**Homework 7 - due Wed. Apr. 1st**
**High-Dimensional Approximation, Probability, and Statistical Learning**

**Instructor**: Mauro Maggioni
**Office**: 302D Whitehead Hall
**Web page**: https://mauromaggioni.duckdns.org
**E-mail**: mauro.maggioni *at* youknowwhat.edu

**Homework Policies**. Unlike other homework sets, for this set you will work by yourself with no help from anyone else. It is open book, open notes. This homework will count as two homework sets. Make sure your answers are detailed and well-motivated. The code and corresponding plots should be submitted together with everything else. Submission is electronic via e-mail to me, with scanned/imaged pages in one .pdf (or .zip file if multiple files are included) with name 553.738_FirstLastName.pdf, sent to me in an e-mail with the title "Homework 7 for 553.738". Thank you.

# Assignment

### Exercises

*Exercise* 1 (25pt). Suppose you have $n$ samples $x_1, \ldots, x_n$ i.i.d. from $\mathcal{N}(0, I_D)$. Consider the empirical mean $m_n$: how many samples are needed so that with high probability $||m_n - 0||_\infty < \epsilon$? How many samples are needed so that with high probability $||m_n - 0||_2 < \epsilon$? What if instead of $\mathcal{N}(0, I_D)$ the distribution was $\mathcal{N}(0, \Sigma)$ with $\Sigma$ of rank $r$?

*Exercise* 2 (25pt). Let $A \in \mathbb{R}^{m\times m}$ and assume we observe $\{A + E_i\}_{i=1}^n$, where the $E_i$'s are i.i.d., each being a matrix with subGaussian i.i.d. entries, with $\mathbb{E}[E_{ij}] = 0$ and $||E_{ij}||_{\psi_2} \leq K$. We are interested in estimating $A$.

a. Let $\hat{A}$ be given by the empirical mean of the observations. What can you say about $\mathbb{E}[||\hat{A} - A||^2]$ and $\mathbb{P}[||\hat{A} - A|| > t]$ as a function of $n$?

b. Suppose we knew that $A$ has only $k$ non-zero entries, and we knew the locations of those entries. What estimator $\hat{A}$ would you use? What can you say about $\mathbb{E}[||\hat{A} - A||^2]$ and $\mathbb{P}[||\hat{A} - A|| > t]$ as a function of $n$? What if you did not know a priori which $k$ entries were non-zero?

c. Suppose we knew that $A$ is symmetric and of rank $k$. What estimator $\hat{A}$ would you use? What can you say about the angle between the range of $A$ and the range of $\hat{A}$?

*Exercise* 3 (25pt). Consider two Gaussian distributions in $\mathbb{R}^D$, $\mathcal{N}(0, I_D)$ and $\mathcal{N}((m, 0, \ldots, 0), I_D)$. Sample points $x_1, \ldots, x_n$ i.i.d. uniformly at random from the first Gaussian, and $n$ points $y_1, \ldots, y_n$ i.i.d. from the second Gaussian. Estimate the distance between pairs of $x_i$'s, and then between $x_i$'s and the $y_j$'s, both in expectation and with high probability. Discuss how your bounds behave as $m$ and $D$ change.

*Exercise* 4 (25pt). Suppose $M$ is a subset of $\mathbb{S}^{D-1}$ such that for every $\epsilon > 0$ there is a net of cardinality $\lesssim \epsilon^{-d}$, for $d \leq D$. For example $M$ could be a smooth compact $d$-dimensional manifold (a notion that generalizes 2-dimensional surfaces in $\mathbb{R}^3$ to $d$-dimensional smooth objects in $\mathbb{R}^D$). Let $x_1, \ldots, x_n$ be $n$ points in $M$. Use Johnson-Lindenstrauss' Lemma to construct a projection $P$ of the data onto $d'$ dimensions that almost-preserves Euclidean distances (up to a scaling factor), for some $d'$ as in the standard statement of the J.L.-Lemma. Then try to find the smallest $d'$ such that $C^{-1}(1-\epsilon)||x-y||_2 \leq ||Px - Py|| \leq C(1+\epsilon)||x-y||_2$, or $C(||x-y||_2 - \epsilon) \leq ||Px - Py|| \leq C(||x-y||_2 + \epsilon)$. Can you choose $d'$ independent of $n$?