Outline

1) Interpretations of Probability
2) Where does prior come from?
3) Examples

Interpretations of Probability

Why do we model anything as random?

What does "probability" mean in the real world?

1) Long-run frequency over repeated trials

Ex. repeatedly flipping a coin

shooting electrons at a double slit

"you can never step into the same river twice"

2) Systematic random sampling from a population

Ex. survey of 500 random voters

random assignment to treatment/control

Randomness comes from experimenter's actions

3) Subjective uncertainty about an outcome

chance that...
- President Biden is re-elected
- Higgs boson has a given mass
- $P = NP$
- 100^{th} digit of π is 5

Could be broad intersubjective agreement

These are often intertwined:

Ex. What if survey sampling is pseudo-random?

Probably relying on shared ignorance

Where does Λ come from?

(Bayesian rejoinder: Where does $\mathcal{P}$ come from?)

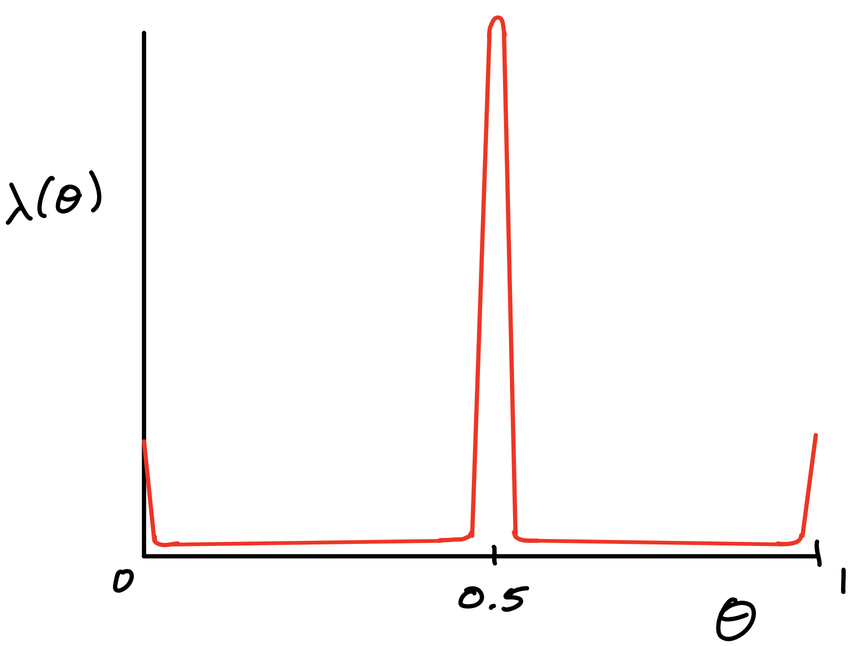
Four main sources for prior on θ

Source #1: Subjective beliefs

Pro: Brings all relevant info. to bear
Straightforward interp. of posterior

Con: Posterior is therefore subjective
Embarrassing to write "I think" in abstract
Hard if θ high-dim or $\mathcal{P}$ nonparametric

Ex: Flip coin 20 times, get 7 heads
0.5 probably a better estimate than 0.35
My subjective prior on coins:



Source #2: "Objective" or "vague" prior

Using default prior removes subjectivity
(But then what does the posterior mean?)

Flat prior $\lambda(\theta) \propto_\theta 1$ on Θ

"Indifference" (in θ parameterization)

Often improper ($\Lambda(\Theta) = \infty$) but usually ok

Ex: $\theta \sim$ flat prior on $\mathbb{R}$

$X \mid \theta \sim N(\theta, \sigma^2)$

$$\lambda(\theta \mid x) \propto_\theta p_\theta(x)$$
$$= \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2\sigma^2}$$
$$\propto_\theta N(x, \sigma^2)$$

$\Lambda([\theta, \theta+\varepsilon])$

$\approx \varepsilon \lambda(\theta) \propto \varepsilon \sqrt{J(\theta)}$

$\approx \sqrt{D_{KL}(P_\theta \| P_{\theta+\varepsilon})}$

Jeffreys prior $\lambda(\theta) \propto_\theta |J(\theta)|^{1/2}$

Higher density where P_θ "changing faster"

Invariant to parameterization

Ex. $X \mid \theta \sim \mathrm{Binom}(n, \theta)$

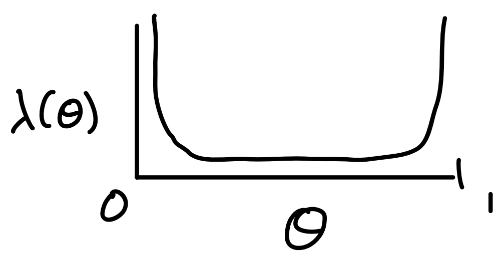
$$\lambda(\theta) \propto_\theta J(\theta)^{1/2} = \left(\frac{n}{\theta(1-\theta)}\right)^{1/2} \propto_\theta \mathrm{Beta}\left(\tfrac{1}{2}, \tfrac{1}{2}\right)$$

$\lambda(\theta) \to \infty$ as $\theta \to 0$ or 1:

(33x)

$D_{KL}(0.001 \| 0.01) \gg D_{KL}(0.49 \| 0.5)$

$7n \cdot 10^{-3}$ $\quad$ $2n \cdot 10^{-4}$



Intersubjective Agreement

Data may effectively rule out most θ values

$\rightsquigarrow$ Makes posterior uncontroversial

Ex. $X \sim \text{Binom}(10^4, \theta)$, observe $X = 3000$

$$SD_\theta(X/n) = \sqrt{\frac{\theta(1-\theta)}{n}} \leq 0.005$$

$$\Rightarrow \text{Lik}(\theta; X) \approx 0 \text{ outside } C = [0.29, 0.31]$$

All "reasonable" priors may be $\approx$ flat on C

$$\Rightarrow \lambda(\theta \mid X) \ \tilde{\propto}_\theta \ \text{Lik}(\theta; X)$$

$$\tilde{\propto}_\theta \ \exp\left\{-\frac{J(0.3)}{2}(\theta - 0.3)^2\right\}$$

$$\propto_\theta \ N(0.3, J(0.3)^{-1})$$

Data "swamps" everyone's prior

Gaussian sequence model

$X|\theta \sim N_d(\mu, I_d) \qquad \mu \in \mathbb{R}^d$

Jeffereys prior is flat: $\lambda(\mu) \propto_\mu 1$

$$\lambda(\mu|x) = N_d(X, I_d) \Rightarrow \mathbb{E}[\mu|x] = X$$

Same as UMVU

What about $\rho^2 = \|\mu\|^2$? Recall

$$\mu \sim N_d(X, I_d) \Rightarrow \mathbb{E}[\|\mu\|^2|X] = \|X\|^2 + d$$

Note $\delta_{UMVU}(x) = \|X\|^2 - d \Rightarrow \delta_\lambda(x) = \delta_{UMVU}(x) + 2d$

$$MSE(\theta; \delta_\lambda) = Var_\theta(\delta_\lambda) + Bias_\theta(\delta_\lambda)^2$$
$$= Var_\theta(\delta_{UMVU}) + 4d^2$$

What went wrong? Examine Jeffreys prior:

$$\mathbb{P}(\rho^2 \le t) = Vol(\text{Ball of radius } \sqrt{t})$$
$$= const(d) \cdot t^{d/2}$$

$$\Rightarrow \lambda(\rho^2) \propto_{\rho^2} (\rho^2)^{d/2 - 1} = \rho^{d-2}$$

Grows rapidly! Prior "expects" ρ^2 to be huge

Source #3: Prior or concurrent experience

May have many "copies" of same problem

Assume corresp. θ values drawn from a population

$\rightsquigarrow$ Hierarchical Bayes / empirical Bayes

Can be hard to choose right reference class

Ex. Estimate same-side bias for $m = 48$ coin flippers

Flipper i has n_i trials, "true" same-side prob θ_i

Hierarchical model: flippers $i = 1, \dots, m$

"hyperparameters" $\alpha, \beta \sim \lambda$ ← "hyperprior"

$$\theta_i \mid \alpha, \beta \overset{iid}{\sim} \text{Beta}(\alpha, \beta)$$

$$X_i \mid \alpha, \beta, \theta \overset{ind.}{\sim} \text{Binom}(n_i, \theta_i)$$

$$\mathbb{E}[\theta_i \mid X, \alpha, \beta] = \frac{X_i + \alpha}{n_i + \alpha + \beta}$$

$$\mathbb{E}[\theta_i \mid X] = \mathbb{E}\left[\, \mathbb{E}[\theta_i \mid X, \alpha, \beta] \mid X \right]$$

$$\iint \frac{X_i + \alpha}{n_i + \alpha + \beta} \lambda(\alpha, \beta \mid X)\, d\alpha\, d\beta$$

If m large α, β may be "almost known"

$\rightsquigarrow$ choice of λ doesn't matter much

Flexibility of Bayes

Any Λ, $\mathcal{P}$, L, $g(\theta)$: δ_Λ defined straightforwardly

$$\delta_\Lambda(x) = \arg\min_d \int L(\theta, d)\, \lambda(\theta | x)\, d\theta$$

Problem reduced to (possibly hard) computation

Posterior is "one stop shop" for all answers

No need for:

- special family structure (exp. fam. / complete s.s.)
- special estimator (u-estimable)
- convex or nice L

$\Rightarrow$ Highly expressive modeling & estimation

Caveat: Limited by ability to do computations (Topic of next lecture)

Source #4: Convenience Priors

Choosing conjugate or other "nice" priors

$\rightsquigarrow$ much faster computations esp. in high-dim.

(But what does the posterior mean?)

Ex. $X_1, \ldots, X_n \overset{iid}{\sim} p$, p unknown density on $\mathbb{R}$

Estimand: $m = \text{median}(p)$

Estimator $\delta(X) = \text{median}(X)$

good estimator: robust, nonparametric

large n: $\delta(X) \approx N\left(m, (4np(m))^{-1}\right)$

not Bayes for any realistic prior

Bayes approach

Step 1. Define prior over p (infinite-dim!)

Step 2. Calculate posterior

Horrific unless we pick special prior

Step 3. Return e.g. $\mathbb{E}[m \mid X]$

If it differs substantially from median(X), do we trust it?

Gaussian Hierarchical Model:

$$\tau^2 \sim \lambda_0$$

$$\theta_i \mid \tau^2 \overset{iid}{\sim} N(0, \tau^2) \qquad i \le d$$

$$X_i \mid \tau^2, \theta \overset{ind.}{\sim} N(\theta_i, 1)$$

Posterior mean:

$$\delta(X_i) = \mathbb{E}\left[\theta_i \mid X\right]$$

$$= \mathbb{E}\left\{ \mathbb{E}\left[\theta_i \mid X, \tau^2\right\} \mid X\right]$$

$$= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2} X_i \mid X\right]$$

$$= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2} \mid X\right] \cdot X_i$$

Linear shrinkage estimator,

Bayes-optimal shrinkage estimated from data

Likelihood for τ^2 : marginalize over θ_i

$$X_i \mid \tau^2 \sim N(0, 1+\tau^2)$$

$$\Rightarrow \frac{1}{d}\|X\|^2 \sim \frac{1+\tau^2}{d} \chi^2_d$$

$$\sim \left(1+\tau^2, \frac{2+2\tau^2}{d}\right)$$

notation (mean, variance)

Define $\zeta(\tau^2) = \frac{1}{1+\tau^2}$

$$\Rightarrow \delta(X) = \left(1 - \mathbb{E}[\zeta \mid X]\right) X_i$$

Conjugate prior:

$$v = \|X\|^2 \mid \zeta \sim \frac{1}{\zeta} \chi^2_d = \frac{\zeta^{d/2}}{\Gamma(d/2)} v^{\frac{d}{2}-1} e^{-\zeta v}$$

$$\zeta \sim \frac{1}{s} \chi^2_k = \frac{s^{k/2}}{\Gamma(k/2)} \zeta^{k/2 - 1} e^{-s\zeta}$$

$$\Rightarrow \zeta \mid \|X\|^2 \propto_\zeta \zeta^{\frac{k+d}{2}-1} e^{-(s+\|X\|^2)\zeta}$$

$$\sim \frac{1}{s+\|X\|^2} \chi^2_{k+d}$$

$$\mathbb{E}\left[\zeta \mid \|X\|^2\right] = \frac{k+d}{s+\underbrace{\|X\|^2}_{\approx d(1+\tau^2) + O(d^{1/2})}}$$

[might want to truncate prior to $[0,1]$ if d small]