Bayesian Inference

- So far we have been working with frequentist inference
    - The interpretation of an event with probability 0.9 is that if we repeat it many times, then the event will occur 90% of the tiem
- Bayesian approach is alternative method to produce estimators / confidence intervals / tests

Bayesian Method

- In the Bayesian world, we first have aprior pdf $f(\theta)$ which indicates our prior belief about θ before seeing the data
- After seeing the data, we update it into a posterior belief

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})}$$

- The numerator is just $L_n(\theta) * f(\theta)$ and the bottom is a normalization constant that does not depend on θ
- Therefore, we write our posterior distribution as:

$$f(\theta|\text{data}) \propto L_n(\theta)f(\theta)$$

- If the prior and posterior turn out to be in the same family of distributions, we call the prior a **conjugate prior**

Bayes Estimator

Bayes Estimator

The Bayes estimator is the mean of the posterior, or mean a posteriori

$$\hat{\theta}^{\text{Bayes}} = \int \theta \cdot f(\theta|\text{data}) d\theta$$

This can be hard to compute explicitly because it requires knowing the normalizing constant

- Can approximately compute it using Markov Chain Monte Carlo
    - Markov Chain: draw $\theta_1, \ldots, \theta_T$ approximately iid from $f(\theta|X_1, \ldots, X_n)$ by simulating a Markov chain
        - This can be done through something like Metropolis Hastings, which requires only the ratio between states, not the actual probability value of the state
    - Monte Carlo: compute the mean of the θ_i drawn

MAP (mode)

MAP

The maximum a posteriori, or MAP, is the mode of the posterior:

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} f(\theta|\text{data})$$

The mode and mean do not necessarily have to be close (i.e. think about a bimodal posterior)

To compute the MAP, we can maximize the log posterior, which turns out to maximize:

- $\ell_n(\theta) + \log f(\theta) - \log c_n$
- Recall that ℓ_n is the log likelihood and c_n is a constant
- This can also be maximized via gradient ascent