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 $0 < a_1 \leq a_2 \leq \cdots \leq a_L$, $\kappa := \frac{a_L}{a_1}$ 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 $\rho^\star_{\mathrm{GD}} = \frac{\kappa-1}{\kappa+1}$. 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.

Solution: The unique minimizer is $w^\star = 0$ and $\nabla f(w) = (a_1 w_1, \ldots, a_L w_L)$, so the GD update becomes, block by block,

$$w_\ell^{t+1} = w_\ell^t - \eta a_\ell w_\ell^t = (1 - \eta a_\ell) w_\ell^t, \qquad \ell = 1, \ldots, L.$$

Hence $w_\ell^t = (1-\eta a_\ell)^t w_\ell^0$. Therefore GD converges if and only if $|1 - \eta a_\ell| < 1$ for all ℓ, i.e., $0 < \eta < \frac{2}{a_L}$. The contraction factor of the iterates is

$$\rho_{\mathrm{GD}}(\eta) = \max_\ell |1 - \eta a_\ell| = \max\{|1-\eta a_1|,\ |1-\eta a_L|\}.$$

because $a_\ell \in [a_1, a_L]$ and $a \mapsto |1 - \eta a|$ is convex. It is easy to see the optimal $\eta^\star = \frac{2}{a_1 + a_L}$. Substituting this into the above expression gives

$$\rho^\star_{\mathrm{GD}} = |1 - \eta^\star a_1| = |1 - \eta^\star a_L| = \frac{a_L - a_1}{a_L + a_1} = \frac{\kappa - 1}{\kappa + 1}.$$

Now consider Adam. Since every coordinate of w^0 equals $+1$ or -1,

$$|\nabla f(w^0)| = (a_1 \mathbf{1}_{d_1}, \ldots, a_L \mathbf{1}_{d_L}), \quad \text{and} \quad D_0 = \mathbf{diag}(a_1 I_{d_1}, \ldots, a_L I_{d_L})$$

Therefore $D_0^{-1}\nabla f(w^t) = (w_1^t, \ldots, w_L^t) = w^t$ and Adam update becomes

$$w^{t+1} = w^t - \alpha w^t = (1-\alpha) w^t,$$

so $w^t = (1-\alpha)^t w^0$. Thus Adam converges if and only if $|1-\alpha| < 1$ or equivalently, $0 < \alpha < 2$. Its exact contraction factor is $\rho_{\mathrm{Adam}} = |1 - \alpha|$. This factor is independent of κ.
Finally, GD needs $O(\kappa \log(1/\varepsilon))$ iterations to reduce the error by a factor ε, while this Adam special case needs only $O(\log(1/\varepsilon))$ iterations; thus Adam removes the dependence on the global heterogeneity parameter κ 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.

Solution: A correct numerical experiment should show the following qualitative pattern.

- In **Case H**, Adam decreases $\|\nabla f(w^t)\|_2$ much faster than GD.
- In **Case M**, the gap between GD and Adam is much smaller.

In Case H each block has a very different spectral scale:

$$\{1, 2, 3\}, \qquad \{99, 100, 101\}, \qquad \{4998, 4999, 5000\},$$

so one single GD stepsize must be chosen conservatively for the hardest block, which slows down the other blocks. By contrast, Adam uses the diagonal preconditioner D_0^{-1}, which rescales coordinates and handles the heterogeneous blocks much better.

In Case M, the same eigenvalues are redistributed across the three blocks:

$$\{1, 99, 4998\}, \qquad \{2, 100, 4999\}, \qquad \{3, 101, 5000\},$$

so the blocks are much more similar to each other. Therefore one global GD stepsize is no longer as badly bottlenecked, and GD becomes much closer to Adam.

Hence the difference between the two cases cannot be explained by the full condition number alone; it is the block heterogeneity that explains why GD is much worse than Adam in Case H.

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

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

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

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

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

Solution: We set $f(x) = \frac{1}{2}x^\top \Sigma x$, $\Omega = \{x \in \mathbb{R}^n : \mu^\top x \geq r,\ x \geq 0\}$, $A = \mathbf{1}^\top$, and $b = 1$. Thus the problem has the desired form. We can also write Ω as a polyhedron $\Omega = \{x : Cx \leq d\}$ with

$$C = \begin{pmatrix} -\mu^\top \\ -I \end{pmatrix}, \qquad d = \begin{pmatrix} -r \\ 0 \end{pmatrix}.$$

By the normal-cone formula for a polyhedron,

$$N_\Omega(x) = \{C^\top \nu : \ \nu \geq 0,\ \nu_j(C_j x - d_j) = 0\}.$$

Writing $\nu = (\alpha, \beta)$, where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^n$, we obtain

$$N_\Omega(x) = \left\{-\alpha\mu - \beta : \ \alpha \geq 0,\ \beta \geq 0,\ \alpha(\mu^\top x - r) = 0,\ \beta_i x_i = 0\ \forall i\right\}.$$

(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.

Solution: Since $\nabla f(x) = \Sigma x$, the optimality condition $-\nabla f(x^*) - A^\top \lambda^* \in N_\Omega(x^*)$ becomes

$$-\Sigma x^* - \lambda^* \mathbf{1} \in N_\Omega(x^*).$$

Using part (a), there exist $\alpha^* \geq 0$ and $\beta^* \geq 0$ such that

$$-\Sigma x^* - \lambda^* \mathbf{1} = -\alpha^* \mu - \beta^*.$$

Therefore the full KKT system is

$$\begin{cases} \Sigma x^* + \lambda^* \mathbf{1} - \alpha^* \mu - \beta^* = 0, \\ \mathbf{1}^\top x^* = 1, \\ \mu^\top x^* \geq r, \\ x^* \geq 0, \\ \alpha^* \geq 0, \qquad \alpha^*(\mu^\top x^* - r) = 0, \\ \beta^* \geq 0, \qquad \beta_i^* x_i^* = 0, \quad i = 1, \ldots, n. \end{cases}$$

(c) (4 points) Form the Lagrangian and derive the dual function $q(\alpha, \lambda, \beta)$, where $\alpha \geq 0$, $\lambda \in \mathbb{R}$, $\beta \geq 0$. 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.

Solution: The Lagrangian is

$$\mathcal{L}(x, \alpha, \lambda, \beta) = \frac{1}{2}x^\top \Sigma x + \alpha(r - \mu^\top x) + \lambda(\mathbf{1}^\top x - 1) - \beta^\top x,$$

where $\alpha \geq 0$, $\lambda \in \mathbb{R}$, and $\beta \geq 0$. To compute the dual function, minimize over x. The stationarity condition is

$$\nabla_x \mathcal{L}(x, \alpha, \lambda, \beta) = \Sigma x - \alpha\mu + \lambda \mathbf{1} - \beta = 0 \implies x(\alpha, \lambda, \beta) = \Sigma^{-1}(\alpha\mu - \lambda\mathbf{1} + \beta).$$

Substituting this into the Lagrangian gives

$$q(\alpha, \lambda, \beta) = -\frac{1}{2}(\alpha\mu - \lambda\mathbf{1} + \beta)^\top \Sigma^{-1}(\alpha\mu - \lambda\mathbf{1} + \beta) + \alpha r - \lambda.$$

Hence the dual problem is

$$\max_{\alpha \geq 0,\ \lambda \in \mathbb{R},\ \beta \geq 0} \quad -\frac{1}{2}(\alpha\mu - \lambda\mathbf{1} + \beta)^\top \Sigma^{-1}(\alpha\mu - \lambda\mathbf{1} + \beta) + \alpha r - \lambda.$$

For every primal feasible x and dual feasible (α, λ, β), we have $\mathcal{L}(x, \alpha, \lambda, \beta) \leq f(x)$, because $r - \mu^\top x \leq 0$, $\mathbf{1}^\top x - 1 = 0$, $-x \leq 0$, and $\alpha, \beta \geq 0$. Therefore

$$q(\alpha, \lambda, \beta) = \inf_z \mathcal{L}(z, \alpha, \lambda, \beta) \leq \mathcal{L}(x, \alpha, \lambda, \beta) \leq f(x).$$

Taking the supremum over dual feasible variables and the infimum over primal feasible x yields weak duality $d^* \leq p^*$. Finally, the strict feasibility assumption is exactly Slater's condition. Since the primal problem is convex, Slater's condition implies strong duality $d^* = p^*$.

(d) (4 points) Now consider

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

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^*.

Solution: Under the assumption $x_i^* > 0$ for all i, complementary slackness gives $\beta^* = 0$. Under the assumption $\mu^\top x^* = r$, the return constraint is active. Thus the KKT system reduces to

$$\Sigma x^* + \lambda^* \mathbf{1} - \alpha^* \mu = 0, \qquad \mathbf{1}^\top x^* = 1, \qquad \mu^\top x^* = r = \tfrac{9}{2}.$$

The stationarity equations are

$$2x_1 + \lambda - 6\alpha = 0, \qquad x_2 + \lambda - 4\alpha = 0, \qquad 3x_3 + \lambda - 2\alpha = 0.$$

Hence

$$x_1 = 3\alpha - \tfrac{\lambda}{2}, \qquad x_2 = 4\alpha - \lambda, \qquad x_3 = \tfrac{2\alpha - \lambda}{3}.$$

Using $x_1 + x_2 + x_3 = 1$, we get

$$3\alpha - \tfrac{\lambda}{2} + 4\alpha - \lambda + \tfrac{2\alpha - \lambda}{3} = 1 \implies 46\alpha - 11\lambda = 6.$$

Using $6x_1 + 4x_2 + 2x_3 = \frac{9}{2}$, we get

$$6(3\alpha - \tfrac{\lambda}{2}) + 4(4\alpha - \lambda) + 2(\tfrac{2\alpha - \lambda}{3}) = \tfrac{9}{2} \implies 212\alpha - 46\lambda = 27.$$

Solving these two equations gives $\alpha^* = \frac{7}{72}$ and $\lambda^* = -\frac{5}{36}$. Substituting back,

$$x_1^* = 3 \cdot \tfrac{7}{72} - \tfrac{1}{2}(-\tfrac{5}{36}) = \tfrac{13}{36}, \quad x_2^* = 4 \cdot \tfrac{7}{72} - (-\tfrac{5}{36}) = \tfrac{19}{36}, \quad x_3^* = \tfrac{2 \cdot \frac{7}{72} - (-\frac{5}{36})}{3} = \tfrac{1}{9}.$$

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.

Solution: The unscaled augmented Lagrangian is

$$\mathcal{L}_\beta(x,y;\lambda) = \frac{1}{2}\|Ax-b\|_2^2 + \gamma\mathbf{1}^\top y - \mu\sum_{i=1}^n \log y_i + \lambda^\top(x-y) + \frac{\beta}{2}\|x-y\|_2^2.$$

Let $u = \lambda/\beta$ be the scaled dual variable. Then the scaled ADMM iteration is

$$\begin{cases} x^{k+1} = \arg\min_x \ \frac{1}{2}\|Ax-b\|_2^2 + \frac{\beta}{2}\|x-y^k+u^k\|_2^2, \\ y^{k+1} = \arg\min_{y\in\mathbb{R}^n_{++}} \ \gamma\mathbf{1}^\top y - \mu\sum_{i=1}^n \log y_i + \frac{\beta}{2}\|x^{k+1}-y+u^k\|_2^2, \\ u^{k+1} = u^k + x^{k+1} - y^{k+1}. \end{cases}$$

For the x-update, the first-order condition gives

$$A^\top(Ax^{k+1}-b) + \beta(x^{k+1}-y^k+u^k) = 0.$$

Hence

$$x^{k+1} = (A^\top A + \beta I)^{-1}(A^\top b + \beta(y^k - u^k)).$$

For the y-update, let $v^k := x^{k+1} + u^k$. Then the problem is separable:

$$y_i^{k+1} = \arg\min_{y_i>0}\left\{\gamma y_i - \mu\log y_i + \frac{\beta}{2}(y_i - v_i^k)^2\right\}.$$

The first-order condition is

$$\gamma - \frac{\mu}{y_i} + \beta(y_i - v_i^k) = 0 \iff \beta y_i^2 + (\gamma - \beta v_i^k)y_i - \mu = 0.$$

Taking the unique positive root, we obtain

$$y_i^{k+1} = \frac{1}{2}\left(v_i^k - \frac{\gamma}{\beta} + \sqrt{\left(v_i^k - \frac{\gamma}{\beta}\right)^2 + \frac{4\mu}{\beta}}\right), \qquad i = 1,\ldots,n.$$

Finally, the dual update is $u^{k+1} = u^k + x^{k+1} - y^{k+1}$.

(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 $(y^{k+1} - y^k)^\top r^{k+1} \geq 0$ for all $k \geq 1$.

Solution: Recall that $r^{k+1} := x^{k+1} - y^{k+1}$. From the x-subproblem,

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

Using dual update $u^{k+1} = u^k + x^{k+1} - y^{k+1}$, we have

$$x^{k+1} - y^k + u^k = u^{k+1} + y^{k+1} - y^k.$$

Therefore, $\nabla f(x^{k+1}) + \beta u^{k+1} + \beta(y^{k+1} - y^k) = 0$. From the y-subproblem,

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

similarly by using dual update, we get $\nabla g(y^{k+1}) - \beta u^{k+1} = 0$. Now let $(x^\star, y^\star, u^\star)$ be any KKT point of the split problem. Then stationarity and feasibility give

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

Finally, for $k \geq 1$, from the second identity above we have

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

Since g is convex and differentiable, ∇g is monotone, so

$$0 \leq (y^{k+1} - y^k)^\top \big(\nabla g(y^{k+1}) - \nabla g(y^k)\big) = \beta(y^{k+1} - y^k)^\top (u^{k+1} - u^k).$$

But $u^{k+1} - u^k = r^{k+1}$, hence $(y^{k+1} - y^k)^\top r^{k+1} \geq 0$ for $k \geq 1$.

(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

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

Solution: Since f is convex and differentiable, ∇f is monotone. Therefore

$$0 \leq (x^{k+1} - x^\star)^\top (\nabla f(x^{k+1}) - \nabla f(x^\star)).$$

Using part (b),

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

so

$$0 \leq -\beta(x^{k+1} - x^\star)^\top (u^{k+1} - u^\star) - \beta(x^{k+1} - x^\star)^\top (y^{k+1} - y^k). \tag{1}$$

Similarly, since g is convex and differentiable, ∇g is monotone:

$$0 \leq (y^{k+1} - y^\star)^\top (\nabla g(y^{k+1}) - \nabla g(y^\star)).$$

Using part (b),

$$\nabla g(y^{k+1}) = \beta u^{k+1}, \qquad \nabla g(y^\star) = \beta u^\star,$$

hence

$$0 \leq \beta(y^{k+1} - y^\star)^\top (u^{k+1} - u^\star). \tag{2}$$

Adding (1) and (2), and using $x^\star = y^\star$, gives

$$-\beta r^{k+1\top}(u^{k+1} - u^\star) - \beta(y^{k+1} - y^\star)^\top (y^{k+1} - y^k) \geq \beta r^{k+1\top}(y^{k+1} - y^k), \tag{3}$$

where $r^{k+1} = x^{k+1} - y^{k+1}$. Compute

$$V^k - V^{k+1} = \beta(\|u^k - u^\star\|_2^2 - \|u^{k+1} - u^\star\|_2^2) + \beta(\|y^k - y^\star\|_2^2 - \|y^{k+1} - y^\star\|_2^2).$$

Using $u^{k+1} - u^k = r^{k+1}$ and $\|a\|_2^2 - \|b\|_2^2 = -2b^\top(b-a) + \|a-b\|_2^2$, we obtain

$$V^k - V^{k+1} = -2\beta(u^{k+1} - u^\star)^\top r^{k+1} - 2\beta(y^{k+1} - y^\star)^\top (y^{k+1} - y^k) + \beta\|r^{k+1}\|_2^2 + \beta\|y^{k+1} - y^k\|_2^2.$$

By (3) and part (b),

$$\begin{aligned} V^k - V^{k+1} &\geq 2\beta r^{k+1\top}(y^{k+1} - y^k) + \beta\|r^{k+1}\|_2^2 + \beta\|y^{k+1} - y^k\|_2^2. \\ &\geq \beta\|r^{k+1}\|_2^2 + \beta\|y^{k+1} - y^k\|_2^2. \end{aligned}$$

(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 $u^k \to u_\mu^\star$.

Solution: Define Φ_μ as the objective function of (P_μ). As any component $x_i \downarrow 0$, the barrier term $-\mu \log x_i \to +\infty$, so $\Phi_\mu(x) \to +\infty$ near the boundary of $\mathbb{R}_{++}^n$. Also, as $\|x\|_2 \to \infty$ with $x > 0$, the linear term $\gamma\mathbf{1}^\top x$ dominates the logarithmic term, so $\Phi_\mu(x) \to +\infty$. Hence Φ_μ attains a minimizer $x_\mu^\star \in \mathbb{R}_{++}^n$. In addition, for $x \in \mathbb{R}_{++}^n$,

$$\nabla^2 \Phi_\mu(x) = A^\top A + \mu\, \mathbf{diag}(x_1^{-2}, \ldots, x_n^{-2}) \succ 0.$$

Thus Φ_μ is strictly convex on $\mathbb{R}_{++}^n$, so the minimizer $x_\mu^\star$ is unique.
The first-order optimality condition for Φ_μ is

$$A^\top(Ax_\mu^\star - b) + \gamma\mathbf{1} - \mu(x_\mu^\star)^{-1} = 0.$$

Hence

$$\nabla f(x_\mu^\star) + \beta u_\mu^\star = 0, \qquad \nabla g(y_\mu^\star) - \beta u_\mu^\star = 0, \qquad x_\mu^\star = y_\mu^\star.$$

So $(x_\mu^\star, y_\mu^\star, u_\mu^\star)$ is a KKT point of the split problem.
Next, part (c) implies that for $k \geq 1$, $V^{k+1} \leq V^k$, so $\{V^k\}$ is decreasing and bounded below by 0. Therefore $\{V^k\}$ converges, and in particular $\{u^k\}$ and $\{y^k\}$ are bounded.
Summing the descent estimate from $k = 1$ to N, we get

$$\beta \sum_{k=1}^N \|x^{k+1} - y^{k+1}\|_2^2 + \beta \sum_{k=1}^N \|y^{k+1} - y^k\|_2^2 \leq V^1 - V^{N+1} \leq V^1.$$

Hence

$$\sum_{k=1}^{\infty} \|x^{k+1} - y^{k+1}\|_2^2 < \infty, \qquad \sum_{k=1}^{\infty} \|y^{k+1} - y^k\|_2^2 < \infty.$$

Therefore $x^{k+1} - y^{k+1} \to 0$ and $y^{k+1} - y^k \to 0$. Since $\{y^k\}$ is bounded, the sequence $\{x^k\}$ is also bounded. Now take any convergent subsequence $x^{k_j} \to \bar{x}$, $y^{k_j} \to \bar{y}$, and $u^{k_j} \to \bar{u}$. Because $x^{k_j} - y^{k_j} \to 0$, we have $\bar{x} = \bar{y}$.

Also, from part (b),

$$\beta u^{k+1} = \nabla g(y^{k+1}) = \gamma \mathbf{1} - \mu (y^{k+1})^{-1}.$$

Since $\{u^k\}$ is bounded, $\{(y^k)^{-1}\}$ is bounded, so $\bar{y} \in \mathbb{R}^n_{++}$.

Passing to the limit in the identities from part (b), and using $y^{k+1} - y^k \to 0$, we obtain

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

Thus $(\bar{x}, \bar{y}, \bar{u})$ is a KKT point of the split problem.

But the KKT point is unique: indeed $x_\mu^\star$ is unique, then $y_\mu^\star = x_\mu^\star$, and finally $u_\mu^\star$ is uniquely determined by $\beta u_\mu^\star = \nabla g(y_\mu^\star)$. Hence every cluster point of (x^k, y^k, u^k) must equal $(x_\mu^\star, y_\mu^\star, u_\mu^\star)$. Since the whole sequence is bounded, it follows that $x^k \to x_\mu^\star$, $y^k \to y_\mu^\star$, and $u^k \to u_\mu^\star$.