**Name**:

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

**IIT Kanpur**
**CS698X (TPMI)**
**Quiz-1**
*Date:* January 31, 2019

---

**Instructions**:

*Total:* **40 marks**

1. Please write your name, roll number, department on **all pages** of this question paper.
2. Write your answers clearly in the provided box. Keep your answer precise and concise.

**Section 1** (16 very short answer questions: 16 x 2 = 32 marks). .

1. Can the integral $\int_{-\infty}^{\infty} \exp[-\lambda(x-\mu)^2]dx$ be computed exactly? If yes, write its value. If no, state why.

   Yes because it is simply the normalization constant of $\mathcal{N}(x|\mu, 1/(2\lambda))$. The value will be $\sqrt{\frac{\lambda}{\pi}}$

2. Consider a weight vector $\boldsymbol{w} \in \mathbb{R}^D$ with a Gaussian prior of the form $p(\boldsymbol{w}) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda_d^{-1})$. The expression of the equivalent regularizer that corresponds to this prior will be

   $R(\boldsymbol{w}) = \sum_{d=1}^{D} \lambda_d w_d^2$

3. Given the observation model $p(x|\theta)$ with parameters $\theta$ and a prior $p(\theta|\lambda)$ on the parameters $\theta$, the general expression for the marginal likelihood $p(x|\lambda)$ will be

   $p(x|\lambda) = \int p(x|\theta)p(\theta|\lambda)d\theta$

4. Does approximating the posterior of a Bayesian Logistic Regression model using Laplace approximation result in a closed-form posterior predictive? If yes, why? If no, why not?

   No because even after Laplace approximation (which gives a Gaussian posterior), the likelihood $p(y|\boldsymbol{w}, \boldsymbol{x})$ is still Bernoulli, and the integral required can't be done in closed form.

5. Suppose you have computed the posterior distribution for some parameter given some observed data. Can you get the MLE from this posterior? Can you get the MAP estimate from this posterior?

   Can't get MLE but can get MAP by using the mode of the obtained posterior.

6. Suppose you have tossed a coin a number of times. Now suppose you want to compute the probability that $\theta \leq 0.4$ where $\theta$ is the probability of heads. Briefly suggest a Bayesian way to do this.

   Once we have compute the posterior (which will be Beta), we can get $p(\theta \leq 0.4)$ by simply computing the CDF of this posterior, i.e., $\int_0^{0.4} p(\theta|X)d\theta$

7. Consider $N$ scalar-valued i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ from $\mathcal{N}(\mu, \sigma^2)$. Assume a prior $p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$ and $\sigma^2, \mu_0, \sigma_0^2$ to be known. Write down the expression for the marginal distribution of $x_1$. (hint: you can also do it without computing integrals)

   Note that $x = \mu + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Thus the marginal will be $p(x|\mu_0, \sigma_0^2, \sigma^2) = \mathcal{N}(\mu_0, \sigma_0^2 + \sigma^2)$

8. Can you turn a Gaussian prior into a uniform prior by choosing the Gaussian's hyperparameters suitably? If yes, how? If no, why not?

   Yes, by making it a very very flat Gaussian by choosing the variance to be very large (tending to infinity)

9. Give two examples of problems where the marginal likelihood of a model can be useful, with a brief explanation of *how/why* marginal likelihood is useful for those problems.

   (1) Model selection by picking the model $m$ that has the largest marginal likelihood $p(X|m)$, (2) Hyperparameter estimation by finding the hyperparameters that maximize the marginal likelihood (which is a function of these hyperparameters).

10. State at least two conditions under which the MAP solution for a probabilistic model would be (at least roughly) equivalent to the MLE solution.

    (1) When the prior is uniform, (2) When we have lots of data for estimating the parameter (the prior's influence would wash out when we have a lot of data)

**Name**:

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

**IIT Kanpur**
**CS698X (TPMI)**
**Quiz-1**
*Date:* January 31, 2019

11. Is posterior predictive distribution more expensive to compute as compared to a plug-in predictive distribution? Briefly justify your answer.

    Usually yes since we have to perform an integral to compute the posterior predictive (it's like taking the plug-in predictive, computing it for each possible parameter, and doing posterior-weighted averaging). In some cases however, we may have closed form solution (recall Bayesian linear regression), and in such cases, it is almost as easy to compute as the plug-in predictive.

12. Briefly explain why posterior predictive has a larger variance as compared to the plug-in predictive?

    Because we are averaging over the posterior over parameters (which itself has uncertainty). Basically, uncertainty in the parameters translates into uncertainty in the predictions. The plug-in predictive only uses a single value of the parameters, due to which it typically has a smaller variance

13. When using MLE for estimating the parameters $\theta$ of an exponential family distribution of the form $p(x|\theta) = h(x)\exp[\theta^\top \phi(x) - A(\theta)]$, are you guaranteed to find a global optima? Briefly justify your answer.

    Yes. $A(\theta)$ is convex and therefore MLE will be a concave maximization (convex minimization) problem.

14. Consider the Poisson distribution $p(x|\lambda)$ over non-negative integers $x \geq 0$. Given $N$ observations $x_1, \ldots, x_N$, we wish to perform MLE for its rate parameter $\lambda > 0$. What is the sufficient statistics in this case?

    Sum of of the observation, i.e., $\sum_{n=1}^{N} x_n$ (easy to see by looking at the form of the likelihood, which is a product of $N$ Poisson PMFs.)

15. Briefly explain why the precision/variance of the Gaussian prior of some weight vector $\boldsymbol{w}$ can be seen as controlling the extent of regularization.

    Consider a zero mean Gaussian prior $\mathcal{N}(\boldsymbol{w}|0, \lambda^{-1}\mathbf{I}_D)$, where $\lambda$ is the precision. This corresponds to a regularizer of the form $\lambda \boldsymbol{w}^\top \boldsymbol{w}$ where $\lambda$ can be seen as the regularization hyperparameter.

16. Briefly explain how hierarchical modeling can be used to obtain a sparse prior on a real-valued weight vector (e.g., the ones used for linear regression/classification models).

    In the hierarchical modeling approach, we can assume a zero-mean Gaussian prior but also put a prior on the precision/variance. Integrating out the precision/variance results in a sparse prior on $\boldsymbol{w}$ (e.g., Student-t, Laplace, etc)

**Section 2** (2 short answer questions: 2 x 4 = 8 marks). .

1. For Bayesian linear regression with $p(y_n|\boldsymbol{w}, \boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$ and $p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$, assuming hyperparameters as known, and responses $y_n$'s to be independent of each other given $\boldsymbol{w}$, can you compute the posterior predictive distribution $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ without first computing the posterior of $\boldsymbol{w}$? If yes, how? If no, why not? You don't need to provide a detailed mathematical derivation or need to obtain the final expression of the posterior predictive. Just a basic augument (for/against) in words and/or a couple of basic equations should suffice.

   I already explained this in the email sent right after the quiz was over. The basic idea is to write down the joint marginal distribution $p(\boldsymbol{y}, y_*|\mathbf{X}, \boldsymbol{x}_*)$ of training and test outputs, which in this case will be a Gaussian. Due to the property of Gaussians, the conditional distribution $p(y_*|\boldsymbol{y})$ of test outputs given training outputs (skipping the inputs from the notation) will also be Gaussian whose mean and variance can be found easily using standard results of Gaussians.

**Name**:

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

**IIT Kanpur**
**CS698X (TPMI)**
**Quiz-1**
*Date:* January 31, 2019

---

2. Assume $N$ observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn i.i.d. from the exponential distribution,which is defined as $p(x_n|\theta) = \theta \exp(-\theta x_n)$ and the prior on the parameter $\theta > 0$ is $p(\theta) = \text{Gamma}(\theta|a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}\exp(-b\theta)$. What is the marginal likelihood $p(\mathbf{X}|a, b)$? Give your answer as a closed-form expression (not an integral). Avoid very detailed derivation; show only the basic steps and write down the final expression.

   This is easy to compute. To see this, first note that $p(\mathbf{X}|\theta) = \prod_{n=1}^{N} p(x_n|\theta) = \theta^N \exp(-\theta \sum_{n=1}^{N} x_n)$. From this, we can get the marginal likelihood as $p(\mathbf{X}|a, b)$ by integrating out $\theta$

$$p(\mathbf{X}|a, b) = \int p(\mathbf{X}|\theta)p(\theta|a, b)d\theta = \int \theta^N \exp(-\theta \sum_{n=1}^{N} x_n)\frac{b^a}{\Gamma(a)}\theta^{a-1}\exp(-b\theta)d\theta$$

   The constant $\frac{b^a}{\Gamma(a)}$ comes out and the remaining integral is nothing but the normalization constant of $\text{Gamma}(\theta|a+N, b+\sum_{n=1}^{N})$. Therefore $p(\mathbf{X}|a, b) = \frac{b^a}{\Gamma(a)}\frac{\Gamma(a+N)}{(b+\sum_{n=1}^{N})^{a+N}}$, which incidentally is also the posterior of $\theta$. (Overall, this marginal likelihood can be seen as a ratio of two normalization constants (of the prior and posterior of $\theta$); you may recall our discussion of exponential family distributions and this property holds not just for this example but for the marginal distributions of all exp-fam distributions).

**Some essential equations/results that may be useful:**

- For $x \in \mathbb{R}$, the univariate Gaussian distribution: $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$.
- For $x \in \{0, 1, 2, \ldots\}$ (non-negative integers), $\text{Poisson}(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$ where $\lambda$ is the rate parameter.
- Some standard results on Gaussians: 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 $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}$.
- Most likely you won't need any other results (the questions themselves contain some of the formulae). :-)

**Name**:

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

**IIT Kanpur**
**CS698X (TPMI)**
**Quiz-1**
*Date:* January 31, 2019