Sample final questions

May 5, 2026

1. (170 points) Suppose we want to deploy an ensemble of d small classifiers on an edge device. Each classifier i gets a weight $x_i \in \mathbb{R}$, and the ensemble's prediction on an input is a weighted combination of the individual predictions; collect the weights into the vector $x \in \mathbb{R}^d$.

   To tune x, we use a held-out validation set of m examples. Let $A \in \mathbb{R}^{m\times d}$ collect the validation predictions (column i holds classifier i's predictions on the m examples), and let $b \in \mathbb{R}^m$ be the true labels. Two side-vectors are given for each classifier: a latency vector $\ell \in \mathbb{R}^d$ (how slow each classifier is) and a disparity score $g \in \mathbb{R}^d$ (how much each classifier widens performance gaps across user subgroups). For regularization parameter $\lambda > 0$, define

   $$f(x) := \frac{1}{2}\|Ax - b\|^2 + \frac{\lambda}{2}\|x\|^2.$$

   i.e., validation squared error plus a ridge penalty on the weights.

   We want x to (i) form a probability distribution over the classifiers (non-negative and summing to one), (ii) keep total ensemble latency under a budget B, and (iii) keep total disparity under a threshold r:

   $$\min_{x\in\mathbb{R}^d} f(x) \quad \text{s.t.} \quad \mathbf{1}^\top x = 1, \quad x \geq 0, \quad \ell^\top x \leq B, \quad g^\top x \leq r, \tag{P}$$

   Assume there exists $\bar{x} > 0$ such that $\mathbf{1}^\top \bar{x} = 1$, $\ell^\top \bar{x} < B$, and $g^\top \bar{x} < r$.

   (a) (55 points) Using the normal-cone of a polyhedron to derive the full KKT system for (P).

   (b) (55 points) Derive the scaled ADMM iteration for

   $$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax - b\|^2 + \frac{\lambda}{2}\|x\|^2 + \tau\|Dx\|_1, \qquad \tau > 0,$$

   where $D \in \mathbb{R}^{q\times d}$ is a graph-incidence matrix that pairs up classifiers expected to behave similarly: each row has a $+1$ and a -1 in the two paired entries (and zeros elsewhere), so $\|Dx\|_1$ shrinks weight differences between paired classifiers.

   (c) (60 points) For the equality-only calibration problem $\min_x f(x)$ subject to $Ex = h$, prove the dual-gradient-ascent rate

   $$g(u^*) - g(u_T) \leq \frac{L_g\|u_0 - u^*\|^2}{2T}, \qquad L_g := \frac{\|E\|_{\text{op}}^2}{\lambda},$$

   where $g(u) := -f^*(-E^\top u) - u^\top h$, $u^* \in \arg\max g$, and $u_{k+1} = u_k + L_g^{-1}\nabla g(u_k)$.

2. (170 points) A social platform sees pairwise interaction scores between n users. Stack these into a symmetric matrix $Y = Y^\top \in \mathbb{R}^{n\times n}$, but the platform only observes entries on a symmetric index set Ω. Two effects shape Y: (i) latent community structure, i.e., users in the same community interact similarly, which is well modeled by a low-rank PSD matrix; (ii) a few manipulated edges that are best caught by a separate sparse regression layer. The data-fit term and the PSD trace-bounded constraint set are

   $$F(X) := \frac{1}{2}\|P_\Omega(X - Y)\|_F^2, \qquad D_\tau := \{X \in \mathbb{S}^n : X \succeq 0,\ \operatorname{tr}(X) \leq \tau\}.$$

   (a) (50 points) For a Frank-Wolfe step over D_τ, derive the linear minimization oracle

   $$S_t \in \arg\min_{S\in D_\tau} \langle \nabla F(X_t), S\rangle.$$

(b) (55 points) To avoid carrying the full $n \times n$ matrix, parameterize $X = UU^\top$ with $U \in \mathbb{R}^{n\times k}$ with $k \ll n$. The data-fit term in this factored form is

$$\Phi(U) := \frac{1}{4}\|P_\Omega(UU^\top - Y)\|_F^2, \qquad U \in \mathbb{R}^{n\times k}.$$

Derive the gradient $\nabla\Phi(U)$.

(c) (65 points) To catch the manipulated edges, regress the residual against p hand-crafted anomaly features. Let $B = [b_1, \ldots, b_p] \in \mathbb{R}^{N\times p}$ stack the features over N candidate edges, $r \in \mathbb{R}^N$ be the residual, $\theta \in \mathbb{R}^p$ be the unknown anomaly coefficients, and $\alpha > 0$ promote sparsity. Derive the exact coordinate-descent update for coordinate i in

$$H(\theta) := \frac{1}{2}\|B\theta - r\|^2 + \alpha\|\theta\|_1.$$

3. (160 points) A cloud provider plays a game against a strategic client over how to route network traffic. The provider has m routing policies; the client has n traffic-shifting patterns. Both sides randomize: the provider picks a distribution $x \in \Delta_m$ over its policies and the client picks $y \in \Delta_n$ over its patterns. The provider's expected cost minus the client's expected gain is

$$\chi(x, y) := x^\top By + a^\top x - c^\top y, \qquad B \in \mathbb{R}^{m\times n},\ a \in \mathbb{R}^m,\ c \in \mathbb{R}^n.$$

(a) (40 points) Derive a closed-form expression for the duality gap $\mathrm{gap}(x, y)$ of χ on $\Delta_m \times \Delta_n$.

(b) (50 points) Now drop the simplex constraints and let both players choose freely in $\mathbb{R}^s$, i.e., assume $m = n = s$ and B is nonsingular, and consider the same $\chi(x, y)$ over $\mathbb{R}^s \times \mathbb{R}^s$ without simplex projection. The unique saddle point is then $x^* = B^{-\top}c$, $y^* = -B^{-1}a$. For simultaneous GDA, prove

$$\|x_{t+1} - x^*\|^2 + \|y_{t+1} - y^*\|^2 = \|x_t - x^*\|^2 + \|y_t - y^*\|^2 + \eta^2\|B(y_t - y^*)\|^2 + \eta^2\|B^\top(x_t - x^*)\|^2.$$

The identity is exact, so the distance to z^* strictly grows every step, unless $(x_0, y_0) = (x^*, y^*)$.

(c) (70 points) To fix the divergence, both players use one step of memory: OGDA updates with $z_{t+1} = z_t - 2\eta F(z_t) + \eta F(z_{t-1})$, where $F = (\nabla_x\chi, -\nabla_y\chi)$ is the saddle gradient field. Prove that OGDA converges linearly from every initialization if and only if

$$0 < \eta < \frac{1}{\sqrt{3}\|B\|_{\mathrm{op}}}.$$