Model Selection

1. Given observed data D and parameters w, complete the equation using the likelihood and the marginal probability of w:
$$p(D) =$$

2. You have 2 biased coins: one that comes up heads with probability .8 and another that comes up with heads with probability .2. You select a coin randomly and you observe 5 heads out of 10 coin flips. Calculate $p(D)$ for the models that represent the first and second coin respectively. Which coin are you more likely to have selected?

3. Following the previous problem, the apriori model selection is now different. $p(M_1) = .2$ where M_1 is the model in which the first coin is selected. Which model is more likely now?

4. In general, when we evaluate the posterior distribution, we do leave out the normalization constant Z (i.e., we have $p(w|D) \propto p(D|w)p(w)$). Suppose we have two models M_1, M_2, and we want to evaluate the posterior probability of a parameter vector w_i under both models. Explain why you cannot just use $p(D|w, M_i)p(w|M_i)$ to evaluate the posterior probabilities.

5. Suppose you had three models, M_1, M_2, M_3, each increasing in complexity. For example, you could imagine that the models represented unregularized polynomial regression, with M_1 linear regression, M_2 quadratic regression, and M_3 cubic regression. Within the context of Bayesian model selection, come up with a way to penalize the complexity of a model so you do not always choose M_3. Additionally, explain why, in many cases, Bayesian model selection will recover the simplest model to explain the data without explicit penalization.

6. Bishop discusses an approximation to the marginal likelihood as $\ln p(D) \approx \ln p(D|\boldsymbol{w}_{map}) + M \ln(a/b)$, where $\boldsymbol{w}_{MAP}$ is the MAP parameter and a, b are constants with $a < b$, and M is the number of parameters (each parameter is $\boldsymbol{w}_i$). Describe how this marginal likelihood naturally penalizes more complex models.