

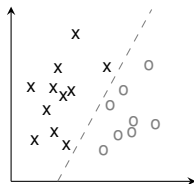
PROBABILISTIC AND UNSUPERVISED LEARNING

PETER ORBANZ • FALL 2026

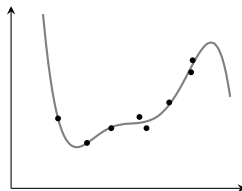
Version: 7 October 2026

SUPERVISED LEARNING

Here are two types of learning problems you may have encountered before:



classification



regression

- These are both **supervised learning** problems.
- In both cases we are fitting a function to data—statistically speaking we are performing regression. The function takes finitely many values (classification) or is continuous-valued (regression).
- The term *supervised* indicates that we have examples of (possibly noisy) solutions, namely input and output values.

THREE LEARNING PROBLEMS

- Predicting new outcomes: generalizing.

- **Supervised learning.** Observe input/output pairs (“teaching”):

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \dots$$

Predict the correct y^* for new input x^* .

- Systematizing (noisy) observations: discovering structure.

- **Unsupervised learning.** Observe (sensory) input alone:

$$x_1, x_2, x_3, x_4, \dots$$

Describe pattern of data $[p(x)]$, identify and extract underlying structural variables $[x_i \rightarrow y_i]$.

- Choosing actions wisely.

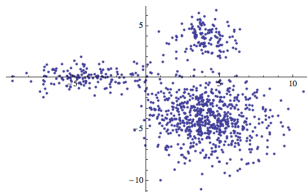
- **Reinforcement learning.** Rewards or payoffs (and possibly also inputs) depend on actions:

$$x_1 : a_1 \rightarrow r_1, x_2 : a_2 \rightarrow r_2, x_3 : a_3 \rightarrow r_3 \dots$$

Find a policy for action choice that maximizes payoff.

SUPERVISED VS UNSUPERVISED

Consider a data source generating points in $\mathbb{R}^2$. Each point belongs to one of three groups. Here is a sample:



We are asked to assign data points to groups. There is a supervised and an unsupervised versions of this problem.

Classification

- Given: The data above, and the group assignment of each point.
- Task: Assign *new* points generated by the same source to their groups.
- Supervised problem.
- Jargon: Groups are called “classes”.
- Solution is typically a function.

Clustering

- Given: The data above only.
- Task: Assign each point *in the data* to a group.
- Unsupervised problem.
- Jargon: Groups are called “clusters”.
- Solution is typically a distribution.

PENDULUM

(WORK OF MARC DEISENROTH AND CARL EDWARD RASMUSSEN)

Task

Balance the pendulum upright by moving the sled left and right.

- The computer can control *only* the motion of the sled.
- Available data: Current state of system (measured 25 times/second).



Formalization

State = 4 variables (sled location, sled velocity, angle, angular velocity)

Actions = sled movements

The system can be described by a function

$$f : \quad \mathcal{S} \times \mathcal{A} \quad \rightarrow \quad \mathcal{S}$$
$$(\text{state}, \text{action}) \mapsto \text{state}$$

PENDULUM



In Part I

- Models and learning approaches
Review of Bayesian learning, maximum likelihood estimation, Gaussian distributions
- Latent variable models
Latent variable models, free energies, the EM algorithm
- Markov and hidden Markov models
A special class of latent variable models for sequence data.
- Sampling algorithms and MCMC
Fitting and evaluating models by randomized simulation can be an alternative to optimization.
- Optimization
Arguably the most widely used way to fit a model to data is by optimization.

In Part II

- (Approximate) inference in graphical models.

BASICS

REPRESENTING A DATA SOURCE

Our approach to learning starts with a model of data production:

$$P(\text{data}|\text{parameters}) \quad P(x|\theta) \text{ or } P(y|x, \theta)$$

Machine learning jargon also calls $P(x|\theta)$ a **generative model**.

In more detail

- We observe data $\mathcal{D} = (x_1, \dots, x_n)$ from a data source, in a **sample space** $\mathcal{X}$.
- Our mathematical description of the source is a probability distribution P on $\mathcal{X}$.
- A **(statistical) model** is a set $\mathcal{M}$ of probability distributions. “I assume model $\mathcal{M}$ ” means “I assume P is contained in $\mathcal{M}$ ”.
- We typically index the elements of $\mathcal{M}$ by the elements of a **parameter space** $\mathcal{T}$. The model then takes the form

$$\mathcal{M} = \{P(\bullet|\theta) \mid \theta \in \mathcal{T}\}$$

- Example: $\mathcal{M} = \{\text{Gaussian distributions with variance 1 and mean } \theta \mid \theta \in \mathbb{R}\}$

Terminology varies, and the term “model” may refer to other mathematical objects.

BASIC RULES OF PROBABILITY

- Probabilities are non-negative $P(x) \geq 0 \forall x$.
- Probabilities normalize: $\sum_{x \in \mathcal{X}} P(x) = 1$ for distributions if x is a discrete variable and $\int_{\mathcal{X}} p(x) dx = 1$ for probability densities over continuous variables
- 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' Formula:** Since $P(x, y) = P(x)P(y|x) = P(y)P(x|y)$, we have

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Warning: We will not be obsessively careful in our use of p and P for probability density and probability distribution. Should be obvious from context.

INDEPENDENCE

Recall that two random variables X_1 with (marginal) distribution P_1 and X_2 with distribution P_2 are **independent** if their joint distribution P factorizes,

$$P(x_1, x_2) = P_1(x_1)P_2(x_2) \quad \text{or equivalently} \quad P(x_1|x_2) = P_1(x_1) .$$

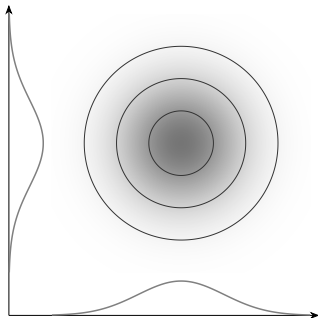
Informally:

Two random variables are independent if knowing the value of one does not provide any information about the outcome of the other.

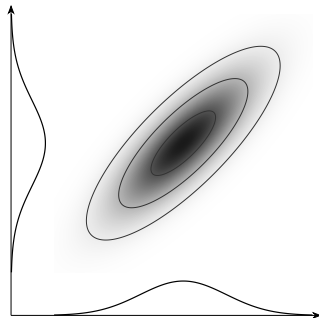
Random variables are called **iid (independent and identically distributed)** if they are independent and have identical marginal distributions: $P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i)$

JOINT VS MARGINAL DISTRIBUTIONS

A joint distribution may contain much more information than its marginals.



2-dimensional Gaussian



another Gaussian with the same marginals

Assuming independence means assuming the marginals carry *all* information about the joint.

EXPONENTIAL FAMILIES

A model $\mathcal{M} = \{p(\bullet | \theta) \mid \theta \in \mathcal{T}\}$ is an **exponential family** if

$$p(x|\theta) = f(x)g(\theta)e^{\phi(\theta)^\top \mathbf{T}(x)} \quad \text{for all } x \in \mathcal{X} \text{ and } \theta \in \mathcal{T}.$$

The components of the model are:

- A function $\mathbf{T} : \mathcal{X} \rightarrow \mathbb{R}^m$, for some dimension $m \in \mathbb{N}$. This function is the **sufficient statistic** of the model.
- A function $\phi : \mathcal{T} \rightarrow \mathbb{R}^m$. Its values $\phi(\theta)$ are called **natural parameters**.
- Functions $f : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ and $g : \mathcal{T} \rightarrow \mathbb{R}_{> 0}$.

Note g is completely determined by $\mathbf{T}, f$ and ϕ .

Relevance

- These are arguably the most common types of distributions in machine learning.
- They often arise as building blocks of more complicated models.
- The class of exponential families includes most elementary probability distributions. As we will see, it is also *very* convenient to work with.

EXAMPLES OF EXPONENTIAL FAMILIES

Model	Sample space	Sufficient statistic
Gaussian	$\mathbb{R}^d$	$\mathbf{T}(\mathbf{x}) = (\mathbf{x}\mathbf{x}^t, \mathbf{x})$
Gamma	$\mathbb{R}_+$	$\mathbf{T}(x) = (\ln(x), x)$
Poisson	$\mathbb{N}_0$	$\mathbf{T}(x) = x$
Categorical	$\{1, \dots, K\}$	$\mathbf{T}(x) = x$
Wishart	Positive definite matrices	(requires more details)
Mallows	Rankings (permutations)	(requires more details)
Beta	$[0, 1]$	$\mathbf{T}(x) = (\ln(x), \ln(1 - x))$
Dirichlet	Probability distributions on d events	$\mathbf{T}(\mathbf{x}) = (\ln x_1, \dots, \ln x_d)$
Bernoulli	$\{0, 1\}$	$\mathbf{T}(x) = x$

A non-example: Student's t -distribution.

Informally

- A given sample space $\mathcal{X}$ often has a “natural” statistic $\mathbf{T}$. (A **statistic** is a function of the data.)
- The exponential family defined by $\mathbf{T}$ is often the “natural” distribution for simple sources generating data in $\mathcal{X}$.

LEARNING PARAMETERS

We distinguish two approaches:

- **Point estimation** of a parameter θ assumes there is a true but unknown value θ_0 . An estimate $\hat{\theta}$ is a function of observed data that approximates θ_0 .
- **Bayesian inference** represents lack of knowledge as randomness; the parameter is a random variable Θ . This means we have to specify the distribution of Θ , called the **prior distribution**. We ask not for a true value of the parameter, but for its distribution $p(\theta|\mathcal{D})$ given the observed data, called the **posterior distribution**.
(We assume that you are familiar with Bayes' equation and basics of Bayesian inference.)

Specific approaches we will discuss:

- **Maximum likelihood estimation** is the most common form of point estimation.
- Bayesian inference may compute a posterior distribution from data (e.g. **conjugate posteriors**), approximate complicated posteriors by simpler distributions (e.g. **variational inference**), or use **sampling algorithms** to generate random draws from the posterior, which are then processed further.
- The **MAP** estimate (**m**aximum **a** posteriori estimate) is the parameter value most probable under the posterior,

$$\hat{\theta}_{\text{MAP}} = \arg \max P(\theta|\mathcal{D}) = \arg \max P(\theta)P(\mathcal{D}|\theta).$$

Although it is defined in terms of a posterior, it is a point estimate.

Modelling assumption

- Model $\mathcal{M} = \{p(\bullet|\theta)|\theta \in \mathcal{T}\}$. Each element $p(\bullet|\theta)$ is the distribution of a *single* random element X of $\mathcal{X}$.
- As prior, we choose a distribution π on $\mathcal{T}$.
- We assume data is generated as follows: Generate

$$\begin{aligned}\Theta &\sim \pi \\ X_1, \dots, X_n | \Theta &\sim_{\text{iid}} p(\bullet | \Theta) .\end{aligned}$$

and explain data $\mathcal{D} = \{x_1, \dots, x_n\}$ as the values assumed by $X_1, \dots, X_n$.

- $X_1, \dots, X_n$ are **conditionally iid**.

The term **Bayesian model** often refers to the pair $(\mathcal{M}, \pi)$.

Posterior

The independence assumption implies $p(\mathcal{D}|\theta) = \prod_{i=1}^n p(x_i|\theta)$. Bayes' rule then takes the form

$$\pi(\theta|\mathcal{D}) = \frac{\prod_{i=1}^n p(x_i|\theta)}{\int_{\mathcal{T}} \prod_{i=1}^n p(x_i|\theta) \pi(\theta) d\theta} \pi(\theta) \propto \prod_{i=1}^n p(x_i|\theta) \pi(\theta)$$

Bayesian inference without the conditional i.i.d. assumption is also possible, but typically more complicated.

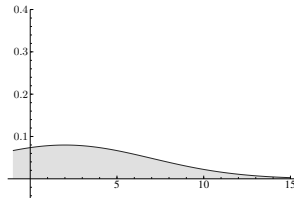
EXAMPLE: UNKNOWN GAUSSIAN MEAN

Model

- Observations are generated by a Gaussian, with unknown mean θ and fixed, known standard deviation σ .
- That means our model is $\mathcal{M} = \{p(\bullet|\theta, \sigma) | \theta \in \mathbb{R}\}$, where p is the Gaussian density on the real line.

Prior

- We choose a Gaussian prior $\pi(\bullet|\mu, \xi)$ with known mean μ and standard deviation ξ .
- In the figure, $\mu = 2$ and $\xi = 5$. Hence, we assume $\mu = 2$ is the most probable value of θ , and that $\theta \in [-3, 7]$ with a probability ~ 0.68 .



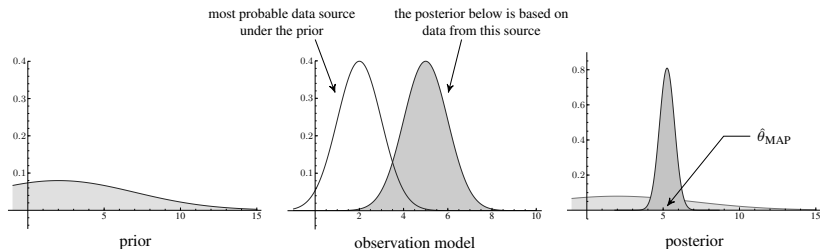
Posterior

Application of Bayes' formula shows the posterior is again a Gaussian,

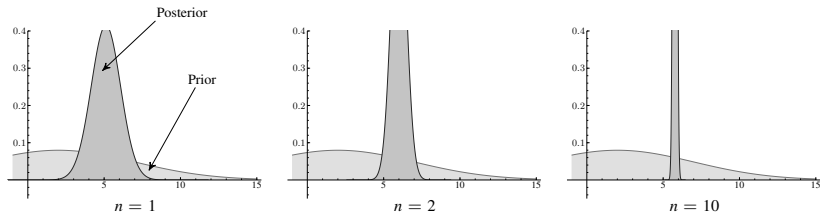
$$\pi(\theta|x_{1:n}, \mu, \xi) = \pi(\theta|\mu_n, \xi_n) \quad \text{where} \quad \mu_n := \frac{\sigma^2\mu + \xi^2 \sum_{i=1}^n x_i}{\sigma^2 + n\xi^2} \quad \text{and} \quad \xi_n^2 := \frac{\sigma^2\xi^2}{\sigma^2 + n\xi^2}$$

EXAMPLE: UNKNOWN GAUSSIAN MEAN

Model



Posterior under increasing sample size

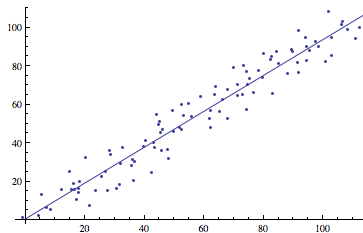


A SLIGHTLY DIFFERENT PERSPECTIVE

Parameters

Intuitively, we can think of θ as the common pattern underlying the data:

$$P(X|\theta) = \text{Probability}(\text{data}|\text{pattern})$$



Inference idea

data = underlying pattern + independent randomness

Broadly speaking, the goal of inference is to extract the underlying pattern from the data. Bayesian statistics models the pattern as a random quantity.

BAYESIAN LEARNING AND MODEL SELECTION

Choosing a model $\mathcal{M}$ from a set $\mathcal{M}_1, \mathcal{M}_2, \dots$ of candidate models is called **model selection**.
Bayesian inference can be extended to model selection.

- Problem specification:

Models: $\mathcal{M}_i = \{P(x|\theta_i, \mathcal{M}_i) | \theta_i \in \mathcal{T}_i\}$.

Prior probability of models: $P(\mathcal{M}_i)$.

Prior probabilities of model parameters: $P(\theta_i | \mathcal{M}_i)$

- Data probability (**likelihood**)

$$P(\mathcal{D} | \theta_i, \mathcal{M}_i) = \prod_{j=1}^n P(x_j | \theta_i, \mathcal{M}_i) \equiv \mathcal{L}(\theta_i)$$

Note we are assuming the data is conditionally i.i.d. given the model and parameter.

- Parameter learning (**posterior**):

$$P(\theta_i | \mathcal{D}, \mathcal{M}_i) = \frac{P(\mathcal{D} | \theta_i, \mathcal{M}_i) P(\theta_i | \mathcal{M}_i)}{P(\mathcal{D} | \mathcal{M}_i)}; \quad P(\mathcal{D} | \mathcal{M}_i) = \int d\theta_i P(\mathcal{D} | \theta_i, \mathcal{M}_i) P(\theta_i | \mathcal{M}_i)$$

$P(\mathcal{D} | \mathcal{M}_i)$ is called the **marginal likelihood** or **evidence** for $\mathcal{M}_i$. It is proportional to the posterior probability of model $\mathcal{M}_i$ being the one that generated the data.

- Model selection:

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

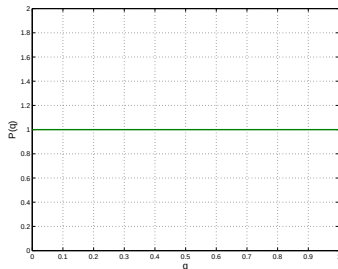
BAYESIAN LEARNING: A COIN TOSS EXAMPLE

Coin toss: One parameter q — the probability of obtaining *heads*.

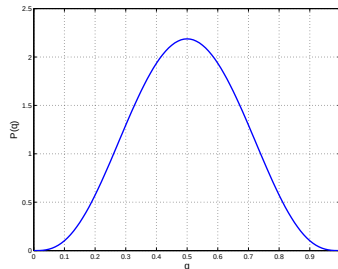
So our space of models is the set of distributions over $q \in [0, 1]$.

Learner A believes (Bayesian) model $\mathcal{M}_A$: all values of q are equally plausible.

Learner B believes (Bayesian) model $\mathcal{M}_B$: “fair” coin ($q \approx 0.5$) more plausible than “biased”.



A: $\alpha_1 = \alpha_2 = 1.0$



B: $\alpha_1 = \alpha_2 = 4.0$

Both **prior beliefs** can be described by the Beta distribution:

$$p(q|\alpha_1, \alpha_2) = \frac{q^{(\alpha_1-1)}(1-q)^{(\alpha_2-1)}}{B(\alpha_1, \alpha_2)} = \text{Beta}(q|\alpha_1, \alpha_2)$$

BAYESIAN LEARNING: THE COIN TOSS (CONT)

Now we observe a toss. Two possible outcomes:

$$p(H|q) = q \qquad p(T|q) = 1 - q$$

Suppose our single coin toss comes out *heads*.

The probability of the observed data (likelihood) is:

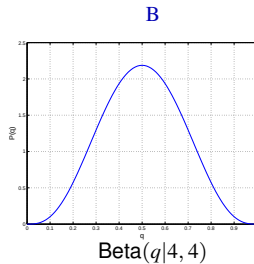
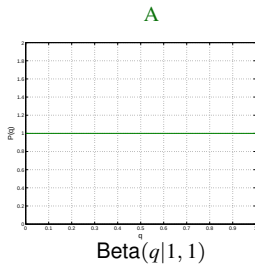
$$p(H|q) = q$$

Using **Bayes' formula**, we multiply the prior, $p(q)$ by the likelihood and renormalize to get the posterior probability:

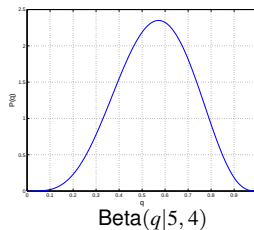
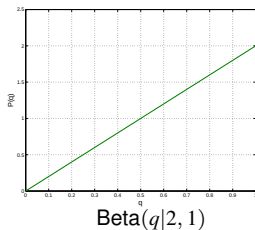
$$\begin{aligned} p(q|H) &= \frac{p(q)p(H|q)}{p(H)} \propto q \text{Beta}(q|\alpha_1, \alpha_2) \\ &\propto q q^{(\alpha_1-1)} (1-q)^{(\alpha_2-1)} \propto \text{Beta}(q|\alpha_1 + 1, \alpha_2) \end{aligned}$$

BAYESIAN LEARNING: THE COIN TOSS (CONT)

Prior



Posterior



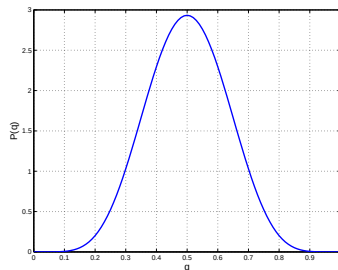
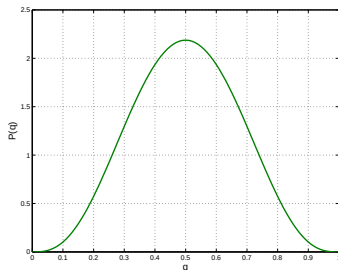
BAYESIAN LEARNING: THE COIN TOSS (CONT)

What about multiple tosses? Suppose we observe $\mathcal{D} = \{\text{H H T H T T}\}$:

$$p(\{\text{H H T H T T}\}|q) = qq(1-q)q(1-q)(1-q) = q^3(1-q)^3$$

This is still straightforward:

$$\begin{aligned} p(q|\mathcal{D}) &= \frac{p(q)p(\mathcal{D}|q)}{p(\mathcal{D})} \propto q^3(1-q)^3 \text{Beta}(q|\alpha_1, \alpha_2) \\ &\propto \text{Beta}(q|\alpha_1 + 3, \alpha_2 + 3) \end{aligned}$$



CONJUGATE PRIORS

Updating the prior to form the posterior was particularly easy in these examples. This is because we used a **conjugate prior** for an exponential family likelihood.

Recall that exponential family distributions take the form:

$$P(x|\theta) = g(\theta)f(x)e^{\phi(\theta)^T \mathbf{T}(x)}$$

with $g(\theta)$ the normalizing constant. Given n conditionally iid observations,

$$P(\{x_i\}|\theta) = \prod_i P(x_i|\theta) = g(\theta)^n e^{\phi(\theta)^T \left(\sum_i \mathbf{T}(x_i) \right)} \prod_i f(x_i)$$

A **conjugate prior** for an exponentially family is of the form

$$P(\theta) = F(\tau, \nu) g(\theta)^\nu e^{\phi(\theta)^T \tau},$$

with normalizer $F(\tau, \nu)$. The posterior is then

$$P(\theta|\{x_i\}) \propto P(\{x_i\}|\theta)P(\theta) \propto g(\theta)^{\nu+n} e^{\phi(\theta)^T \left(\tau + \sum_i \mathbf{T}(x_i) \right)}$$

with the normalizer given by $F(\tau + \sum_i \mathbf{T}(x_i), \nu + n)$.

CONJUGATE PRIORS

The posterior given an exponential family likelihood and conjugate prior is:

$$P(\theta|\{x_i\}) = F\left(\tau + \sum_i \mathbf{T}(x_i), \nu + n\right) g(\theta)^{\nu+n} \exp\left[\phi(\theta)^\top \left(\tau + \sum_i \mathbf{T}(x_i)\right)\right]$$

Here,

- $\tau \in \mathbb{R}^m$ controls the expected value of the prior
(more precisely: τ/ν is the prior mean if the prior is placed on the natural parameter)
- $\nu \in \mathbb{R}_{>0}$ is the **concentration** or **scale** of the prior
(roughly: larger scale $\rightarrow$ smaller variance).

As new data come in, each one increments the sufficient statistics vector and the scale to define the posterior.

Interpretation

- The posterior effectively interpolates between the prior assumption τ and the observational evidence $\sum_i \mathbf{T}(x_i)$.
- The concentration ν specifies how much weight we assign to the prior belief τ . (Large $\nu \rightarrow$ strong prior assumption.)
- Some authors interpret τ as $\tau = \sum \mathbf{T}(x'_i)$ for “fictitious observations” or “pseudo-observations” $x'_1, \dots, x'_\nu$. (If ν is an integer, which it need not be.)

CONJUGACY IN THE COIN FLIP

Distributions are not always written in their natural exponential form.

The Bernoulli distribution (a single coin flip) with parameter q and observation $x \in \{0, 1\}$, can be written:

$$\begin{aligned}P(x|q) &= q^x(1-q)^{(1-x)} \\&= e^{x \log q + (1-x) \log(1-q)} \\&= e^{\log(1-q) + x \log(q/(1-q))} = (1-q)e^{\log(q/(1-q))x}\end{aligned}$$

So the natural parameter is the **log odds** $\log(q/(1-q))$, and the sufficient stats (for multiple tosses) is the number of heads. The conjugate prior is

$$\begin{aligned}P(q) &= F(\tau, \nu) (1-q)^\nu e^{\log(q/(1-q))\tau} \\&= F(\tau, \nu) (1-q)^\nu e^{\tau \log q - \tau \log(1-q)} = F(\tau, \nu) (1-q)^{\nu-\tau} q^\tau\end{aligned}$$

which has the form of the **Beta** distribution $\Rightarrow F(\tau, \nu) = 1/B(\tau+1, \nu-\tau+1)$.

In general, then, the posterior $P(q|\{x_i\})$ is a **Beta**(α_1, α_2), with

$$\alpha_1 = 1 + \tau + \sum_i x_i \qquad \alpha_2 = 1 + (\nu + n) - \left(\tau + \sum_i x_i\right)$$

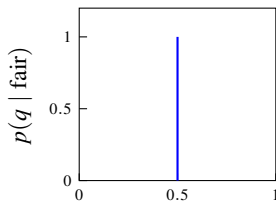
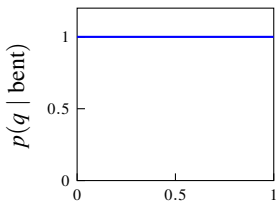
If we observe a head, we add 1 to the sufficient statistic $\sum x_i$, and also 1 to the count n . This increments α_1 . If we observe a tail we add 1 to n , but not to $\sum x_i$, incrementing α_2 .

BAYESIAN COINS – COMPARING MODELS

We have seen how to update posteriors within each model. To study the choice of model, consider two more extreme models: “fair” and “bent”. A priori, we may think that “fair” is more probable, eg:

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

For the bent coin, we assume all parameter values are equally likely, whilst the fair coin has a fixed probability:



We make 10 tosses, and get: $\mathcal{D} = (\text{T H T H T T T T T T})$.

BAYESIAN COINS – COMPARING MODELS

Which model should we prefer *a posteriori* (i.e. after seeing the data)?

The **evidence** for the fair model is:

$$P(\mathcal{D}|\text{fair}) = (1/2)^{10} \approx 0.001$$

and for the bent model is:

$$P(\mathcal{D}|\text{bent}) = \int dq P(\mathcal{D}|q, \text{bent})p(q|\text{bent}) = \int dq q^2(1-q)^8 = B(3, 9) \approx 0.002$$

Thus, the posterior for the models, by Bayes' formula:

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

ie, a two-thirds probability that the coin is fair.

How do we make predictions? Could choose the fair model (model selection).

Or could weight the predictions from each model by their probability (model averaging).

Probability of H at next toss is:

$$P(H|\mathcal{D}) = P(H|\mathcal{D}, \text{fair})P(\text{fair}|\mathcal{D}) + P(H|\mathcal{D}, \text{bent})P(\text{bent}|\mathcal{D}) = \frac{1}{2} \times \frac{2}{3} + \frac{3}{12} \times \frac{1}{3} \approx \frac{5}{12}.$$

MAXIMUM LIKELIHOOD ESTIMATION

Problem specification

- Data $\mathcal{D} = \{x_1, \dots, x_n\}$ is generated iid from model $\mathcal{M} = \{p(x|\theta) \mid \theta \in \mathcal{T}\}$.
We assume iid rather than conditionally iid since the parameter is not assumed to be random.
- Objective: Find the distribution in $\mathcal{M}$ that best explains the data.
That means we have to identify a "best" parameter value $\hat{\theta}$.

Maximum Likelihood approach

The **maximum likelihood estimator (MLE)** is defined as

$$\hat{\theta}_{\text{ML}} := \arg \max_{\theta \in \mathcal{T}} p(x_1, \dots, x_n | \theta) .$$

We hence assume the data is best explained by that distribution in $\mathcal{M}$ under which it is most likely to occur (has the highest probability or density value).

If differentiability holds, $\hat{\theta}_{\text{ML}}$ is the solution of the **maximum likelihood equation**

$$\nabla_{\theta} p(\mathcal{D} | \theta) = \nabla_{\theta} \left(\prod_{i=1}^n p(x_i | \theta) \right) = 0$$

LOGARITHM TRICK

Instead of the likelihood $p(\mathcal{D}|\theta)$, we often work with the **log-likelihood** $\log p(\mathcal{D}|\theta)$.

Recall two properties of the logarithm function:

- It turns products into sums: $\log(\prod_i f_i) = \sum_i \log(f_i)$
- Since it is monotonically increasing on $\mathbb{R}_+$, it does not change the *location* of maxima and minima:

$$\max_y \log(g(y)) \neq \max_y g(y) \quad \text{The *value* changes.}$$

$$\arg \max_y \log(g(y)) = \arg \max_y g(y) \quad \text{The *location* does not change.}$$

MLE with the log-likelihood: Since

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \prod_{i=1}^n p(x_i|\theta) = \arg \max_{\theta} \log\left(\prod_{i=1}^n p(x_i|\theta)\right) = \arg \max_{\theta} \sum_{i=1}^n \log p(x_i|\theta) ,$$

the MLE $\hat{\theta}_{\text{ML}}$ is the solution of

$$\sum_{i=1}^n \nabla_{\theta} \log p(x_i|\theta) = 0$$

BACKGROUND: LAW OF LARGE NUMBERS

For iid random variables $X_1, X_2, \dots$ and any function f with $\mathbb{E}[|f(X_1)|] < \infty$,

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \xrightarrow{n \rightarrow \infty} \mathbb{E}[f(X_1)] \quad \text{with probability 1.}$$

In short: We know how to estimate expectations.

Consequences for parameter estimation

Idea: Reduce parameter estimation to the problem of estimating an expectation.

- Suppose θ_0 is the parameter value that has generated the data.
- Find some function g such that

$$\mathbb{E}_{P_{\theta_0}}[g(X, \theta)] = 0 \quad \text{if and only if } \theta = \theta_0$$

- By the law of large numbers, we can then estimate θ_0 as the solution of

$$\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) = 0$$

MLE DERIVED FROM THE LAW OF LARGE NUMBERS

- It can be shown (under some conditions on the model) that the function

$$g(x, \theta) := \nabla_{\theta} \log p(x|\theta)$$

satisfies the property above. That is,

$$\mathbb{E}_{P_{\theta_0}} [\nabla_{\theta} \log p(X|\theta)] = 0 \quad \text{if and only if } \theta = \theta_0$$

- The function $\theta \mapsto \nabla_{\theta} \log p(x|\theta)$ is called the **score function**, or the **Fisher score**.
- Substituting the law of large numbers estimate for the expectation gives

$$\frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \log p(X_i|\theta) = 0 .$$

The solution of the equation is exactly the maximum likelihood estimator.