Name: _____ Student ID: _____ Scores: _____

AIE6003 Homework 3
Instructor: Xiaopeng Li **Deadline**: April 12

Notice: If you are trying to invoke a result outside of this course, then you should put a reference to it. The questions with 1 star are easy, with 2 stars are medium, and with 3 stars are hard. Easy and medium level of questions are likely to occur in the exams, while hard questions are not likely to occur in the exams.

1*. (14 points) This problem studies why Adam can be faster than GD on block-heterogeneous quadratic problems. Consider the Adam special case with $\beta_1 = 0$, $\beta_2 = 1$, $\epsilon = 0$, and no bias correction. In this case, Adam uses the fixed diagonal preconditioner and the update is

$$w^{t+1} = w^t - \alpha D_0^{-1} \nabla f(w^t), \quad D_0 = \mathbf{diag}(|\nabla f(w^0)|),$$

where $|\cdot|$ denotes entry-wise absolute value.

(a) (7 points) Let $0 < a_1 \le a_2 \le \cdots \le a_L$, $\kappa := \frac{a_L}{a_1}$ and

$$f(w) = \frac{1}{2} \sum_{\ell=1}^{L} a_\ell \|w_\ell\|_2^2, \qquad w = (w_1, \ldots, w_L), \quad w_\ell \in \mathbb{R}^{d_\ell}.$$

Assume every coordinate of w^0 equals $+1$ or -1. Prove that the GD with constant step size $\eta > 0$ converges iff $0 < \eta < 2/a_L$ and the best constant stepsize is $\eta^\star = \frac{2}{a_1 + a_L}$, with exact contraction factor of the iterates $\rho_{\mathrm{GD}}^\star = \frac{\kappa - 1}{\kappa + 1}$. Next, prove that Adam converges iff $0 < \alpha < 2$, with exact contraction factor $\rho_{\mathrm{Adam}} = |1 - \alpha|$, which is independent of κ. Explain in one sentence why this shows acceleration of Adam over GD on this model.

(b) (7 points) Let $w \in \mathbb{R}^9$, partitioned into three blocks $w_1, w_2, w_3 \in \mathbb{R}^3$. Define

$$f(w) = \frac{1}{2} w^\top H w, \qquad H = \mathbf{diag}(H_1, H_2, H_3), \qquad H_\ell = Q_\ell \Lambda_\ell Q_\ell^\top,$$

where each $Q_\ell \in \mathbb{R}^{3 \times 3}$ is an orthogonal matrix obtained from the QR factorization of a Gaussian random matrix. Use the same Q_1, Q_2, Q_3 in both cases.

Case H:

$$\Lambda_1 = \mathbf{diag}(1, 2, 3), \qquad \Lambda_2 = \mathbf{diag}(99, 100, 101), \qquad \Lambda_3 = \mathbf{diag}(4998, 4999, 5000).$$

Case M:

$$\Lambda_1 = \mathbf{diag}(1, 99, 4998), \qquad \Lambda_2 = \mathbf{diag}(2, 100, 4999), \qquad \Lambda_3 = \mathbf{diag}(3, 101, 5000).$$

Use the same random initialization $w^0 \sim \mathcal{N}(0, I_9)$ for both methods in each case. Run GD and Adam in part (a), with the best η and α you can find. Compare the convergence of the two algorithms numerically and briefly explain why the gap between GD and Adam is much larger in Case H than in Case M, even though the two problems have the same condition number.

2**. (16 points) Consider the minimum-variance portfolio problem

$$\min_{x \in \mathbb{R}^n} \ \frac{1}{2} x^\top \Sigma x \quad \text{s.t.} \quad \mu^\top x \ge r, \qquad \mathbf{1}^\top x = 1, \qquad x \ge 0,$$

where $\Sigma \succ 0$, $\mu \in \mathbb{R}^n$, and $r > 0$. Assume there exists a strictly feasible portfolio $\bar{x}$ such that

$$\bar{x} > 0, \qquad \mathbf{1}^\top \bar{x} = 1, \qquad \mu^\top \bar{x} > r.$$

(a) (4 points) Write the problem in the form

$$\min\ f(x) \qquad \text{s.t.} \qquad x \in \Omega,\ \ Ax = b,$$

and identify f, Ω, A, b. Also write Ω as a polyhedron $\Omega = \{x : Cx \le d\}$. Using the normal-cone formula from the lecture, write $N_\Omega(x)$ explicitly.

(b) (4 points) Using the optimality condition

$$-\nabla f(x^*) - A^\top \lambda^* \in N_\Omega(x^*),$$

derive the full KKT system for this problem.

(c) (4 points) Form the Lagrangian and derive the dual function $q(\alpha, \lambda, \beta)$, where $\alpha \ge 0$, $\lambda \in \mathbb{R}$, $\beta \ge 0$. State the dual problem. Explain why every dual feasible point gives a lower bound on the primal optimum, and use the strict-feasibility assumption to conclude strong duality.

(d) (4 points) Now consider

$$\Sigma = \operatorname{diag}(2, 1, 3), \qquad \mu = \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix}, \qquad r = \frac{9}{2}.$$

Assume the optimal solution satisfies $x_i^* > 0$ for all i and $\mu^\top x^* = r$. Under this assumption, solve the KKT system and find x^*, α^*, and λ^*.

3***. (20 points) For a fixed $\mu > 0$ consider the nonnegative Lasso problem (barrier version)

$$(P_\mu) \qquad \min_{x \in \mathbb{R}^n_{++}} \frac{1}{2}\|Ax - b\|_2^2 + \gamma \mathbf{1}^\top x - \mu \sum_{i=1}^n \log x_i.$$

Introduce a positive dummy variable $y \in \mathbb{R}^n_{++}$ and rewrite (P_μ) as

$$\min_{x \in \mathbb{R}^n,\ y \in \mathbb{R}^n_{++}} f(x) + g(y) \qquad \text{s.t.} \qquad x - y = 0,$$

where

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2, \qquad g(y) = \gamma \mathbf{1}^\top y - \mu \sum_{i=1}^n \log y_i.$$

Let $\beta > 0$ be the penalty parameter in the augmented Lagrangian and u be the scaled dual variable.

(a) (5 points) Write down the augmented Lagrangian for the above problem, and then write the ADMM update in scaled form, with the x-update and the y-update in closed form, that is, do not leave arg min in the update.

(b) (5 points) Let $r^{k+1} := x^{k+1} - y^{k+1}$. Show that the ADMM iterates satisfy

$$\nabla f(x^{k+1}) + \beta u^{k+1} + \beta(y^{k+1} - y^k) = 0, \qquad \nabla g(y^{k+1}) - \beta u^{k+1} = 0.$$

Show also that any KKT point $(x^\star, y^\star, u^\star)$ of the split problem satisfies

$$\nabla f(x^\star) + \beta u^\star = 0, \qquad \nabla g(y^\star) - \beta u^\star = 0, \qquad x^\star = y^\star.$$

Finally, show $(y^{k+1} - y^k)^\top r^{k+1} \ge 0$ for all $k \ge 1$.

(c) (5 points) Let $(x^\star, y^\star, u^\star)$ be any KKT point and define

$$V^k := \beta\|u^k - u^\star\|_2^2 + \beta\|y^k - y^\star\|_2^2.$$

Use part (b) and the monotonicity of ∇f and ∇g to prove the descent estimate

$$V^{k+1} \le V^k - \beta\|x^{k+1} - y^{k+1}\|_2^2 - \beta\|y^{k+1} - y^k\|_2^2, \qquad k \ge 1.$$

(d) (5 points) Show that (P_μ) has a unique minimizer $x_\mu^\star \in \mathbb{R}^n_{++}$. Let $y_\mu^\star := x_\mu^\star$, $u_\mu^\star := \beta^{-1}(\gamma \mathbf{1} - \mu(x_\mu^\star)^{-1})$, where the inverse is taken componentwise. Use part (c) to prove that the ADMM iterates converge, i.e., $x^k \to x_\mu^\star$, $y^k \to y_\mu^\star$ and $u^k \to u_\mu^\star$.