Name: _____

Roll No.: _____ **Dept.**: _____

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

Instructions: *Total:* **60 marks**

1. Total duration: **2 hours**. Please write your name, roll number, department on **all pages**.
2. This booklet has 8 pages (6 pages + 2 pages for rough work). No part of your answers should be on pages designated for rough work. Additional rough sheets may be provided if needed.
3. Write/mark your answers clearly in the provided space. Please keep your answers precise and concise.
4. Avoid showing very detailed derivations (you may use the rough sheet for that). In some cases, you may directly use the standard results/expressions provided on page 6 of this booklet.

Section 1 (6 Multiple Choice Questions: 5 x 2 = 10 marks). (Tick/circle all options that you think are true)

1. Which of the following is true about probabilistic linear regression with Gaussian likelihood, Gaussian prior on weights $\boldsymbol{w}$, and gamma priors on the precision hyperparametes of the likelihood and Gaussian prior? (1) The joint posterior distribution over all the unknowns can be computed in closed form, (2) The joint posterior is intractable, (3) The PPD can be computed in closed form,(4) The PPD is intractable.
(The posterior and PPD are tractable only when hyperparams are known and $\boldsymbol{w}$ is the only unknown).

2. Which of the following is true about Gaussian Processes, assuming the hyperparameters are fixed/known: (1) The PPD is available in closed form for the regression setting with Gaussian likelihood, (2) The PPD is available in closed form for settings when the likelihood is form exponential family, (3) The PPD computation, in general, does not require the posterior to be computed, (4) The label prediction for a test input only depends on the labels of a subset of training examples.
(For GLMs, the likelihood and the Gaussian prior aren't always conjugate (recall logistic regression)).

3. Which of the following is true about the marginal likelihood? (1) It is available in closed form if the likelihood and prior are a conjugate pair from exponential family, (2) It is the expectation of the likelihood w.r.t. the posterior distribution of the model parameters, (3) It is the expectation of the likelihood w.r.t. the prior distribution of the model parameters, (4) If the marginal likelihood is intractable for a model, the posterior will also be intractable.

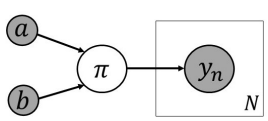
4. Which of the following is true about MAP estimation: (1) Its solution is more robust against overfitting as compared to the MLE solution, (2) Assuming the log-posterior function is differentiable, computing the MAP solution is not much harder as compared to computing the MLE solution, (3) The MAP estimate is equal to the mean of the posterior, (4) When the likelihood and prior are a conjugate pair from exponential family, MAP and MLE solutions are identical.

5. Which of the following is true about the expectation maximization (EM) algorithm used for models that contain both parameters Θ as well as latent variables $\mathbf{Z}$? (1) It can be used to compute the MLE solution of the parameters, (2) It can be used to compute the MAP solution of the parameters, (3) It can be used to compute the joint posterior of the latent variables and the parameters, (4) The maximization (M) step estimates the parameters Θ by maximizing the log-likelihood $\log p(\mathbf{X}|\Theta)$.

Section 2 (6 short answer questions: 6 x 3 = 18 marks). .

1. Draw the complete plate notation diagram for a beta-Bernoulli model of N observations $y_1, y_2, \ldots, y_N$, with each $y_n \sim \text{Bernoulli}(y_n|\pi)$ and $\pi \in (0, 1)$ given a $\text{Beta}(a, b)$ prior. Assume a, b to be known.

   Plate diagram shown below. Here π, a, b are global parameter and thus shown outside the plate. Observations $y_1, y_2, \ldots, y_N$ are given (their values known) and are thus shaded. Likewise, a, b are assumed known and hence shaded. The parameter π is unknown and thus unshaded.



Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

2. The gradient expression for canonical GLM is of the form $\boldsymbol{g} = \sum_{n=1}^{N}(y_n - \mu_n)\boldsymbol{x}_n$, where μ_n is the conditional mean of response y_n, defined as $f(\boldsymbol{w}^\top \boldsymbol{x}_n)$. Briefly explain why this expression makes intuitive sense.

   This is a form of "corrective" update. Training examples $(\boldsymbol{x}_n, y_n)$ on which the current estimate of $\boldsymbol{w}$ makes a large error, i.e., the difference between the true label y_n in its prediction μ_n is large, will contribute more to the gradient expression.

3. Briefly describe how we can construct a scale-mixture of Gaussian distributions, and also give the mathematical expression for this construction.

   A scale-mixture of Gaussians can be obtained by taking a Gaussian with some mean and variance and integrating out the variance parameter using a suitable conjugate prior (gamma in this case). This basically means that we are "mixing" infinite number of Gaussians, all having the same mean but all possible values of its variance parameter. One example of such a scale-mixture, in form of a mathematical expression, would be $p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \tau_d^2)\text{Gamma}(\tau_d^2|a, b)d\tau_d^2$.

4. Consider the following distribution: $p(\theta|m_0, \phi_0) = \exp(\phi_0^\top \theta - m_0 g(\theta) - A(m_0, \phi_0))$. Is this an exponential family distribution? If yes, write down its natural parameters and sufficient statistics. It not, state why it is not an exponential family distribution.

   Yes, it is, since it can be expressed in the form of an exp-fam distribution where natural parameters $\eta = [\phi_0, m_0]$, suff-stats $\phi(\theta) = [\theta, -g(\theta)]$, $h(\theta) = 1$ and log-partition function $A(m_0, \phi_0)$.

5. In a latent variable model, suppose we want to perform hybrid estimation of the unknowns, i.e., compute the posterior for some and compute the point estimate for the others. How would you decide which of these approaches to use for which unknowns?

   Typically, for unknowns for which we have very little data to do their estimation, there is likely to be a fair degree of uncertainty. For such unknowns (e.g., local latent variables which are only associated with a single observation), it is desirable to compute the posterior. On the other hand, for unknowns for which we have plenty of data to do their estimation, the estimates will usually not have that much uncertainty. For such unknowns, we can just do point estimation.

6. Why are the Gaussian Process based models for regression/classification slow at test time?

   It is because the prediction function has a form that depends on computing the similarity of the test input with each of the training inputs. For example, in GP regression, the predictive mean of the response for a new test input x_* is of the form $\mu = \sum_{n=1}^{N} \alpha_n k(x_n, x_*)$. For large N, the prediction will be slow. This issue is similar to kernel methods like kernel SVM or kernel ridge regression.

Section 3 (4 not-so-short answer questions: 10+8+8+6 = 32 marks). .

1. Consider a linear regression model with the likelihood $p(y_n|\boldsymbol{w}, \boldsymbol{x}_n) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta_n^{-1})$ and prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{\Lambda}^{-1})$, where $\boldsymbol{\Lambda}$ is a diagonal precision matrix with its d^{th} diagonal entry being λ_d. Assume $\beta, \boldsymbol{\Lambda}$ to be known. The goal is to estimate $\boldsymbol{w}$ from training data $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$. Write down the final expressions for MLE and MAP objective functions (no need to solve for $\boldsymbol{w}$). Looking at these expressions, what roles do β_n and λ_d play here? Also write down the final expression for the posterior distribution of $\boldsymbol{w}$. You need not show the full derivations; only the final expressions are required.

   Because of the Gaussian likelihood, the log-likelihood term for each observation will be of the form $-\beta_n(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$ (ignoring terms that don't depend on $\boldsymbol{w}$). Because of the Gaussian process, the log-prior term will be of the form $-\boldsymbol{w}^\top \Lambda \boldsymbol{w} = -\sum_{d=1}^{D} \lambda_d w_d^2$ (ignoring terms that don't depend on $\boldsymbol{w}$).

Name:

Roll No.: Dept.:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

The MLE objective is simply the negative log-likelihood over all the observations

$$\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^{N} \beta_n (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

The MAP objective is the MLE objective plus the negative log-prior:

$$\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^{N} \beta_n (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \sum_{d=1}^{D} \lambda_d w_d^2$$

In the above expressions, the Gaussian noise's precision β_n plays the role of controlling the importance of the n-th training example. The Gaussian prior's precision λ_d plays the role of the importance of the d-th feature in the data. The nice aspect of the probabilistic approach is that we can even learn these hyperparameters, thereby learning the importance of each training example and each feature in the data.

Because of the conjugacy between the Gaussian likelihood and the Gaussian prior (note that the hyper-parameters are assumed known), using the results of linear Gaussian model, the posterior distribution of $\boldsymbol{w}$ will also be a Gaussian given by $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \boldsymbol{\beta}, \Lambda) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$ where

$$\boldsymbol{\Sigma}_N = (\Lambda + \sum_{n=1}^{N} \beta_n \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} \quad \text{and} \quad \boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N (\sum_{n=1}^{N} \beta_n y_n \boldsymbol{x}_n)$$

2. Consider a generative classification model with training data $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$ where $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0, 1\}$. Assume each class-conditional distribution to be a Gaussian with its covariance matrix being known. Which quantities would you need to estimate for training this model? Derive the expressions for the MLE solutions of these quantities. Please do not show very detailed steps of derivations; only write the key equations and the final answers. Would the expressions for MAP solutions of these quantities, in general, be the same as the MLE solution? Briefly justify the answer.

This is generative classification where we model the conditional distribution of the labels given the inputs as $p(y|\boldsymbol{x}) = \frac{p(y)p(\boldsymbol{x}|y)}{p(\boldsymbol{x})}$. To estimate this model, we need to estimate the class-prior (a.k.a. class marginal) $p(y)$ and the class-conditional distribution of the inputs, i.e., $p(\boldsymbol{x}|y)$.

Since the labels are binary, we can model the class-prior $p(y)$ using Bernoulli distribution. Assume $\pi \in (0, 1)$ to denote the parameters of this Bernoulli, it is straightforward to estimate π. This is identical to the coin-toss problem where the observations are simply the class labels $\{y_1, y_2, \ldots, y_N\}$ in the training data. The MLE solution for π is simply $\pi_{MLE} = \frac{\sum_{n=1}^N y_n}{N}$.

The class-conditional distributions of inputs is Gaussian. There will be two such Gaussians (one for class 1 and the other for class 0). Since the covariance matrices are assumed to be known, we only need to estimate the means of these two Gaussians. The MLE solution for the mean of a Gaussian is simply the empirical mean of the observations drawn from that Gaussian. For our problem, the mean of the class 1 Gaussian will be $\boldsymbol{\mu}_1 = \frac{\sum_{n:y_n=1} \boldsymbol{x}_n}{N_1}$ and the mean of the class 0 Gaussian will be $\boldsymbol{\mu}_0 = \frac{\sum_{n:y_n=0} \boldsymbol{x}_n}{N_0}$, where N_1 and N_0 denote the number of training examples with class 1 and class 0, respectively.

The MAP solution in general will be different from the above MLE solutions for π, and the Gaussian means because of the prior distributions on π, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, which will result in slightly different expressions for the MAP estimates of these quantities (as we had seen in the case of beta-Bernoulli and Gaussian parameter estimation problems in the class).

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

3. Consider a logistic regression model $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \frac{1}{1+\exp(-y_n \boldsymbol{w}^\top \boldsymbol{x}_n)}$, with a zero-mean Gaussian prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \lambda^{-1}\mathbf{I})$. Note that this loss function for logistic regression assumes $y_n \in \{-1, +1\}$ instead of $\{0, 1\}$. Show that the MAP estimate for $\boldsymbol{w}$ can be written as $\boldsymbol{w} = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n$ where each α_n itself is a function of $\boldsymbol{w}$. Based on the expression of α_n, you would see that it has a precise meaning. Briefly state what α_n means, and also briefly explain why the result $\boldsymbol{w} = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n$ makes sense for this model.

   Taking the derivative of the log-likelihood w.r.t. $\boldsymbol{w}$ and setting it to zero gives

$$\boldsymbol{w} = \frac{1}{\lambda}\sum_{n=1}^N \frac{1}{1+\exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n)} y_n \boldsymbol{x}_n = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n$$

   In the above expression, we can think of $\alpha_n = \frac{1}{1+\exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n)} = 1 - \frac{1}{1+\exp(-y_n \boldsymbol{w}^\top \boldsymbol{x}_n)}$ as denoting the importance of the n-th training example in the final solution $\boldsymbol{w}$. Note that this term is also the probability of **mis-classification**. Thus the training examples that are harder (thus having a higher probability of being mis-classified) are deemed as more important and contribute more to the solution of $\boldsymbol{w}$. Note that solving for $\boldsymbol{w}$ in logistic regression requires an iterative optimization (no closed form solution), and in each iteration, the estimate of $\boldsymbol{w}$ uses these importances to calculate the value of $\boldsymbol{w}$.

4. Consider N scalar-valued observations $x_1, \ldots, x_N$ drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. Consider their empirical mean $\bar{x} = \frac{1}{N}\sum_{n=1}^N x_n$. Expressing $\bar{x}$ as a linear transformation of a random variable, derive the probability distribution of $\bar{x}$.

   Consider the $N \times 1$ vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^\top$. Note that the distribution of $\boldsymbol{x}$ will be the N-dimensional Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \sigma^2\mathbf{I}_N)$, where $\boldsymbol{\mu}$ is the $N \times 1$ vector with each entry being equal to μ.

   We can then express $\bar{x}$ as the linear transformation $\bar{x} = \frac{1}{N}\mathbf{1}_N{}^\top \boldsymbol{x}$ where $\mathbf{1}_N$ is the $N \times 1$ vector of all 1s. Using the properties of linear transformation of a Gaussian random variable, it is easy to see that the distribution of $\bar{x}$ is also a Gaussian with:

   - Mean $= \mathbb{E}[\frac{1}{N}\mathbf{1}_N{}^\top \boldsymbol{x}] = \frac{1}{N}\mathbf{1}_N{}^\top \boldsymbol{\mu} = \mu$
   - Variance $= \text{var}[\frac{1}{N}\mathbf{1}_N{}^\top \boldsymbol{x}] = \frac{1}{N}\mathbf{1}_N{}^\top \text{cov}[\boldsymbol{x}]\frac{1}{N}\mathbf{1}_N = \frac{1}{N}\mathbf{1}_N{}^\top \sigma^2\mathbf{I}_N\frac{1}{N}\mathbf{1}_N = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

Some distributions and their properties:

- For $x \in (0,1)$, $\text{Beta}(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and Γ denotes the gamma function s.t. $\Gamma(x) = (x-1)!$ for a positive integer x. Expectation of a Beta r.v.: $\mathbb{E}[x] = \frac{a}{a+b}$.
- For $x \in \{0,1,2,\ldots\}$ (non-negative integers), $\text{Poisson}(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$ where λ is the rate parameter.
- For $x \in \mathbb{R}_+$, $\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$ (shape and rate parameterization), and $\text{Gamma}(x|a,b) = \frac{1}{\Gamma(a)b^a} x^{a-1} \exp(-\frac{x}{b})$ (shape and scale parameterization)
- For $x \in \mathbb{R}$, Univariate Gaussian: $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$
- For $x \in \mathbb{R}^D$, D-dimensional Gaussian: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$.
  Trace-based representation: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\Big\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\Big\}$, $\mathbf{S} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top$.
  Information form: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x} + \boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi} - 2\boldsymbol{x}^\top\boldsymbol{\xi}\right)\right]$ where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$
- For $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$, s.t. $\sum_{k=1}^K \pi_k = 1$, $\text{Dirichlet}(\boldsymbol{\pi}|\alpha_1,\ldots,\alpha_K) = \frac{1}{B(\alpha_1,\ldots,\alpha_K)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$
  where $B(\alpha_1,\ldots,\alpha_K) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}$, and $\mathbb{E}[\pi_k] = \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}$
- For $x_k \in \{0, N\}$ and $\sum_{k=1}^K x_k = N$, $\text{multinomial}(x_1,\ldots,x_K|N,\boldsymbol{\pi}) = \frac{N!}{\boldsymbol{x}_1!\ldots,x_K!}\pi_1^{x_1}\ldots\pi_K^{x_K}$
  where $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$, s.t. $\sum_{k=1}^K \pi_k = 1$. The multinoulli is the same as multinomial with $N = 1$.

Some other useful results:

- If $\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon$, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{L}^{-1})$ then $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z} + \boldsymbol{b}, \mathbf{L}^{-1})$,
  $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top + \mathbf{L}^{-1})$, and $p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top\mathbf{L}(\boldsymbol{x}-\boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma})$,
  where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}$.
- Marginal and conditional distributions for Gaussians: $p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$,
  $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ where $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$, $\boldsymbol{\mu}_{a|b} = \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\right\} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$, where symbols have their usual meaning. :)
- $\frac{\partial}{\partial\boldsymbol{\mu}}[\boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}] = [\mathbf{A} + \mathbf{A}^\top]\boldsymbol{\mu}$, $\frac{\partial}{\partial\mathbf{A}}\log|\mathbf{A}| = \mathbf{A}^{-\top}$, $\frac{\partial}{\partial\mathbf{A}}\text{trace}[\mathbf{A}\mathbf{B}] = \mathbf{B}^\top$
- For a random variable vector $\boldsymbol{x}$, $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\top + \text{cov}[\boldsymbol{x}]$

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023