IIT Roorkee, Department of Electrical Engineering **Set-A**

EEC-351: Fundamentals of AI/ML
End-Term Examination (Autumn 2026–27)

Timing: 3 hrs | Date: 17 Nov 2026 | Time: 9:15 am – 12:15 pm | Max. Marks: 60

Instructions

- Answer all questions. Each question must begin on a **new page**.
- Provide complete justifications for your answers. Partial/unjustified answers may not receive full credit.
- Mention assumptions & conditions clearly wherever necessary. Natural log throughout.
- **No partial marking:** each part earns either the full marks shown against it or zero.

***All things must pass* — and so will you if you think of one step at a time.**

1. A learning algorithm searches a *finite* hypothesis set $\mathcal{H}$ with $|\mathcal{H}| = M$ and evaluates every $h \in \mathcal{H}$ on the same N i.i.d. examples. Each per-example error is an indicator in $\{0,1\}$, so Hoeffding's inequality and the union bound give $\mathbb{P}\big[\exists h \in \mathcal{H} : |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\big] \le 2Me^{-2\epsilon^2 N}$. ($\epsilon = 0.1$; $\delta = 0.05$.)
   (a) With $N = 1000$, find the largest M for which the bound still certifies that this probability is at most δ. **[1]**
   (b) Let $N = 2000$ and recompute the largest M. By what factor did it grow? In one line, state how the number of hypotheses you can afford to search grows with N, and name the feature of the bound responsible. **[1]**
   (c) Let $\mathcal{H}$ be the *positive rays* on $\mathbb{R}$, $h_a(x) = \text{sign}(x - a)$, one hypothesis for every $a \in \mathbb{R}$. Find $m_{\mathcal{H}}(N)$ of this $\mathcal{H}$. **[1]**

2. Let $\mathcal{H}_0$ be a set of perceptrons in $\mathbb{R}^2$ (no bias term & no data point at origin), $\mathcal{H}_0 = \big\{ h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x}) : \mathbf{w} \in \mathbb{R}^2,\ \mathbf{w} \neq \mathbf{0} \big\}$.
   (a) Take N points at distinct polar angles. **(i)** If no two of them are diametrically opposite, show that $m_{\mathcal{H}_0}(N) = 2N$. **(ii)** Now suppose exactly p pairs of them are diametrically opposite. How many dichotomies does $\mathcal{H}_0$ produce on these N points? **Justify**. **(iii)** Find $d_{\text{vc}}(\mathcal{H}_0)$ and prove your answer. **[2+1+2]**

3. A regression model is evaluated with the *clipped* squared error $\ell(h, \mathbf{x}, y) = \min\{(h(\mathbf{x}) - y)^2, 4\}$. You evaluate on N fresh test examples and report $\bar{L}_N = \frac{1}{N}\sum_{n=1}^{N} \ell_n$. Let $L = \mathbb{E}[\ell]$.
   (a) Use Hoeffding's inequality to derive a bound on $\mathbb{P}\big[|\bar{L}_N - L| > \epsilon\big]$, & find smallest N for $\mathbb{P}\big[|\bar{L}_N - L| > 0.2\big] \le 0.05$. **[2]**
   (b) Suppose you know only that $L \le 1$. Use Markov's inequality to bound $\mathbb{P}[\bar{L}_N > 1.2]$. Name two facts about the ℓ_n that Hoeffding exploits and Markov does not. **[1]**

4. Let $\mathbf{x} \sim P$ on $\mathcal{X}$, let f be the target, and let $y = f(\mathbf{x}) + \varepsilon$ where ε has mean 0 and variance σ^2, independent of $\mathbf{x}$ and across examples. The hypothesis set is the constant functions $\mathcal{H} = \{h(\mathbf{x}) = b :\ b \in \mathbb{R}\}$, and least squares returns $g^{(\mathcal{D})}(\mathbf{x}) = \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$. Let $\mu_f = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ and $v_f = \text{Var}_{\mathbf{x}}[f(\mathbf{x})]$.
   (a) Find $\bar{g}(\mathbf{x})$. Find the bias and variance in terms of μ_f, v_f, σ^2, N, with **justification**. **[2+2]**
   (b) Write $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}]$. Which part does not decrease with N & which does; give $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}]$ at $N = 1$ and its limit as $N \to \infty$, and sketch it against N. Evaluate each term for $f(x) = x$, $x \sim U[-1, 1]$, $\sigma^2 = 0.1$, $N = 5$. **[1+2+1]**

5. Let E be the energy of the correct answer and $\bar{E}$ the energy of the *most offending incorrect* answer. For a margin $m > 0$ and a temperature $\beta > 0$ we can define $L_\beta = \frac{1}{\beta}\ln\Big(1 + \exp\big(\beta\,(E - \bar{E} + m)\big)\Big)$, where $u := E - \bar{E} + m$, and $[u]_+ = \max(0, u)$.
   (a) Show that $L_\beta \ge [u]_+$ for every $\beta > 0$, and that $L_\beta \to [u]_+$ as $\beta \to \infty$. **[2]**
   (b) If $\beta \to 0$ then $L_\beta = \frac{\ln 2}{\beta} + \frac{u}{2} + O(\beta)$. Does the loss stop pushing once the margin is satisfied? Give the reason. **[1]**
   (c) Show that L_β is convex in $(E, \bar{E})$, and verify that $\big(\nabla_{(E,\bar{E})} L_\beta\big)^\top [1, -1]^\top > 0$ wherever $E + m > \bar{E}$. Does it hold only there, or everywhere? Conclude whether L_β is a generalised margin loss. **[3]**
   (d) Take $\beta = 2$, $m = 1$, $E = 0.5$. Compute L_β and $\partial L_\beta / \partial E$ for $\bar{E} = 1.0$ and for $\bar{E} = 3.0$. What does the comparison show? **[1]**

6. Let $L(\mathbf{w}) = \frac{1}{N}\|X\mathbf{w} - \mathbf{y}\|_2^2$ with $X \in \mathbb{R}^{N \times d}$ of full column rank, and let $\mathbf{w}^*$ be its minimiser. Gradient descent runs $\mathbf{w}_{k+1} = \mathbf{w}_k - \eta\,\nabla L(\mathbf{w}_k)$. *Use without proof the MTE result:* for a quadratic with symmetric positive definite Hessian A, the error $\mathbf{e}_k = \mathbf{w}_k - \mathbf{w}^*$ satisfies $\mathbf{e}_{k+1} = (I - \eta A)\mathbf{e}_k$; in the eigenbasis of A each coordinate evolves as $(1 - \eta\lambda_i)^k$, and GD converges from every start iff $0 < \eta < 2/\lambda_{\max}$.
   (a) Find the gradient and the Hessian A of L, and write $\nabla L(\mathbf{w})$ in terms of A and $\mathbf{w} - \mathbf{w}^*$. **[1]**
   (b) Let $\eta = 1/\lambda_{\max}(A)$ and $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$. Show that every eigen-coordinate of the error contracts per step by a factor of at most $1 - 1/\kappa(A)$. Hence show that $k \ge \dfrac{\ln(1/\tau)}{-\ln(1 - 1/\kappa(A))} \approx \kappa(A)\ln(1/\tau)$ iterations suffice to reach $\|\mathbf{e}_k\| \le \tau\|\mathbf{e}_0\|$ from any start, and are needed when $\mathbf{e}_0$ lies along the slowest eigenvector. **[3]**

7. Let $f : \mathbb{R} \to \mathbb{R}$ be twice differentiable with $|f''(w)| \le L$ for every w. You may use the *descent lemma*: $f(v) \le f(u) + f'(u)(v - u) + \frac{L}{2}(v - u)^2$ for all u, v. At a point w write $g = f'(w)$. Now, *Backtracking line search* takes the gradient step $w - \eta g$ and starts from a trial step η_0. It halves the step until the *sufficient-decrease test* $f(w - \eta g) \le f(w) - \frac{\eta}{2}g^2$ is passed. Prove that it always stops, and that the step it accepts satisfies $\eta \ge \min\big(\eta_0,\ 1/(2L)\big)$. **[4]**

8. Four wingmates plot $x =$ nights left before the EEC-351 end-term against $y =$ cups of canteen chai drunk that night, and fit a line through the origin (zero nights left, zero chai: by then it is too late anyway). They fit $y \approx wx$ by least squares on the four points $x = (1, 2, 3, 4)$, $y = (0, 3, 4, 3)$, with $L(w) = \frac{1}{4}\sum_{n=1}^{4}(wx_n - y_n)^2$. The gradient of example n alone is $g_n(w) = 2x_n(wx_n - y_n)$, so $\nabla L(w) = \frac{1}{4}\sum_n g_n(w)$. SGD picks an index I uniformly at random and uses $g_I(w)$, an unbiased estimate of $\nabla L(w)$.
   (a) At $w = 0$, compute the four $g_n(0)$ and $\nabla L(0)$. Compute the variance σ^2 of $g_I(0)$. **[1]**
   (b) A minibatch of B indices is drawn independently with replacement and the B gradients are averaged. Show that the variance of the minibatch gradient is σ^2/B. Find the smallest B for which its standard deviation at $w = 0$ is at most 10% of $|\nabla L(0)|$, and the smallest B for 5%. What law does the ratio of the two values of B show? **[2]**
   (c) The minimiser is $w^* = 1$. Compute the four $g_n(1)$ and show that $\nabla L(1) = 0$ although no $g_n(1)$ is zero. Starting exactly at w^*, one SGD step with constant $\eta = 0.01$ and $B = 1$ moves w by a random amount; find its standard deviation. In one line each: why does constant-step SGD never settle at w^*, and what fixes it? **[2]**

9. In an energy-based model the answers $Y \in \mathcal{Y} = \{0, 1, 2\}$ for an input x have energy $E(W, Y, x) = \frac{1}{2}(Y - Wx)^2$, with a scalar parameter $W \in \mathbb{R}$. With the Gibbs distribution $P(Y \mid x, W) = e^{-\beta E(W,Y,x)} / \sum_{y\in\mathcal{Y}} e^{-\beta E(W,y,x)}$ ($\beta > 0$), the NLL loss on a training pair (x, Y^i) is $L(W) = E(W, Y^i, x) + \frac{1}{\beta}\ln\sum_{y\in\mathcal{Y}} e^{-\beta E(W,y,x)}$.
   (a) Show that $L = -\frac{1}{\beta}\ln P(Y^i \mid x, W)$. Hence $L > 0$. **[1]**
   (b) Take $x = 1$, $Y^i = 2$, $W = 0.6$, $\beta = 1$. Compute the three energies, $P(Y)$, L, the three $\partial L/\partial E(Y)$ and $\partial L/\partial W$. Which wrong answer does the loss push up hardest, i.e. which has the most negative $\partial L/\partial E(Y)$, and why that one? **[2]**
   (c) Keep $x = 1$, $Y^i = 2$, $\beta = 1$ but let W vary. Show that $\partial L/\partial W = \mathbb{E}_P[Y] - 2$ for every W. Deduce that gradient descent on L never stops at a finite W. Compare with the generalised perceptron loss $E(W, Y^i, x) - \min_Y E(W, Y, x)$: for which W is it zero? Also, show that $-\min_Y E \le \frac{1}{\beta}\ln\sum_y e^{-\beta E(y)} \le -\min_Y E + \frac{\ln|\mathcal{Y}|}{\beta}$. **[2]**

10. The mess secy of Vivekanand Hostel knows only the mean of the waiting time $X \ge 0$ (in min.) at the dinner counter, and in exam week also its standard deviation. He wants guarantees on long waits to hold whatever the distribution of X is.
   (a) Let $X \ge 0$, $\mathbb{E}[X] < \infty$ & $a > 0$. Prove $\mathbb{P}(X \ge a) \le \mathbb{E}[X]/a$, and find which distributions make it an equality. **[2]**
   (b) On normal days the mean wait is 1 minute. What is the largest possible probability that a day's wait is at least 2 min, and which distributions achieve it? A waiting time has a density in practice: is this largest value achieved by some X with a density? If not, what is the supremum over such X? Justify. **[2]**
   (c) In exam week the mean is 10 min and the standard deviation 2 min. What is the largest possible $\mathbb{P}(|X - 10| \ge 4)$? Give a distribution of X that achieves it, and for that distribution compute $\mathbb{P}(|X - 10| > 4)$ and $\mathbb{P}(|X - 10| < 4)$. **[2]**
   (d) A random Prof. claims that $\mathbb{P}(|X - \mu| < \varepsilon) \ge 1 - \sigma^2/\varepsilon^2$ is a stronger guarantee than the bound used in (c), because it has a strict inequality. Is he right? **[1]**

11. Inputs are 3-bit strings, $\mathcal{X} = \{0, 1\}^3$, and labels are ± 1. The unknown target f is seen on five points:

$\mathbf{x}$	001	010	100	110	111	000	011	101
$f(\mathbf{x})$	-1	-1	-1	$+1$	$+1$	?	?	?

   (a) How many target functions on $\mathcal{X}$ agree with this data? **[1]**
   (b) A classmate notices that every label above is the *majority vote* of the three bits, and proposes g_{maj}. Show that for *every* hypothesis g that agrees with the data (g_{maj} included), exactly $\binom{3}{k}$ of the targets in (a) agree with g on exactly k of the unseen points, $k = 0, 1, 2, 3$. What does this say about g_{maj} compared with a g that outputs $+1$ on all three unseen points? **[2]**

12. PCA as an energy-based model. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be centred data with $C = \frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i^\top$ and eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d \ge 0$. PCA picks the unit vector $\mathbf{v}$ that maximises the variance of the projections, $\mathbf{v}^\top C\mathbf{v}$. Read PCA as an energy-based model with parameter $\mathbf{v}$ and energy $E(\mathbf{v}, \mathbf{x}) = \left\|\mathbf{x} - (\mathbf{v}^\top\mathbf{x})\,\mathbf{v}\right\|_2^2$ (squared reconstruction error; low energy = compatible). The *energy loss* of an energy-based model is the average energy of the training examples, $L_{\text{energy}}(\mathbf{v}) = \frac{1}{n}\sum_i E(\mathbf{v}, \mathbf{x}_i)$: it pushes the energies of the data down and pulls nothing up.
   (a) Show that $E(\mathbf{v}, \mathbf{x}) = \|\mathbf{x}\|^2 - (\mathbf{v}^\top\mathbf{x})^2$, and sketch what E measures for a point in $\mathbb{R}^2$. **[1]**
   (b) Show that $L_{\text{energy}}(\mathbf{v}) = \operatorname{tr} C - \mathbf{v}^\top C\mathbf{v}$ and prove that PCA *is* energy-loss minimization. Identify the minimizing $\mathbf{v}$, and show that the minimum energy loss is $\lambda_2 + \cdots + \lambda_d$. **[3]**
   (c) Replace $\mathbf{v}$ by $V \in \mathbb{R}^{d\times k}$ with orthonormal columns and $E = \|\mathbf{x} - VV^\top\mathbf{x}\|^2$. What is L_{energy} at $k = d$? Explain why this is a collapse, why $k < d$ avoids it, and what plays the role of the missing contrastive term. **[2]**

Help me help you! (And no, today is not **With a Little Help from My Friends**.)