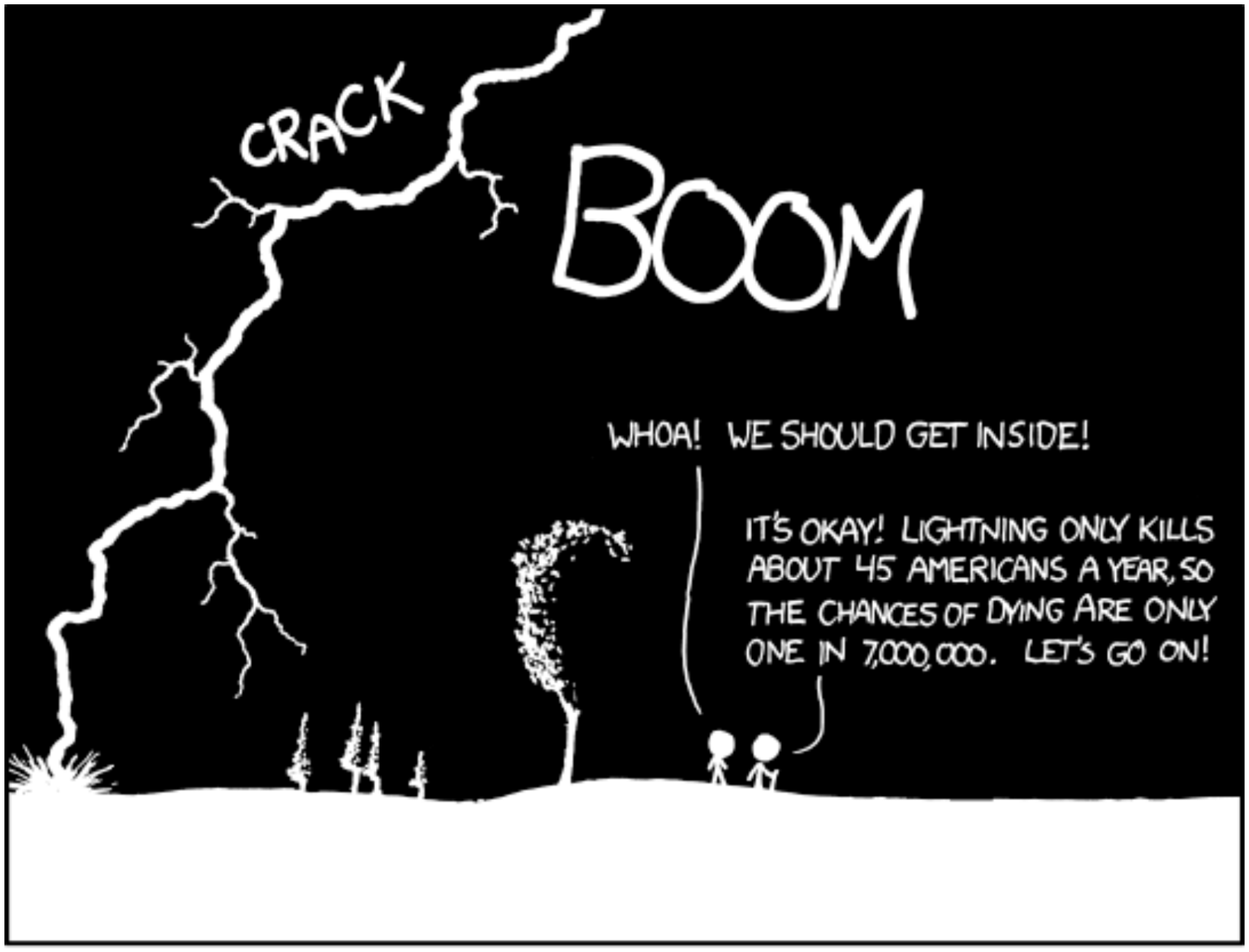
Parameter estimation

Conditional risk



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Formalizing the problem

- Specify random variables we care about
  - e.g., Commute Time
  - e.g., Heights of buildings in a city
- We might then pick a particular distribution over these random variables
  - e.g., Say we think our variable is Gaussian
- Now want a way to use the data to inform the model choice
  - e.g., Let data tell us the parameters for that Gaussian

Parameter estimation

- Assume that we are given some model class, M,
  - e.g., Gaussian with parameters mu and sigma $\mathcal{N}(\mu, \sigma^2)$
  - 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 explains the data (or most likely given the data)
  - Might want most likely model, with preference for “important” samples
  - 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 identify the parameters for a distribution?

- All we get is access to samples x1, …, xn from the true underlying distribution p(x)
- We want to find parameters theta so that p_theta approximates the true distribution p well
  - e.g., theta is the mean and variance for a Gaussian
  - e.g., theta is the parameters to a Gamma distribution (commute times)
- Any ideas?

What about in the prediction setting? That feels easier…

- Here, the goal is to learn a function f for some inputs x to make predictions about targets y

- How do we identify the "best" f in our set of functions F?

$$\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \mid f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \text{ for some } \mathbf{w} \in \mathbb{R}^d\}$$

- Any ideas?
  - Maybe $\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$

  - What does that entail? What if y_i is always 0 or 1?

Using the language of probability

- It can be hard to guess an objective function that tells you what is the best model in your class

- A natural objective is: find the model that is the most likely given the data

$$\arg\max_{\theta} p(\theta|\mathcal{D})$$

- Reflects the inherent uncertainty in identifying the model, since many models are possible given only samples

Some notation

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

$$\prod_{i=1}^{n} x_i = x_1 x_2 \ldots x_n$$

$$c : \mathbb{R}^d \to \mathbb{R}$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w} \in \mathbb{R}^d} c(\mathbf{w})$$

$$c(\mathbf{w}^*) = \max_{\mathbf{w} \in \mathbb{R}^d} c(\mathbf{w})$$

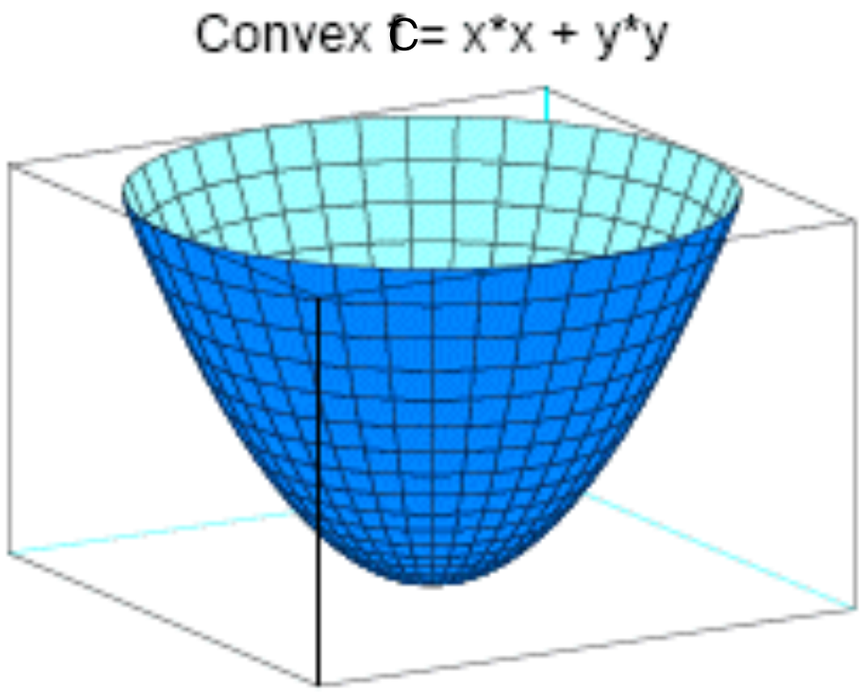
$\mathcal{F}$ is a set of models

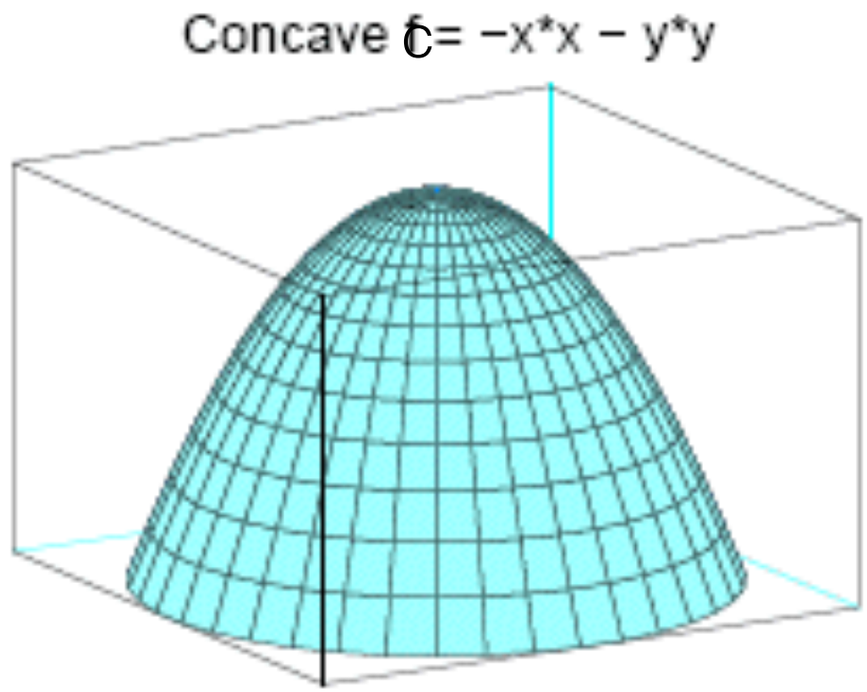
e.g., $\mathcal{F} = \{\mathcal{N}(\mu, \sigma) \mid (\mu, \sigma) \in \mathbb{R}^2, \sigma > 0\}$

e.g., $\mathcal{F} = \{\mathbf{w} \in \mathbb{R}^d \mid f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}\}$

Definition of optimization

- We select some (error) function c we care about





- Maximizing c means we are finding largest point
- Minimizing c means we are finding smallest point

Maximum a posteriori (MAP) estimation

$$f_{\mathrm{MAP}} = \arg\max_{f \in \mathcal{F}} p(f|\mathcal{D})$$

- Want the f that is most likely, given the data
- p(f | D) is the **posterior distribution** of the model given data
  - e.g., F could be the space of Gaussian distributions, the model is f and f(x) returns probability/density of a point x
  - e.g., we could assume x is Gaussian distributed with variance = 1, and so F could be the reals, and the model f is the mean

Question: What is the function we are optimizing and what are the parameters we are learning?

MAP

$$f_{\mathrm{MAP}} = \arg\max_{f \in \mathcal{F}} p(f|\mathcal{D})$$

- p(f | D) is the **posterior distribution** of the model given data
  - e.g., we could assume x is Gaussian distributed with variance = 1, and so F could be the reals, and the model f is the mean

Question: What is the function we are optimizing and what are the parameters we are learning?

$$c(f) = p(\text{mean is } f|\mathcal{D})$$

$$\max_{f \in \mathbb{R}} c(f)$$

Maximum a posteriori (MAP)

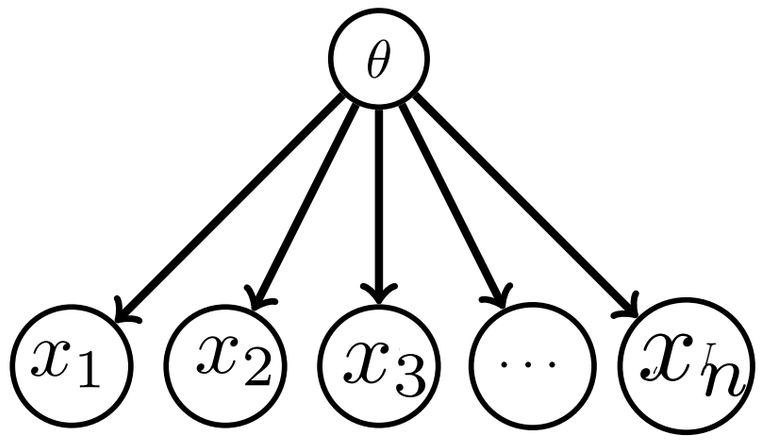
$$f_{\mathrm{MAP}} = \arg\max_{f \in \mathcal{F}} p(f|\mathcal{D})$$

- p(f | D) is the **posterior distribution** of the model given data
- In discrete spaces: p(f | D) is a PMF
  - the MAP estimate is exactly the most probable model
  - e.g., bias of coin is 0.1, 0.5, or 0.7, p(f = 0.1 | D), …
- In continuous spaces: p(f | D) is a PDF
  - the MAP estimate is the model with the largest value of the posterior density function
  - e.g., bias of a coin is in [0, 1]
- But what is p(f | D)? Do we pick it? If so, how?

Example: Posterior for discrete distributions

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

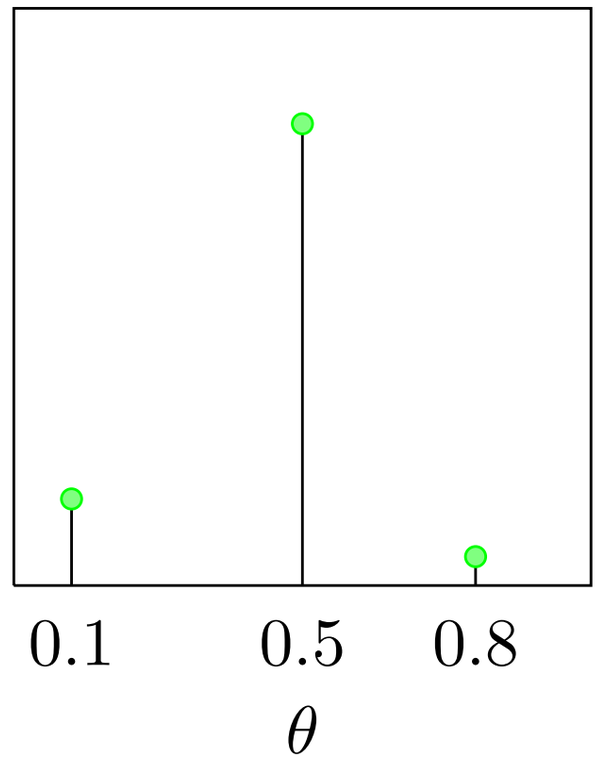
- You get a dataset D = {x_1, …, x_n} of coin flips, where x_i = 1 if it was heads, and x_i = 0 if it was tails



Example: Prior before seeing data

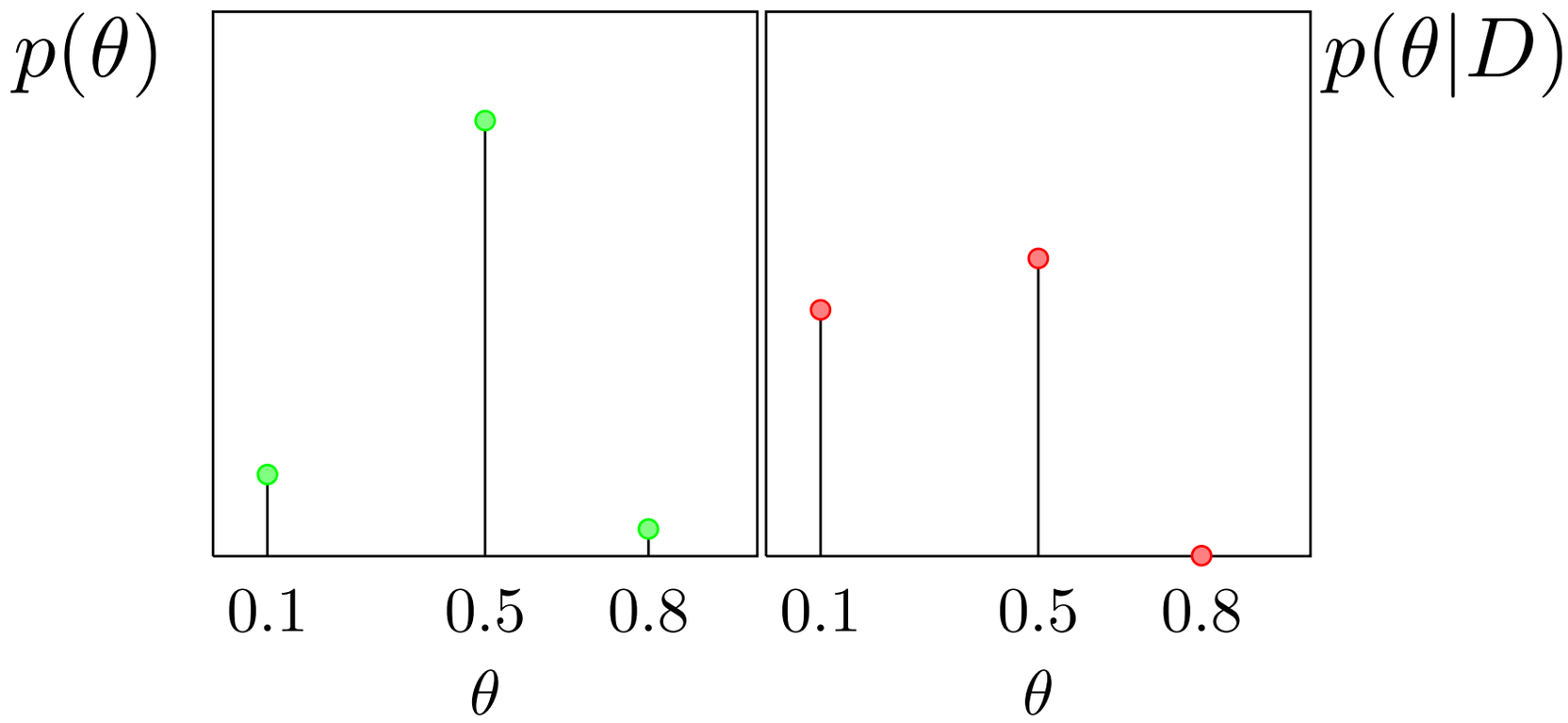
To avoid complexities resulting from continuous variables, we'll consider a discrete θ with only three possible states, $\theta \in \{0.1, 0.5, 0.8\}$. Specifically, we assume

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



Example: Posterior after data

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



But how do we get this posterior?

MAP calculation

- Start by applying Bayes rule

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

- p(D l f) is the **likelihood** of the data, under the model
- p(f) is the **prior** of the model
- p(D) is the marginal distribution of the data
  - we will often be able to ignore this term

Why is this conversion important?

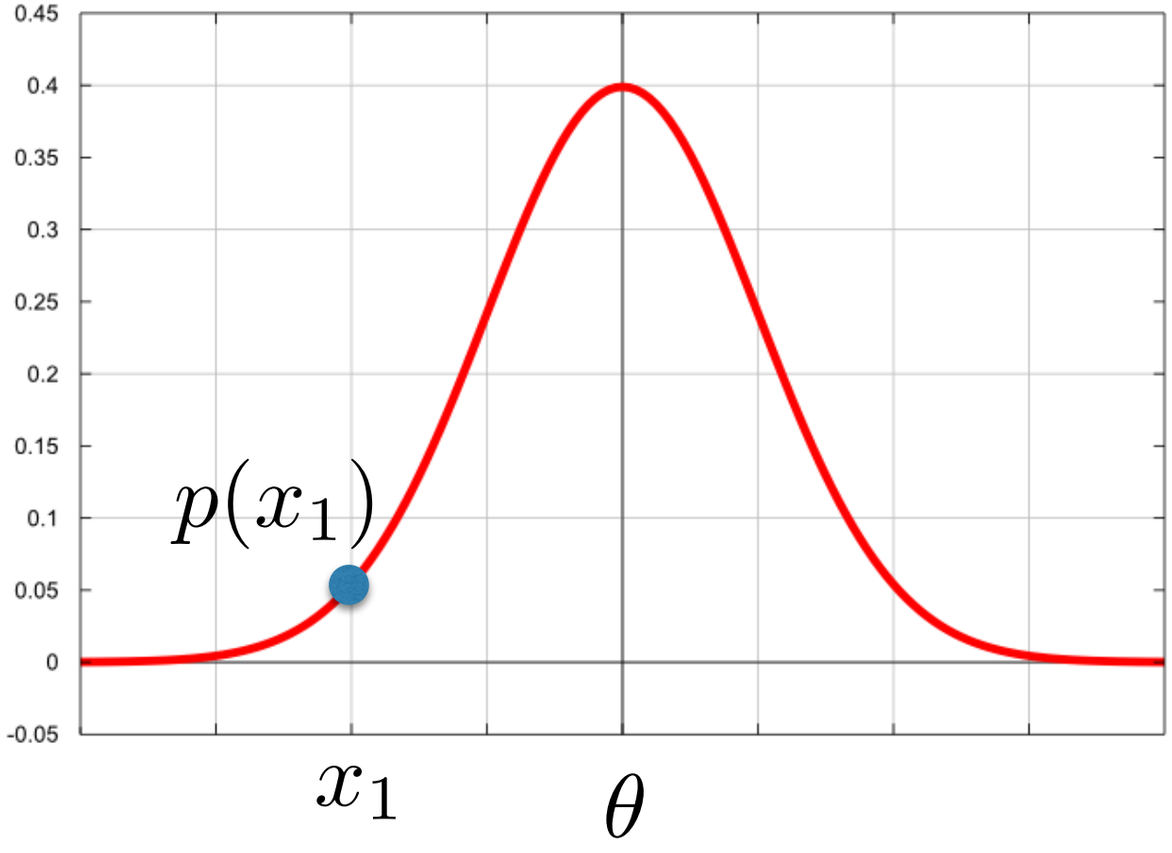
- Do not always have a known form for p(f | D)
- We usually have chosen (known) forms for p(D | f) and p(f)
- **Example**: Let D = {x1} (one sample). Then, if model class is the set of Gaussians: p(D | f) = p(x1 | mu, sigma)
  - p(f | D) is not obvious, since specified our model class for p(x | f)
 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

Example of p(D I theta)

$$= \mathcal{N}(x_1 | \mu = \theta, \sigma^2 = 1)$$

$$= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1 - \theta)^2\right)$$

Probabilities for fixed theta



*Note: we can use variable name f or theta

September 12, 2019

- Thought questions and Assignment coming up
- Clarified biased coin example on eClass
- Any questions?

Recap

$$f_{\mathrm{MAP}} = \arg\max_{f \in \mathcal{F}} p(f|\mathcal{D})$$

- We talked about finding the MAP solution
- This means we are thinking of the parameters as a random variable
- First step: understand p(D l theta)
- Second step: understand the role of the prior

What is the prior?

- The prior gives a chance to incorporate expert knowledge
  - The knowledge or information **prior** to seeing any data

- Example: maybe ahead of time you know what you think are reasonable values for the heights of people

- The prior could be a Gaussian distribution around theta = height, where p(theta) is a Gaussian with mean = 170 (cm) and variance = 60
  - What does the variance mean here, for the prior?

p(D I theta) with n samples

- Imagine we get iid samples x1, …, xn
- We could choose a Gaussian distribution for P(xi I theta), with theta = {mu,sigma}
- Recall iid = independent and identically distributed

What does it mean to say X1 and X2 are independent?

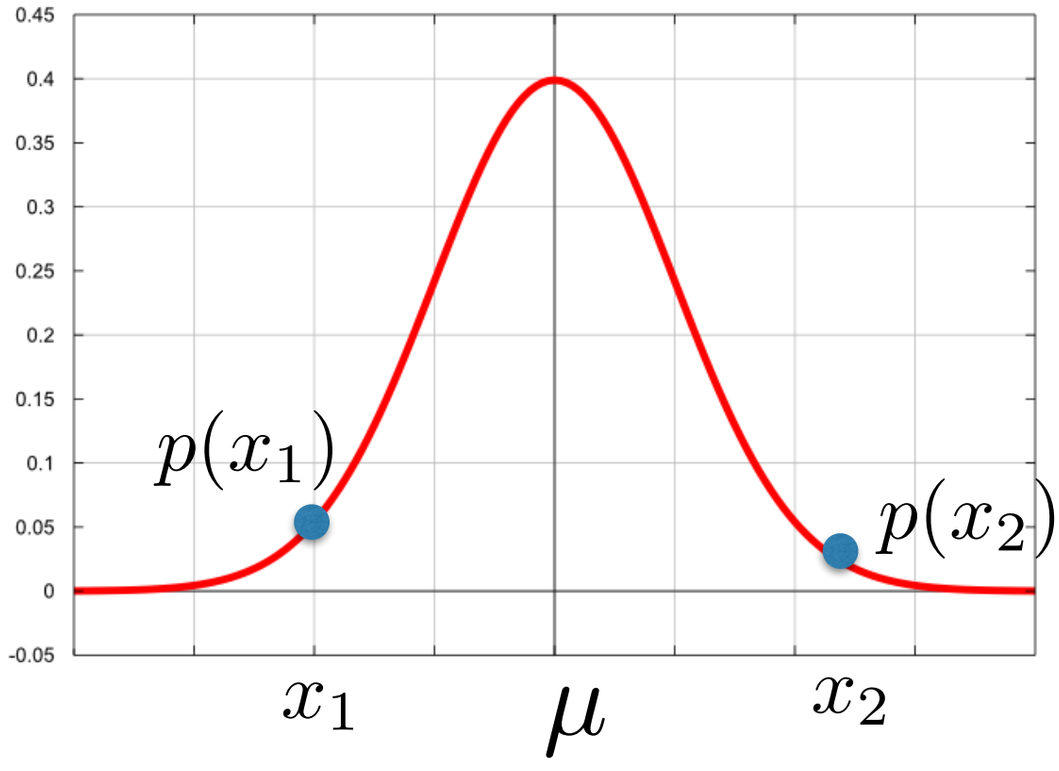
- Independent samples X1, …, Xn
- e.g., P(X1,X2 | theta) = P(X1 | theta) P(X2 | theta)
- X1 and X2 are random variables, from sampling the distribution twice
- Because of randomness, X1 could have been many things
  - it represents the random variable that is the first sample

With n samples

- For iid samples x1, …, xn, we could choose a Gaussian distribution for P(xi | theta), with theta = {mu,sigma}

- P(x1, …, xn | theta) = P(x1 | theta) … P(xn | theta)

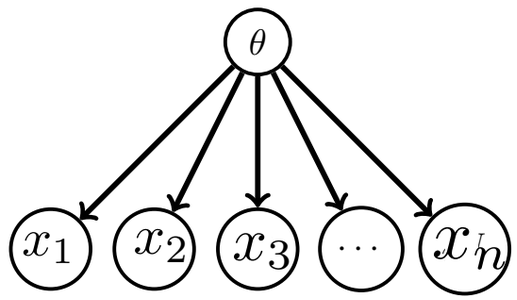
- Example: D = {x1, x2}



$$p(\mathcal{D}|\theta = (\mu, \sigma^2)) = p(\{x_1, x_2\}|\theta = (\mu, \sigma^2))$$
$$= p\Big(x_1|\theta = (\mu, \sigma^2)\Big) \;\; p\Big(x_2|\theta = (\mu, \sigma^2)\Big)$$

Example: MAP for discrete distributions

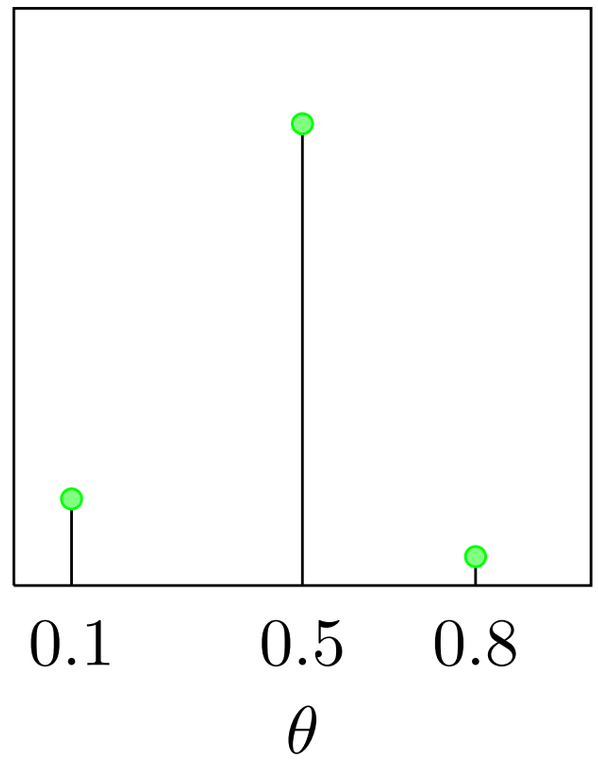
- Imagine you are flipping a biased coin; the model parameter is the bias of the coin, theta
- You get a dataset D = {x_1, …, x_n} of coin flips, where x_i = 1 if it was heads, and x_i = 0 if it was tails
- How do we specify p(theta)?
- What is the MAP estimate?



Example: MAP for discrete distributions

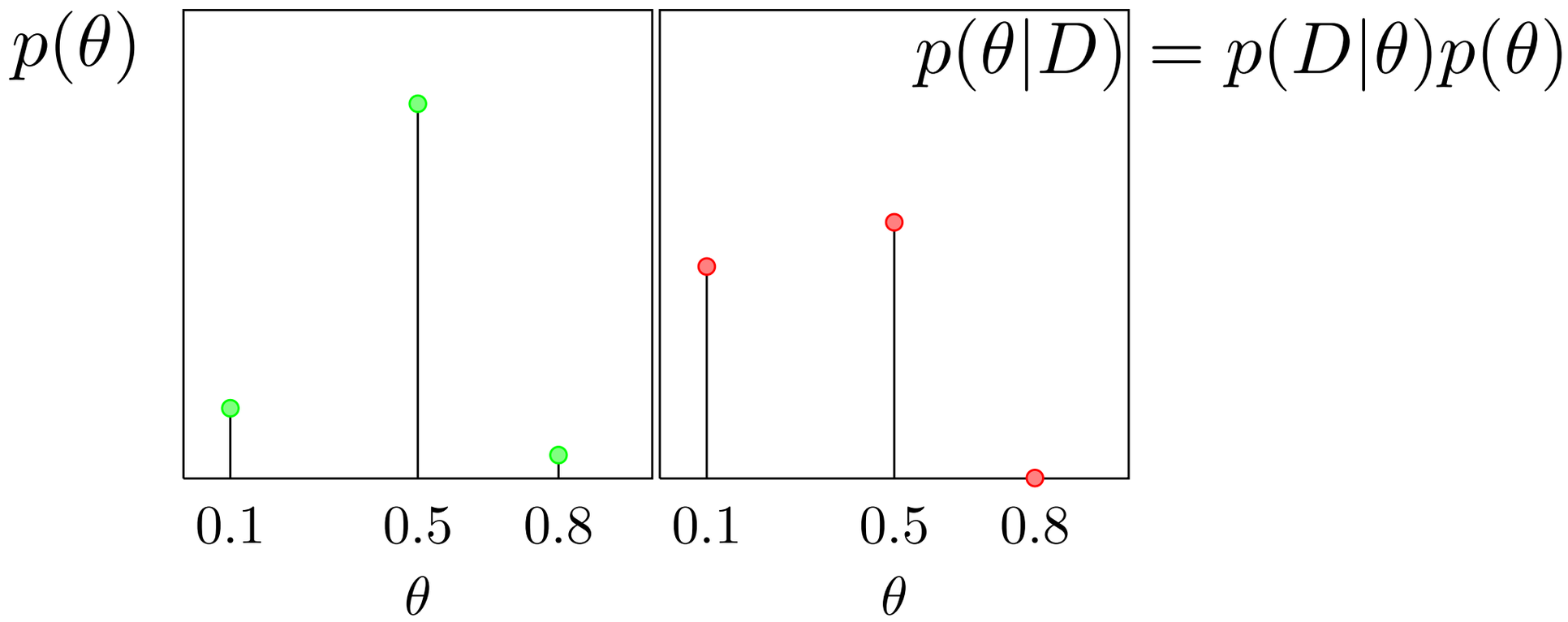
We still need to fully specify the prior $p(\theta)$. To avoid complexities resulting from continuous variables, we'll consider a discrete θ with only three possible states, $\theta \in \{0.1, 0.5, 0.8\}$. Specifically, we assume

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



Example: MAP for discrete distributions

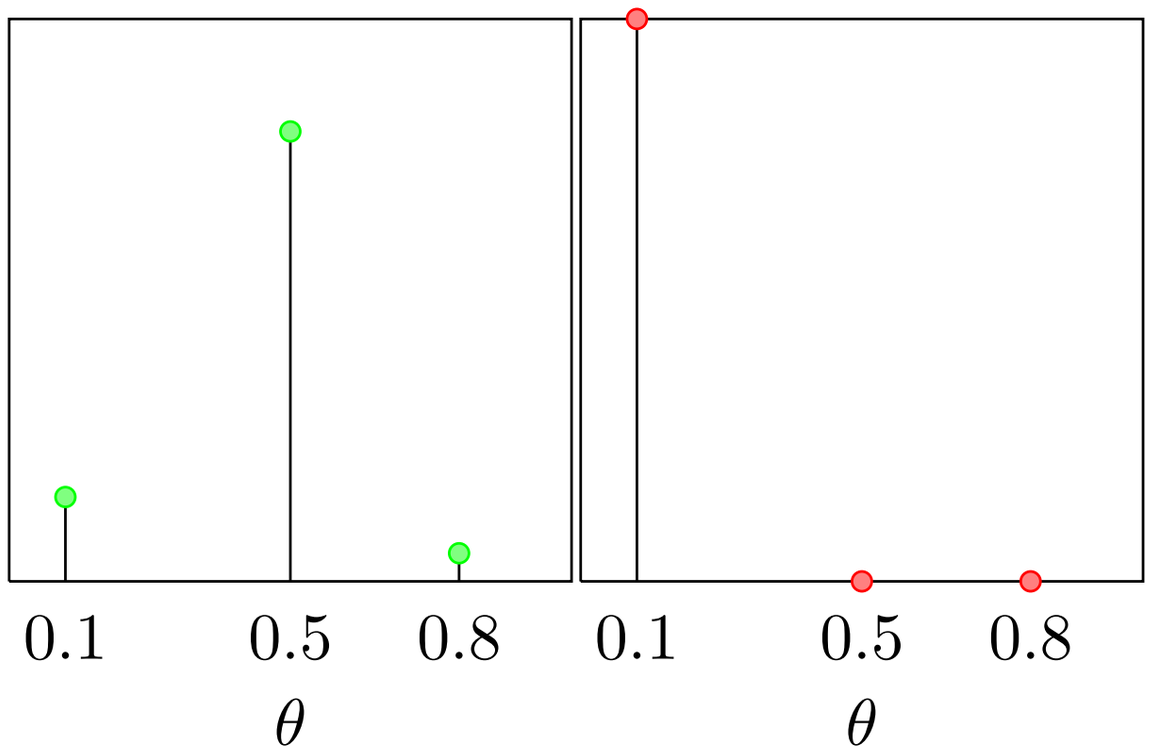
For an experiment with $N_H = 2$, $N_T = 8$, the posterior distribution is



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: MAP 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.

Exercise: Solve for the MAP estimate

Reformulating MAP

The constant is irrelevant, as it is the same for all D

$$\begin{aligned}\theta_{\mathrm{MAP}} &= \arg\max_{\theta} \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ &= \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta)\end{aligned}$$

Will often write:

$$\begin{aligned}p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|\theta)p(\theta)\end{aligned}$$

Maximum likelihood

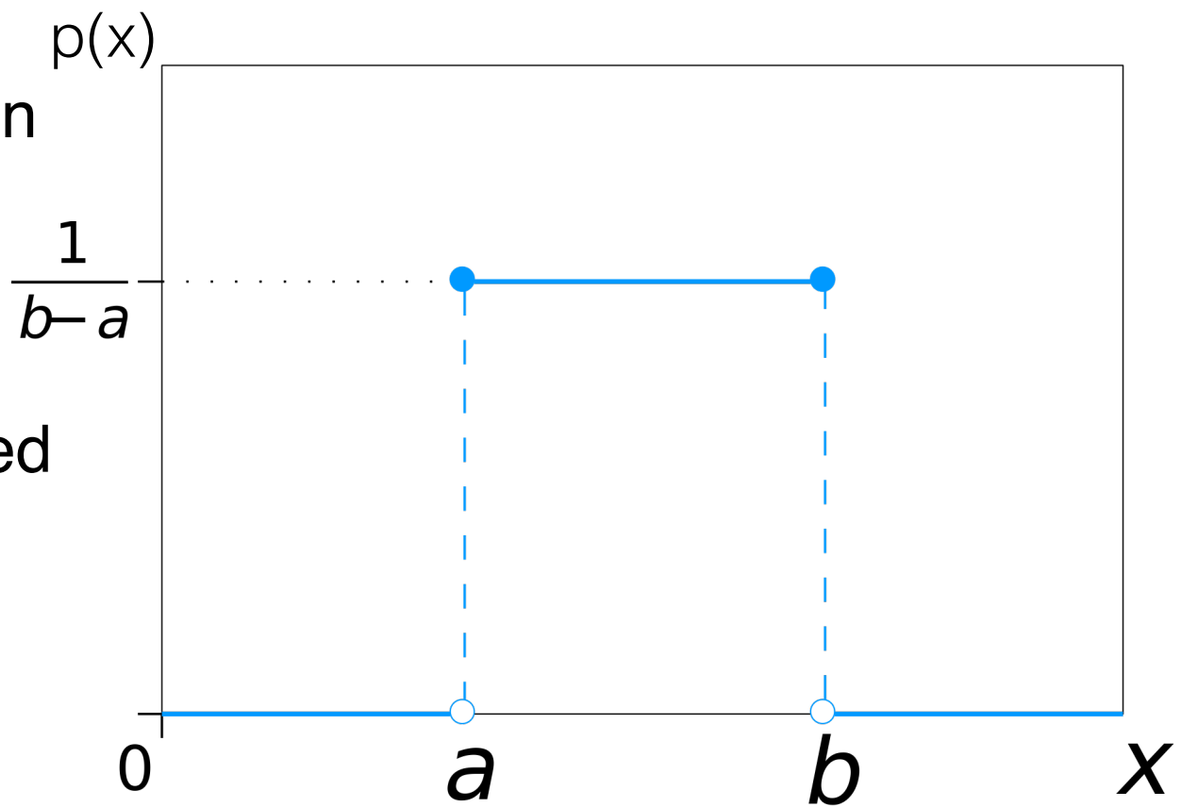
$$\theta_{\mathrm{ML}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D}|\theta)p(\theta)$$

- 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 p(theta) = u for some constant u

- 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

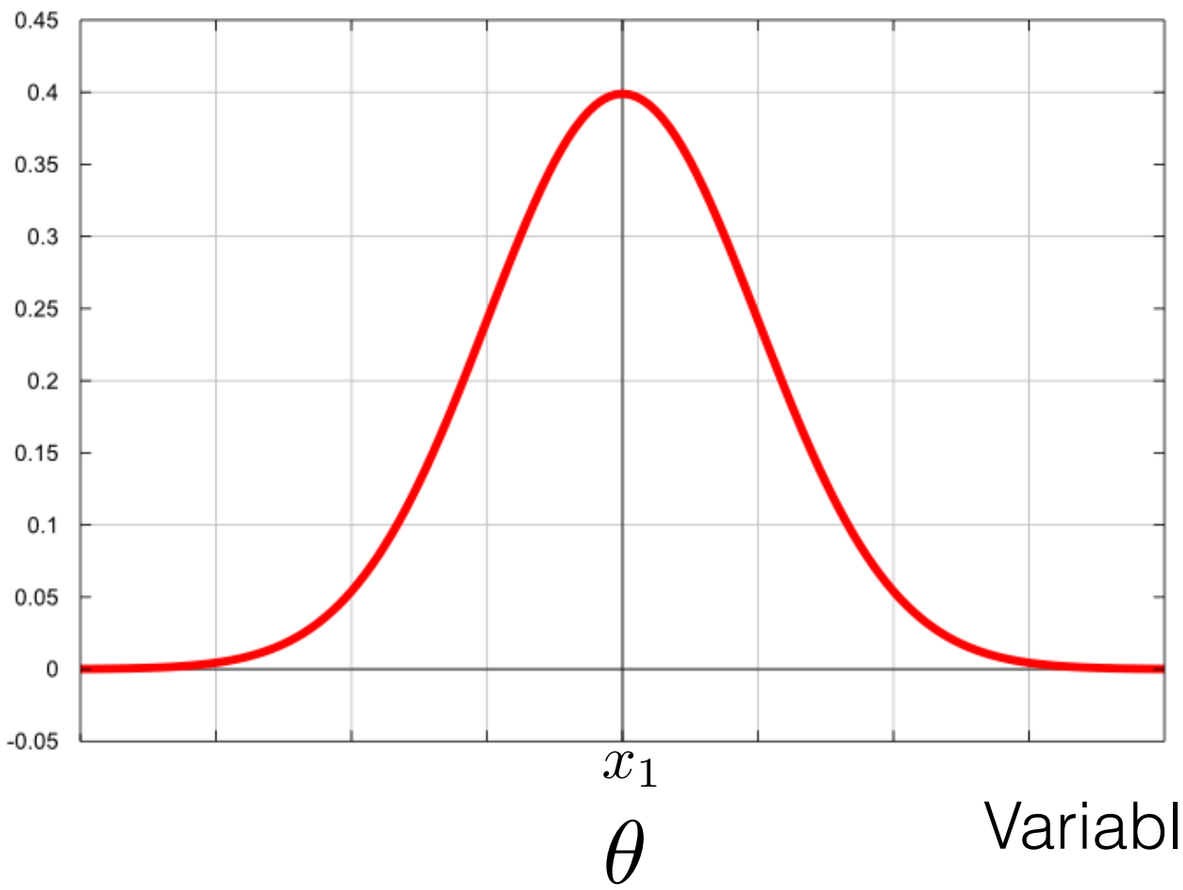


ML example

e.g., $\mathcal{F} = \mathbb{R}, \theta$ is the mean of a Gaussian, fixed $\sigma = 1$

$$
\begin{aligned}
c(\theta) &= p(\mathcal{D}|\theta) \\
&= \mathcal{N}(x_1|\mu = \theta, \sigma^2 = 1) \\
&= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1 - \theta)^2\right)
\end{aligned}
$$

$c(\theta) = p(\mathcal{D}|\theta)$



Variable is theta, optimizing it

Maximizing the log-likelihood

- We want to maximize the **likelihood**, but often instead maximize the **log-likelihood**

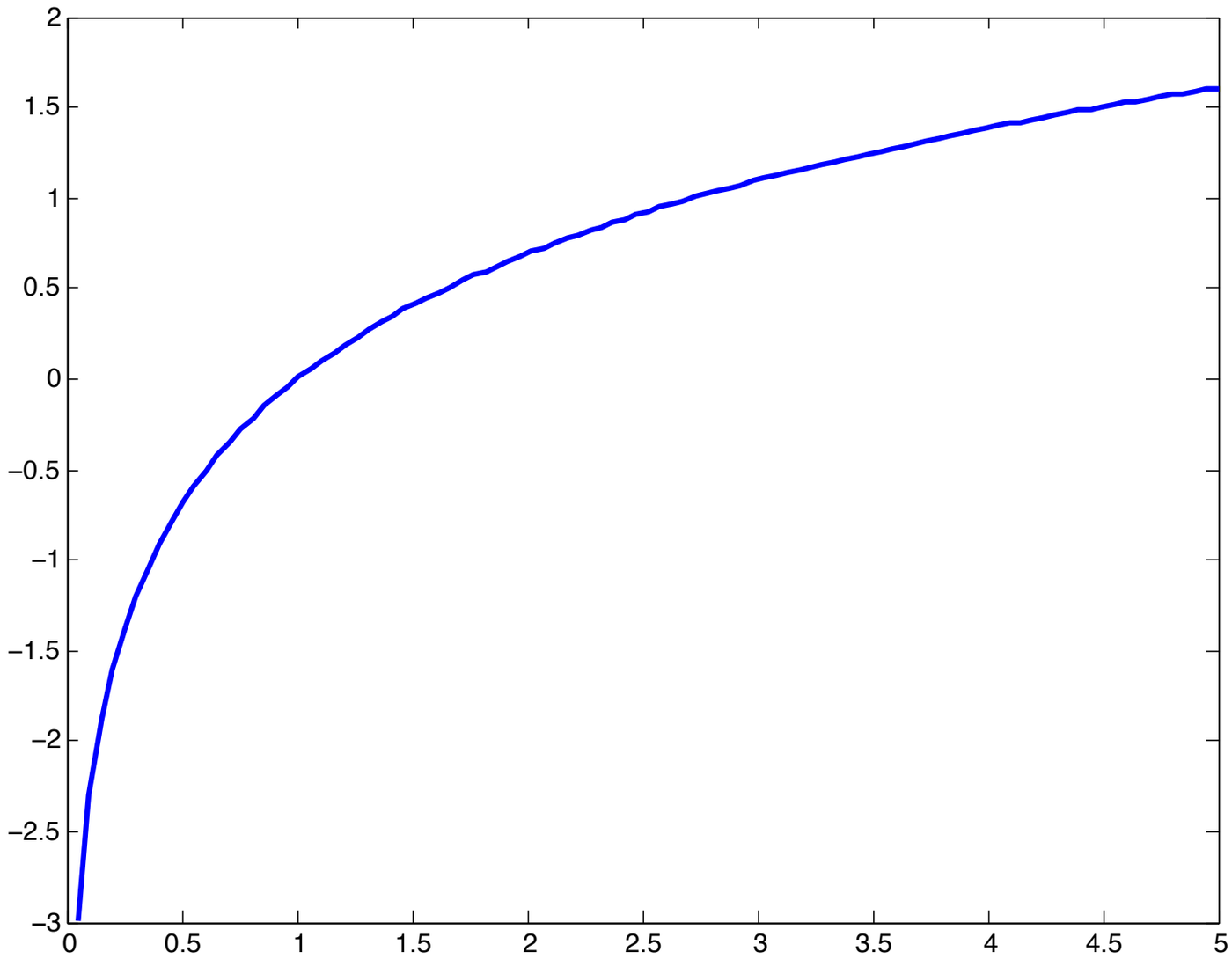
$$\arg\max_{\theta \in \mathcal{F}} p(\mathcal{D}|\theta) = \arg\max_{\theta \in \mathcal{F}} \log p(\mathcal{D}|\theta)$$

- Why? Or maybe first, is this equivalent?
  - The Why is that it makes the optimization much simpler, when we have more than one sample

Why can we shift by log?

$$c(\theta_1) > c(\theta_2) \iff \log c(\theta_1) > \log c(\theta_2) \quad \text{for } c(\theta) > 0$$

Monotone increasing



Likelihood values always > 0

Maximizing the log-likelihood

e.g., $\mathcal{F} = \mathbb{R}, \theta$ is the mean of a Gaussian, fixed $\sigma = 1$

$\log(ab) = \log a + \log b$

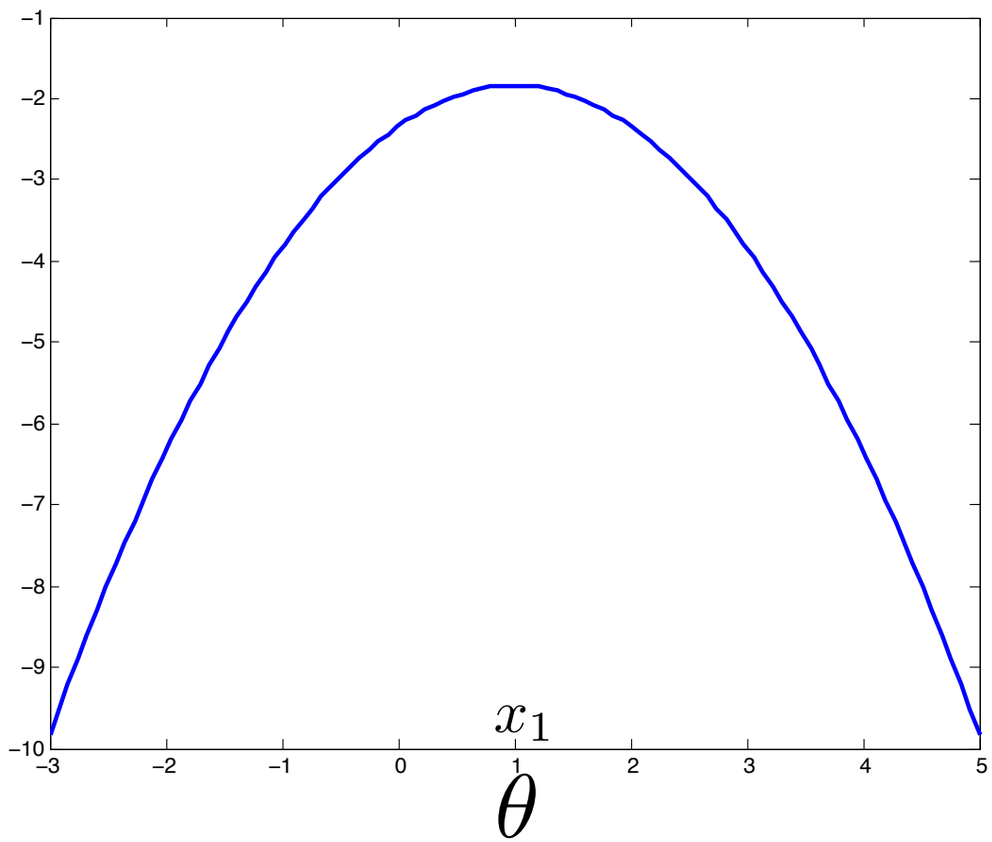
$\log(a^c) = c \log a$

$$c(\theta) = \log p(\mathcal{D}|\theta)$$

$$= \log\left(\frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_1 - \theta)^2\right)\right)$$

$$= -\log(2\pi) - \frac{1}{2}(x_1 - \theta)^2$$

$$c(\theta) = \log p(\mathcal{D}|\theta)$$



This conversion is even more important when we have more than one sample

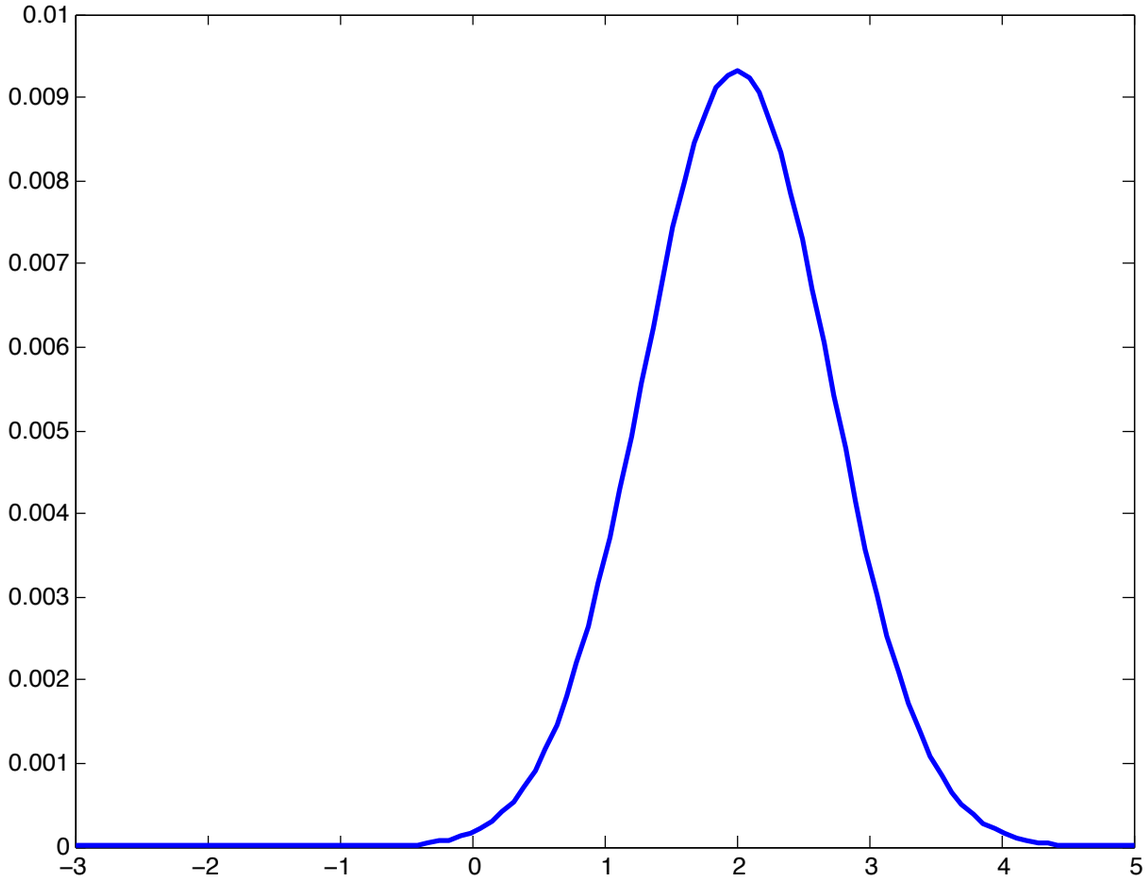
- **Example**: Let D = {x1,x2} (two samples).

- If x1 and x2 are independent samples from same distribution (same model), then P(x1, x2 I theta) = P(x1 I f) P(x2 I theta)

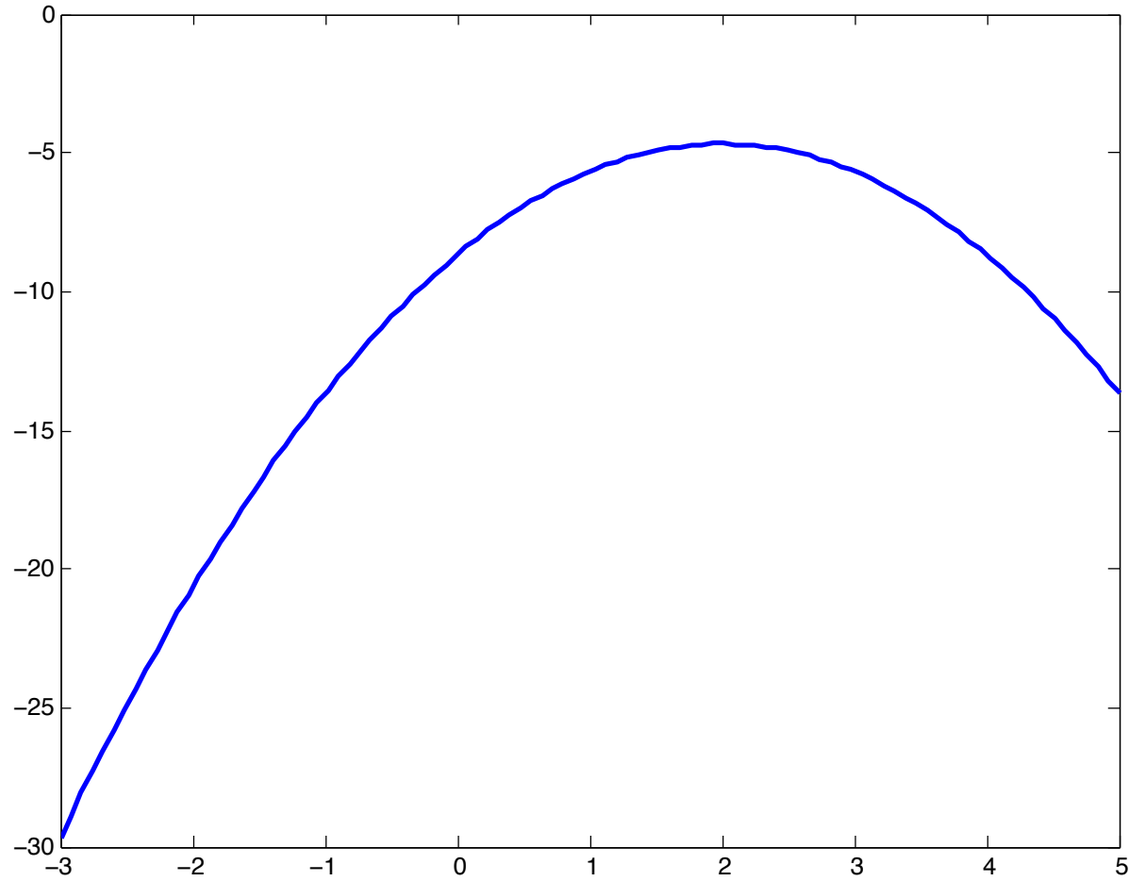
$$p(x_1|\theta)p(x_2|\theta) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_1-\theta)^2\right) \times \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_2-\theta)^2\right)$$

$$\begin{aligned}\log\left(p(x_1|\theta)p(x_2|\theta)\right) &= \log p(x_1|\theta) + \log p(x_2|\theta) \\ &= -2\log(2\pi) - \frac{1}{2}(x_1-\theta)^2 - \frac{1}{2}(x_2-\theta)^2\end{aligned}$$

Visualizing the objective



$p(x_1|\theta)p(x_2|\theta)$



$\log\left(p(x_1|\theta)p(x_2|\theta)\right)$

Exercise: about the marginal over the data p(D)

- Why do we avoid computing p(D)?

- How do we compute p(D)? We have the tools, since we know about marginals

How do we compute p(D)?

- If we have p(D, f), can we obtain p(D)?
  - Marginalization

$$p_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_k} p_{\boldsymbol{X}}(x_1, \ldots, x_k)$$

$$p_{X_i}(x_i) = \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_k} p_{\boldsymbol{X}}(\boldsymbol{x})\, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_k$$

- If we have p(Dlf) and p(f), do we have p(D, f)?

Data marginal

- Using the formula of total probability

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

- Fully expressible in terms of likelihood and prior

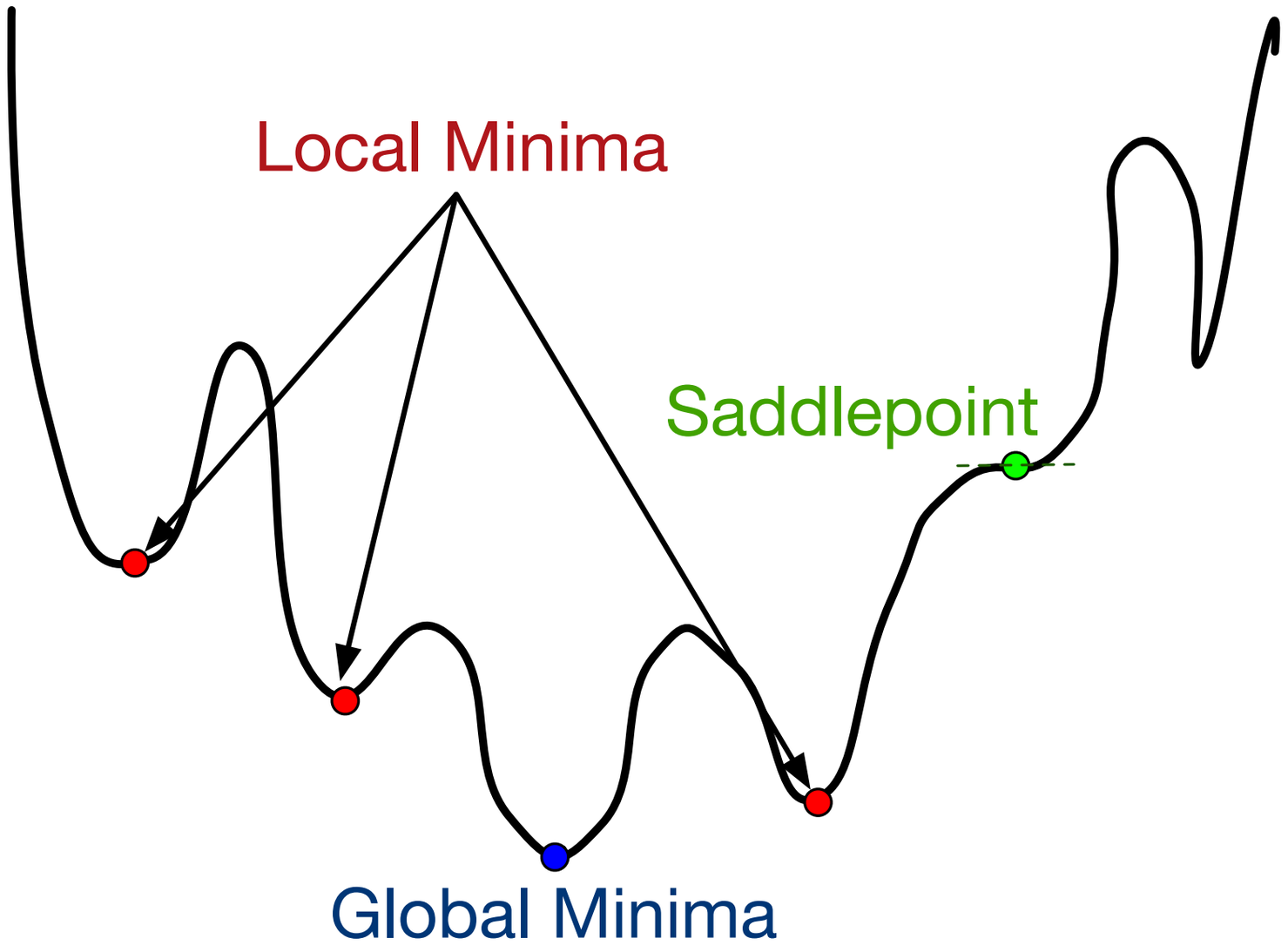
How do we solve the MAP and ML maximization problems?

- Naive strategy:
- 1. Guess 100 solutions theta
- 2. Pick the one with the largest value
- Can we do something better?

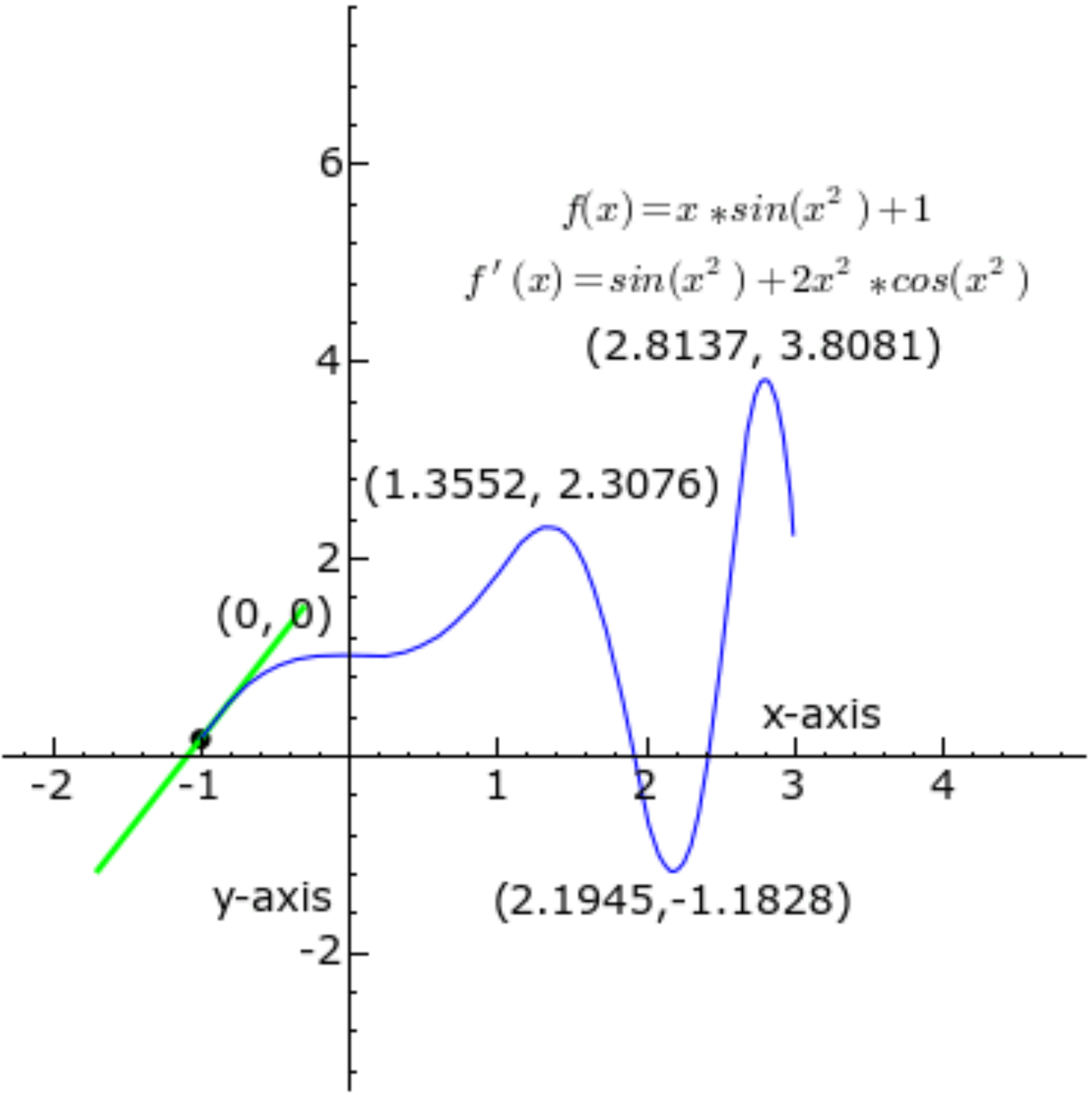
Crash course in optimization

- Goal: find maximal (or minimal) points of functions
- Generally assume functions are smooth, use gradient descent
- Derivative: direction of ascent from a scalar point $\frac{d}{dx}c(x)$
- Gradient: direction of ascent from a vector point $\nabla c(\mathbf{x})$

Function surface



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, $\frac{df}{dx}(x)$ gives the slope of the tangent to the function at f(x)

*Note: f is c in this slide

Why don't constants matter?

$$\max_x c(x)$$

$$\frac{d}{dx} c(x) = 0$$

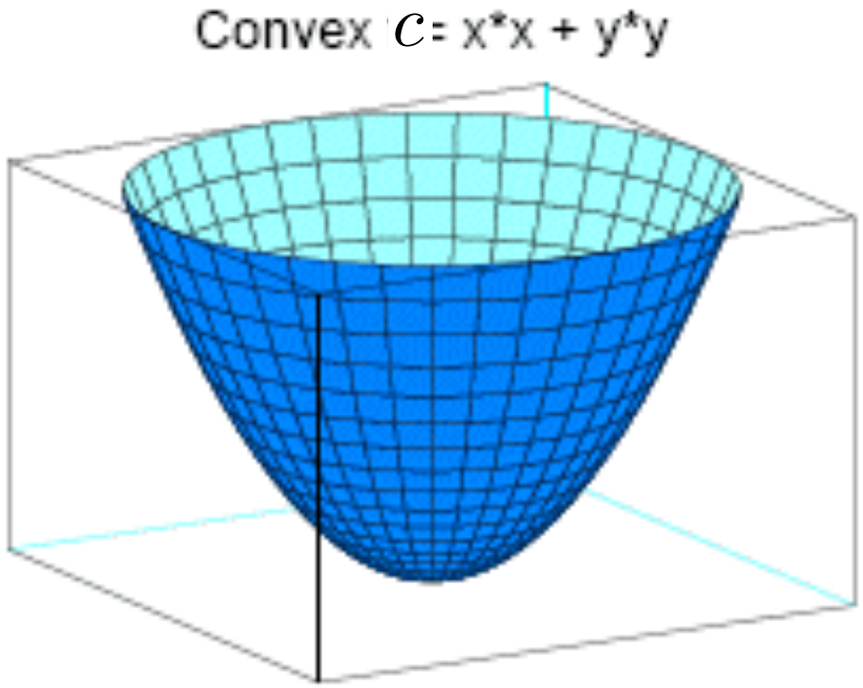
$$\max u c(x), \quad u > 0$$

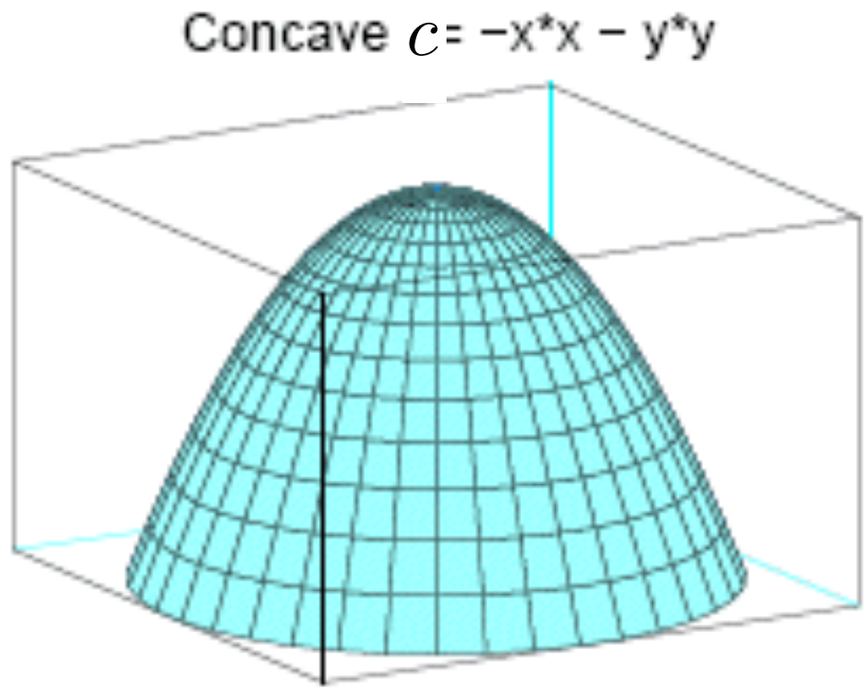
$$\frac{d}{dx} u c(x) = u \frac{d}{dx} c(x) = 0$$

Both have derivative zero under same condition regardless of $u > 0$

Can either minimize or maximize

$$\arg\min_{\theta} c(\theta) = \arg\max_{\theta} -c(\theta)$$





$$c(t\mathbf{w}_1 + (1-t)\mathbf{w}_2) \leq tc(\mathbf{w}_1) + (1-t)c(\mathbf{w}_2)$$

Example: maximum likelihood for discrete distributions

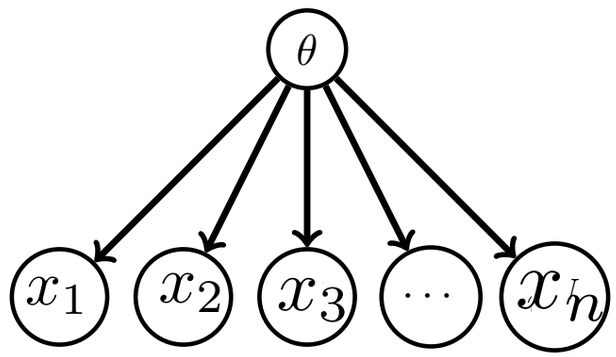
- Imagine you are flipping a biased coin; the model parameter is the bias of the coin, theta

- 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 theta)?

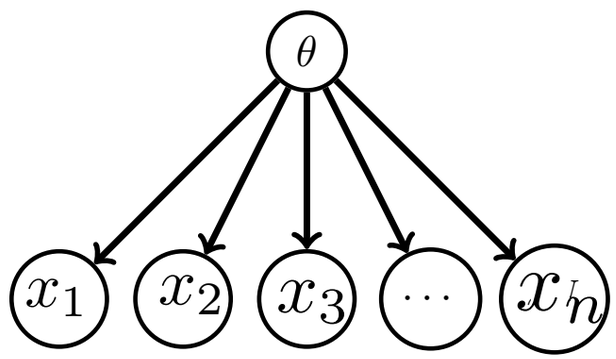
$$p(\mathcal{D}|\theta) = p(x_1, \ldots, x_n|\theta)$$
$$= \prod_{i=1}^{n} p(x_i|\theta)$$

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$



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



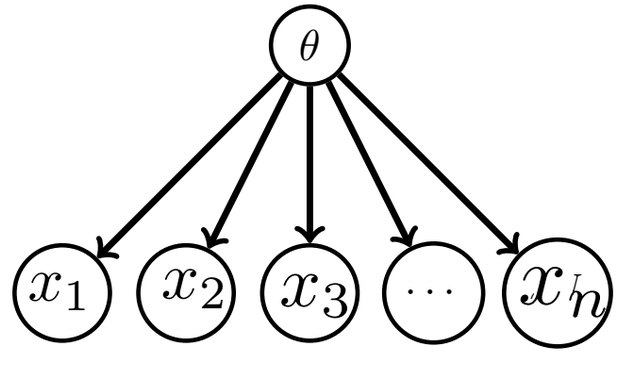
$$p(\mathcal{D}|\theta) = p(x_1, \ldots, x_n|\theta)$$

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

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

Example: maximum likelihood for discrete distributions

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



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

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

$$\arg\max_\theta c\,(\theta) = \arg\max_\theta \log\, c(\theta)$$

$$p(\mathcal{D}|\theta) = p(x_1, \ldots, x_n|\theta)$$

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

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

Example: maximum likelihood for discrete distributions

$$\arg\max_{\theta} c(\theta) = \arg\max_{\theta} \log c(\theta)$$

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

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

$$\log(ab) = \log a + \log b$$

$$\log(a^c) = c \log a$$

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

$$\log p(x|\theta) = \log\left(\theta^x\right) + \log\left((1-\theta)^{1-x}\right)$$

$$= x \log(\theta) + (1-x)\log(1-\theta)$$

Example: maximum likelihood for discrete distributions

$$\sum_{i=1}^{n} \log p(x_i|\theta) = \sum_{i=1}^{n} x_i \log(\theta) + \sum_{i=1}^{n} (1 - x_i) \log(1 - \theta)$$

$$= \log(\theta) \left(\sum_{i=1}^{n} x_i \right) + \log(1 - \theta) \left(\sum_{i=1}^{n} (1 - x_i) \right)$$

$$\bar{x} = \sum_{i=1}^{n} x_i$$

$$\frac{d}{d\theta} = \frac{1}{\theta} \bar{x} - \frac{1}{1 - \theta} (n - \bar{x}) = 0$$

Example: maximum likelihood for discrete distributions

$$\bar{x} = \sum_{i=1}^{n} x_i$$

$$\frac{d}{d\theta} = \frac{1}{\theta}\bar{x} - \frac{1}{1-\theta}(n - \bar{x}) = 0$$

$$\begin{aligned}
&\Longrightarrow \quad \frac{\bar{x}}{\theta} = \frac{n - \bar{x}}{1-\theta} \\
&\Longrightarrow \quad (1-\theta)\bar{x} = \theta(n - \bar{x}) \\
&\Longrightarrow \quad \bar{x} - \theta\bar{x} = \theta n - \theta\bar{x} \\
&\Longrightarrow \quad \theta = \frac{\bar{x}}{n}
\end{aligned}$$

We solved for a stationary point

- But how do we know that this is the optimal solution to the optimization problem we specified?

Univariate optimization

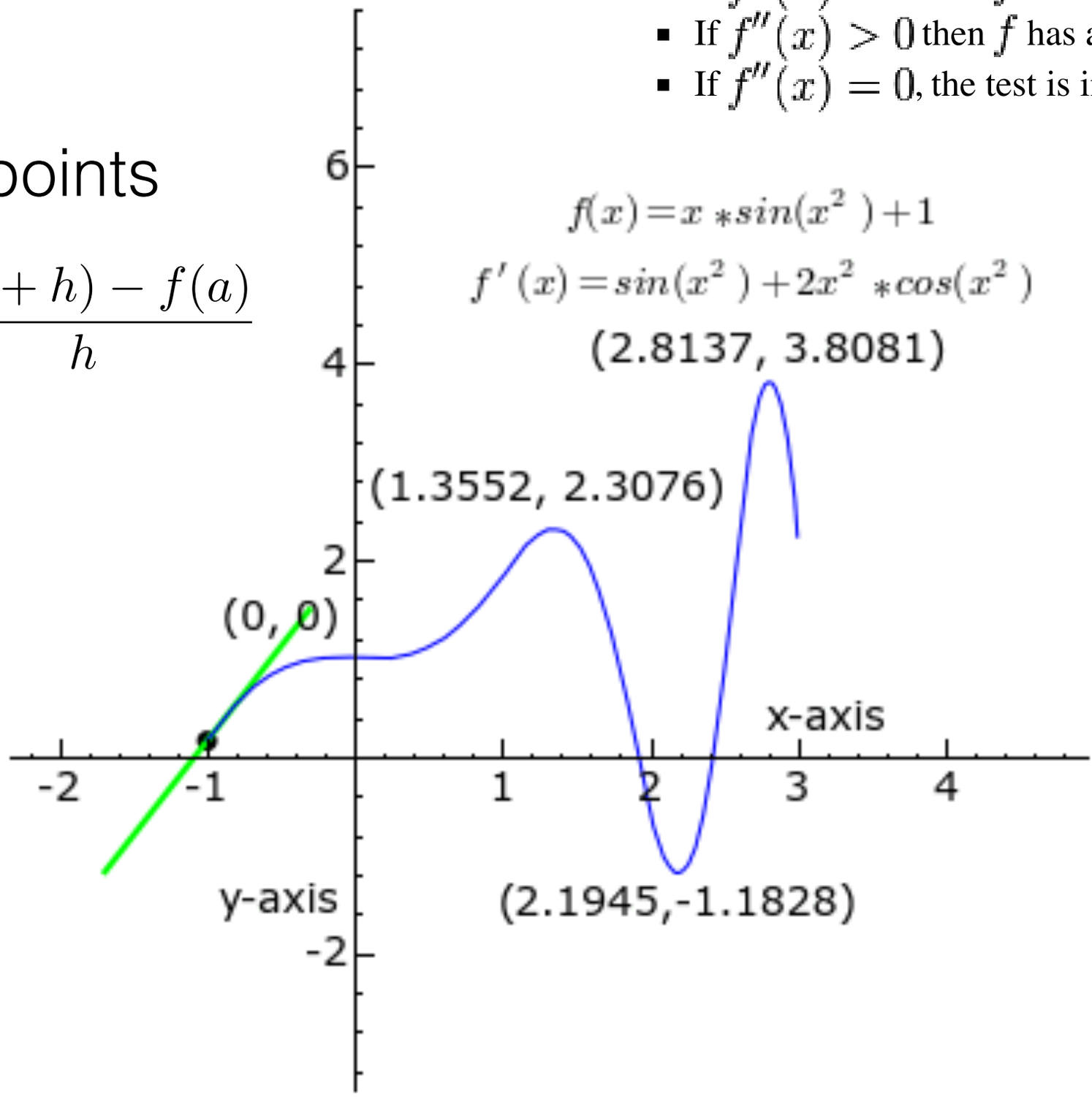
- Minima
- Maxima
- Saddle points

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

- If $f''(x) < 0$ then f has a local maximum at x.
- If $f''(x) > 0$ then f has a local minimum at x.
- If $f''(x) = 0$, the test is inconclusive.

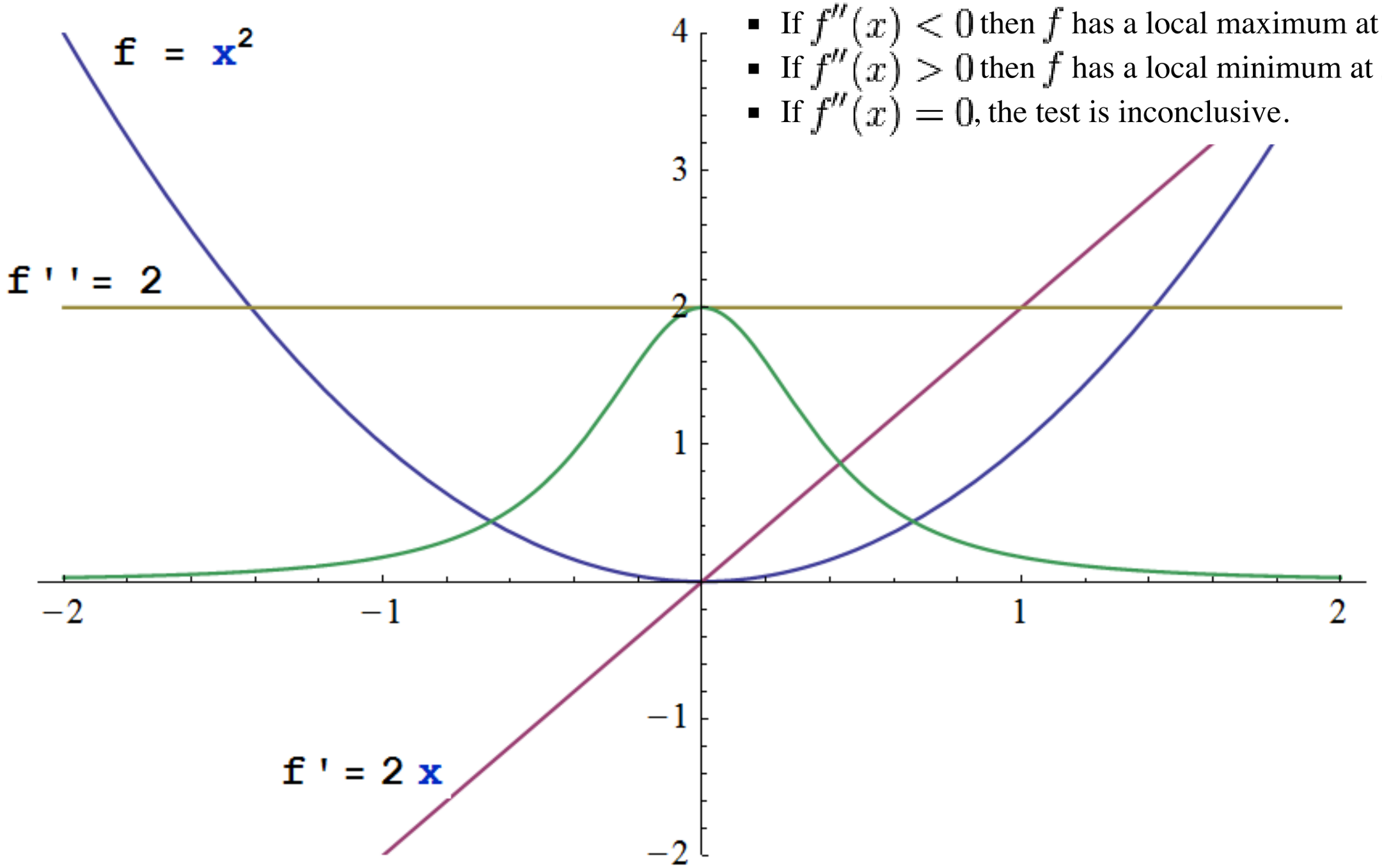
$$f(x) = x * sin(x^2) + 1$$

$$f'(x) = sin(x^2) + 2x^2 * cos(x^2)$$



Second derivative test

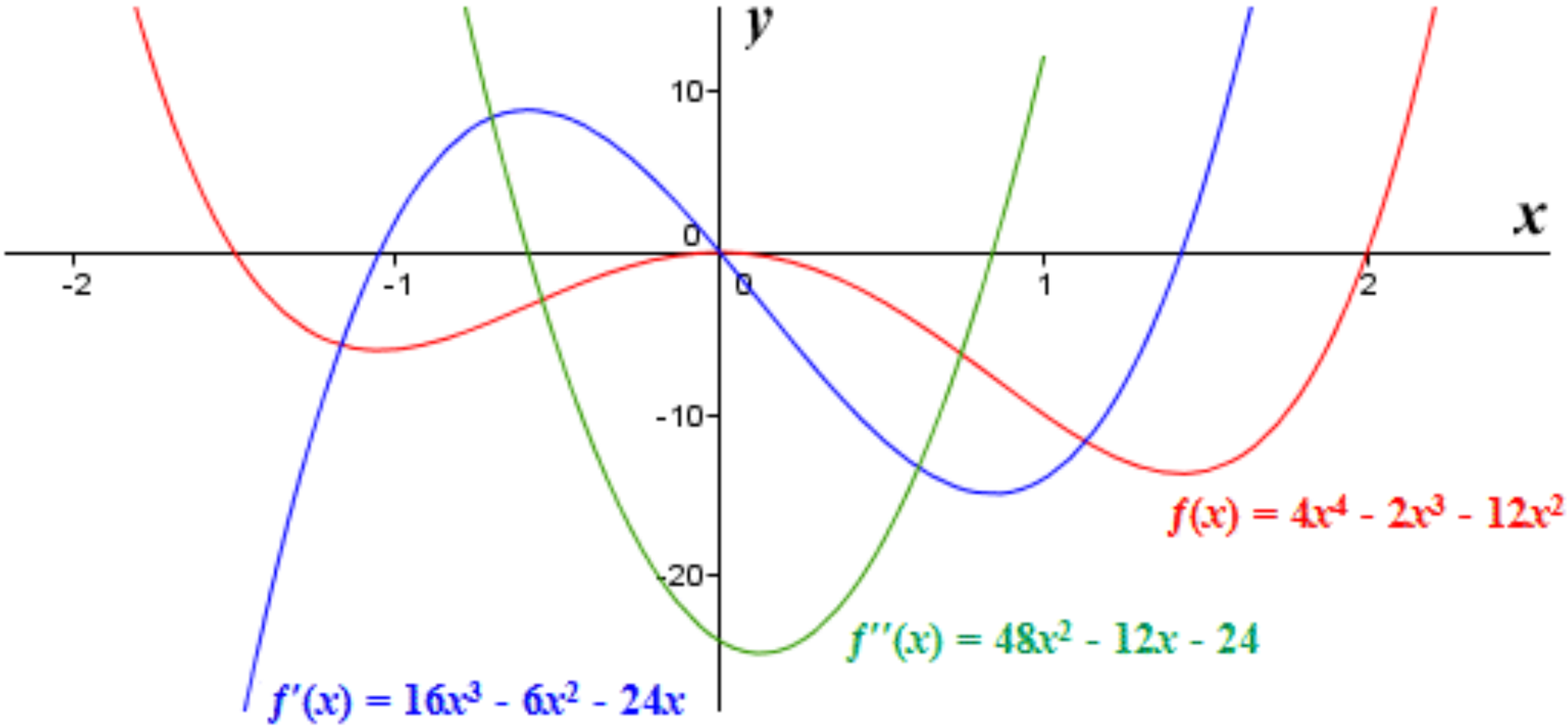
- If $f''(x) < 0$ then f has a local maximum at x.
- If $f''(x) > 0$ then f has a local minimum at x.
- If $f''(x) = 0$, the test is inconclusive.



Ignore the green line

Second derivative test

- If $f''(x) < 0$ then f has a local maximum at x.
- If $f''(x) > 0$ then f has a local minimum at x.
- If $f''(x) = 0$, the test is inconclusive.



Now on to some careful examples of MAP!

- Whiteboard for Examples 8, 9, 10, 11
- More fun with derivatives and finding the minimum of a function
- Next:
  - introduction to prediction problems for ML