A Few More Comments on Bayesian and NN Models

1. [Page 11] In the definition of E_D and E_W, you included a scalar 1/2, Is there particular reason to do this? If there is no particular reason, why not use a simple form?

2. [Page 16] `The approximated version of Bayesian inference looks something like this: ...` - *approximate Bayesian inference* means using a **valid distribution** to approximate the posterior distribution and then making inferences. The well-known Laplace approximation and variational method are such approximations. The MCMC-related sampling methods are numerical approximations of the normalizing component so one still uses the original posterior distribution. Note that the expression of the left-hand side of the *approximated Bayes rule* $p(\theta|D) \approx p(D|\theta)p(\theta)/Z$ is not a valid distribution if the approximated Z is used. You also mentioned KL divergence which assesses the goodness of approximation of two *valid** distributions. The definition of the KL *distance** requires two valid distribution functions - I suggest rephrasing the statement and dropping the expression of the approximated Bayes rule (since it is not the way of expressing the approximation of two different distributions).

3. [Page 16] 'although it is unclear how the work was divided. Marshall Rosenbluth and Edward Teller are co-authors, along with their wives, who performed much of the calculations - this comment seems to be irrelevant to the theme of your thesis. I suggest not commenting on individual authors' contributions to the work.

4. [Page 17] `An essential element to Bayesian inference is marginalization` - marginalization has some merits in Bayesian analysis but is not an essential element of it. Among several applications, marginalizing out nuisance parameters using a prior of these nuisance parameters to reduce the dimension of the parameter space (hence, simplifies the inference) is commonly used. Note that the marginalization method is not developed for Bayesian analysis although one can always name the distribution of the parameter to be integrated out as a prior. The marginal likelihood (obtained from marginalization) is a standard method used in frequentist inferences. I suggest rephrasing this argument to avoid over-emphasizing the importance of marginalization in Bayesian inference.

5. [Page 18] The expression of the normalizing integral $Z(\alpha, \beta) = \int d_k w e^{-(\alpha E_D + \beta E_W)}$ is mathematically incorrect simply because the left-hand side does not have the index k. I suggest simply stating that $Z(\alpha, \beta)$ is a normalizing constant if you don't want to search for the notations used in the original paper.

6. [Page 20] `In Bayes, there are no point estimates.` - this argument is wrong. There are both point and interval estimates in Bayesian analysis. For example, one can find the Bayesian point estimate of the success probability p in a binomial distribution by assuming the prior distribution of p to be the beta distribution with parameters α and β (conjugate prior). The Bayesian point estimate of p is given by $w\hat{p} + (1-w)\alpha/(\alpha+\beta)$, the weighted average of the MLE of p and the mean of the prior distribution of p. One can find the credible interval of p through the posterior distribution of p.

7. *An overall comment*: You tried various methods on Earthquake data. What are the conclusions about the predictive power of individual models/algorithms? If they differ, justify these differences based on your understanding of the models/algorithm. Please provide a summary table with a few paragraphs of interpretation to make a complete case study (example).