# Probability theory and Bayesian statistics

## Machine learning vs Statistics

[If the statistics field had] incorporated computing methodology from its inception as a fundamental tool, as opposed to simply a convenient way to apply our existing tools, many of the other data related fields [such as ML] would not have needed to exist — they would have been part of statistics.

– Jerome Friedman (Stanford)

## Goals

- Review probability theory and random variables (with minimal measure theory).
- Probability theory as an extension of logic.
- Introduce Bayesian statistics.
- Exponential family of distributions and conjugacy.

## Probability theory review

- A probability triple: $(\Omega, \mathcal{F}, P)$
- $\Omega$ is a set of outcomes
- $\mathcal{F}$ is a set of events
- $P : \mathcal{F} \to [0, 1]$ is a function that assigns probability to events (a "measure").

A measure assigns a size or weight to a set, a probability measure is a special type of measure that assigns a value in the range $[0, 1]$.

## Random variable: mapping outcome to values

A random variable $X : \Omega \to \mathcal{X}$ is a function that maps an event $\omega \in \Omega$ to $x \in \mathcal{X}$.

- $\mathcal{X}$ denotes the support set (set of possible values) for $X$.

### Random variable: real and discrete-valued

In most real problems, we are concerned with two broad class of random variables:

- $X$ is a real-valued random variable if $\mathcal{X} \subseteq \mathbb{R}$.
- $X$ is discrete if $\mathcal{X}$ is countable.

Note: Discrete RVs include those defined on countable non-numerical sets, such as categorical variables (e.g., heads/tails)

### Random variable: discrete-valued

- In phylogenetics, the inferential goal is to learn the tree structure to explain the ancestral relationship based on the observed traits of taxa: tree structured random variable.
- Similarly, networks or graph structures can be seen as random variables that do not take on real values.
- But in many instances, it may be convenient to encode non-numerical discrete variables using real-values. For example, head/tail in a coin flip is not real-valued but we may encode head as 0 and tail as 1.

### Random variable: inducing probability measure

$X$ is said to induce a probability measure on $\mathcal{X}$ via $\nu = P \circ X^{-1}$. For a measurable set $A \subset \mathbb{R}$,

$$\nu(A) = P(X \in A) = P(X^{-1}(A)).$$

### Random variable: CDF

Distribution function of a real-valued RV $X$ is defined as $F(x) = P(X \leq x)$, also known as cumulative distribution function (CDF).

### Random variable: density

If $F$ can be expressed as

$$F(x) = \int_{-\infty}^{x} p(y)dy, \tag{1}$$

then we say $X$ has a density function $p$. Necessarily $p(x) \geq 0$ and $\int p(x)dx = 1$.

Note: replace $\int$ with $\sum$ for discrete valued $x \in \mathbb{R}$ but with probability mass function rather than probability density function.

### Random variable: vector and matrix-valued

- The definition for real-valued random variable can be suitably extended to define such things as random vector, $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$ or even a random matrix, $\mathcal{X} \subseteq \mathbb{R}^{d_1 \times d_2}$ for $d_1, d_2 \in \mathbb{N}$.
- Any measurable function $g$ of RVs is a RV: e.g., sum of RVs, transformations like $\log, \exp$.

### Random variable: summary

- RVs provide a way to map canonical events to a more convenient and structured representation $\mathcal{X}$ where computation and analyses are more straight forward.
- Real-valued random variables are particularly natural for describing distributions and quantifying uncertainty, as they align well with our intuition and mathematical tools.
- Much of the mathematics we use is centered around the real line and discrete structures, offering well-developed tools such as derivatives, integrals, sums, and products for analyzing and manipulating functions and data.

### Expectation: definition

The expectation of a RV is defined as,

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) dP(\omega).$$

The above notation emphasizes the underlying probability space and the base measure, $P$.

Here $dP(\omega)$ denotes the contribution of each outcome to the expected value.

### Expectation: definition on RV

For practical purposes, we want to express the expectation over $\mathcal{X}$ where we can do some calculation.

For $\nu = P \circ X^{-1}$ and measurable function $g$,

$$\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(x) d\nu(x).$$

Again, $d\nu(x)$ denotes contribution of $x$ on the integral measured by $\nu$.

### Expectation: RV with density function

If a real-valued RV has density $p$, we can write $\nu(dx) = p(x)dx$ so,

$$\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(x)d\nu(x) = \int_{\mathcal{X}} g(x)p(x)dx.$$

- This is the form of integral that we will deal with for the continuous RVs considered.
- The definition stems from the notion of absolute continous wrt Lebesgue measure and Radon-Nykodym theorem – we won't get into this level of detail in this course.

### Expectation: discrete-valued RV

If $\nu$ has the counting measure $\lambda$ as reference measure, the integral can be expressed using summation:

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x).$$

- Counting measure: $\lambda(\{x\}) = 1$ (counts the number of elements in a set).
- $\nu(dx) = p(x)d\lambda(x) = p(x)$. Here "infinitesimal" is just one single element, which by the counting measure has measure of 1.

### Expectation: indicator function

An indicator function is a special type of a random variable (very useful for various purposes):

$$1_A(x) = \begin{cases} 0 & \text{if } x \notin A \\ 1 & \text{if } x \in A. \end{cases} \tag{2}$$

The expectation of an indicator function is defined as,

$$\mathbb{E}[1_A(X)] = P(X \in A).$$

## Probability theory to generative models

Our goal is to represent the data by a set of random variables, involving one or more hidden (latent) variables.

- Denote a set of observed RVs by $X_O$ and hidden variables by $X_H$.
- Use $\theta$ to denote the model parameters but in general notation, we can assume that $\theta$ is part of $X_H$.

$$p_\theta(X_O, X_H) = p_\theta(X_O|X_H)p_\theta(X_H).$$

## Uncertainty: incomplete knowledge vs randomness

- Model uncertainty: ignorance of the underlying hidden causes or mechanism generating data (epistemic uncertainty).
- Data uncertainty: this is the inherent uncertainty/randomness in the outcome (aleatoric uncertainty).

## Uncertainty: Epistemic

- Predicting tomorrow's weather based on incomplete data.
- Suppose you're trying to predict the outcome of a soccer match between Team A and Team B. You want to model these probabilities, but the process introduces epistemic uncertainty due to incomplete knowledge about factors influencing the match.

## Uncertainty: Aleatoric

- Coin flip. There is inherent uncertainty in the outcome, which we model using a probability distribution, in this case, Bernoulli.
- The result of rolling a die.

### Probability as an extension of logic: binary logic

Boolean logic: an event $\omega$ can only be TRUE (1) or FALSE (0).

- e.g., "it will rain tomorrow" or "the parameter $\theta \in [0.5, 1.5]$", can only be either TRUE or FALSE under boolean logic.
- These statements can be combined using AND, OR, NOT, but they will all evaluate to either TRUE or FALSE.

Boolean logic does not deal with data uncertainty as $P(\omega)$ is either 0 (FALSE) or 1 (TRUE).

### Probability as an extension of logic: probability represents beliefs

Probability theory is a way to express a belief about an event that allows values in between 0 and 1, accompanied by a set of probability calculus that let's you formulate complex beliefs on top of simpler ones.

Probability extends logic to situations where truth is uncertain, as in Bayesian inference.

### Probability as an extension of logic: AND

For events $A, B$,

$$P(A \text{ and } B)$$

is written as

$$P(A, B).$$

If $A$ and $B$ are independent events,

$$P(A, B) = P(A)P(B)$$

### Probability as an extension of logic: OR

The set of event $A$ OR $B$ can be expressed as,

$$P(A \text{ OR } B) = P(A) + P(B) - P(A, B)$$

## Probability as an extension of logic: conditioning

The conditional probability of an event:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

## Probability as an extension of logic: conditional independence

The events $A, B$ are conditionally independent given $C$ if,

$$P(A,B|C) = P(A|C)P(B|C),$$

we write $A \perp B|C$.

## Probability as an extension of logic: RVs

For RVs $X, Y, Z$,

- $p(X,Y)$ denotes the joint distribution of $X$ and $Y$
- $p(X|Y)$ denotes the conditional distribution
- Independence: $p(X,Y) = p(X)p(Y)$
- $X \perp Y|Z$: $p(X,Y|Z) = p(X|Z)p(Y|Z)$
- Marginalization: $p(X) = \int_y p(X,Y)dy$

## Probability as an extension of logic: chain rule

- Product rule: $p(X,Y) = p(X)p(Y|X)$ or $p(Y)p(X|Y)$.
- Chain rule:

$$p(X_{1:N}) = p(X_1)\prod_{n=2}^{N} p(X_n|X_{1:n-1})$$

## Probability as an extension of logic: Bayes theorem

Deriving expression for conditional distribution using Chain rule:

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

so,

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}.$$

This is Bayes' Theorem.

## Statistics: solving the inverse problem

- Probability theory is concerned with generative model or the forward model.
- Statistics is concerned with the inverse problem: infer the state of the world, i.e., the unknown parameters and the latent variables, that generated the observed data.

## Statistics: frequentist vs Bayesian

- Frequentist: treats data as random and parameters as fixed. The inference over the hidden variables utilizes the sampling distribution of the estimator due to randomness in the data. E.g., asymptotic distribution of $\widehat{\theta}_{MLE}$.
- Bayesian: treats data as fixed (observed), and conditioned on the data, formulate a probability distribution over the hidden variables (parameters) given the data: $\theta|y$. It naturally allows to incorporate prior knowledge.

## Components of Bayesian statistics

- Prior distribution over the hidden variables.
- Data likelihood specified via conditional probability distributions.
- Compute posterior distribution.

Prior specifies initial belief on the possible values of the hidden variables, expressed via a distribution.

The posterior serves to update the belief after having seen the data.

## Bayes theorem

$$p(x_H|x_O) = \frac{p(x_O, x_H)}{p(x_O)} \tag{3}$$

$$= \frac{p(x_O|x_H)p(x_H)}{p(x_O)}. \tag{4}$$

- $p(x_H|x_O)$: posterior
- $p(x_H)$: prior
- $p(x_O)$: marginal likelihood of the observed data.
- $p(x_H, x_O)$: joint likelihood.
- $p(x_O|x_H)$: observed data likelihood.

## Example: coin flip

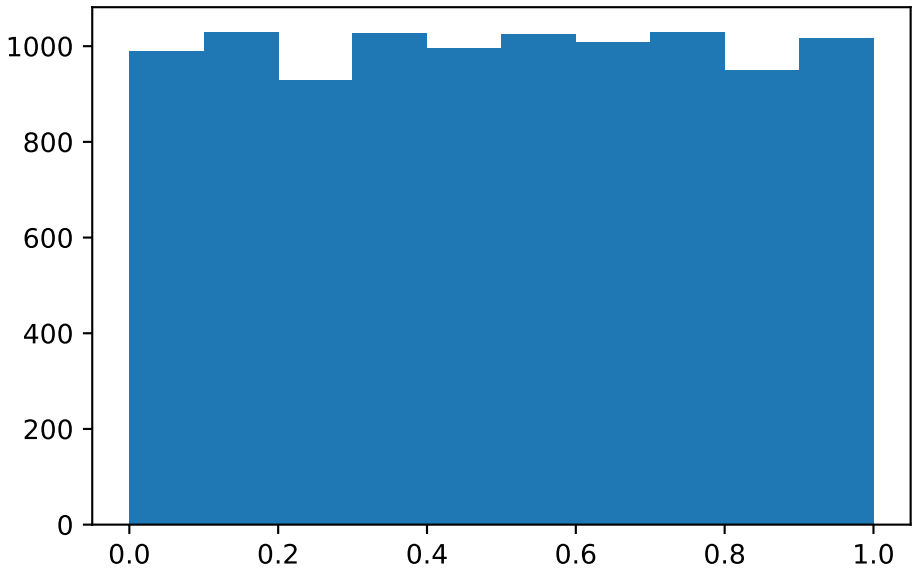
Consider a coin flip. The random variable $X$ takes value 0 if head and 1 if tail. We impose Bernoulli likelihood for $p(x|\theta)$.

Since the support for $\theta \in [0, 1]$; we use Beta distribution to model $\theta$, with parameters $a, b$.

Fair coin: choose $a = b = 1$, yielding uniform distribution over $[0, 1]$.

## Example: coin flip

```
(array([ 989., 1030.,  930., 1027.,  995., 1026., 1008., 1030.,  949.,
        1016.]),
 array([1.26352283e-04, 1.00110131e-01, 2.00093909e-01, 3.00077687e-01,
        4.00061465e-01, 5.00045244e-01, 6.00029022e-01, 7.00012800e-01,
        7.99996578e-01, 8.99980357e-01, 9.99964135e-01]),
 <BarContainer object of 10 artists>)
```

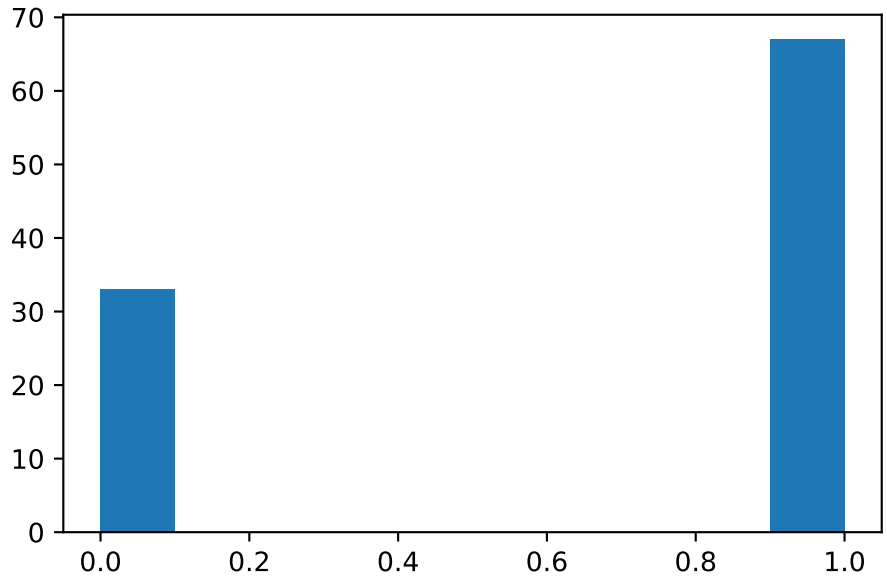



**Example: coin flip**

```
theta = 0.7
N = 100
# 0: tail, 1: head
y = np.random.binomial(n = 1, p = theta, size=N)
print(y)
plt.hist(y)
```

```
[1 0 1 0 1 1 0 1 0 1 1 1 1 0 1 0 1 0 1 0 0 1 1 1 0 1 1 1 0 1 1 1 0 0 1 1 1
 0 1 1 1 0 1 1 1 1 0 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 0 1 0 0 0 1 1 1 1 1 1 1
 1 0 1 0 1 0 0 1 1 1 1 0 1 0 1 1 0 1 1 1 1 0 1 0 0 1]
```

```
(array([33.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0., 67.]),
 array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
 <BarContainer object of 10 artists>)
```




### Example: coin flip

Compute the posterior:

$$p(\theta|y_{1:N}) = \frac{\prod_n p(y_n|\theta)p(\theta)}{p(y_{1:N})}$$

How?

### Example: coin flip

Numerator (let $s_n = \sum_n y_n$):

$$p(y_{1:N}|\theta)p(\theta) = \prod_n p(y_n|\theta)p(\theta) \tag{5}$$

$$= \theta^{\sum_n y_n}(1-\theta)^{N-\sum y_n}\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \tag{6}$$

$$= \frac{1}{B(a,b)}\theta^{s_n+a-1}(1-\theta)^{N-s_n+b-1}. \tag{7}$$

This form looks a lot like a Beta distribution with parameters $\tilde{a} = s_n + a$ and $\tilde{b} = N - s_n + b$.

## Example: coin flip

Denominator (marginal likelihood):

$$p(y_{1:N}) = \int_0^1 \prod_n p(y_n|\theta)p(\theta)d\theta \tag{8}$$

$$= \frac{1}{B(a,b)} \int_0^1 \theta^{\sum_n y_n + a - 1}(1-\theta)^{N - \sum y_n + b - 1} d\theta \tag{9}$$

$$\tag{10}$$

## Example: coin flip

So the posterior is,

$$p(\theta|y_{1:N}) = \frac{\theta^{\sum_n y_n + a - 1}(1-\theta)^{N - \sum y_n + b - 1}}{\int_0^1 \theta^{\sum_n y_n + a - 1}(1-\theta)^{N - \sum y_n + b - 1} d\theta}. \tag{11}$$

$B(a,b)$ cancels.

The denomiator is a constant but what is it?

## Example: coin flip

Beta function:

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta.$$

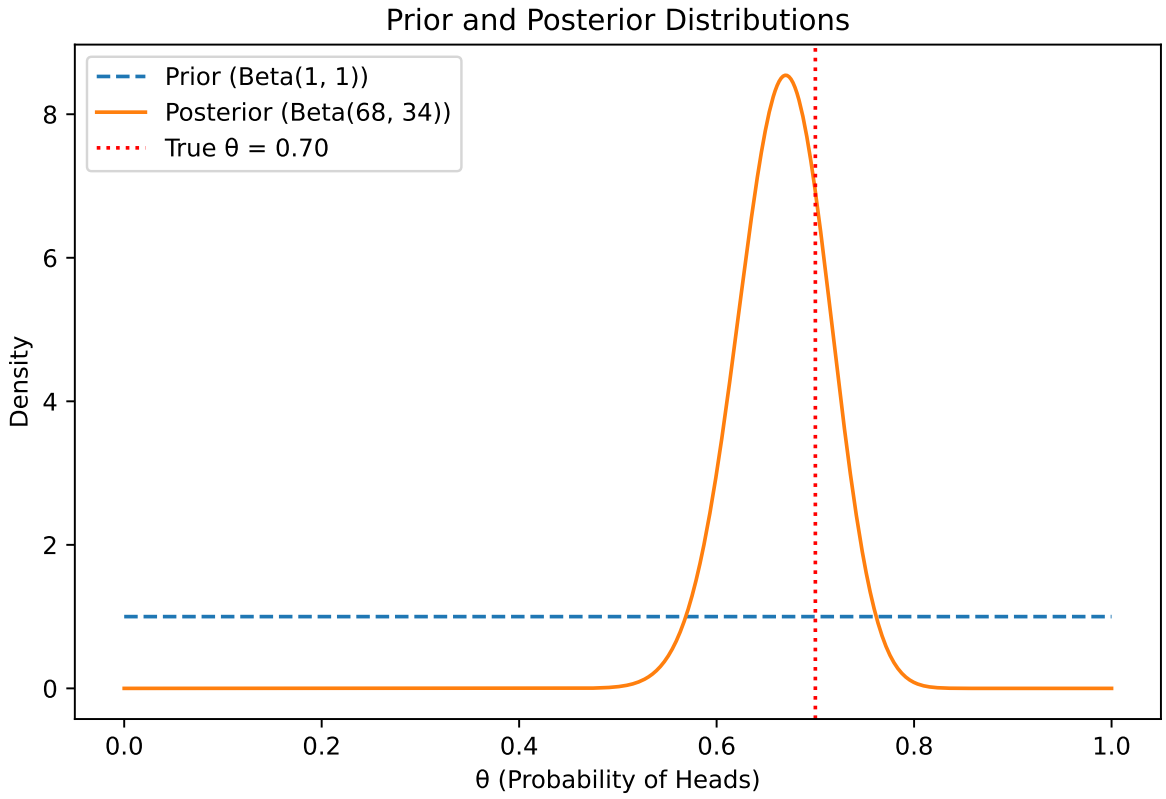
The denominator:

$$\int_0^1 \theta^{s_n + a - 1}(1-\theta)^{N - s_n + b - 1} d\theta,$$

so the denominator is $B(s_n + a, N - s_n + b)$. The posterior is indeed Beta distributed with parameters $\tilde{a} = s_n + a$ and $\tilde{b} = N - s_n + b$.

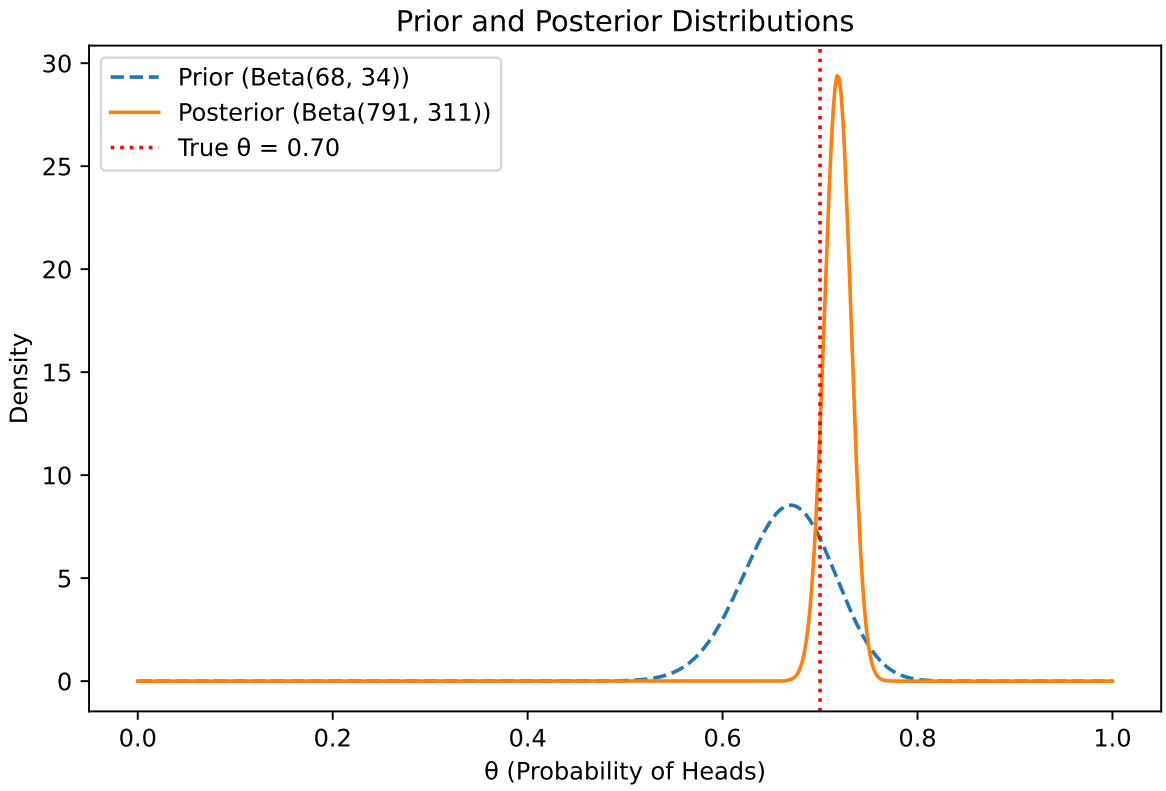
### Example: coin flip

```
Posterior parameters: alpha = 68, beta = 34
```




### Example: coin flip

```
Posterior parameters: alpha = 791, beta = 311
```




### Example: coin flip

- Bernoulli likelihood.
- Beta prior.
- Closed form expression for the marginal likelihood.
- The posterior was also Beta distribution.

Is it true that the prior and posterior are always in the same *family*?

What do we do if the closed form for marginal is not available?

### Example: Monty Hall problem

Famous game show:

- A car is hidden behind one of $N$ doors ,
- Goats behind the remaining $N - 1$ doors.

You get to make an initial guess and select a door.

- The show host opens $N-2$ doors, except the door that you chose and another door where the car is hidden.

Should you switch your selection?

### Example: Monty Hall problem

Let $N = 10,000$.

Prior belief: you do not have any information so the car is equally likely to be behind each door: $1/N$.

You choose door no. 1.

Observation: host opens all of the doors except door no. 1 and no. 387.

Exercise: is your belief on where the car is hidden updated? how would you update your belief?

### Computation

Bayesian inference requires computing the posterior $p(x_H|x_O)$. Why is this difficult?

$$p(x_H|x_O) = \frac{p(x_O|x_H)p(x_H)}{p(x_O)}. \tag{12}$$

$p(x_O) = \int p(x_O|x_H)p(x_H)dx_H$.

- To obtain the closed form for the posterior, we need to compute the integral over $x_H$.
- In many cases, the integral is intractable.

### Conjugacy: conjugate prior

- We say the prior is a conjugate prior to the likelihood if the resulting posterior is in the same family as the prior.
- The prior-posterior are conjugate distributions with respect to the likelihood.

Beta-Bernoulli used for modeling the coin flip is an example. Can we generalize?

## Conjugacy: family of distributions

- The same functional form: the prior and posterior distributions are described by the same type of probability distribution, characterized by a set of parameters.
- A consistent parameterization: the posterior distribution's parameters are updated versions of the prior's parameters, derived from the likelihood and observed data.

## Exponential family: definition

A random variable $Y$ with parameter(s) $\theta$ with the probability distributional form:

$$p(y|\theta) = h(y) \exp(\theta^T T(y) - A(\theta))$$

- All members of the exponential family have a conjugate prior.
- Exponential family plays a central role in Variational Inference and Generalized Linear Models.

## Exponential family: definition

A random variable $Y$ with parameter(s) $\theta$ with the probability distributional form:

$$p(y|\eta) = h(y) \exp(\eta^T T(y) - A(\eta))$$

- $h(x)$: scaling constant,
- $T(x) \in \mathbb{R}^D$: are sufficient statistics,
- $\eta$: are the natural or canonical parameters,
- $Z(\eta)$: partition function (or normalization constant) such that $A(\eta) = \log(Z(\eta))$.

## Exponential family: Bernoulli example

$Y \sim \text{Bernoulli}(\theta)$:

$$\begin{aligned} p(y|\theta) &= \theta^y (1-\theta)^{1-y} && (13) \\ &= \exp(y \log(\theta) + (1-y) \log(1-\theta)) && (14) \\ &= \exp(T(y)^T \eta), && (15) \end{aligned}$$

where

$$T(y) = (1[y = 1], 1[y = 0]) \tag{16}$$
$$\eta = (\log(\theta), \log(1 - \theta)). \tag{17}$$

## Exponential family: minimal representation

- Over-complete representation arises when the sufficient statistics are not linearly independent, meaning there are redundancies in how the statistics describe the data.
- For example, the sufficient statistic $T(y) = (1[y = 1], 1[y = 0])$ are not linearly independent.

## Exponential family: linear indepence

- Linear independence of functions $f_k(x)$: if $\sum_k c_k f_k(x) = 0$ for all $x$, then $c_k = 0$ for all $k$.
- To show the functions are linearly dependent: find non-trivial $c_k$ such that $\sum_k c_k f_k(x) = 0$ for all $x$.

Note: $1[y = 1] + 1[y = 0] = 1$ so $c_1 = c_2 = 1$.

## Exponential family: Bernoulli example

Minimal representation for Beta:

$$p(y|\theta) = \exp\left[y \log\left(\frac{\theta}{1-\theta}\right) - \log(1 - \theta)\right] \tag{18}$$
$$= \exp(y\eta + \log(1 + e^{\eta})), \tag{19}$$

- $T(y) = y$,
- $\eta = \log(\theta/(1 - \theta))$,
- $A(\eta) = \log(1 + e^{\eta})$.

## Exponential family: Discrete-valued

- Bernoulli/Binomial.
- Poisson.
- Geometric.
- Negative Binomial (mean-dispersion parameterization).

## Exponential family: Continuous-valued

- Normal.
- Gamma.
- Beta.
- Exponential.

## Exponential family: Multivariate distributions

- Dirichlet.
- Multinomial/Categorical.
- Multivariate Normal.
- Wishart.
- Inverse Wishart.

## Summary

- Probability theory: modeling forward process (generation).
- Statistics: inverse problem of inferring the model state that produced the observed data.
- Inverse problem is challenging: many possible states could have led to the observed data so we need to represent uncertainty.
- Bayes theorem provides a natural mechanism to update your belief (representation of uncertainty) in accordance with the observation.

## Summary

- Even though a likelihood is in the exponential family, we may not know the conjugate form of the prior (we just know it exists).
- Many problems can modelled using Exponential Family and coupled with conjugacy, we can compute the posterior distribution. Table of conjugate distributions.
- Even when conjugacy does not apply, it may be possible to exactly compute the marginal and hence, perform posterior inference.
- When conjugacy does not apply, for complex models, we need to approximate the posterior via sampling or optimization.

## Appendix

- $d\nu(x)$ is used mostly in the context of an integral depicting the measure and the space. $d$ can be thought of as "infinitesimal" but in general, the notion of infinitesimal is vague so it can be treated more as a symbolic representation connected to integral.

- $dx$: denotes Lebesgue measure but also a set.
- $\nu(dx)$: measure of the infinitesimal set $dx$.