Stat 238, Fall 2025
Project Minis

Alexander Strang
Due: by **5:00 pm** Friday, March 7, 2025

Mini Project 2: Consistency, Inconsistency, Ishconsistency

In class we sketched the general asymptotic theory for posterior inference under "mild" regularity conditions. We've only discussed these informally.

In this project, you will be asked to identify and test the precise regularity conditions needed for the key theorems in BDA Chapter 4 and Appendix B. These theorems are, loosely:

1. *"Posterior point estimators (e.g. MAP) are consistent and asymptotically unbiased."*
2. *"The posterior concentrates about the truth, when available, and about a best fit to the truth when misspecified."*
3. *"The posterior distribution is asymptotically normal."*
4. *"The posterior distribution is asymptotically exponential in the KL divergence between the true data generating distribution and the conditional distribution proposed for the data under the model."*[1]

In this project, you will explore the limits of these statements.

Policies

- You may work with up to **2 partners**. To find partners, talk to your friends in class, or post on the Ed thread associated to this project mini.
- All submissions must be properly type-set and sourced.
- Project minis should be uploaded to Gradescope under the appropriate mini-assignment. If you work in a group (eligible projects are marked), you should submit one assignment, and link your partners. Please do not submit duplicated assignments.

Part I: Regularity Conditions

(a) Find, and state precisely, the minimal, sufficient, regularity guarantees needed for each of the four statements provided above. Make sure you clearly define any element that has been left loose in class, or that is new. Plainly cite the sources you use.

[1] (see Gelman and Shalizi, Philosophy and Practice of Bayesian Statistics.)

(b) Comment on the plausibility of each regularity condition. If you disagree as a group, record some of your debate. Try to identify an example where each might usually fail.[2]

(c) Elaborate (1 to 2 paragraphs) on one of your proposed examples. You will be asked to try and demonstrate this example at the end of the project.

Part II: Misspecification Gaurantees

The generic asymptotic theory applies in both the well and misspecified case. However, when misspecified, the posterior cannot concentrate about the truth. Instead, it concentrates about a model that is as hard to distinguish from the truth as possible.

In this part, you will explore the sense in which the posterior concentrates about the "best fit" parameters, and, whether the resulting "best fit" guarantees ensure that inferences made using misspecificed models will have small errors. In most tellings, the "best fit" model is "best fit" in the sense that it minimizes a Kullbach-Liebler (KL) divergence. Let's try to make sense of this notion.

(a) Is concentrating to the minimizer of the KL equivalent to concentrating to the MLE? Do the Frequentist and Bayesian inference procedures still agree in an infinite data limit?

(b) Relate the KL divergence between distributions p and q to the expected value of the log-likelihood-ratio test statistic used in a binary hypothesis test between p and q. In what sense does your answer imply that, if the KL divergence between two distributions is large, then they are easy to distinguish?

(c) The asymptotic theory states that the posterior concentrates about the parameters θ_* which minimize the KL divergence between the true data generating distribution, f_Y, and the modeled distribution $p_{Y|\theta}$. Use the Chernoff-Stein Lemma to translate this result into a conclusion regarding the distinguishability of the two distributions in a hypothesis test. Specifically, identify the sense in which minimizing the KL divergence selects the model $p_{Y|\theta}$ which is hardest to distinguish from the f_Y.

(d) Since we can always test for the difference between two distributions by comparing a hypothesized sampling distribution for a test statistics to the observed test statistic, argue that, if we consider a sequence of nested model families that approach a family including the true model, then, as the model families expand, the amount the sampling distribution of any test statistic under the true data generating process differs from its sampling distribution under the KL minimizing model, must vanish.

(e) Put your arguments together to argue that, if we use the KL between $p_{Y|\theta_*}$ and the f_Y to measure the degree of misspecification, then mild misspecification will induce mild errors in inference in the sense that the proposed generating distributions for the data, and sampling distribution for any tests statistic, are hard to distinguish.[3]

[2]Note: consistency is an asymptotic guarantee. We never collect an infinite amount of data, so the actual utility of an asymptotic theory follows from its usefulness in providing approximations that hold for sufficiently large data sets. The closer a problem gets to violating the regularity conditions, the more samples are usually needed before the asymptotic theory provides useful approximations.

[3]As usual, beware the allure of asymptotics... really we should want to know what KL divergence is *small enough* to ensure all other errors are small.

(f) Does this guarantee also imply that the best fit parameters are close to the true parameters if the model could be extended to include a parameterization of the f_Y? Try to identify a counter-example if your answer is no. You will be asked to try and test this idea using your counter-example in Part III. Discuss when errors in the inferred unknowns do or do not matter.[4]

(g) Assume, for this part, that θ_* is on the interior of the set of possible parameters. Taylor expand the KL divergence between $p_{Y|\theta}$ and the f_Y to second order in θ about θ_*. Then substitute into statement 4 on page 1 to derive an asymptotic approximation to the posterior. Does this approximation agree with our standard asymptotic approximation?

Part III: Exploration

Now that we have a rich idea for the theory, let's inspect it through example. Propose two Bayesian models and use them to demonstrate some of the ideas you discussed above.

- One should be a conjugate model, where, to study misspecification, you fix a subset of the prior parameters to the wrong values, and let the rest vary. Your model should be a member of a broader conjugate class which does include the truth. For example, consider a conjugate Bayesian model with a normal likelihood, normal prior on the mean, and with fixed, but incorrect covariance.
- One should be a non-conjugate model. It helps, as before, to choose a model parameterization that could include the truth, but, when you perform inference, you incorrectly fix some parameters. For example, you could try a logistic generalized linear model where you exclude necessary predictor variables, or, a Gaussian mixture model where you use the wrong number of mixture modes.

(a) Propose a pair of models and explain why you think they are interesting.

(b) Propose some characteristics of your posterior distributions that you might care about, or might return as inferential summaries (e.g. the point estimators, a notion of an interval estimator, a notion of spread). You can use discrepancies in these characteristics as both measures of the difference in two distributions, and, as practical measures of the inferential error incurred when using the wrong distribution.

(c) Implement code that can (a) generate data from your models, (b) perform posterior inference, and (c) compute or approximate the quantities you identified as your posterior summaries. You are welcome to use existing packages to implement your models.

(d) Try to illustrate the importance of the regularity condition you identified as suspect in Part I. Was your concern justified in the case you implemented? How strongly does your example violate the asymptotic theory? If it doesn't violate the asymptotic theory exactly, but is close to violating the asymptotic theory, discuss how many samples are needed before the asymptotic theory gives reasonable approximations.[5]

(e) Try to illustrate the biasing effects of misspecification. Does a small misspecification (in the sense of small KL) imply that the ensuing errors in posterior inference are small?

[4] Is your focus inference or prediction?

[5] See, for example, HW 2 problem 1.