/2} = \begin{pmatrix} \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$. Every column sums to 1, so $(G^\mathsf{T}\frac{1}{3}\mathbf{1})_j = \frac{1}{3}\sum_i G_{ij} = \frac{1}{3}$. (c) 1, $\alpha e^{2\pi i/3}$, $\alpha e^{-2\pi i/3}$: the eigenvalue 1 stays, every other eigenvalue is multiplied by α. Output `[0.5 0.5 1.]`. (d) $2 \cdot 2^{-t} < 0.01 \iff 2^t > 200$: $t = 8$ ($2^7 = 128$, $2^8 = 256$). For $\alpha = 0.85$: $t > \log 200 / \log(1/0.85) \approx 32.6$, so $t = 33$.

A3. (a) $P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$. $\pi_2 = \frac{1}{2} \cdot 0 + \frac{1}{6} = \frac{1}{6}$. $\pi_1 = \frac{1}{2}\pi_0 + \frac{1}{6}$, $\pi_0 = \frac{1}{2}(\pi_1 + \pi_2) + \frac{1}{6} = \frac{1}{2}\pi_1 + \frac{1}{4}$. Substituting: $\pi_0 = \frac{1}{4}\pi_0 + \frac{1}{3}$, so $\pi = (\frac{4}{9}, \frac{7}{18}, \frac{1}{6})$ (sum 1). (b) $v = e_2$: $\pi_2 = \frac{1}{2}$, $\pi_1 = \frac{1}{2}\pi_0$, $\pi_0 = \frac{1}{2}(\pi_1 + \pi_2) = \frac{1}{4}\pi_0 + \frac{1}{4}$: $\pi = (\frac{1}{3}, \frac{1}{6}, \frac{1}{2})$. $v = e_0$: $\pi_2 = 0$, $\pi_1 = \frac{1}{2}\pi_0$, $\pi_0 = \frac{1}{2}\pi_1 + \frac{1}{2}$: $\pi = (\frac{2}{3}, \frac{1}{3}, 0)$. Page 2 gets zero: no page links to it and no restart lands on it. Its rank is exactly the floor $(1 - \alpha)v_2$. (c) $v = e_1$: $\pi_2 = 0$, $\pi_0 = \frac{1}{2}\pi_1$, $\pi_1 = \frac{1}{2}\pi_0 + \frac{1}{2}$: $\pi = (\frac{1}{3}, \frac{2}{3}, 0)$. Average: $\frac{1}{3}(\frac{2}{3} + \frac{1}{3} + \frac{1}{3},\ \frac{1}{3} + \frac{2}{3} + \frac{1}{6},\ 0 + 0 + \frac{1}{2}) = (\frac{4}{9}, \frac{7}{18}, \frac{1}{6})$. The equation $\pi = \alpha P^\mathsf{T}\pi + (1 - \alpha)v$ is linear in v, so $\pi_v = \sum_j v_j \pi_{e_j}$.

Part B — Sampling a given measure

B1. (a) `(10000, 400)`. One row is one path: the positions of one walker after $1, 2, \ldots, 400$ steps of size $1/20$. The last column is the position of every walker at time $t = 1$. (b) The position at time $t = k/n$ is a sum of k independent steps of variance $1/n$: variance $k/n = t$. Line 2 ≈ 1.00. Column 99 is after 100 steps, $t = 0.25$: line 3 ≈ 0.25. (c) By the central limit theorem the position at time 1 is approximately $N(0, 1)$, and $\mathbb{P}(|Z| > 1.96) = 0.05$: line 4 ≈ 0.05. (d) $B_3 = B_1 + (B_3 - B_1)$ with $B_3 - B_1 \sim N(0, 2)$ independent of B_1: $B_3 \sim N(2, 2)$.

B2. (a) Invariance: $\sum_i \pi_i P_{ij} = \pi_j$ for every j (that is, $P^\mathsf{T}\pi = \pi$): the mass arriving at a vertex equals the mass sitting there, hence the mass leaving. Detailed balance: $\pi_i P_{ij} = \pi_j P_{ji}$ for every pair: on each edge the traffic in the two directions is equal. (b) Sum $\pi_i P_{ij} = \pi_j P_{ji}$ over i: $\sum_i \pi_i P_{ij} = \pi_j \sum_i P_{ji} = \pi_j$, because rows of P sum to 1. (c) $d = (2,3,3)$, $\pi = (2,3,3)/8$. $\pi_i P_{ij} = \frac{d_i}{8}\frac{W_{ij}}{d_i} = \frac{W_{ij}}{8}$, symmetric: edge 0–1: $\frac{1}{8}$ both ways; edge 1–2: $\frac{2}{8}$; edge 0–2: $\frac{1}{8}$. (d) Invariant: every column of R sums to 1, so $\frac{1}{3}\sum_i R_{ij} = \frac{1}{3}$. Not in detailed balance: $\pi_0 R_{01} = \frac{1}{4}$ but $\pi_1 R_{10} = \frac{1}{12}$. Mass $\frac{1}{4}$ moves $0 \to 1$ and $\frac{1}{12}$ moves $1 \to 0$: a net circulation of $\frac{1}{6}$ around the triangle. Invariance does not imply detailed balance.

B3. (a) $P_{01} = \frac{1}{2}\min(1, \frac{0.3}{0.2}) = \frac{1}{2}$; $P_{10} = \frac{1}{2}\cdot\frac{2}{3} = \frac{1}{3}$; $P_{12} = \frac{1}{2}\cdot 1 = \frac{1}{2}$; $P_{21} = \frac{1}{2}\cdot\frac{3}{5} = \frac{3}{10}$; the diagonal is what is left: $P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ 0 & \frac{3}{10} & \frac{7}{10} \end{pmatrix}$. (b) $j=0$: $0.2\cdot\frac{1}{2} + 0.3\cdot\frac{1}{3} = 0.2$. $j=1$: $0.2\cdot\frac{1}{2} + 0.3\cdot\frac{1}{6} + 0.5\cdot\frac{3}{10} = 0.1 + 0.05 + 0.15 = 0.3$. $j=2$: $0.3\cdot\frac{1}{2} + 0.5\cdot\frac{7}{10} = 0.5$. Pair $(1,2)$: $0.3\cdot\frac{1}{2} = 0.15 = 0.5\cdot\frac{3}{10}$. (c) The ratio is $\pi_0/\pi_1 = \frac{2}{3}$. $U = 0.8 > \frac{2}{3}$: refused, the walker stays at 1. $U = 0.5 < \frac{2}{3}$: accepted, the walker is at 0. (d) They approach $\pi = (0.2, 0.3, 0.5)$. Only ratios are needed, so π may be known up to a constant factor (for example e^{-E} without its normalizing sum).

Part C — Vectors, orthonormal bases, and projection

C1. (a) $\|u\|^2 = 9$, $\|v\|^2 = 9$, $u\cdot v = 2 - 4 + 2 = 0$, $\|u+v\|^2 = \|(3,0,3)\|^2 = 18$. In general $\|u+v\|^2 = \|u\|^2 + 2\,u\cdot v + \|v\|^2$; the energies add exactly when $u\cdot v = 0$. $\cos(u,w) = \frac{2+2+4}{3\cdot 3} = \frac{8}{9}$. (b) $\sqrt{2}\,c = (5+1, 5-1)$: line 1 `[6.  4.]`. Line 2: $c\cdot c = \frac{36+16}{2} = 26 = \|w\|^2$ (Parseval: an orthonormal change of coordinates keeps the energy). Line 3: `[5.  1.]`: $Q^\mathsf{T}c = Q^\mathsf{T}Qw = w$, the coordinates rebuild the vector.

C2. (a) $c_0 = \frac{20}{\sqrt{8}} = 5\sqrt{2}$, $c_1 = \frac{16-4}{\sqrt{8}} = 3\sqrt{2}$, $c_2 = \frac{12-4}{2} = 4$, $c_3 = 0$, $c_4 = \cdots = c_7 = 0$ (x is constant on each pair). Parseval: $50 + 18 + 16 = 84 = 36 + 36 + 4 + 4 + 1 + 1 + 1 + 1$. (b) Span of q_0: $c_0 q_0 = (2.5, \ldots, 2.5)$, the average; error energy $c_1^2 + c_2^2 = 34$. Span of q_0, q_1: $(4,4,4,4,1,1,1,1)$, the averages of the two halves; error energy $c_2^2 = 16$. Span of $q_0, \ldots, q_3$: x itself; error 0. The error energy is the sum of the squares of the dropped coordinates. (c) $\mathbb{E}[(q\cdot\xi)^2] = \sum_{j,k} q_j q_k \mathbb{E}[\xi_j\xi_k] = \sigma^2 \sum_j q_j^2 = \sigma^2$ (independence and mean zero remove the terms $j \neq k$). Four directions are kept: $4\sigma^2$ survives, half of the total $8\sigma^2$.

C3. (a) $E(c) = \|x\|^2 - 2\sum_k c_k\, q_k\cdot x + \sum_{k,l} c_k c_l\, q_k\cdot q_l = \|x\|^2 + \sum_k \left(c_k^2 - 2b_k c_k\right)$ with $b_k = q_k\cdot x$: orthonormality removes the cross terms. $\partial E/\partial c_k = 2c_k - 2b_k = 0$ gives $c_k = b_k$. Each term is an upward parabola in its own variable ($c_k^2 - 2b_k c_k = (c_k - b_k)^2 - b_k^2$), so this is the unique global minimum. (b) $q_1\cdot q_2 = \frac{1}{9}(2+2-4) = 0$, $\|q_1\|^2 = \|q_2\|^2 = \frac{9}{9} = 1$. $b_1 = \frac{15}{3} = 5$, $b_2 = \frac{3}{3} = 1$. $y = 5q_1 + q_2 = \frac{1}{3}(7, 11, 8)$, $r = \frac{1}{3}(2, -2, 1)$. $r\cdot q_1 = \frac{1}{9}(2-4+2) = 0$, $r\cdot q_2 = \frac{1}{9}(4-2-2) = 0$. $\|x\|^2 = 27$, $\|y\|^2 = b_1^2 + b_2^2 = 26$, $\|r\|^2 = 1$. $\cos(x,y) = \|y\|/\|x\| = \sqrt{26/27} \approx 0.981$. (c) $A^\mathsf{T}A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $A^\mathsf{T}x = (1,1)$, so $c = (\frac{1}{3}, \frac{1}{3})$, $y = \frac{1}{3}(1,2,1)$, $r = \frac{1}{3}(2,-2,2)$, and $r\cdot a_1 = r\cdot a_2 = 0$. Code: `c = np.linalg.solve(A.T @ A, A.T @ x)`.

C4. (a) Every row has $\|u\|^2 = \|v\|^2 = N$ exactly (entries ± 1), so $\cos = \frac{u\cdot v}{\|u\|\|v\|} = \frac{u\cdot v}{N}$, and `(U * V).sum(axis=1)` is the row-by-row dot product. (b) Line 1 `(5000,)`. Line 2 ≈ 0.00. Line

$3 \approx 0.05$: $u \cdot v$ is a sum of N independent ± 1 terms, mean 0, variance N, so the cosine has standard deviation $\sqrt{N}/N = 1/\sqrt{N} = 1/20$. (c) $1/\sqrt{784} = 1/28 \approx 0.036$, so 0.5 is 14 standard deviations. Two random directions in $\mathbb{R}^{784}$ are almost orthogonal; images of the same digit are far from random: the data line up with few directions.

Part D — Quadratic energies, diagonalization, and PCA

D1. (a) $\lambda^2 - 9\lambda + 14 = 0$: $\lambda_1 = 7$, $\lambda_2 = 2$. $(S - 7I)q = 0$ gives $-4a + 2b = 0$: $q_1 = (1,2)/\sqrt{5}$; $q_2 = (2,-1)/\sqrt{5}$. (b) $E = 7u^2 + 2v^2$, $u = \frac{a+2b}{\sqrt{5}}$, $v = \frac{2a-b}{\sqrt{5}}$. Check: $\frac{1}{5}\big[7(a^2 + 4ab + 4b^2) + 2(4a^2 - 4ab + b^2)\big] = \frac{1}{5}(15a^2 + 20ab + 30b^2)$. (c) For a unit vector $u^2 + v^2 = 1$, so $E = 2 + 5u^2$: maximum 7 at $x = \pm q_1$, minimum 2 at $x = \pm q_2$. (d) $x^\mathsf{T} S^{-1} x = u^2/7 + v^2/2 = 1$: half-axis $\sqrt{7} \approx 2.65$ along q_1 (major), $\sqrt{2} \approx 1.41$ along q_2. (e) `[2.  7.]` (increasing order), then `[[2. 0.]  [0.  7.]]`: $Q^\mathsf{T} S Q = \Lambda$, with the columns of Q in the same order.

D2. (a) For fixed q each term $\|x_i - c_i q\|^2 = \|x_i\|^2 - 2c_i\, q \cdot x_i + c_i^2$ is a parabola in c_i, minimized at $c_i = q \cdot x_i$, with value $\|x_i\|^2 - (q \cdot x_i)^2$. Averaging, the error is $\frac{1}{4}\sum_i \|x_i\|^2 - \frac{1}{4}\sum_i (q \cdot x_i)^2 = \frac{1}{4}\sum_i \|x_i\|^2 - q^\mathsf{T} S q$. The first term does not depend on q, so the best q maximizes $q^\mathsf{T} S q$ over unit vectors: the top eigenvector of S. (b) $S = \frac{1}{4}X^\mathsf{T} X = \begin{pmatrix} 6.5 & 4.5 \\ 4.5 & 18.5 \end{pmatrix}$. $Sx_1 = (40, 120) = 20x_1$ and $Sx_3 = (15, -5) = 5x_3$: $\lambda_1 = 20$, $q_1 = (1,3)/\sqrt{10}$; $\lambda_2 = 5$, $q_2 = (3,-1)/\sqrt{10}$. Best $q = q_1$; $c = (2\sqrt{10}, -2\sqrt{10}, 0, 0)$; minimal error $\frac{1}{4}(40 + 40 + 10 + 10) - 20 = 5 = \lambda_2$; share $20/25 = 80\%$. (c) $\hat{\Sigma} = S$ (the data are centered). Major axis along $q_1 = (1,3)/\sqrt{10}$, standard deviation $\sqrt{20} \approx 4.47$ (and $\sqrt{5}$ along q_2). (d) Line 1 `[ 5.  20.]`; line 2 `[6.325 6.325 0. 0.]`; line 3 `5.0`; line 4 `[20.  5.]` ($\sigma_k^2/n = \lambda_k$, decreasing order). `np.abs`: an eigenvector is defined up to sign, so `eigh` may return q or $-q$, and c changes sign with it.

Part E — The SVD, best approximation, and the power method

E1. (a) $Y^\mathsf{T} Y = \begin{pmatrix} 6 & 2 \\ 2 & 6 \end{pmatrix}$: eigenvalues $8, 4$ with $v_1 = (1,1)/\sqrt{2}$, $v_2 = (1,-1)/\sqrt{2}$. $\sigma_1 = 2\sqrt{2}$, $\sigma_2 = 2$. $Yv_1 = (2\sqrt{2}, 0, 0)$, so $u_1 = (1,0,0)$; $Yv_2 = (0, \sqrt{2}, \sqrt{2})$, so $u_2 = (0,1,1)/\sqrt{2}$. Orthonormal: $u_1 \cdot u_2 = 0$, both of length 1. (b) $\sigma_1 u_1 v_1^\mathsf{T} = \begin{pmatrix} 2 & 2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$, $\sigma_2 u_2 v_2^\mathsf{T} = \begin{pmatrix} 0 & 0 \\ 1 & -1 \\ 1 & -1 \end{pmatrix}$; their sum is Y. (c) $Y - Y_1$ is the second piece: $1 + 1 + 1 + 1 = 4 = \sigma_2^2$. (d) For every matrix B of rank at most 1, $\|Y - B\|_F^2 \geq \sigma_2^2$, with equality at $B = Y_1 = \sigma_1 u_1 v_1^\mathsf{T}$. No: $3 < 4 = \sigma_2^2$.

E2. (a) U: 100×7; Σ: 7×7; V^T: 7×7; $X^\mathsf{T} X$: 7×7; XX^T: 100×100. $X^\mathsf{T} X = nS$. (b) $\|Xq\|^2 = q^\mathsf{T} X^\mathsf{T} X q = 900$: $\|Xq\| = 30 = \sigma$, and $\lambda = 900/100 = 9$. $XX^\mathsf{T} u = \frac{1}{30} X(X^\mathsf{T} X q) = \frac{900}{30} Xq = 900\, u$: eigenvalue 900. (c) $\sigma_k = \sqrt{n\lambda_k}$: $30, 20, 10, 0, 0, 0, 0$. Best line: error $\lambda_2 + \lambda_3 = 5$. Best plane: $\lambda_3 = 1$. No: by Eckart–Young no 2-dimensional subspace has average error below $\lambda_3 + \cdots + \lambda_7 = 1$.

E3. (a) $x_1 = (5,2)$, $x_2 = (29,14)$. $R(x_0) = 5$. $Sx_1 = (29,14)$: $R(x_1) = \frac{145+28}{29} = \frac{173}{29} \approx 5.966$. $Sx_2 = (173,86)$: $R(x_2) = \frac{29 \cdot 173 + 14 \cdot 86}{841 + 196} = \frac{6221}{1037} \approx 5.999$. (b) $c_1 = q_1 \cdot x_0 = \frac{2}{\sqrt{5}}$, $c_2 = q_2 \cdot x_0 = \frac{1}{\sqrt{5}}$. Tangent: $\frac{1}{2}$, $\frac{1}{12}$, $\frac{1}{72}$: it shrinks by $\lambda_2/\lambda_1 = \frac{1}{6}$ per step. The gap is 1, $\frac{1}{29}$, $\frac{1}{1037}$: it shrinks by about $(\lambda_2/\lambda_1)^2 = \frac{1}{36}$ per step. (c) $Sq_2 = q_2$: the iteration stays on q_2 forever and never finds

q_1. With a perturbation of size δ along q_1, that component is multiplied by 6 at every step relative to the q_2 component, so the iterate turns to q_1 after about $\log(1/\delta)/\log 6$ steps. (d) S is $d \times d$ and costs nd^2 multiplications to form; the step only needs the two products Xx (nd) and $X^\mathsf{T}(Xx)$ (nd). One step: $2nd \approx 1.9 \cdot 10^7$. Forming S: $nd^2 \approx 7.4 \cdot 10^9$, the cost of $d/2 = 392$ steps.

Part F — Nonlinear features, overfitting, and gradient descent

F1. (a) $\Phi = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$, $\Phi^\mathsf{T}\Phi = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$, $\Phi^\mathsf{T} y = (2, 0)$: $c = (\frac{2}{3}, 0)$, $\hat{y} = \frac{2}{3}$. Training error $\frac{1}{3}(\frac{1}{9} + \frac{4}{9} + \frac{1}{9}) = \frac{2}{9}$. (b) Three points, three coefficients: the parabola interpolates. $c = (0, 0, 1)$, $\hat{y} = x^2$, training error 0. (c) The larger model contains the smaller one (set the new coefficient to 0), so its minimum over more candidates cannot be larger. (d) Line: $(\frac{2}{3} - 2)^2 = \frac{16}{9} \approx 1.78$. Parabola: $(4 - 2)^2 = 4$. The model with zero training error is worse on the new point. The test error (data not used for fitting) should choose; this is overfitting.

F2. (a) Suppose all seven are classified correctly. Then $2w_1 + b > 0$ and $-2w_1 + b > 0$; adding, $b > 0$. But $(0, 0)$ needs $\hat{y} = b < 0$. Contradiction. (b) $r^2 = x_1^2 + x_2^2$ is $0, \frac{1}{4}, \frac{1}{4}$ on class -1: $\hat{y} = -1, -\frac{7}{8}, -\frac{7}{8} < 0$. On class $+1$, $r^2 = 4$: $\hat{y} = 1 > 0$. All correct. Boundary $\hat{y} = 0$: the circle $r^2 = 2$, radius $\sqrt{2}$. (c) Linear in the coefficients c (so least squares and the normal equations apply); not linear in the input x.

F3. (a) $E'(c) = 2c$, so $c \leftarrow (1 - 2h)c$. Factors 0.5, -0.5, -1.5: after three steps $c = 0.125$, -0.125, -3.375. Converges iff $|1 - 2h| < 1$, i.e. $0 < h < 1$ ($= 2/a$ with curvature $a = 2$). (b) $c_k \leftarrow (1 - h\lambda_k)c_k$. $h = 0.2$: factors -0.8 and 0.8: output `[0.64 0.64]`. $h = 0.25$: factors -1.25 and 0.75: `[1.5625 0.5625]`; the first coordinate diverges since $0.25 > 2/9$. (c) $h < 2/\lambda_{\max}$. $\lambda_{\max}$ limits the step; $\lambda_{\min}$ sets the speed (its coordinate shrinks by $1 - h\lambda_{\min}$ per step, the slowest). $\nabla E = \frac{2}{n}\Phi^\mathsf{T}(\Phi c - y)$; the curvatures are the eigenvalues of the Hessian $\frac{2}{n}\Phi^\mathsf{T}\Phi$.

F4. (a) With $s = wx + b$ and $r = \hat{y} - y$: $\partial E/\partial c = 2r \tanh(s)$; $\partial E/\partial w = 2r\, c\,(1 - \tanh^2 s)\, x$; $\partial E/\partial b = 2r\, c\,(1 - \tanh^2 s)$. (b) $s = 0$, $\hat{y} = 0$, $r = -1$: $\partial E/\partial c = 0$, $\partial E/\partial w = 2(-1)(1)(1)(2) = -4$, $\partial E/\partial b = -2$. Step: $c = 1$, $w = 0.4$, $b = 0.2$. (c) The equations $\nabla E = 0$ have w, b inside tanh in every term; they are not linear in (w, b), so there is no formula. $\nabla_w E = \frac{2}{n}\sum_i (\hat{y}_i - y_i)(1 - \hat{y}_i^2) x_i$. At $w = 0, b = 0$: $\hat{y}_i = 0$, so $\nabla_w E = -\frac{2}{n}\sum_i y_i x_i$ and the step is $w = \frac{2h}{n}\sum_i y_i x_i$. With n_+, n_- images in the two classes and class means m_+, m_-: $\sum_i y_i x_i = n_+ m_+ - n_- m_-$, essentially the difference of the two class means.

F5. (a) $\nabla E = \big((ab - 1)b,\ (ab - 1)a\big)$. Minima: $E = 0$ on the curve $ab = 1$ (a hyperbola: a whole curve of minima, not a point). (b) Trace $a^2 + b^2$, determinant $a^2 b^2 - a^2 b^2 = 0$: eigenvalues $a^2 + b^2$ and 0. Sharpness: 2 at $(1, 1)$; 4.25 at $(2, \frac{1}{2})$; $9 + \frac{1}{9} \approx 9.11$ at $(3, \frac{1}{3})$. (c) $2/h = 5$: it can stay near $(1, 1)$ and $(2, \frac{1}{2})$, not near $(3, \frac{1}{3})$. Started near $(3, \frac{1}{3})$ the oscillation across the valley grows, and the iterate moves along the curve toward flatter minima until the sharpness is below $2/h$ (or diverges if h is too large). (d) $ab - 1 = 1$, $\nabla E = (1, 2)$, new point $(1.9, 0.8)$. E goes from $\frac{1}{2}$ to $\frac{1}{2}(1.52 - 1)^2 = 0.1352$. (e) Yes: the batch is drawn at random, so the expected batch gradient is the full gradient. Cost $2Bd$ multiplications: $2 \cdot 12\,000 \cdot 785 \approx 1.9 \cdot 10^7$ for $B = n$, $2 \cdot 100 \cdot 785 \approx 1.6 \cdot 10^5$ for $B = 100$ (120 times cheaper; 120 steps for the price of one). The iterate does not settle: it fluctuates in a ball whose variance is proportional to h/B

(radius $\propto \sqrt{h/B}$); a smaller step or a larger batch shrinks it.