

# Bayesian philosophy, non-informative priors, exchangeability

Spring 2024

Instructor: Shandian Zhe

[zhe@cs.utah.edu](mailto:zhe@cs.utah.edu)

School of Computing



# Outline

- Bayesian vs. frequentist
- Uninformative priors
- Exchangeability, de Finetti's theorem

# Outline

- Bayesian vs. frequentist
- Uninformative priors
- Exchangeability, de Finetti's theorem

# Bayesian vs. Frequentist

- Let us consider to estimate a parameter  $\theta$ , e.g., the chance of head (tossing a coin), from observed data  $\mathbf{x}_1, \dots, \mathbf{x}_N$

# Bayesian vs. Frequentist

- Let us consider to estimate a parameter  $\theta$ , e.g., the chance of head (tossing a coin), from observed data  $\mathbf{x}_1, \dots, \mathbf{x}_N$
- Frequentist:  $\theta$  is some fixed parameter, no randomness

# Bayesian vs. Frequentist

- Let us consider to estimate a parameter  $\theta$ , e.g., the chance of head (tossing a coin), from observed data  $\mathbf{x}_1, \dots, \mathbf{x}_N$
- Frequentist:  $\theta$  is some fixed parameter, no randomness
  - We want to estimate it from observations

$$\theta_{ML} = \operatorname{argmax}_{\theta} \sum_{i=1}^N \log p(\mathbf{x}_i | \theta)$$

- How to quantify your uncertainty?
  - confidence level, note that  $\theta_{ML}$  is a R.V., but  $\theta$  is not.

# Bayesian vs. Frequentist

- Let us consider to estimate a parameter  $\theta$ , e.g., the chance of head (tossing a coin), from observed data  $\mathbf{x}_1, \dots, \mathbf{x}_N$
- Bayesian:  $\theta$  is a random variable as well!

# Bayesian vs. Frequentist

- Let us consider to estimate a parameter  $\theta$ , e.g., the chance of head (tossing a coin), from observed data  $\mathbf{x}_1, \dots, \mathbf{x}_N$

- Bayesian:  $\theta$  is a random variable as well!
  - We want to estimate it from observations

$$p(\theta|\mathcal{D}) \propto p(\theta) \prod_{i=1}^N p(\mathbf{x}_i|\theta)$$

- How to quantify your uncertainty?

Posterior distribution!  $p(\theta|\mathcal{D})$



# Bayesian vs. Frequentist

- In Bayesian world, every thing is random! (every variable is a random variable)

# Bayesian vs. Frequentist

- In Bayesian world, every thing is random! (every variable is a random variable)
- Why is random  $\theta$  is important?

# Bayesian vs. Frequentist

- In Bayesian world, every thing is random! (every variable is a random variable)
- Why is random  $\theta$  is important?
  - We can encode our **beliefs, previous experience and desires** in the prior  $p(\theta)$

# Bayesian vs. Frequentist

- In Bayesian world, every thing is random! (every variable is a random variable)
- Why is random  $\theta$  is important?
  - ❑ We can encode our **beliefs, previous experience and desires** in the prior  $p(\theta)$
  - ❑ We can make probabilistic statements about  $\theta$  (mean, variance, quantiles, etc.).

# Bayesian vs. Frequentist

- In Bayesian world, every thing is random! (every variable is a random variable)
- Why is random  $\theta$  is important?
  - ❑ We can encode our **beliefs, previous experience and desires** in the prior  $p(\theta)$
  - ❑ We can make probabilistic statements about  $\theta$  (mean, variance, quantiles, etc.).
  - ❑ We can make Bayesian prediction that integrates all the possible outcomes

$$p(\mathbf{x}^* | \mathbf{x}_1, \dots, \mathbf{x}_N) = \int p(\mathbf{x}^* | \theta) p(\theta | \mathbf{x}_1, \dots, \mathbf{x}_N)$$

# Bayesian vs. Frequentist

- Is Bayesian analysis subjective?
  - Not necessary: Bayesian provides a convenient way to incorporate subjective beliefs (important for AI!) But it can also use uninformative priors (this is objective Bayesian!)
  - Frequentist models make assumptions, too!
  - Whether using frequent or Bayesian models, always **check the assumptions you make**

# Outline

- Bayesian vs. frequentist
- **Uninformative priors**
- Exchangeability, de Finetti's theorem

# Uninformative priors

- In many cases, we have little idea of what form the distribution should take
- Though conjugate priors are computationally nice, objective Bayesians instead prefer priors which *has little influence* on the posterior distribution. Such a prior is called an *uninformative prior*.
- Let the *data speak for themselves*



# Uninformative priors

- What priors do you have immediately in mind?

# Uninformative priors

- What priors do you have immediately in mind?

Uniform distribution!

# Uninformative priors

- What priors do you have immediately in mind?

**Uniform distribution!**

Now that I do not know which parameter is more likely to be sampled, let us just assume the chances are equal!

# Uninformative priors

- Uniform distribution

For finite states:  $p(\lambda) = 1/K$

For finite interval:  $p(\lambda) = 1/(b - a)$

$$p(\mathbf{x}|\lambda)$$

# Uninformative priors

- Uniform distribution

What about unbounded domains?  $\lambda \in \mathbb{R}$

# Uninformative priors

- Uniform distribution

What about unbounded domains?  $\lambda \in \mathbb{R}$

$$p(\lambda) \propto \text{const}$$

# Uninformative priors

- Uniform distribution

What about unbounded domains?  $\lambda \in \mathbb{R}$

$$p(\lambda) \propto \text{const}$$

This is an *improper* prior, because normalization diverges

We can still use it as long as the posterior is *proper*

# Uninformative priors

- Problem of uniform distribution: NOT *translation invariance*

$$p(\lambda) \propto \text{const}$$

$$\lambda = \eta^2$$



# Uninformative priors

- Problem of uniform distribution: NOT *translation invariance*

$$p(\lambda) \propto \text{const}$$

$$\lambda = \eta^2$$

$$p_\eta(\eta) = p_\lambda(\lambda) \left| \frac{d\lambda}{d\eta} \right| = p_\lambda(\eta^2) 2\eta \propto \eta$$

# Uninformative priors

- Problem of uniform distribution: NOT *translation invariance*

$$p(\lambda) \propto \text{const}$$

$$\lambda = \eta^2$$

$$p_\eta(\eta) = p_\lambda(\lambda) \left| \frac{d\lambda}{d\eta} \right| = p_\lambda(\eta^2) 2\eta \propto \eta$$

When we do variable transformations, the prior is no longer uninformative!

# Uninformative priors

- Let us take *translation invariance* into account

If the likelihood takes the form

$$p(x|\lambda) = f(x - \lambda)$$

# Uninformative priors

- Let us take *translation invariance* into account

If the likelihood takes the form

$$p(x|\lambda) = f(x - \lambda)$$

$\lambda$  is *location* parameter, and then density exhibits *shift invariance*

$$\hat{x} = x + c \quad \hat{\lambda} = \lambda + c$$

$$p(\hat{x}|\hat{\lambda}) = f(\hat{x} - \hat{\lambda})$$

$$p(\mathbf{x}|\lambda)$$

# Uninformative priors

- We want to construct a prior that reflects this shift invariance (why: more consistent with the likelihood, less influence on the posterior!)

$$p(\mathbf{x}|\lambda)$$

# Uninformative priors

- We want to construct a prior that reflects this shift invariance (why: **more consistent with the likelihood, less influence on the posterior!**)
- How? We choose a prior that assigns equal probability mass to an arbitrary interval  $[A, B]$  as to the shifted interval  $[A+c, B+c]$

# Uninformative priors

- We want to construct a prior that reflects this shift invariance (why: **more consistent with the likelihood, less influence on the posterior!**)
- How? We choose a prior that assigns equal probability mass to an arbitrary interval  $[A, B]$  as to the shifted interval  $[A+c, B+c]$

$$\int_A^B p(\lambda) d\lambda = \int_{A+c}^{B+c} p(\lambda) d\lambda$$

# Uninformative priors

$$\int_A^B p(\lambda) d\lambda = \int_{A+c}^{B+c} p(\lambda) d\lambda = \int_A^B p(\lambda + c) d\lambda$$

$\lambda = \bar{\lambda} + c$

$$p(\lambda) = p(\lambda + c)$$

$$p(\lambda) \propto \text{const}$$



# Uninformative priors

Example: for a Gaussian likelihood

$$p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} \underline{(x - \mu)^2}\right)$$

# Uninformative priors

Example: for a Gaussian likelihood

$$p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} \underbrace{(x - \mu)^2}_{\text{shift invariance density}}\right)$$

shift invariance density

Conjugate prior

$$p(\mu|\alpha, v^2) = \mathcal{N}(\mu|\alpha, v^2) = \frac{1}{\sqrt{2\pi}v} \exp\left(-\frac{1}{2v^2}(\mu - \alpha)^2\right)$$

# Uninformative priors

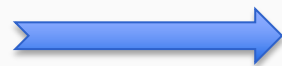
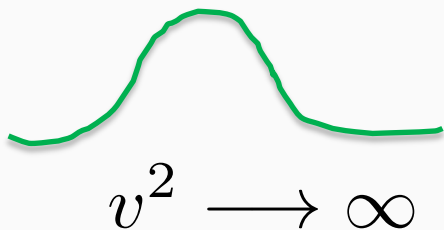
Example: for a Gaussian likelihood

$$p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

shift invariance density

Conjugate prior

$$p(\mu|\alpha, v^2) = \mathcal{N}(\mu|\alpha, v^2) = \frac{1}{\sqrt{2\pi}v} \exp\left(-\frac{1}{2v^2}(\mu - \alpha)^2\right)$$



$$p(\mu) \propto \text{const}$$

# Uninformative priors

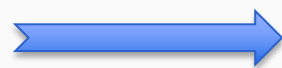
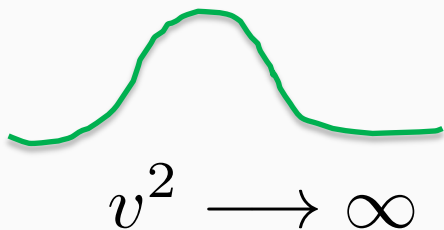
Example: for a Gaussian likelihood

$$p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

shift invariance density

Conjugate prior

$$p(\mu|\alpha, v^2) = \mathcal{N}(\mu|\alpha, v^2) = \frac{1}{\sqrt{2\pi}v} \exp\left(-\frac{1}{2v^2}(\mu - \alpha)^2\right)$$



$$p(\mu) \propto \text{const}$$

Limit of the conjugate prior

# Uninformative priors

- Let us take *translation invariance* into account

If the likelihood takes the form

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \quad \sigma > 0 \quad f \text{ normalizes regularly}$$

# Uninformative priors

- Let us take *translation invariance* into account

If the likelihood takes the form

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \quad \sigma > 0 \quad f \text{ normalizes regularly}$$

$\sigma$  is *scale* parameter, and the density exhibits *scale invariance*

$$\hat{x} = cx \quad \hat{\sigma} = c\sigma$$

$$p(\hat{x}|\hat{\sigma}) = \frac{1}{\hat{\sigma}} f\left(\frac{\hat{x}}{\hat{\sigma}}\right) \quad \text{Verify it by yourself}$$

# Uninformative priors

- We want to construct a prior that reflects this scale invariance (why: **more consist with the likelihood, less influence on the posterior!**)

# Uninformative priors

- We want to construct a prior that reflects this scale invariance (why: **more consist with the likelihood, less influence on the posterior!**)
- How, consider an arbitrary interval  $[A, B]$ , the prior should assign equal mass over an arbitrary scaled interval  $[A/c, B/c]$

$$\int_A^B p(\sigma) d\sigma = \int_{A/c}^{B/c} p(\sigma) d\sigma$$



# Uninformative priors

$$\bar{\sigma} = c\sigma$$

$$\int_A^B p(\sigma) d\sigma = \int_{A/c}^{B/c} p(\sigma) d\sigma = \int_A^B p\left(\frac{1}{c}\sigma\right) \frac{1}{c} d\sigma$$

$$p(\sigma) = p\left(\frac{1}{c}\sigma\right) \frac{1}{c}$$

$$p(\sigma) \propto 1/\sigma$$

# Uninformative priors

Example: for a Gaussian likelihood

$$p(x|\sigma) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2} \left[\frac{x - \mu}{\sigma}\right]^2\right)$$

Uninformative prior

$$p(\sigma) \propto 1/\sigma \quad \xrightarrow{\lambda = 1/\sigma^2} \quad p(\lambda) \propto 1/\lambda$$

Conjugate prior

$$p(\lambda|a, b) = \text{Gam}(\lambda|a, b) \propto \lambda^{a-1} \exp(-b\lambda)$$

$$a = 0, b = 0 \quad \xrightarrow{\quad} \quad p(\lambda) \propto 1/\lambda$$

# Uninformative Priors

- Jeffreys priors

$$\pi_J(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$$

Fisher information

$$I(\theta) = -\mathbb{E}_X \left[ \frac{d^2 \log p(X|\theta)}{d\theta^2} \right]$$

# Uninformative Priors

- Jeffreys priors

$$\pi_J(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$$

Fisher information

$$I(\theta) = -\mathbb{E}_X \left[ \frac{d^2 \log p(X|\theta)}{d\theta^2} \right]$$

Expectation w.r.t  $p(X|\theta)$



# Uninformative Priors

- Jeffreys priors

$$\pi_J(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$$

Fisher information

$$I(\theta) = -\mathbb{E}_X \left[ \frac{d^2 \log p(X|\theta)}{d\theta^2} \right]$$

Expectation w.r.t  $p(X|\theta)$



Note, for vector case, it becomes the Hessian

# Jeffreys priors - example

Binomial likelihood

$$X \sim \text{Bin}(n, \theta), 0 \leq \theta \leq 1$$

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

# Jeffreys priors - example

Binomial likelihood

$$X \sim \text{Bin}(n, \theta), 0 \leq \theta \leq 1$$

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

Let's construct a Jeffreys prior over  $\theta$

# Jeffreys priors - example

Binomial likelihood

$$X \sim \text{Bin}(n, \theta), 0 \leq \theta \leq 1$$

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

Let's construct a Jeffreys prior over  $\theta$

$$\log p(x|\theta) = x \log \theta + (n - x) \log(1 - \theta)$$

$$\frac{d}{d\theta} \log p(x|\theta) = \frac{x}{\theta} - \frac{n - x}{1 - \theta}$$

$$\frac{d^2}{d\theta^2} \log p(x|\theta) = -\frac{x}{\theta^2} - \frac{n - x}{(1 - \theta)^2}$$



# Jeffreys priors - example

$$\frac{d^2}{d\theta^2} \log p(x|\theta) = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}$$



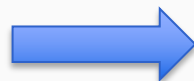
$$\mathbb{E}[x] = n\theta$$

$$I(\theta) = -\mathbb{E}_x \left[ \frac{d^2 \log p(x|\theta)}{d\theta^2} \right]$$

$$= \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2}$$

$$= \frac{n}{\theta} + \frac{n}{1-\theta}$$

$$= \frac{n}{\theta(1-\theta)}$$



$$\pi_J(\theta) = I(\theta)^{\frac{1}{2}} \propto \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}}$$

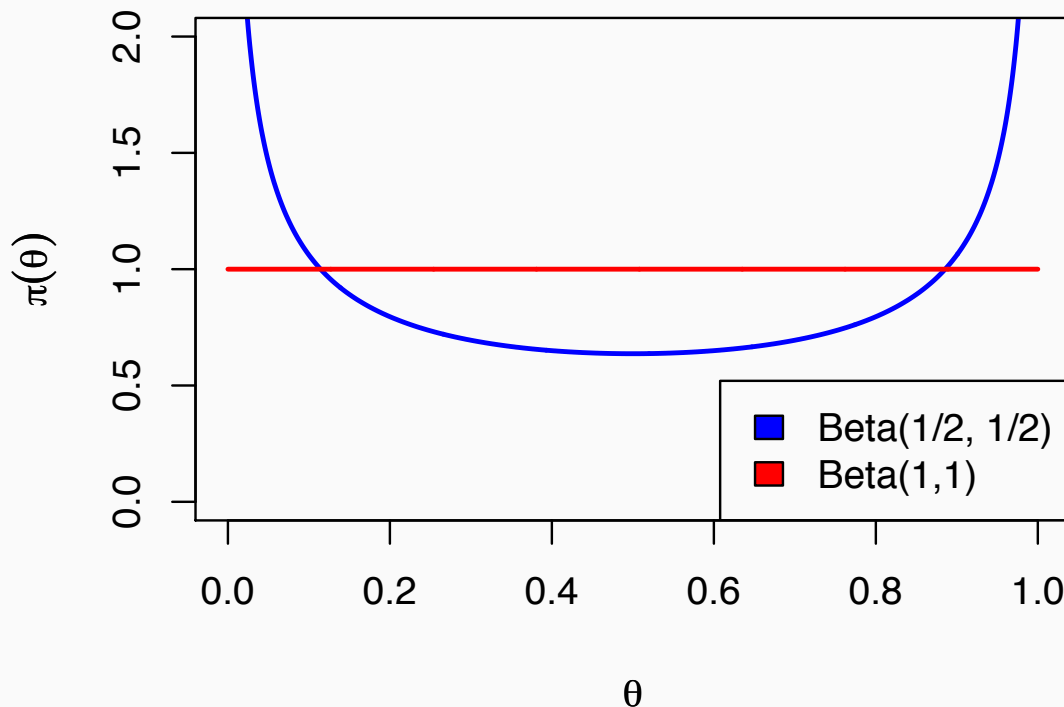
Beta( $\frac{1}{2}$ ,  $\frac{1}{2}$ )

# Jeffreys priors - example

Binomial likelihood

$$X \sim \text{Bin}(n, \theta), 0 \leq \theta \leq 1$$

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



Data takes least effect

$$\theta = \frac{1}{2}$$

Data takes greatest effect

$$\theta = 0 \text{ or } 1$$

Prior is consistent with the data effect!

# Jeffreys priors – translation invariance

- Let us consider a general translation  $\phi = h(\theta)$

What is the Jeffreys prior over  $\phi$  ?

$$\pi_J(\phi) \propto |\mathbf{I}(\phi)|^{\frac{1}{2}}$$

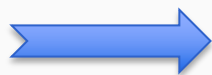
Use Chain rule

$$\begin{aligned}\mathbf{I}(\phi) &= -\mathbb{E} \left[ \frac{d^2 \log p(X|\phi)}{d\phi^2} \right] \\ &= -\mathbb{E} \left[ \frac{d^2 \log p(X|\theta)}{d\theta^2} \left( \frac{d\theta}{d\phi} \right)^2 + \frac{d \log p(X|\theta)}{d\theta} \frac{d^2 \theta}{d\phi^2} \right]\end{aligned}$$

# Jeffreys priors – translation invariance

We know  $\mathbb{E} \left[ \frac{d \log p(X|\theta)}{d\theta} \right] = 0$  Why?

$$\forall \theta, \int p(X|\theta) dX = 1$$



$$\begin{aligned} 0 &= \frac{d}{d\theta} \int p(X|\theta) dX \\ &= \int \frac{dp(X|\theta)}{d\theta} \frac{p(X|\theta)}{p(X|\theta)} dX \\ &= \int \left[ \frac{dp(X|\theta)}{d\theta} \frac{1}{p(X|\theta)} \right] p(X|\theta) dX \\ &= \int \left[ \frac{d \log p(X|\theta)}{d\theta} \right] p(X|\theta) dX \\ &= \mathbb{E} \left[ \frac{d \log p(X|\theta)}{d\theta} \right] \end{aligned}$$

# Jeffreys priors – translation invariance

$$\mathbf{I}(\phi) = -\mathbb{E} \left[ \underbrace{\frac{d^2 \log p(X|\theta)}{d\theta^2}}_{\mathbf{I}(\theta)} \left( \frac{d\theta}{d\phi} \right)^2 + \underbrace{\frac{d \log p(X|\theta)}{d\theta} \frac{d^2 \theta}{d\phi^2}}_0 \right]$$



$$\mathbf{I}(\phi) = \mathbf{I}(\theta) \left( \frac{d\theta}{d\phi} \right)^2$$



$$\sqrt{\mathbf{I}(\phi)} = \sqrt{\mathbf{I}(\theta)} \left| \frac{d\theta}{d\phi} \right|$$

# Jeffreys priors – translation invariance

$$\sqrt{\mathbf{I}(\phi)} = \sqrt{\mathbf{I}(\theta)} \left| \frac{d\theta}{d\phi} \right|$$

Now, we can see

When we directly construct Jeffreys prior

$$\pi_J(\phi) \propto \sqrt{\mathbf{I}(\phi)} = \sqrt{\mathbf{I}(\theta)} \left| \frac{d\theta}{d\phi} \right| \longrightarrow \text{The same!}$$

When we derive the prior via variable transformation

$$\pi_J(\phi) \propto \sqrt{\mathbf{I}(h^{-1}(\phi))} \left| \frac{d\theta}{d\phi} \right| = \sqrt{\mathbf{I}(\theta)} \left| \frac{d\theta}{d\phi} \right|$$

# Jeffreys priors – translation invariance

- Now we can show, for a Gaussian likelihood

$$p(x|\mu, \sigma) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

$$\pi_J(\mu) \propto 1$$

$$\pi_J(\sigma) \propto \frac{1}{\sigma}$$

Leave it as your  
exercise

# Jeffreys prior: cons

- Usually not conjugate
  - If you choose Jeffreys prior over  $\mu, \sigma$  for a Gaussian likelihood

The posterior of  $\mu$  will be a student  $t$  distribution
- Works well for single parameter, but not for models with multidimensional parameters (e.g., poor convergence properties, not very reasonable estimates)



# Reference priors

- formalize what exactly we mean by an “uninformative prior”: *a function that maximizes some measure of distance or divergence between the posterior and prior, as data observations are made.*
- A commonly used divergence is KL divergence

$$\text{KL}(p(\theta|t) \| p(\theta)) = \int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta$$


Observations



# Reference priors

- We choose the prior that maximizes the expected KL divergence between the posterior and the prior

Observations


$$\begin{aligned} I(\Theta, T) &= \int p(t) \int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta dt \\ &= \int \int p(\theta, t) \log \frac{p(\theta, t)}{p(\theta)p(t)} d\theta dt \end{aligned}$$



$$p^*(\theta) = \arg \max_{p(\theta)} I(\Theta, T)$$

Mutual information

# Outline

- Bayesian vs. frequentist
- Uninformative priors
- Exchangeability, de Finetti's theorem

# Bayesian vs. Frequentist

- Given a distribution  $p(x|\theta)$  governed by  $\theta$
- Frequentist: I believe  $\theta$  is objective constant, I need to estimate it from IID samples  $x_1, \dots, x_N$
- Bayesian: I believe  $\theta$  is some *latent random variable* – it was first sampled from a *prior distribution*  $p(\theta)$  , then given  $\theta$  , we sample the observations  $x_1, \dots, x_N$

# Bayesian vs. Frequentist

- Bayesian: I believe  $\theta$  is some *latent random variable* – it was first sampled from a *prior distribution*  $p(\theta)$  , then given  $\theta$  , we sample the observations  $x_1, \dots, x_N$
- Although it sounds a philosophical choice, can we justify Bayesian modeling with some mathematical evidence?

# Exchangeability

- Most statistical analysis are based on IID observations  $x_1, \dots, x_N$

$$p(X_1 = x_1, \dots, X_N = x_N) = \prod_{n=1}^N p(X_n = x_n)$$

- While the assumption is convenient, it may not be reasonable in many problems: weather conditions, stock prices, precipitation, disease rate, ...
- Exchangeability is a much *weaker* assumption

# Exchangeability

- Finite exchangeability: Given  $N$  random variables, and arbitrary permutation  $\pi(1), \dots, \pi(N)$

$$X_1, \dots, X_N \stackrel{d}{=} X_{\pi(1)}, \dots, X_{\pi(N)}$$



$\forall x_1, \dots, x_N$  in the domain

$$p(X_1 = x_1, \dots, X_N = x_N) = p(X_1 = x_{\pi(1)}, \dots, X_N = x_{\pi(N)})$$

**e.g.**  $p(X_1 = 1, X_2 = 2, X_3 = 3) = p(X_1 = 2, X_2 = 3, X_3 = 1)$   
 $= p(X_1 = 3, X_2 = 1, X_3 = 2) = \dots$

# Exchangeability – infinite sequence

- An infinite sequence of random variables  $\{X_i\}_{i=1}^{\infty}$  is exchangeable if  $\forall n = 1, 2, \dots$

$$X_1, \dots, X_n \stackrel{d}{=} X_{\pi(1)}, \dots, X_{\pi(n)}, \quad \forall \pi \in S(n),$$

where  $S(n)$  are all possible permutations over the first  $n$  variables



# Exchangeability

- Essentially assume the *symmetry* of the density

$$p(X_1 = x_1, \dots, X_N = x_N) = p(X_1 = x_{\pi(1)}, \dots, X_N = x_{\pi(N)})$$



# Exchangeability - one specific example

- Polya's Urn

- Given an urn with  $B_0$  black and  $W_0$  white balls, draw balls with the following procedure

- (1) Draw a ball at random from the urn and note its color
    - (2) If the ball is black then  $X_i = 1$ ; otherwise  $X_i = 0$
    - (3)  $i = i + 1$
    - (4) Place  $\alpha$  balls of the same color in the urn
    - (5) Goto (1)

# Exchangeability - one specific example

- Polya's Urn

- Given an urn with  $B_0$  black and  $W_0$  white balls, draw balls with the following procedure

- (1) Draw a ball at random from the urn and note its color
- (2) If the ball is black then  $X_i = 1$ ; otherwise  $X_i = 0$
- (3)  $i = i + 1$
- (4) Place  $a$  balls of the same color in the urn
- (5) Goto (1)

$$p(1, 1, 0, 1) = \frac{B_0}{B_0 + W_0} \times \frac{B_0 + a - 1}{B_0 + W_0 + a - 1} \times \frac{W_0}{B_0 + W_0 + 2a - 2} \times \frac{B_0 + 2a - 2}{B_0 + W_0 + 3a - 3}$$

$$p(1, 0, 1, 1) = \frac{B_0}{B_0 + W_0} \times \frac{W_0}{B_0 + W_0 + a - 1} \times \frac{B_0 + a - 1}{B_0 + W_0 + 2a - 2} \times \frac{B_0 + 2a - 2}{B_0 + W_0 + 3a - 3}$$

The sequence  $\{X_i, i \geq 1\}$  is exchangeable but not IID

# De Finetti's theorem

**(de Finetti 1931)** A binary sequence  $\{X_i\}_{i=1}^{\infty}$  is *exchangeable* iff there exists a *distribution function*  $F$  on  $[0, 1]$  such that for all  $n$ ,

$$p(x_1, \dots, x_n) = \int_0^1 \theta^{t_n} (1 - \theta)^{n - t_n} dF(\theta),$$

where  $p(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$  and  $t_n = \sum_{i=1}^n x_i$ .

$p(\theta)d\theta$

1. There is a latent random variable  $\theta$
2. It has a prior distribution

# De Finetti's theorem

It further holds that  $F$  is the distribution function of the limiting frequency:

$$\Theta = \hat{X}_\infty = \lim_{n \rightarrow \infty} \sum_i X_i/n, \quad p(\Theta \leq \theta) = F(\theta)$$

and the Bernoulli distribution is obtained by conditioning with  $\Theta = \theta$ :

$$p(X_1 = x_1, \dots, X_n = x_n | \Theta = \theta) = \theta^{t_n} (1 - \theta)^{n - t_n} = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1 - x_i}$$

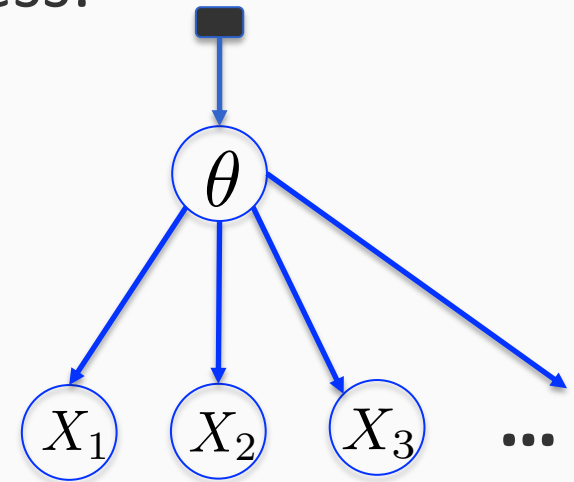
# De Finetti's theorem – the underlying sampling process

- If our binary observations  $\{X_i\}_{i=1}^{\infty}$  are exchangeable, it implies a hierarchical sampling process:

$$\theta \sim p(\theta)$$

Conditional independent

$$X_1, X_2, \dots | \theta \sim \prod_{i=1}^{\infty} p(X_i | \theta)$$



This justifies Bayesian modeling --- prior distribution objectively exists!

# Exchangeability

- Very widely used assumption in Bayesian modeling
- More flexible than IID, but is also restrictive
- Some classical/popular models

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "**Latent dirichlet allocation.**" *Journal of machine Learning research* 3.Jan (2003): 993-1022.

Airoldi, Edoardo M., et al. "**Mixed membership stochastic blockmodels.**" *Journal of machine learning research* 9.Sep (2008): 1981-2014

Lloyd, J., Orbanz, P., Ghahramani, Z., & Roy, D. M. (2012). **Random function priors for exchangeable arrays with applications to graphs and relational data.** In *Advances in Neural Information Processing Systems* (pp. 998-1006).

# What you need to know

- Bayesian vs. Frequentist
- What is uninformative prior
- What are shift invariance, scale invariance in likelihood? How to derive the corresponding uninformative prior?
- What is Jeffery's prior? Arbitrary translation invariance
- Exchangeability
- De-Finette theorem (how does it justify Bayesian )