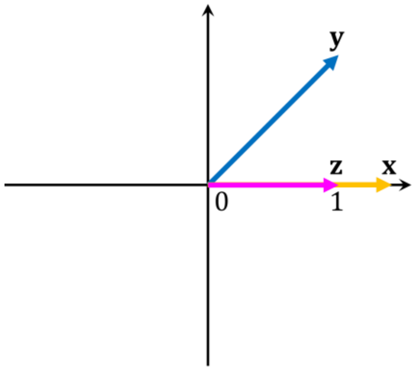
CS 771A: Intro to Machine Learning, IIT Kanpur				**Midsem Exam** (26 Feb 2023)
Name	MELBO			**40 marks**
Roll No	230001	Dept.	AWSM	Page **1** of **4**

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 will get 0 marks.

Q1. Write **T** or **F** for True/False **in the box**. Also, **give justification**. **(4 x (1+2) = 12 marks)**

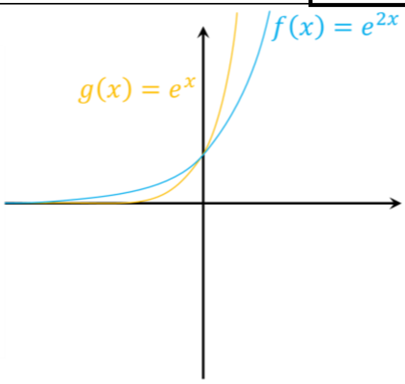
1	For $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$ s.t. $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = \sqrt{2}$, $\|\mathbf{z}\|_2 = 1$ and $\mathbf{x}^\top\mathbf{y} \geq \mathbf{x}^\top\mathbf{z}$, we always have $\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \|\mathbf{x}-\mathbf{z}\|_2^2$. Give a brief proof if True else give a counter example if False.	F

Consider the following counterexample:
$\mathbf{x} = [\sqrt{2}, 0], \mathbf{y} = [1,1], \mathbf{z} = [1,0]$.
We have $\mathbf{x}^\top\mathbf{y} = \sqrt{2} \geq \sqrt{2} = \mathbf{x}^\top\mathbf{z}$ however we also have
$\|\mathbf{x}-\mathbf{y}\|_2^2 = (\sqrt{2}-1)^2 + 1 > (\sqrt{2}-1)^2 = \|\mathbf{x}-\mathbf{z}\|_2^2$



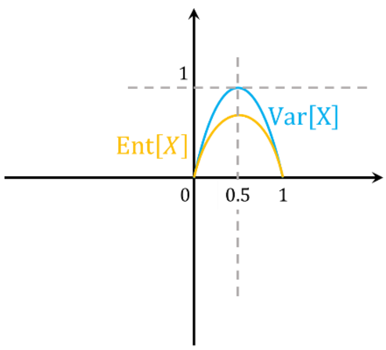
2	Let $f, g: \mathbb{R} \to \mathbb{R}$ be two distinct, non-constant, convex functions i.e., $f \neq g$ and it is not the case that for some $c, d \in \mathbb{R}$, $f(x) = c, g(x) = d$ for all $x \in \mathbb{R}$. Then $h: \mathbb{R} \to \mathbb{R}$ defined as $h(x) \stackrel{\text{def}}{=} f(x)/g(x)$ can never be convex. Give a brief proof if True else if False, give a counter example using two distinct non-constant, convex functions. It is okay to give a counter example where h has isolated, removable discontinuities.	F

Consider the following counterexample:
$f(x) = e^{2x}, g(x) = e^x$.
Both are distinct, non-constant, convex functions.
Note that $f(x) = (g(x))^2$. However,
$f(x)/g(x) = e^x$,
which is a convex function itself.



3	X is a discrete random variable that takes value -1 with probability p and 1 with probability $1-p$. The value of p at which X has maximum entropy is the same as the value of p at which X has maximum variance.	T

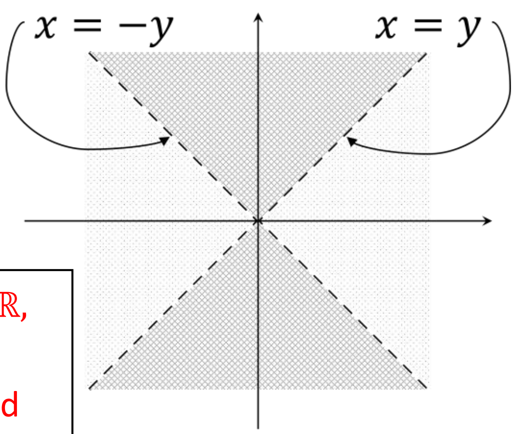
$\mathbb{E}[X] = 1-2p, \mathbb{E}[X^2] = 1$ i.e., $\text{Var}[X] = 4p(1-p)$. Applying FOO and the second-derivative test tells us that the maximum variance is achieved at $p = 1/2$. The entropy of X is defined as
$\text{Ent}[X] = -p\ln p - (1-p)\ln(1-p)$. Applying FOO tells us that entropy is maximized at $1/2$ as well.



4	Y is a Boolean random variable $\mathbb{P}[Y=1]=1/(1+\exp(-t))$. Then Y's entropy is maximized as $t\to\infty$. Justify your answer by giving brief calculations.	F

Let $\mathbb{P}[Y=1]=1/(1+\exp(-t))\overset{\text{def}}{=}p$. As Y is Boolean, this gives us $\mathbb{P}[Y=0]=1-p$. Thus, the entropy of Y is $\text{Ent}[Y]=-p\ln p-(1-p)\ln(1-p)$. The derivative of the entropy is $\ln((1-p)/p)$ which is maximized as $p\to 1/2$. However, as $t\to\infty, p\to 1$ i.e., Y's entropy is not maximized as $t\to\infty$. Note that entropy goes to 0 as $t\to\infty$ or $t\to-\infty$. In fact, entropy is maximized as $t\to 0$.

Q2. (X marks the split) Create a feature map $\phi:\mathbb{R}^2\to\mathbb{R}^D$ for some $D>0$ so that for any $\mathbf{z}=(x,y)\in\mathbb{R}^2$, $\text{sign}\left(\mathbf{1}^\top\phi(\mathbf{z})\right)$ takes value -1 if $\mathbf{z}$ is in the dark cross-hatched region and $+1$ if $\mathbf{z}$ is in the light dotted region (see fig). $\mathbf{1}=(1,1,\dots,1)\in\mathbb{R}^D$ is the D-dimensional all-ones vector. The dashed lines in the fig are $x=y$ and $x=-y$. No derivation needed – just give the final map below. **(3 marks)**



Several solutions are possible e.g., $[x^2,-y^2]\in\mathbb{R}^2$, $[x^2-y^2]\in\mathbb{R}$, $[|x|,-|y|]\in\mathbb{R}^2$, $[|x|-|y|]\in\mathbb{R}$.
Incorrect solutions include $[|xy|,-1]\in\mathbb{R}^2$, $[|xy|,-y^2]\in\mathbb{R}^2$ and $[y^2-x^2,xy]\in\mathbb{R}^2$. Note that all these solutions give a wrong label on the point $(1,0)$. The label should be $+1$ on this point but we have $|xy|-1=|xy|-y^2=y^2-x^2+xy=-1$ for $x=1,y=0$.

Q3. (Maximum stretch) Consider the optimization problem $\min_{\mathbf{x}\in\mathbb{R}^3}\frac{1}{2}\|\mathbf{x}\|_2^2\quad \text{s.t.}\quad \mathbf{c}^\top\mathbf{x}\geq p$ which has a single constraint and $\mathbf{c}\in\mathbb{R}^3$ is a constant vector and $p\in\mathbb{R}$ is a real constant. **(3+2 = 5 marks)**

(a) Give brief derivation solving the problem for $\mathbf{c}=(1,2,3)$ and $p=7$. Write the value of $\mathbf{x}$ at which the optimum is achieved. (*Hint: try orthogonal decomposition or some other trick*)

Decompose $\mathbf{x}=\mathbf{x}_\parallel+\mathbf{x}_\perp$ where $\mathbf{x}_\parallel$ is along $\mathbf{c}$ and $\mathbf{x}_\perp$ is perpendicular to $\mathbf{c}$. Note that $\mathbf{c}^\top\mathbf{x}=\mathbf{c}^\top\mathbf{x}_\parallel$ but by Pythagoras's theorem, $\|\mathbf{x}\|_2^2=\left\|\mathbf{x}_\parallel\right\|_2^2+\|\mathbf{x}_\perp\|_2^2>\left\|\mathbf{x}_\parallel\right\|_2^2$ unless $\|\mathbf{x}_\perp\|_2=0$. This means that having $\mathbf{x}_\parallel\neq\mathbf{0}$ does not contribute to the constraint but increases the objective value. This means that the optimum must be achieved at $\mathbf{x}_\perp=\mathbf{0}$. This means $\mathbf{x}=\lambda\cdot\mathbf{c}$. We want $\mathbf{c}^\top\mathbf{x}\geq p$ i.e., $\lambda\geq p/\|\mathbf{c}\|_2^2=7/14=1/2$. Since we wish to minimize $\frac{1}{2}\|\mathbf{x}\|_2^2$, we choose the smallest value of λ that satisfies the constraint i.e., the optimal value of $\mathbf{x}=(0.5,1,1.5)$

(b) Give brief derivation solving the problem for $\mathbf{c}=(-1,-2,-3)$ and $p=-7$. Write the value of $\mathbf{x}$ at which the optimum is achieved.

The optimal value of $\mathbf{x}=(0,0,0)$. To see this, notice that this value achieves $\frac{1}{2}\|\mathbf{x}\|_2^2=0$ which is the smallest possible value since norms always take non-negative values. Moreover, this also satisfies the constraint since $\mathbf{c}^\top\mathbf{x}=0\geq-7$.

Q4 (Elastic-net regression) Given n pts $(\mathbf{x}^i, y^i)$ $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$, we wish to solve $\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2 + \|\mathbf{w}\|_1 + \frac{1}{2}\sum_{i\in[n]}\left(y^i - \mathbf{w}^\top\mathbf{x}^i\right)^2$. To create its dual, we introduce variables $\mathbf{z} = [z_1, \ldots, z_d] \in \mathbb{R}^d$ and $\mathbf{r} = [r_1, \ldots, r_n] \in \mathbb{R}^n$ to give us the constrained problem in the box on the right. Note that $\mathbf{1} \in \mathbb{R}^d$ is the all-ones vector.

$$\min_{\substack{\mathbf{w},\mathbf{z}\in\mathbb{R}^d\\ \mathbf{r}\in\mathbb{R}^n}} \frac{1}{2}\|\mathbf{w}\|_2^2 + \mathbf{z}^\top\mathbf{1} + \frac{1}{2}\|\mathbf{r}\|_2^2 \quad \text{s.t.}$$
$$w_j - z_j \leq 0 \text{ for all } j \in [d]$$
$$-w_j - z_j \leq 0 \text{ for all } j \in [d]$$
$$y^i - \mathbf{w}^\top\mathbf{x}^i - r_i = 0 \text{ for all } i \in [n]$$

We introduce dual variables α_j for the constraints $w_j - z_j \leq 0$, β_j for $-w_j - z_j \leq 0$ and λ_i for $y^i - \mathbf{w}^\top\mathbf{x}^i - r_i = 0$. For simplicity, we collect the dual variables as vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\lambda} \in \mathbb{R}^n$. For each part, give your answers in the space demarcated for that part. **(3+2+6+5+4=20 marks)**

a. `Fill in the circle indicating the correct constraint for the dual variables $\alpha_j, \beta_j, \lambda_i$. *(3x1 marks)*

$\alpha_j \leq 0$ ○	$\beta_j \leq 0$ ○	$\lambda_i \leq 0$ ○
$\alpha_j \geq 0$ ●	$\beta_j \geq 0$ ●	$\lambda_i \geq 0$ ○
No constraint on α_j ○	No constraint on β_j ○	No constraint on λ_i ●

b. Write down the Lagrangian $\mathcal{L}(\mathbf{w}, \mathbf{z}, \mathbf{r}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda})$ – no derivation needed. *(2 marks)*

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

c. The dual problem is $\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda}}\left\{\min_{\mathbf{w},\mathbf{z},\mathbf{r}} \mathcal{L}(\mathbf{w}, \mathbf{z}, \mathbf{r}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda})\right\}$. To simplify it, solve the 3 inner problems $\min_{\mathbf{w}} \mathcal{L}$, $\min_{\mathbf{z}} \mathcal{L}$ and $\min_{\mathbf{r}} \mathcal{L}$. In each case, give brief derivation and write the expression you get while solving the inner problem (e.g., in CSVM $\min_{\mathbf{w}}\mathcal{L}$ gives $\mathbf{w} = \sum_i \alpha_i y^i \mathbf{x}^i$). *(3x(1+1) marks)*

Expression + derivation for $\min_{\mathbf{w}} \mathcal{L}$.

Applying FOO and setting $\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = \mathbf{0}$ gives us $\mathbf{w} = X^\top\boldsymbol{\lambda} + \boldsymbol{\beta} - \boldsymbol{\alpha}$

Expression + derivation for $\min_{\mathbf{z}} \mathcal{L}$.

The term in the Lagrangian involving $\mathbf{z}$ is $\mathbf{z}^\top(\mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\beta})$ which is linear. The minimization of a linear function always yields $-\infty$ unless the linear function is identically 0. This means that at the optimum, we must have $\boldsymbol{\alpha} + \boldsymbol{\beta} = \mathbf{1}$

Expression + derivation for $\min_{\mathbf{r}} \mathcal{L}$.

Applying FOO and setting $\frac{\partial \mathcal{L}}{\partial \mathbf{r}} = \mathbf{0}$ gives us $\mathbf{r} = \boldsymbol{\lambda}$.

d. Use the expressions obtained above and eliminate $\boldsymbol{\beta}$. Fill in the 5 blank boxes below to show us the simplified dual you get. $X \in \mathbb{R}^{n \times d}$ is the feature matrix with the ith row being $\mathbf{x}^i$. We have turned the max dual problem into a min problem by negating the objective. *(5x1 marks)*

$$\min_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^d \\ \boldsymbol{\lambda} \in \mathbb{R}^n}} \frac{1}{2} \left\| X^\top(\boxed{\boldsymbol{\lambda}}) + (\boxed{\mathbf{1} - 2\boldsymbol{\alpha}}) \right\|_2^2 + \frac{1}{2} \|\boldsymbol{\lambda}\|_2^2 - \boldsymbol{\lambda}^\top (\boxed{\mathbf{y}})$$

s.t.

$\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1}$ ⟸ Write constraint for $\boldsymbol{\alpha}$ here.

No constraint or equivalently $\boldsymbol{\lambda} \in \mathbb{R}^n$ ⟸ Write constraint for $\boldsymbol{\lambda}$ here.

e. For the simplified dual obtained above, let us perform block coordinate minimization.

1. For any fixed value of $\boldsymbol{\alpha} \in \mathbb{R}^d$, obtain the optimal value of $\boldsymbol{\lambda} \in \mathbb{R}^n$.
2. For any fixed value of $\boldsymbol{\lambda} \in \mathbb{R}^n$, obtain the optimal value of $\boldsymbol{\alpha} \in \mathbb{R}^d$.

Note: the optimal value for a variable must satisfy its constraints (if any). Show brief calculations. You may use the QUIN trick and invent shorthand notation to save space e.g., $\mathbf{m} \stackrel{\text{def}}{=} X\boldsymbol{\alpha}$. *(2+2 marks)*

For any fixed value of $\boldsymbol{\alpha} \in \mathbb{R}^d$, obtain the optimal value of $\boldsymbol{\lambda} \in \mathbb{R}^n$**:** Applying FOO (since there are no constraints on $\boldsymbol{\lambda}$) gives us $X(X^\top \boldsymbol{\lambda} + \mathbf{1} - 2\boldsymbol{\alpha}) + \boldsymbol{\lambda} - \mathbf{y} = 0$ i.e.,

$$\boldsymbol{\lambda} = (XX^\top + I_n)^{-1}\left(\mathbf{y} + X(2\boldsymbol{\alpha} - \mathbf{1})\right)$$

where I_n is the $n \times n$ identity matrix.

For any fixed value of $\boldsymbol{\lambda} \in \mathbb{R}^n$, obtain the optimal value of $\boldsymbol{\alpha} \in \mathbb{R}^d$: The optimization problem becomes $\min_{\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1}} \frac{1}{2} \|X^\top \boldsymbol{\lambda} + \mathbf{1} - 2\boldsymbol{\alpha}\|_2^2$ which splits neatly into d separate coordinate-wise problems as shown below:

$$\min_{\alpha_i \in [0,1]} \frac{1}{2}(k_i + 1 - 2\alpha_i)^2$$

where $\mathbf{k} = [k_1, k_2, \ldots, k_d] \stackrel{\text{def}}{=} X^\top \boldsymbol{\lambda} \in \mathbb{R}^d$. The above problem can be solved in a single step using the QUIN trick i.e.,

$$\alpha_i = \Pi_{[0,1]}\left(\frac{k_i + 1}{2}\right)$$