Problem 1. Consider the parametric family of probability densities $\mathcal{P} := \{p_{\mu, \sigma^2} : \mu \in \mathbb{R}^d, \sigma^2 > 0\}$ over $\mathbb{R}^d$, where $p_{\mu, \sigma^2}$ is the density for $N(\mu, \sigma^2 I)$. Specify the maximum likelihood estimator (MLE) for the parameters $(\mu, \sigma^2)$ based on a data set $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$. Also give a precise condition on the data set under which the MLE produces legal parameter estimates (in particular, a necessary and sufficient condition on the data set under which the MLE for $\sigma^2$ is strictly positive).

Problem 2. Prove that there exists an absolute constant $C > 0$ such that for any $d, k, s \in \mathbb{N}$ where $s$ divides $k$, $k \geq C$, and $s > \sqrt{6k \ln(d)}$, there exists vectors $c_1, c_2, \ldots, c_d \in \{0, 1\}^k$ satisfying both of the following properties.

1. Each $c_i$ has exactly $s$ non-zero entries.

2. $\langle c_i, c_j \rangle \leq 2s^2$ for all $1 \leq i < j \leq d$.

Hint: It is possible to choose the $c_i$ so that each vector $c_i$ can be thought of as being comprised of $s$ blocks of coordinates, with each block containing $k/s$ coordinates, and there is only a single non-zero entry within each block.

Problem 3. Let $X \sim N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ (for some $d \gg 2$) has exactly two non-zero eigenvalues, both equal to one. Consider a random two-dimensional subspace $T \subseteq \mathbb{R}^d$ produced by choosing a unit vector $U$ uniformly at random from $S^{d-1}$, then choosing a second unit vector $U'$ uniformly at random among unit vectors orthogonal to $U$, and finally setting $T := \text{span}\{U, U'\}$. Also let $W \subseteq \mathbb{R}^d$ be the top two-dimensional PCA subspace for $X$. What is the expected squared-length of the orthogonal projection of $X$ to $T$? And what is the expected squared-length of the orthogonal projection of $X$ to $W$?

Problem 4. This problem is about $k$-means clustering and low-dimensional linear maps.

For any $k, n \in \mathbb{N}$, a $k$-partitioning of $[n] := \{1, 2, \ldots, n\}$ is a collection of subsets $C_1, C_2, \ldots, C_k \subseteq [n]$ such that $C_i \cap C_j = \emptyset$ and $\bigcup_{i=1}^k C_i = [n]$. Let $\Pi(k, n)$ denote the set of all $k$-partitionings of $[n]$.

For any $d, k, n \in \mathbb{N}$, any $(C_1, C_2, \ldots, C_k) \in \Pi(k, n)$, and any data set $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$, define

$$\text{cost}(x_1, x_2, \ldots, x_n; C_1, C_2, \ldots, C_k) := \sum_{i=1}^k \min_{c_i \in \mathbb{R}^d} \sum_{j \in C_i} \|x_j - c_i\|_2^2.$$ 

The $k$-means clustering problem concerns finding $(C_1, C_2, \ldots, C_k) \in \Pi(k, n)$ so as to minimize $\text{cost}(x_1, x_2, \ldots, x_n; C_1, C_2, \ldots, C_k)$.

(a) Prove that for any iid random vectors $X$ and $Y$, and any other vector $c$,

$$\mathbb{E}\|X - c\|_2^2 = \mathbb{E}\|X - \mathbb{E}X\|_2^2 + \|c - \mathbb{E}X\|_2^2 = \frac{1}{2} \mathbb{E}\|X - Y\|_2^2 + \|c - \mathbb{E}X\|_2^2.$$ 

Use this result to prove that for any $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$,

$$\min_{c \in \mathbb{R}^d} \sum_{i=1}^m \|x_i - c\|_2^2 = \sum_{i=1}^m \|x_i - \mu\|_2^2 = \frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^m \|x_i - x_j\|_2^2,$$

where $\mu := \frac{1}{m} \sum_{i=1}^m x_i$. 

1
(b) Prove that for any \( \varepsilon \in (0, 1/2) \) and any data set \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d \), there exists a positive integer \( p \leq 100 \ln(n)/\varepsilon^2 \) and a matrix \( \mathbf{M} \in \mathbb{R}^{p \times d} \) such that for any \( k \in \mathbb{N} \) and any \( (C_1, C_2, \ldots, C_k) \in \Pi(k,n) \),

\[
(1 - \varepsilon) \text{cost}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n; C_1, C_2, \ldots, C_k) \\
\leq \text{cost}(\mathbf{M} \mathbf{x}_1, \mathbf{M} \mathbf{x}_2, \ldots, \mathbf{M} \mathbf{x}_n; C_1, C_2, \ldots, C_k) \\
\leq (1 + \varepsilon) \text{cost}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n; C_1, C_2, \ldots, C_k).
\]

(c) Use the result of (b) to prove that for any \( \varepsilon \in (0, 1/2) \) and data set \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d \), if

(i) \( \mathbf{M} \in \mathbb{R}^{p \times d} \) satisfies the property from (b) for the given value of \( \varepsilon \) and data set \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \);

(ii) \( (\tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_k) \in \arg\min \text{cost}(\mathbf{M} \mathbf{x}_1, \mathbf{M} \mathbf{x}_2, \ldots, \mathbf{M} \mathbf{x}_n; C_1, C_2, \ldots, C_k) \);

then

\[
\text{cost}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n; \tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_k) \leq \frac{1 + \varepsilon}{1 - \varepsilon} \min_{(C_1, C_2, \ldots, C_k) \in \Pi(k,n)} \text{cost}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n; C_1, C_2, \ldots, C_k).
\]

**Problem 5.** Let \( \Phi : \mathbb{R} \to [0, 1] \) denote the cumulative distribution function for \( N(0, 1) \), i.e., \( \Phi(t) = \mathbb{P}(Z \leq t) \) where \( Z \sim N(0, 1) \). Prove the Diaconis-Freedman result on projection pursuit: specifically, prove that there are absolute positive constants \( c_0, C_1, c_2, C_3, c_4 > 0 \) such that for any \( d \in \mathbb{N} \) with \( d^c_0 \geq \ln(C_1)/c_2 \), if

1. \( X_1, X_2, \ldots, X_d \) are independent random variables;
2. \( \mathbb{E} X_i = 0 \) and \( \mathbb{E} X_i^2 = 1 \) for all \( i \in [d] \);

then for a \( 1 - C_1 \exp(-c_2d^{c_0}) \) fraction of unit vectors \( \mathbf{u} \in S^{d-1} \), the random vector \( \mathbf{X} = (X_1, X_2, \ldots, X_d) \) satisfies

\[
\max_{t \in \mathbb{R}} \left| \mathbb{P}(\langle \mathbf{u}, \mathbf{X} \rangle \leq t) - \Phi(t) \right| \leq \frac{C_3 \rho}{d^{c_4}},
\]

where \( \rho := \max_{i \in [d]} \mathbb{E} |X_i|^3 \).

You may/should use the following version of the central limit theorem:

**Theorem 1.** There is an absolute positive constant \( C > 0 \) such that the following holds. Let \( Y_1, Y_2, \ldots, Y_n \) be independent random variables with \( \mathbb{E} Y_i = 0 \), \( \sigma_i^2 := \mathbb{E} Y_i^2 < \infty \). Define \( v_n := \sum_{i=1}^n \sigma_i^2 \) and \( \rho_n := \sum_{i=1}^n \mathbb{E} |Y_i|^3 \). Then

\[
\max_{t \in \mathbb{R}} \left| \mathbb{P} \left( \frac{Y_1 + Y_2 + \cdots + Y_n}{\sqrt{v_n}} \leq t \right) - \Phi(t) \right| \leq \frac{C \rho_n}{v_n^{3/2}}.
\]

You don’t need to prove Theorem 1.