Homework 13
36-705
Due: Thursday December 3 by 3pm.

1. Let $X_1, \ldots, X_n \sim N(\mu, 1)$. Let μ have a $N(a, b^2)$ prior.

   (a) Find the posterior.

   (b) Find c_1 and c_2 so that $P(c_1 < \mu < c_2 | X_1, \ldots, X_n) = 1 - \alpha$.

   (c) Show that, from the frequentist perspective, $P(\mu \in C) \to 1 - \alpha$ as $n \to \infty$ where $C = [c_1, c_2]$.

2. Consider a model $(p_\theta : \ \theta \in \Theta)$ where the parameter space $\Theta = \{\theta_1, \ldots, \theta_k\}$ is finite. The π be prior for θ. Let θ_* denote that true value of θ. In other words, we observe $X_1, \ldots, X_n \sim p_{\theta_*}$. Show that a necessary and sufficient condition for the posterior to concentrate at θ_* is $\pi(\theta_*) > 0$. That is, let $b_n = P(\theta = \theta_* | X_1, \ldots, X_n)$. Then $b_n \xrightarrow{P} 1$ if and only if $\pi(\theta_*) > 0$.

3. Let $X_1, \ldots, X_n \sim p_\theta$ where $\theta \in \mathbb{R}$. The posterior density is

$$p(\theta | X_1, \ldots, X_n) \propto \mathcal{L}(\theta) \pi(\theta)$$

   where $\mathcal{L}(\theta)$ is the likelihood and $\pi(\theta)$ is the prior. Give an informal, heuristic argument to show that the posterior is approximately Normal.

   Hint: write $\mathcal{L}(\theta) = e^{\ell(\theta)}$ where $\ell(\theta)$ is the log-likelihood. Now Taylor expand the log-likelihood around the mle.

4. We saw that AIC adds a correction because the log-likelihood is a biased estimate of how well the log-likelihood approximates the KL distance. Another way to say this is that the likelihood of the observed data is a biased estimate of the likelihood of a future observation. We'll take a closer look at that problem in this question. Suppose that $X_1, \ldots, X_n \sim N(\theta, I)$ where $\theta \in \mathbb{R}^d$. Let $\ell_n(\theta) = \sum_i \ell(\theta; X_i)$ denote the log-likelihood and $\ell(\theta; X_i) = \log p_\theta(X_i)$. Let $\widehat{\theta}$ be the mle. We want to show that $\ell_n(\widehat{\theta})$ is a biased estimate of $\ell(\widehat{\theta}; X) = \log p_{\widehat{\theta}}(X_i)$, where X is a new observation. The negative of the training log-likelihood (up to constants) is:

$$\ell_n(\widehat{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \|X_i - \widehat{\theta}\|_2^2.$$

   The MLE is $\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

   (a) Show that,

$$\mathbb{E}[\ell_n(\widehat{\theta})] = \frac{d}{2} - \frac{d}{2n}.$$

   (b) For a new observation X, the negative of the true log-likelihood of that point is

$$\ell(\widehat{\theta}; X) = \frac{1}{2} \|X - \widehat{\theta}\|_2^2.$$

   Show that this has mean

$$\frac{d}{2} + \frac{d}{2n}.$$

   (c) This shows that the training log-likelihood has an upward bias (or the training loss on average appears lower than it really is) and that this bias depends on the complexity of the parameter we are estimating (in this case, we are estimating a d-dimensional mean). Use this to justify the AIC correction to the training log-likelihood in this setting.