Name: _____ Student ID: _____ Scores: _____

AIE6003 Final
Instructor: Xiaopeng Li **Time Limit**: 2 hours

Notice: You can invoke results in the lecture and homework without proving it.

1. (170 points) A linear model has n tunable coefficients $x \in \mathbb{R}^n$, fitted by minimizing a squared-error loss on a training set. After expanding the loss, the training objective is a convex quadratic in the coefficients, $f(x) := x^\top Q x - 2p^\top x$, where $Q = Q^\top \succ 0$ is the feature Gram matrix and $p \in \mathbb{R}^n$ collects the feature–label correlations. On a small or noisy training set, the fitted coefficients can blow up and produce an unstable model; to keep things well-conditioned, we impose a hard cap on their total magnitude by requiring the coefficient vector to lie inside the unit Euclidean ball. Let $\mathbb{B} := \{x \in \mathbb{R}^n : \|x\| \le 1\}$. Consider
$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad x^\top x \le 1. \tag{P}$$
   (a) (45 points) Derive the normal cone $N_{\mathbb{B}}(x)$ for every feasible $x \in \mathbb{B}$.
   (b) (70 points) Derive the KKT system and the one-dimensional Lagrange dual problem for (P).
   (c) (55 points) Assume $p \neq 0$ and $\lambda_0 \ge 0$. Let $g(\lambda)$ be the dual function from part (b), let $\lambda^* \in \operatorname{argmax}_{\lambda \ge 0} g(\lambda)$, and define projected dual-gradient ascent update
$$\lambda_{t+1} = \left[\lambda_t + L_d^{-1} g'(\lambda_t)\right]_+, \quad \text{where} \quad L_d := \frac{2\|p\|^2}{\lambda_{\min}(Q)^3}.$$
   Prove
$$g(\lambda^*) - g(\lambda_T) \le \frac{L_d |\lambda_0 - \lambda^*|^2}{2T}.$$
   Hint: first compute $g'(\lambda)$ and $g''(\lambda)$, then apply projected gradient descent to $h = -g$ on $[0, \infty)$.

2. (170 points) After deployment, a large model is fine-tuned by adding a correction matrix $X \in \mathbb{R}^{m \times n}$ to its frozen weights. Storing or transmitting the full correction is expensive, so we want X to be effectively low rank, i.e., expressible as the product of two thin matrices. To keep the correction's overall magnitude under control, we cap its nuclear norm by a budget $\tau > 0$: this is a convex surrogate for a hard rank cap. During training, only a subset of entries $\Omega \subseteq [m] \times [n]$ is monitored, and $Y \in \mathbb{R}^{m \times n}$ records the desired correction values on those entries. The data-fit term and the nuclear-norm ball are
$$F(X) := \frac{1}{2}\|P_\Omega(X - Y)\|_F^2, \qquad \mathcal{N}_\tau := \{X \in \mathbb{R}^{m \times n} : \|X\|_* \le \tau\}.$$
   On top of this, a separate correction layer cleans up the residual: it splits the residual into a smooth low-rank piece and a sparse piece.
   (a) (60 points) Given a residual matrix $R \in \mathbb{R}^{m \times n}$, the low-rank + sparse decomposition is captured by the convex objective
$$J(L, S) := \frac{1}{2}\|L + S - R\|_F^2 + \lambda_L \|L\|_* + \alpha \|S\|_1, \qquad \lambda_L, \alpha > 0.$$
   where L is the smooth drift component and S is the spike component. Introduce suitable auxiliary matrices and derive the scaled ADMM iteration with penalty parameter $\rho > 0$, using singular-value thresholding and entrywise soft-thresholding.

(b) (55 points) To avoid storing the full $m \times n$ matrix X, parametrize $X = UV^\top$ with $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ ($k \ll \min(m,n)$). The factorization is non-unique, so we add a balancing regularizer that pushes U and V to have the same scale. Derive the gradients of

$$\Psi(U,V) := \frac{1}{2}\|P_\Omega(UV^\top - Y)\|_F^2 + \frac{\gamma}{4}\|U^\top U - V^\top V\|_F^2, \qquad \gamma > 0.$$

(c) (55 points) An alternative to factorization is to keep X in matrix form but update it by Frank-Wolfe over $\mathcal{N}_\tau$, which never forms a dense matrix: each linear minimization returns a single rank-1 outer product, so after t iterations the iterate is a convex combination of at most t such outer products. Assume F is L-smooth on $\mathcal{N}_\tau$. Starting with $X_0 = 0$, consider Frank-Wolfe with

$$S_t \in \operatorname*{argmin}_{S\in\mathcal{N}_\tau} \langle \nabla F(X_t), S\rangle, \qquad X_{t+1} = (1-\eta_t)X_t + \eta_t S_t, \qquad \eta_t = \frac{2}{t+2}.$$

Denote $X^* \in \operatorname{argmin}_{X\in\mathcal{N}_\tau} F(X)$. Prove

$$F(X_t) - F(X^*) \le \frac{8L\tau^2}{t+2}, \qquad t \ge 1.$$

Hint: Use smoothness, the Frank-Wolfe oracle, convexity of F, and $\operatorname{diam}(\mathcal{N}_\tau) \le 2\tau$.

3. (160 points) A power-grid operator must dispatch resources in real time while defending against worst-case demand fluctuations. The operator chooses a load-shifting vector $u \in [-1,1]^m$ ($u_i \in [-1,1]$ is how much of the i-th flexible resource to ramp up or down, in normalized units, capped at full intensity in either direction). An adversary chooses a demand perturbation $v \in [-1,1]^n$ ($v_j \in [-1,1]$ is the percentage shock applied to the j-th load bucket, again capped). The operator's loss combines a coupling term $u^\top Cv$ (how the dispatch responds to the demand shock through the network), a fixed dispatch cost $a^\top u$ on the operator side, and a fixed reward $b^\top v$ for the adversary:

$$\psi(u,v) := u^\top Cv + a^\top u - b^\top v, \qquad C \in \mathbb{R}^{m\times n},\ a \in \mathbb{R}^m,\ b \in \mathbb{R}^n.$$

The operator minimizes ψ over its action set; the adversary maximizes ψ over its action set.

(a) (40 points) The duality gap measures how far a candidate strategy pair (u,v) is from a saddle-point equilibrium: it adds the operator's regret to the adversary's regret. Derive a closed-form expression for the duality gap $\operatorname{gap}(u,v)$ of ψ on $[-1,1]^m \times [-1,1]^n$.

(b) (70 points) On a tight market day, both sides become risk averse and pay a quadratic penalty for large actions, which is captured by adding $\frac{\mu}{2}\|u\|^2$ to the operator's loss and $\frac{\mu}{2}\|v\|^2$ to the adversary's loss. The action sets are also relaxed: instead of the box, the operator moves freely in $\mathbb{R}^m$ and the adversary moves freely in $\mathbb{R}^n$. The regularized unconstrained payoff becomes

$$\psi_\mu(u,v) := \frac{\mu}{2}\|u\|^2 + u^\top Cv - \frac{\mu}{2}\|v\|^2 + a^\top u - b^\top v, \qquad \mu > 0.$$

The regularization makes the game strongly convex in u and strongly concave in v, so simultaneous GDA contracts toward the saddle point as long as the stepsize is in the right range. Prove that simultaneous GDA is a global contraction exactly when

$$0 < \eta < \frac{2\mu}{\mu^2 + \|C\|_{\mathrm{op}}^2}.$$

(c) (50 points) Drop the regularizers and the action box, leaving only the pure coupling term: the operator and adversary play the centered bilinear game $\min_u \max_v u^\top Cv$ over $\mathbb{R}^s \times \mathbb{R}^s$, with $m = n = s$ and C nonsingular. Bilinear games of this type are notoriously bad for explicit GDA, so consider instead the extragradient (EG) update,

$$z_{t+1/2} = z_t - \eta F(z_t), \qquad z_{t+1} = z_t - \eta F(z_{t+1/2}),$$

where $z = (u,v)$ and $F = (\nabla_u(u^\top Cv), -\nabla_v(u^\top Cv))$, prove that if $0 < \eta < 1/\|C\|_{\mathrm{op}}$, then EG converges linearly to $(0,0)$, and derive the exact SVD-mode contraction factor.
Hint: apply SVD to C and decouple the original problem into independent 2-d bilinear games.