1. *Matrix sensing and sequential convex optimization.* In the matrix sensing problem, we receive observations of the form

$$b_i = \mathbf{tr}\, A_i^T X, \quad i = 1, \ldots, m,$$

where $X \in \mathbf{R}^{n_1 \times n_2}$ is assumed to be low rank, and $A_i \in \mathbf{R}^{n_1 \times n_2}$ as well. In this problem, we will assume that this (unknown) X is rank 1, so that we may write $X = x_\star y_\star^T$, where $x_\star \in \mathbf{R}^{n_1}$ and $y_\star \in \mathbf{R}^{n_2}$. To estimate X, we therefore wish to solve the (non-smooth, non-convex) problem

$$\text{minimize} \quad \sum_{i=1}^m \left| b_i - \mathbf{tr}\, A_i^T x y^T \right| = \sum_{i=1}^m \left| b_i - x^T A_i y \right| \tag{1}$$

in variables $x \in \mathbf{R}^{n_1}, y \in \mathbf{R}^{n_2}$. Luckily, this is an instance of a *composite optimization problem*, which we can (approximately) minimize by sequentially minimizing convex approximations.

The general composite problem is as follows: we wish to minimize $f(x) = h(c(x))$, where $h : \mathbf{R}^m \to \mathbf{R}$ is convex and $c : \mathbf{R}^n \to \mathbf{R}^m$ is a smooth function. Abusing notation and letting $\nabla c(x) \in \mathbf{R}^{n \times n}$ denote the matrix $[\nabla c_1(x) \ \cdots \ \nabla c_m(x)]$ of the gradients of the component functions making up c, a natural model of f at a point x is

$$f_x(x + \Delta) := h(c(x) + \nabla c(x)^T \Delta),$$

which is evidently convex in Δ. Then one version of *sequential convex optimization* iterates via

$$\begin{aligned} \Delta x &:= \underset{\Delta}{\text{argmin}} \left\{ f_{x^k}(x^k + \Delta) + \frac{1}{2\alpha} \|\Delta\|_2^2 \right\}, \\ x^{k+1} &:= x^k + \Delta x \end{aligned}$$

where $\alpha > 0$ is a stepsize chosen to guarantee that the objectives $f(x^k)$ are decreasing. We'll specialize this iteration to the matrix sensing problem (1).

For the rest of the problem, we will assume the sensing matrices A_i take the form $A_i = u_i v_i^T$ for vectors $u_i \in \mathbf{R}^{n_1}, v_i \in \mathbf{R}^{n_2}$, so the observations in the sensing problem now take the form

$$b_i = x_\star^T u_i v_i^T y_\star, \quad i = 1, \ldots, m.$$

(a) Write problem (1) in the form

$$f(x, y) = \sum_{i=1}^m h(c_i(x, y))$$

where $h : \mathbf{R} \to \mathbf{R}_+$ is a convex function and $c_i : \mathbf{R}^{n_1 + n_2} \to \mathbf{R}$. Your formulation should satisfy

$$\nabla c_i(x, y) = \begin{bmatrix} (v_i^T y) u_i \\ (u_i^T x) v_i \end{bmatrix}.$$

Define the linear mapping $\mathcal{A} : \mathbf{R}^{n_1 \times n_2} \to \mathbf{R}^m$ by

$$\mathcal{A}(X) = \begin{bmatrix} \mathbf{tr}\, A_1^T X & \mathbf{tr}\, A_2^T X & \cdots & \mathbf{tr}\, A_m^T X \end{bmatrix}^T.$$

and define the matrices

$$U = [u_1 \ u_2 \ \cdots \ u_m]^T \in \mathbf{R}^{m \times n_1} \quad \text{and} \quad V = [v_1 \ v_2 \ \cdots \ v_m]^T \in \mathbf{R}^{m \times n_2}.$$

Then assuming the normalization condition $\|u_i\|_2 = \|v_i\|_2 = 1$ for each $i = 1, \ldots, m$, it is not hard to show that if we define the model

$$f_{(x,y)}(x + \Delta_x, y + \Delta_y) := \left\| b - \mathcal{A}(xy^T) - \mathbf{diag}(Vy)U\Delta_x - \mathbf{diag}(Ux)V\Delta_y \right\|_1,$$

which is centered at (x, y) and has variables $\Delta_x \in \mathbf{R}^{n_1}$ and $\Delta_y \in \mathbf{R}^{n_2}$, then

$$f(x + \Delta_x, y + \Delta_y) \le f_{(x,y)}(x + \Delta_x, y + \Delta_y) + \frac{1}{2}\Delta_x^T U^T U \Delta_x + \frac{1}{2}\Delta_y^T V^T V \Delta_y \tag{2}$$

for any $\Delta_x \in \mathbf{R}^{n_1}, \Delta_y \in \mathbf{R}^{n_2}$.

(b) Using the bound (2), argue that from any initial points (x^0, y^0), the iteration

$$(\Delta x^k, \Delta y^k) := \operatorname*{argmin}_{\Delta_x, \Delta_y} \left\{ f_{(x^k,y^k)}(x^k + \Delta_x, y^k + \Delta_y) + \frac{1}{2}\Delta_x^T U^T U \Delta_x + \frac{1}{2}\Delta_y^T V^T V \Delta_y \right\}$$
$$(x^{k+1}, y^{k+1}) := (x^k + \Delta x^k, y^k + \Delta y^k)$$

is a descent (or at least, non-ascent) method, that is, $f(x^{k+1}, y^{k+1}) \le f(x^k, y^k)$.

(c) Using the data in the file `matrix_sco_data.*`, which defines a U matrix, V matrix, and b vector, implement the procedure from part (b). Have your procedure iterate until the change between iterations satisfies

$$\left\|\Delta x^k\right\|_2^2 + \left\|\Delta y^k\right\|_2^2 \le \epsilon^2 \quad \text{where} \quad \epsilon = 10^{-4}.$$

Run your procedure on 10 different random initializations x^0, y^0, choosing the entries of each vector to be independent $\mathcal{N}(0, 1)$ random variables. Plot the error $f(x^k, y^k)$ on a semilog plot (logarithmic vertical axis) over iterations for each of the runs of your method. (All 10 runs should be in the same plot.)