CS698S

ASSIGNMENT-1

NAME: PRATIK MISHRA

ROLL: 14493

Ques 1 :

Calculation of MGF of $\int P(x|\eta)P(\eta|r)d\eta$:

$$M_x(t) = \int_{-\infty}^{\infty} e^{tx}\left[\int_0^{\infty} \frac{1}{\sqrt{2\pi\eta}} e^{-\frac{x^2}{2\eta}} \cdot \frac{r^2}{2} e^{-\frac{r^2\eta}{2}} d\eta\right]dx$$

$$= \int_0^{\infty} \frac{r^2 e^{-r^2\eta/2}}{2\sqrt{2\pi\eta}}\left[\int_{-\infty}^{\infty} e^{tx - x^2/2\eta}\, dx\right]d\eta$$

$$= \int_0^{\infty} \frac{r^2 e^{-r^2\eta/2}}{2\sqrt{2\pi\eta}}\left[\int_{-\infty}^{\infty} e^{-\left(\frac{x}{\sqrt{2\eta}} - \frac{t\sqrt{2\eta}}{2}\right)^2} \cdot e^{\frac{t^2\eta}{2}} dx\right]d\eta$$

Substitute $y = \frac{x}{\sqrt{2\eta}} - \frac{t\sqrt{2\eta}}{2}$, then, $dy = \frac{dx}{\sqrt{2\eta}}$

$$\therefore M_x(t) = \int_0^{\infty} \frac{r^2 e^{-r^2\eta/2} \cdot e^{t^2\eta/2}}{2\sqrt{2\eta\pi}} \cdot \sqrt{2\eta}\left[\int_{-\infty}^{\infty} e^{-y^2} dy\right]d\eta$$

$$= \int_0^{\infty} \frac{r^2}{2} e^{-\eta\left(\frac{r^2 - t^2}{2}\right)}\, d\eta$$

$$= \frac{\gamma^2}{2}\left[\frac{e^{-\eta\left(\frac{\gamma^2}{2}-t^2\right)}}{-\left(\frac{\gamma^2}{2}-t^2\right)}\right]_0^\infty \quad , \quad \gamma^2 > t^2$$

$$= \frac{1}{1-\frac{t^2}{\gamma^2}}$$

$$= L\left(0, \frac{1}{\gamma}\right) \rightarrow \underline{\text{Laplace distribution}}.$$

Thus, the distribution function that was required is:

$$p(x|\gamma) = \frac{\gamma}{2} e^{-\gamma|x|}$$

<u>Mean = 0</u> and <u>Variance</u> $= \frac{2}{\gamma^2}$

The plot of $p(x|\gamma)$ which is a laplace pdf is more concentrated towards the mean than the Gaussian $p(x|\eta)$ and also has a sharp corner which is not present in $p(x|\eta)$.

Ques 2

$$\sigma_N^2(x_*) = \beta^{-1} + x_*^T \Sigma_N x_*$$
$$= \beta^{-1} + x_*^T \left(\beta \sum_{n=1}^{N} x_n x_n^T + \lambda I\right)^{-1} x_*$$

$$\sigma_{N+1}^2(x_*) = \beta^{-1} + x_*^T \left(\beta \sum_{n=1}^{N+1} x_n x_n^T + \lambda I\right)^{-1} x_*$$
$$= \beta^{-1} + x_*^T \left(\beta \sum_{n=1}^{N} x_n x_n^T + \lambda I + \beta x_{N+1} x_{N+1}^T\right)^{-1} x_*$$

Let $M = \beta \sum_{n=1}^{N} x_n x_n^T + \lambda I$

$$\therefore \sigma_{N+1}^2(x_*) = \beta^{-1} + x_*^T \left(M + \beta x_{N+1} x_{N+1}^T\right)^{-1} x_*$$

$$\left(M + \beta x_{N+1} x_{N+1}^T\right)^{-1} = M^{-1} - \frac{\beta (M^{-1} x_{N+1})(x_{N+1}^T M^{-1})}{1 + \beta x_{N+1}^T M^{-1} x_{N+1}}$$

$$\Rightarrow \sigma_{N+1}^2(x_*) = \beta^{-1} + x_*^T M^{-1} x_* - \frac{x_*^T \beta (M^{-1} x_{N+1})(x_{N+1}^T M^{-1}) x_*}{1 + \beta x_{N+1}^T M^{-1} x_{N+1}}$$

$$= \sigma_N^2(x_*) - \frac{x_*^T \beta (M^{-1} x_{N+1})(x_{N+1}^T M^{-1}) x_*}{1 + \beta x_{N+1}^T M^{-1} x_{N+1}}$$

Clearly, $\frac{x_*^T \beta (M^{-1} x_{N+1})(x_{N+1}^T M^{-1}) x_*}{1 + \beta x_{N+1}^T M^{-1} x_{N+1}} > 0$

Hence,

$$\sigma_{N+1}^2 (x_*) < \sigma_N^2 (x_*)$$

Thus, variance decreases with each data point and therefore the Model gets more certain.

ans 3

(1) Given that $x_1, \ldots, x_N \in \mathbb{R}^D, \forall n$

such that x_n is generated from the observation model $Az_n + \varepsilon_n$

where $A = [a_1, \ldots, a_k]$ is a $D \times K$ real valued matrix and $z_n \in R^K$.

Therefore $x_n \sim N(Az_n, \Psi)$.

Also, prior on $a_k \sim N(0, D^{-1} I_D)$

~~Posterior~~ $\cancel{= P(a_k | x, z,) \propto P(a_k) \cdot p(x | z, A)}$

$$\cancel{\propto N(0, D^{-1} I_D) \cdot \prod_{n=1}^{N} p(x_n | z, A)}$$

$$\cancel{\propto N(0, D^{-1} I_D) \cdot \prod_{n=1}^{N} N(Az_n, \Psi)}$$

$\therefore$ Posterior $= P(a_k | x, z, a_l) \propto p(a_k) \cdot p(x | z, A)$

$$\propto N(0, D^{-1} I_D) \cdot \prod_{n=1}^{N} P(x_n | z_n, A)$$

$$\propto N(0, D^{-1} I_D) \cdot \prod_{n=1}^{N} N(Az_n, \Psi)$$

② Posterior distribution of z_n -

$$p(z_n \mid X, A, Z_\ell) \propto p(z_n) \cdot p(X \mid A, Z)$$

$$\propto N(0, \gamma^{-1} I_K) \cdot \prod_{n=1}^{N} p(x_n \mid A, z_n)$$

$$\propto N(0, \gamma^{-1} I_K) \cdot \prod_{n=1}^{N} (A z_n, \varphi)$$

Ques 4

- Given that $p(w \mid b, \sigma^2_{spike}, \sigma^2_{slab})$

$$= \begin{cases} N(w \mid 0, \sigma^2_{spike}), b=0 \\ N(w, 0, \sigma^2_{slab}), b=1 \end{cases}$$

Prior $p(b=1) = \pi = \frac{1}{2}$

$\therefore$ $p(w \mid \sigma^2_{spike}, \sigma^2_{slab}) = \frac{1}{2} p(w \mid b=1, \sigma^2_{spike}, \sigma^2_{slab})$

$$+ \frac{1}{2} p(w \mid b=0, \sigma^2_{spike}, \sigma^2_{slab})$$

$$= \frac{1}{2} N(0, \sigma^2_{slab}) + \frac{1}{2} N(0, \sigma^2_{spike})$$

$\therefore p(w \mid \sigma^2_{spike}, \sigma^2_{slab}) = N\left(0, \frac{1}{4}(\sigma^2_{slab} + \sigma^2_{spike})\right)$

- For $\sigma^2_{spike} = 1$ and $\sigma^2_{slab} = 100$,

$p(w \mid \sigma^2_{spike}, \sigma^2_{slab}) = N(0, 25.25)$

[Plot is attached as a jpg file].

- Let $x = w + \varepsilon$ where $\varepsilon \sim N(\varepsilon, 0, \rho^2)$

$\therefore p(x \mid w, \rho^2) = N(x \mid w, \rho^2)$

$\therefore p(b=1 \mid x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)$

$$\propto p(b=1) \cdot p(x \mid b=1, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)$$

$$\propto \frac{1}{2} \cdot p(x \mid b=1, \rho^2, \sigma^2_{spike}, \sigma^2_{slab})$$

Integrating out 'w', we get

$$p(x \mid b=1, \rho^2, \sigma^2_{spike}, \sigma^2_{slab}) = \int_{-\infty}^{\infty} p(x \mid b=1, \rho^2, \sigma^2_{spike}, \sigma^2_{slab}) \cdot p(w \mid b=1, \sigma^2_{slab})\, dw$$

$$= \int_{-\infty}^{\infty} N(w, \rho^2) \cdot N(0, \sigma^2_{slab})\, dw$$

$$= \frac{1}{2\pi \rho \sigma^2_{slab}} e^{-\frac{x^2}{2\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{w^2}{2\rho^2} + \frac{xw}{\rho^2}} \cdot e^{-\frac{w^2}{2\sigma^2_{slab}}}\, dw$$

$$= \frac{1}{2\pi \rho \sigma_{slab}} e^{-\frac{x^2}{2\rho^2}} \int_{-\infty}^{\infty} e^{\left(-\frac{w^2}{2\rho^2} + \frac{xw}{\rho^2} - \frac{w^2}{2\sigma^2_{slab}}\right)} dw$$

$$= \frac{1}{2\pi \rho \sigma_{slab}} e^{-\frac{x^2}{2\rho^2}} \int_{-\infty}^{\infty} e^{-w^2\left(\frac{1}{2\rho^2} + \frac{1}{2\sigma^2_{slab}}\right) + \frac{xw}{\rho^2}} dw$$

$$\boxed{\text{Let } M^2 = \frac{1}{\left(\frac{1}{2\rho^2} + \frac{1}{2\sigma^2_{slab}}\right)}}$$

$\therefore p(x \mid b=1, \rho^2, \sigma^2_{slab}, \sigma^2_{spike})$

$$= \frac{1}{2\pi \rho \sigma_{slab}} e^{-\frac{x^2}{2\rho^2}} \cdot e^{\frac{x^2 M^2}{4\rho^4}} \int_{-\infty}^{\infty} e^{-\left(\frac{w}{k} - \frac{xM}{2\rho^2}\right)^2} dw$$

$$\boxed{\text{Let } y = \frac{w}{k} - \frac{xM}{2\rho^2} \text{, such that } dy = dw}$$

$$= \frac{1}{2\pi \rho \sigma_{slab}} e^{-\frac{x^2}{2\rho^2}\left(1 - \frac{M^2}{2\rho^2}\right)} \int_{-\infty}^{\infty} e^{-y^2} dy$$

$$= \frac{1}{2\sqrt{\pi}\, \rho \sigma_{slab}} e^{-\frac{x^2}{2\rho^2}\left(1 - \frac{M^2}{2\rho^2}\right)}$$

$$= \frac{1}{\sqrt{4\pi \rho^2 \sigma^2_{slab}}} e^{-\frac{x^2}{2\rho^2}\left(\frac{\rho^2}{\rho^2 + \sigma^2_{slab}}\right)}$$

$$= \frac{1}{\sqrt{4\pi \rho^2 \sigma^2_{slab}}} e^{-\frac{x^2}{\rho^2 + \sigma^2_{slab}}}$$

So,

$$p(b=1 \mid x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) \propto \frac{1}{2\sqrt{4\pi\rho^2\sigma^2_{slab}}} e^{-\frac{x^2}{2(\rho^2+\sigma^2_{slab})}}$$

Hence,

$$p(b=1 \mid x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) = N(0, \rho^2+\sigma^2_{slab})$$

- Let $x = w + \varepsilon$ as defined in the previous part

$$p(w \mid x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) \propto p(w \mid \sigma^2_{spike}, \sigma^2_{slab}) \cdot p(x \mid w, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2)$$

$$\propto N\left(0, \tfrac{1}{4}(\sigma^2_{spike}+\sigma^2_{slab})\right) \cdot N(w, \rho^2)$$

Hence,

$$p(w \mid x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) = N(w \mid \mu, \sigma^2)$$

where,

$$\mu = \frac{x}{1+\frac{4\rho^2}{\sigma^2_{spike}+\sigma^2_{slab}}} = \frac{x(\sigma^2_{spike}+\sigma^2_{slab})}{(\sigma^2_{spike}+\sigma^2_{slab}+4\rho^2)}$$

$$\sigma^2 = \frac{1}{\frac{1}{\rho^2}+\frac{4}{\sigma^2_{spike}+\sigma^2_{slab}}} = \frac{\rho^2(\sigma^2_{spike}+\sigma^2_{slab})}{\sigma^2_{spike}+\sigma^2_{slab}+4\rho^2}$$

• Given, $\sigma^2_{spike} = 1$, $\sigma^2_{slab} = 100$, $\rho^2 = 0.01$, $x = 3$,

$$p(w \mid x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2) = \mathcal{N}(2.99,\ 0.0099)$$

Summary of Edward's Key Features and Capabilities

Edwardlib.org is the official website for Edward which is a Python library for Probabilistic modelling, inference and criticism named after the statistician George Edward Pelham Box.

The short introduction on the website says "It is a testbed for fast experimentation and research with probabilistic models, ranging from classical hierarchical models on small data sets to complex deep probabilistic models on large data sets. Edward fuses three fields: Bayesian statistics and machine learning, deep learning, and probabilistic programming."

Edward is built on top of TensorFlow, a library for numerical computing using data flow graphs built by Google. TensorFlow enables Edward to speed up computation with hardware such as GPUs, to scale up computation with distributed training, and to simplify engineering effort with automatic differentiation.

Edward is built keeping in mind an iterative process for probabilistic modelling which could be described as follows:

Given data from some unknown phenomena, first, formulate a model of the phenomena;
second, use an algorithm to infer the model's hidden structure, thus reasoning about the phenomena;
third, criticize how well the model captures the data's generative process.
With criticism of the model's fit to the data, the components of the model are revised and the process is repeated forming an iterative loop.

Edward supports **modelling** with Directed graphical models, neural networks (via libraries such as Keras and TensorFlow Slim), conditionally specified undirected models and Bayesian nonparametric and probabilistic programs.

It supports **inference** with Variational inference in the form of Black box variational inference, stochastic variational inference, Inclusive KL divergence and Maximum a posteriori estimation.
It supports **inference** with Monte Carlo methods such as Hamiltonian Monte Carlo, stochastic gradient Langevin dynamics and Metropolis-Hastings. It also allows for Compositions of inference like Expectation-Maximization, Pseudo-marginal and ABC methods and Message passing algorithms.

It also allows for **Criticism** of the model and inference via point based evaluations and posterior predictive checks (PPCs). PPCs analyze the degree to which data generated from the model deviate from data generated from the true distribution. They can be used either numerically to quantify this degree, or graphically to visualize this degree.