EECS208 Written HW4

Issued: Nov. 4. **Due:** Nov. 14, 11:59 PM via Gradescope

Reading: Chapters 7 of *High-Dim Data Analysis with Low-Dim Models*.

Problem 1 (Complete Dictionary Learning via ℓ^4 Norm Maximization)

Exercise 7.2 of High-Dim Data Analysis with Low-Dim Models. In this exercise, we derive and practice an algorithm to solve ℓ^4 norm maximization problem (7.3.14) for complete dictionary learning.

1. Derive the gradient $\varphi(\boldsymbol{A}) = \|\boldsymbol{A}\boldsymbol{Y}\|_4^4$ with respect to $\boldsymbol{A}$.
2. Derive a project gradient ascent for maximizing $\varphi(\boldsymbol{A})$:
$$\boldsymbol{A}_{k+1} = \mathcal{P}_{\mathsf{O}(n)}[\boldsymbol{A}_k + \gamma \cdot \nabla\varphi(\boldsymbol{A}_k)]. \tag{0.1}$$
Hint: you may directly apply the following claim (without proof):

   Claim 0.1 (Projection onto Orthogonal Group) *$\forall \boldsymbol{A} \in \mathbb{R}^{n\times n}$, the orthogonal matrix which has minimum Frobenius norm with $\boldsymbol{A}$ is the following*
$$\mathcal{P}_{\mathsf{O}(n;\mathbb{R})}(\boldsymbol{A}) = \operatorname*{arg\,min}_{\boldsymbol{M}\in\mathsf{O}(n;\mathbb{R})} \|\boldsymbol{M} - \boldsymbol{A}\|_F^2 = \boldsymbol{U}\boldsymbol{V}^\top, \tag{0.2}$$
where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \mathsf{SVD}(\boldsymbol{A})$.

   If you can prove Claim 0.1, you will **receive extra credits** for this problem.
3. Let $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ and suppose each entry of $\boldsymbol{Y}$ is iid drawn from a standard Gaussian distribution $\mathcal{N}(0,1)$ with probability θ, and equals to 0 with probability $1-\theta$:
$$\bar{y}_{i,j} = \begin{cases} z \sim \mathcal{N}(0,1) & \text{with probability } \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{0.3}$$
and $\boldsymbol{A}_0$ is a randomly initialized orthogonal matrix. Conduct simulation of the algorithm and play with different step size γ of the gradient ascent, what is the relationship between the step size γ and the convergence speed? What happens if you make the step size to be infinite? That is,
$$\boldsymbol{A}_{k+1} = \mathcal{P}_{\mathsf{O}(n)}[\nabla\varphi(\boldsymbol{A}_k)]. \tag{0.4}$$

Problem 2 (Low-rank Regularization via the $\log\det(\cdot)$ Function)

Exercise 7.4 of High-Dim Data Analysis with Low-Dim Models. When a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ is symmetric and positive semi-definite, the nuclear norm $\|\boldsymbol{X}\|_*$ is the same as its trace of the matrix. In this exercise, we try to study the connection of the convex nuclear norm (or the trace norm) with another popular smooth but nonconvex surrogate for minimizing $\mathsf{rank}(\boldsymbol{X})$ is to minimize the quantity
$$\min_{\boldsymbol{X}\in\mathsf{C}} f(\boldsymbol{X}) \doteq \log\det(\boldsymbol{X} + \delta\boldsymbol{I}), \tag{0.5}$$
where $\delta > 0$ is a small regularization constant and $\boldsymbol{X}$ belongs to some constraint set C. To see how this objective is related to the trace norm:

1. First, show that $\nabla_{\boldsymbol{X}} f(\boldsymbol{X}) = (\boldsymbol{X} + \delta \boldsymbol{I})^{-1}$.

2. Second, the first-order expansion of $f(\boldsymbol{X})$ around a point $\boldsymbol{X}_k$ is given by:

$$f(\boldsymbol{X}) \approx f(\boldsymbol{X}_k) + \operatorname{tr}\left((\boldsymbol{X}_k + \delta \boldsymbol{I})^{-1}(\boldsymbol{X} - \boldsymbol{X}_k)\right) + o(\|\boldsymbol{X} - \boldsymbol{X}_k\|). \tag{0.6}$$

Then to minimize $f(\boldsymbol{X})$, we can use a greedy descent algorithm with the iteration

$$\boldsymbol{X}_{k+1} = \arg\min_{\boldsymbol{X} \in \mathsf{C}} \operatorname{tr}\left((\boldsymbol{X}_k + \delta \boldsymbol{I})^{-1} \boldsymbol{X}\right). \tag{0.7}$$

Argue that when $\boldsymbol{X}_k$ is initialized around $\boldsymbol{X}_o = \boldsymbol{I}$, then the above iteration becomes minimizing the trace norm $\boldsymbol{X}_{k+1} = \arg\min_{\boldsymbol{X} \in \mathsf{C}} \operatorname{tr}(\boldsymbol{X})$.