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

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

   (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 $B = [b_1, \dots, b_p] \in \mathbb{R}^{N\times p}$ 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 $\operatorname{gap}(x, y)$ 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

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

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

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

Solution 1:

(a) The inequality set is

$$\Omega = \{x : -x \le 0,\ \ell^\top x - B \le 0,\ g^\top x - r \le 0\}.$$

For $x \in \Omega$, the polyhedral normal cone is

$$N_\Omega(x) = \{-s + \alpha\ell + \beta g : s \ge 0,\ \alpha \ge 0,\ \beta \ge 0,\ s_i x_i = 0,\ \alpha(\ell^\top x - B) = 0,\ \beta(g^\top x - r) = 0\}.$$

Let $\nu \in \mathbb{R}$ be the multiplier for $\mathbf{1}^\top x = 1$. Since $\nabla f(x) = A^\top(Ax - b) + \lambda x$, the KKT system is

$$\begin{cases} A^\top(Ax^* - b) + \lambda x^* + \nu^*\mathbf{1} + \alpha^*\ell + \beta^* g - s^* = 0, \\ \mathbf{1}^\top x^* = 1, \qquad x^* \ge 0, \qquad \ell^\top x^* \le B, \qquad g^\top x^* \le r, \\ \alpha^* \ge 0, \qquad \beta^* \ge 0, \qquad s^* \ge 0, \\ \alpha^*(\ell^\top x^* - B) = 0, \qquad \beta^*(g^\top x^* - r) = 0, \qquad s_i^* x_i^* = 0,\ i = 1, \dots, d. \end{cases}$$

The strict feasibility assumption is Slater's condition, so these conditions are also sufficient.

(b) Introduce $z = Dx$ and use the scaled dual variable w. The split problem is

$$\min_{x,z}\ \frac{1}{2}\|Ax - b\|^2 + \frac{\lambda}{2}\|x\|^2 + \tau\|z\|_1 \quad \text{s.t.} \quad Dx - z = 0.$$

The scaled augmented Lagrangian, up to the constant $-\rho\|w\|^2/2$, is

$$\frac{1}{2}\|Ax - b\|^2 + \frac{\lambda}{2}\|x\|^2 + \tau\|z\|_1 + \frac{\rho}{2}\|Dx - z + w\|^2.$$

Thus ADMM is

$$\begin{aligned} x^{k+1} &= (A^\top A + \lambda I + \rho D^\top D)^{-1}\big(A^\top b + \rho D^\top(z^k - w^k)\big), \\ z^{k+1} &= \mathrm{S}_{\tau/\rho}(Dx^{k+1} + w^k), \\ w^{k+1} &= w^k + Dx^{k+1} - z^{k+1}, \end{aligned}$$

where $\mathrm{S}_c(t) = \mathrm{sign}(t)\max\{|t| - c, 0\}$ is applied entrywise.

(c) Since f is λ-strongly convex, f^* is $1/\lambda$-smooth. The dual function is

$$g(u) = \min_x\{f(x) + u^\top(Ex - h)\} = -f^*(-E^\top u) - u^\top h.$$

Moreover, $\nabla g(u) = E\nabla f^*(-E^\top u) - h$. Hence, for any u, v,

$$\|\nabla g(u) - \nabla g(v)\| \le \|E\|_{\mathrm{op}}\frac{1}{\lambda}\|E^\top(u - v)\| \le \frac{\|E\|_{\mathrm{op}}^2}{\lambda}\|u - v\| = L_g\|u - v\|.$$

Let $H = -g$. Then H is convex and L_g-smooth, and the stated dual ascent is exactly gradient descent on H:

$$u_{k+1} = u_k - \frac{1}{L_g}\nabla H(u_k).$$

For gradient descent on a convex L_g-smooth function,

$$H(u_{k+1}) - H(u^*) \le \frac{L_g}{2}\big(\|u_k - u^*\|^2 - \|u_{k+1} - u^*\|^2\big).$$

Summing from $k = 0$ to $T - 1$ gives

$$\sum_{k=1}^{T}\big(H(u_k) - H(u^*)\big) \le \frac{L_g}{2}\|u_0 - u^*\|^2.$$

Since $H(u_k)$ is nonincreasing,

$$T\big(H(u_T) - H(u^*)\big) \le \frac{L_g}{2}\|u_0 - u^*\|^2.$$

Using $H = -g$ gives

$$g(u^*) - g(u_T) \le \frac{L_g\|u_0 - u^*\|^2}{2T}.$$

Solution 2:

(a) Let $G_t := \nabla F(X_t) = P_\Omega(X_t - Y)$. Because Y, X_t, and Ω are symmetric, G_t is symmetric. Let $\lambda_{\min}(G_t)$ be its smallest eigenvalue and let v_t be a unit eigenvector for it. For any $S \succeq 0$,

$$\langle G_t, S\rangle \geq \lambda_{\min}(G_t)\,\mathrm{tr}(S).$$

If $\lambda_{\min}(G_t) < 0$, the minimum is obtained by using all trace budget in the most negative eigendirection: $S_t = \tau v_t v_t^\top$. If $\lambda_{\min}(G_t) \geq 0$, the minimum is obtained by $S_t = 0$. Thus

$$S_t = \begin{cases} \tau v_t v_t^\top, & \lambda_{\min}(G_t) < 0, \\ 0, & \lambda_{\min}(G_t) \geq 0. \end{cases}$$

(b) Define $R(U) := P_\Omega(UU^\top - Y)$. Because Ω and Y are symmetric, $R(U)$ is symmetric. For a perturbation H,

$$R(U + \varepsilon H) = R(U) + \varepsilon P_\Omega(HU^\top + UH^\top) + O(\varepsilon^2).$$

Therefore

$$\begin{aligned} d\Phi(U)[H] &= \frac{1}{2}\langle R(U), P_\Omega(HU^\top + UH^\top)\rangle \\ &= \frac{1}{2}\langle R(U), HU^\top + UH^\top\rangle = \langle R(U)U, H\rangle. \end{aligned}$$

Hence $\nabla\Phi(U) = P_\Omega(UU^\top - Y)U$.

(c) Fix all coordinates except θ_i and define

$$r_{-i} := r - \sum_{j\neq i} b_j\theta_j = r - B\theta + b_i\theta_i.$$

The one-dimensional subproblem is

$$\min_{\xi\in\mathbb{R}} \ \frac{1}{2}\|b_i\xi - r_{-i}\|^2 + \alpha|\xi|.$$

Expanding the quadratic gives

$$\min_{\xi\in\mathbb{R}} \ \frac{\|b_i\|^2}{2}\xi^2 - (b_i^\top r_{-i})\xi + \alpha|\xi| + \text{constant}.$$

If $\|b_i\| > 0$, the exact coordinate update is

$$\theta_i^+ = \frac{1}{\|b_i\|^2}\,\mathrm{S}_\alpha(b_i^\top r_{-i}) = \frac{1}{\|b_i\|^2}\,\mathrm{S}_\alpha(\|b_i\|^2\theta_i - b_i^\top e),$$

where we use the residual $e = B\theta - r$. If $b_i = 0$, then the subproblem is minimized at $\theta_i^+ = 0$.

Solution 3:

(a) For fixed $x \in \Delta_m$,

$$\max_{y\in\Delta_n} \chi(x,y) = a^\top x + \max_{1\leq j\leq n}\left((B^\top x)_j - c_j\right).$$

For fixed $y \in \Delta_n$,

$$\min_{x\in\Delta_m} \chi(x,y) = \min_{1\leq i\leq m}\left((By)_i + a_i\right) - c^\top y.$$

Therefore

$$\mathrm{gap}(x,y) = a^\top x + c^\top y + \max_{1\leq j\leq n}\left((B^\top x)_j - c_j\right) - \min_{1\leq i\leq m}\left((By)_i + a_i\right).$$

(b) In the unconstrained case,

$$\nabla_x \chi(x,y) = By + a, \qquad \nabla_y \chi(x,y) = B^\top x - c.$$

The stated point satisfies $By^* + a = 0$ and $B^\top x^* - c = 0$, so it is the unique saddle point. Define centered variables $\xi_t := x_t - x^*$ and $\upsilon_t := y_t - y^*$. Simultaneous GDA gives

$$\xi_{t+1} = \xi_t - \eta B \upsilon_t, \qquad \upsilon_{t+1} = \upsilon_t + \eta B^\top \xi_t.$$

Hence

$$\begin{aligned}\|\xi_{t+1}\|^2 + \|\upsilon_{t+1}\|^2 &= \|\xi_t - \eta B\upsilon_t\|^2 + \|\upsilon_t + \eta B^\top \xi_t\|^2 \\ &= \|\xi_t\|^2 + \|\upsilon_t\|^2 + \eta^2\|B\upsilon_t\|^2 + \eta^2\|B^\top \xi_t\|^2.\end{aligned}$$

Thus the exact identity is

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

Since B is nonsingular, the last two terms are positive unless $(x_t, y_t) = (x^*, y^*)$.

(c) OGDA in centered variables is

$$\begin{aligned}\xi_{t+1} &= \xi_t - 2\eta B\upsilon_t + \eta B\upsilon_{t-1}, \\ \upsilon_{t+1} &= \upsilon_t + 2\eta B^\top \xi_t - \eta B^\top \xi_{t-1}.\end{aligned}$$

Let $B = U\Sigma V^\top$ be an SVD, where $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_s)$ and $\sigma_j > 0$. Set $\bar{\xi}_t = U^\top \xi_t$ and $\bar{\upsilon}_t = V^\top \upsilon_t$. Each singular mode evolves independently:

$$\begin{aligned}\bar{\xi}_{j,t+1} &= \bar{\xi}_{j,t} - 2\eta\sigma_j \bar{\upsilon}_{j,t} + \eta\sigma_j \bar{\upsilon}_{j,t-1}, \\ \bar{\upsilon}_{j,t+1} &= \bar{\upsilon}_{j,t} + 2\eta\sigma_j \bar{\xi}_{j,t} - \eta\sigma_j \bar{\xi}_{j,t-1}.\end{aligned}$$

Define $w_{j,t} := \bar{\xi}_{j,t} + i\bar{\upsilon}_{j,t}$ and $c_j := \eta\sigma_j$. Then

$$w_{j,t+1} = (1 + 2ic_j)w_{j,t} - ic_j w_{j,t-1}.$$

The characteristic equation is

$$r^2 - (1 + 2ic_j)r + ic_j = 0,$$

with roots

$$r_\pm(c_j) = \frac{1 + 2ic_j \pm \sqrt{1 - 4c_j^2}}{2}.$$

For $0 < c_j \le 1/2$, let $d_j := \sqrt{1 - 4c_j^2}$. Then

$$|r_+(c_j)|^2 = \frac{1 + d_j}{2} < 1, \qquad |r_-(c_j)|^2 = \frac{1 - d_j}{2} < 1.$$

For $c_j > 1/2$, write $\sqrt{1 - 4c_j^2} = i\sqrt{4c_j^2 - 1}$. The larger root has squared modulus

$$\left|\frac{1}{2} + i\left(c_j + \frac{1}{2}\sqrt{4c_j^2 - 1}\right)\right|^2.$$

This is strictly less than 1 if and only if

$$c_j + \frac{1}{2}\sqrt{4c_j^2 - 1} < \frac{\sqrt{3}}{2},$$

which is equivalent to $c_j < 1/\sqrt{3}$. Therefore the recurrence for mode j is linearly stable exactly when $0 < \eta\sigma_j < \frac{1}{\sqrt{3}}$. All modes are linearly stable exactly when

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

If this condition fails, the mode corresponding to $\|B\|_{\mathrm{op}}$ has a characteristic root with modulus at least 1, so OGDA cannot converge linearly from every initialization.