Homework 3, due Monday November 9

**Problem 1.** Let $M \in \mathbb{R}^{d \times d}$ be a symmetric psd matrix with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ and corresponding orthonormal eigenvectors $v_1, v_2, \ldots, v_d$. For any positive integer $p \in \mathbb{N}$, the matrix power $M^p$ is equal to $\sum_{i=1}^{d} \lambda_i^p v_i v_i^\top$, which is close to $\lambda_1^p v_1 v_1^\top$ when $p$ is large enough. How large is “large enough”? We need to be able to identify a column of $M^p$ that is a non-zero scaling of $v_1$, so that if the column is rescaled to have unit length, we approximately recover $v_1$ (or $-v_1$). The column we’ll take is $M^p e_{j^*}$, where

$$j^* \in \arg \max_{j \in [d]} \| M^p e_j \|_2.$$

Let $q := M^p e_{j^*} / \| M^p e_{j^*} \|_2$. State a lower bound on $(\langle v_1, q \rangle)^2$ in terms of $p$, $d$, and $\gamma := 1 - \lambda_2 / \lambda_1$; and prove the correctness of the bound.

*Hint:* If you like, you can assume that $p > \log(2d)/(2\gamma)$.

**Problem 2.** Let $A$ be any matrix whose singular values are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

(a) Prove that for any $k \in [r]$, $\sigma_k \leq \| A \|_F / \sqrt{k}$.

(b) Prove that for any $k \in [r]$, there exists a matrix $B$ with rank($B$) $\leq k$ satisfying $\| A - B \|_2 \leq \| A \|_F / \sqrt{k}$.

**Problem 3.** Suppose $A, B \in \mathbb{R}^{n \times d}$ each have rank $d$. Specify an algorithm (that can use things like SVD, Gram-Schmidt, etc. as black-box subroutines) that find all solutions $v \in S^{d-1}$ satisfying

$$\exists \lambda \in \mathbb{R} \setminus \{ 0 \} \text{ s.t. } A^\top A v = \lambda B^\top B v.$$

(If there is an entire subspace of solutions, you just need to return an orthonormal basis for this subspace.) Explain why the algorithm is correct.

**Problem 4.** Let $A \in \mathbb{R}^{n \times d}$ be a data matrix whose rows are $a_1, a_2, \ldots, a_n \in \mathbb{R}^d$. Let $D \in \mathbb{R}^{n \times n}$ be the matrix whose $(i, j)$-th entry is the squared Euclidean distance $D_{i,j} = \| a_i - a_j \|_2^2$. Suppose you are given the squared Euclidean distance matrix $D$ as input, and you are asked to recover the set of original points $\{ a_1, a_2, \ldots, a_n \}$. You do not have access to the original data matrix $A$.

(a) Explain why it is not generally possible to accomplish this task exactly.

(Why is the solution $\{ a_1, a_2, \ldots, a_n \}$ not unique?)

(b) Let $s \in \mathbb{R}^n$ be the vector whose $i$-th entry is $\| a_i \|_2^2$. Prove that $D = s 1_n^\top - 2 A A^\top + 1_n s^\top$, where $1_n \in \mathbb{R}^n$ is the all-ones vector.

(c) Let $\Pi \in \mathbb{R}^{n \times n}$ be the orthogonal projector for the $(n-1)$-dimensional subspace $\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 0 \}$. Prove that $- (1/2) \Pi D \Pi = \Pi A A^\top \Pi$.

*Hint:* Explain why $\Pi 1_n = 0$.

(d) Explain how to determine a set of points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ from $D$ such that:

- $D_{i,j} = \| x_i - x_j \|_2^2$ for all $i, j \in [n]$; and
- $\sum_{i=1}^{n} x_i = 0$.

(You may assume that you are told the original dimension $d$.)
Problem 5. Let $S \subset \Delta^{d-1}$ be data points, each of which represents a probability distribution over $[d]$. For instance, each $x_i \in S$ could be a bag-of-words histogram representation of a document (where the vocabulary is identified with $[d]$). We would like to find a small set of representatives $C \subset \Delta^{d-1}$, so as to minimize

$$
cost_{\text{RE}}(S, C) := \sum_{x \in S} \min_{y \in C} \text{RE}(x, y),$$

where $\text{RE}(x, y) := \sum_{i=1}^{d} x_i \ln(x_i/y_i)$ is the relative entropy of $x$ from $y$. It turns out Lloyd’s algorithm can be generalized to accommodate this objective.

(a) Let $f(x) := \sum_{i=1}^{d} x_i \ln x_i$, and let $a_y(x) := f(y) + \langle \nabla f(y), x - y \rangle$ be the affine approximation of $f(x)$ around $y$. Show that $\text{RE}(x, y) = f(x) - a_y(x)$ and that $\text{RE}(x, y) \geq 0$.

(You may assume that $f$ is differentiable at $y$, and also use the fact that $f$ is convex.)

(b) Prove that for any random vector $X$ in $\Delta^{d-1}$ and any other vector $y \in \Delta^{d-1}$ (where, again, you may assume that $f$ is differentiable at $y$),

$$
\mathbb{E} \text{RE}(X, y) = \mathbb{E} \text{RE}(X, \mu) + \text{RE}(\mu, y) \geq \mathbb{E} \text{RE}(X, \mu)
$$

where $\mu := \mathbb{E}(X)$.

Hint: Use the result from (a) and linearity of expectation.

(c) Derive a variant of Lloyd’s algorithm for finding a local minimizer of $\text{cost}_{\text{RE}}(S, \cdot)$ (over $C \subset \Delta^{d-1}$ with $|C| = k$), and explain why the objective monotonically decreases with the iteration number until convergence.