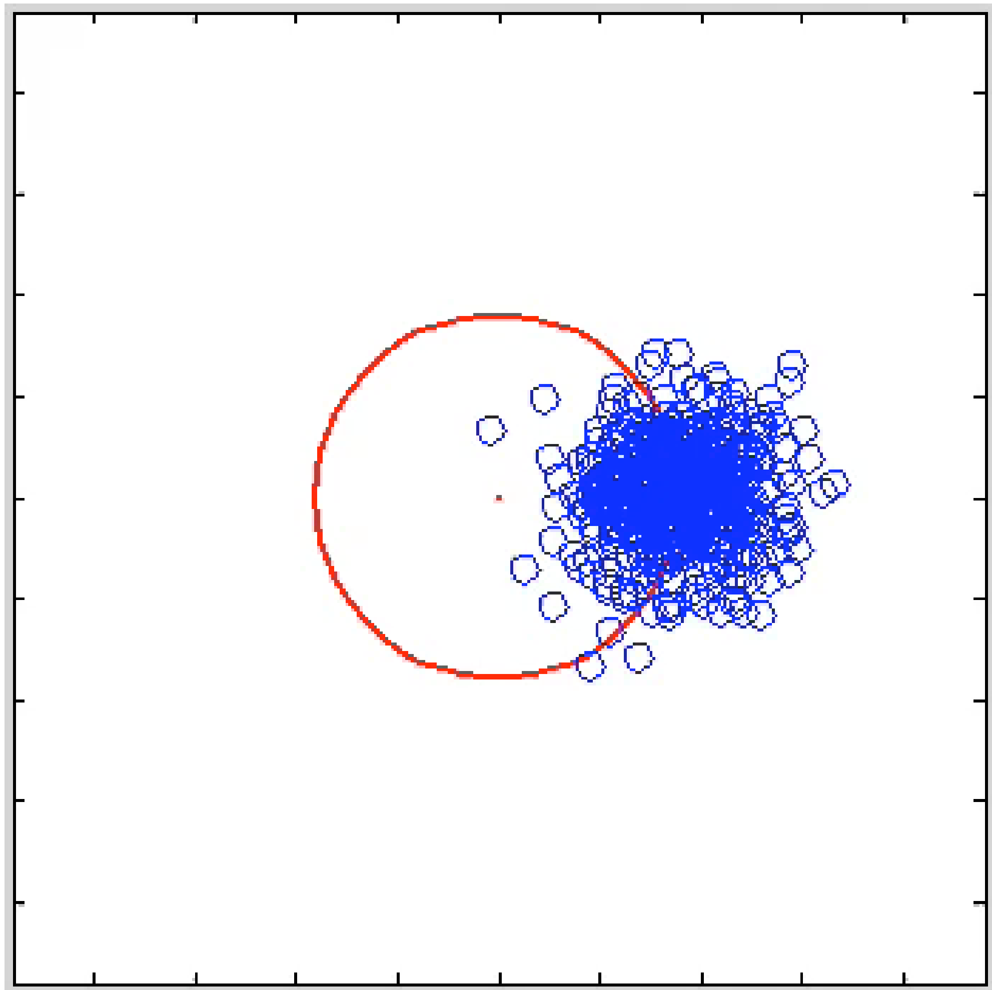
15-780: Graduate AI

Lecture 19. Learning

Geoff Gordon (this lecture)
Tuomas Sandholm
TAs Sam Ganzfried, Byron Boots

Review

Stationary distribution



Stationary distribution

$$Q(\mathbf{x}_{t+1}) = \int \mathbb{P}(\mathbf{x}_{t+1} \mid \mathbf{x}_t) Q(\mathbf{x}_t) d\mathbf{x}_t$$

MH algorithm

- *Proof that MH algorithm's stationary distribution is the desired P(**x**)*
- *Based on* ***detailed balance****: transitions between* ***x*** *and* ***x****' happen equally often in each direction*

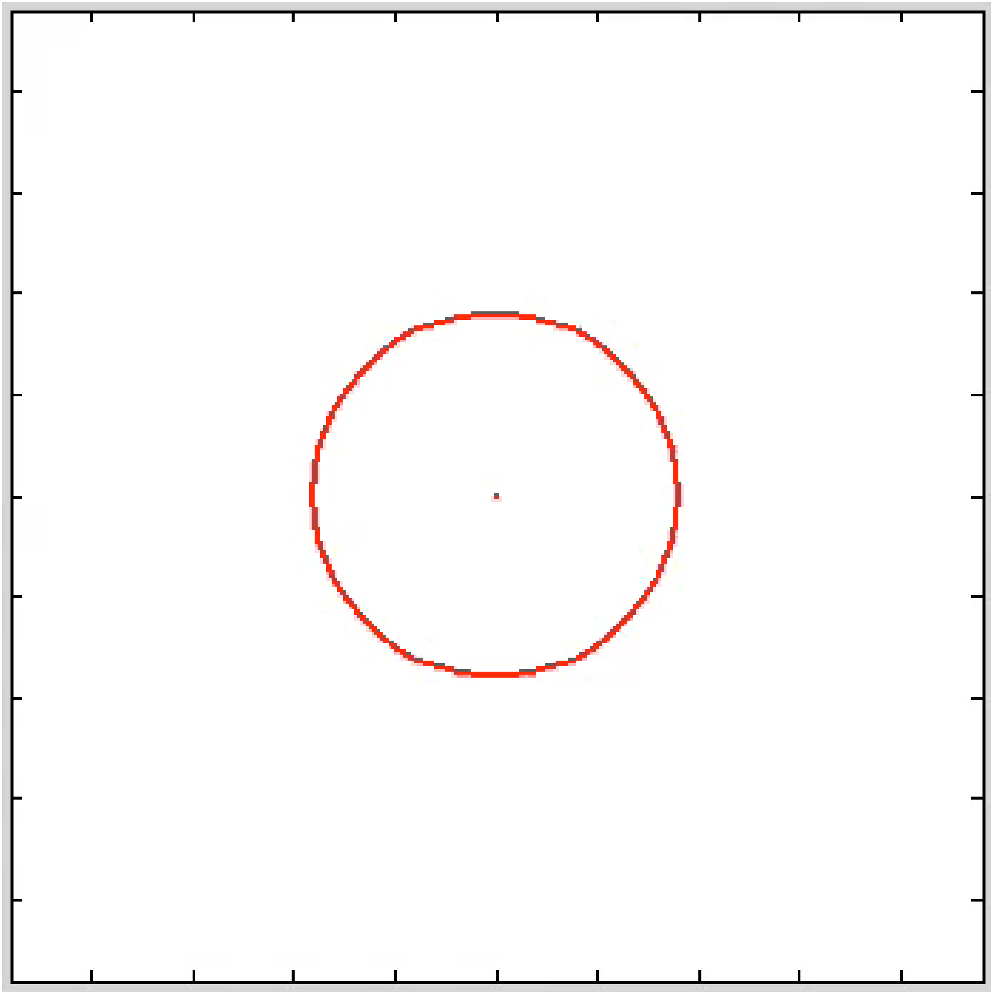
Gibbs

- *Special case of MH*
- *Proposal distribution: conditional probability of block i of* $\boldsymbol{x}$, *given rest of* $\boldsymbol{x}$
- *Acceptance probability is always 1*

Sequential sampling

- *Often we want to keep a sample of belief at current time*
- *This is the sequential sampling problem*
- *Common algorithm: particle filter*
  - *Parallel importance sampling for* $P(\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t)$

Particle filter example



Learning

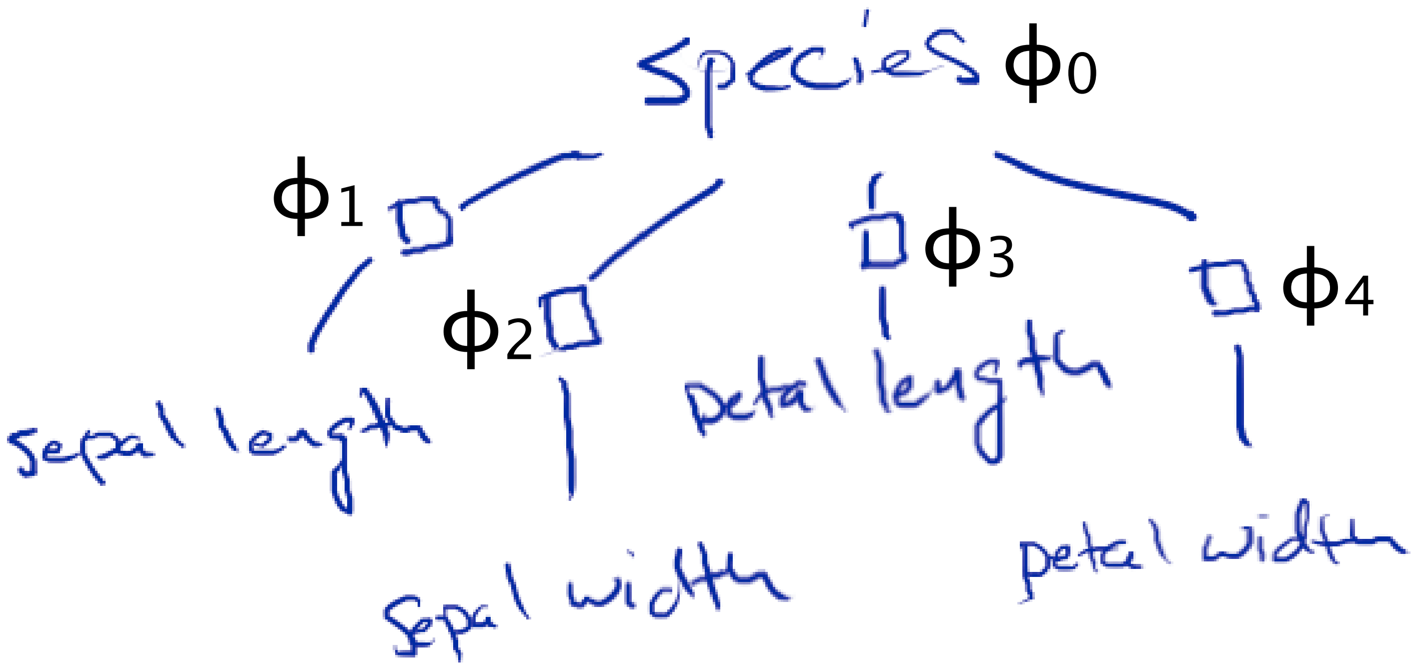
- *Improve our model, using sampled data*
- *Model = factor graph, SAT formula, …*
- *Hypothesis space = { all models we'll consider }*
- *Conditional models*

Version space algorithm

- *Predict w/ majority of still-consistent hypotheses*
- *Mistake bound analysis*

Bayesian Learning

Recall iris example



- $\mathscr{H}$ = *factor graphs of given structure*
- *Need to specify entries of* ϕ*s*

Factors

ϕ_0

setosa	p
versicolor	q
virginica	$1-p-q$

$\phi_1 - \phi_4$

	lo	*m*	*hi*
set.	p_i	q_i	$1-p_i-q_i$
vers.	r_i	s_i	$1-r_i-s_i$
vir.	u_i	v_i	$1-u_i-v_i$

Continuous factors

ϕ_1

	lo	*m*	*hi*
set.	p_1	q_1	$1-p_1-q_1$
vers.	r_1	s_1	$1-r_1-s_1$
vir.	u_1	v_1	$1-u_1-v_1$

Discretized petal length

$$\Phi_1(\ell, s) = \exp(-(\ell - \ell_s)^2/2\sigma^2)$$

parameters $\ell_{\mathrm{set}}, \ell_{\mathrm{vers}}, \ell_{\mathrm{vir}}$;
constant σ^2

Continuous petal length

Simpler example

H	*p*
T	*1–p*

Coin toss

Parametric model class

- $\mathcal{H}$ *is a* ***parametric*** *model class: each H in* $\mathcal{H}$ *corresponds to a vector of parameters* $\theta = (p)$ *or* $\theta = (p, q, p_1, q_1, r_1, s_1, \ldots)$
- H_θ: $\boldsymbol{X} \sim P(\boldsymbol{X} \mid \theta)$ *(or,* $Y \sim P(Y \mid \boldsymbol{X}, \theta)$*)*
- *Contrast to* ***discrete*** $\mathcal{H}$*, as in version space*
- *Could also have* ***mixed*** $\mathcal{H}$*: discrete choice among parametric (sub)classes*

Prior

- *Write* $\mathbf{D} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)$
- H_θ *gives* $P(\mathbf{D} \mid \theta)$
- *Bayesian learning also requires* ***prior***
  - *distribution over* $\mathscr{H}$
  - *for parametric classes,* $P(\theta)$
- *Together,* $P(\mathbf{D} \mid \theta)\, P(\theta) = P(\mathbf{D}, \theta)$

Prior

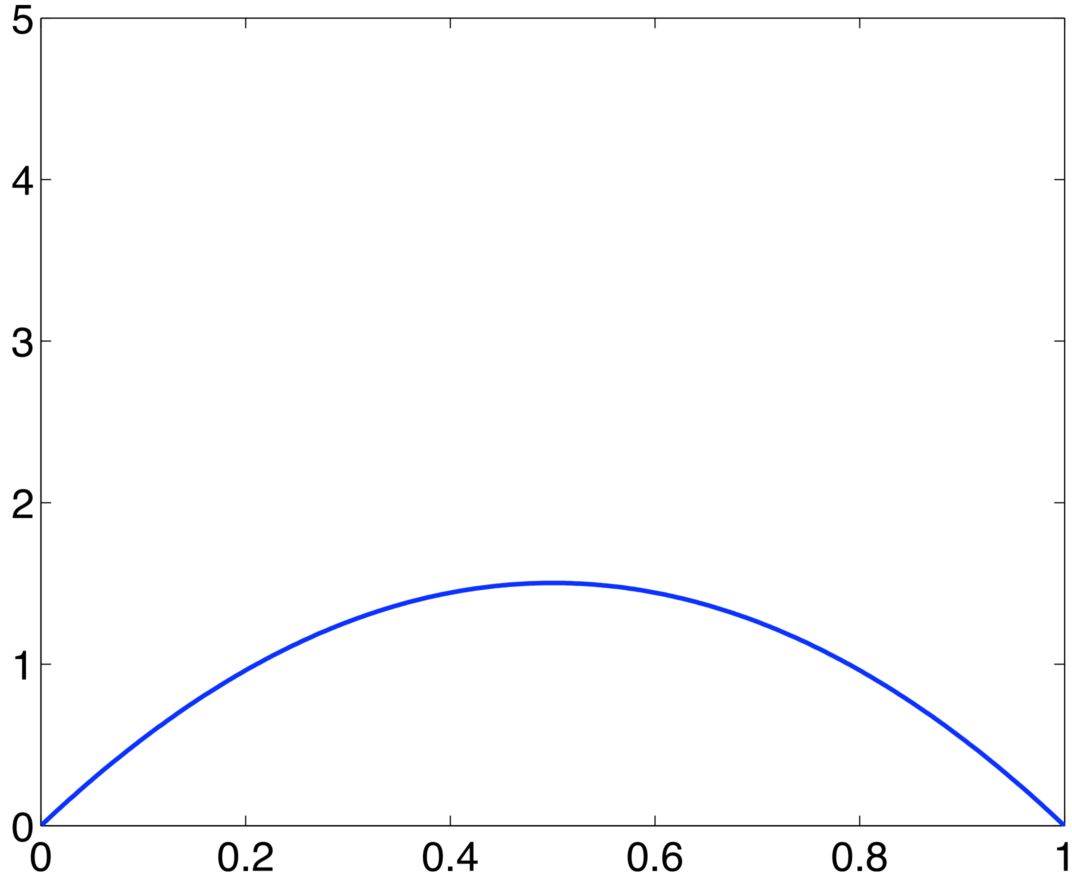
- *E.g., for coin toss, p ~ Beta(a, b):*

$$P(p \mid a, b) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$$

- *Specifying, e.g., a = 2, b = 2:*

$$P(p) = 6p(1-p)$$

Prior for p



Coin toss, cont'd

- *Joint dist'n of parameter p and data x_i:*

$$\begin{aligned} P(p, \mathbf{x}) &= P(p) \prod_i P(x_i \mid p) \\ &= 6p(1-p) \prod_i p^{x_i} (1-p)^{1-x_i} \end{aligned}$$

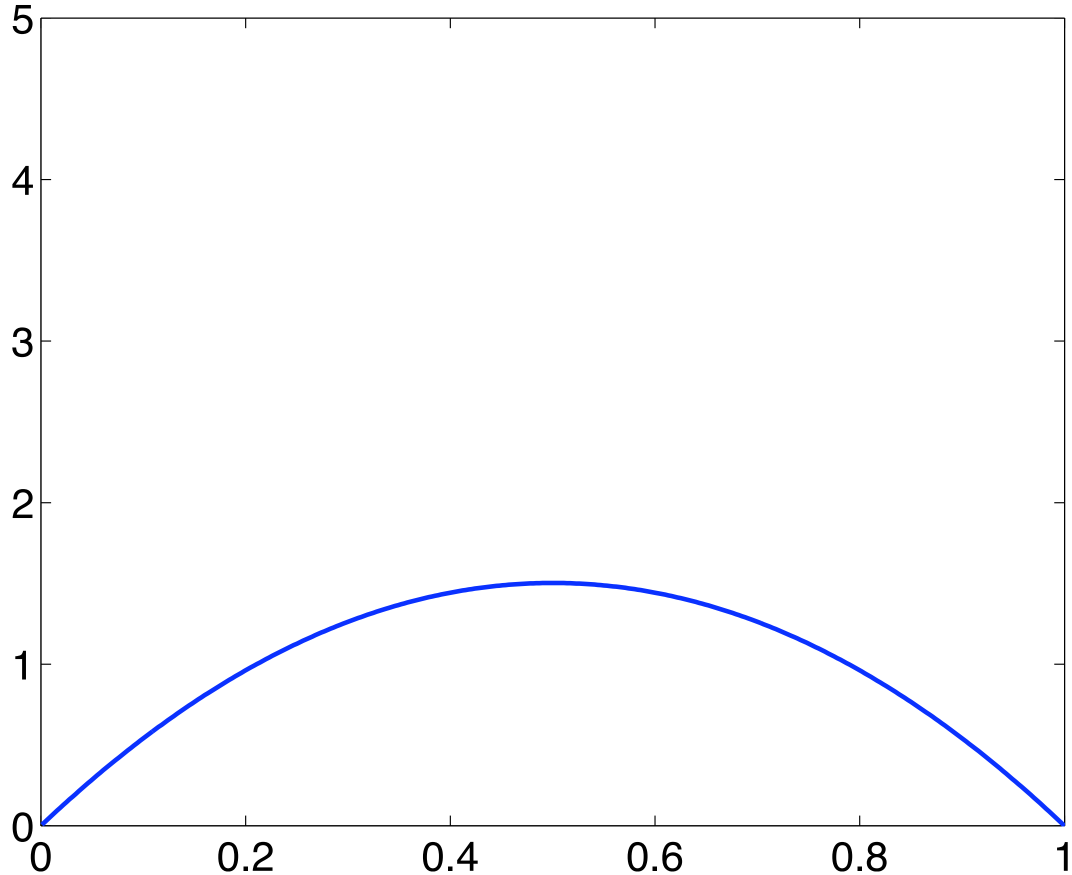
Posterior

- $P(\theta \mid \mathbf{D})$ *is* ***posterior***
- *Prior says what we know about* θ *before seeing* $\mathbf{D}$*; posterior says what we know after seeing* $\mathbf{D}$
- *Bayes rule:*
  - $P(\theta \mid \mathbf{D}) = P(\mathbf{D} \mid \theta)\, P(\theta) \,/\, P(\mathbf{D})$
- $P(\mathbf{D} \mid \theta)$ *is (data or sample)* ***likelihood***

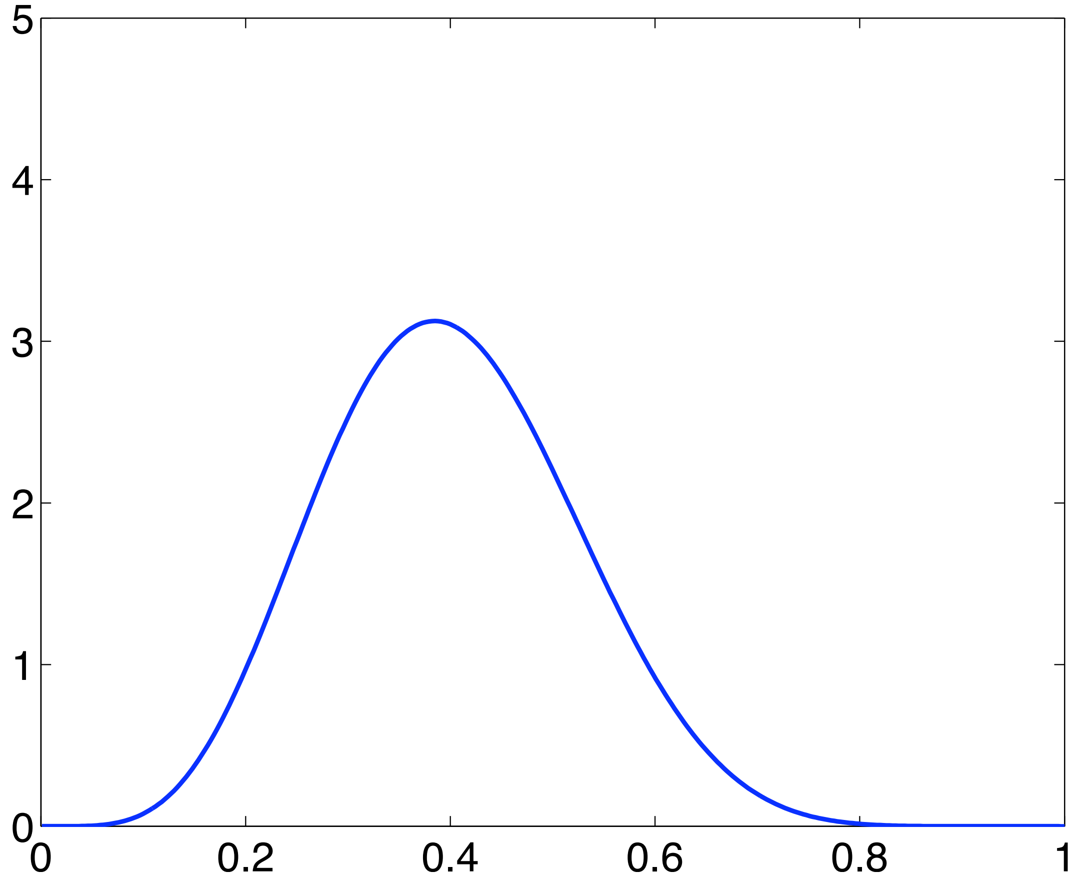
Coin flip posterior

$$
\begin{aligned}
P(p \mid \mathbf{x}) &= P(p) \prod_i P(x_i \mid p) / P(\mathbf{x}) \\
&= \frac{1}{Z} p(1-p) \prod_i p^{x_i} (1-p)^{1-x_i} \\
&= \frac{1}{Z} p^{1+\sum_i x_i} (1-p)^{1+\sum_i (1-x_i)} \\
&= \mathrm{Beta}(2 + \textstyle\sum_i x_i,\ 2 + \sum_i (1-x_i))
\end{aligned}
$$

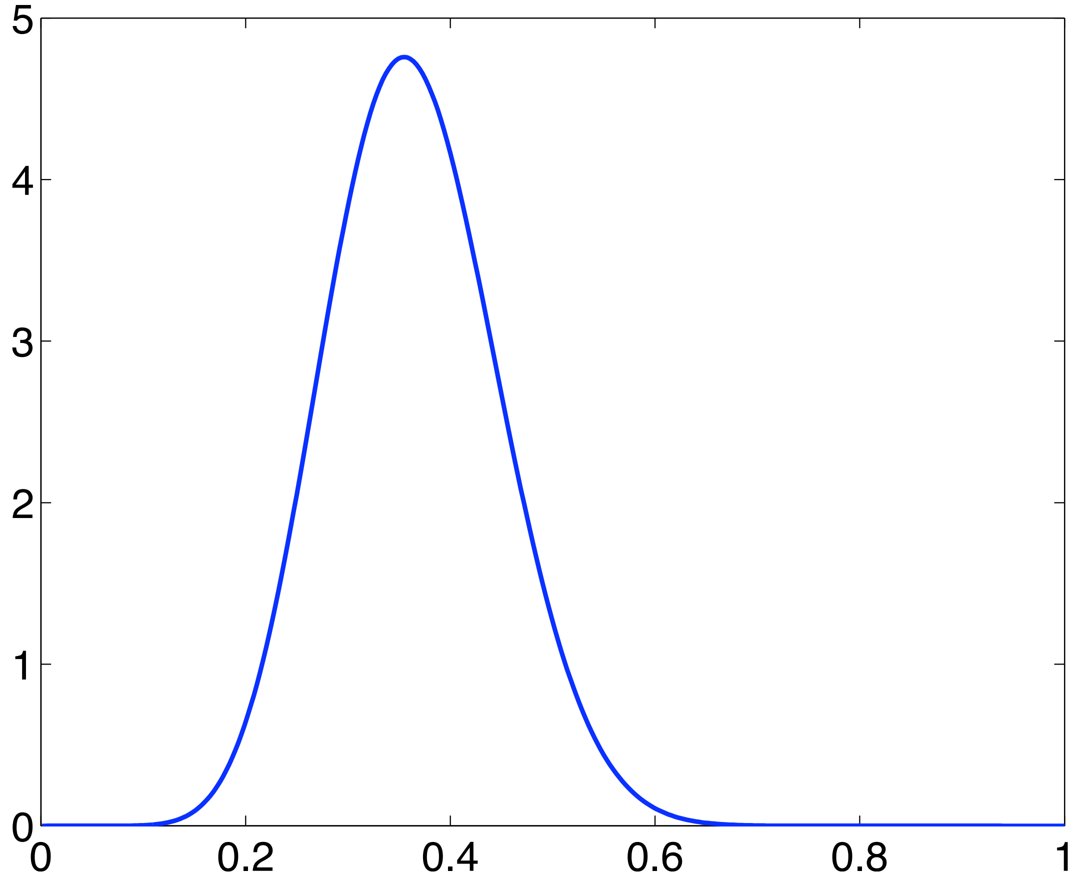
Prior for p



Posterior after 4 H, 7 T



Posterior after 10 H, 19 T



Where does prior come from?

- *Sometimes, we know something about θ ahead of time*
  - *in this case, encode knowledge in prior*
  - *e.g.,* $\|\theta\|$ *small, or* θ *sparse*
- *Often, we want prior to be* ***noninformative*** *(i.e., not commit to anything about* θ*)*
  - *in this case, make prior "flat"*
  - *then* $P(\mathbf{D} \mid \theta)$ *typically overwhelms* $P(\theta)$

Predictive distribution

- *Posterior is nice, but doesn't tell us directly what we need to know*
- *We care more about $P(x_{N+1} \mid x_1, \ldots, x_N)$*
- *By law of total probability, conditional independence:*

$$\begin{aligned} P(x_{N+1} \mid \mathbf{D}) &= \int P(x_{N+1}, \theta \mid \mathbf{D}) d\theta \\ &= \int P(x_{N+1} \mid \theta) P(\theta \mid \mathbf{D}) d\theta \end{aligned}$$

Coin flip example

- *After 10 H, 19 T:* $p \sim \mathit{Beta}(12, 21)$
- $E(x_{N+1} \mid p) = p$
- $E(x_{N+1} \mid \theta) = E(p \mid \theta) = a/(a+b) = 12/33$
- *So, predict 36.4% chance of H on next flip*

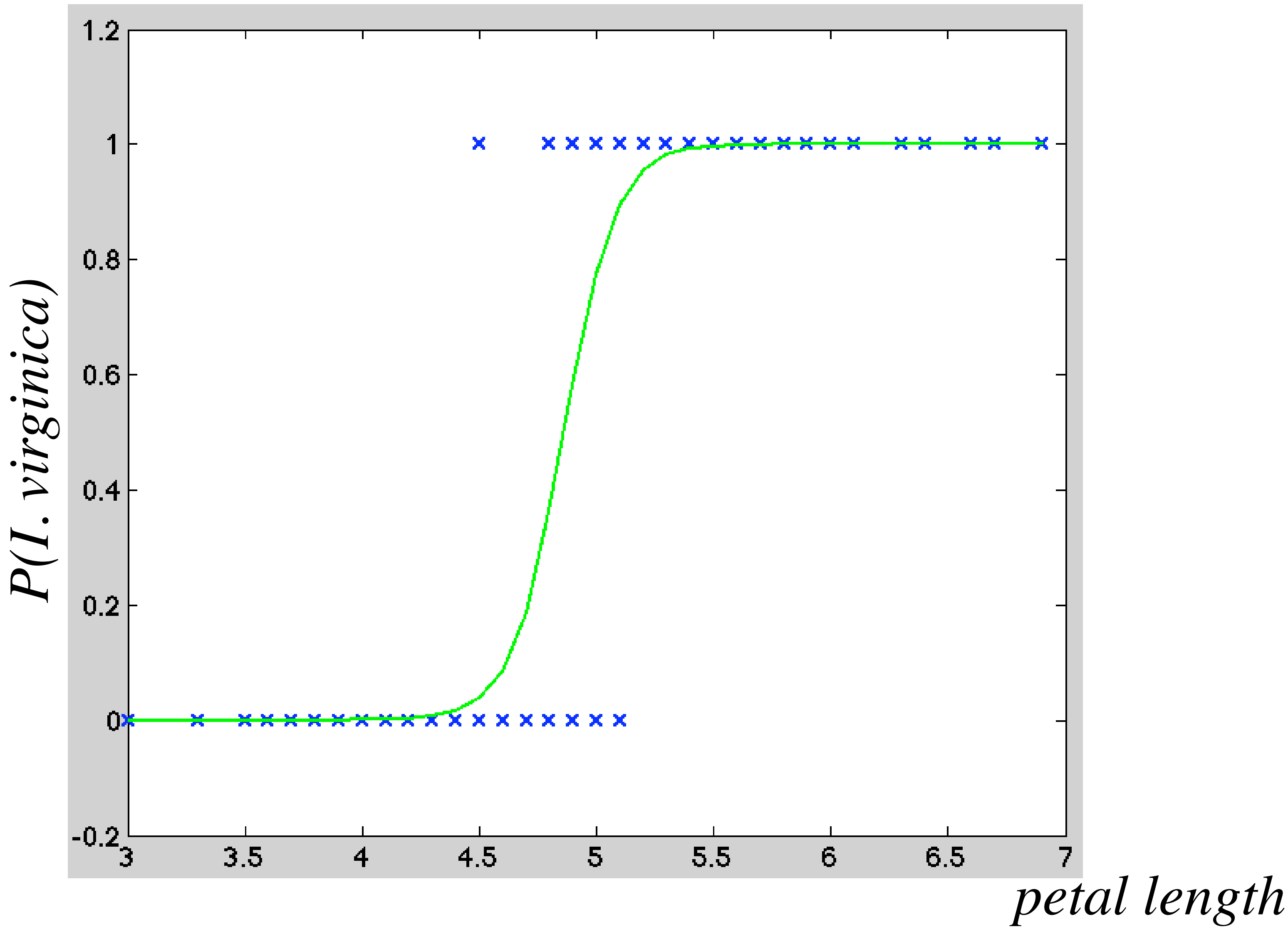
Approximate Bayes

Approximate Bayes

- *Coin flip example was easy*
- *In general, computing posterior (or predictive distribution) may be hard*
- *Solution: use the approximate integration techniques we've studied!*

Bayes as numerical integration

- *Parameters* θ, *data* $\boldsymbol{D}$
- $P(\theta \mid \boldsymbol{D}) = P(\boldsymbol{D} \mid \theta)\, P(\theta) \,/\, P(\boldsymbol{D})$
- *Usually,* $P(\theta)$ *is simple; so is* $Z\, P(\boldsymbol{D} \mid \theta)$
- *So,* $P(\theta \mid \boldsymbol{D}) \propto Z\, P(\boldsymbol{D} \mid \theta)\, P(\theta)$
- *Perfect for MH*



$$P(y \mid x) \quad = \quad \sigma(ax + b)$$
$$\sigma(z) \quad = \quad 1/(1 + exp(-z))$$

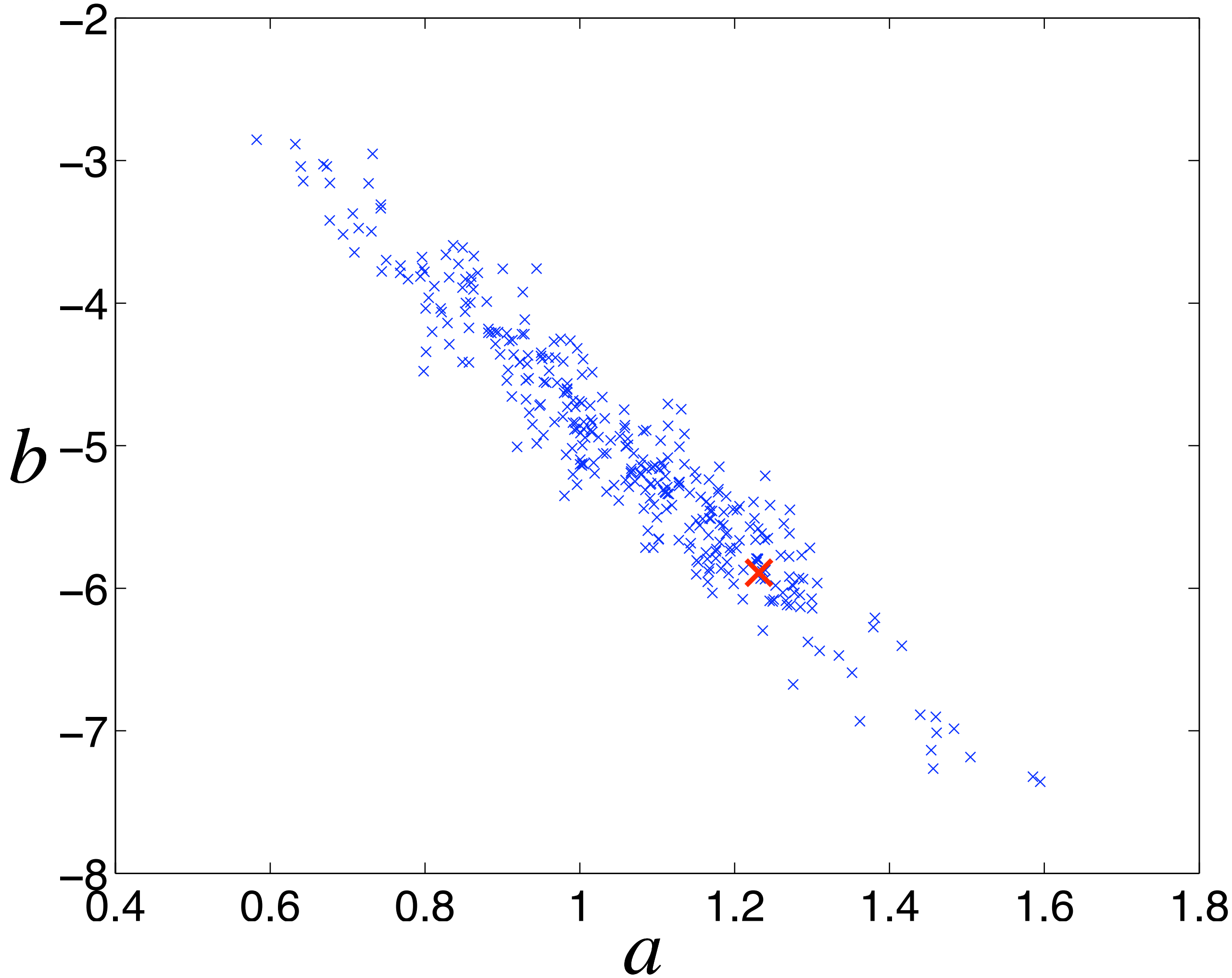
Posterior

$$P(a, b \mid x_i, y_i) =$$

$$Z P(a, b) \prod_i \sigma(a x_i + b)^{y_i} \sigma(-a x_i - b)^{1 - y_i}$$

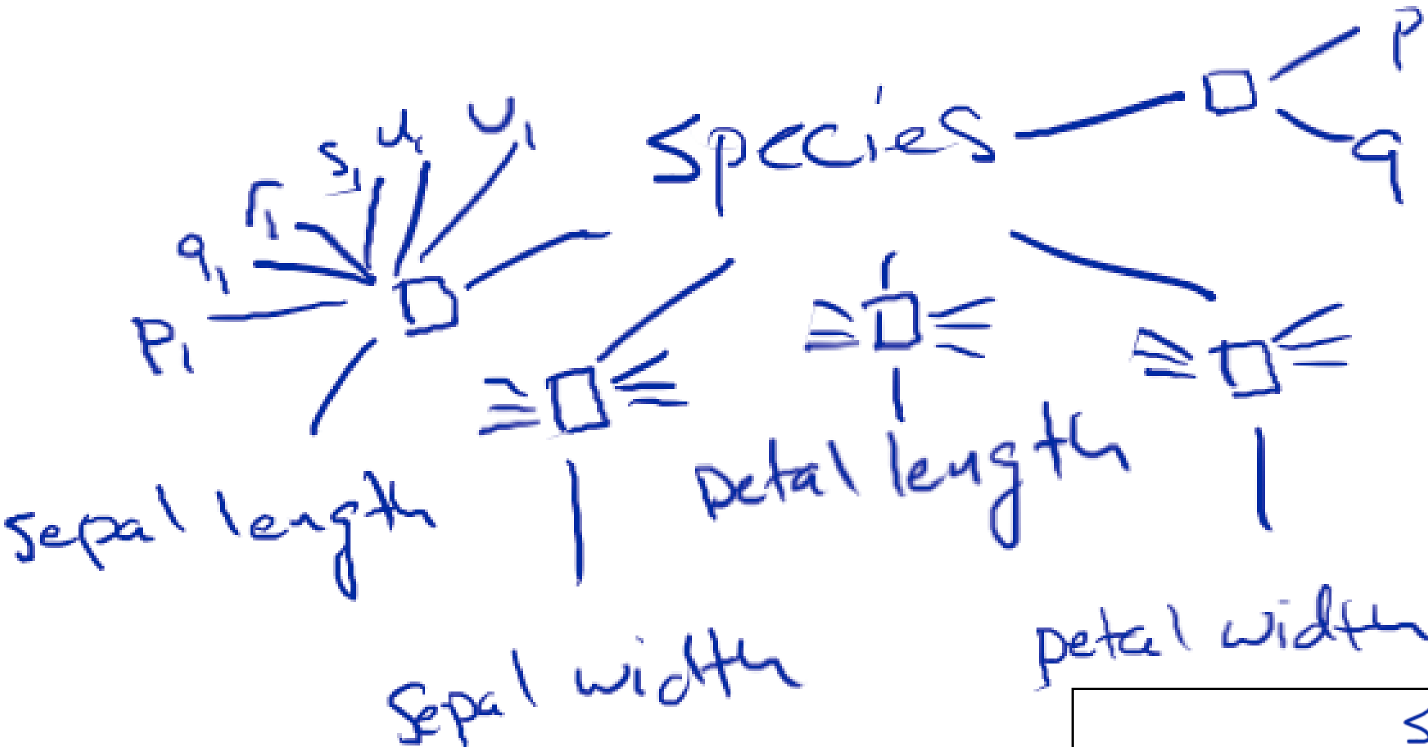
$$P(a, b) = N(0, I)$$

Sample from posterior

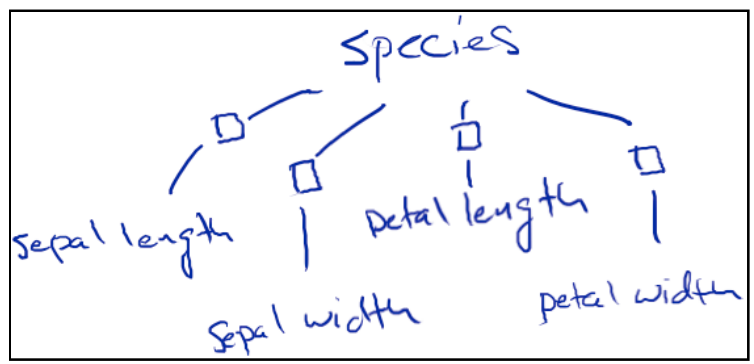


Bayes discussion

Expanded factor graph



original factor graph:



Inference vs. learning

- *Inference on expanded factor graph = learning on original factor graph*
  - *aside: why the distinction between inference and learning?*
  - *mostly a matter of algorithms: parameters are usually continuous, often high-dimensional*

Why Bayes?

- *Recall: we wanted to ensure our agent doesn't choose too many mistaken actions*
- *Each action can be thought of as a bet: e.g., eating X = bet X is not poisonous*
- *We choose bets (actions) based on our inferred probabilities*
- *E.g., R = 1 for eating non-poisonous, –99 for poisonous: eat iff P(poison) < 0.01*

Choosing bets

- *Don't know which bets we'll need to make*
- *So, Bayesian reasoning tries to set probabilities that result in reasonable betting decisions **no matter what** bets we are choosing among*
- *I.e., works if betting against an **adversary** (with rules defined as follows)*

Bayesian bookie

- *Bookie (our agent) accepts bets on any event (defined over our joint distribution)*
  - *A: next I. versicolor has petal length ≥ 4.2*
  - *B: next three coins in a row come up H*
  - *C: A ^ B*

Odds

- *Bookie can't refuse bets, but can set* ***odds****:*
  - *A: 1:1 odds (stake of \$1 wins \$1 if A)*
  - *¬B: 1:7 odds (stake of \$7 wins \$1 if ¬B)*
- *Must accept same bet in either direction*
  - *no "house cut"*
  - *e.g., 7:1 odds on B ⇔ 1:7 odds on ¬B*

Odds vs. probabilities

- *Bookie should choose odds based on probabilities*
- *E.g., if coin is fair,* $P(B) = 1/8$
- *So, should give 7:1 odds on B (1:7 on ¬B)*
  - *bet on B:* $(1/8)\,(7) + (7/8)\,(-1) = 0$
  - *bet on ¬B:* $(7/8)\,(1) + (1/8)\,(-7) = 0$
- *In general: odds* $x{:}y \Leftrightarrow p = y/(x+y)$

Conditional bets

- *We'll also allow conditional bets: "I bet that, if we go to the restaurant, Ted will order the fries"*
- *If we go and Ted orders fries, I win*
- *If we go and Ted doesn't order fries, I lose*
- *If we don't go, bet is called off*

How can adversary fleece us?

- *Method 1: by knowing the probabilities better than we do*
  - *if this is true, we're sunk*
  - *so, assume no informational advantage for adversary*
- *Method 2: by taking advantage of bookie's non-Bayesian reasoning*

Example of Method 2

- *Suppose I give probabilities:*

  $P(A)=0.5 \quad P(A \wedge B)=0.333 \quad P(B \mid A)=0.5$

- *Adversary will bet on A at 1:1, on ¬(A^B) at 1:2, and on B | A at 1:1*

Result of bet

A	B	$\$_1$	$\$_2$	$\$_3$	$\$_{ttl}$
T	T	1	-2	1	0
T	F	1	1	-1	1
F	T	-1	1	0	0
F	F	-1	1	0	0

- A at $1:1$ $\neg(A \wedge B)$ at $1:2$ $B|A$ at $1:1$

Dutch book

- *Called a "Dutch book"*
- *Adversary can print money, with no risk*
- *This is bad for us…*
  - *we shouldn't have stated* ***incoherent*** *probabilities*
  - *i.e., probabilities inconsistent with Bayes rule*

Theorem

- *If we do all of our reasoning according to Bayesian axioms of probability, we will never be subject to a Dutch book*
- *So, if we don't know what decisions we're going to need to make based on learned hypothesis H, we should use Bayesian learning to compute posterior P(H)*

Cheaper approximations

Getting cheaper

- *Maximum a posteriori (MAP)*
- *Maximum likelihood (MLE)*
- *Conditional MLE / MAP*

- *Instead of true posterior, just use single most probable hypothesis*

MAP

$$\arg\max_{\theta} P(D \mid \theta) P(\theta)$$

- *Summarize entire posterior density using the maximum*

MLE

$$\arg\max_{\theta} P(D \mid \theta)$$

- *Like MAP, but ignore prior term*

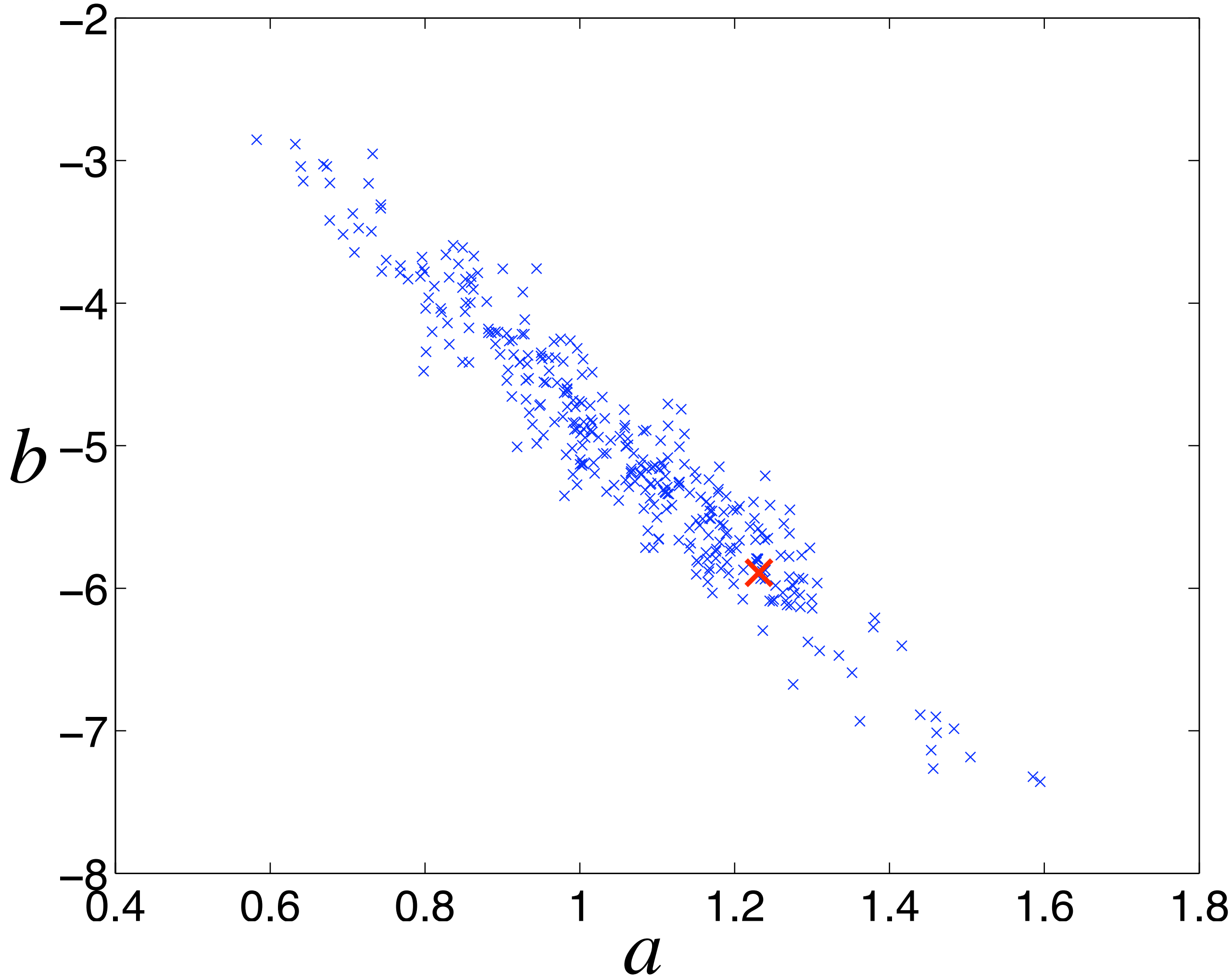
Conditional MLE, MAP

$$\arg\max_{\theta} P(\mathbf{y} \mid \mathbf{x}, \theta)$$

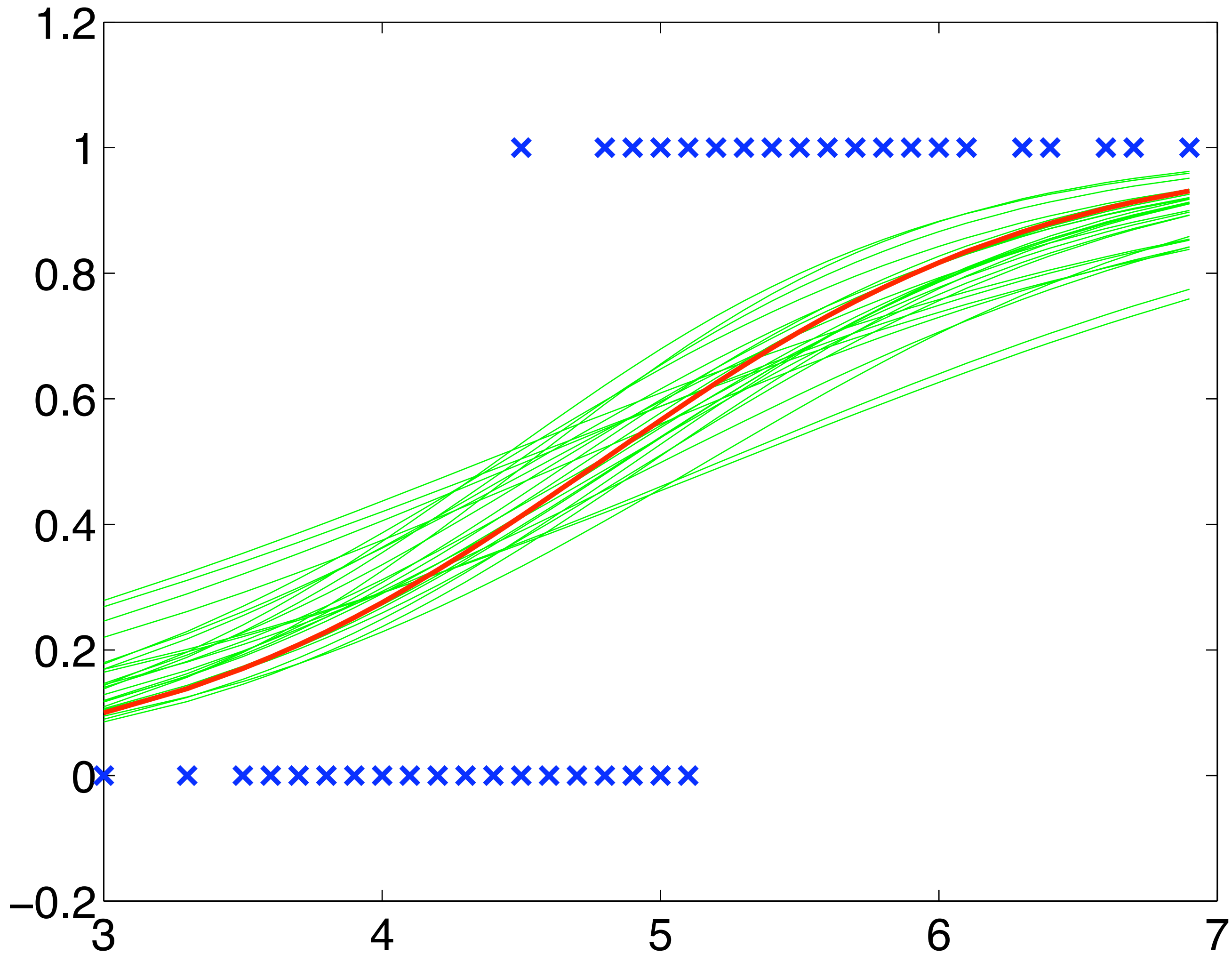
$$\arg\max_{\theta} P(\mathbf{y} \mid \mathbf{x}, \theta) P(\theta)$$

- *Split D = (**x**, **y**)*
- *Condition on **x**, try to explain only **y***

Iris example: MAP vs. posterior



Irises: MAP vs. posterior



Too certain

- *This behavior of MAP (or MLE) is typical: we are too sure of ourselves*
- *But, often gets better with more data*
- *Theorem: MAP and MLE are consistent estimates of true θ, if "data per parameter" $\rightarrow \infty$*