Probabilistic Machine Learning

Day 1: Introduction, Probability Theory, Information Theory and Bayesian Inference

Joaquin Quiñonero Candela
joaquinc@microsoft.com

Applied Games Group
Microsoft Research Cambridge, UK

PASCAL Bootcamp in Machine Learning
Vilanova i la Geltrú, July 5 and 6, 2007

Why Probabilistic Models for Learning?

Probabilities can be used to quantify information gained from the training set

Probability theory gives a framework for computing with such quantities

- A probabilistic model can be used to
  - make predictions (with an indication of uncertainty)
  - make decisions (which minimize expected loss)
  - make inferences about missing inputs
  - find good representations of data
  - phantasize data (generative model)

- Probabilistic models are equivalent to other views of learning:
  - information theoretic: find a compact representation of the data
  - physical analogies: minimizing free energy of corresponding statistical mechanical system

Probabilities and Ensembles

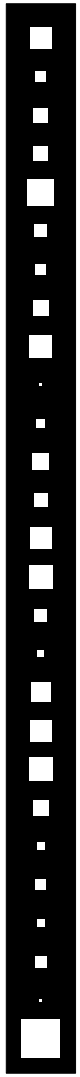
An ensemble is a triple $(x, \mathcal{A}_\mathcal{X}, \mathcal{P}_\mathcal{X})$:

- the outcome x is the value of a random variable,
- x takes values from set $\mathcal{A}_\mathcal{X} = \{a_1, a_2, \ldots, a_L\}$,
- with probabilities $\mathcal{P}_\mathcal{X} = \{p_1, p_2, \ldots, p_L\}$.
- $P(x = a_i) = p_i, \quad p_i \geq 0$
- $\sum_{a_i \in \mathcal{A}_\mathcal{X}} P(x = a_i) = \sum_{i=1}^{L} p_i = 1.$

Simpler notation:

$$P(x = a_i) = P(x) = P(a_i)$$

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928



(from David MacKay)

Basic Rules of Probability

Probabilities are non-negative $p(x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(x) = 1$ (discrete) or $\int_{-\infty}^{+\infty} p(x)dx = 1$ (continuous)

The joint probability of x and y is: $P(x, y)$.

The marginal probability of x is: $P(x) = \sum_y P(x, y)$, assuming y is discrete.

The conditional probability of x given y is: $P(x|y) = P(x, y)/P(y)$

Bayes Rule:

$$p(x, y) \; = \; p(x)p(y|x) \; = \; p(y)p(x|y) \; \Rightarrow \; p(y|x) = \frac{p(x|y)p(y)}{p(x)} \; \propto \; p(x|y)p(y)$$

Expectation and Variance (Moments)

The expectation (mean, average) of a random variable is:

$$\mu \;=\; \mathbb{E}[x] \;=\; \int x\, p(x)\, \mathrm{d}x \;=\; \langle x \rangle_{p(x)} \;.$$

The variance (or second central moment) is:

$$\sigma^2 \;=\; \mathbb{V}[x] \;=\; \int (x-\mu)^2\, p(x)\, \mathrm{d}x \;=\; \mathbb{E}[x^2] - \mathbb{E}[x]^2 \;.$$

The covariance between x and y:

$$\mathrm{cov}(x,y) \;=\; \mathbb{E}[(x-\mathbb{E}[x])(y-\mathbb{E}[y])] \;.$$

If x and y are independent, then their covariance is zero, since:

$$p(x,y) = p(x)\, p(y) \;.$$

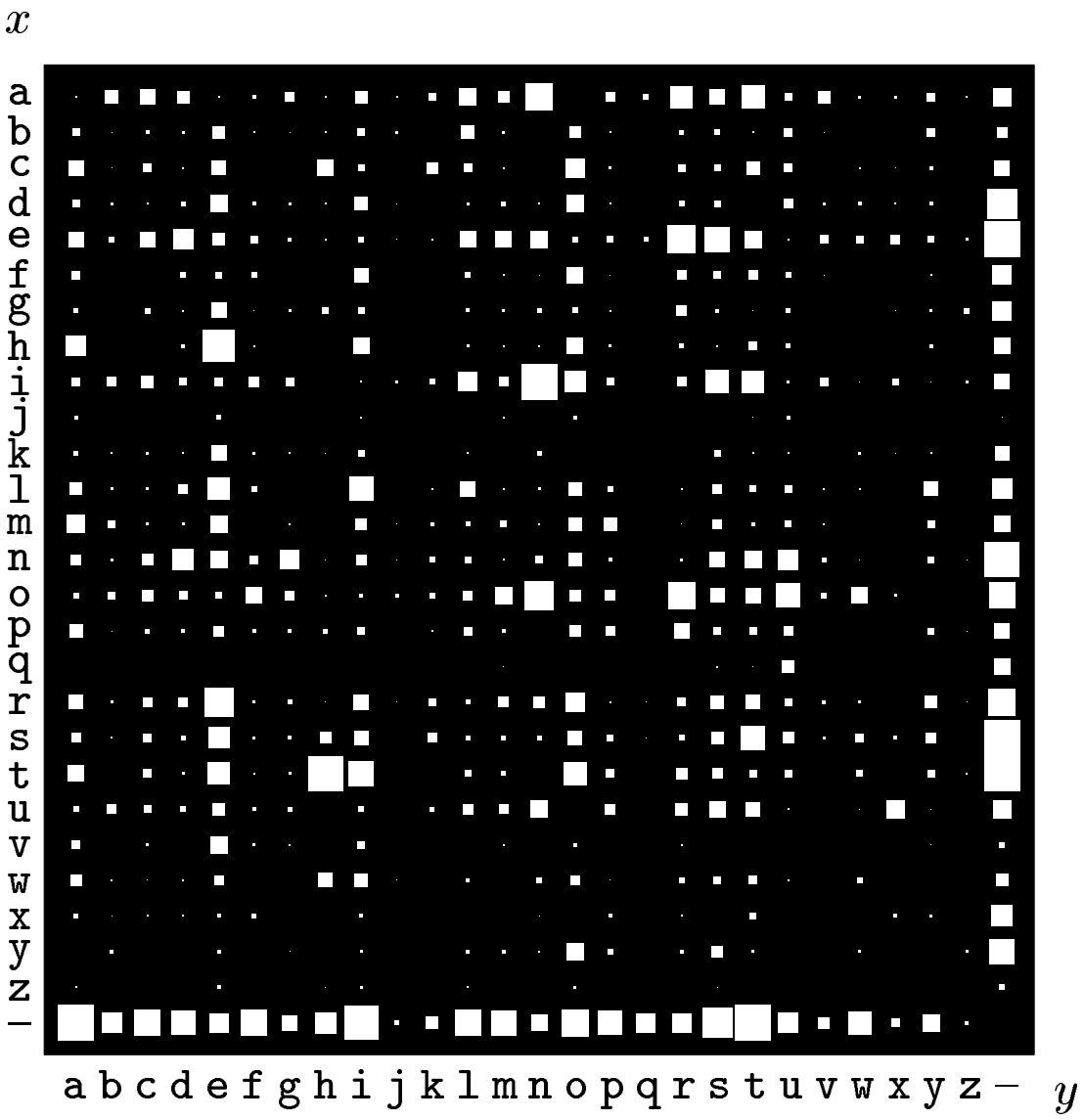
Example of Joint Probability - Bigrams

- Bigrams: probability of letter x followed by leter y
- Marginal probability from joint:

$$P(x = a_i) \;=\; \sum_{y \in \mathcal{A}_\mathcal{Y}} P(x = a_i,\, y)\;.$$

- Similarly

$$P(y) \;=\; \sum_{x \in \mathcal{A}_\mathcal{X}} P(x,\, y)\;.$$



(figure from David MacKay)

An Exercise on Mammographies

The facts:

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9,6% of women without breast cancer also get positive mammography

The question:

A woman has a positive mammography. What is the probability that she has breast cancer?

Quick guess:

a) less than 1%

b) somewhere between 1% and 70%

c) between 70% and 80%

d) more than 80%

Solving the Mammography Exercise
Writing Down Probabilities

Write down the probabilities of everything (Steve Gull):

- Define: C = presence of breast cancer, $\bar{C}$ = no cancer
- Define: M = positive mammography, $\bar{M}$ = negative mammography
- The **prior** probability of cancer for scanned women is $p(C) = 1\%$
- If there **is** cancer, the probability of a positive mammography is $p(M|C) = 80\%$
- If there is no cancer, we still have $p(M|\bar{C}) = 9,6\%$

The question is: what is $p(C|M)$?

Solving the Mammography Exercise
Playing with Concrete Numbers

With a little help from concrete numbers: consider 10000 subjects of screening

- $p(C) = 1\%$, therefore 100 of them have cancer, of which
  - $p(M|C) = 80\%$, therefore 80 get a positive mammography
  - 20 get a negative mammography

- $p(C) = 1\%$, therefore 9900 of them do not have cancer, of which
  - $p(M|\bar{C}) = 9,6\%$, therefore 950 get a positive mammography
  - 8950 get a a negative mammography

Let us see where our 10000 subjects fall:

	M	$\bar{M}$
C	80	20
$\bar{C}$	950	8950

Solving the Mammography Exercise
Apply Bayes' Rule

- A very natural way of obtaining Bayes' Rule:

	M	$\bar{M}$
C	80	20
$\bar{C}$	950	8950

- Marginal: total number of positive mammographies $p(M) = p(C, M) + p(\bar{C}, M)$

- $p(C|M)$ is the proportion of all positive mammographies for which there is breast cancer:

$$p(C|M) = \frac{p(C,M)}{p(M)} = \frac{p(C,M)}{p(C,M) + p(\bar{C},M)} = \frac{80}{80 + 950} \approx 7,8\%$$

Bayes' rule

$$p(C|M) = \frac{P(M,C)}{P(M)} = \frac{P(M|C)P(C)}{P(M)}$$

Do you trust your doctor?

Although the probability of a positive mammography given cancer is 80%, the probability of cancer given a positive mammography is 7.8%.

85% of the doctors would have said: c) between 70% and 80%
This is totally wrong!

Common mistakes

- "the probability that a woman with positive mammography has cancer" is not at all the same as "the probability that a woman with cancer has a positive mammography".

- One must also consider the original fraction of women with breast cancer, and those without breast cancer who receive false positives.

from `http://yudkowsky.net/bayes/bayes.html`:

An Intuitive Explanation of Bayesian Reasoning

Bayes' Theorem for the curious and bewildered; an excruciatingly gentle introduction.

By Eliezer Yudkowsky

Information, Probability and Entropy

Information is the reduction of uncertainty:

> How much information do we gain by observing the outcome $x = a_i$ of a random variable? As much as the corresponding reduction of the uncertainty.

How do we measure uncertainty?

Some axioms (informal):

- if something is certain its uncertainty $= 0$
- uncertainty should be maximum if all choices are equally probable
- uncertainty (information) should add for independent sources

Exercise: Consider 12 balls of equal weight, one of which is either lighter or heavier. You have a two-pan balance that outputs *left* is heavier, *right* is heavier, or *balanced*. How many uses of the balance do you need to find the odd ball?

(exercise 4.1 in David MacKay's book)

Entropy

Let X be a random variable X whose outcome x takes values in $\{a_1, \ldots, a_L\}$ with probabilities $\{p_1, \ldots, p_L\}$

The Shannon information content of the outcome $x = a_i$ is:

$$h(x = a_i) = \log_2 \frac{1}{p_i}$$

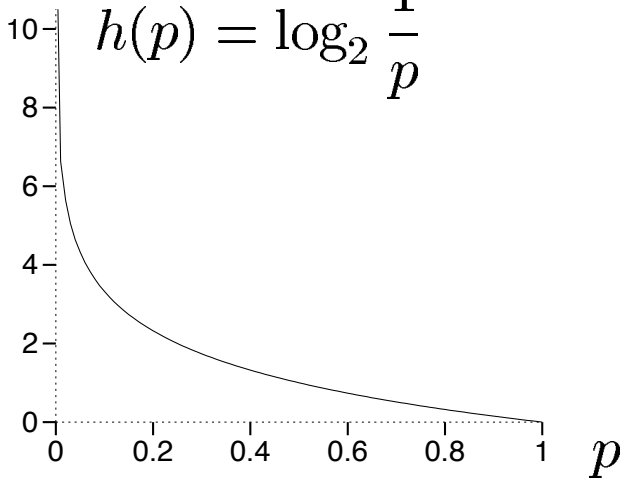
The entropy of the random variable X is the average information content:

$$H(X) = \sum_i p_i \log \frac{1}{p_i} = -\sum_i p_i \log p_i$$

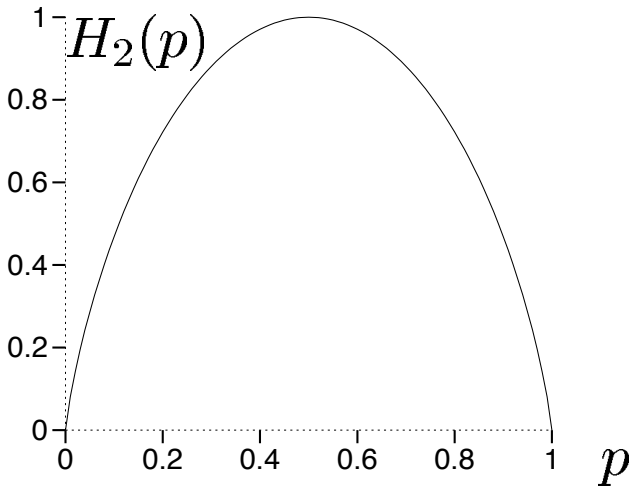
measured in *bits* (**bi**nary digi**ts**): base 2 log or *nats* (**na**tural digi**ts**): natural (base e) log.

Entropy of a Binary Random Variable

Consider a binary random variable, that can take two values with probabilities p and $1-p$.



p	$h(p)$	$H_2(p)$
0.001	10.0	0.011
0.01	6.6	0.081
0.1	3.3	0.47
0.2	2.3	0.72
0.5	1.0	1.0



(fig 4.1 in David MacKay's book)

Improbable events are more informative, but less frequent on average.

The entropy satisfies the two first axioms

- observation of a certain event carries no information
- maximum information is carried by uniformly probable events

Information Between Two Random Variables

- **Joint entropy H(X,Y):**
  If X and Y are independent, then $p(x,y) = p(x)p(y)$ and the Shannon information content

$$h(x,y) = \log \frac{1}{p(x,y)} = \log \frac{1}{p(x)} + \log \frac{1}{p(y)}$$

  is additive. As a consequence $H(X,Y) = H(X) + H(Y)$ and the third axiom is satisfied.

- **Conditional entropy:** average uncertainty remaining about x if we have observed y

$$H(X|Y) = -\sum_x \sum_y p(x,y) \log_2 p(x|y) = H(X,Y) - H(Y)$$

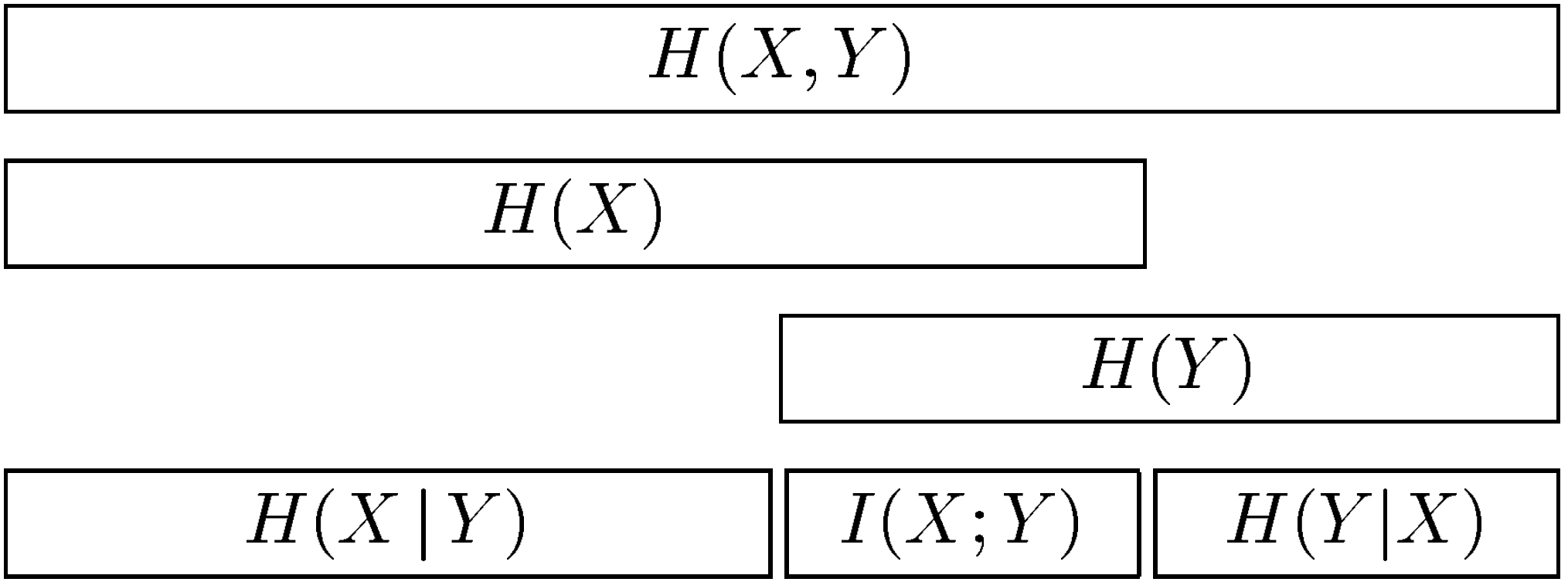
  (if X and Y are independent $H(X|Y) = H(X)$)

- **Mutual information:** average reduction in uncertainty about x if we observe y

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

  or vice versa. (if X and Y are independent $I(X;Y) = 0$)

Information Between Two Random Variables (2)



(from David MacKay's book)

$H(X,Y)$ is the **joint entropy** of X, Y

$H(X|Y)$ is the **conditional entropy** of X given Y

$I(X;Y)$ is the **mutual information** between X and Y

Kullback-Leibler Divergence

Kullback-Leibler divergence (relative entropy)

$$KL(p(x)\|q(x)) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Note that this is not a distance, since it is not necessarily symmetric:

In general $KL(p(x)\|q(x)) \neq KL(q(x)\|p(x))$

The KL divergence is very important in probabilistic machine learning. We will encounter it often again, for example in the expectation-maximization (EM) algorithm.

Relation between mutual information and KL:

$$I(X;Y) = KL(p(x,y)\|p(x)p(y))$$

(this is symmetric)

Shannon's Source Coding Theorem

A discrete random variable X, distributed according to $p(x)$ has entropy equal to:

$$H(X) = -\sum_x p(x) \log p(x)$$

Shannon's source coding theorem: n independent samples of the random variable X, with entropy $H(X)$, can be compressed into minimum expected code of length $n\mathcal{L}$, where

$$H(X) \leq \mathcal{L} < H(X) + \frac{1}{n}$$

If each symbol is given a code length $l(x) = -\log_2 q(x)$ then the expected per-symbol length $\mathcal{L}_q$ of the code is

$$H(X) + KL(p\|q) \leq \mathcal{L}_q < H(X) + KL(p\|q) + \frac{1}{n},$$

where the relative-entropy or Kullback-Leibler divergence is

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0$$

What is Probability?

Two possible interpretations:

- long run frequencies (frequentist, classical)
- subjective degrees of **belief** (Bayesian)

How can we represent the beliefs of a learning agent (robot)?

"Is Luke going to need that light saber right now?"

"Am I being too rude to C3P0?"

"How sure am I about my present location?"

We want to represent the **strength** of beliefs numerically in the brain of the robot, and we want to know what rules (calculus) we should use to manipulate those beliefs.

Beliefs and Probability

Let $b(x)$ be the degree of belief in proposition x. The degree of belief in a conditional proposition, 'x, assuming that y is true' is $b(x|y)$, and

$0 \leq b(x) \leq 1$, $\quad b(x) = 0$ [x is definitely **not true**] , $\quad b(x) = 1$ [x is definitely **true**]

Degrees of belief can be mapped onto probabilities if they satisfy simple consistency rules: **Cox axioms** (Cox, 1946):

1. Degrees of belief can be ordered. If $b(z)$ is greater than $b(y)$ and $b(y)$ is greater than $b(x)$, then $b(z)$ is greater than $b(x)$.

2. The degree of belief in x and in its negation $\bar{x}$ are related: $b(x) = f[b(\bar{x})]$.

3. The degree of belief in x AND y is related to the degree of belief in the conditional proposition $x|y$ and the degree of belief in the proposition y.

Consequence: Belief functions (e.g. $b(x)$, $b(x|y)$, $b(x, y)$) must satisfy the rules of probability theory, including Bayes rule. (Jaynes, *Probability Theory: The Logic of Science*)

Bayesian Learning

Apply the basic rules of probability to learning from data.

Data set: $\mathcal{D} = \{x_1, \ldots, x_n\}$ Models: m, m' etc. Model parameters: θ
Prior probability of models: $P(m)$, $P(m')$ etc.
Prior probabilities of model parameters: $P(\theta|m)$
Model of data given parameters (likelihood model): $P(x|\theta, m)$

If the data are independently and identically distributed then:

$$P(\mathcal{D}|\theta, m) = \prod_{i=1}^{n} P(x_i|\theta, m)$$

Posterior probability of model parameters:

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

Posterior probability of models:

$$P(m|\mathcal{D}) = \frac{P(m)P(\mathcal{D}|m)}{P(\mathcal{D})}$$

Bayesian Learning: A coin toss example

The likelihood **of the parameters given the data** is the probability of the observed data given the parameters.

- Parameter: $\theta = \pi$ (probability of heads), data: tail $x = 0$, heads $x = 1$. Given π, independent Bernoulli likelihood:

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

- Observations: $\mathcal{D} = \{x_i \in \{T, H\} | i = 1, \ldots, n\}$, summary: n and number of heads k

- Likelihood of the parameter given the data $\mathcal{D}$:

$$p(\mathcal{D}|\pi) = \pi^k(1-\pi)^{n-k}$$

- Maximum Likelihood estimation:

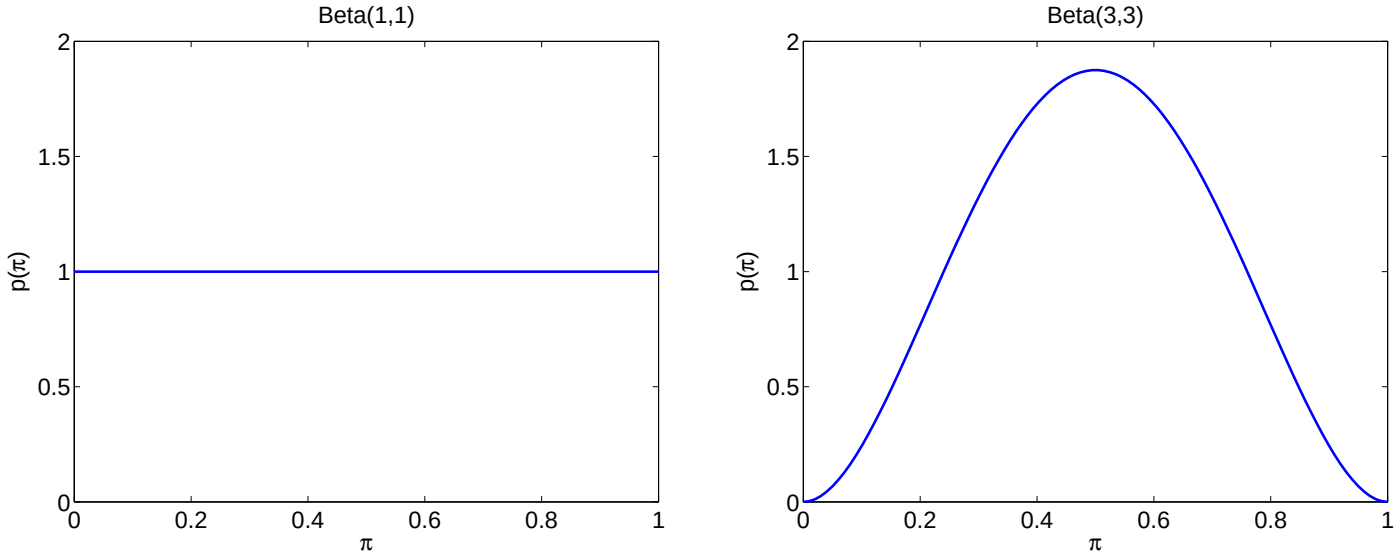
$$\frac{\partial \log p(\mathcal{D}|\pi)}{\partial \pi} = 0 \Rightarrow \pi = \frac{k}{n}$$

  Is this always a good answer?

Priors for Coin Tossing

Compare two different models (priors):

- Learner A believes all values of π are equally plausible.
- Learner B believes that it is more plausible that the coin is "fair" ($\pi \approx 0.5$) than "biased".



A B

- We can write these prior beliefs using a Beta distribution

$$p(\pi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}$$

for A: $\alpha = \beta = 1.0$ and B: $\alpha = \beta = 3.0$.

Posterior for Coin Tossing

Two possible outcomes:

$$p(\text{heads}|\pi) = \pi \qquad p(\text{tails}|\pi) = 1 - \pi$$

Imagine we observe a single coin toss and it comes out *heads*
The probability of the observed data (likelihood) is:

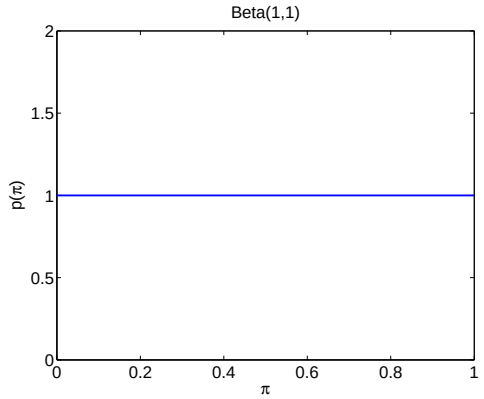
$$p(\text{heads}|\pi) = \pi$$

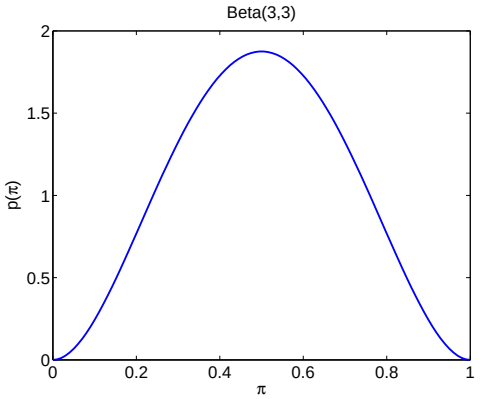
Using Bayes Rule, we multiply the prior, $p(\pi)$ by the likelihood and renormalise to get the posterior probability:

$$\begin{aligned} p(\pi|\text{heads}) &= \frac{p(\pi)p(\text{heads}|\pi)}{p(\text{heads})} \propto \pi \, \text{Beta}(\pi|\alpha, \beta) \\ &\propto \pi \, \pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} = \text{Beta}(\pi|\alpha+1, \beta) \end{aligned}$$

Before and After Observing One Head

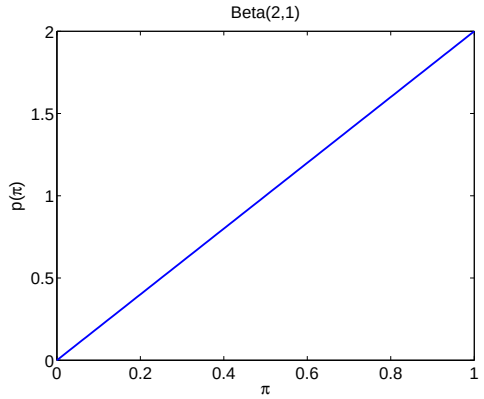
Prior

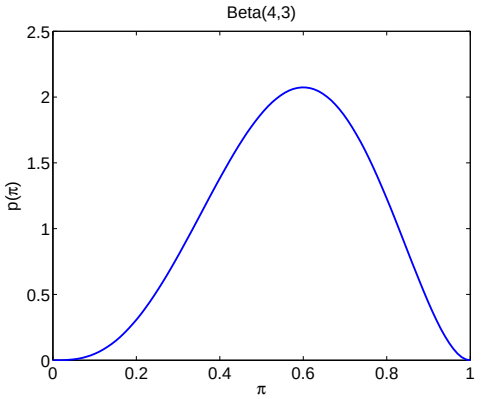




A

B





Posterior

Making Predictions

As opposed to the Maximum Likelihood approach, **average** over all possible parameter settings:

$$p(x = 1|\mathcal{D}) = \int p(x = 1|\pi)\, p(\pi|\mathcal{D})\, \mathrm{d}\pi$$

Learner A predicts $p(x = 1|\mathcal{D}) = \frac{2}{3}$

Learner B predicts $p(x = 1|\mathcal{D}) = \frac{4}{7}$

Some Terminology

Maximum Likelihood (ML) Learning: Does not assume a prior over the model parameters. Finds a parameter setting that maximises the likelihood of the data: $P(\mathcal{D}|\theta)$.

Maximum a Posteriori (MAP) Learning: Assumes a prior over the model parameters $P(\theta)$. Finds a parameter setting that maximises the posterior: $P(\theta|\mathcal{D}) \propto P(\theta)P(\mathcal{D}|\theta)$.

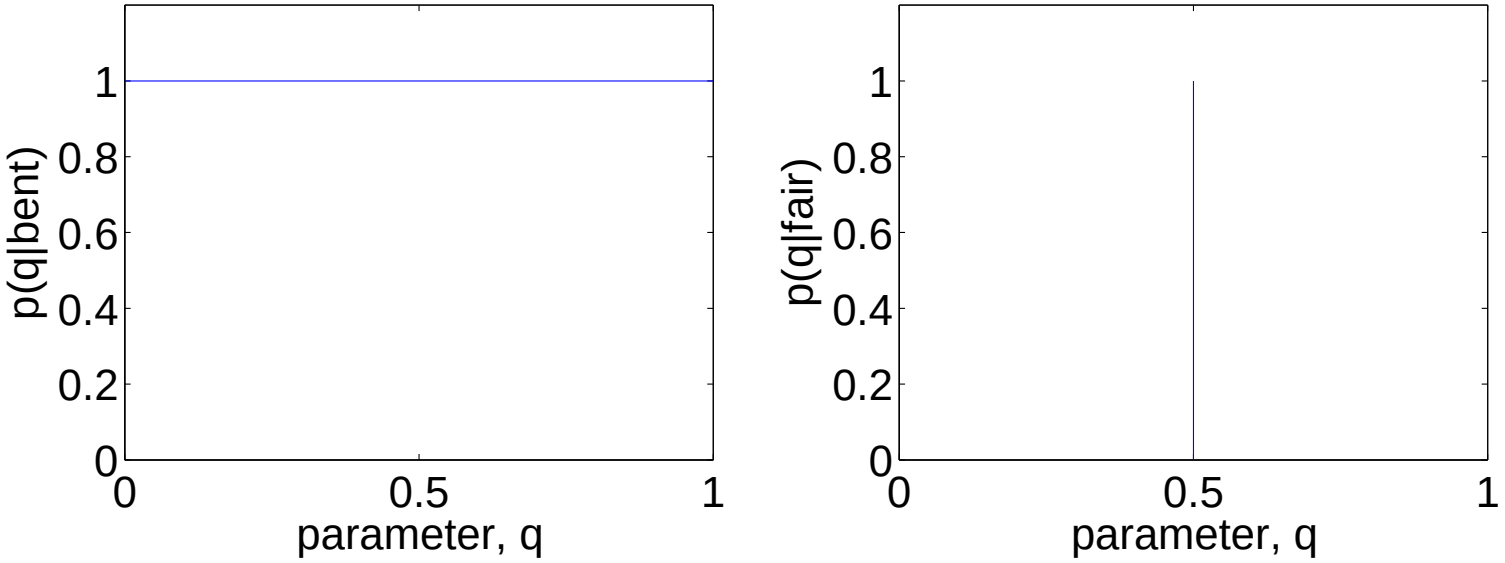
Bayesian Learning: Assumes a prior over the model parameters. Computes the posterior distribution of the parameters: $P(\theta|\mathcal{D})$.

Learning about a coin II

Consider two alternative models of a coin, "fair" and "bent". A priori, we may think that "fair" is more probable, eg:

$$p(\text{fair}) = 0.8, \qquad p(\text{bent}) = 0.2$$

For the bent coin, (a little unrealistically) all parameter values could be equally likely, where the fair coin has a fixed probability:



We make 10 tosses, and get: T H T H T T T T T T

Learning about a coin. . .

The evidence for the fair model is: $p(\mathcal{D}|\text{fair}) = (1/2)^{10} \simeq 0.001$
and for the bent model:

$$p(\mathcal{D}|\text{bent}) = \int d\pi \; p(\mathcal{D}|\pi, \text{bent}) p(\pi|\text{bent}) = \int d\pi \; \pi^2 (1-\pi)^8 = \text{B}(3,9) \simeq 0.002$$

The posterior for the models, by Bayes rule:

$$p(\text{fair}|\mathcal{D}) \propto 0.0008, \qquad p(\text{bent}|\mathcal{D}) \propto 0.0004,$$

ie, two thirds probability that the coin is fair.
How do we make predictions? By weighting the predictions from each model by their probability. Probability of Head at next toss is:

$$\frac{2}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{3}{12} = \frac{5}{12}.$$

Bayesian Classification

Example: Linear Classification (example from Radford Neal's NIPS*04 tutorial)

- Binary classification in 2D input space $y \in \{-1, +1\}$
- Model: linear decision function, "Hard" likelihood:

$$p(y = +1|x, w, M) \;=\; \mathrm{sign}\big(u\,(w^\top x - \|w\|^2)\big)$$

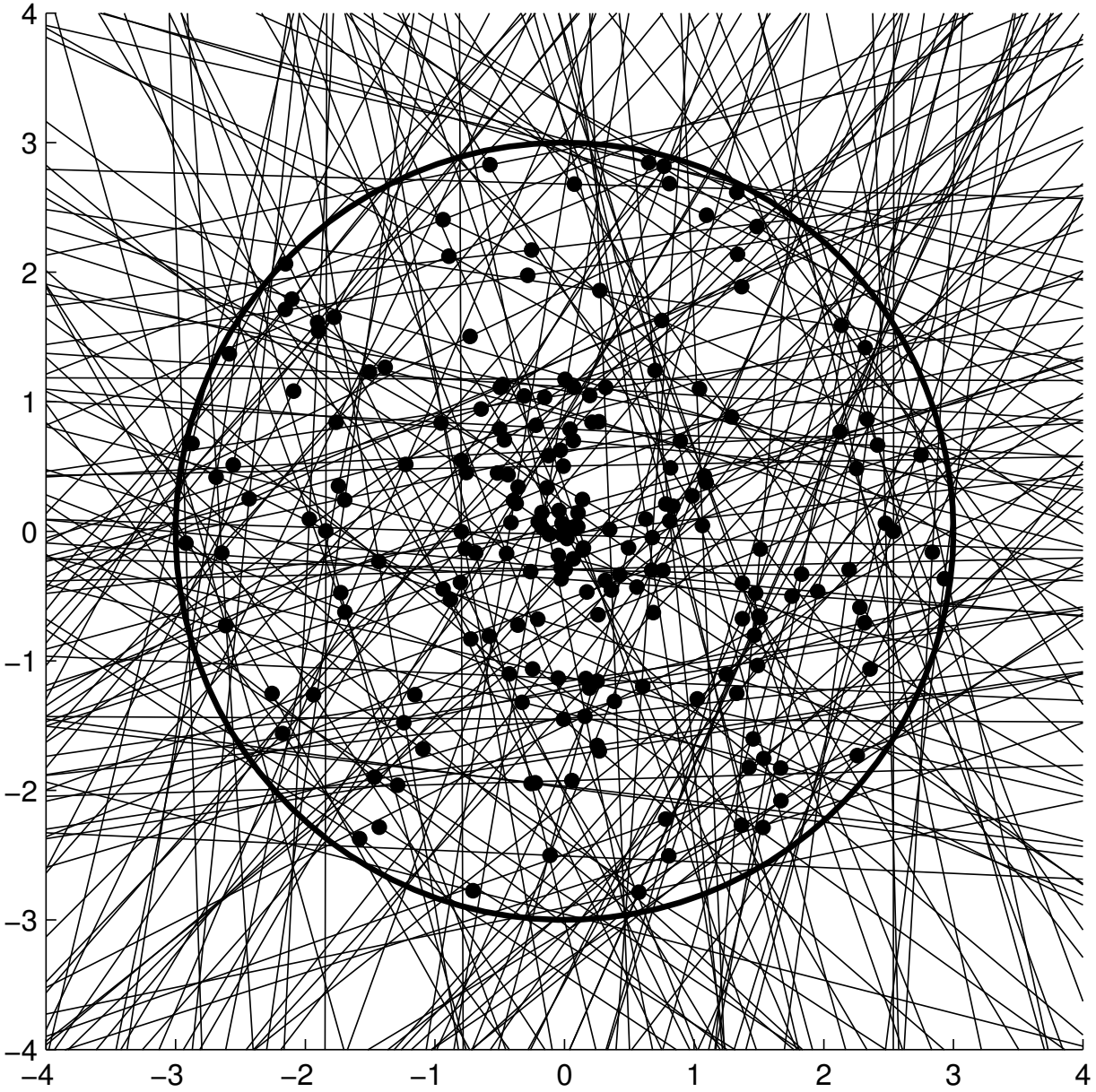
  (where $w = [w_1, w_2]^\top$ and $u \in \{-1, +1\}$ are model parameters)

- Prior: decision boundary anywhere distant at most 3 from origin:

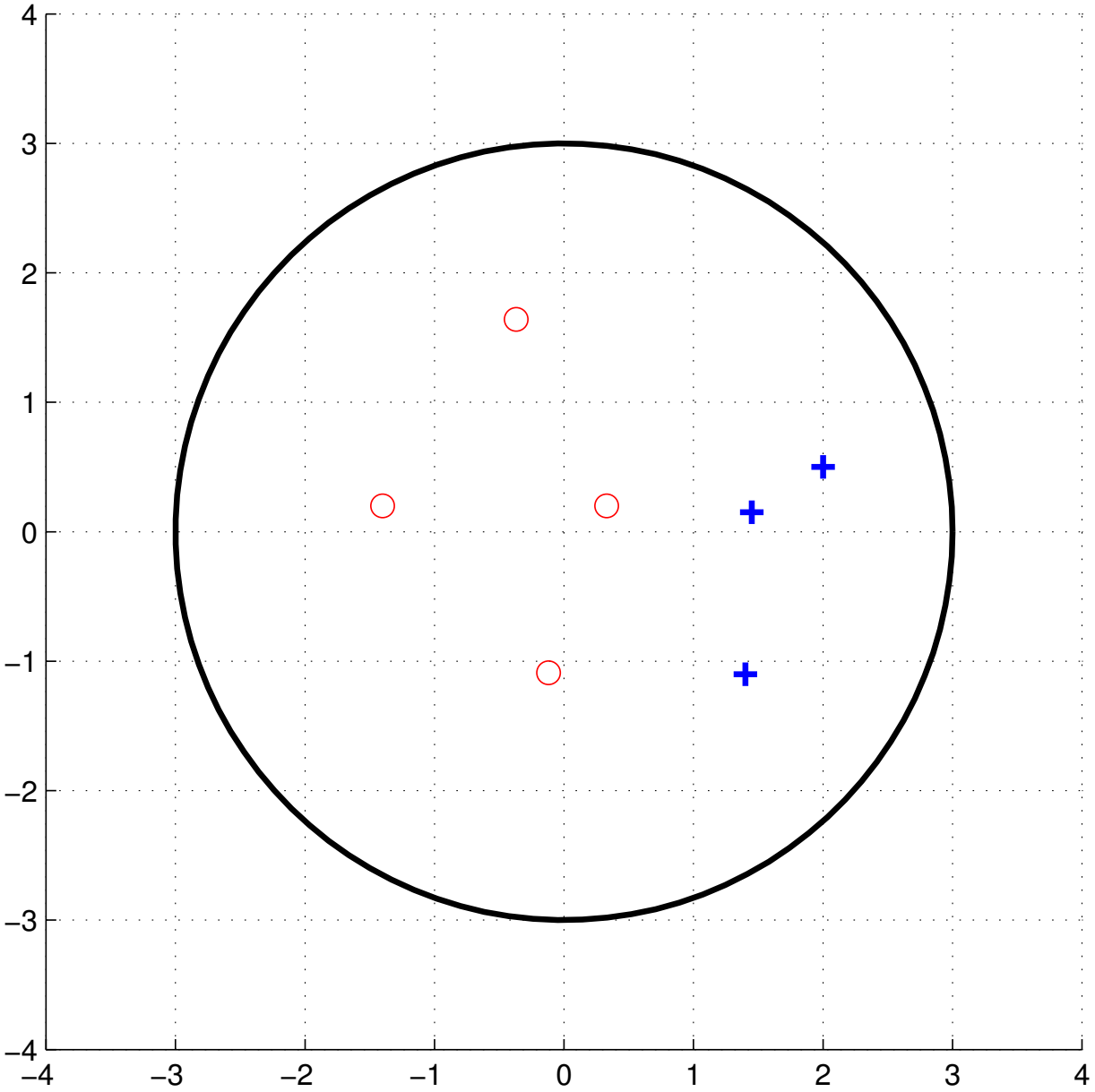
$$w \;=\; r \left[\begin{array}{c} \cos\theta \\ \sin\theta \end{array} \right] ,\; r \;\sim\; \mathrm{Uniform}(0, 3) \;,\; \theta \;\sim\; \mathrm{Uniform}(0, 2\pi)$$

  and u chosen from $\{-1, +1\}$ with equal probability

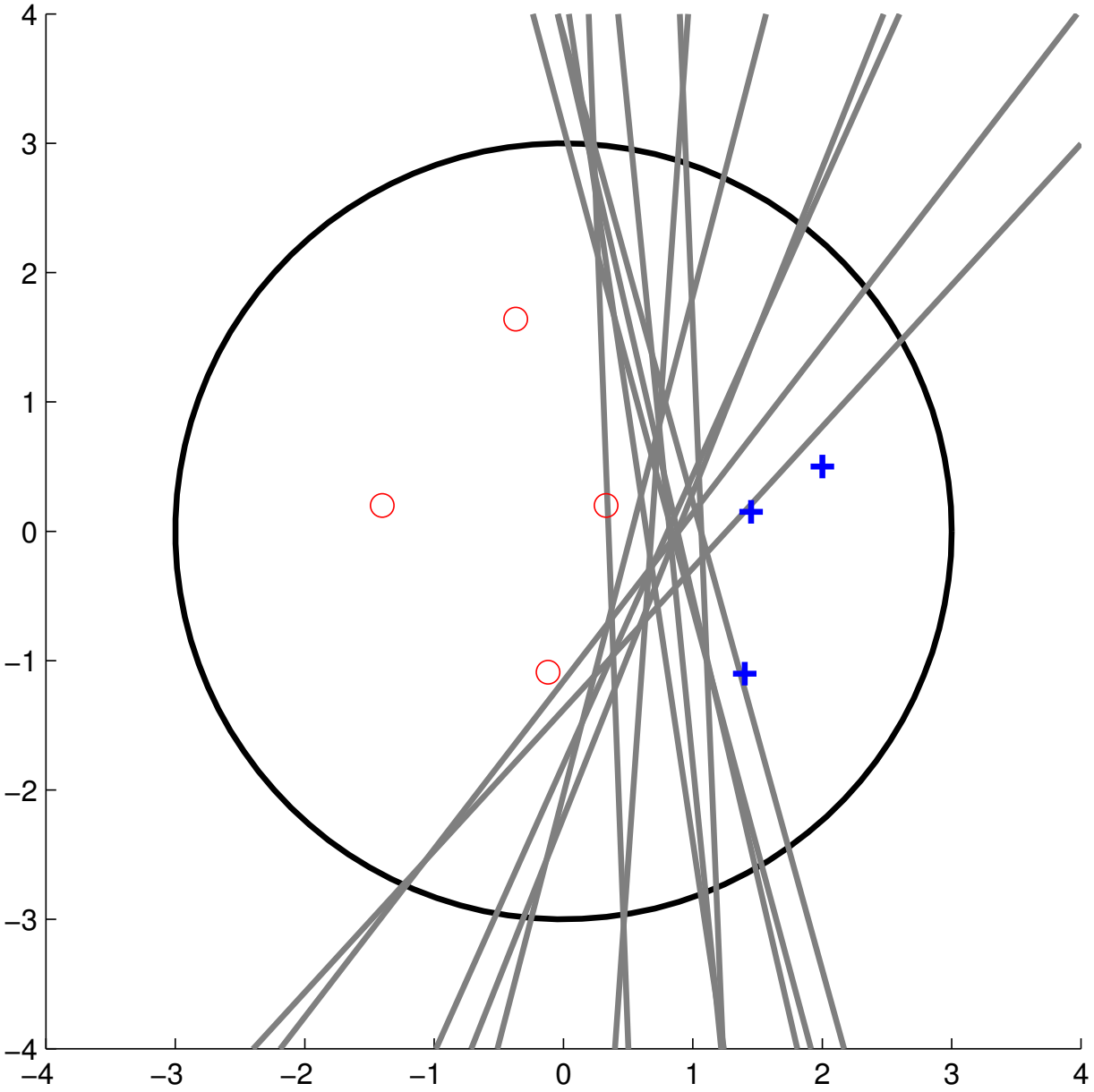
Verifying the Prior



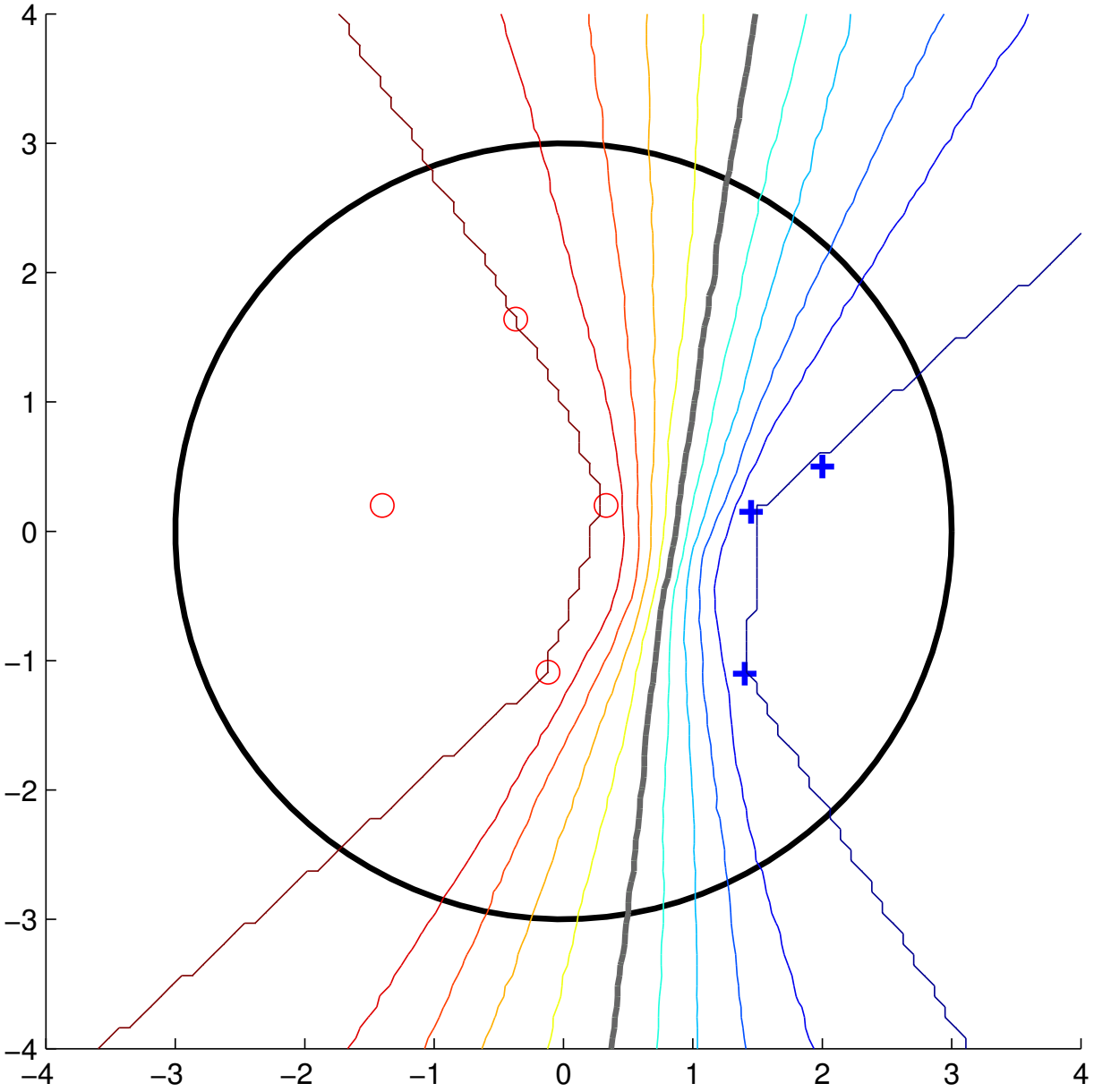
Observing Some Data



Posterior Samples



Bayesian Predictions



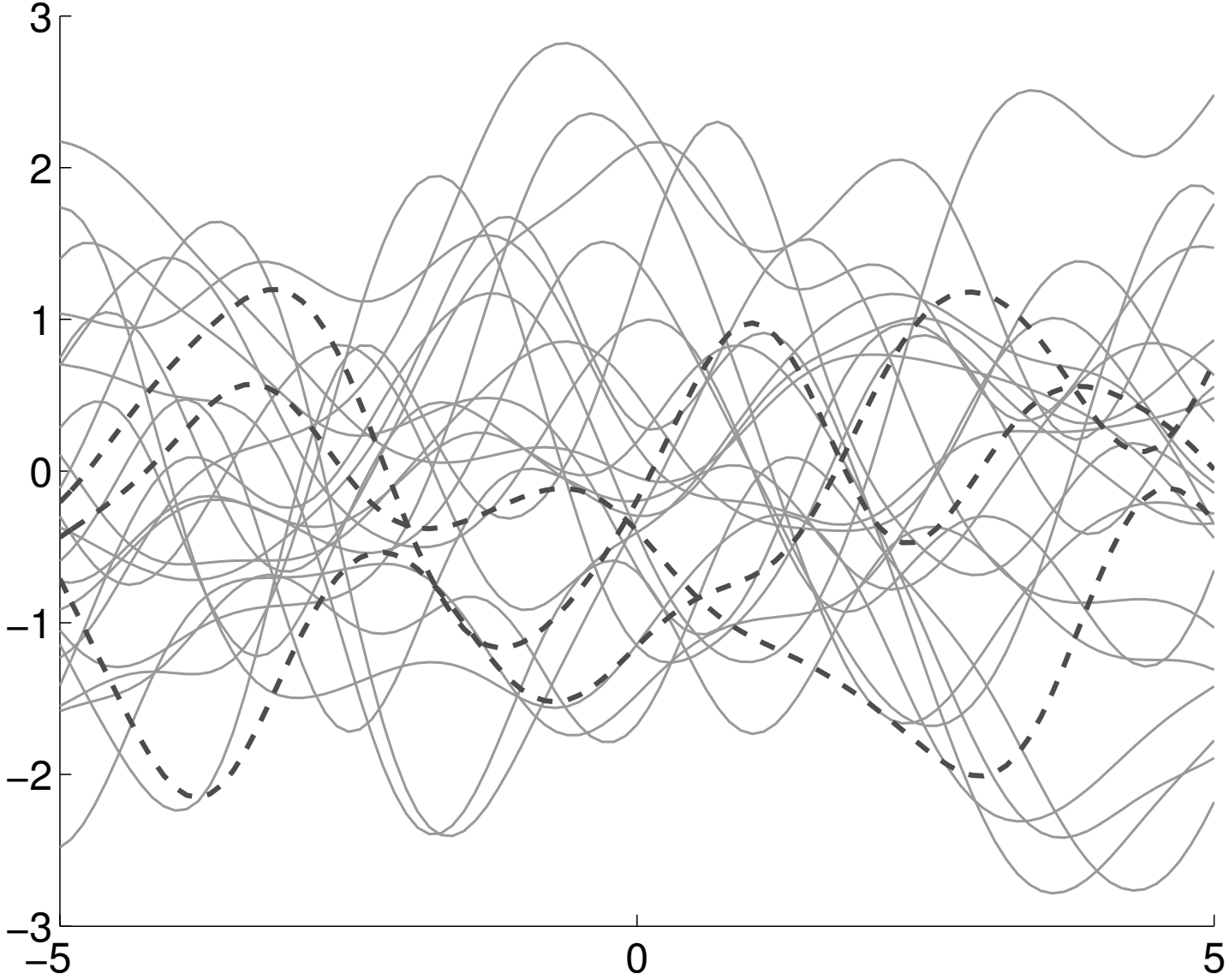
The Evidence or Marginal Likelihood Revisited

$$p(D|M) \;=\; \int p(D|w, M)\, p(w|M) \mathrm{d}w$$

- Probability of the data given M and $p(w|M)$, it's an average likelihood
- Volume of the unnormalized posterior: agreement of prior with likelihood
- Let's compute it for the previous example:
  The likelihood of a given line is 1 if it separates the data and 0 otherwise. We draw 10000 posterior samples, of which only 513 separate the data.
- The evidence is the fraction of the lines drawn from the prior that are compatible with the data: $513/10000 = 0.0513$

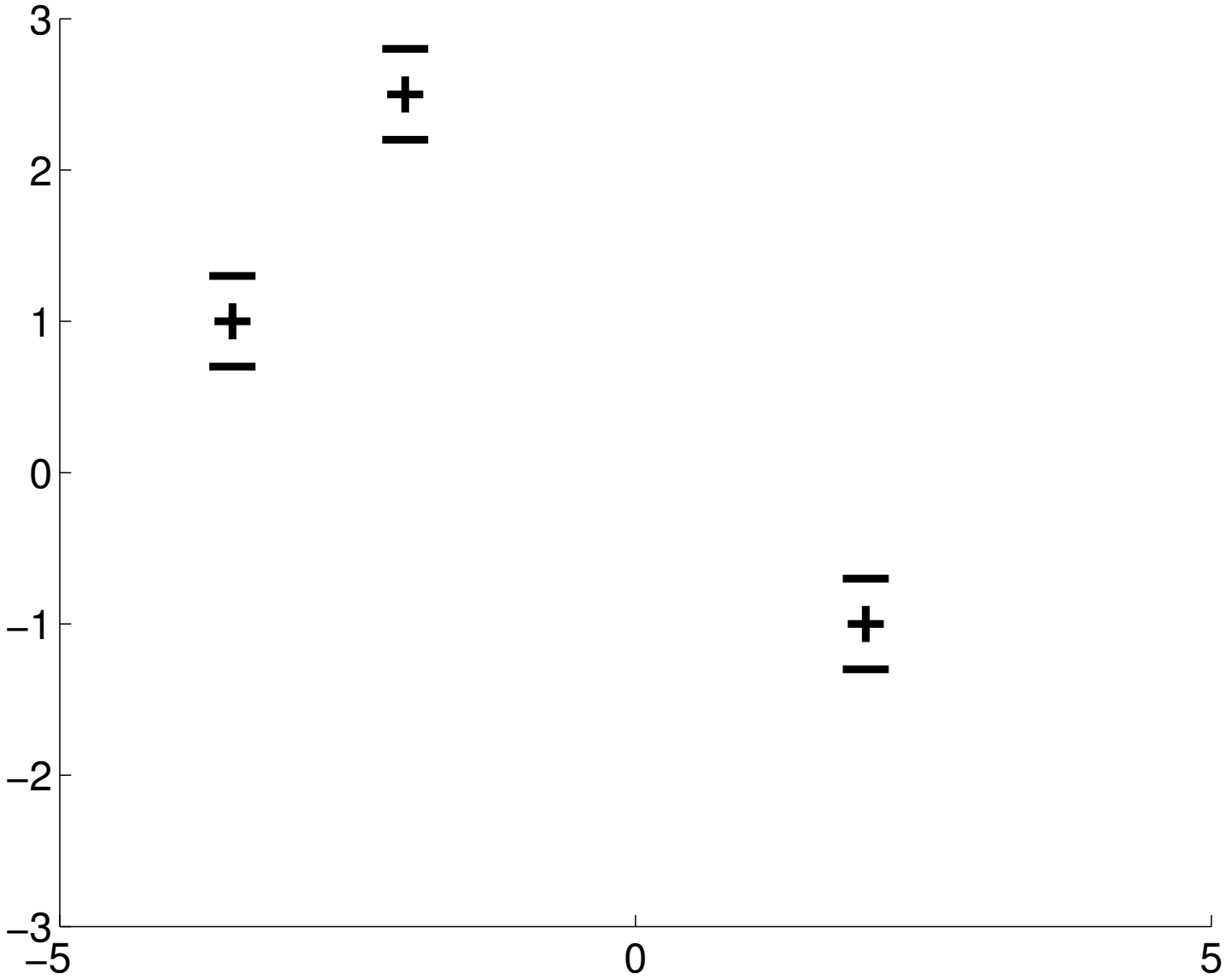
Bayesian Regression

Assume we were able to impose a prior over functions, we could draw some samples



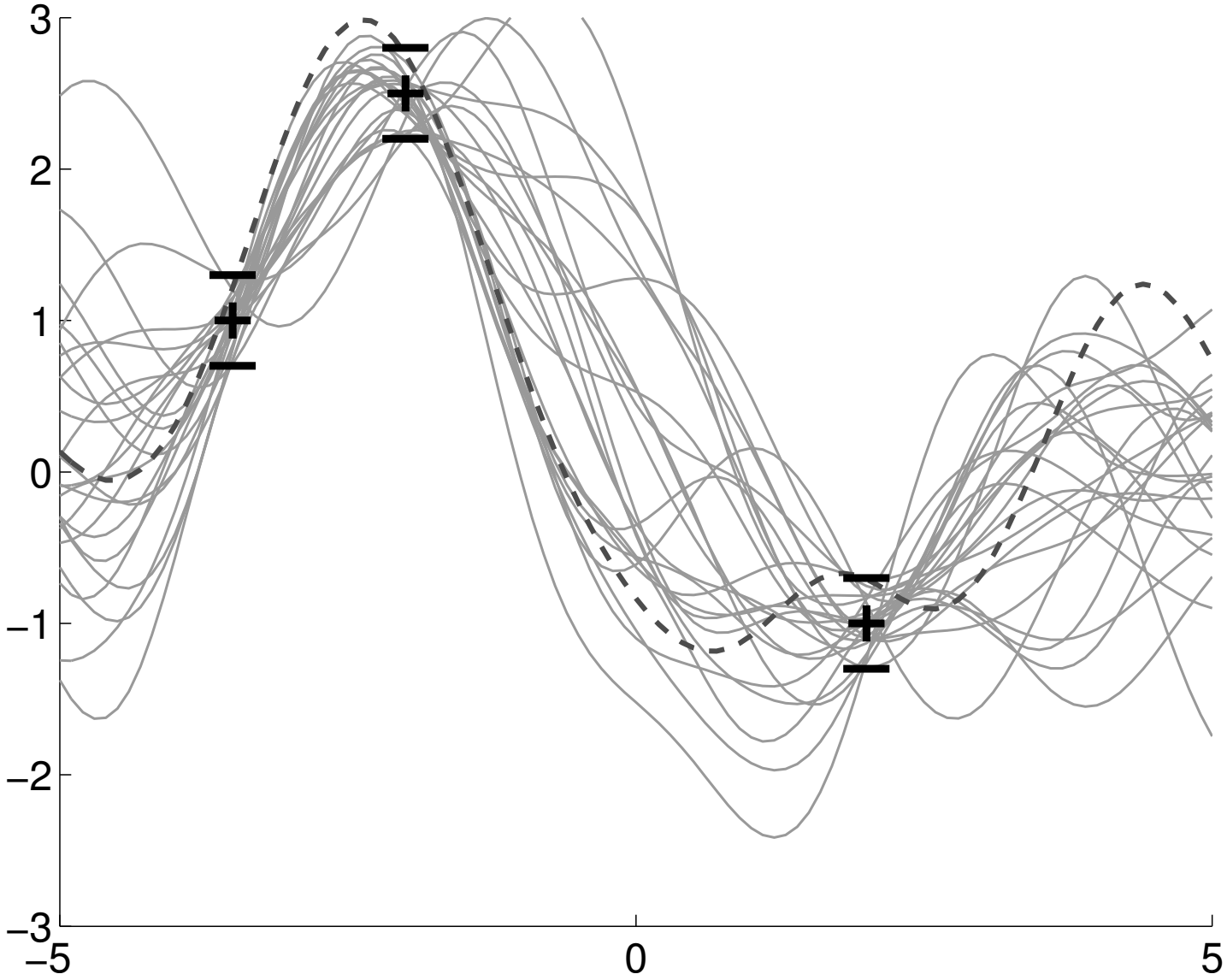
Bayesian Regression - Observe Some Data

Now let us assume we observe three data points, and decide to define take a uniform noise model. This gives us the likelihood of the function values given the observed data:



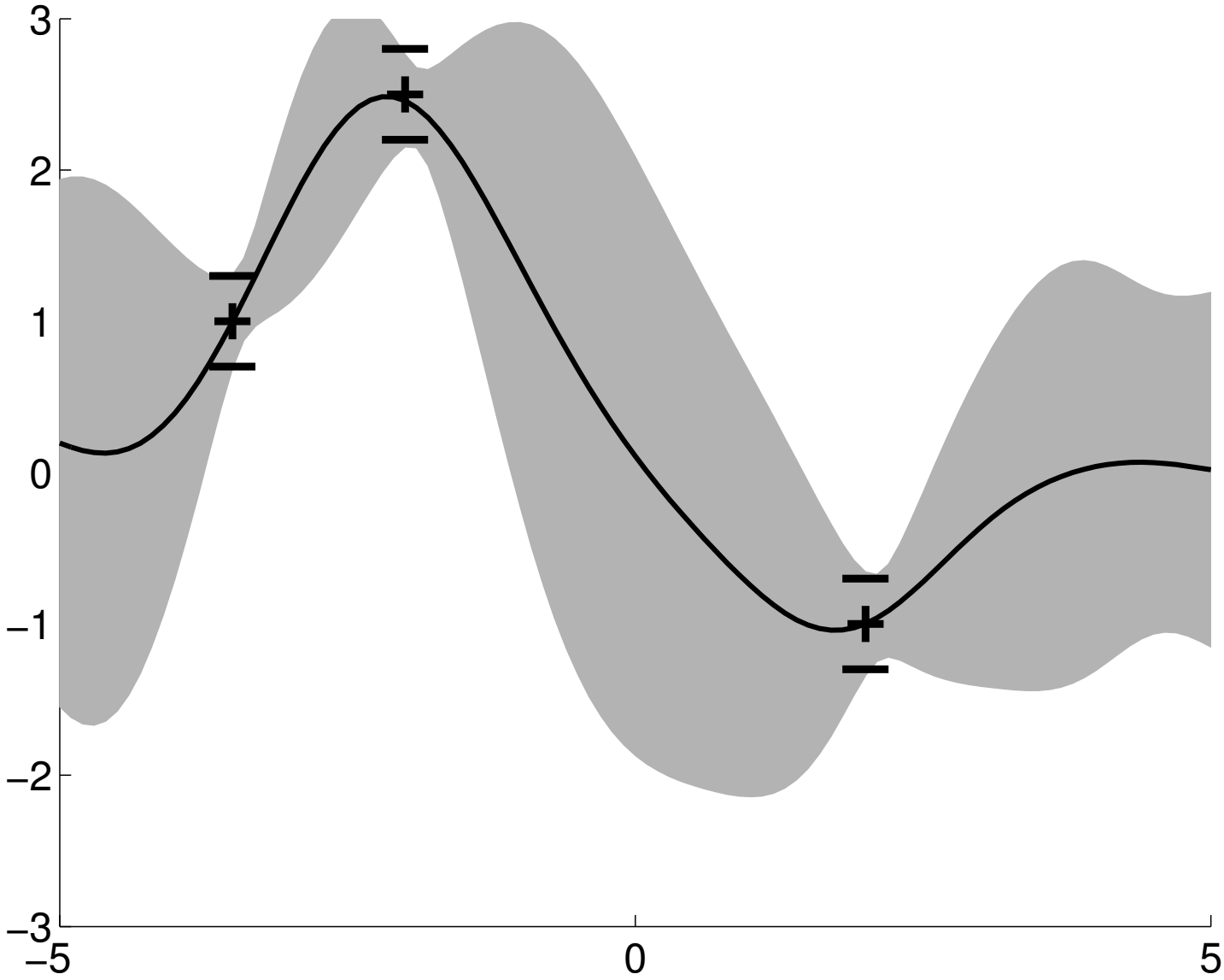
Bayesian Regression - Posterior Distribution

Given the prior and the likelihood we can now draw samples from the posterior distribution:

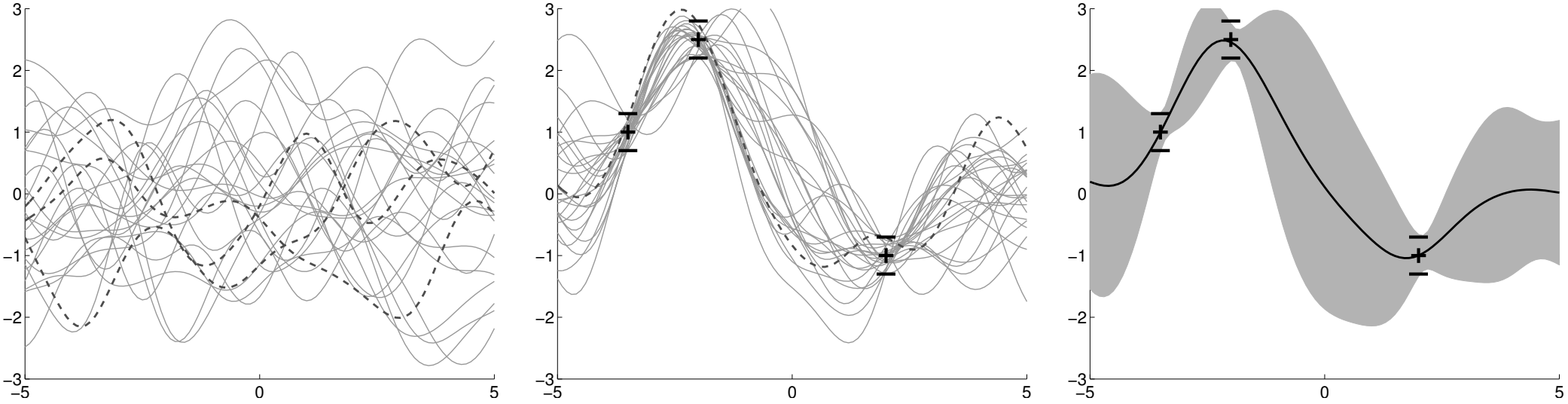


Bayesian Regression - Predictive Distribution

If are now asked to make predictions, we average over the posterior distribution to obtain the predictive distribution:



Bayesian Regression - Summary



left samples from our the prior (could be all of you!)

middle samples from the posterior, data observed (crosses) and uniform noise model (horizontal bars)

right predictive distribution, empirically computed from the posterior samples. Here mean and 2 std dev given

parameters of the prior? Either specify hyperprior on, or learn the parameters of the prior by maximizing the evidence