Instructions:
1. This question paper contains 2 pages (4 sides of paper). Please verify.
2. Write your name, roll number, department in **block letters** with **ink** on **each page**.
3. Write your final answers neatly **with a blue/black pen**. Pencil marks may get smudged.
4. Don't overwrite/scratch answers especially in MCQ – ambiguous cases may get 0 marks.

Q1. (**True-False**) Write **T** or **F** for True/False (write **only in the box on the right-hand side**). You must also give a brief justification for your reply in the space provided below.**(3 x (1+2) = 9 marks)**

1	EM run on data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^2$ s.t. $\|\mathbf{x}_i\|_2 \leq 2$ for all $i \in [N]$ to learn mixture of two Gaussians $\mathcal{N}(\boldsymbol{\mu}_k, I), k \in [2]$ will always ensure that the means satisfy $\|\boldsymbol{\mu}_k\|_2 \leq 2$.	T

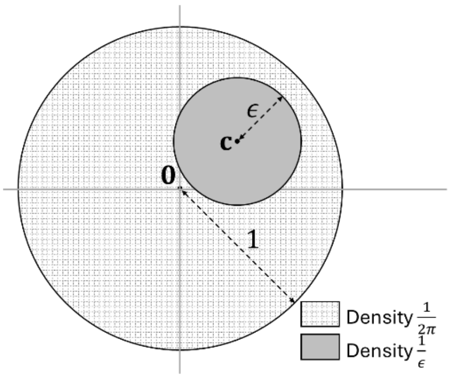
In any iteration of the EM algorithm, the means $\boldsymbol{\mu}_k$ are updated as $\boldsymbol{\mu}^c = \sum_{i \in [N]} \eta_{ic} \cdot \mathbf{x}^i$ where $\eta_{ic} \overset{\text{def}}{=} \frac{q_c^i}{\sum_{i \in [N]} q_c^j}$ i.e. $\sum_{i \in [N]} \eta_{ic} = 1$ i.e., a weighted average of the points is used to update the mean. However, convex sets $\mathcal{C}$ satisfy the property that if $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, then $\eta \cdot \mathbf{x} + (1 - \eta) \cdot \mathbf{y} \in \mathcal{C}$ as well for any $\eta \in [0,1]$. Since the set $\mathcal{B}_2(0,2) \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 \leq 2\}$ is convex, the result follows.

2	The difference of two Mercer kernels can never be Mercer. If True, give a proof. If False, construct two Mercer kernels K_1, K_2 with maps ϕ_1, ϕ_2 s.t. the difference $K_3 \overset{\text{def}}{=} K_1 - K_2$ is also a Mercer kernel with map ϕ_3. Give maps ϕ_1, ϕ_2, ϕ_3 explicitly.	F

Let $K_1, K_2 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be defined as $K_1(x, y) \overset{\text{def}}{=} 25xy$ and $K_2(x, y) \overset{\text{def}}{=} 16xy$. The corresponding feature maps are $\phi_1(x) = [5x]$ and $\phi_2(x) = [4x]$. Note the feature maps are unidimensional. We have $K_3(x, y) = 9xy$ for which the feature map $\phi_3(x) = [3x]$ works.

3	For convex differentiable $f : \mathbb{R} \to \mathbb{R}$, if $f\left(\frac{x+y}{2}\right) > 1$ for some $x, y \in \mathbb{R}$, then we must have $\max\{f(x), f(y)\} > 1$. Justify either using a proof or counter example.	T

Convex functions satisfy $f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}$. If $f(x) \leq 1$ as well as $f(y) \leq 1$ then we will have $f\left(\frac{x+y}{2}\right) \leq \frac{1+1}{2}$ i.e. $f\left(\frac{x+y}{2}\right) \leq 1$ which is a contradiction. Thus, at least one of $f(x)$ or $f(y)$ must be strictly greater than 1 which implies that $\max\{f(x), f(y)\} > 1$.



Q2 (Almost Uniform) Melbo is constructing a distribution $\mathcal{D}$ with support over 2D vectors of length up to 1 i.e. $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 \le 1\}$. $\mathcal{D}$ has two parameters $\mathbf{c} \in \mathbb{R}^2, \epsilon \in [0,1]$ and assigns a *high* density $\frac{1}{\epsilon}$ in a "dense ball" of radius ϵ centered at $\mathbf{c}$ i.e., in $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 \le 1, \|\mathbf{x}-\mathbf{c}\|_2 \le \epsilon\}$ and a *low* density of $\frac{1}{2\pi}$ in the rest of the support i.e., in $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 \le 1, \|\mathbf{x}-\mathbf{c}\|_2 > \epsilon\}$. We have $\|\mathbf{c}\|_2 \le 1-\epsilon$ i.e., the dense ball stays within the support.

a. For which values of ϵ will $\mathcal{D}$ be a proper distribution? Find them and show calculations. You may find the fact that $\pi - \sqrt{\pi^2 - 2} \in [0,1]$ and $\pi - \sqrt{\pi^2-1} \in [0,1]$ to be useful.
b. Find out the mean vector $\boldsymbol{\mu} \in \mathbb{R}^2$ of this distribution. Show calculations. **(5 + 7 = 12 marks)**

Hint: the mean of a uniform distribution over a circle is its centre.

Find value(s) of ϵ for which $\mathcal{D}$ is a proper distribution.

Distributions are normalized i.e., $\frac{1}{\epsilon}\cdot\pi\epsilon^2 + \frac{1}{2\pi}\cdot(\pi - \pi\epsilon^2) = 1$ i.e. $\epsilon^2 - 2\pi\epsilon + 1 = 0$. Solving the quadratic gives us the candidate values as $\pi \pm \sqrt{\pi^2-1}$. However, the larger root would result in $\epsilon = \pi + \sqrt{\pi^2-1} > \pi > 1$ which is absurd since that would in-turn force $\|\mathbf{c}\|_2 \le 1-\epsilon < 0$. Thus, the only value ϵ can take is $\pi - \sqrt{\pi^2-1}$. Note that this value satisfies $\epsilon \in [0,1]$ using the fact provided in the question statement.

Find out the mean vector of the distribution $\mathcal{D}$.

Let $\mathcal{U}$ denote the unit ball $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 \le 1\}$ and $\mathcal{H}$ be the dense ball $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}-\mathbf{c}\|_2 \le \epsilon\}$. We have $\boldsymbol{\mu} = \int_{\mathcal{U}} \mathbf{x}\cdot\mathcal{D}(\mathbf{x})\, d\mathbf{x} = \underbrace{\int_{\mathcal{H}} \mathbf{x}\cdot\mathcal{D}(\mathbf{x})\, d\mathbf{x}}_{(A)} + \underbrace{\int_{\mathcal{U}\setminus\mathcal{H}} \mathbf{x}\cdot\mathcal{D}(\mathbf{x})\, d\mathbf{x}}_{(B)}$.

$(A) = \frac{1}{\epsilon}\int_{\mathcal{H}} \mathbf{x}\, d\mathbf{x}$. Now $\int_{\mathcal{H}} \mathbf{x}\, d\mathbf{x} = \pi\epsilon^2 \cdot \int_{\mathcal{H}} \mathbf{x}\cdot\mathcal{P}(\mathbf{x})\, d\mathbf{x}$ where $\mathcal{P}(\mathbf{x}) = \frac{1}{\pi\epsilon^2}$ is the (conditional) uniform distribution inside the heavy ball. As the mean of a uniform distribution over a circle is its centre, we have $\int_{\mathcal{H}} \mathbf{x}\cdot\mathcal{P}(\mathbf{x})\, d\mathbf{x} = \mathbf{c}$ which gives us $(A) = \frac{1}{\epsilon}\cdot\pi\epsilon^2\cdot\mathbf{c} = \pi\epsilon\cdot\mathbf{c}$.

$(B) = \frac{1}{2\pi}\int_{\mathcal{U}\setminus\mathcal{H}} \mathbf{x}\, d\mathbf{x} = \frac{1}{2\pi}\left(\underbrace{\int_{\mathcal{U}} \mathbf{x}\, d\mathbf{x}}_{(C)} - \underbrace{\int_{\mathcal{H}} \mathbf{x}\, d\mathbf{x}}_{(D)}\right)$. Using the same argument as above, we get $(C) = \pi 1^2 \cdot \mathbf{0}$ and $(D) = \pi\epsilon^2\cdot\mathbf{c}$ which gives us $(B) = -\frac{\epsilon^2}{2}\cdot\mathbf{c}$ giving us $\boldsymbol{\mu} = \left(\pi\epsilon - \frac{\epsilon^2}{2}\right)\cdot\mathbf{c}$.
However, recall that ϵ satisfies $\epsilon^2 - 2\pi\epsilon + 1 = 0$ which means $\boldsymbol{\mu} = \frac{1}{2}\cdot\mathbf{c}$.

Can you simplify these calculations? What if the *low* density is some general value $p_l \ne \frac{1}{2\pi}$?

Q3 (Positive Linear Regression) We have data features $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ and labels $y_1, \ldots, y_N \in \mathbb{R}$ stylized as $X \in \mathbb{R}^{N \times D}, \mathbf{y} \in \mathbb{R}^N$. We wish to fit a linear model with positive coefficients:

$$\underset{\mathbf{w}\in\mathbb{R}^D,}{\arg\min} \frac{1}{2}\|X\mathbf{w} - \mathbf{y}\|_2^2 \text{ s.t. } w_j \geq 0 \text{ for all } j \in [D]$$

1. Write the Lagrangian for this problem by introducing dual variables (no derivation needed).
2. Simplify the dual problem (eliminate $\mathbf{w}$) – show major steps. Assume $X^\top X$ is invertible.
3. Give a coordinate descent/ascent algorithm to solve the dual. **(2 + 4 + 6 = 12 marks)**

Write down the Lagrangian here (you will need to introduce dual variables and give them names)

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|X\mathbf{w} - \mathbf{y}\|_2^2 - \boldsymbol{\alpha}^\top \mathbf{w}$$

which can be rewritten for convenience as

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^\top X^\top X \mathbf{w} - \mathbf{w}^\top X^\top \mathbf{y} - \mathbf{w}^\top \boldsymbol{\alpha} + \frac{1}{2}\|\mathbf{y}\|_2^2$$

Derive and simplify the dual. Show major calculations steps.

The dual is $\max_{\boldsymbol{\alpha}\geq\mathbf{0}} \left\{\min_{\mathbf{w}}\{\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha})\}\right\}$. Solving the inner problem by applying first-order optimality (since it is an unconstrained problem) gives us $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow X^\top(X\mathbf{w} - \mathbf{y}) - \boldsymbol{\alpha} = \mathbf{0}$. Putting this in the Lagrangian and neglecting constant terms gives us

$$\min_{\boldsymbol{\alpha}\geq\mathbf{0}} \left\{\frac{1}{2}\boldsymbol{\alpha}^\top C \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{s}\right\}$$

where $C = [c_{ij}] \stackrel{\text{def}}{=} (X^\top X)^{-1} \in \mathbb{R}^{D\times D}$ and $\mathbf{s} = [s_i] \stackrel{\text{def}}{=} C X^\top \mathbf{y} \in \mathbb{R}^D$.

Give a coordinate descent/ascent algorithm to solve the dual problem.

Consider a single coordinate of the dual variable, say α_i (the coordinate may have been chosen cyclically or via random permutation, etc. The optimization problem restricted to α_i is

$$\min_{\alpha_i \geq 0} \frac{1}{2} c_{ii} \alpha_i^2 + \alpha_i \left(s_i + \sum_{j \neq i} c_{ij} \alpha_j \right)$$

Using the QUIN trick tells us that the optimal value is $\max \left\{ 0, -\frac{1}{c_{ii}} \left(s_i + \sum_{j \neq i} c_{ij} \alpha_j \right) \right\}$

Q4. (Kernel Smash) $K_1, K_2, K_3 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are Mercer kernels i.e., for any $x, y \in \mathbb{R}$, we have $K_i(x, y) = \langle \phi_i(x), \phi_i(y) \rangle$ with $\phi_1(x) = (1, x), \phi_2(x) = (x, x^2), \phi_3(x) = (x^2, x^4, x^6)$. Design a map $\phi_4 : \mathbb{R} \to \mathbb{R}^7$ for kernel K_4 s.t. $K_4(x, y) = \left(K_1(x, y) - K_2(x, y) \right)^2 + 3K_3(x, y)$ for all $x, y \in \mathbb{R}$. No derivation needed. **Note that ϕ_4 must not use more than 7 dimensions. If your solution does not require 7 dimensions then leave the rest of the dimensions blank or fill with zero. (7 marks)**

$$\phi_4(x) =$$

$$\left(\boxed{1}, \boxed{x^2}, \boxed{2x^4}, \boxed{x^6\sqrt{3}}, \boxed{0}, \boxed{0}, \boxed{0} \right)$$