STCS 6701 Recitation (9/15)

Bayesian Statistics Review

Probability Fundamentals

Random Variables

A random variable X stochastically takes values according to either a:

- PMF: probability mass function (discrete)
- PDF: probability density function (continuous)

Can also think of a random variable as a function:

$$p : \Omega \rightarrow [0, 1], s.t. \int_{\Omega} p(x)dx = 1$$

Moments

Expectation:

$$\mathbb{E}[X] = \int_{\Omega} xp(x)dx$$

Variance:

$$\operatorname{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Joint Distributions

A new r.v. induces a joint distribution:

$$p(X, Y)$$

The joint can always be broken up as:

$$p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X)$$

If they are independent, then you can reduce this to:

$$p(X, Y) = p(X)p(Y)$$

Conditioning

Conditioning on another random variable:

$$p(X|Y) = \frac{p(X,Y)}{p(Y)}$$

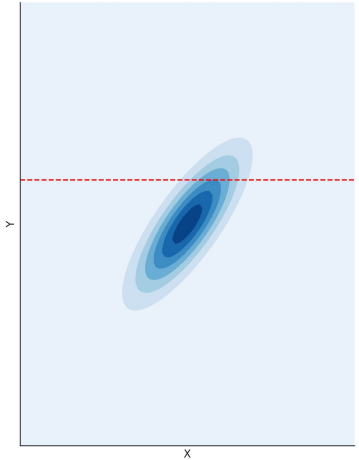
Conditional Expectation:

$$\mathbb{E}[X|Y=y] = \int_{\Omega} xp(X|Y=y)dx$$

Tower Rule:

$$\mathbb{E}_{Y\sim p(y)}[\mathbb{E}_{X\sim p(x)}[X|Y=y]] = \mathbb{E}[X]$$

Marginalization



If you want to get the marginal distribution of one r.v.

given the joint:

$$p(X = x) = \int p(X = x, Y = y)dy$$

A handy trick is to introduce an r.v. then marginalizing it out:

$$p(X = x) = \int p(X = x, Y = y)dy = \int p(X = x|Y = y)p(Y = y)dy$$

Conditional Independence

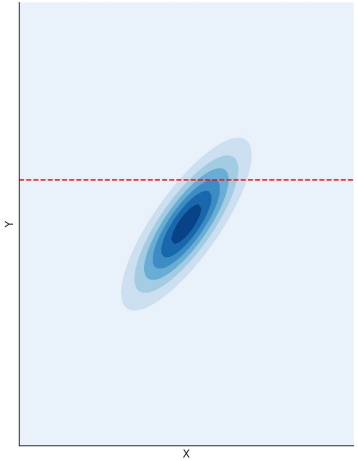
Concretely:

$$X \perp Y|Z \rightarrow p(X,Y|Z) = p(X|Z)p(Y|Z) \neq X \perp Y$$

Take trials of a coin with unknown probability p of hitting heads:

- Is one flip independent of another?
- How about conditioned on knowing the probability of heads?

Dave's Notation on Conditioning

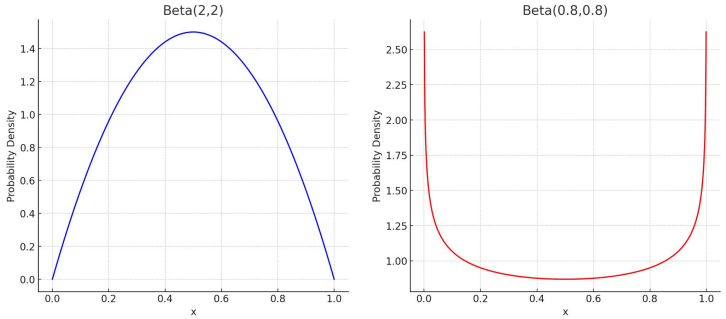


Note the difference between a | and a ;

A bar implies the conditioning on a random variable:

A semi-colon implies a dependence on a non-stochastic variable (usually set by the statistician):

$$X \sim \text{Beta}(\alpha, \beta)$$



$p(X; \alpha = 2, \beta = 2)$ $p(X; \alpha = 0.8, \beta = 0.8)$

Bayes' Theorem

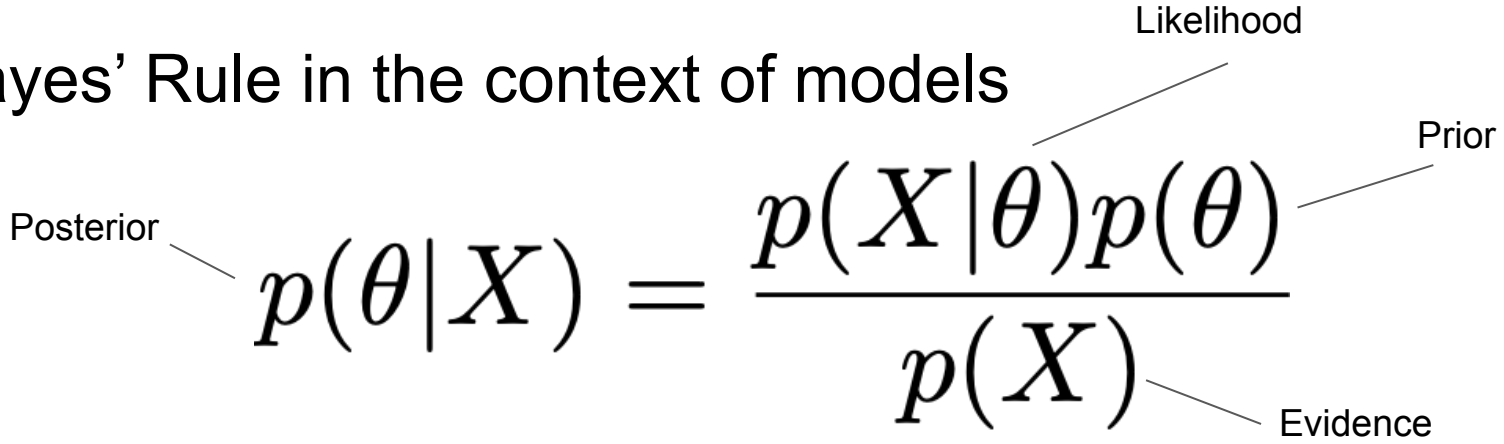
$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

Notice this is just an expansion of the conditional probability definition.

Importantly, this theorem helps us understand unknown quantities given known observations in a probabilistic framework.

What is Bayesian Statistics?

Bayes' Rule in the context of models



Let X represent the data we observe.

Let θ represent the parameters of our model of the world.

Objectives of Bayesian Statistics

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

We want to understand θ

We construct a prior that reflects our prior beliefs.

We construct a model that reflects our understanding of how the unknown parameters relate to the data.

Apply Bayes' to compute the posterior!

….. ?????.... profit?

Bayesian vs. Frequentist schools of thought

A frequentist also wants to understand θ

So what's the big deal?

Frequentists consider the parameter to be fixed.	Bayesians consider the parameter to be randomly sampled from a distribution.
Mainly consider what action to take as a result of the data.	Consider how the data affects their understanding of the parameter distribution.
The statistical tests are tuned according to what level of risk you would like to take.	Hope your model fits and compare likelihoods.
	Need to choose a prior.

Credible Intervals (how this differs from CIs)

Confidence intervals (frequentist concept)

- 95% CI denotes an interval for which 95% of the time across repeated trials, would contain the true parameter.

Credible Intervals (bayesian concept)

- 95% Credible interval denotes an interval which the parameter will exist in 95% of the time according to the model.
- Notice, there can be multiple credible intervals.

Modern Problems with Bayesian Inference

Beta-Binomial Revisited

We were able to compute the posterior of the Beta-Binomial in class analytically. So when does Bayesian statistics become difficult?

Let’s revisit what happened with the Beta-Binomial (where is the trick?):

-> do on the board

Problems with more complicated models

When does this become difficult to compute?

$$p(X) = \int p(X|\theta)p(\theta)d\theta$$

- Do we even need it?
- Cannot do integral analytically
  - Then why not use a computer?
- Cannot compute the integral numerically in a reasonable amount of time

Posterior Approximation Methods

Objectives and things to consider

If we can’t compute the posterior, can we approximate it?

- What is the source of the approximation gap?
- Can we understand how close we are to the true posterior?
- How much compute does it take to get a good approximation?
- Does our approximation method introduce unwanted bias?

Posterior Approximation Methods

Some examples of posterior approximation methods are:

- Markov Chain Monte Carlo (MCMC)
  - Gibbs Sampling
  - Metropolis-Hastings
- Variational Inference
  - Variational EM
  - Auto-encoding Variational Bayes

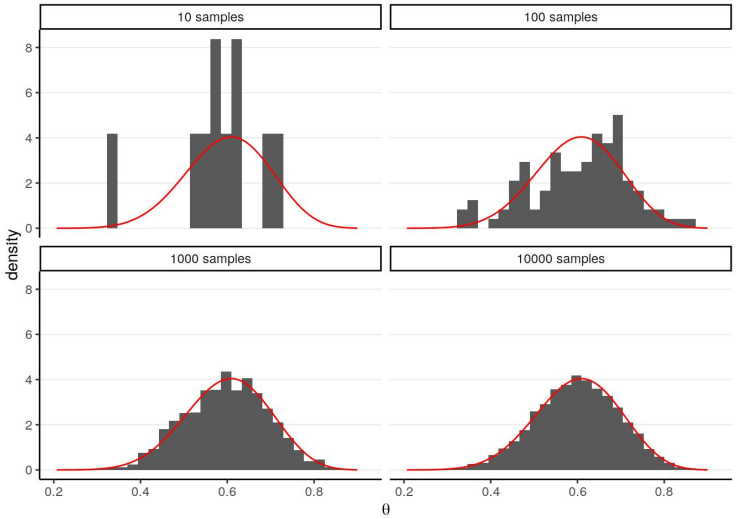
Markov Chain Monte-Carlo (MCMC)

Main idea:

Construct a markov chain (set of nodes and transition probabilities) such that after sampling for a long time, the sampling distribution converges to the true posterior.

Gibbs sampling is one such method that

will be covered in the class.

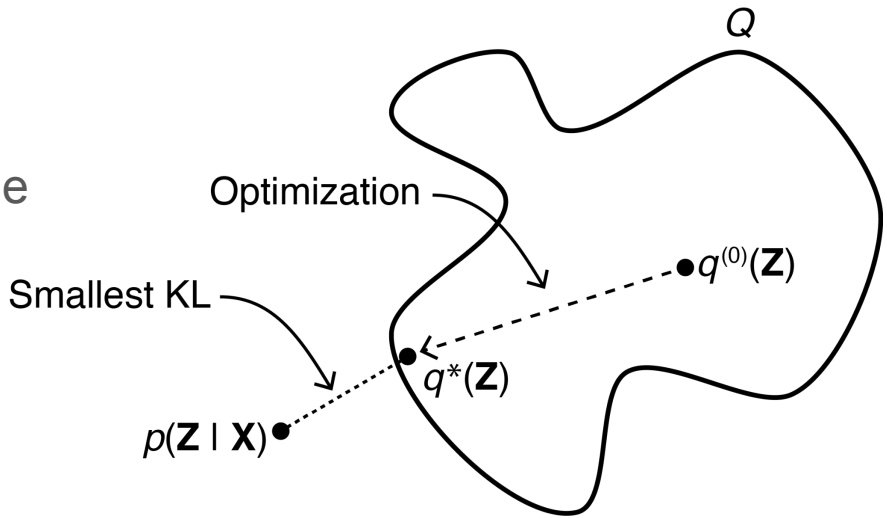


Variational Inference

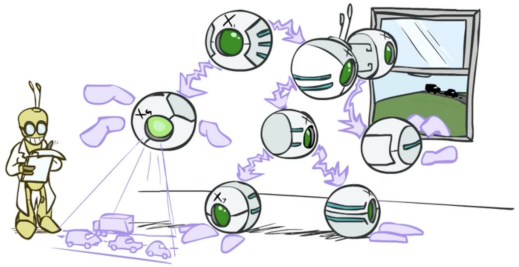
Main idea:

Within a parametrized family of distributions, optimize the parameters to get as close to the true posterior.

The measure comparing the distributions is called the Kullback-Leibler divergence (KL).



Pros and Cons



MCMC

- Pros:
  - Can sample from the true posterior (assuming convergence).
- Cons:
  - Can be compute intensive.
  - Hard to determine whether the chain has converged.

Pros and Cons

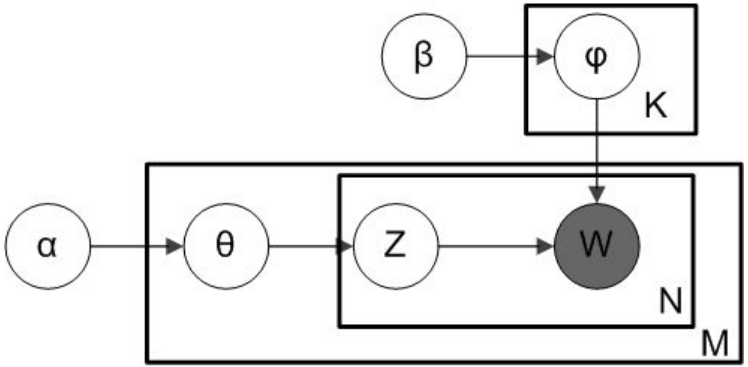
VI

- Pros:
  - Usually faster than MCMC.
  - Can employ tricks like stochastic variational inference to operate over huge datasets.
- Cons:
  - True posterior inferred only if the family Q contains the true posterior…which is almost never.

How this ties into the class

What are PGMs good for?

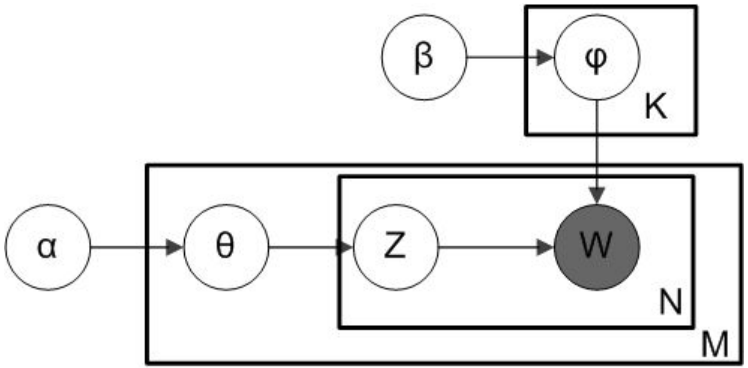
- We can express complicated models and conditional independences of a model easily.
- Coupled with posterior approximation models, we can create and fit complex models quickly.
    - Modern algorithms allow us to do this without writing out the math!



Conditional Independence in PGMs

Just from the diagram below, what are some conditional independences we can identify?

What does this tell us about the model?



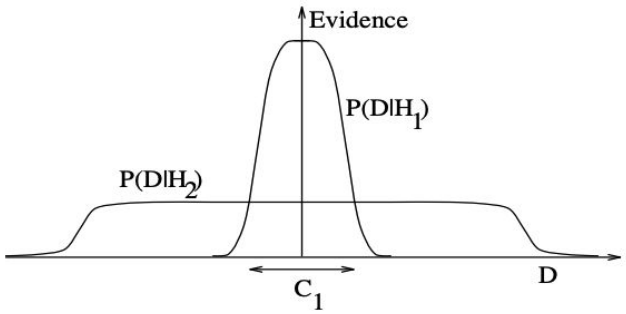
More variables, more problems…

Be warned, having more parameters is not always a good thing.

Bayesian Occam's Razor states that you should choose the simplest model that fits the data.

Bayesian statistics automatically penalizes unnecessarily complicated models due to the stretching of mass over the prior distribution.

Optimization, is a whole nother thing..

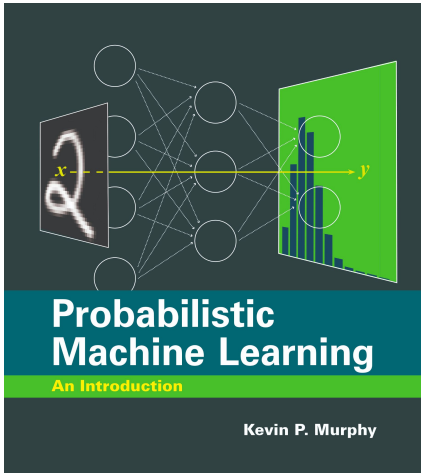


Probabilistic Programming Languages

- Provides a high-level framework for expressing probabilistic models and performing inference.
- With probabilistic models, one can:
  - Represent unknowns as random variables.
  - Incorporate prior beliefs into the model.
  - Formally define the uncertainty of parameter estimates.

Resources for further reading

- Bertsekas and Tsitsiklis, Introduction to Probability
- Kevin Murphy, Probabilistic Machine Learning: An Introduction
- Dave’s notes!



Thanks for listening!