COMPUTER SCIENCE

INDIANA UNIVERSITY

School of Informatics and Computing
Bloomington

Parameter estimation

CSCI-B 555

Reminders

- No class on Monday
  - my office hours are moved to Wednesday
- Thought questions #1 due next week (on Wednesday)
- Assignment #1 is due in two weeks (on Wednesday)
  - Derive Figure 1.5 (B)
- Introduction to probability done
  - I will sprinkle in exercises during class to give you more practice
  - It will make even more sense when you apply it
  - Note: I will be repetitive because its useful for learning

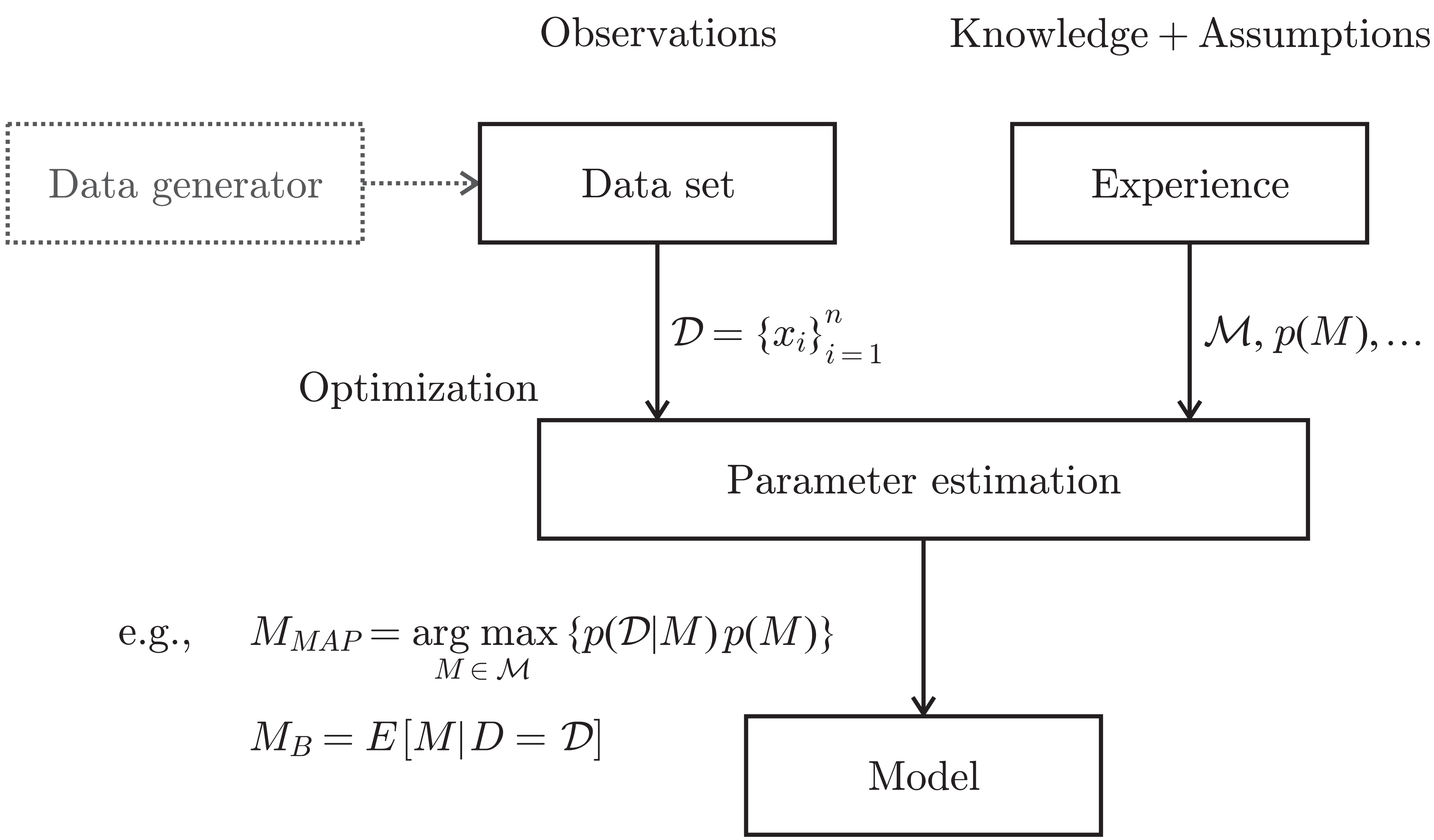
Summary on models

- We specify random variables and corresponding distributions
  - enables precise definition of uncertainty in the world, measurements, etc.
- Joint distributions and conditional distributions
  - generative versus discriminative (we will discuss these more)
- Parametric and non-parametric
  - If parametric, then many known/useful PDFs and PMFs
- Now want a way to use data to inform models
- **Note**: I do not expect you to be an expert in all the PMFs and PDFs discussed; they are mostly introduced as options
  - we will continue to deal with pdfs more abstractly, with general principles

Parameter estimation

- Assume that we are given some model class, M,
  - e.g., Gaussian with parameters mu and sigma
  - selection of model from the class corresponds to selecting mu, sigma
- Now want to select “best” model; how do we define best?
  - Generally assume data comes from that model class; might want to find model that best explain the data (or is most likely given the data)
  - Might want most likely model, that also matches expert prior info
  - Might want most likely model, that is the simplest (least parameters)
- These additional requirements are usually in place to enable better generalization to unseen data

How can we incorporate data?



Model inference: Observations + Knowledge and Assumptions + Optimization

Maximum a posteriori (MAP) estimation

$$M_{MAP} = \arg\max_{M \in \mathcal{M}} \{p(M|\mathcal{D})\}$$

- p(M | D) is the **posterior distribution** of the model given data

- In discrete spaces: p(M | D) is the PMF
  - the MAP estimate is exactly the most probable model

- In continuous spaces: p(M | D) is the PDF
  - the MAP estimate is the model with the largest value of the posterior density function

MAP calculation

- Start by applying Bayes rule

$$p(M|\mathcal{D}) = \frac{p(\mathcal{D}|M) \cdot p(M)}{p(\mathcal{D})},$$

- p(D I M) is the **likelihood** of the data, under the model
- P(M) is the **prior** of the model
- P(D) is the marginal distribution of the data (also called the evidence)
  - we will often be able to ignore this term

Why is this conversion important?

- Do not always have a known form for P(M | D)
- We usually have chosen (known) forms for P(D | M) and P(M)
- **Example**: Let D = {x1} (one sample). Then one common choice is a Gaussian over x1: P(D | M) = P(x1 | mu, sigma)
  - p(M | D) is not obvious, since specified our model class for P(D | M)
  - What is p(M) in this case? We may put some prior “preferences” on mu and sigma, e.g., normal distribution around mu, specifying that really large positive or negative mu is unlikely
  - Specifying and using p(M) is related to regularization and Bayesian parameter estimation, which will will discuss more later

Why is this conversion important?

- **Example**: Let D = {x1,x2} (two samples).
- If x1 and x2 are independent samples from same distribution (same model), then P(x1, x2 | M) = P(x1 | M) P(x2 | M)
- For many iid samples x1, …, xn, we could choose (e.g.,) a Gaussian distribution for P(xi | M), with M = {mu,sigma}
  - iid = independent and identically distributed
  - P(x1, …, xn | M) = P(x1 | M) … P(xn | M)

Data marginal

- Using the formula of total probability

$$p(\mathcal{D}) = \begin{cases} \sum_{M \in \mathcal{M}} p(\mathcal{D}|M)p(M) & M : \text{discrete} \\ \\ \int_{\mathcal{M}} p(\mathcal{D}|M)p(M)dM & M : \text{continuous} \end{cases}$$

- Fully expressible in terms of likelihood and prior

Chain rule:

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

$$P(X,Y,Z) = P(X|Y,Z)P(Y,Z) = P(X|Y,Z)P(Y|Z)P(Z)$$

$$P(X,Y,Z) = P(X,Y|Z)P(Z) = P(X|Y,Z)P(Y|Z)P(Z)$$

Exercise: conditional independence

- Recall the example with the coin: Z is an RV that is the bias of a coin, and X and Y are two independent flips of the coin
- With your improved knowledge of marginals/probability rules, write P(X=1, Y=1) in terms of P(X|Z=c), P(Y|Z=c) and P(Z=c)
- Now for simplicity, let C = {0.2, 0.8} with p(0.2) = 0.5 = p(0.8).
  - Compare P(X=1, Y=1) to P(X=1) P(Y=1). What do you notice?

Total probability:

$$P(X) = \sum_y P(X, Y = y) = \sum_y P(X|Y = y)P(Y = y)$$

Chain rule:

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

Optimization to get model

$$M_{MAP} = \underset{M \in \mathcal{M}}{\arg\max} \left\{ \frac{p(\mathcal{D}|M)p(M)}{p(D)} \right\}$$

$$= \frac{1}{p(D)} \underset{M \in \mathcal{M}}{\arg\max} \{p(\mathcal{D}|M)p(M)\}$$

$$= \underset{M \in \mathcal{M}}{\arg\max} \{p(\mathcal{D}|M)p(M)\}$$

Will often write:

$$p(M|\mathcal{D}) = \frac{p(\mathcal{D}|M) \cdot p(M)}{p(\mathcal{D})}$$

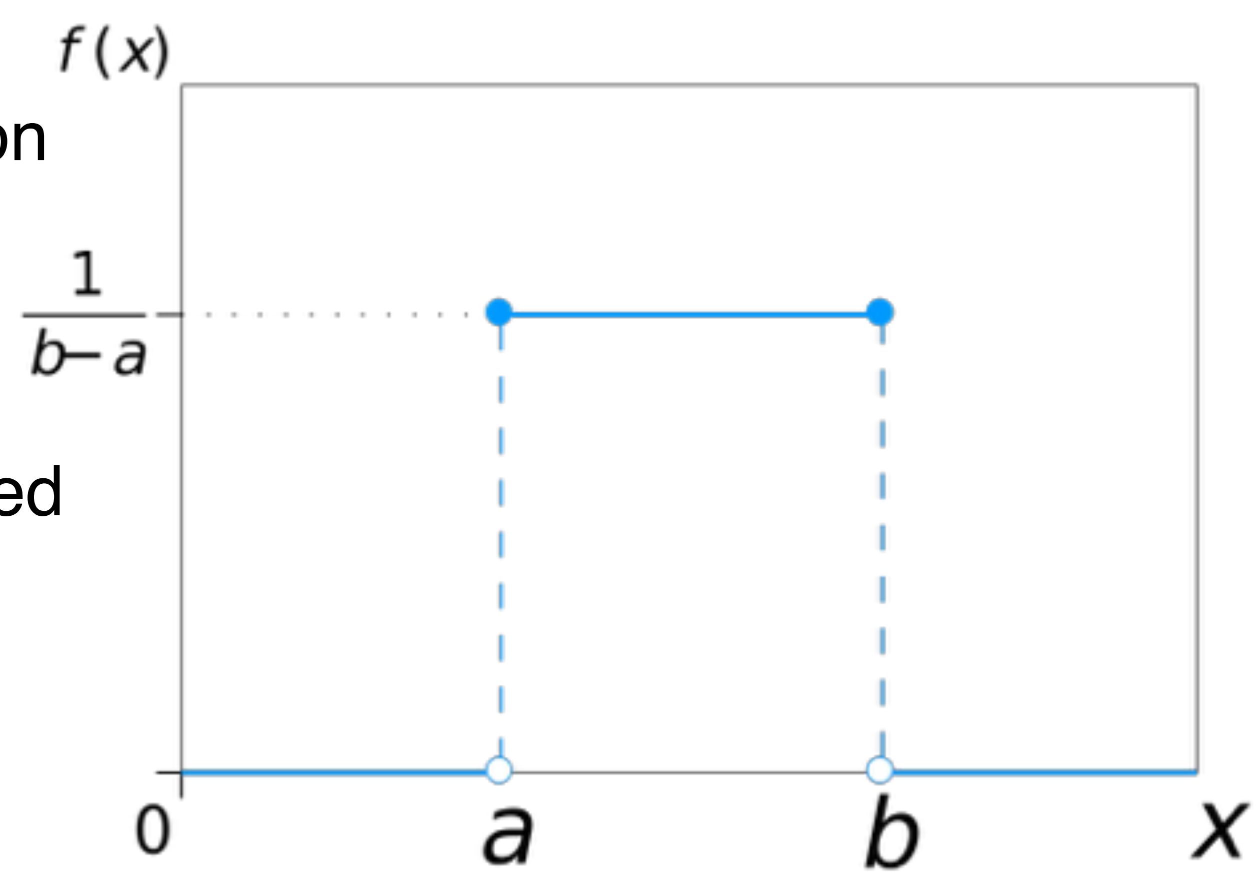
$$\propto p(\mathcal{D}|M) \cdot p(M),$$

Maximum likelihood

$$M_{ML} = \arg\max_{M \in \mathcal{M}} \{p(\mathcal{D}|M)\}.$$

- In some situations, may not have a reason to prefer one model over another (i.e., no prior knowledge or preferences)

- Can loosely think of maximum likelihood as instance of MAP, with uniform prior
  - If domain is infinite (example, the set of reals), the uniform distribution is not defined!
  - but the interpretation is still similar
  - in practice, typically have a bounded space in mind for the model class



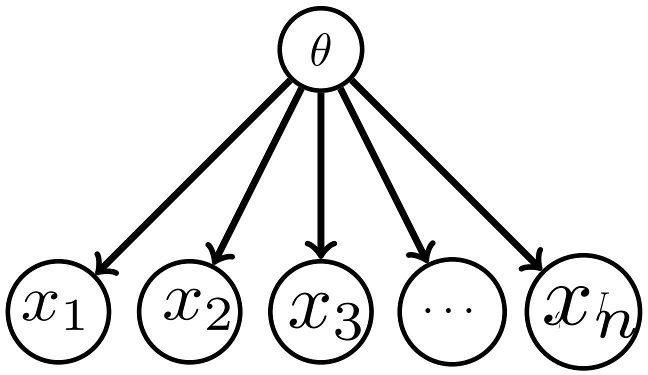
Example: maximum likelihood for discrete distributions

- Imagine you are flipping a biased coin; the model parameter is the bias of the coin, theta

- You get a dataset D = {x_1, …, x_n} of coin flips, where x_i = 1 if it was heads, and x_i = 0 if it was tails

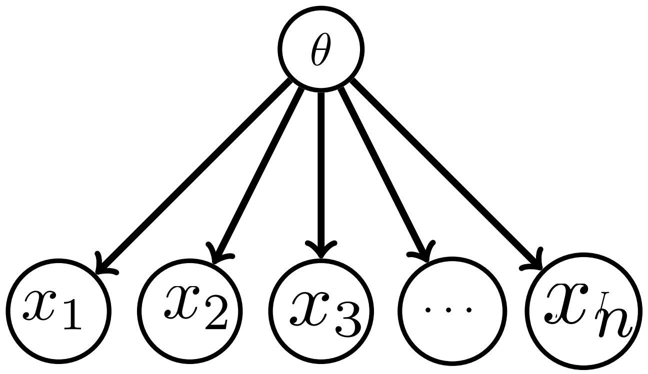
- What is p(D I M)?

$$p(D|M) = p(x_1, \ldots, x_n|\theta)$$
$$= \prod_{i=1}^{n} p(x_i|\theta)$$
$$p(x_i|\theta) = \theta$$



Example: maximum likelihood for discrete distributions

- How do we estimate theta?



- Counting:
  - count the number of heads Nh
  - count the number of tails Nt
  - normalize: theta = Nh/ (Nh + Nt)

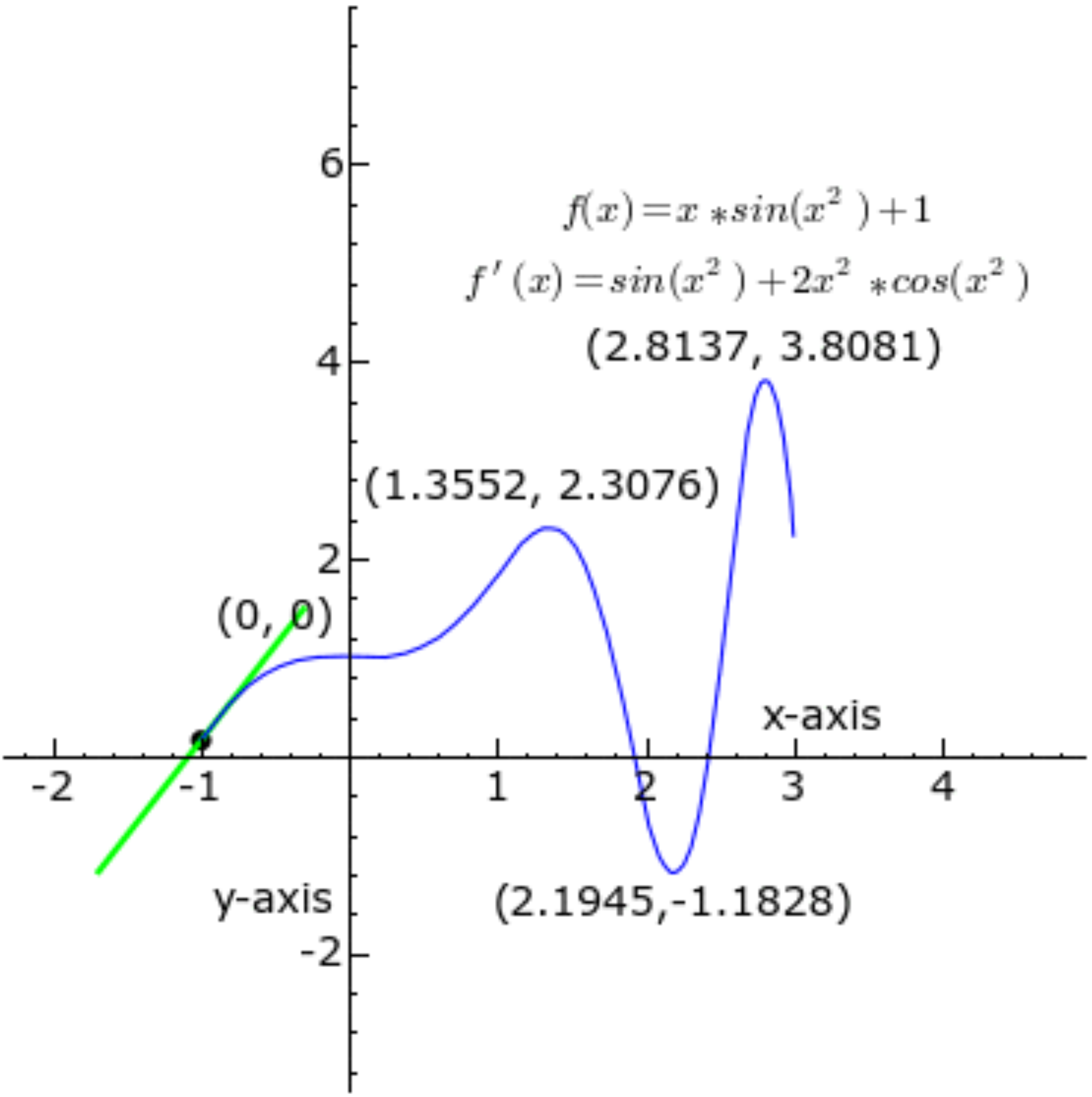
$$p(D|M) = p(x_1, \ldots, x_n|\theta)$$

$$= \prod_{i=1}^{n} p(x_i|\theta)$$

$$p(x_i|\theta) = \theta$$

- What if you actually try to maximize the likelihood?
  - i.e., solve argmax p(D I theta)

Single-variate calculus



GIF from Wikipedia: Tangent

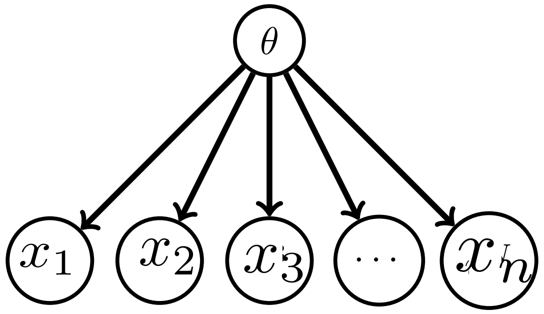
For a function f defined on a scalar x, the derivative is

$$\frac{df}{dx}(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

At any point, x, $\dfrac{df}{dx}(x)$ gives the slope of the tangent to the function at f(x)

Example: MAP for discrete distributions

- Imagine you are flipping a biased coin; the model parameter is the bias of the coin, theta
- You get a dataset D = {x_1, …, x_n} of coin
1 if it was heads, and x_i = 0 if it was tails

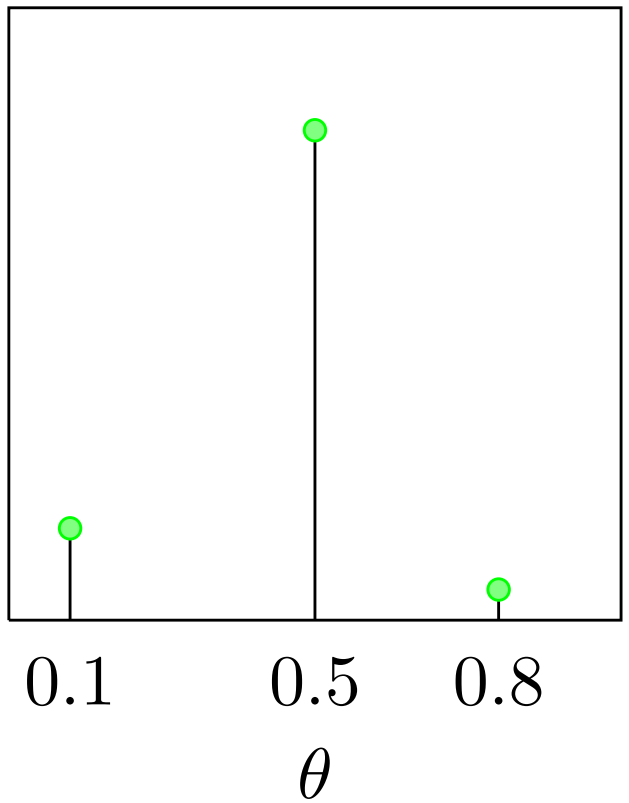


- What if we also specify p(M)?
- What is the MAP estimate?

Example: maximum likelihood for discrete distributions

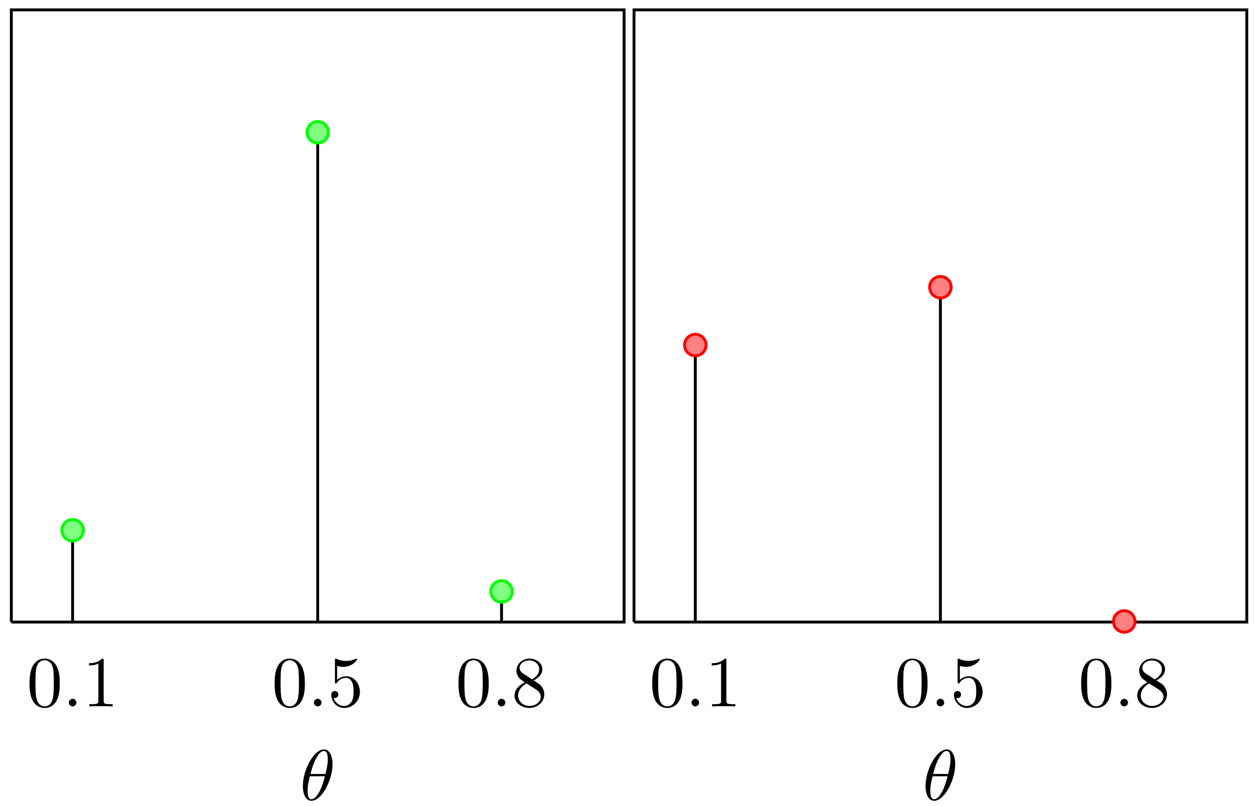
We still need to fully specify the prior $p(\theta)$. To avoid complexities resulting from continuous variables, we'll consider a discrete θ with only three possible states, $\theta \in \{0.1, 0.5, 0.8\}$. Specifically, we assume

$$p(\theta = 0.1) = 0.15,\ p(\theta = 0.5) = 0.8,\ p(\theta = 0.8) = 0.05$$



Example: maximum likelihood for discrete distributions

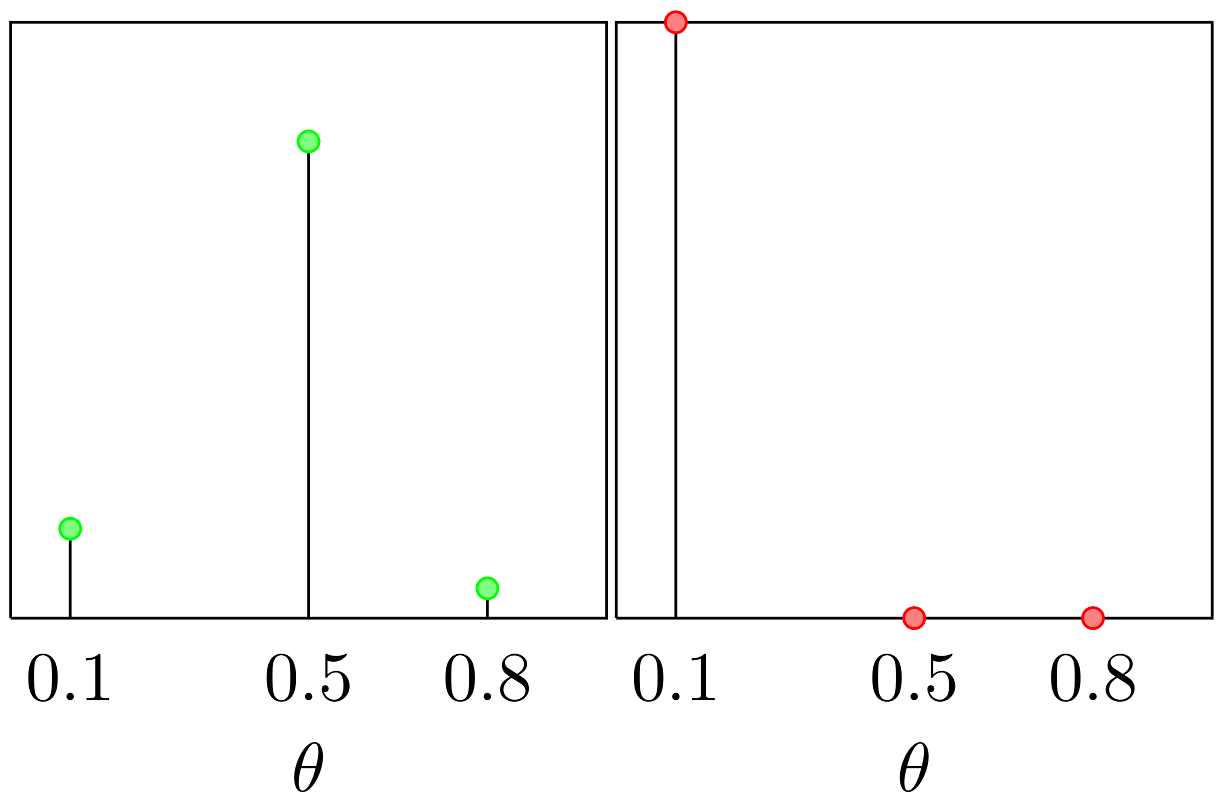
For an experiment with $N_H = 2$, $N_T = 8$, the posterior distribution is



If we were asked to choose a single *a posteriori* most likely value for θ, it would be $\theta = 0.5$, although our confidence in this is low since the posterior belief that $\theta = 0.1$ is also appreciable. This result is intuitive since, even though we observed more Tails than Heads, our prior belief was that it was more likely the coin is fair.

Example: maximum likelihood for discrete distributions

Repeating the above with $N_H = 20$, $N_T = 80$, the posterior changes to



so that the posterior belief in $\theta = 0.1$ dominates. There are so many more tails than heads that this is unlikely to occur from a fair coin. Even though we *a priori* thought that the coin was fair, *a posteriori* we have enough evidence to change our minds.

Now on to some careful examples of MAP!

- Whiteboard time for Examples 8, 9, 10
- More fun with derivatives and finding the minimum of a function
- Next class:
  - finish off parameter estimation
  - introduction to prediction problems for ML