Math 341: Bayesian Modeling

Darshan Patel

Spring 2017

Definition 0.1. Random Variable: realizes to a data "x," denoted by X

Definition 0.2. Supports: all possible realization values, denoted by Supp(X)

Note: Real variables have "supports."

Two Types of Random Variables:

- Discrete:
$$|\text{Supp}\,[X]\,| \leq |\mathbb{N}|$$
where it is countable,
If Supp(X) = 1, then $X \sim \text{Deg}(c) = \{1 \text{ outcome}\}$.

  There exists $p(x) = P(X = x)$ called the probability mass function or pmf which relates $\text{Supp}(X) \to \big(0, 1\big)$.

  $F(x) = P(X \leq x)$ is called the cumulative density function (cdf)

- Continuous:
$$|\text{Supp}\,[X]\,| \leq |\mathbb{R}|$$
There exists $f(x) = F'(x)$ called the probability density function (pdf) where f : $\text{Supp}\,[X] \to \big(0, 1\big)$. The cumulative density function is denoted $P(X \in \big[a, b\big])$ which is equal to
$$\int_a^b \underbrace{f(x)}_{F'(x)}\,dx = F(b) - F(a)$$

Note: Discrete random variables are defined by their pmf and cdf whereas continuous random variables are defined by their pdf and cdf.

Types of Distributions:

- Discrete
  - $X \sim \text{Bern}(x) = p^x(1-p)^{1-x}$ where $x \in \text{Supp}[X] = \{0, 1\}$.
  - $X \sim \text{Bern}(n, x) = \binom{n}{p} p^x 1 - p^{1-x}$ where $x \in \text{Supp}[X] = \{0, 1, 2, \dots, n\}$.
- Continuous
  - $X \sim \text{Exp}(\lambda) = \lambda e^{-\lambda x}$ where $x \in \text{Supp}[X] = [0, \infty)$.
  - $X \sim \text{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ where $x \in \text{Supp}[X] = \big(-\infty, \infty\big)$.

From now on, parameters will be denoted by θ and parameter spaces will be denoted Θ (capital θ). This transforms the above distributions to the following:

- $X \sim \text{Bern}(\theta) = \theta^x(1-\theta)^{1-x}$
- $X \sim \text{Bern}(n, \theta) = \binom{n}{x} \theta^x 1 - \theta^{1-x}$
- $X \sim \text{Exp}(\theta) = \theta e^{-\theta x}$
- $X \sim \text{N}(\theta_1, \theta_2^2) = \frac{1}{\sqrt{2\pi\theta_2^2}} e^{-\frac{1}{2\theta_2^2}(x-\theta_1)^2}$

Definition 0.3. Parametric Models: a set of random variable models with finite parameters, denoted by $\mathcal{F}$

$$\mathcal{F} : \{p(x;\theta) : \theta \in \Theta\}$$

where $p(x;\theta)$ is the probability of assuming the value of the parameter θ.

Example 0.1. Let's say we want to model the parameters for a normal distribution. We can represent this as follows:

$$\hat{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$$

Note: Parametric models can be either pmf or pdf.

If $x_1, x_2, \dots, x_n$ are realizable, then

$$p(x_1, x_2, \dots, x_n; \theta) = p(x_1;\theta)p(x_2;\theta)\dots p(x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta)$$

In the real world, let's say we "observe" data as follows: $x = \langle 0, 0, 1, 0, 1, 0 \rangle$ and we assume IID. Then you pick a parametric model, $\mathcal{F}$, but θ is not known. Figuring out θ is the point of statistical inference.

Three Main Types:

- Point Estimation: best guess of θ
- Confidence Set: a set of "likely" θ's
- Theory Testing: θ value testing, also called hypothesis testing

Let's say we assume a Bernoulli distribution for the data set $x = \langle 0,0,1,0,1,0\rangle$. Then

$$p(0,0,1,0,1,0) = \prod_{i=1}^{6} \theta^x (1-\theta)^{1-x}$$

For example. let's take $\theta = \frac{1}{2}$, then

$$p(x_1, x_2, \dots, x_6; \frac{1}{2}) = 0.5^6 = 0.0156$$

Let's take $\theta = \frac{1}{4}$, then

$$p(x_1.x_2.\dots, x_6; \frac{1}{4}) = (\frac{1}{4})^2 (\frac{3}{4})^4 = 0.0198$$

Out of the two choices for θ, the second one is more likely since the second model has a higher probability than the first one. But we can take an infinite number of guess for θ. There has to be a better way to figure out θ.

Definition 0.4. Likelihood Function:

$$p(x_1, x_2, \dots, x_n; \theta) = \mathcal{L}(\theta; x_1, x_2, \dots, x_n)$$

where the joint density function on the left hand side is in perspective of $x_1, x_2, \dots, x_n$ and allowing it to change whereas the likelihood function on the right hand side is in perspective of θ and allowing it to change.

To get the best model, we must optimize argmax$\{\mathcal{L}(\theta; x_1, x_2, \dots, x_n)\}$.

Definition 0.5. $\hat{\theta}_{MLE}$: maximum likelihood estimate or maximum likelihood estimate, must be within Θ

Example 0.2. If $f(x) = 1 - x^2$, then $\max\{f(x)\} = 1$ but argmax$\{f(x)\} = 0$.

Note: If you taken an increasing 1-1 function of $\mathcal{L}$, then θ_{MLE} won't change.

Example 0.3. Let $l(\theta; x_1, x_2, \dots, x_n) = \ln(\mathcal{L}(\theta; x_1, x_2, \dots, x_n))$ be a log-likelihood function. Then

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\text{argmax}}\{l(\theta; x_1, x_2, \dots, x_n)\}$$

or

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\text{argmax}}\{\ln(\mathcal{L}(\theta; x_1, x_2, \dots, x_n))\}$$

Example 0.4. Let $x_1, \dots, x_6 \overset{iid}{\sim} \text{Bern}(\theta)$ be the data set $\langle 0, 0, 1, 0, 1, 0 \rangle$. Then:

$$\begin{aligned}
l(\theta; x) &= \ln(\prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{1-x_i}) \\
&= \sum_{i=1}^{6} \ln(\theta^{x_i}(1-\theta)^{1-x_i}) \\
&= \sum_{i=1}^{6} x_i \ln(\theta) + (1-x_i)\ln(1-\theta) \\
&= \ln(\theta)\sum_{i=1}^{6} x_i + (6 - \sum_{i=1}^{6} x_i)\ln(1-\theta) \\
&= \ln(\theta)6\bar{x} + (6 - 6\bar{x})\ln(1-\theta) \\
&= 6(\bar{x}\ln(\theta) + (1-\bar{x})\ln(1-\theta))
\end{aligned}$$

Now let's differentiate this to maximize it:

$$\frac{d}{dt}6(\bar{x}\ln(\theta) + (1-\theta)\ln(1-\theta)) = 6(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta})$$

If we set it equal to 0,

$$(1-\theta)\bar{x} - \theta(1-\bar{x}) = 0 \to \hat{\theta}_{MLE} = \bar{x}$$

Note: For our convenience, we use the natural log to differentiate $\prod$ to $\sum$. It is easier to differentiate sums rather than products.

Definition 0.6. Maximum Likelihood Estimation: $\hat{\theta}_{MLE} = \bar{X}$ where $\bar{X}$ is a random variable and has properties

Definition 0.7. Maximum Likelihood Estimate: $\hat{\theta}_{MLE} = \bar{x}$ where $\bar{x}$ has a numerical value

Example 0.5. Let $x_1, \dots, x_n \overset{iid}{\sim} \text{Geom}(\theta) = (1-\theta)^x\theta$ where x is the number of failures before stopping success. $\text{Supp}(X) = \{0, 1, \dots\} = \mathbb{N}$ and $\Theta = (0, 1)$. Then:

$$\begin{aligned}
p(x_i, \dots, x_n) &= \mathcal{L}(\theta; x_i, \dots, x_n) \\
&= \prod_{i=1}^{n}(1-\theta)^{x_i}\theta
\end{aligned}$$

Therefore

$$\begin{aligned}
l(\theta; x) &= \sum \ln(1-\theta)^{x_i}\theta \\
&= \ln(1-\theta)\sum x_i + n\ln(\theta)
\end{aligned}$$

We will now differentiate this function to solve for $\hat{\theta}_{MLE}$.

$$\begin{aligned}
l'(\theta; x) &= \frac{n}{\theta} - \frac{n\bar{x}}{1-\theta} = 0 \\
\frac{1}{\theta} &= \frac{\bar{x}}{1-\theta} \\
\frac{1}{\theta - 1} &= \bar{x} \\
\hat{\theta}_{MLE} &= \frac{1}{\bar{x}+1}
\end{aligned}$$

Properties of MLE:

1. Consistency: there exists $\varepsilon > 0$ such that
$$\lim_{n\to\infty} P(|\hat{\theta}_{MLE} - \theta| \geq \varepsilon) = 0$$

2. Asymptotic Normaling: As n increases, the the parameters behave like a normal distribution
$$\hat{\theta}_{MLE} \xrightarrow{d} N(\hat{\theta}_{MLE}, SE(\hat{\theta}_{MLE})^2)$$

3. Efficiency: $\hat{\theta}_{MLE}$ has the lowest standard error theoretically possible

Inference with MLE:

- Point Estimate: $\hat{\theta}_{MLE}$
- Confidence Set: $CI_{\theta,1-\alpha} = [\hat{\theta}_{MLE} \pm z_{\frac{\alpha}{2}} SE(\hat{\theta}_{MLE})]$
  Here, θ is the parameter of interest whereas $1-\alpha$ is the confidence level.
- Hypothesis Testing: $H_0 : \theta = \theta_0$, $H_A : \theta \neq \theta$ - fail to reject if $\hat{\theta}_{MLE}$ is in the region of $[\theta_0 \pm z_{\frac{alpha}{2}} SE(\hat{\theta}_{MLE})]$

We must observe data, then pick a parametric model $\mathcal{F}$, do inference with MLE. The problem with this is that

1. If all data values taken are 0 and we take $\mathcal{F} = \text{Bern}(\theta)$, then $\hat{\theta}_{MLE} = \bar{x} = 0$ and $SE(\bar{\theta}_{MLE}) = \sqrt{\bar{\theta}_{MLE}(1-\bar{\theta}_{MLE})} = 0$. This gives no information and thus is a big problem. No confidence set, no hypothesis testing.

2. What if we have prior knowledge about Θ? We can't use it because only data set can be used.

3. Frequentist Confidence Interval Interpretation: Let's say we found $CI_{\theta,1-\alpha} = [0.42, 0.47]$. If the experiment is repeated "many" times, then a confidence level of 95% will cover θ and $1-\alpha$ is contained in the set. But given just an interval, we can only say that a certain value will either fall in the interval or not. We can't claim that the probability that the interval contains θ is $1-\alpha$.

4. Hypothesis testing: not satisfactory since we do not know if data values are far from being retained yet rejected or near rejection (extremeness). How good is the rejection? What is $P(H_0|x)$, or H_0 given x?

5. Boundary Issues: Let's say $x = \langle 0,0,1,0,1,0\rangle$ and $\hat{\theta}_{MLE} = \frac{1}{3}$. We want a confidence set at the 95% confidence level: $CI_{\theta,95\%} = (\frac{1}{3} \pm 2\sqrt{\frac{1}{3}\frac{2}{3}}) = (-0.6, 1.26)$. In this confidence interval, we have both a negative value and one that's greater than 1. This is no good. This happened because our data set is only composed of 6 values. Thus it cannot converge to normality. We cannot use the normal distribution to construct the interval and since we did, it came out looking wrong.

Good news: The Bayesian approach will not cause any of these issues.

Definition 0.8. Conditional Probability: $P(B|A)$, the probability of B occurring given A occurs

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

Note: There is a proportionality between $P(A,B)$, the intersection of two events, and $P(B|A)$, the probability of B occurring given A occurs. Thus we can write

$$P(A,B) \propto P(B|A)$$

or

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

Definition 0.9. Baye's Rule:

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

We know from previous probability courses that $P(A,B) = P(B,A)$. We also know that $P(A,B) = P(B|A)P(A)$ and $P(B,A) = P(A|B)P(B)$. Let's set them equal to each other.

$$\begin{aligned} P(A,B) &= P(B,A) \\ P(B|A)P(A) &= P(A|B)P(B) \end{aligned}$$

This is another form of Baye's rule.

Definition 0.10. Law of Total Probability: the probability of event A occurring is sum of the probability of the intersection of event A and event B and the probability of the intersection of event A and not event B (complement of B)

$$P(A) = P(A,B) + P(A,B^C)$$

Let's combine the two equations from above.

$$\begin{aligned} P(A) &= P(A,B) + P(A,B^C) \\ &= P(A|B)P(B) + P(A|B^C)P(B^C) \\ P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)} \end{aligned}$$

This is another form of Baye's rule.

Note:

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

The LHS is the posterior probability where B is the parameter of interest, A is the evidence/data, and $B|A$ is the targeted estimation. On the RHS, $P(A|B)$ is the likelihood or probability of data/effect and $P(B)$ is a prior probability, a prior model or theory.

Finding $P(B|A)$ using A(data) and applying it to $P(B)$ is called Bayesian conditionalism.

Definition 0.11. Law of Total Probability: Let $B_1, \dots, B_k$ be mutually exclusive events and collectively exhaustive. Then

$$P(A) = \sum_{i=1}^{k} P(A, B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Theorem 0.1. Baye's Theorem:

$$P(B|A) = \frac{P(A|B)P(B)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$$

Definition 0.12. Bayesian Conditionalism is taking $P(B)$, adding A, or data, to it, to find $P(B|A)$

Another way to think about probability of A is: Odds$(A) := \frac{P(A)}{P(A^C)} = \frac{P(A)}{1-P(A)}$.

Example 0.6. Let's say an event has an odds of 4, or "4 to 1" odds. Then the event has a probability of occurring of 0.8 since for each 4 +1, or 5, chances, the odds of it occurring is 4.

Note: To get odds against,

$$\text{Odd}(A)^{-1} = \frac{P(A^C)}{P(A)} = \frac{1-P(A)}{P(A)}$$

Example 0.7. Let A represent the event of a person being a smoker and B be the event that a person has lung cancer.

$$P(A) = 0.2, P(B) = 0.0.06, P(A, B) = 0.036$$

Then $P(A|B) = \frac{P(A,B)}{P(B)} = \frac{0.36}{0.06} = 0.06$. That's easy.

$$P(A|B^C) = \frac{P(A, B^C)}{P(B^C)} = \frac{P(A) - P(A, B)}{1 - P(B)} = \frac{0.2 - 0.036}{1 - 0.06} = 0.174$$

What's the ratio of $\frac{P(B|A)}{P(B^C)|A}$? Well we know, $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ and $P(B^C|A) = \frac{P(A|B^C)P(B^C)}{P(A)}$. Thus,

$$\underbrace{\frac{P(B|A)}{P(B^C|A)}}_{\text{posterior odds}} = \overbrace{\frac{P(A|B)}{P(A|B^C)}}^{\text{likelihood ratio}} \left(\overbrace{\frac{P(B)}{P(B^C)}}^{\text{prior odds}} \right)$$

Plugging in the numbers, that gives us

$$\frac{P(B|A)}{P(B^C|A)} = \frac{0.6}{0.174}\left(\frac{0.06}{0.94}\right) = 0.22$$

This tells us that the odds of getting lung cancer given that a person smokes is 0.22.

Let X, Y be two random variables. We can represent the joint probability mass function as follows:

$P(X=x, Y=y)$				Supp(Y)		
Supp(X)		1	2	3	4	5
	1					
	2					
	3					
	4					
	5					

Then

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

This is the shorthand form of

$$P(X=x|Y=y) = \frac{P(Y=y|X=x)P(X=x)}{P(Y=y)}$$

For this specific joint PMF,

$$P(Y=y) = P(Y=1|X=1) + \cdots + P(Y=1|X=5)$$

In general,

$$P(Y=y) = \sum_{x \in \text{ Supp}(X)} P(Y=y|X=x) = \sum_{x \in \text{ Supp}(X)} P(Y=y|X=x)P(X=x)$$

This is called marginalization, where we are margining out x.

For a probability density function,

$$f_Y(y) = \int_{x \in \text{ Supp}(X)} f(x,y)\,dx = \int_{x \in \text{ Supp(X)}} f_{y|x} f(x)\,dx$$

Consider $P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$ where x is the data and θ is the parameter of a model where $\mathcal{L}(\theta; X) = P(X;\theta)$. The LHS is the probability of cause given effect whereas $P(X|\theta)$ is the probability of effect given cause. We say $P(\theta) = \text{Deg}(\theta_0) = \{0, 1\}$. We don't know what θ is exactly so $P(\theta)$ is degenerate. Also, for $P(X)$, we can't find the probability of the data values X without knowing θ. If we did, then $P(X) = \sum_{\theta \in \Theta} P(X|\theta_0)P(\theta_0)$. But $P(\theta_0)$ can only be zero or one (in the case $\theta_0 = \theta$). Thus $P(X) = P(X|\theta)$. This problem began when we assumed $P(\theta)$ is 0 or 1. There was only one true value of θ, call it θ_0.

In the frequentist approach, $P(\theta)$ is degenerate, In the Bayesian approach, we allow $P(\theta)$ to repress our prior knowledge, or prior information.

In the Bayesian approach,

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\sum_{\theta_i \in \Theta} P(X|\theta_i)P(\theta_i)} = \frac{P(X|\theta)P(\theta)}{\int_{\theta_i \in \Theta} P(X|\theta_i)P(\theta_i)\, d\theta_i}$$

Example 0.8. Let's assume $\mathcal{F}$ is a Bernoulli model where $X = \langle 0, 1, 1 \rangle$ and assume IID. If we estimate θ to be 0.75,

$$P(X|\theta = 0.75) = 0.25 \times 0.75^2 = 0.141$$

If we estimate θ to be 0.25,

$$P(X|\theta = 0.25) = 0.75 \times 0.25^2 = 0.047$$

Here we assumed $\Theta = \{0.25, 0.75\}$. But what's $P(\theta = 0.75|X)$?

$$P(\theta = 0.75|X) = \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X)}$$

We know that $P(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.25 \\ 0.5 & \text{if } \theta = 0.75 \end{cases}$. This is the principle of inference; we take all models to be equally likely. Then

$$\begin{aligned} P(\theta = 0.75|X) &= \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X)} \\ &= \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X|\theta = 0.75) + P(X|\theta = 0.25)} \\ &= \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X|\theta = 0.75)P(\theta = 0.75) + P(X|\theta = 0.25)P(\theta = 0.25)} \\ &= \frac{0.141 \times 0.5}{0.141 \times 0.5 + 0.047 \times 0.5} \\ &= 0.75 \end{aligned}$$

If we know this, what is $P(\theta = 0.25|X)$?

$$P(\theta = 0.25|X) = 1 - P(\theta = 0.75|X) = 1 - 0.75 = 0.25$$

Let X and θ be two random variables having a joint distribution. The "dim space" (of all possible realizations) if X can be 0 or 1 and there's three trials is:

$$x \in X = \{\langle 0,0,0\rangle, \langle 0,0,1\rangle, \langle 0,1,0\rangle, \langle 1,0,0\rangle, \langle 0,1,1\rangle, \langle 1,0,1\rangle, \langle 1,1,0\rangle, \langle 1,1,1\rangle\}$$

Then

$$\begin{aligned} P(x = \langle 0,0,0\rangle, \theta = 0.25) &= P(x = \langle 0,0,0\rangle | \theta = 0.25)P(\theta = 0.25) \\ &= 0.75^3 \times 0.5 = 0.211 \\ P(x = \langle 1,0,0\rangle, \theta = 0.25) &= 0.25 \times 0.75^2 \times 0.5 = 0.070 \\ P(x = \langle 1,1,0\rangle, \theta = 0.25) &= 0.25^2 \times 0.75 \times 0.5 = 0.023 \\ P(x = \langle 1,1,1\rangle, \theta = 0.25) &= 0.25^3 \times 0.5 = 0.008 \end{aligned}$$

What if we want to do it for the case where $\theta = 0.75$? Then $P(\langle 0,0,0\rangle, \theta = 0.75) = 0.008$. In fact, it'll be all the above probabilities, but reversed.

Is θ independent of X? No. Knowing θ tells you something about X and known x tells you something about θ.

Let's look at the case where $\Theta = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Then $P(\theta) = \begin{cases} 0.2 & \text{if } \theta = 0.1 \\ 0.2 & \text{if } \theta = 0.25 \\ 0.2 & \text{if } \theta = 0.5 \\ 0.2 & \text{if } \theta = 0.75 \\ 0.2 & \text{if } \theta = 0.9 \end{cases}$.

Let $X = \langle 0,1,1\rangle$. Then

$$\begin{aligned} P(X|\theta = 0.1) &= 0.09 \\ P(X|\theta = 0.25) &= 0.047 \\ P(X|\theta = 0.5) &= 0.125 \\ P(X|\theta = 0.75) &= 0.141 \\ P(X|\theta = 0.9) &= 0.061 \end{aligned}$$

What we have found that is that

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \Big(\frac{1}{P(X)}\Big)P(X|\theta)P(\theta) \propto P(X|\theta)P(\theta) \propto P(X|\theta)$$

We have previously calculated that $\hat{\theta}_{MLE} = 0.66$ for $x = \langle 0,1,1\rangle$ using the point estimate. But according to our best guess here, it is 0.75.

Let $\mathcal{F}$ be Bernoulli where $x = \langle 0,1,1\rangle$ and $\Theta = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ ($\theta \sim U(\Theta_0)$, discrete uniform). We want $P(\theta|X)$, the probability of likelihood. If we use Θ, we find

$$\begin{aligned} P(X|\theta = 0.1) &= 0.09 \\ P(X|\theta = 0.25) &= 0.047 \\ P(X|\theta = 0.5) &= 0.125 \\ P(X|\theta = 0.75) &= 0.141 \\ P(X|\theta = 0.9) &= 0.061 \end{aligned}$$

The best model here is the biggest slice, $\theta = 0.75$.
Idea to find "best" θ:

$$\hat{\theta}_{\text{MAP}} = \underset{\theta \in \Theta_0}{\text{argmax}}\{P(\theta|x)\}$$

where $\hat{\theta}_{\text{MAP}}$ is the maximum a posterior or posterior mode. Let's simplify it.

$$\begin{aligned}
\hat{\theta}_{\text{MAP}} &= \underset{\theta \in \Theta_0}{\text{argmax}}\{P(\theta|x)\} \\
&= \underset{\theta \in \Theta_0}{\text{argmax}}\{\frac{P(X|\theta)P(\theta)}{P(X)}\} \\
&= \underset{\theta \in \Theta_0}{\text{argmax}}\{P(X|\theta)P(\theta)\} \text{ } (P(X) \text{ is a constant and not based on } \theta) \\
&= \underset{\theta \in \Theta_0}{\text{argmax}}\{P(X|\theta)\} \text{ } (P(\theta) \text{ is a constant due to principle of indifference}) \\
&= \hat{\theta}_{\text{MLE}}
\end{aligned}$$

We find that

$$\begin{aligned}
P(\theta|X) &= P(X|\theta) \overbrace{P(\theta)}^{*} \overbrace{\frac{1}{P(X)}}^{**} \\
&= \frac{P(X|\theta)P(\theta)}{P(X)} \\
&= \frac{P(X|\theta)P(\theta)}{\sum_{\theta_0 \in \Theta} P(X, \theta_0)} \\
&= \frac{P(X|\theta)P(\theta)}{\sum_{\theta_0 \in \Theta} P(X|\theta_0)P(\theta_0)} \\
&\text{under principle of indifference} \\
&= \frac{P(X|\theta)}{P(X|\theta_1) + \cdots + P(X|\theta_m)} \text{ where } m = |\Theta|
\end{aligned}$$

In the above, $*$ is a scale by prior belief and $**$ is a normalization constant so that all $P(\theta|X)$'s add up to 1. In the Bernoulli model for $x = \langle 0, 1, 1 \rangle$,

$$P(\theta = 0.75|X) = \frac{0.141}{0.009 + 0.047 + 0.125 + 0.141 + 0.061} = \frac{0.141}{0.363} = 0.38$$

Thus we found that if $\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}$, then $0.75 = 0.66$ which is absurd. This is because our prior did not cover the entire parameter space ($\Theta_0 \neq \Theta = (0, 1)$).
Main reason to be skeptic: prior could be wrong!

Let's say $\Theta = \{0.25, 0.75\}$ and $x = \langle 0, 1, 1 \rangle$ and we assumed $\mathcal{F}$ is a Bernoulli model. Then for $x_1 = 0$:

$$P(\theta = 0.25|X_1 = 0) = \frac{P(X_1 = 0|\theta = 0.25)}{P(X_1 = 0|\theta = 0.25) + P(X_1 = 0|\theta = 0.75)} = \frac{0.75}{0.75 + 0.25} = 0.75$$

If $P(\theta = 0.25|X_1) = 0.75$, then it is clear that $P(\theta = 0.75|X_1 = 0) = 0.25$.
Now let's look at $X_2 = 1$. Let's let our prior be its posterior from the previous data. Then

$$\begin{aligned}
&P(\theta = 0.25|X_2 = 1) \\
&\quad = \frac{P(X = 1|\theta = 0.25)P(\theta = 0.25|X_1 = 0)}{P(X_2 = 1|\theta = 0.25)P(\theta = 0.25|X_1 = 0) + P(X_2 = 1|\theta = 0.75)P(\theta = 0.75|X_1 = 0)} \\
&\quad = \frac{0.25 \cdot 0.75}{0.25 \cdot 0.75 + 0.75 \cdot 0.25} = 0.5
\end{aligned}$$

In the similar logic as before, $P(\theta = 0.75|X_2 = 1) = 0.5$.
Now let's look at $X_3 = 1$.

$$\begin{aligned}
&P(\theta = 0.25|X_3 = 1) = \\
&\frac{P(X_3 = 1|\theta = 0.25)P(\theta = 0.25|X_1 = 0, X_2 = 1)}{P(X_3 = 1|\theta = 0.25)P(\theta = 0.25|X_1 = 0, X_2 = 1) + P(X_3 = 1|\theta = 0.75)P(\theta = 0.75|X_1 = 0, X_2 = 1)} \\
&\quad = \frac{0.25 \cdot 0.5}{0.25 \cdot 0.5 + 0.75 \cdot 0.5} = 0.25
\end{aligned}$$

In fact, this result is indeed $P(\theta = 0.25|X = \langle 0, 1, 1\rangle)$.

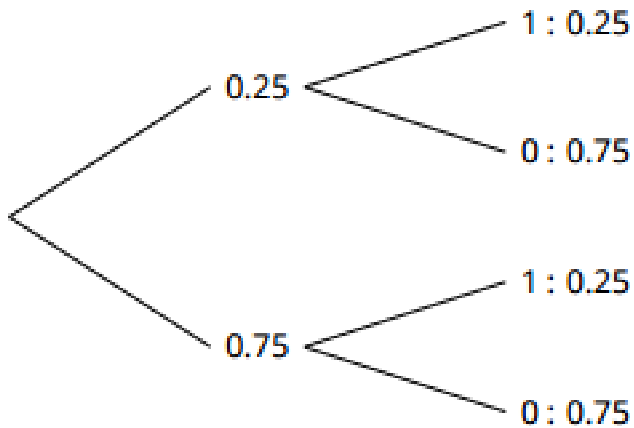
Proof.

$$\begin{aligned}
P(\theta|X_1, \dots, X_n) &= \frac{P(X_1, \dots, X_n|\theta)P(\theta)}{P(X_1, \dots, X_n)} \\
&= \frac{P(X_n|\theta) \cdot \dots \cdot P(X_2|\theta)P(X_1|\theta)P(\theta)}{P(X_n, \dots, X_2|X_1)P(X_1)} = P(\theta|X_1) \\
&= \frac{P(X_n|\theta) \cdot \dots \cdot P(X_3|\theta)P(X_1, X_2|\theta)P(\theta)}{P(X_n, \dots, X_3|X_1, X_2)P(X_1, X_2)} = P(\theta|X_1, X_2) \text{ and keep going forward}
\end{aligned}$$

□

Using the same model as before, let's introduce X^*, the next unseen observation. What is its distribution? $X \sim \text{Bern}(?)$.
Based on the frequentist approach, $P(X^*|X_1, X_2, X_3) \approx P(X^*|\theta = \hat{\theta}_{\text{MLE}}) = \text{Bern}(0.66)$. But $\hat{\theta}_{\text{MLE}}$ is inaccurate and does not account for uncertainty. Thus we must use a posterior

predictive distribution: $P(X^*|X_1, X_2, X_3)$.



In this tree diagram, we assign the same probabilities to the possible outcomes of X^*(0 or 1) that we found for $X_1.X_2.X_3$. This gives:

$P(X^*\|X_1, X_2, X_3)$
$0.25 \cdot 0.25 = 0.0625$
$0.25 \cdot 0.75 = 0.1875$
$0.75 \cdot 0.25 = 0.1875$
$0.75 \cdot 0.75 = 0.5625$

For example, $P(X^* = 1|X_1, X_2, X_3) = 0.0625 + 0.5625 = 0.625$ and so $X^*|X_1, X_2, X_3 \sim$ Bern(0.625). What we did here was that we used the posterior to predict the next and add up the probabilities. We incorporated all uncertainties of θ assuming the prior.

Marginalization:

$$\begin{aligned}
P(X^*|X_1, X_2, X_3) &= \sum_{\theta \in \Theta_0} P(X^*, \theta|X_1, X_2, X_3) \\
&= \sum_{\theta \in \Theta_0} P(X^*|\theta, X_1, X_2, X_3)P(\theta|X_1, X_2, X_3) \\
&= \sum_{\theta \in \Theta_0} P(X^*|\theta)P(\theta|X_1, X_2, X_3) \\
&= \sum_{\theta \in \Theta_0} P(X^*|\theta)P(\theta|X_1, X_2, X_3) \\
&= \sum_{\theta \in \Theta_0} P(X^*|\theta)\frac{P(X_1, X_2, X_3|\theta)P(\theta)}{P(X_1, X_2, X_3)}
\end{aligned}$$

What this is saying is that we look at all possible models and average them. Thus,

$$P(X^*|X_1, X_2, X_3) = \sum_{\theta \in \Theta_0} P(X^*|\theta)\frac{P(X_1, X_2, X_3|\theta)P(\theta)}{P(X_1, X_2, X_3)}$$

Procedure for Posterior Predictive Distribution:

1. Draw θ from posterior
2. Examine $X^*|\theta$
3. Repeat for all θ's and average them up

Proof.

$$\begin{aligned} P(X^*|\theta) &= P(X^*|\theta, X_1, X_2, X_3) \\ &= \frac{P(X^*, X_1, X_2, X_3, \theta)}{P(X_1, X_2, X_3, \theta)} \\ &= \frac{P(X^*, X_1, X_2, X_3|\theta)P(\theta)}{P(X_1, X_2, X_3|\theta)P(\theta)} \\ &= \frac{P(X^*|\theta)P(X_1|\theta)P(X_2|\theta)P(X_3|\theta)}{P(X_1|\theta)P(X_2|\theta)P(X_3|\theta)} \\ &= P(X^*|\theta) \end{aligned}$$

□

In general,

$$P(X^*|X_1, \dots, X_n) = \sum_{\theta \in \Theta_0} P(X^*|\theta)P(\theta|X_1, \dots, X_n) = \int_{\theta \in \Theta_0} P(X^*|\theta_0)P(\theta_0|X_1, \dots, X_n)\, d\theta$$

Note: $P(X^*|X_1, \dots, X_n) \neq P(X^*|\hat{\theta}_{\text{MLE}})$.

What we have now found is that if $\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}$, then $0.75 = 0.66$. This is still inaccurate. This is because Θ_0 does not cover $\Theta = (0, 1)$.

What prior should we use? Supp(θ) = parameter space of $\mathcal{F}$ = (0, 1).
Idea: Let $\theta \sim U(0, 1)$ where all numbers from 0 to 1 are equally likely.

Let $X = \langle 0, 1, 1 \rangle$. Then

$$P(\theta|X) = P(X|\theta)\frac{P(\theta)}{P(X)} \propto P(X|\theta)$$

if $\hat{\theta}_{\text{MAP}}$ matters. In this example,

$$P(\theta|X) = (1 - \theta)(\theta)(\theta) = \theta^2 - \theta^3$$

Then

$$\hat{\theta}_{\text{MAP}} = \operatorname*{argmax}_{\theta \in \Theta}\{P(\theta|X)\} = \operatorname*{argmax}_{\theta \in \Theta}\{P(X|\theta)\}(\text{ if principle of indifference}) = \operatorname*{argmax}_{\theta \in \Theta}\{\theta^2 - \theta^3\}$$

To find the maximum of that function, differentiate it and set it equal to 0.

$$\frac{d}{d\theta}(\theta^2 - \theta^3) = 2\theta - 3\theta^2$$

If we set it equal to 0, we find that $\hat{\theta}_{\text{MAP}} = 0.67$ which is $\hat{\theta}_{\text{MLE}}$.

What about $P(\theta = [0.6, 0.7]|X)$?

$$P(\theta = [0.6, 0.7]|X) = \int_{0.6}^{0.7} P(\theta|X)\,d\theta$$

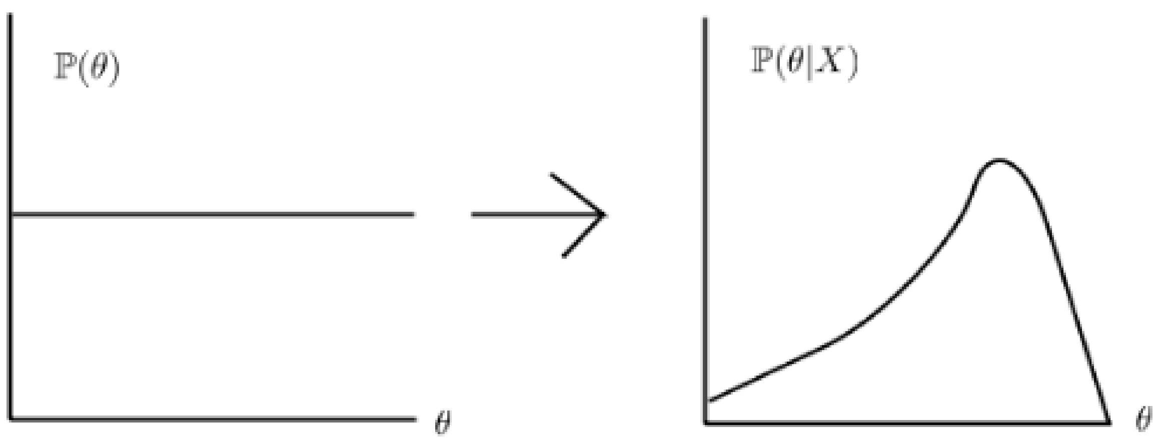
$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{\theta^2 - \theta^3}{\int_0^1 P(X|\theta)P(\theta)\,d\theta} = \frac{\theta^2 - \theta^3}{\int_0^1 (\theta^2 - \theta^3)\,d\theta} = 12(\theta^2 - \theta^3)$$

Thus

$$\int_{0.6}^{0.7} 12(\theta^2 - \theta^3)\,d\theta = 0.1765 = P(\theta = [0.6, 0.7]|X)$$

All this is saying is that the probability θ is between 0.6 and 0.7 is 0.1765, assuming the prior.

We let $\mathcal{F}$ be Bernoulli with $X = \langle 0, 1, 1\rangle$ and $\theta \sim U(0,1)$. This means that we give equal weightage to all values for θ in between 0 and 1. If $\mathbb{P}(\theta \mid X) = 12\theta^2(1-\theta)$, then we went from $\mathbb{P}(\theta)$, the prior distribution, to $\mathbb{P}(\theta \mid X)$, the posterior distribution, or,



This shows a skewness towards 1 because $\hat{\theta}_{\text{MAP}} = \frac{2}{3} = \hat{\theta}_{\text{MLE}}$.

Note: Under the principle of indifference,

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}$$

Let $\mathcal{F}$ be Bernoulli with $X = \langle 0, 1, 1\rangle$ and $\theta \sim U(0,1)$. Then

$$\overbrace{\mathbb{P}\left(\theta \mid X\right)}^{\text{all data}} = \frac{\mathbb{P}\left(X \mid \theta\right)\mathbb{P}\left(\theta\right)}{\mathbb{P}\left(X\right)} = \frac{\mathbb{P}\left(X \mid \theta\right)\mathbb{P}\left(\theta\right)}{\int_{\Theta_0}\mathbb{P}\left(X \mid \theta\right)\mathbb{P}\left(\theta\right)\,d\theta}$$

where $\mathbb{P}\left(\theta\right) = 1$. Then, for this model,

$$\begin{aligned}\mathbb{P}\left(X \mid \theta\right) &= \prod_{i=1}^{n}\mathbb{P}\left(x_i \mid \theta\right)\\ &= \prod_{i=1}^{n}\theta^{x_1}(1-\theta)^{1-x_i}\\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\\ &= \theta^{x}(1-\theta)^{n-x} \text{ where } x = \sum_{i}^{n} x_i\end{aligned}$$

Plugging this back into $\mathbb{P}\left(\theta \mid X\right)$ gives:

$$\mathbb{P}\left(\theta \mid X\right) = \frac{\theta^x(1-\theta)^{n-x}}{\int_0^1 \theta^x(1-\theta)^{n-x}\,d\theta}$$

which can only be computed numerically.

Definition 0.13. Beta Function:

$$\mathrm{B}(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

Using the beta function, we get

$$\mathbb{P}\left(\theta \mid X\right) = \frac{\theta^x(1-\theta)^{n-x}}{\mathrm{B}(x+1, n-x+1)}$$

Let's look at the random variable $X \sim \text{Beta}(\alpha,\beta)$ and its distribution.

$$X \sim \text{Beta}(\alpha,\beta) := \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

Its support is $(0,1)$.

If $f(x)$ is a pdf, then $\int_{\text{Supp}[X]} f(x)\,dx = 1$. Using this information, show that $\text{Beta}(\alpha,\beta)$ is a pdf.

$$\int_0^1 \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{1}{\mathrm{B}(\alpha,\beta)}\overbrace{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx}^{\mathrm{B}(\alpha,\beta)} = 1\checkmark$$

Its parameter space is $\alpha > 0$ and $\beta > 0$ where its finite.

Definition 0.14. Gamma Function:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt$$

which can only be computed numerically.

Properties of the Gamma Function:

1. $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
2. $\Gamma(x) = (x-1)!$ where $x \in \mathbb{N}$
3. $\Gamma(x) = (x-1)\Gamma(x-1)$ valid $\forall x$
4. $\Gamma(x+1) = x\Gamma(x)$

What's the expected value of a Beta distribution?

$$\begin{aligned}
\mathrm{E}[X] &= \int_{\Theta_0} x f(x)\, dx \\
&= \int_0^1 x \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}\, dx \\
&= \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx \\
&= \frac{\mathrm{B}(\alpha+1, \beta)}{\mathrm{B}(\alpha, \beta)} \\
&= \frac{[\Gamma(\alpha+1)\Gamma(\beta)]/[\Gamma(\alpha+\beta+1)]}{[\Gamma(\alpha)\Gamma(\beta)]/[\Gamma(\alpha+\beta)]} \\
&= \frac{\alpha\Gamma(\alpha)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

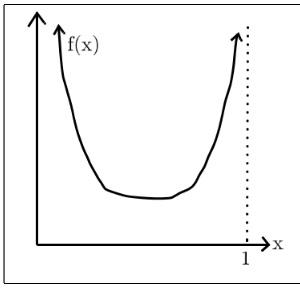
What's the mode of X if X is Beta?

$$\begin{aligned}
\mathrm{Mode}[X] &= \underset{x \in \mathrm{Supp}[X]}{\mathrm{argmax}} \{f(x)\} \\
&= \mathrm{argmax}\{\frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}\} \\
&= \mathrm{argmax}\{x^{\alpha-1}(1-x)^{\beta-1}\} \\
&= \mathrm{argmax}\{(\alpha-1)\ln(x) + (\beta-1)\ln(1-x)\}
\end{aligned}$$

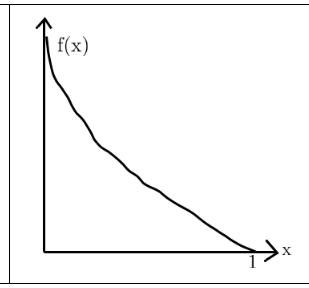
If we differentiate this function and set it equal to 0, we will find x.

$$\begin{aligned}
\frac{d}{dx}\Big[(\alpha-1)\ln(x) + (\beta-1)\ln(1-x)\Big] &= \frac{\alpha-1}{x} - \frac{\beta-1}{1-x} = 0 \\
x &= \frac{\alpha-1}{\alpha+\beta-2} \text{ only for } \alpha > 1, \beta > 1
\end{aligned}$$

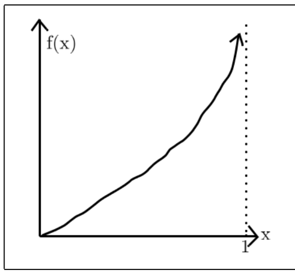
Different Types of Gamma Distributions



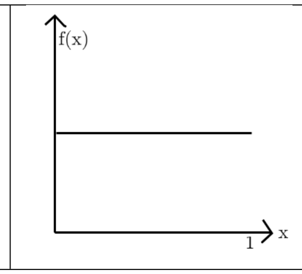
$\alpha = \beta = 0.5$



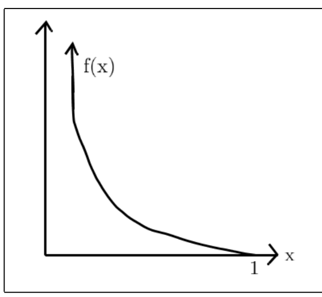
$\alpha = 1, \beta > 1$



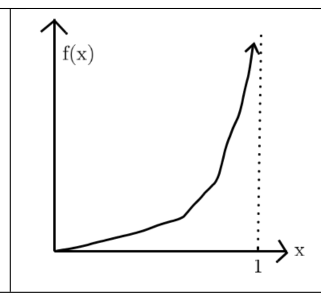
$\alpha > 1, \beta = 1$



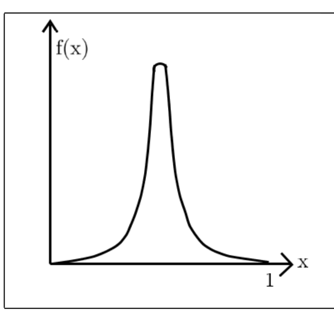
$\alpha = \beta = 1$



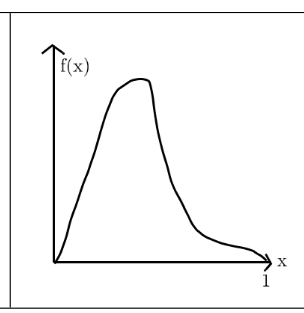
$\alpha < 1, \beta > 1$



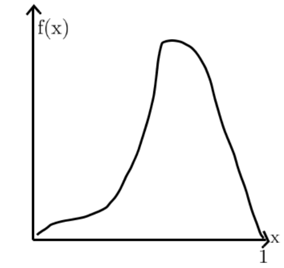
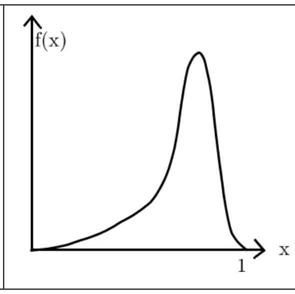
$\alpha > 1, \beta < 1$



$\alpha = \beta = 100$



$\alpha = 1.01, \beta >> 1$

	$\alpha >> 1, \beta = 1.01$		$\alpha = 100, \beta = 10$

Let's say $\mathcal{F}$ is Binomial with n known and $\theta \sim U(0,1) = \text{Beta}(1,1)$. Refresher: $\text{Binom}(n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. Then:

$$\begin{aligned}\mathbb{P}(\theta \mid X) &= \frac{\mathbb{P}(X \mid \theta)\overbrace{\mathbb{P}(\theta)}^{1}}{\underbrace{\mathbb{P}(X)}_{\int_{\Theta_0}\mathbb{P}(X \mid \theta)\,d\theta}} \\ &= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{\int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}\,d\theta} \\ &= \text{Beta}(x+1, n-x+1)\end{aligned}$$

Before we transformed $\mathbb{P}(\theta) \to \mathbb{P}(\theta \mid X)$ using X (the data). Here we transformed $\text{Beta}(1,1) \to \text{Beta}(x+1, n-x+1)$ where the first value is α and the second is β. For example, if $n = 10$ and $x = 7$, then $\theta|X \sim \text{Beta}(8,4)$. What's $\hat{\theta}_{\text{MLE}}$?

$$\hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MAP}} = \text{Mode}[\theta|X] = \frac{\alpha-1}{\alpha+\beta-1} = \frac{7}{10} = 0.7$$

Definition 0.15. Minimum Mean Square Error:

$$\hat{\theta}_{\text{MMSE}} := \text{E}[\theta|X]$$

where E is the posterior mean or expectation.

What's $\hat{\theta}_{\text{MMSE}}$ of the above distribution?

$$\hat{\theta}_{\text{MMSE}} = \text{E}[\theta|X] = \frac{\alpha}{\alpha+\beta} = \frac{2}{3} = 0.67$$

Definition 0.16. Mean Absolute Error:

$$\hat{\theta}_{\text{MAE}} = \text{Med}[\theta|X]$$

where Med is the posterior median.

Note: MAE can only be computed numerically using a computer. If using R, the command is: qbeta(0.5, α, β).
In this distribution, $\hat{\theta}_{\text{MAE}}$ comes out to be 0.676.

Definition 0.17. Quantile: If X is a continuous random variable,

$$\text{Quantile}[X, p] = F^{-1}(p)$$

Thus we say that $\text{Med}[X] = \text{Quantile}[X,\, 0.5] = F^{-1}(\frac{1}{2})$.

Let say $\mathcal{F}$ is Binomial and $\theta \sim \text{Beta}(\alpha, \beta)$ with appropriately chosen α and β. Then:

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &= \frac{\mathbb{P}(X \mid \theta)\, \mathbb{P}(\theta)}{\mathbb{P}(X)} \\
&= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \frac{1}{\text{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{1}{\text{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\, d\theta} \\
&= \frac{\theta^{x-\alpha-1}(1-\theta)^{n-x+\beta-1}}{\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\, d\theta} \\
&= \frac{1}{\text{B}(x+\alpha, n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \\
&= \text{Beta}(x+\alpha, n-x+\beta)
\end{aligned}$$

Here we have went from Beta to Beta using X. We call this conjugacy, where the prior and posterior are of the same family. In other words, the beta is conjugate prior for the binomial model.

Let $\mathcal{F}$ be a Binomial model where n is fixed and $\theta \sim \text{Beta}(\alpha, \beta) = \frac{1}{\text{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$. It turns out that

$$\text{E}[\theta] = \frac{\alpha}{\alpha+\beta}$$

and

$$\text{Var}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Then

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &= \frac{\mathbb{P}(X \mid \theta)\, \mathbb{P}(\theta)}{\mathbb{P}(X)} \\
&= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \frac{1}{\text{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{1}{\text{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\, d\theta} \\
&= \frac{\theta^{x-\alpha-1}(1-\theta)^{n-x+\beta-1}}{\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\, d\theta} \\
&= \frac{1}{\text{B}(x+\alpha, n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \\
&= \text{Beta}(x+\alpha, n-x+\beta)
\end{aligned}$$

What we have done here is that we went from $\theta \to \theta|X$. We went from $\text{Beta}(\alpha, \beta))$ to $\text{Beta}(x-\theta, n-x+\beta)$. The beta is the conjugate prior for the binomial likelihood model.

Note:

- $\hat{\theta}_{\text{MMSE}} = \text{E}[\theta|X] = \frac{x+\alpha}{n+\alpha+\beta}$
- $\hat{\theta}_{\text{MAP}} = \text{Mode}[\theta|X] = \frac{x+\alpha-1}{n+\alpha+\beta-2}$ if $x+\alpha > 1$ and $n-x+\beta > 1$
- $\hat{\theta}_{\text{MAE}} = \text{Med}[\theta|X]$ which is done by a computer

Let's look at X^*, a future observation. This means $n^* = 1$. Then

$$\begin{aligned}
\mathbb{P}\left(X^* \mid X\right) &= \int_{\Theta)} \mathbb{P}\left(X^* \mid \theta\right)\mathbb{P}\left(\theta \mid X\right)\, d\theta \\
&= \int_0^1 \underbrace{\theta^{x^*}(1-\theta)^{1-x^*}}_{PMF} \cdot \underbrace{\frac{1}{\text{B}(x+\alpha, n-x+\beta-1)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}_{PDF}\, d\theta \\
&= \frac{1}{\text{B}(x+\alpha, n-x+\beta)}\int_0^1 \theta^{x^*+x+\alpha-1}(1-\theta)^{-x^*+n-x+\beta}\, d\theta \\
&= \frac{\text{B}(x^*+x+\alpha, -x^*+n-x+\beta+1)}{\text{B}(\alpha+\beta, n-x+\beta-1)} \\
&= \frac{\Gamma(x^*+x+\alpha)\Gamma(-x^*+n-x+\beta+1)/\Gamma(n+\alpha+\beta+1)}{(\Gamma(x+\alpha)\Gamma(n-x+\beta))/\Gamma(n+\alpha+\beta)}
\end{aligned}$$

If we let $X^* = 1$:

$$\begin{aligned}
\mathbb{P}\left(X^* = 1 \mid X\right) &= \frac{\Gamma(1+x+\alpha)\Gamma(n-X+\beta)/\Gamma(n+\alpha+\beta+1)}{(\Gamma(x+\alpha)\Gamma(n-x+\beta))/\Gamma(n+\alpha+\beta)} \\
&= \frac{(x+\alpha)\Gamma(x+\alpha)/(n+\alpha+\beta)\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)/\Gamma(n+\alpha+\beta)} \\
&= \frac{x+\alpha}{n+\alpha+\beta}
\end{aligned}$$

Here we went from θ to $\theta|X$ using X, or Beta(α,β) to Beta$(x+\alpha, n-x+\beta)$ where x is the number of successes in the data and $n-x$ is the number of failures in the data. Thus we say α is the number of prior successes (pseudosuccesses) and β is the number of prior failures (pseudofailures) Together, α and β represent pseudocounts.

When we assumed $\theta \sim U(0,1)$, we assumed Beta$(\alpha,\beta) =$ Beta$(1,\ 1)$. Thus $\text{E}[\theta] = \frac{1}{1+1} = \frac{1}{2}$. We think we assumed nothing but actually we assumed 0.5. This is a criticism of Bayesian inference.

In a conjugate model, the prior parameter α,β are "usually" interpreted as pseudocounts.

$$\begin{aligned}
\theta_{\text{MMSE}} = \text{E}[\theta|X] = \frac{x+\alpha}{n+\alpha+\beta} &= \frac{n}{n}\cdot\frac{x}{n+\alpha+\beta} + \frac{\alpha+\beta}{\alpha+\beta}\cdot\frac{\alpha}{n+\alpha+\beta} \\
&= \frac{n}{n+\alpha+\beta}\hat{\theta}_{\text{MLE}} + \frac{\alpha+\beta}{n+\alpha+\beta}\text{E}[\theta] \\
&= (1-\rho)\hat{\theta}_{\text{MLE}} + \rho(\text{E}[\theta])
\end{aligned}$$

If n is high, then ρ is low and thus θ_{MLE} dominates. If n is low, then ρ is high and $\text{E}[\theta]$ dominates. ($\lim_{n\to\infty} \rho = 0$).

$\text{E}[\theta|X]$ is called a "shrinkage estimation" because it shrinks to $\text{E}[\theta]$.

Let's say $n = 2, x = 0$, and $\theta \sim U(0,1)$, meaning $\alpha = \beta = 1$. Thus $\text{E}[\theta] = 0.5$, as shown above, Then $\theta_{\text{MLE}} = 0$. If $\rho = 0.5$, then

$$\text{E}[\theta|X] = (1-\rho)\theta_{\text{MLE}} + \rho\text{E}[\theta] = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$$

Here we have shrunk $\text{E}[\theta|X]$ closer to $\text{E}[\theta]$. If α and β are bigger, it shrinks harder.

Wilson Estimate:

$$\text{E}[\theta|X] = \frac{x+\alpha}{n+\alpha+\beta} = \frac{x+1}{n+2}$$

when $\alpha = \beta = 1$.

Confidence Interval:

$$CI_{\theta,1-\alpha} = \left[\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta}_{\text{MLE}})\right]$$

Let's say $x = 1, n = 2, \hat{\theta} = \bar{x} = 0.5$. Then the confidence interval at the 95% confidence level is

$$CI_{\theta,95\%} = \left[0.5 \pm 2\sqrt{\frac{0.5(1-0.5)}{2}}\right] = (-0.21, 1.21)$$

This is absurd because one value is negative and the other is more than 1. We can say $[0,1]$ but that is just useless.

Let $\theta \sim U(0,1)$, then $\theta|X \sim \text{Beta}(x+1, n-x+1) = \text{Beta}(2,\ 2)$. Here we won't make a best guess but a range.

Credible Region (CR) for θ of size $1-\alpha$:

$$CR_{\theta,1-\alpha} = [\text{Quantile}[\theta|X, \frac{\alpha}{2}], \text{Quantile}(\theta|X, 1-\frac{\alpha}{2})]$$

For this example,

$$\begin{aligned}&= [\text{qbeta}(0.025, 2, 2), \text{qbeta}(0.975, 2, 2)]\\ &= [0.094, 0.906]\end{aligned}$$

Let's say we have a distribution such that there are three peaks. To find a credible region of it, we would have to find the the union of three different peaks, or the HDR (higher density region). This is a disadvantage because it is not plausible to have non contiguous regions and it is computationally expensive.

Let $\mathcal{F}$ be Binomial, with $\theta \sim U(0,1)$, $n = 2$ and $x = 1$. Then $\theta|X \sim \text{Beta}(2,2)$. At an alpha level of 5%, the 2 sided is $CR_{\theta,1-\alpha} = [\text{Quantile}[\theta|X, \frac{\alpha}{2}], \text{Quantile}(\theta|X, 1-\frac{\alpha}{2})] =$

$[\text{qbeta}(0.025, 2, 2), \text{qbeta}(0.975, 2, 2)] = [0.094, 0.906]$. However since $n = 2$, asymptotic normaling breaks down and we can't do this.

One Sided Credible Region:

$$CR_{L,\theta,1-\alpha} = [0, \text{Quantile}[\theta|X, 1-\alpha]]$$

$$CR_{R,\theta,1-\alpha} = [\text{Quantile}[\theta|X, 1], 1]$$

The left credible region is for the lower 95% while the right credible region is for the higher 95%.

In the above example,

$$\begin{aligned} CR_{L,\theta,1-\alpha} &= [0, \text{qbeta}(0.95, 2, 2)] \\ &= [0, 0.865] \end{aligned}$$

and

$$\begin{aligned} CR_{R,\theta,1=\alpha} &= [\text{qbeta}(0.05, 2, 2), 1] \\ &= [0.135, 1] \end{aligned}$$

.

Hypothesis Test (Theory Testing): "theory" - research hypothesis or alternative hypothesis - H_A
Null hypothesis - assuming the theory is opposite - H_0
We reject the null hypothesis (accept theory) if "overwhelming" evidence. "Overwhelming" is the "level" of α that is chosen. If data is sufficient at α, reject H_0 and accept H_A. If it is not sufficient, retain H_0 (fail to reject).

One Sided Hypothesis Test: $H_0 : \theta \leq \theta_0 = 0.5$, $H_A : \theta > \theta_0 = 0.5$ where $\hat{P} = N(\theta_0, (\sqrt{\frac{\theta(1-\theta)}{n}})^2)$.
If $\theta \in$ retainment region, retain H_0 (fail to reject). If $\theta \notin$ retainment region, reject H_0.
P-value = P(seeing the data or more extreme $|H_0$ true) = $\underset{\alpha}{\text{argmax}}\{\hat{\theta} \in \text{Retainment region}\}$
If the p-value $< \alpha$, reject H_0. If the p-value $> \alpha$, retain H_0.

Two Sided Hypothesis Test: $H_0 : \theta = \theta_0 = 0.5$, $H_A : \theta \neq \theta_0 = 0.5$. This is the same as asking if $\{\theta > 0.5 \bigcup \theta < 0.5\}$.
Note:

- p-value $\neq \mathbb{P}(H_0)$
- p-value $\neq \mathbb{P}(H_A)$
- p-value $\neq \mathbb{P}(H_0 \mid X)$
- p-value $\neq \mathbb{P}(H_A \mid X)$

Let's say $H_0 : \theta \leq \theta_0 = 0.5$, $H_A : \theta > \theta_0 = 0.5$ and $\alpha = 5\%$, $n = 2$, $x = 1$ and $\theta \sim U(0,1)$.
Bayesian P-value:

$$\begin{aligned} \text{p-value} &= \mathbb{P}\left(H_0 \mid X\right) = \mathbb{P}\left(\theta \leq \theta_0 \mid X\right) \\ &= \int_0^1 \frac{1}{\mathrm{B}(\alpha + x, \beta + n - x)} \theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1}\, d\theta = \text{pbeta}(\theta_0, x+\alpha, n-x+\beta) \end{aligned}$$

For this example, p-value $= \mathbb{P}\left(\theta < 0.5 \mid X\right) = \int_0^{0.5} \text{Beta}(2,2)\, d\theta = \text{pbeta}(0.5, 2, 2) = 0.5$
Since this is $\not< \alpha = 5\%$, retain H_0. Note that here, we said $U(0,1) = \text{Beta}(2,2)$.

$$\mathbb{P}\left(H_0 \mid X\right) = \frac{\mathbb{P}\left(X \mid H_0\right)\mathbb{P}\left(H_0\right)}{\mathbb{P}\left(X\right)} = \frac{\mathbb{P}\left(X \mid H_0\right)\mathbb{P}\left(H_0\right)}{\mathbb{P}\left(X \mid H_0\right)\mathbb{P}\left(H_0\right) + \mathbb{P}\left(X \mid H_A\right)\mathbb{P}\left(H_A\right)}$$

This puts more weight on H_A than desired.
Point Null: $H_0 : \theta = \theta_0 = 0.5$, $H_A : \theta = \theta \neq 0.5$. Then

$$\text{p-value} = \mathbb{P}\left(H_0 \mid X\right) = \mathbb{P}\left(\theta = 0.5 \mid X\right) = \int_{0.5}^{0.4} \text{Beta}(2,2)\, d\theta = 0$$

This integral will always be zero..

Solution: (1) $H_0 : \theta \in (\theta_0 \pm \delta)$, $H_A : \theta \notin (\theta_0 \pm \delta)$. The parenthesis is the region of equivalence. (2) $H_0 : \theta = \theta_0 = 0.5$, $H_A : \theta = \theta_0 \neq 0.5$, if $\theta_0 \in CR_{\theta, 1-\alpha}$, retain H_0

Let's say $\alpha = 5\%$, $n = 100$ and $x = 61$.
In the frequentist approach: Retainment Region =

$$[\theta_0 \pm z_{\alpha/1}\sqrt{\frac{\theta_0(1-\theta_0)}{n}}] = [0.5 \pm 2\sqrt{\frac{0.5^2}{100}}] = [0.4, 0.6]$$

Since $\hat{\theta} = \frac{61}{100} = 0.61$, $0.61 \in$ retainment region, thus reject H_0.
P-value $= \mathbb{P}\left(|z| > \frac{0.61-0.5}{0.05}\right) = 2\mathbb{P}\left(z > 2.2\right) = 2(1 - \text{pnorm}(2.2)) = 0.278$. This is less than $\alpha = 5\%$ thus reject H_0.

In the Bayesian approach, $\theta \sim U(0,1)$ and $\delta = 0.01$. Then $H_0 : \theta \in (0.49, 0.51)$ and $H_A : \theta \notin (0.49, 0.51)$. Since $\theta | X \sim \text{Beta}(62, 40)$,

$$\begin{aligned} \text{p-value} &= \mathbb{P}\left(H_0 \mid X\right) \\ &= \mathbb{P}\left(\theta \in (0.49, 0.51) \mid X\right) \\ &= \int_{0.49}^{0.51} \text{Beta}(62, 40)\, d\theta \\ &= \text{qbeta}(.51, 62, 40) - \text{qbeta}(0.49, 62, 40) = 0.0147 \end{aligned}$$

This value is $< \alpha - .05$. Thus retain H_0.

$$CR_{\theta, 1-\alpha} = [\text{qbeta}(0.025, 62, 40), \text{qbeta}(0.975, 62, 40)] = (0.511, 0.700)$$

Thus $\theta_0 = 0.5 \notin CR$, therefore reject H_0.

Let's say $H_0 : \theta = \theta_0 = 0.5$ and $H_A : \theta \neq \theta_0 = 0.5$ with $\theta \sim U(0,1)$.

Bayesian Factor: tells the relativity of $P_{H_A}(X)$ to $P_{H_0}(X)$

$$\begin{aligned} B &= \frac{P_{H_A}(X)}{P_{H_0}(X)} \\ &= \frac{\int_{\Theta \in H_A} \mathbb{P}\left(X \mid \theta\right) P_{H_A}(\theta)\, d\theta}{\int_{\Theta \in H_0} \mathbb{P}\left(X \mid \theta\right) P_{H_0}(\theta)\, d\theta} \\ &= \frac{\int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x}\, d\theta}{\int_{0.5} \binom{n}{x} \theta^x (1-\theta)^{n-x}\, d\theta} \\ &= \frac{\int_0^1 \theta^{0.61} (1-\theta)^{0.39}\, d\theta}{0.5^{0.61}(1-0.5)^{0.39}} \\ &= \frac{\mathrm{B}(62,40)}{0.5^{100}} = 1.39 \end{aligned}$$

This tells us that P_{H_A} is not too far from P_{H_0}.

Bayes Factor:

$$B := \frac{P_{H_A}(X)}{P_{H_0}(X)} = \frac{\int_{\Theta_{H_A}} P_{H_A}(X \mid \theta) P_{H_A}(\theta)\, d\theta}{\int_{\Theta_{H_0}} P_{H_0}(X \mid \theta) P_{H_0}(\theta)\, d\theta}$$

Note: If $B > 1$, H_A is supported. The bigger B is, the better H_A is.

Let $H_0 : \theta = 0.5$ and $H_A : \theta \neq 0.5$. Assume $\mathcal{F}$ is Binomial. For H_0: $\theta \sim \text{Deg}(0.5)$ and for H_A: $\theta \sim U(0,1)$. $n = 100$ and $x = 61$. In the frequentist approach, H_0 is rejected because $p = 0.61$ which is too far from 0.5.

$$B = \frac{\int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot (1)\, d\theta}{\int_{\{0.5\}} \binom{n}{x} 0.5^x (1-0.5)^{n-x} \cdot (1)\, d\theta} = \frac{B(x+1, n-x+1)}{0.5^n} = \frac{B(62,98)}{0.5^{100}} = 1.39$$

Difference Conclusions:

- If $B < 1$, then no evidence
- If $B \in [1 : 1.3 : 1]$, then barely worth mentioning
- If $B \in [3:1, 10:1]$, then substantial
- If $B \in [10:1, 30:1]$, then strong
- If $B \in [30:1, 100:1]$, then very strong
- If $B > 100\%$, then decisive

Suppose $H_0 : \theta = 0.5$ and $H_A : \theta \neq 0.5$. Let $n = 104490000$, $x = 52263920$ and $\hat{\theta} = 0.50001768$. In the frequentist approach, the p-value is 0.0003, which is less than 0.05 and thus H_0 is rejected. In the Bayesian approach, assuming $\theta \sim \text{Beta}(1,1)$,

$$B = \frac{B(52263921, 104490000 - 52263920 + 1)}{0.50001768^{104490000}} = \frac{1}{12}$$

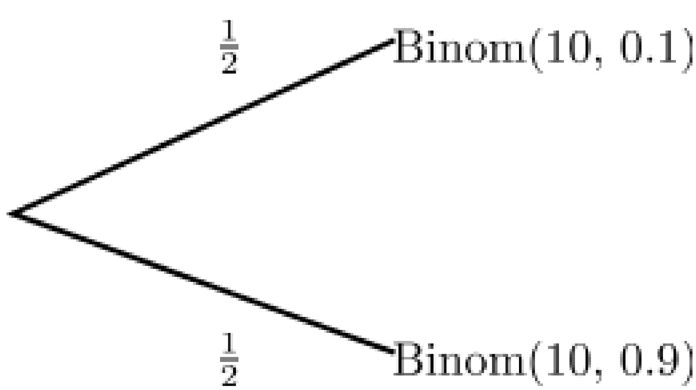
According to this, since $B < 1$, there is no evidence. This gives conflicting results. This happened because as n becomes large, H_0 cannot be true and thus is rejected.

End of Midterm 1 Material

Mixture Distribution: Let $X \sim \begin{cases} N(0,1)^2 & 0.5 \\ N(10,1^2) & 0.5 \end{cases}$.

$$\begin{aligned} P(X) &= \sum_{\theta \in \Theta} \mathbb{P}(X \mid \theta) \mathbb{P}(\theta) \\ &= \mathbb{P}(X \mid \theta = 0) \mathbb{P}(\theta = 0) + \mathbb{P}(X \mid \theta = 10) \mathbb{P}(\theta = 10) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{2} + \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2} \cdot \frac{1}{2} \end{aligned}$$

Suppose the following:



Then

$$\begin{aligned} \mathbb{P}(X) &= \sum_{\theta \in \Theta} \mathbb{P}(X \mid \theta) \mathbb{P}(\theta) \\ &= \mathbb{P}(X \mid \theta = 0.1) \mathbb{P}(\theta = 0.1) + \mathbb{P}(X \mid \theta = 0.9) \mathbb{P}(\theta = 0.9) \\ &= \binom{10}{x} 0.1^x (1-0.1)^{10-x} \cdot \frac{1}{2} + \binom{10}{x} 0.9^x (1-0.9)^{10-x} \cdot \frac{1}{2} \end{aligned}$$

What we did here is that we went from $\theta \sim \text{Beta}(\alpha, \beta)$ to $X \mid \theta \sim \text{Binom}(n, \theta)$. Since θ is continuous:

$$\begin{aligned}
\mathbb{P}(X) &= \int_\Theta \mathbb{P}(X \mid \theta)\, \mathbb{P}(\theta)\, d\theta \\
&= \int_0^1 \left(\binom{n}{x} \theta^x (1-\theta)^{n-x} \right) \cdot \left(\frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \right) d\theta \\
&= \binom{n}{x} \frac{1}{B(\alpha,\beta)} \int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\, d\theta \\
&= \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha,\beta)} \\
&= \text{BetaBinom}(n, \alpha, \beta)
\end{aligned}$$

This is the Beta-Binomial model. Let X is a random variable of this model; then $X \sim$BetaBinom(n, α, β). Supp$[X] = \{0.1.\dots, n\}$ and the parameter spaces are: $n \in \mathbb{N}$, $\alpha > 0$ and $\beta > 0$.

$$\begin{aligned}
\text{E}[X] &= n \frac{\alpha}{\alpha+\beta} \\
\text{Var}[X] &= \frac{n\alpha\beta}{(\alpha+\beta)^2} \underbrace{\frac{\alpha+\beta+n}{\alpha+\beta+1}}_{\in[1,n]}
\end{aligned}$$

Thus the variance is an inflated binomial variance. Let $\theta = \frac{\alpha}{\alpha+\beta}$, then $\text{E}[X] = n\theta$. Let $B = \frac{\alpha}{\theta} - \alpha$. Then

$$\begin{aligned}
\lim_{\alpha\to\infty} \text{E}[X] &= n\theta \\
\lim_{\alpha\to\infty} \text{Var}[X] &= \lim_{\alpha\to\infty} n \overbrace{\frac{\alpha}{\alpha+\beta}}^{\theta} \overbrace{\frac{\beta}{\alpha-\beta}}^{1-\theta} \\
&= \frac{\alpha+\beta+n}{\alpha+\beta+1} \\
&= \underbrace{n\theta(1-\theta)}_{\text{variance of binom}} \lim_{\alpha\to\infty} \frac{\alpha + \frac{\alpha}{\theta} - \alpha + n}{\alpha + \frac{\alpha}{\theta} - \alpha + 1} \\
&= n\theta(1-\theta) \lim_{\alpha\to\infty} \frac{\alpha + n\theta}{\alpha+\theta} = n\theta(1-\theta)\cdot 1 \\
&= n\theta(1-\theta)
\end{aligned}$$

From this, as α gets higher, θ gets tighter and becomes degenerate and more like a binomial model.

Suppose $X \mid \theta \sim \text{Binom}(n, \theta)$, $\theta \sim \text{Beta}(\alpha, \beta)$ and $\theta \mid X \sim \text{Beta}(\alpha + x, \beta + n - x)$. Suppose

$X^* \mid X \sim \text{Bern}(\frac{x+\alpha}{n+\alpha+\beta})$ where $n^* = 1$. Then:

$$\begin{aligned}\mathbb{P}(X^* \mid X) &= \int_\Theta \underbrace{\mathbb{P}(X^* \mid \theta)}_{\text{binom}} \underbrace{\mathbb{P}(\theta \mid X)}_{\text{beta}} \, d\theta \\ &= \int_0^1 \binom{n^*}{x^*} \theta^{x^*}(1-\theta)^{n^*-x^*} \cdot \frac{1}{B(\alpha+x, \beta+n-x)} \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \, d\theta \\ &= \text{BetaBinom}(n^*, \alpha+x, \beta+n-x)\end{aligned}$$

Let $X|\theta \sim \text{Binom}(n,\theta)$, $\theta \sim \text{Beta}(\alpha,\beta)$ and $\theta|X \sim \text{Beta}(\overbrace{\alpha+x}^{\alpha'}, \overbrace{\beta+n-x}^{\beta'})$. Then

$$X^*|X \sim \text{BetaBinom}(n^*, \alpha', \beta') = \binom{n^*}{x^*} \frac{B(\overbrace{\alpha+x}^{\alpha'} + x^*, \overbrace{\beta+n-x}^{\beta'} + n^* - x^*)}{B(\underbrace{\alpha+x}_{\alpha'}, \underbrace{\beta+n-x}_{\beta'})}$$

Posterior Predictive Distribution: $\mathbb{P}(X^* \mid X) = \int_\Theta \mathbb{P}(X^* \mid \theta) \mathbb{P}(\theta \mid X) \, d\theta$ (the distribution of function X^* given data x)
$\mathbb{P}(X)$ is the distribution of data observed $= \int_\Theta \mathbb{P}(X \mid \theta) \mathbb{P}(\theta) \, d\theta$
Prior Predictive Distribution: $\mathbb{P}(X \mid \{\}) = \int \mathbb{P}(X \mid \theta) \mathbb{P}(\theta \mid \{\}) \, d\theta$

Let $X \sim \text{BetaBinom}(n, \alpha, \beta)$. If $\theta \sim U(0,1) = \text{Beta}(1,1)$, this is an uninformative prior, as well as a indifference or Laplace prior. It says there is one success and one failure. The most uninformative prior is $\theta \sim \text{Beta}(0,0)$. However, this is "illegal" because α and β are not in the parameter space and thus do not form a true PDF. This prior is called an improper prior, as well as Haldane prior.

Let's say we go along with $\theta \sim \text{Beta}(0,0)$. Then $\theta|X \sim \text{Beta}(x, n-x)$. From this,

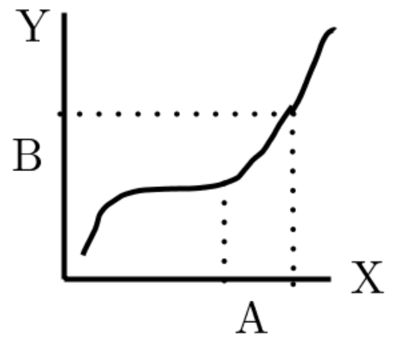
$$\hat{\theta}_{\text{MMSE}} = \frac{x}{n} = \hat{\theta}_{\text{MLE}}$$

This posterior could be improper if $x = 0$ (no successes) or if $x = n$ (no failures). Therefore, be careful when using "improper" priors as your posterior could also be improper.
Note: Beta(0, 0) and Beta(1, 1) are both uninformative but only Beta(1, 1) is indifferent.

Reparameterization: $R = \text{Odds}(\theta) = \frac{\theta}{1-\theta}$. For example, $R = \text{Odds}(0.9) = \frac{0.9}{1-0.9} = 9$. Note that $\theta = (0,1)$ and $R = (0,\infty)$.

Let X and Y be two random variables related by a 1-1 inverse transform. This means

$Y = t(X)$ and $X = t^{-1}(Y)$. We know $f_X(x)$, the PDF of X. We want the PDF of Y, $f_Y(y)$.



Since $\mathbb{P}(X \in A) \approx f_X(x)A$ and $\mathbb{P}(Y \in B) \approx f_Y(y)B$

$$f_X(x)|dx| = f_Y(y)|dy| \to f_Y(y) = f_X(x)\Big|\frac{dx}{dy}\Big|$$

By the above equations, we can substitute for X:

$$f_Y(y) = f_X(t^{-1}(y))\Big|\frac{d}{dy}[t'(x)]\Big|$$

Since $R = t(\theta) = \frac{\theta}{1-\theta}$, then $\theta = t^{-1}(R) = \frac{R}{R+1}$ Therefore

$$f_R(r) = f_\theta(t^{-1}(r))\Big|\frac{d}{dr}[t^{-1}(r)]\Big| = f_Y(\frac{r}{r+1})\Big|\frac{d}{dr}p\frac{r}{r+1}\Big| = (1)\Big| - \frac{1}{(r+1)^2}\Big| = \frac{1}{(r+1)^2}$$

Let $\theta \sim U(0,1)$ or $\theta \sim \text{Beta}(0,0)$ (uninformative). If under a reparameterization $\phi = t(\theta)$, what if I had a protocol which allows us to pick a priors given $\mathcal{F}$:

$$\mathbb{P}(X \mid \theta) \overset{\text{pick}}{\to} \mathbb{P}(\theta) \text{ and } \mathbb{P}(X \mid \phi) \overset{\text{pick}}{\to} \mathbb{P}(\phi)$$

such that we have $P(\phi) = p(t^{-1}(\phi))\Big|\frac{d}{dt}t^{-1}(\phi)\Big|$ (Jeffrey's prior).

$\mathbb{P}(\theta \mid X) = \frac{\mathbb{P}(X \mid \theta)\mathbb{P}(\theta)}{\mathbb{P}(X)} \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta)$
in fact, $f(x;\theta) \propto g(x;\theta)$ where g is a kernel of f. This means $f(x;\theta) = \frac{1}{c}g(x;\theta)$.

$$\int f(x)\,dx = 1 \to \int g(x)\,dx = \int cf(x)\,dx = c\underbrace{\int f(x)\,dx}_{1} \to c = \int g(x)\,dx$$

Note: f and g are 1-1.

Let $X|\theta \sim \text{Binom}(n,\theta)$ and $\theta \sim \text{Beta}(\alpha,\beta)$.

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &\propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{\overbrace{x+\alpha-1}^{a}}(1-\theta)^{\overbrace{n-x+\beta-1}^{b}} \\
&= \text{Beta}(x+\alpha, n-x+\beta)
\end{aligned}$$

$$\theta \sim \text{Beta}(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \underbrace{\theta^a(1-\theta)^b}_{\text{kernel of the beta}}$$

$X|\theta \sim \text{Binom}(n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = (\frac{n!}{x!(n-x)!})\theta^x(1-\theta)^n(1-\theta)^{-x} \propto \frac{1}{x!(n-x)!}(\frac{\theta}{1-\theta})^x$

Likelihood: $\mathcal{L}(\theta;x) = \mathbb{P}(x;\theta)$
Log-Likelihood: $l(\theta;x) = \ln(\mathcal{L}(\theta;x))$
Score Function: $s(\theta;x) = l'(\theta;x)$
Fisher Information: $I(\theta) = \text{Var}_x[s(\theta;x)] = \cdots = \text{E}_x[s(\theta;x)^2] = \cdots = \text{E}_x[-l''(\theta;x)]$

The Fisher Information measures the information in X about θ.

Let $X \sim \text{Binom}(n;\theta)$ Then

$$\begin{aligned}
X \sim \text{Binom}(n;\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\
l(\theta;x) &= ln\frac{n}{x} + x\ln\theta + (n-x)\ln(1-\theta) \\
l'(\theta;x) &= \frac{x}{\theta} - \frac{n-x}{1-\theta} \\
l''(\theta;x) &= \frac{-x}{\theta^2} - \frac{n-x}{(1-\theta)^2} \\
I(\theta) &= \text{E}_x[-l''(\theta;x)] \\
&= \text{E}[\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}] \\
&= \frac{\text{E}[X]}{\theta^2} + \frac{n-\text{E}[X]}{(1-\theta)^2} \\
&= \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} \\
&= n(\frac{1}{\theta} + \frac{1}{1-\theta}) \\
&= n\frac{1}{\theta(1-\theta)}
\end{aligned}$$

The Fisher information for the Binomial distribution is $n\frac{1}{\theta(1-\theta)}$.
For example, if $X \sim \text{Binom}(1,0.5)$, $I(\theta) = 4$; if $X \sim \text{Binom}(1,0.01)$, $I(\theta) = 101.01$.

Given $\mathcal{F} = \mathbb{P}(X \mid \theta)$, pick $\mathbb{P}(\phi)$ where $\phi = t(\theta)$ and t is 1-1 and smooth.

$$\mathbb{P}(X \mid \theta) \overset{\text{pick}}{\to} \mathbb{P}(\theta) \text{ and } \mathbb{P}(X \mid \phi) \overset{\text{pick}}{\to} \mathbb{P}(\phi)$$

But we want $\mathbb{P}(\theta)$ and $\mathbb{P}(\phi)$ to be related via change of variables.
Jeffrey's Prior: $\mathbb{P}(\theta) \propto \sqrt{I(\theta)}$

Let $X \sim \text{Binom}(n,\theta)$ Then

$$\begin{aligned}\mathbb{P}\left(\theta\right) &\propto \sqrt{n(\frac{1}{\theta(1-\theta)})} \\ &\propto \frac{1}{\theta(1-\theta)} \\ &= \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \\ &\propto \text{Beta}(\frac{1}{2},\frac{1}{2}) \\ &= \underbrace{\frac{1}{B(\frac{1}{2},\frac{1}{2})}}\pi\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \\ &= \frac{1}{\pi\sqrt{\theta(1-\theta)}}\end{aligned}$$

This is the arcsin distribution. It is equidistant from Beta(0,0) and Beta(1,1). It is also called Jeffrey's prior (uninformative).

$$\mathbb{P}\left(X \mid \theta\right) \to \mathbb{P}\left(\theta\right) = \frac{1}{\pi\sqrt{\theta(1-\theta)}}$$

Recall that $R = t(\theta) = \frac{\theta}{1-\theta}$ and $\theta = t^{-1}(R) = \frac{R}{R+1}$.
Let $X \sim \text{Binom}(n,\theta)$. Then

$$\begin{aligned}\mathbb{P}\left(X \mid R\right) &= \binom{n}{x}(\frac{R}{R+1})^x(\underbrace{1-\frac{R}{R+1}}_{\frac{1}{R}+1})^{n-x} \\ &= \binom{n}{x}\frac{R^x}{(R+1)^n} \\ l(X;R) &= \ln\binom{n}{x} + x\ln R - n\ln(R+1) \\ l'(X;R) &= \frac{X}{R} - \frac{n}{R+1} \\ l''(X;R) &= -\frac{X}{R^2} + \frac{n}{(R+1)^2} \\ I(R) &= \text{E}[-l''(X;R)] = \text{E}[\frac{X}{R^2} - \frac{n}{(R+1)^2}] \\ &= \frac{\text{E}[X]}{R^2} - \frac{n}{(R+1)^2} \\ &= \frac{n\frac{R}{R+1}}{R^2} - \frac{n}{(R+1)^2} \\ &= n\Big(\frac{1}{R(R+1)} + \frac{1}{(R+1)^2}\Big) \\ &= n\frac{1}{R(R+1)^2}\end{aligned}$$

Therefore
$$\mathbb{P}(R) \propto \sqrt{n}R(R+1)^2 \propto \frac{1}{\sqrt{R}}\frac{1}{R+1} \propto \frac{1}{\pi}\frac{1}{\sqrt{R}}\frac{1}{R+1} = \mathbb{P}(\phi)$$
By change of variables,
$$\begin{aligned}\mathbb{P}_R(R) &= \mathbb{P}_\theta((t^{-1}(R)))\Big|\frac{d}{dr}[t^{-1}(R)]\Big| \\ &= \frac{1}{\pi}(\frac{R}{R+1})^{-\frac{1}{2}}(\frac{1}{R+1})^{-\frac{1}{2}} \cdot \frac{1}{(R+1)^2} \\ &= \frac{1}{\pi}R^{-\frac{1}{2}}(R+1)\frac{1}{(R+1)^2} \\ &= \frac{1}{\pi}\frac{1}{\sqrt{R}}\frac{1}{R+1}\end{aligned}$$
General Case: Given $\mathbb{P}(X \mid \theta)$, $\mathbb{P}(X \mid \phi)$, and that
$$\begin{aligned}\mathbb{P}(\theta) &\propto \sqrt{I(\theta)} \\ \mathbb{P}(\phi) &\propto \sqrt{I(\phi)}\end{aligned}$$
Then
$$\begin{aligned}\mathbb{P}(\phi) &= \mathbb{P}_\theta(\underbrace{t^{-1}(\phi)}_{\theta})\Big|\frac{d}{d\phi}t^{-1}(\phi)\Big| \propto \sqrt{I(\phi)} \\ &= \mathbb{P}_\theta(\theta)\Big|\frac{d\theta}{d\phi} \\ &\propto \sqrt{I(\theta)}\Big|\frac{d\theta}{d\phi}\Big| \\ &= \sqrt{I(\theta)\frac{d\theta}{d\phi}\frac{d\theta}{d\phi}} \\ &= \sqrt{\mathrm{E}[s(\theta;X)^2]\frac{d\theta}{d\phi}\frac{d\theta}{d\phi}} \\ &= \sqrt{\mathrm{E}[\frac{dl}{d\theta}\frac{dl}{d\theta}\frac{d\theta}{d\phi}\frac{d\theta}{d\phi}]} \\ &= \sqrt{\mathrm{E}[(\frac{dl}{dt})^2]} \\ &= \sqrt{\mathrm{E}[s(\phi;X)^2]} \\ &= \sqrt{I(\phi)}\end{aligned}$$
A baseball player's true batting average is given as follows:
$$\hat{\theta} = BA := \frac{\text{\# hits}}{\text{\# at bats}} = \frac{x}{n} = \hat{\theta}_{\text{MLE}}$$
Say # of hits $\propto$ Binom(# bats, θ). For $n = 2$, if $x = 0$, then BA $= 0$. If $x = 1$, BA $= \frac{1}{2}$. If $x = 2$, BA $= 1$. This is absurd. Thus let's use $\theta \sim \text{Beta}(\alpha, \beta)$ to shrink. Fix a beta to the

prior data. Let's say $\hat{\alpha}_{\text{MLE}} = 78.7$ and $\hat{\beta}_{\text{MLE}} = 224.8$. Then $\hat{\alpha} + \hat{\beta} = 303.5$ which is strong. It also follows that $\hat{\theta}_{\text{MMSE}} = \frac{x+\alpha}{n+\alpha+\beta} = \frac{x+78.7}{n+303.5}$. For n large, use this estimation. This is called Empirical Bayes.
Steps

1. Get all data.
2. Fit prior to all data using MLE.
3. Use this fit's hyperparameters for inference.

Let $\mathcal{F}$ = Geometric. Then $X|\theta \sim (1-\theta)^x\theta$ where X is number of failures. $\text{Supp}[X] = \{0, 1, \dots\}$. $\Theta = (0,1)$ and $\text{E}[X] = \frac{1}{\theta} - 1$. If θ is large, then x is small; if θ is small, then x is large. Let's say $X_1 \sim \theta_1, \dots, X_n \sim \theta_n \overset{iid}{\sim} \text{Geom}(\theta)$. Then

$$\mathbb{P}(X \mid \theta) = \prod_{i=1}^n (1-\theta_i)^n\theta_i = (1-\theta)^{\sum x_i}\theta^n$$

Furthermore,

$$\begin{aligned}\mathbb{P}(\theta \mid X) &\propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta) \\ &= \underbrace{(1-\theta)^{\sum x_i}\theta^n}_{\text{kernel of beta}}\,\mathbb{P}(\theta) \\ &\propto \theta^n(1-\theta)^{\sum x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{n+\alpha-1}(1-\theta)^{\sum x_i+\beta-1} \\ &= \text{Beta}(n+\alpha, \sum x_i + \beta)\end{aligned}$$

This is done using $\mathbb{P}(\theta) = \text{Beta}(\alpha, \beta)$. What we found here is that beta is also the conjugate prior for the geometric random variable.

If $X_1|\theta, \dots, X_n|\theta \overset{iid}{\sim} \text{Geom}(\theta)$ and $\theta \sim \text{Beta}(\overbrace{\alpha, \beta}^{\text{hyperparameters}})$, then

$$\theta|X_1, \dots, X_n \sim \text{Beta}(\underbrace{n+\alpha}_{\alpha'}, \underbrace{\sum x_i + \beta}_{\beta'})$$

Furthermore

$$\begin{aligned}\hat{\theta}_{\text{MMSE}} &= \frac{n+\alpha}{n+\alpha+\sum x_i+\beta} \\ \hat{\theta}_{\text{MAE}} &= \text{qbeta}(0.5, n+\alpha, \sum x_i + \beta) \\ \hat{\theta}_{\text{MAP}} &= \frac{n+\alpha-1}{n+\alpha+\sum x_i+\beta-2}\end{aligned}$$

α = pseudo number of trials, β = seen total number of failures. If $\theta \sim \text{Beta}(0,0)$, Haldone, where $\alpha = 0$ and $\beta = 0$, this is complete ignorance. If $\theta \sim U(0,1) = \text{Beta}(1,1)$, Laplace, where $\alpha = 1$ and $\beta = 1$, this is indifference prior which gives no special preference. What's Jeffrey's prior?

$$\begin{aligned}
\mathcal{L}(\theta; X) &= (1-\theta)^{\sum x_i}\theta^n \\
l(\theta; X) &= \sum x_i \ln(1-\theta) + n\ln\theta \\
l'(\theta; X) &= -\frac{\sum x_i}{1-\theta} + \frac{n}{\theta} \\
l''(\theta; X) &= -\frac{\sum x_i}{(1-\theta)^2} - \frac{n}{\theta^2} \\
I(\theta) &= \mathrm{E}[-l''(\theta; X)] = \mathrm{E}\Big[\frac{\sum x_i}{(1-\theta)^2} + \frac{n}{\theta^2}\Big] \\
&= \frac{\mathrm{E}[x_i]}{(1-\theta)^2} + \frac{n}{\theta^2} \\
&= \frac{n\mathrm{E}[X]}{(1-\theta)^2} + \frac{n}{\theta^2} \\
&= n\Big(\frac{\frac{1}{\theta}-1}{(1-\theta)^2} + \frac{1}{\theta^2}\Big) \\
&= n\Big(\frac{\frac{1-\theta}{\theta}}{(1-\theta)^2} + \frac{1}{\theta^2}\Big) \\
&= n(\frac{1}{\theta(1-\theta)} + \frac{1}{\theta^2}) \\
&= n(\frac{1}{\theta^2(1-\theta)})
\end{aligned}$$

Therefore

$$\mathbb{P}(\theta) \propto \sqrt{I(\theta)} = \sqrt{n\frac{1}{\theta^2(1-\theta)}} \propto \theta^{-1}(1-\theta)^{-\frac{1}{2}} \propto \mathrm{Beta}(0, \frac{1}{2})$$

Jeffrey's prior is $\theta \sim \mathrm{Beta}(0, \frac{1}{2})$, with $\alpha = 0$ and $\beta = \frac{1}{2}$. This is an improper prior and similar to Wilson's estimate.

Let $X_1, \dots, X_n | \theta \overset{iid}{\sim} \mathrm{Geom}(\theta)$, $\theta \sim \mathrm{Beta}(\alpha, \beta)$. Then $\theta | X_1, \dots, X_n \sim \mathrm{Beta}(n+\alpha, \sum x_i + \beta)$ where α is the number of pseudotrials and β is the number of pseudofailures.

$$\hat{\theta}_{\text{MMSE}} = \frac{n+\alpha}{n+\alpha+\sum x_i + \beta}$$

Haldane Prior: if $\theta \sim \mathrm{Beta}(0,0), \hat{\theta}_{\text{MMSE}} = \frac{n}{n+\sum x_i} = \frac{1}{1+\frac{\sum x_i}{n}} = \frac{1}{1+\bar{x}} = \hat{\theta}_{\text{MLE}}$

Laplace Prior: if $\theta \sim \mathrm{Beta}(1,1), \hat{\theta}_{\text{MMSE}} = \frac{n+1}{n+1+\sum x_i + 1} = \frac{1}{1+\frac{\sum x_i + 1}{n+1}}$

Jeffrey's Prior: if $\theta \sim \mathrm{Beta}(0, \frac{1}{2}), \hat{\theta}_{\text{MMSE}} = \frac{n}{n+\sum x_i + \frac{1}{2}} = \frac{1}{1+\frac{\sum x_i + \frac{1}{2}}{n}}$

Note: Harmonic average: $\frac{1}{\bar{x}} = \frac{1}{n}\sum_i \frac{1}{x}$

In the general case, is there a shrinkage interpretation?

$$\begin{aligned}\frac{1}{\hat{\theta}_{\text{MMSE}}} &= \frac{n+\alpha+\sum x_i+\beta}{n+\alpha} \\ &= \frac{\alpha+\beta}{n+\alpha}\cdot\frac{\alpha}{\alpha}+\frac{\sum x_i+n}{n+\alpha}\cdot\frac{n}{n} \\ &= \frac{\alpha+\beta}{\alpha}\cdot\frac{\alpha}{n+\alpha}+\frac{n+\sum x_i}{n}\cdot\frac{n}{n+\alpha} \\ &= \frac{1}{\text{E}[\theta]}\rho+\frac{1}{\hat{\theta}_{\text{MLE}}}(1-\rho)\end{aligned}$$

Note, if n is small, then there is huge shrinkage; if n is large, $\hat{\theta}_{\text{MMSE}} = \hat{\theta}_{\text{MLE}}$.
Under $n^* = 1$,

$$\begin{aligned}\mathbb{P}(X^* \mid X) &= \int_\Theta \mathbb{P}(X^* \mid \theta)\,\mathbb{P}(\theta \mid X)\,d\theta \\ &= \int_0^1 \Big((1-\theta)^{x^*}\theta\Big)\Big(\frac{1}{B(n+\alpha,\sum x_i+\beta)}\theta^{n+\alpha-1}(1-\theta)^{\sum x_i+\beta-1}\Big)\,d\theta \\ &= \frac{1}{B(n+\alpha,\sum x_i+\beta)}\int_0^1 \theta^{n+\alpha+1-1}(1-\theta)^{x^*+\sum x_i+\beta-1}\,d\theta \\ &= \frac{B(n+\alpha+1,x^*+\sum x_i+\beta)}{B(n+\alpha,\sum x_i+\beta)} \\ &= \text{BetaGeom}(n+\alpha,\sum x_i+\beta)\end{aligned}$$

Let $X_1,\dots,X_n \overset{iid}{\sim} \text{NegBinom}(r,\theta) = \binom{x+r-1}{x}(1-\theta)^x\theta^r$ and $\theta \sim \text{Beta}(\alpha,\beta)$. Then $\theta|X_1,\dots,X_n \sim \text{Beta}(r+\alpha,\sum x_i+\beta)$ and $\mathbb{P}(X^* \mid X) = \text{BetaGeom}(n+\alpha,\sum x_i+\beta)$.

Let $X \sim \text{Binom}(n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. If n is large and θ is small, let $\lambda = n\theta$. Then

$$\begin{aligned}\lim_{n\to\infty}\frac{n!}{x!(n-x)!}(\frac{\lambda}{n})^x(1-\frac{\lambda}{n})^{1-x} &= \frac{\lambda^x}{x!}\lim_{n\to\infty}\frac{n\cdot n-1\cdot n-2\cdot\dots\cdot n-x+1}{n\cdot n\cdot n\cdot\dots\cdot n}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-x} \\ &= \frac{\lambda^x e^{-\lambda}}{x!}\end{aligned}$$

Let $X \sim \text{Poisson}(\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$. $\text{Supp}[X] = \{0,1,\dots\}$, $\lambda \in (0,\infty)$. $\text{E}[X] = \lambda$, $\text{Var}[X] = \lambda$.

Let $X|\theta \sim \text{Poisson}(\theta) = \frac{e^{-\theta}\theta^x}{x!}$.

$$\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta) = \frac{e^{-\theta}\theta^x}{x!}\mathbb{P}(\theta) \propto e^{-\theta}\theta^x\mathbb{P}(\theta)$$

Therefore $\mathbb{P}(\theta) \propto e^{-b\theta}\theta^a$.

$$\mathbb{P}(\theta) = \frac{b^{a+1}}{\Gamma(a+1)}e^{-b\theta}\theta^a$$

Then

$$\theta \sim \text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\theta}\theta^{\alpha-1}$$

Supp$[\theta] = (0, \infty)$, parameter space: $\alpha > 0, \beta > 0$. E$[\theta] = \frac{\alpha}{\beta}$, Var$[\theta] = \frac{\alpha}{\beta^2}$, Mode$[\theta] = \frac{\alpha-1}{\beta}$ if $\alpha \geq 1$ and Med$[\theta] = \text{qgamma}(0.5,\ \alpha, \beta)$.

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &\propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta) \\
&= \frac{e^{-\theta}\theta^x}{x!} \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\theta}\theta^{\alpha-1} \\
&\propto e^{-\theta}\theta^x e^{-\beta\theta}\theta^{\alpha-1} \\
&= e^{-(\beta+1)\theta}\theta^{x+\alpha-1} \\
&\propto \text{Gamma}(x+\alpha, \beta+1)
\end{aligned}$$

Therefore when $X|\theta \sim \text{Poisson}(\theta)$ and $\theta \sim \text{Gamma}(\alpha, \beta)$, $\theta|X \sim \text{Gamma}(x+\alpha, \beta+1)$. We say that the gamma is conjugate prior for the Poisson likelihood.

Let $X_1, \dots, X_n|\theta \overset{iid}{\sim} \text{Poisson}(\theta)$ and $\theta \sim \text{Gamma}(\alpha, \beta)$.

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &\propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta) \\
&= \Big(\prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!}\Big)\Big(\frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\theta}\theta^{\alpha-1}\Big) \\
&= \frac{e^{-\sum_{i=1}^n \theta_i}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\theta}\theta^{\alpha-1} \\
&\propto e^{-n\theta}\theta^{\sum x_i} e^{-\beta\theta}\theta^{\alpha-1} \\
&\propto \text{Gamma}(\sum x_i + \alpha, n+\beta)
\end{aligned}$$

Here α is the total number of successes seen previously and β is the number of pseudotrials performed.

$$\hat{\theta}_{\text{MMSE}} = \frac{\sum x_i + \alpha}{n+\beta}\ \hat{\theta}_{\text{MAE}} = \text{qgamma}(0.5, \sum x_i + \alpha, n+\beta)\ \hat{\theta}_{\text{MAP}} = \frac{\sum x_i + \alpha - 1}{n+\beta} \text{ if } \sum x_i + \alpha \geq 1$$

Can we say that the Laplace prior is $\theta \sim U$? No because the support in infinity and thus not an integratable region. Let's say $\mathbb{P}(\theta) \propto 1$. This is clearly improper and indifferent.

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &\propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta) \\
&\propto e^{-n\theta}\theta^{\sum x_i}\mathbb{P}(\theta) \\
&\propto e^{-n\theta}\theta^{\sum x_i} \\
&= \text{Gamma}(\sum x_i, n)
\end{aligned}$$

Thus if $\theta \sim \text{Gamma}(0,0)$, then the Haldane prior equals the Laplace prior, both of which are improper.

$$\begin{aligned}
\mathcal{L}(\theta; x) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod x_i!} \\
l(\theta; x) &= -n\theta + \sum x_i \ln\theta - \ln(\prod x_i!) \\
l'(\theta; x) = -n + \frac{\sum x_i}{\theta} \overset{\text{set}}{=} 0 \to \frac{\sum x_i}{\theta} = n \to \hat{\theta}_{\text{MLE}} = \bar{x} l''(\theta; x) &= -\frac{\sum x_i}{\theta^2} \\
I(\theta) &= \text{E}[-l''(\theta; x)] = \text{E}[\frac{\sum x_i}{\theta^2}] \\
&= \frac{\text{E}[\sum x_i]}{\theta^2} \\
&= \frac{\sum \text{E}[x_i]}{\theta^2} = \frac{\sum \theta}{\theta^2} = \frac{n\theta}{\theta^2} = \frac{n}{\theta} \\
\mathbb{P}\left(\theta\right) &\propto \sqrt{I(\theta)} = \sqrt{\frac{n}{\theta}} \propto \sqrt{\frac{1}{\theta}} = \theta^{-\frac{1}{2}} \\
&\propto \text{Gamma}(\frac{1}{2}, 0)
\end{aligned}$$

This Jeffrey's prior is improper.

$$\hat{\theta}_{\text{MMSE}} = \frac{\sum x_i + \alpha}{n = \beta} = \frac{\sum x_i}{\beta + n} \cdot \frac{n}{n} + \frac{\alpha}{n + \beta} \cdot \frac{\beta}{\beta} = \frac{n}{n + \beta} \frac{\sum x_i}{n} + \frac{\beta}{n + \beta} \frac{\alpha}{\beta} = \hat{\theta}_{\text{MLE}}(1 - \rho) + \rho\text{E}[\theta]$$

For $n^* = 1$,

$$\begin{aligned}
\mathbb{P}(X^* \mid X) &= \int_\alpha \mathbb{P}(X^* \mid \theta)\,\mathbb{P}(\theta \mid X)\; d\theta \\
&= \int_0^\infty \Big(\frac{e^{-\theta}\theta^{x^*}}{x^*!}\Big)\Big(\frac{\beta'^{\alpha'}}{\Gamma(\alpha')}e^{-\beta'\theta}\theta^{\alpha'-1}\Big)\, d\theta \\
&= \frac{\beta'^{\alpha'}}{\Gamma(\alpha')x^*!}\int_0^\infty \underbrace{e^{-(\beta'+1)\theta}\theta^{x^*+\alpha'-1}}_{\text{kernel of Gamma}(x^*+\alpha',\beta'+1)}\, d\theta \\
&= \frac{\Gamma(\alpha'+x^*)}{(\beta+1)^{x^*+\alpha'}}\int_0^\infty \frac{(\beta'+1)^{x^*+\alpha'}}{\Gamma(\alpha'+x^*)}e^{-(\beta'+1)\theta}\theta^{x^*+\alpha'-1}\, d\theta \\
&= \Big(\frac{\beta'}{\beta'+1}\Big)^{\alpha'}\Big(\frac{1}{\beta'+1}\Big)^{x^*}\frac{\Gamma(x^*+\alpha')}{x^*!\Gamma(\alpha')} \\
&\text{Note that } \frac{\beta'}{\beta'+1} \in (0,1), 1-\frac{\beta'}{\beta'+1} = \frac{1}{\beta'+1} \in (0,1) \\
&\text{Let } p = \frac{\beta'}{\beta'+1}, 1-p = \frac{1}{\beta'+1} \\
&= \frac{\Gamma(x^*+\alpha')}{x^*!\Gamma(\alpha')}(1-p)^{x^*}p^{\alpha} \\
&\text{If } \alpha' \in \mathbb{N}, \Gamma(x^*+\alpha') = (x^*+\alpha'-1)!, \Gamma(\alpha') = (\alpha'1)! \\
&= \frac{(x^*+\alpha'-1)!}{x^*!(\alpha'-1)!}(1-p)^{x^*}p^{\alpha'} \\
&= \binom{x^*+\alpha'-1}{x^*}(1-p)^{x^*}p^{\alpha'} \\
&= \text{NegBinom}(\alpha', p) \\
&= \text{NegBinom}(\sum x_i + \alpha, \frac{n+\beta}{n+\beta+1})
\end{aligned}$$

Let $X|\theta \sim \text{Gamma}(1,\theta) = \frac{\theta^1}{\Gamma(1)}e^{-\theta x}\theta^{1-1} = \text{Exp}(\theta)$. Let $\theta \sim \text{Gamma}(\alpha,\beta)$.

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &\propto \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta) \\
&= \underbrace{\theta e^{-\theta x}}_{\text{gamma kernel}} \underbrace{\mathbb{P}(\theta)}_{\text{should also be gamma kernel}} \\
&= \theta e^{-\theta x}\frac{\beta^\alpha}{\Gamma(\alpha)}e^{-\beta\theta}\theta^{\alpha-1} \\
&\propto e^{-(\beta+x)\theta}\theta^{\alpha+1-1} \\
&\propto \text{Gamma}(\alpha+1, \beta+x)
\end{aligned}$$

Therefore if $X|\theta \sim \text{Exp}(\theta)$, $\theta \sim \text{Gamma}(\alpha,\beta)$, then $\theta|X \sim \text{Gamma}(\alpha+1,\beta+x)$. In addition, $\theta|X_1,\dots,X_n \sim \text{Gamma}(\alpha+n,\beta+\sum x_i)$.
Gamma is conjugacy for the exponential likelihood.

Let $X|\theta \sim \text{Gamma}(r,\theta) = \frac{\theta^r}{\Gamma(r)}e^{-\theta x}x^{r-1} = \frac{\theta^r}{(r-1)!}e^{-\theta x}x^{r-1} = \text{Erlang}(r,\theta)$. Then

$$\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta) = \Big(\frac{\theta^r}{(r-1)!}e^{-\theta x}\theta^{r-1}\Big)\mathbb{P}(\theta) \propto \theta^r e^{-\theta x}\mathbb{P}(\theta)$$

Gamma is conjugate for the gamma likelihood with fixed α.

$$\mathbb{P}(\theta \mid X, r) \propto \mathbb{P}(X \mid \theta, r)\mathbb{P}(\theta, r) \text{ because } r \text{ is considered known.}$$

Let $X|\theta,\sigma^2 \sim N(\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma}(x-\theta)^2}$. $\text{E}[X] = \theta$. $\text{Var}[X] = \theta^2$. $\text{Supp}[X] = \mathbb{R}$. Parameter space: $\theta \in \mathbb{R}$ and $\sigma^2 \in (0,\infty)$.

$$\begin{aligned} X|\theta,\sigma^2 &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma}(x-\theta)^2} \\ &\propto e^{-\frac{1}{2\sigma^2}(x-\theta)^2} \\ &= e^{-\frac{x^2}{2\sigma^2}+\frac{\theta x}{\sigma^2}-\frac{\theta^2}{2\sigma^2}} \\ &= e^{-\frac{x^2}{2\sigma^2}}e^{\frac{\theta x}{\sigma^2}}e^{-\frac{\theta^2}{2\sigma^2}} \\ &\propto e^{-\frac{x^2}{2\sigma^2}}e^{\frac{\theta x}{\sigma^2}} \end{aligned}$$

Given $X_1,\dots,X_n|\theta,\sigma^2 \overset{iid}{\sim} N(\theta,\sigma^2)$ and assuming σ^2 is known,

$$\begin{aligned} \mathcal{L}(\theta;x,\sigma^2) &= \prod_{i=1}^n \mathbb{P}\left(X \mid \theta,\sigma^2\right) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\theta)^2} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x-\theta_i)^2} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i^2-2\theta x_i+\theta^2)} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n e^{-\frac{1}{2\sigma^2}(\sum x_i^2+2\theta\sum x_i+n\theta^2)} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n e^{-\frac{1}{2\sigma^2}(\sum x_i^2-2\theta n\bar{x}+n\theta^2)} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n e^{-\frac{\sum x_i^2}{2\sigma^2}}e^{\frac{\theta\bar{x}n}{\sigma^2}}e^{-\frac{n\theta^2}{2\sigma^2}} \\ l(\theta;x,\sigma^2) &= n\ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big) - \frac{\sum x_i^2}{2\sigma^2} + \frac{\theta\bar{x}n}{\sigma^2} - \frac{n\theta^2}{2\sigma^2} \\ l'(\theta;x,\sigma^2) &= \frac{\bar{x}n}{\sigma^2} - \frac{n\theta}{\sigma^2} \\ &\overset{\text{set}}{=} 0 \\ \hat{\theta}_{MLE} &= \bar{x} \end{aligned}$$

$$\begin{aligned}
\mathbb{P}\left(\theta \mid X, \sigma^2\right) &= \mathbb{P}\left(X \mid \theta, \sigma^2\right)\mathbb{P}\left(X \mid \sigma^2\right) \\
&\propto \mathbb{P}\left(X \mid \theta, \sigma^2\right)\mathbb{P}\left(\theta \mid \sigma^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum x_i^2}{2\sigma^2}} e^{\frac{\theta\bar{x}n}{\sigma^2}} e^{-\frac{n\theta^2}{2\sigma^2}} \mathbb{P}\left(\theta \mid \sigma^2\right) \\
&\propto e^{\frac{\theta\bar{x}n}{\sigma^2}} e^{-\frac{n\theta^2}{2\sigma^2}} \mathbb{P}\left(\theta \mid \sigma^2\right) \\
&= \underbrace{e^{-\frac{n}{2\sigma^2}} e^{\frac{\bar{x}n}{\sigma^2}\theta} e^{-\frac{n}{2\sigma^2}\theta^2}}_{\text{kernel for normal}} \mathbb{P}\left(\theta \mid \sigma^2\right)
\end{aligned}$$

What's $\mathbb{P}\left(\theta \mid \sigma^2\right)$?

$$\begin{aligned}
\mathbb{P}\left(\theta \mid \sigma^2\right) &= N(\mu_0, \tau^2) \\
&= \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(x-\mu_0)^2} \\
&\propto e^{-\frac{1}{2\tau^2}(\theta^2 - 2\mu_0\theta + 2\mu_0)} \\
&\propto e^{-\frac{1}{2\tau^2}\theta^2} e^{\frac{\mu_0}{\tau^2}\theta}
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{P}\left(\theta \mid X, \sigma^2\right) &\propto \left(e^{-\frac{n}{2\sigma^2}\theta^2} e^{\frac{\bar{x}n}{\sigma^2}\theta}\right)\left(e^{-\frac{1}{2\tau^2}\theta^2} e^{\frac{\mu_0}{\tau^2}\theta}\right) \\
&= e^{-(\frac{n}{2\sigma^2} + \frac{1}{2\tau^2})\theta^2} e^{(\frac{\bar{x}n}{\sigma^2} + \frac{\mu_0}{\tau^2})\theta}
\end{aligned}$$

Let c and v^2 be constants. Then

$$\begin{aligned}
N(c, v^2) &= \frac{1}{\sqrt{2\pi v^2}} e^{-\frac{1}{2v^2}(x-c)^2} \\
&\propto e^{-\frac{1}{2v^2}\theta^2} e^{\frac{c}{v^2}\theta} e^{-\frac{c^2}{2v^2}} \\
-\frac{1}{2v^2} &= -(\frac{n}{2\sigma^2} + \frac{1}{2\tau^2}) \to \frac{1}{v^2} = \frac{n}{\sigma^2} + \frac{1}{\tau^2} \\
v^2 &= \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \\
\frac{c}{v^2} &= \frac{\bar{x}n}{\sigma^2} + \frac{\mu_0}{\tau^2} \\
c &= \left(\frac{\bar{x}n}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)v^2 = \frac{\frac{\bar{x}n}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}
\end{aligned}$$

Therefore if $X_1, \dots, X_n | \theta \overset{iid}{\sim} N(\theta, \sigma^2)$ and $\theta | X_1, \dots, X_n, \sigma^2 \sim N(\mu_0, \tau^2)$ then

$$\theta | X_1, \dots, X_n, \tau^2 \sim N(\underbrace{\frac{\frac{\bar{x}n}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}}_{\theta_p}, \underbrace{\frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}}_{\sigma_p'})$$

This is the normal-normal conjugacy model. The normal is conjugate for the normal likelihood when σ^2 is known. μ_0 is the prior mean and τ^2 is the prior variance.

$$\hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MAE}} = \hat{\theta}_{\text{MAP}} = \frac{\frac{\bar{x}n}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

Using $\hat{\theta}_{\text{MMSE}}$ as a shrinkage estimator

$$\begin{aligned}
\hat{\theta}_{\text{MMSE}} &= \frac{\frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \cdot \frac{\tau^2}{\tau^2} + \frac{\frac{\bar{x}n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \cdot \frac{\frac{\sigma^2}{n}}{\frac{\sigma}{n}} \\
&= \frac{1}{\frac{n\tau^2}{\sigma^2} + 1}\mu_0 + \frac{1}{1 + \frac{\sigma^2}{n\tau^2}}\bar{x} \\
&= \frac{\sigma^2}{n\tau^2 + \sigma^2}\mu_0 + \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{x} \\
&= \rho \mathrm{E}[\theta] + (1-\rho)\hat{\theta}_{\text{MLE}}
\end{aligned}$$

This is a weighed arithmetic average shrinkage.

$$\lim_{n\to\infty} \rho = 0$$

Imagine you see n_0 previous trials with σ^2 known. Let $\mu_0 = \bar{y} = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i$. Let $\tau^2 = \frac{\sigma^2}{n_0}$. Then

$$\begin{aligned}
\theta_p &= \frac{\frac{\bar{x}n}{\sigma^2} + \frac{\bar{y}n_0}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{n_0}{\tau^2}} \\
&= \frac{\bar{x}n + \bar{y}n_0}{n + n_0} \\
&= \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n_0} y_0}{n + n_0}
\end{aligned}$$

Therefore if $X_1, \dots, X_n | \theta, \sigma^2 \sim N(\theta, \sigma^2)$ then $\theta \sim \sigma^2 \sim N(\mu_0, \frac{\sigma^2}{n_0})$. This is the posterior average of all prior data. Furthemore,

$$\theta \sim X_1, \dots, X_n, \sigma^2 \sim N(\frac{\bar{x}n + \bar{y}n_0}{n + n_0}, \Big(\frac{\sigma}{\sqrt{n + n_0}}\Big)^2)$$

Laplace prior for $\theta|\sigma^2$ - $\mathbb{P}(\theta \mid \sigma^2) \propto 1$ - improper.

$$\begin{aligned}
\mathbb{P}\left(\theta \mid X, \sigma^2\right) &\propto \mathbb{P}\left(X \mid \theta, \sigma^2\right)\mathbb{P}\left(\theta \mid \sigma^2\right) \\
&\propto \mathbb{P}\left(X \mid \theta, \sigma^2\right) \\
&\propto \underbrace{e^{\frac{\bar{x}n}{\sigma^2}\theta}}_{\frac{c}{v^2}} \underbrace{e^{-\frac{n}{2\sigma^2}\theta^2}}_{\overline{2v^2}} \\
\frac{1}{2v^2} &= \frac{n}{2\sigma^2} \to v^2 = \frac{\sigma^2}{n} \\
\frac{c}{v^2} &= \frac{\bar{x}n}{\sigma^2} \to c = \frac{\bar{x}n}{\sigma^2}v^2 = \frac{\bar{x}n}{\sigma^2} \cdot \frac{\sigma^2}{n} = \bar{x} \\
\mathbb{P}\left(\theta \mid X, \sigma^2\right) &\propto N(\bar{x}, \frac{\sigma^2}{n})
\end{aligned}$$

This is always a proper posterior. In addition, under the Laplace prior,

$$\hat{\theta}_{\text{MMSE}} = \hat{\theta}_{\text{MAE}} = \hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}} = \bar{x}$$

What's the Jeffrey's prior?

$$\begin{aligned}
l'(\theta; X, \sigma^2) &= \frac{\bar{x}n}{\sigma^2} - \frac{n\theta}{\sigma^2} \\
l''(\theta; X, \sigma^2) &= -\frac{n}{\sigma^2} \\
I(\theta) &= \mathrm{E}[-l''(\theta; X, \sigma^2)] = \mathrm{E}[\frac{n}{\sigma^2}] = \frac{n}{\sigma^2} \\
\mathbb{P}\left(\theta \mid \sigma^2\right) &\propto \sqrt{I(\theta)} = \sqrt{\frac{n}{\sigma^2}} \propto 1
\end{aligned}$$

This is the Laplace prior.
Note that improper priors can be thought as limits of proper priors.

Let $X|\theta \sim \text{Binom}(n, \theta)$, $\theta \sim \text{Beta}(\alpha, \beta)$ and $\theta|X \sim \text{Beta}(x + \alpha, n - x + \beta)$. Then

$$\lim_{\alpha \to 0 \beta \to 0} \mathbb{P}(\theta \mid X) = \text{Beta}(x, n - x)$$

Let $X_1, \dots, X_n|\theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2), \theta|\sigma^2 \sim N(\mu_0, \tau^2)$ and $\theta|X_1, \dots, X_n, \sigma^2 \sim N\left(\frac{\frac{\bar{x}n}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right) = N(\hat{\theta}_{\text{MMSE}}, \sigma_p^2)$. Then

$$\begin{aligned}
\lim_{\tau^2 \to \infty} \mathbb{P}\left(\theta \mid X_1, \dots, X_n, \sigma^2\right) &= N(\bar{x}, \frac{\sigma^2}{n}) \\
\lim_{\tau^2 \to \infty} \frac{\frac{\bar{x}n}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \cdot \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}} &= \lim_{\tau^2 \to \infty} \frac{\bar{x} + \frac{\mu_0 \sigma^2}{\tau^2 n}}{1 + \frac{\sigma^2}{n\tau^2}} = \bar{x} \\
\lim_{\tau^2 \to \infty} \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} &= \frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n} \\
\lim_{\tau^2 \to \infty} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\theta - \mu_0)^2} &= 0 \\
\mathbb{P}\left(\theta \mid \sigma^2\right) &\propto 1
\end{aligned}$$

For $n^* = 1$,

$$\begin{aligned}
\mathbb{P}\left(X^* \mid X, \sigma^2\right) &= \int_\Theta \mathbb{P}\left(X^* \mid \theta, \sigma^2\right) \mathbb{P}\left(\theta \mid X, \sigma^2\right) \, d\theta \\
&= \int_R \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^* - \theta)^2} \cdot \frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{1}{2\sigma_p^2}(\theta - \theta_p)^2} \, d\theta \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_p^2}} \int_R e^{-\frac{1}{2\sigma^2}(x^* - \theta)^2 - \frac{1}{2\sigma_p^2}(\theta - \theta_p)^2} \, d\theta
\end{aligned}$$

Let $X_1, X_2 \overset{iid}{\sim} U(\{1,2,3,4,5,6\})$. What is $S = X_1 + X_2 \sim ?$

$$\mathbb{P}(S=1) = 0$$

$$\mathbb{P}(S=1) = \mathbb{P}(X_1 = 1) \cdot \mathbb{P}(X_2 = 1) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

$$\mathbb{P}(S=3) = \mathbb{P}(X_1 = 1)\,\mathbb{P}(X_2 = 2) + \mathbb{P}(X_1 = 2)\,\mathbb{P}(X_2 = 1) = \sum_{x \in \text{Supp}[X]} \mathbb{P}(X_1 = x)\,\mathbb{P}(X_2 = 3 - x)$$

$$\begin{aligned} \mathbb{P}(S = s) &= \sum_{x \in \text{Supp}[X]} \mathbb{P}(X_1 = x)\,\mathbb{P}(X_2 = s - x) \\ &= \sum_{x \in \text{Supp}[X]} \mathbb{P}(X_2 = x)\,\mathbb{P}(X_1 = s - x) \end{aligned}$$

Since it is iid, order does not matter.

For continuous random variables

$$S = X_1 + X_2 \sim \int_{\text{Supp}[X]} f_{x_1}(x) f_{x_2}(s-x)\,dx = f_{x_1} * f_{x_2}$$

Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. Then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Furthermore

$$f_{x_1} * f_{x_2} = \int_R f_{x_1}(x) f_{x_2}(s-x)\,dx = \int_R \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{3\sigma_1^2}(x-\mu_1)^2} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(s-x-\mu_2)^2}\,dx = \int_R \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-\frac{}{2(\sigma}}$$

Hence

$$\begin{aligned} \mathbb{P}\left(X^* \mid X, \sigma^2\right) &= \int_R \frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{1}{2\sigma_p^2}(\theta-\theta_p)^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^*-\theta-0)}\,d\theta \\ &= N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\ &= N(\theta_p, \sigma_p^2 + \sigma^2) \end{aligned}$$

If Jeffrey's prior, the posterior predictive distribution is

$$\mathbb{P}\left(X^* \mid X, \sigma^2\right) = N(\theta_p, \sigma_p^2 + \sigma^2) = N(\bar{x}, \frac{\sigma^2}{n} + \sigma^2)$$

Let $X_1, \dots, X_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, with θ known and σ^2 unknown. What's the MLE for σ^2?

$$\begin{aligned}
\mathcal{L}(\sigma^2; X, \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2} \\
&= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^2 e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2} \\
l(\sigma^2; X, \theta) &= n\ln(\frac{1}{\sqrt{2\pi}}) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum (x_i-\theta)^2 \\
l'(\sigma^2; X, \theta) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (x_i-\theta)^2 = 0
\end{aligned}$$

$$-n + \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\theta)^2 = 0$$

$$\hat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\underbrace{\sum_{i=1}^n (x_i-\theta)^2}_{\text{sum of squared error}} = \frac{SSE}{n}$$

Let $\theta \sim \text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\theta}\theta^{\alpha-1}$. If $Y = \frac{1}{\theta} = t(\theta)$, what is $Y \sim$? $\theta = t^{-1}(y) = \frac{1}{y}$. Then

$$\begin{aligned}
f_Y(y) &= f_\theta(t^{-1}(y))\Big|\frac{d}{dy}[t^{-1}(y)]\Big| \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\frac{\beta}{y}}(\frac{1}{y})^{\alpha-1}\Big|\frac{d}{dy}[y^{-1}]\Big| \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\frac{\beta}{y}} y^{-\alpha-1} \\
&= \text{InvGamma}(\alpha, \beta)
\end{aligned}$$

If $Y \sim \text{InvGamma}(\alpha, \beta)$,

$$\begin{aligned}
\text{E}[y] &= \frac{\beta}{\alpha-1} \text{ if } \alpha > 1 \\
\text{Med}(y) &= \text{qinvgamma}(0.5, \alpha, \beta) \\
\text{Mode}(y) &= \frac{\beta}{\alpha+1} \\
\text{Supp}[Y] &= (0, \infty) \\
\text{Parameter Space } &: \alpha,\ \beta > 0
\end{aligned}$$

What's $\mathbb{P}(\sigma^2 \mid X, \theta)$?

$$\begin{aligned}
\mathbb{P}\left(\sigma^2 \mid X, \theta\right) &\propto \mathbb{P}\left(X \mid \theta, \sigma^2\right)\mathbb{P}\left(\sigma^2 \mid \theta\right) \\
&= \Big(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2}\Big)\mathbb{P}\left(\sigma^2 \mid \theta\right) \\
&= \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \Big(\sigma^2\Big)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2}\mathbb{P}\left(\sigma^2 \mid \theta\right) \\
&\propto \underbrace{(\sigma^2)^{-\frac{n}{2}} e^{-\frac{n\hat{\sigma^2}_{\text{MLE}}}{2\sigma^2}}}_{\text{kernel of InvGamma}} \mathbb{P}\left(\sigma^2 \mid \theta\right) \\
&\propto \text{InvGamma}(\frac{n}{2} - 1, \frac{n\hat{\sigma^2}_{\text{MLE}}}{2})
\end{aligned}$$

Therefore if $\sigma^2|\theta \sim \text{InvGamma}(\alpha, \beta)$,

$$\begin{aligned}
\mathbb{P}\left(\sigma^2 \mid X, \theta\right) &\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{\frac{n\hat{\sigma^2}}{2}}{\sigma^2}} \cdot (\sigma^2)^{-\alpha-1} e^{-\frac{\beta}{\sigma^2}} \\
&= (\sigma^2)^{-\frac{n}{2}-\alpha-1} e^{-\frac{(\frac{n\hat{\sigma^2}}{2}+\beta)}{\sigma^2}} \\
&\propto \text{InvGamma}\Big(\frac{n}{2} + \alpha, \frac{n\hat{\sigma^2_{\text{MLE}}}}{2} + \beta\Big)
\end{aligned}$$

If we let $\sigma^2 \sim \text{InvGamma}(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2})$, then

$$\mathbb{P}\left(\sigma^2 \mid X, \theta\right) = \text{InvGamma}\Big(\frac{n+n_0}{2}, \frac{n\hat{\sigma^2} + n_0\hat{\sigma_0^2}}{2}\Big)$$

Here n_0 is the number of prior trials and $n_0\sigma_0^2$ is the prior SSE. Therefore if

$$\sigma_0^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(Y_i - \theta)^2$$

, then

$$n_0\sigma_0^2 = \sum_{i=1}^{n_0}(Y_i - \theta)^2 = \text{SSE}_0$$

Hence

$$\sigma^2|X, \theta \sim \text{InvGamma}(\underbrace{\frac{n+n_0}{2}}_{\alpha'}, \underbrace{\frac{SSE + SSE_0}{2}}_{\beta'})$$

Imagine prior data: $Y_1, \dots, Y_{n_0} | \theta, \sigma^2 \sim N(\theta, \sigma^2)$, where θ is known, then

$$\begin{aligned}
\hat{\sigma^2}_{\text{MMSE}} &= \text{E}[\sigma^2 | X, \theta] = \frac{\alpha}{\beta - 1} = \frac{\frac{n\sigma^2_{\text{MLE}} + n_0\sigma_0^2}{2}}{\frac{n+n_0}{2} - 1} \\
&= \frac{n\hat{\sigma^2}_{\text{MLE}} + n_0\sigma_0^2}{n + n_0 - 2} \\
\hat{\sigma^2}_{\text{MAP}} &= \frac{n\hat{\sigma^2}_{\text{MLE}} + n_0\sigma_0^2}{n + n_0 - 2} \\
\hat{\sigma^2}_{\text{MAE}} &= \text{qinvgamma}(0.5, \frac{n + n_0}{2}, \frac{n\hat{\sigma^2} + n_0\sigma_0^2}{2})
\end{aligned}$$

Uninformative prior: Let $n_0 = 0$. Then $\sigma^2 \sim \text{InvGamma}(0, 0)$ - which is improper. But if we go along with it, $\sigma^2 | X, \theta \sim \text{InvGamma}(\frac{n}{2}, \frac{n\hat{\sigma^2}_{textMLE}}{2})$ which is always proper.

$$\hat{\sigma^2}_{\text{MMSE}} = \frac{\frac{n\hat{\sigma^2}}{2}}{\frac{n}{2} - 1} = \frac{n\hat{\sigma^2}}{n - 2} = \frac{n - 2}{\sum}(x_i - \theta)^2 \approx \hat{\sigma^2}_{\text{MLE}}$$

Another uninformative prior is $\sigma^2 | \theta \sim \text{InvGamma}(2, 0)$. Continue with it.

$$|sigma^2 | X_1, \dots, X_n, \theta \sim \text{InvGamma}(\frac{n + 2}{2}, \frac{n\hat{\sigma^2}}{2})$$

Furthermore

$$\hat{\sigma^2}_{\text{MMSE}} = \frac{\frac{n\sigma^2}{2}}{\frac{n+2}{2} - 1} = \hat{\sigma^2}_{\text{MLE}}$$

What's Jeffrey's prior?

$$\begin{aligned}
\mathbb{P}\left(\sigma^2 \mid \theta\right) &\propto \sqrt{I(\sigma^2)} \\
l'(\sigma^2; X, \theta) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}SSE = -\frac{n}{2}(\sigma^2)^{-1} + \frac{SSE}{2}(\sigma^2)^{-2} \\
l''(\sigma^2; X, \theta) &= \frac{n}{2}(\sigma^2)^{-2} - SSE(\sigma^2)^{-3} \\
I(\sigma^2) &= \text{E}[-l''(\sigma^2; X, \theta)] = \text{E}[-\frac{n}{2}(\sigma^2)^{-2} + SSE(\sigma^2)^{-3}] \\
&= -\frac{n}{2}(\sigma^2)^{-2} + (\sigma^2)^{-3}\text{E}[SSE] \\
\text{E}[SSE] &= \text{E}[\sum_{i=1}^n (x_i - \theta)^2] = \sum_{i=1}^n \text{E}[(x_i - \theta)^2] = n\text{E}[(X - \theta)^2] = n\text{Var}[X] = n\sigma^2 \\
I(\sigma^2) &= -\frac{n}{2}(\sigma^2)^{-2} + (\sigma^2)^{-3}(n\sigma^2) = -\frac{n}{2}(\sigma^2)^{-2} = n(\sigma^2)^{-2} = (n - \frac{n}{2})(\sigma^2)^{-2} \\
\mathbb{P}\left(\sigma^2 \mid X\right) &\propto \sqrt{\frac{n}{2}(\sigma^2)^{-2}} \propto (\sigma^2)^{-1} = \text{InvGamma}(0, 0)
\end{aligned}$$

This is an improper prior.

End of Midterm 2 Material

Let $X_1, \dots, X_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$. Let both θ and σ^2 be unknown.

$$\begin{aligned}
\mathbb{P}\left(\theta, \sigma^2 \mid X_1, \dots, X_n\right) &\propto \mathbb{P}\left(X_1, \dots, X_n \mid \theta, \sigma^2\right) \mathbb{P}\left(\theta, \sigma^2\right) \\
&\propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2} \mathbb{P}\left(\theta, \sigma^2\right) \\
&\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2} \mathbb{P}\left(\theta, \sigma^2\right)
\end{aligned}$$

This is not the kernel of InvGamma. Consider the following:

$$\begin{aligned}
SSE &= \sum_{i=1}^n (x_i - \theta)^2 \\
&= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \theta)^2 \\
&= \sum_{i=1}^n \left((x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \theta) + (\bar{x} - \theta)^2\right) \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 + 2\sum_{i=1}^n (x_i\bar{x} - x_i\theta - \bar{x}^2 + \bar{x}\theta) + n\sum_{i=1}^n (x_i - \theta)^2 \\
&\text{Note that } s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2 \\
&= (n-1)s^2 + 2(\bar{x}\sum x_i - \theta\sum x_i - \sum \bar{x}^2 + \theta\sum x_i) + n(\bar{x} - \theta)^2 \\
&= (n-1)s^2 + 2(n\bar{x}^2 - \theta\bar{x}n - n\bar{x}^2 + \theta\bar{x}n) + n(\bar{x} - \theta)^2 \\
&= (n-1)s^2 + n(\bar{x} - \theta)^2 \\
&\propto \mathbb{P}\left(X \mid \theta, \sigma^2\right) \mathbb{P}\left(\theta, \sigma^2\right) \\
\mathbb{P}\left(\sigma^2, \theta \mid X\right) &= (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{x}-\theta)^2\right)} \\
&= \underbrace{(\sigma^2)^{-\frac{n}{2}} e^{-\frac{\frac{(n-1)s^2}{2}}{\sigma^2}} e^{-\frac{1}{2\frac{\sigma^2}{n}}(\bar{x}-\theta)^2}}_{\propto \text{NormInvGamma}(\mu = \bar{x},\ \lambda = n,\ \alpha = \frac{n}{2}+1,\ \beta = \frac{(n-1)s^2}{2})} \mathbb{P}\left(\theta, \sigma^2\right)
\end{aligned}$$

Therefore $\mathbb{P}(\theta, \sigma^2)$ should also be NormInvGamma (conjugacy). Note that NormInvGamma is the conjugate prior for normal likelihood where both θ and σ^2 are unknown.

Jeffrey's prior: $\mathbb{P}(\theta, \sigma^2) = \mathbb{P}(\theta \mid \sigma^2)\mathbb{P}(\sigma^2) \propto (1)(\frac{1}{\sigma^2}) = \frac{1}{\sigma^2}$. Then

$$\mathbb{P}\left(\theta, \sigma^2 \mid X\right) \propto \text{NormInvGamma}(\bar{x}, n, \frac{n}{2}, \frac{(n-1)s^2}{2})$$

How to simulate from NormInvGamma distribution? Assuming Jeffrey's prior,

$$\begin{aligned}
\mathbb{P}\left(\theta \mid X, \sigma^2\right) &= \frac{\mathbb{P}\left(\theta, \sigma^2 \mid X\right)}{\mathbb{P}\left(\sigma^2 \mid X\right)} \propto \mathbb{P}\left(\theta, \sigma^2 \mid X\right) \\
&= (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{(n-1)s^2}{2\sigma^2}} e^{-\frac{1}{2\frac{\sigma^2}{n}}(\bar{x}-\theta)^2} \\
&\propto e^{-\frac{1}{2\frac{\sigma^2}{n}}(\bar{x}-\theta)^2} \\
&\propto N(\bar{x}, \frac{\sigma^2}{n}) \\
\mathbb{P}\left(\sigma^2 \mid X\right) &= \frac{\mathbb{P}\left(\theta, \sigma^2 \mid X\right)}{\mathbb{P}\left(\theta \mid X, \sigma^2\right)} \\
&\propto \frac{(\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{(n-1)s^2}{2\sigma^2}} e^{-\frac{1}{2\frac{\sigma^2}{n}}(\bar{x}-\theta)^2}}{\frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} e^{-\frac{1}{2\frac{\sigma^2}{n}}(\bar{x}-\theta)^2}} \\
&\propto \frac{(\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{(n-1)s^2}{2\sigma^2}}}{(\sigma^2)^{-\frac{1}{2}}} \\
&= (\sigma^2)^{-\frac{n}{2}-\frac{1}{2}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \\
&\propto \text{InvGamma}(\frac{n-1}{2}, \frac{(n-1)s^2}{2})
\end{aligned}$$

Note that

$$\mathbb{P}\left(\sigma^2 \mid X, \theta\right) = \text{InvGamma}(\frac{n}{2}, \frac{n\hat{\sigma^2}_{\text{MLE}}}{2})$$

Let $X_1, \dots, X_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$ and $\mathbb{P}(\theta, \sigma^2) \propto \frac{1}{\sigma^2}$. Let θ and σ^2 be unknown.
If σ^2 is known, $\mathbb{P}(\theta \mid X, \sigma^2) = N\Big(\bar{x}, \Big(\frac{\sigma}{\sqrt{n}}\Big)^2\Big)$.
If θ is known, $\mathbb{P}(\sigma^2 \mid X, \theta) = \text{InvGamma}\Big(\frac{n}{2}, \frac{n\hat{\sigma^2}_{\text{MLE}}}{2}\Big)$.
If both are unknown,

$$\begin{aligned}
\mathbb{P}\left(\theta, \sigma^2 \mid X\right) &\propto \mathbb{P}\left(X \mid \theta, \sigma^2\right) \mathbb{P}\left(\theta, \sigma^2\right) \\
&= \Big(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2}\Big)\Big(\frac{1}{\sigma^2}\Big) \\
&\propto (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{(n-1)s^2}{2\sigma^2}} e^{-\frac{n}{2\sigma^2}(\bar{x}-\theta)} \\
&\propto \text{NormInvGamma}(\mu = \bar{x}, \lambda = n, \alpha = \frac{n}{2}, \beta = \frac{(n-1)s^2}{2})
\end{aligned}$$

Sampling:

- How do you sample $X \sim \text{Bern}(0.5)$? Toss a coin.
- How do you sample $X \sim \text{Binom}(10, 0.5)$? Toss 10 coins.

Recalling that $F(x) = \mathbb{P}(X \leq x)$ (cdf), for a continuous random variable, what is the distribution of $Y = F(X)$?

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(x)\frac{1}{\left|\frac{dy}{dx}\right|} = f_X(x)\left|\frac{1}{\frac{d}{dx}[F(x)]}\right| = f_X(x)\frac{1}{|f_X(x)|} = 1$$

Note that $\text{Supp}(Y) = [0,\, 1]$ and $f_Y(y) = 1$, then $Y \sim U(0,1)$. Furthermore, $X + F^{-1}(Y)$. To sample x^*,

1. Sample y_0^* from $U(0,1)$
2. Compute $x_0 = F^{-1}(y_0)$
3. Return x_0

What if F^{-1} is not available in closed form? Pick a x_{min}, x_{max} and Δx. Using this, create a "grid"

$$\mathcal{G} = \langle x_{\text{min}}, x_{\text{min}} + \Delta x, x_{\text{min}} + 2\Delta x, \dots, x_{\text{max}} \rangle$$

Express $F(x) \forall x \in \mathcal{G}$. Approximate $x_0 \approx \min_{x^* \in \mathcal{G}} F(x^*) \geq y$. What if X is discrete? Let $\mathcal{G} = \text{Supp}[X]$ where X is not approximate.
We know how to sample from $f(x)$ but how do we sample from $f(x,y)$? Recall Bayes Rule: $f(x,y) = f(y|x)f(x)$.
To sample,

1. Draw x_0 from $f(x)$
2. Draw y_0 from $f(y|x = x_0)$
3. return $\langle x_0, y_0 \rangle$

Can we do this with the NormInvGamma?

$$\begin{aligned}
\mathbb{P}\left(\theta, \sigma^2 \mid X\right) &= \mathbb{P}\left(\theta \mid X, \sigma^2\right)\mathbb{P}\left(\sigma^2 \mid X\right) \\
\mathbb{P}\left(\theta \mid X, \sigma^2\right) &= N\Big(\bar{x}, \Big(\frac{\sigma^2}{\sqrt{n}}\Big)^2\Big) \\
\mathbb{P}\left(\sigma^2 \mid X\right) &= \frac{\mathbb{P}\left(\theta, \sigma^2 \mid X\right)}{\mathbb{P}\left(\theta \mid \sigma^2, X\right)} \\
&\propto \frac{(\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{(n-1)s^2}{2\sigma^2}} e^{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2}}{\frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} e^{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2}} \\
&\propto \frac{(\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{(n-1)s^2}{2\sigma^2}}}{(\sigma^2)^{-\frac{1}{2}}} \\
&= (\sigma^2)^{-\frac{n}{2}-\frac{1}{2}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \\
&\propto \text{InvGamma}\Big(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\Big)
\end{aligned}$$

Thus to sample from $N(\theta, \sigma^2|X)$

1. Sample σ_0^2 from $\text{InvGamma}\Big(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\Big)$

2. Sample θ_0 from $\mathbb{P}(\theta \mid X, \sigma^2 = \sigma_0^2) = N\Big(\bar{x}, \Big(\frac{\sigma_0}{\sqrt{n}}\Big)^2\Big)$

3. Return $\langle \theta_0, \sigma_0^2 \rangle$

Note: No need to ever work with NormInvGamma.
What about the other term? If $\mathbb{P}(\theta, \sigma^2) = \frac{1}{\sigma^2}$,

$$\mathbb{P}\left(\sigma^2 \mid X\right) = \text{InvGamma}\Big(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\Big) \neq \mathbb{P}\left(\sigma^2 \mid X, \theta\right) = \text{InvGamma}\Big(\frac{n}{2}, \frac{n\hat{\sigma^2}_{\text{MLE}}}{2}\Big)$$

$$\mathbb{P}\left(\sigma^2 \mid X\right) = \int_R \mathbb{P}\left(\sigma^2, \theta \mid X\right)\, d\theta$$

It is the posterior of σ^2 with the uncertainty unknown in ignorance of θ "averaged" over or margined over. In the other scenario, $\mathbb{P}(\theta \mid X)$ is the posterior of θ with the uncertainty in σ^2 averaged or margined out. σ^2 is a "nuisance parameter." Thus

$$\mathbb{P}(\theta \mid X) = \int_0^\infty \mathbb{P}\left(\theta, \sigma^2 \mid X\right)\, d\sigma^2 = \frac{\mathbb{P}(\theta, \sigma^2 \mid X)}{\mathbb{P}(\sigma^2 \mid \theta, X)}$$

If $X_1, \dots, X_n | \theta, \sigma^2 \stackrel{iid}{\sim} N(\theta, \sigma^2)$, $\frac{\bar{x}-\theta}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. What about $\frac{\bar{x}-\theta}{\frac{s}{\sqrt{n}}} \sim$? Use student T distribution.
Let $V \sim T_n := \frac{\Gamma\Big(\frac{n+1}{2}\Big)}{\sqrt{\pi n}\Gamma\Big(\frac{n}{2}\Big)}\Bigg(1 + \frac{v^2}{n}\Bigg)$ be Student T's distribution, or the Standard T distribution. It can be shown that

$$\frac{\bar{x}-\theta}{\frac{s}{\sqrt{n}}} \sim T_{n-1}$$

Let $W = \sigma V + \mu = t(v)$. Then $v = t^{-1}(w) = \frac{w-\mu}{\sigma}$.

$$\begin{aligned} f_W(w) &= f_V(t^{-1}(w))\Big|\frac{d}{dw}[t^{-1}(w)]\Big| \\ &= \frac{\Gamma\Big(\frac{n+1}{2}\Big)}{\sqrt{\pi n}\Gamma\Big(\frac{n}{2}\Big)}\Bigg(1 + \frac{(\frac{w-\mu}{\sigma})^2}{n}\Bigg)^{-\frac{n+1}{2}}\frac{1}{\sigma} \\ &= \frac{\Gamma\Big(\frac{n+1}{2}\Big)}{\sqrt{\pi n}\Gamma\Big(\frac{n}{2}\Big)}\Bigg(1 + \frac{1}{n}(\frac{w-\mu}{\sigma})^2\Bigg)^{-\frac{n+1}{2}} \\ &:= T_n(\mu, \sigma) \end{aligned}$$

Now solve for $\mathbb{P}(\theta \mid X)$. Recall that $n\hat{\sigma^2} = \dots = (n-1)s^2 + n(\bar{x}-\theta)^2$.

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &= \frac{\mathbb{P}(\theta, \sigma^2 \mid X)}{\mathbb{P}(\sigma^2 \mid \theta, X)} \\
&= \frac{(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2})(\frac{1}{\sigma^2})}{\frac{(\frac{n\hat{\sigma^2}}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}} (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{n\hat{\sigma^2}}{2\sigma^2}} \\
&\propto \frac{(\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{n\hat{\sigma^2}}{2\sigma^2}}}{(\frac{n\hat{\sigma^2}}{2})^{\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{n\hat{\sigma^2}}{2\sigma^2}}} \\
&= \left(\frac{n\hat{\sigma^2}}{2}\right)^{-\frac{n}{2}} \\
&= \left(\frac{(n-1)s^2}{2} + \frac{n(\bar{x}-\theta)^2}{2}\right)^{-\frac{n}{2}} \\
&\propto \left(\frac{1}{\frac{(n-1)s^2}{2}}\right)^{-\frac{n}{2}} \left(\frac{(n-1)s^2}{2} + \frac{n(\bar{x}-\theta)^2}{2}\right)^{-\frac{n}{2}} \\
&= \left(1 + \frac{\frac{n(\bar{x}-\theta)^2}{2}}{\frac{(n-1)s^2}{2}}\right)^{-\frac{n}{2}} \\
&= \left(1 + \frac{1}{n-1}\left(\frac{\bar{x}-\theta}{\frac{s}{\sqrt{n}}}\right)^2\right)^{-\frac{n}{2}} \\
&\propto T_{n-1}\left(\bar{x}, \frac{s}{\sqrt{n}}\right)
\end{aligned}$$

Let $X_1, \dots, X_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ and σ^2 are unknown and so $\mathbb{P}(\theta, \sigma^2) = \frac{1}{\sigma^2}$. Then

$$\begin{aligned}
\mathbb{P}\left(\theta, \sigma^2\right) &\propto \frac{1}{\sigma^2} \\
\mathbb{P}\left(\theta \mid X, \sigma^2\right) &= N(\bar{x}, (\frac{\sigma}{\sqrt{n}})^2) \\
\mathbb{P}\left(\sigma^2 \mid X, \theta\right) &= \text{InvGamma}(\frac{n}{2}, \frac{n\hat{\sigma^2}}{2}) \\
\mathbb{P}(\theta \mid X) &= T_{n-1}(\bar{x}, \frac{s}{\sqrt{n}}) \\
\mathbb{P}\left(\sigma^2 \mid X\right) &= \text{InvGamma}(\frac{n-1}{2}, \frac{(n-1)s^2}{2})
\end{aligned}$$

Use the last two for hypothesis testing and making credible regions.

What's $\mathbb{P}(X^* \mid X)$?

$$\begin{aligned}
\mathbb{P}(X^* \mid X) &= \int_0^\infty \int_{-\infty}^\infty \mathbb{P}\left(X^* \mid \theta, \sigma^2\right) \mathbb{P}\left(\theta, \sigma^2 \mid X\right) \, d\theta d\sigma^2 \\
&\propto \int_0^\infty \int_{-\infty}^\infty \left((\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x^*-\theta)^2}\right)\left((\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2}\right) d\theta d\sigma^2 \\
&= \int_0^\infty (\sigma^2)^{-(\frac{n+1}{2})-1} \, d\sigma^2 \int_{-\infty}^\infty e^{-\frac{1}{2\sigma^2}((x^*-\theta)^2+\sum(x_i-\theta)^2)} \, d\theta \\
&= \int_0^\infty (\sigma^2)^{-(\frac{n+1}{2})-1} e^{-\frac{{x^*}^2+n\bar{x}^2+(n-1)s^2}{2\sigma^2}} \, d\sigma^2 \int_{-\infty}^\infty \underbrace{e^{\frac{x^*+n\bar{x}}{\sigma^2}\theta} e^{-\frac{n+1}{2\sigma^2}\theta^2}}_{\text{kernel for normal}} \, d\theta \\
&\propto T_{n-1}(\bar{x}, \sqrt{s^2 \frac{n+1}{n}})
\end{aligned}$$

When n is large, $T_{n-1} \approx N$, $\frac{n+1}{n} \approx 1$ and so $X^*|X \approx N(\bar{x}, s^2)$.

$$\mathbb{P}(X^* \mid X) = \iint \underbrace{\mathbb{P}\left(X^* \mid \theta, \sigma^2\right)}_{N(\theta,\sigma^2)} \underbrace{\mathbb{P}\left(\theta \mid X, \sigma^2\right)}_{N(\bar{x},(\frac{\sigma}{\sqrt{n}})^2)} \quad \underbrace{\mathbb{P}\left(\sigma^2 \mid X\right)}_{\text{InvGamma}(\frac{n-1}{2},\frac{(n-1)s^2}{2})} \quad d\theta d\sigma^2$$

Sampling from $X^*|X$:

1. Sample σ_0^2 from InvGamma$(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$
2. Sample θ_0 from $N(\bar{x}, (\frac{\sigma}{\sqrt{n}})^2)$
3. Sample x^* from $N(\theta_0, \sigma_0^2)$
4. Repeat step 1 - 3 S times and return $x_1^*, \dots, x_S^*$

Let $X_1, \dots, X_n|\theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$ and $\mathbb{P}(\theta, \sigma^2) \propto \frac{1}{\sigma^2}$ Then $\mathbb{P}(\theta, \sigma^2 \mid X) =$ NormInvGamma(...). Let $\mathbb{P}(\theta) = N(\mu_0, \tau^2)$ and $\mathbb{P}(\sigma^2) =$ InvGamma$(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2})$ such that $\tau^2 \neq \frac{\sigma^2}{n_0}$. This means that $\mathbb{P}(\theta, \sigma^2) = \mathbb{P}(\theta)\mathbb{P}(\sigma^2)$ or, θ and σ^2 are independent. Then

$$\begin{aligned}
\mathbb{P}\left(\theta, \sigma^2 \mid X\right) &\propto \mathbb{P}\left(X \mid \theta, \sigma^2\right) \mathbb{P}(\theta) \mathbb{P}\left(\sigma^2\right) \\
&\propto \mathbb{P}\left(\theta \mid X, \sigma^2\right) \mathbb{P}\left(\sigma^2 \mid X\right) \\
&\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}((n-1)s^2+n(\bar{x}-\theta)^2)} e^{-\frac{1}{2\tau^2}(\theta-\mu_0)^2} (\sigma^2)^{-(\frac{n_0}{2}+1)} e^{-\frac{n_0\sigma_0^2}{2\sigma^2}} \\
&= (\sigma^2)^{-\frac{n}{2}-(\frac{n_0}{2}+1)} e^{-\frac{1}{2\sigma^2}((n-1)s^2+n_0\sigma_0^2)} e^{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2-\frac{1}{2\tau^2}(\theta-\mu_0)^2} \\
&\propto (\sigma^2)^{-\frac{n}{2}-(\frac{n_0}{2}+1)} e^{-\frac{1}{2\sigma^2}((n-1)s^2+n_0\sigma_0^2+n\bar{x}^2)} \underbrace{\exp(-(\frac{n}{2\sigma^2}+\frac{1}{2\tau^2})\theta^2+(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau^2})\theta)}_{\propto N(\theta_p,\sigma_p^2)} \\
&= (\sigma^2)^{-\frac{n}{2}-(\frac{n_0}{2}+1)} e^{-\frac{1}{2\sigma^2}((n-1)s^2+n_0\sigma_0^2+n\bar{x}^2)} \cdot \underbrace{\sqrt{2\pi\sigma_p^2}}_{\sqrt{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}} \quad \underbrace{e^{-\frac{\theta_p^2}{2\sigma_p^2}}}_{\exp(-\frac{1}{2}\frac{(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau^2})^2}{(\frac{n}{\sigma^2}+\frac{1}{\tau^2})^3})} \quad \underbrace{N(\theta_p, \sigma_p^2)}_{\frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{11}{2\sigma_p^2}(\theta-\theta_p)^2}}
\end{aligned}$$

This is not proportional to any distribution.

Sampling from the posterior $\mathbb{P}(\theta, \sigma^2 \mid X)$:

1. Sample σ_0^2 from $K(\sigma^2 \mid X)$ where
$$K(\sigma^2 \mid X) = (\sigma^2)^{-\frac{n}{2}-(\frac{n_0}{2}+1)} e^{-\frac{1}{2\sigma^2}((n-1)s^2 + n_0\sigma_0^2 + n\bar{x}^2)} \cdot \sqrt{2\pi\sigma_p^2} e^{-\frac{\theta_p^2}{2\sigma_p^2}}$$
2. Sample θ_0 from $N(\theta_p, \sigma_p^2 = \frac{1}{\frac{n}{\sigma_0^2} + \frac{1}{\tau^2}})$
3. Record $\langle \theta_0, \sigma_0^2 \rangle$
4. Repeat step 1- 3 S times

Sampling from $K(\sigma^2 \mid X)$:

1. Pick σ_{min}^2, σ_{max}^2 and $\Delta\sigma^2$
2. Create grid $\mathcal{G} = \langle \sigma_{\text{min}}^2, \sigma_{\text{min}}^2 + \Delta\sigma^2, \sigma_{\text{min}}^2 + 2\Delta\sigma^2, \dots, \sigma_{\text{max}}^2 \rangle$
3. Compute c where
$$c \approx \frac{1}{\sum_{\sigma^2 \in \mathcal{G}} K(\sigma^2 \mid X)}$$
4. Compute $F(\sigma_0^2 \mid X)$ where
$$F(\sigma_0^2 \mid X) = \sum_{\{\sigma^2 \in \mathcal{G} : \sigma^2 < \sigma_0^2\}} c \cdot K(\sigma^2 \mid X)$$
5. Draw y from $U(0,1)$
6. Compute $\sigma_0^2 = \min_{\sigma^2 < \mathcal{G}} F(\sigma^2) \geq y$

Grid Sampling Disadvantages:

- Numerically assemble - computers have minimum and maximum values of numbers
- How to pick θ_{min}, θ_{max} and $\Delta\theta$? A bad decision for θ_{min} and θ_{max} will lead to missing a part of the support of the parameter A bad decision for $\Delta\theta$ means bad boundaries and so non-realistic samples.
- Let's say $\theta_{\text{min}} = 0$, $\theta_{\text{max}} = 1$, $\Delta\theta = 0.0001$ and $|\mathcal{G}| = 10,000 = 10^5$. What if θ had 10 dimensions? Then $|\mathcal{G}| = 10^{5^{10}} = 10^{50}$ which is impossible for a computer.

Therefore, grid sampling is only good in low dimensions where you know the effective support of θ(where most of the support lies) and if you know the shape so you can pick a reasonable $\Delta\theta$.

Let $X_1, \dots, X_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, $\theta \sim N(\mu_0, \tau^2)$ and $\sigma^2 \sim \text{InvGamma}(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2})$. Then $\mathbb{P}(\theta, \sigma^2 \mid X) = N(\theta_p, \sigma_p^2)K(\sigma^2 \mid X)$.

Let $X|\theta \sim \text{Binom}(n, \theta)$ and $\theta | X \sim \text{Beta}(\alpha + x, \beta + n - x)$. What if you want to use a irregular distribution for θ that has wacky ups and downs that cannot be represented using a Beta distribution? If you know the function $\mathbb{P}(\theta)$, then you can compute $\mathbb{P}(\theta \mid X) \propto \mathbb{P}(X \mid \theta)\mathbb{P}(\theta) = K(\theta \mid X)$ and use a grid search $\mathcal{G} = \langle \theta_{\min}, \theta_{\min} + \Delta\theta, \theta_{\min} + 2\Delta\theta, \dots, \theta_{\max} \rangle$. Can we still use conjugacy? Imagine $\mathbb{P}(\theta)$ is a mixture/compound distribution of a discrete number of beta compounds: $\mathbb{P}(\theta) = \sum_{m=1}^M \gamma_m \underbrace{\mathbb{P}_m(\theta)}_{\text{Beta}(\alpha,\beta)}$ where $\sum \gamma_m = 1$. ex: $\mathbb{P}(\theta) = \frac{1}{2}\text{Beta}(3,3) + \frac{1}{2}\text{Beta}(2,7)$.

Let $X|\theta \sim \text{Binom}(n, \theta)$. Let $\mathbb{P}(\theta) = \sum_{m=1}^M \gamma_m \mathbb{P}_m(\theta)$. Then

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &= \frac{\mathbb{P}(X \mid \theta)\mathbb{P}(\theta)}{\mathbb{P}(X)} \\
&= \frac{\mathbb{P}(X \mid \theta)\sum \gamma_m \mathbb{P}_m(\theta)}{\mathbb{P}(X)} \\
&= \sum_{m=1}^M \gamma_m \frac{\mathbb{P}(X \mid \theta)\mathbb{P}_m(\theta)}{\mathbb{P}(X)} \\
&= \sum_{m=1}^M \gamma_m \underbrace{\frac{\mathbb{P}(X \mid \theta)\mathbb{P}_m(\theta)}{\mathbb{P}(X)}}_{\gamma_m'} \cdot \underbrace{\frac{\mathbb{P}(X \mid \theta)\mathbb{P}_m(\theta)}{\mathbb{P}_m(X)}}_{\mathbb{P}_m(\theta \mid X)} \\
&= \sum_{m=1}^M \gamma_m' \underbrace{\mathbb{P}_m(\theta \mid X)}_{\text{Beta}(\alpha+x, \beta+n-x)}
\end{aligned}$$

What's $\mathbb{P}(X)$?

$$\begin{aligned}
\mathbb{P}(X) &= \int_\Theta \mathbb{P}(X \mid \theta)\mathbb{P}(\theta)\, d\theta \\
&= \int_\Theta \mathbb{P}(X \mid \theta) \sum \gamma_m \mathbb{P}_m(\theta)\, d\theta \\
&= \sum_{m=1}^m \gamma_m \underbrace{\int_\Theta \mathbb{P}(X \mid \theta)\mathbb{P}_m(\theta)\, d\theta}_{\text{BetaBinom}(n, \alpha_m, \beta_m)}
\end{aligned}$$

If $\gamma_m = \frac{1}{M}$ for all m,

$$\gamma_m' = \frac{\gamma_m \mathbb{P}_m(X)}{\mathbb{P}(X)} = \frac{\gamma_m \mathbb{P}_m(X)}{\sum \gamma_m \mathbb{P}_m(X)} = \frac{\mathbb{P}_m(X)}{\sum \mathbb{P}_m(X)}$$

Let $X|\theta \sim \text{Binom}(n,\theta)$, and $\mathbb{P}(\theta) = \sum_{m=1}^M \gamma_m \mathbb{P}_m(\theta)$. What $\theta|X$? Let $\gamma_1 = \gamma_2 = \frac{1}{2}$, $\alpha_1 = 3$, $\beta_1 = 3$, $\alpha_2 = 2$, $\beta = 4$, $n = 10$ and $x = 5$.

$$\begin{aligned}
\mathbb{P}(\theta \mid X = 5) &= \sum_{m=1}^M \gamma_m \mathbb{P}_m(\theta \mid X) \\
&= \frac{1}{\mathbb{P}_1(5) + \mathbb{P}_2(5)} (\mathbb{P}_1(5)\mathbb{P}_1(\theta \mid X = 5) + \mathbb{P}_2(5)\mathbb{P}_2(\theta \mid X = 5)) \\
&= \frac{1}{\text{dbetabinom}(5,10,3,3) + \text{dbb}(5,10,2,4)} \\
&\cdot \Big(\text{dbb}(5,10,3,3) \cdot \text{dbeta}(\theta,8,8) + \text{dbb}(5,10,2,4) \cdot \text{dbeta}(\theta,7,9)\Big) \\
&= 0.57\text{dbeta}(\ ) + 0.43\text{dbeta}()
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{P}(X) &= \text{BetaBinom}(n,\alpha_m,\beta_m) \\
\mathbb{P}_1(5) &= \text{dbetabinom}(5,10,3,3) = 0.147 \\
\mathbb{P}_2(5) &= \text{dbetabinom}(5,10,2,4) = 0.112
\end{aligned}$$

The first one should be higher since $\alpha = 3$ and $\beta = 3$ is centered at 5 and so it splits off evenly.

Sample from $\mathbb{P}(\theta \mid X)$:

1. Sample $\theta_{0,1}$ from Beta(8, 8) using rbeta(8,8) which pulls a sample from Beta
2. Sample $\theta_{0,2}$ from Beta (7, 9) using rbeta(7,9)
3. Retain $\theta_0 = \gamma_1'\theta_{0,1} + \gamma_2'\theta_{0,2}$
4. Repeat Steps 1-3 many times

Point Estimation:

$$\begin{aligned}
\hat{\theta}_{\text{MMSE}} &= \text{E}[\theta \mid X] \\
&= \int_\Theta \theta \sum \gamma_m' \mathbb{P}_m(\theta \mid X)\, d\theta \\
&= \sum \gamma_m' \int_\Theta \theta \mathbb{P}_m(\theta \mid X)\, d\theta \\
&= \sum \gamma_m' \text{E}_m(\theta \mid X) \\
&= \sum_{m=1}^M \gamma_m' \frac{\alpha_m'}{\alpha_m' + \beta_m'}
\end{aligned}$$

In the above example,

$$\hat{\theta}_{\text{MMSE}} = 0.57(\frac{8}{16}) + 0.43(\frac{7}{16})$$

$$\begin{aligned}
\hat{\theta}_{\text{MAE}} &= \ldots \text{Sample median} \\
\hat{\theta}_{\text{MAP}} &= \text{argmax}\{\mathbb{P}(\theta \mid X)\} = \text{argmax}\{K(\theta \mid X)\}
\end{aligned}$$

Find $\hat{\theta}_{\text{MLE}}$.

$$\begin{aligned}\mathbb{P}\left(\theta \mid X\right) &= \sum \gamma_m \mathbb{P}_m(X)\mathbb{P}_m(\theta \mid X) = K(\theta \mid X) \\ &= \sum \gamma_m \Big(\binom{n}{x} \frac{B(x|\alpha_m, n-x+\beta_m)}{B(\alpha_m,\beta_m)}\Big)\Big(\frac{1}{B(x+\alpha_, n-x+\beta_m)}\theta^{x+\alpha_m-1}(1-\theta)^{n-x+\beta_m-1}\Big)\end{aligned}$$

$$\frac{d}{d\theta}\mathbb{P}\left(\theta \mid X\right) = 0$$

Doesn't matter, cannot be solved.

Assume $f(x)$ is continuous and differentiable and has one zero on X. We want x^* such that $f(x^*) = 0$.
Newton's Method

1. Guess $x_0 = x^*$
2. Draw tangent line
3. Set $x_1 = x$-intercept of the tangent line
4. Repeat until $|x_{t+1} - x_t| < \epsilon$ by setting $x_0 = x_t$ and letting ϵ be your accuracy/tolerance level

In Step 2, $y - b = m(x-a) \to y - f(x_0) = f'(x_0)(x - x_0)$
In Step 3, Solve for x-intercept (x_1): $-f(x_0) = f'(x_0)(x_1 - x_0)$ and so $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ and thus $x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$.

Gibb Sampling: if prior is a known mixture, what if likelihood model is a mixture?
$X_1,\dots,X_n|\theta \overset{iid}{\sim} \sum_{m=1}^M \gamma_m \mathbb{P}_m(X \mid \theta)$.
Goal: Get the posterior or function of posterior

$$\mathbb{P}\left(\theta_1,\sigma_1^2,\theta_2,\sigma_2^2,\rho \mid X\right) \propto \Big(\prod_{i=1}^n \mathbb{P}\left(X_i \mid \theta_1,\sigma_1^2,\theta_2,\sigma_2^2,\rho\right)\Big)\mathbb{P}\left(\theta_1,\sigma_1^2,\theta_2,\sigma_2^2,\rho\right)$$

Consider the mixture model
$X_1,\dots,X_n|\vec{\theta}_1,\dots,\vec{\theta}_n,\gamma_1,\dots,\gamma_M \overset{iid}{\sim} \sum_{m=1}^M \gamma_m \mathbb{P}_m(\vec{\theta}_m)$ such that $\gamma_1 + \gamma_2 + \dots + \gamma_M = 1$.
For example, $X_1,\dots,X_n|\theta_1,\sigma_1^2,\theta_2,\sigma_2^2 \overset{iid}{\sim} \underbrace{\rho}_{\gamma_1} N(\theta_1,\sigma_1^2) + \underbrace{(1-\rho)}_{\gamma_2} N(\theta_2,\sigma_2^2)$

Then

$$\begin{aligned}\mathbb{P}\left(\theta_1,\sigma_1^1,\theta_2,\sigma_2^2,\rho \mid X\right) &\propto \mathbb{P}\left(X \mid \theta_1,\sigma_1^2,\theta_2,\sigma_2^2,\rho\right) \underbrace{\underbrace{\mathbb{P}\left(\theta_1,\sigma_1^2,\theta_2,\sigma_2^2,\rho\right)}_{\mathbb{P}\left(\theta_1\right)\mathbb{P}\left(\sigma_1^2\right)\mathbb{P}\left(\theta_2\right)\mathbb{P}\left(\sigma_2^2\right)\mathbb{P}\left(\rho\right)}}_{1\cdot\frac{1}{\sigma_1^2}\cdot 1\cdot\frac{1}{\sigma_2^2}\cdot 1} \\ &= \Big(\prod_{i=1}^n \rho\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{1}{2\sigma_1^2}(x_i-\theta_1)^2} + (1-\rho)\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{1}{2\sigma_2^2}(x_i-\theta_2)^2}\Big)\cdot\frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2} \\ &= K(\theta_1,\sigma_1^2,\theta_2,\sigma_2^2,\rho \mid X)\end{aligned}$$

How to get inference?
Grid search: $\mathcal{G}_{\theta_1} = \langle \theta_{1,\text{min}}, \theta_{1,\text{min}} + \Delta\theta_1, \dots, \theta_{1,\text{max}} \rangle$ and similarly for other parameters. This is inaccurate and too large.
What if we know which components each x_i belonged to?
Let $I = \{I_1, I_2, \dots, I_n\}$. Define

$$\begin{aligned} I_1 &:= I_{x_1} \text{ is in } m = 1 \\ I_2 &:= I_{x_2} \text{ is in } m = 2 \\ &\vdots \\ I_n &:= I_{x_n} \text{ is in } m = n \end{aligned}$$

These are called "latent variables/information" because the I_i's are unobserved but still important (can't seem them).
Recall that $f(z) = \int f(z,y)\,dy = \int f(z \mid y) f(y)\,dy$.Then

$$\mathbb{P}(X \mid \theta) = \int \mathbb{P}(X, I \mid \theta)\ dI = \int \mathbb{P}(X \mid I, \theta)\, \mathbb{P}(I \mid \theta)\ dI$$

This is called Data Augmentation. It is augmenting X with the I_i's, or adding more data to the data. Thus

$$\begin{aligned} \mathbb{P}\left(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \mid X\right) &\propto \int \mathbb{P}\left(X \mid I, \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho\right) \mathbb{P}\left(I \mid \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho\right) \mathbb{P}\left(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho\right)\ dI \\ &= K(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \mid X) \\ &= \int K(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \mid X, I)\, dI \end{aligned}$$

Model Goal: Get $\hat{\theta}_{\text{MAP}} = \text{argmax}\{K(\theta \mid X)\}$, the most likely value of the 5 parameters.

Expectation-Maximization Algorithm

1. Guess $\hat{\theta}_{\text{MAP}} = \theta_0$ to start
2. Compute $I_0 = \text{E}[I_0 \mid X, \theta = \theta_0]$ (expectation step)
3. Consider $\mathcal{L}(\theta; I_0, X) = K(\theta \mid X, I = I_0)\, dI$ and find $\hat{\theta}_1 = \text{argmax}\{\mathcal{L}(\theta; I, X)\}$ (maximization step)
4. Repeat steps 2-3 until $||\theta_{t+1} - \theta_t|| < \epsilon$ where ϵ is the predefined tolerance level

E-M Implementation for our Two-Normal Mixture:

1. Initialize

$$\begin{aligned} \theta_{1,0} &= 0 \\ \sigma_{1,0}^2 &= 1 \\ \theta_{2,0} &= 0 \\ \sigma_{2,0}^2 &= 1 \\ \rho &= 0.5 \end{aligned}$$

2.
$$\begin{aligned} I_{1,0} &= \mathrm{E}[I_1 \mid X, \theta_1 = \theta_{1,0}, \sigma_1^2 = \sigma_{1,0}^2, \theta_2 = \theta_{2,0}, \sigma_2^2 = \sigma_{2,0}^2, \rho = \rho_0] \\ &= \mathbb{P}\left(I_1 = 1 \mid X, \dots\right) \\ &= \frac{\mathbb{P}\left(X \mid I_1 = 1, \dots\right)\mathbb{P}\left(I_1 = 1 \mid \dots\right)}{\underbrace{\mathbb{P}\left(X \mid \dots\right)}_{\mathbb{P}(X \mid I_1 = 1, \dots) + \mathbb{P}(X \mid I_1 = 0, \dots)}} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma_{1,0}}} e^{-\frac{1}{2\sigma_{1,0}^2}(x_i - \theta_{1,0})^2} \cdot \rho}{\rho \frac{1}{\sqrt{2\pi\sigma_{1,0}^2}} e^{-\frac{1}{2\sigma_{1,0}^2}(x_i - \theta_{1,0})^2} + (1-\rho)\frac{1}{\sqrt{2\pi\sigma_{2,0}^2}} e^{-\frac{1}{2\sigma_{2,0}^2}(x_i - \theta_{2,0})^2}} \end{aligned}$$

Then
$$\begin{aligned} I_{2,0} &= \mathrm{E}[I_2 \mid X_2, \dots] \\ I_{3,0} &= \mathrm{E}[I_2 \mid X_3, \dots] \\ &\vdots \\ I_{n,0} &= \mathrm{E}[I_n \mid X_n, \dots] \end{aligned}$$

3. Consider

$$\begin{aligned} \mathcal{L}(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho; I, X) &= \mathbb{P}\left(X \mid I, \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho\right)\mathbb{P}\left(I \mid \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho\right)\mathbb{P}\left(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho\right) \\ &= \Big(\prod_{i=1}^n \Big(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i-\theta_1)^2}\Big)^{I_i} \cdot \Big(\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(x_i-\theta_2)^2}\Big)^{1-I_i}\Big) \cdot \Big(\prod_{i=1}^n \rho^{I_i}(1-\rho)^{1-I_i}\Big) \cdot \left((\sigma_1^2)^{-1}(\sigma_2^2)^{-1}\right) \\ &= \Big(\frac{1}{\sqrt{2\pi}}\Big)^n (\sigma_1^2)^{-1}(\sigma_2^2)^{-1}(\sigma_1^2)^{-\frac{1}{2}\sum I_i} e^{-\frac{1}{2\sigma_1^2}\sum I_i(x_i-\theta_1)^2 - \frac{1}{2\sigma_2^2}\sum(1-I_i)(x_i-\theta_2)^2} \cdot \rho^{\sum x_i}(1-\rho)^{\sum(1-I_i)} \\ &\text{By taking log,} \\ &= l(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho; I, x) \\ &= n\ln\Big(\frac{1}{\sqrt{2\pi}}\Big) - (1 + \frac{1}{2}\sum I_i)\ln(\sigma_1^2) - (1 + \frac{1}{2}\sum(1-I_i))\ln(\sigma_2^2) - \frac{1}{2\sigma_1^2}\sum I_i(x_i-\theta_1)^2 - \frac{1}{2\sigma_2^2}\sum(1-I_i)(x_i-\theta_2)^2 \end{aligned}$$

Take derivatives.

- Get $\hat{\theta}_1$ by $\frac{\partial}{\partial\theta_1}$[log likelihood] $= 0$

$$\frac{\sum x_i I_i}{\sigma_1^2} - \frac{2\theta_1 \sum I_i}{2\sigma_1^2} = 0$$
$$\hat{\theta}_1 = \frac{\sum x_i I_i}{\sum I_i} \text{ like } \bar{x}_{\text{mixture 1}}$$

- Get $\hat{\theta}_2$ by $\frac{\partial}{\partial\theta_2}$[log likelihood] $= 0$

$$\hat{\theta}_2 = \frac{\sum x_i(1-I_i)}{\sum(1-I_i)} \text{ like } \bar{x}_{\text{mixture 2}}$$

- Get $\hat{\sigma^2}_1$ by $\frac{\partial}{\partial\sigma_1^2}$[log likelihood] $= 0$

$$\begin{aligned} -\frac{1+\frac{1}{2}\sum I_i}{\sigma_1^2} + \frac{1}{2(\sigma_1)^2}\sum I_i(x_i-\theta_1)^2 = 0 \\ 1+\frac{1}{2}\sum I_i &= \frac{1}{2\sigma_1^2}\sum I_i(x_i-\theta_1)^2 \\ \hat{\sigma_1}^2 &= \frac{\sum I_i(x_i-\theta_1)^2}{2+\sum I_i} \\ &\text{similar to sample variance when } m=1 \end{aligned}$$

- Get $\hat{\sigma^2}_2$ by $\frac{\partial}{\partial\sigma_2^2}$[log likelihood] $= 0$

$$\hat{\sigma^2}_2 = \frac{\sum(1-I_i)(x_i-\theta_2)^2}{2+\sum(1-I_i)} \text{ similar to sample variance when } m=2$$

- Get $\hat{\rho}$ by $\frac{\partial}{\partial\rho}$[log likelihood] $= 0$

$$\begin{aligned} \frac{\sum I_i}{\rho} - \frac{1-I_i}{1-\rho} &= 0 \\ \sum I_i - \rho\sum I_i &= \rho n - \rho\sum I_i \\ \hat{\rho} &= \frac{\sum I_i}{n} \end{aligned}$$

4. Iterate through the previous two steps until better versions of I's are found and there's convergence

Recall $X_1,\dots,X_n|\theta,\sigma^1 \overset{iid}{\sim} N(\theta,\sigma^2)$, $\theta\sim N(\mu_0,\tau^2)$ and $\sigma^2\sim \text{InvGamma}(\frac{n_0}{2},\frac{n_0\sigma_0^2}{2})$. Therefore $\mathbb{P}(\theta,\sigma^2\mid X)\propto K(\theta,\sigma^2\mid X)$ which is non-conjugate.
But

$$\begin{aligned} \mathbb{P}\left(\theta\mid X,\sigma^2\right) &= N(\theta_p,\sigma_p^0) \\ \mathbb{P}\left(\sigma^2\mid X,\theta\right) &= \text{InvGamma}(\frac{n_0+n}{2},\frac{n_0\sigma_0^2+n\hat{\sigma}^2}{2}) \end{aligned}$$

Can you use $\mathbb{P}(\theta\mid X,\sigma^2)$ and $\mathbb{P}(\sigma^2\mid X,\theta)$ to solve for $\mathbb{P}(\theta,\sigma^2\mid X)$?

$$\mathbb{P}\left(\theta,\sigma^2\mid X\right) = \mathbb{P}\left(\theta\mid\sigma^2\right)\mathbb{P}\left(\sigma^2\mid X\right) = \mathbb{P}\left(\sigma^2\mid\theta,X\right)\mathbb{P}\left(\theta\mid X\right)$$

Not possible without either $\mathbb{P}(\theta\mid X)$ or $\mathbb{P}(\sigma^2\mid X)$.
What if you use an iterative algorithm?

1. Draw an arbitrary value of θ_0
2. Draw σ_0^2 from $\mathbb{P}(\sigma^2\mid X,\theta=\theta_0)$
3. Draw θ_1 from $\mathbb{P}(\theta\mid X,\sigma^2=\sigma_0^2)$

4. Draw σ_1^2 from $\mathbb{P}\left(\sigma^2 \mid X, \theta = \theta_1\right)$

5. Repeat steps 3-4 until there is convergence

This algorithm is called Gibbs sampling or Gibbs sampler. This is different from the N-R and E-M algorithms because for NR, you solve for $f(x) = 0$ which gives one value and for E-M, you solve for $\hat{\theta}_{\text{MAP}}$ which is also one value (or vector). The iteration will then look like:

$$\langle \begin{pmatrix} \theta_0 \\ \sigma_0^2 \end{pmatrix}, \begin{pmatrix} \theta_1 \\ \sigma_1^2 \end{pmatrix}, \begin{pmatrix} \theta_2 \\ \sigma_2^2 \end{pmatrix}, \dots, \begin{pmatrix} \theta_t \\ \sigma_t^2 \end{pmatrix}, \dots \rangle$$

where t is the iteration number. This is called the Gibbs chain. Where does the algorithm converge? It converges at the burn in point, $t = B$ where you start to get nearly constant values for θ and σ^2.

Disadvantages of Gibbs Sampling:

- Bad mixture: lacks ability to traverse Supp$[\hat{\theta}]$ well.

- $\hat{\theta}$ may be a part of a set of distributions with multiple modes. The sampler will get stuck in any of the modes and then not discover the other ones. Solution: Merge all chains that start from all different starting points. This is problematic though with big dimensions of θ. Therefore you are unsure if it's solved adequately.

- Is θ_1 related to θ_0? Yes. Is θ_{1000} related to θ_{999}? Yes. After the burn in point, they're all related to each other. Thus θ_{1000} and θ_{999} are not "independent samples." In fact, $\text{Corr}[\theta_{1000}, \theta_{999}] \neq 0$.

$$\text{Corr}[X,Y] = \frac{\text{Cor}[X,Y]}{SE[X]SE[Y]} = \frac{\text{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

  By

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

  we can have autocorrelation.

Autocorrelation for lag 1 estimates $\text{Corr}[\theta_t, \theta_{t+1}]$:

$$r_{a1} = \frac{\sum_{t=B}^{B+S-1} (\theta_t - \bar{\theta})(\theta_{t+1} - \bar{\theta})^2}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2}$$

such that $\bar{\theta} = \frac{1}{S}\sum_{t=B}^{B+S} \theta_t$.
Autocorrelation for lag 2:

$$r_{a2} = \frac{\sum_{t=B}^{B+S-2} (\theta_t - \bar{\theta})(\theta_{t+2} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2}$$

Thus autocorrelation for lag k:

$$r_{ak} = \frac{\sum_{t=B}^{B+S-k}(\theta_t - \bar{\theta})(\theta_{t+k} - \bar{\theta})}{\sum_{t=B}^{B+S}(\theta_t - \bar{\theta})^2}$$

At some k¡ $r_{ak} \approx 0$ because eventually the dependency is gone. This is seen in an autocorrelation plot for k vs r_k At some value $k = t$, r_k levels off to zero. Around t, the draws are independent. In order to make the chain represent all independent samples from the posterior, we need to throw out all samples except those that are multiples of t after B. This is known as "thinning."

$$\left\{ \begin{pmatrix} \theta_B \\ \sigma_B^2 \end{pmatrix}, \begin{pmatrix} \theta_{B+t} \\ \sigma_{B+t}^2 \end{pmatrix}, \begin{pmatrix} \theta_{B+2t} \\ \sigma_{B+2t}^2 \end{pmatrix}, \dots \right\}$$

This is called the burned out thinned chain.
Let $l = 1, \dots, L$ be the index on the burned out thinned chain. This is almost as good as having $\mathbb{P}(\theta \mid X)$ directly. Then

$$\begin{aligned} \hat{\theta}_{\text{MMSE}} &= \text{E}[\theta \mid X] \approx \bar{\theta} = \frac{1}{L}\sum_{l=1}^{L} \theta_L \\ \hat{\theta}_{\text{MAE}} &= \text{Mode}[\theta \mid X] = \text{order all } \theta\text{'s from smallest to largest and then pick } \theta_{L/2} \\ CR_{\theta,1-\alpha} &= [\theta_{\frac{\alpha}{2}L}, \theta_{(1-\frac{\alpha}{2})L}] \end{aligned}$$

What is $\mathbb{P}(X^* \mid X)$?

$$\mathbb{P}(X^* \mid X) = \int_{\Theta} \mathbb{P}(X^* \mid \theta)\, \mathbb{P}(\theta \mid X)\; d\theta$$

To Sample from this:

1. Pick $l \in \{1, \dots, L\}$
2. Draw x^* from $\mathbb{P}(X^* \mid \theta = \theta_l)$
3. Repeat steps 1-2 over and over

Algorithm: Systematic Sweep/ Gibbs Sampler for $\mathbb{P}(\theta_1, \dots, \theta_p \mid X)$, the unknown posterior with p parameters
Here all conditions, $\mathbb{P}(\theta_j \mid \theta_{ij})$, where $\theta_{-j} = \{\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_p\}$ are known and can be "easily" sampled from.

1. Initialize $\theta = \hat{\theta} = \langle \theta_{0,1}, \theta_{0,2}, \dots, \theta_{0,p} \rangle$
2. Sample $\theta_{1,1}$ from $\mathbb{P}(\theta_1 \mid \theta_2 = \theta_{0,2}, \dots, \theta_p = \theta_{0,p})$.
   Sample $\theta_{1,2}$ from $\mathbb{P}(\theta_2 \mid \theta_1 = \theta_{1,1}, \theta_3 = \theta_{0,3}, \dots, \theta_p = \theta_{0,p})$.

   $$\vdots$$

   Sample $\theta_{1,p}$ from $\mathbb{P}(\theta_p \mid \theta_1 = \theta_{1,1}, \dots, \theta_{p-1} = \theta_{1,p-1})$
3. Repeat step 2 until "convergence"

Proof. Consider $X_0, X_1, X_2, \dots,$ a sample of random variables. Each has a Sample X. If $\mathbb{P}\left(\theta_t \in A \mid X_{t-1}, X_{t-2}, \dots, X_0\right) = \mathbb{P}\left(X_t \in A \mid X_{t-1}\right) \forall t, \forall A \in X$ then the sample sequence is called a "discrete-time Markov chain." The Gibbs sampler is a Markov chain. This is why the Gibbs sampler is a form of "Markov Chain Monte Carlo" or MCMC.

$$\mathbb{P}\left(X_{t+1}\right) = \int_X \mathbb{P}\left(X_{t+1}, X_t\right)\, dx = \int_X \mathbb{P}\left(X_{t+1} \mid X_t\right) \mathbb{P}\left(X_t\right)\, dt$$

If $\mathbb{P}\left(X_{t+1}\right) = \mathbb{P}\left(X_t\right)$, then this distribution is deemed the invariant, equilibrium, stationary or long term. Let

$$\mathbb{P}\left(X_{t+1}\right) = \mathbb{P}\left(X_t \mid X_{t-1}\right) \mathbb{P}\left(X_{t-1} \mid X_{t-2}\right) \dots \mathbb{P}\left(X_1 \mid X_0\right) \mathbb{P}\left(X_0\right)$$

Then you can get an invariant distribution by

$$\begin{aligned}\mathbb{P}(X) &= \lim_{t \to \infty} \int_X \mathbb{P}\left(X_t \mid X_{t-1}\right) \mathbb{P}\left(X_{t-1} \mid X_{t-2}\right) \dots \mathbb{P}\left(X_1 \mid X_0\right) \mathbb{P}\left(X_0\right)\, dx_0 \\ &= \mathbb{P}\left(\theta_{t+1,1} \mid \theta_{t-2}, \dots, \theta_{t,p}\right) \cdot \mathbb{P}\left(\theta_{t+1,2} \mid \theta_{t+1,1}, \theta_{t,3}, \dots, \theta_{t,p}\right) \cdot \mathbb{P}\left(\theta_{t+1,p-1} \mid \theta_{t+1,1}, \dots, \theta_{t+1,p-1}, \theta_{t,p}\right) \cdot \mathbb{P}\left(\theta_{t+\right.\end{aligned}$$

In vector notation,

$$\mathbb{P}\left(\hat{\theta}_{t+1}\right) = \int \mathbb{P}\left(\hat{\theta}_{t+1} \mid \hat{\theta}_t\right) \cdot \mathbb{P}\left(\hat{\theta}\right)_t d\hat{\theta}$$

In scalar notation,

$$\mathbb{P}\left(\theta_{t+1,1}, \dots, \theta_{t+1,p}\right) =$$

Fill in at a later time.. □

Change Point Model:
Parameters:

- λ_1 - mean of "first process"
- λ_2 - mean of "second process"
- m - "change point"

Priors:

$$\begin{aligned}\mathbb{P}\left(\lambda_1\right) &= \text{Gamma}(\alpha, \beta) \\ \mathbb{P}\left(\lambda_2\right) &= \text{Gamma}(\alpha, \beta) \\ \mathbb{P}(m) &= \text{Uniform}\{0, \dots, n\} = \frac{1}{n} \forall m\end{aligned}$$

Posterior:

$$\begin{aligned}\mathbb{P}\left(\lambda_1, \lambda_2, m \mid X_1, \dots, X_n\right) &\propto \mathbb{P}\left(X_1, \dots X_n \mid \lambda_1, \lambda_2, m\right) \cdot \underbrace{\mathbb{P}\left(\lambda_1, \lambda_2, m\right)}_{\mathbb{P}(\lambda_1)\mathbb{P}(\lambda_2)\mathbb{P}(m)} \\ &\propto \Big(\prod_{i=1}^m \frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!}\Big)\Big(\prod_{i=m+1}^n \frac{e^{-\lambda_2}\lambda_2^{x_i}}{x_i!}\Big)\left(\lambda_1^{\alpha-1}e^{-\beta\lambda_1}\right)\left(\lambda_2^{\alpha-1}e^{-\beta\lambda_2}\right) \\ &\propto e^{-m\lambda_1}\lambda^{\sum_{i=1}^m x_i} e^{-(n-m+1)\lambda_2}\lambda_2^{\sum_{i=m+1}^n x_i}\lambda_1^{\alpha-1}e^{-\beta\lambda_1}\lambda_2^{\alpha-1}e^{-\beta\lambda_2} \\ &= e^{-(m+\beta)\lambda_1}\lambda_1^{(\sum_{i=1}^m x_i)+\alpha-1} e^{-(n-m+1)\lambda_2}\lambda_2^{(\sum_{i=m+1}^n x_i)+\alpha-1}\end{aligned}$$

This is an unknown distribution and the best we can do. We need the following conditionals:

$$\mathbb{P}(\lambda_1 \mid X_1,\dots,X_n,\lambda_2,m) \propto e^{-(m+\beta)\lambda_1}\lambda_1^{(\sum_{i=1}^m x_i)+\alpha-1} \propto \text{Gamma}(\alpha+\sum_{i=1}^m x_i, \beta+m)$$

$$\mathbb{P}(\lambda_2 \mid X_1,\dots,X_n,\lambda_1,m) = e^{-(n-m+\beta)\lambda_2}\lambda_2^{(\sum_{i=m+1}^n x_i)+\alpha-1} \propto \text{Gamma}(\alpha+\sum_{i=m+1}^n x_i, \beta+n-m)$$

$$\begin{aligned}\mathbb{P}(m \mid X_1,\dots,X_n,\lambda_1,\lambda_2) &\propto \underbrace{e^{-m(\lambda_1-\lambda_2)}\lambda_1^{\sum_{i=1}^m x_i}\lambda_2^{\sum_{i=m+1}^n x_i}}_{h(m)} \\ &\propto \frac{h(m)}{\sum_{k=0}^m h(k)}\end{aligned}$$

After this, pick λ_1 and a starting point. Plug in to get the next round and keep repeating.

$$\left\langle \begin{pmatrix}\lambda_{0,1}\\ \lambda_{0,2}\\ m_0\end{pmatrix}, \begin{pmatrix}\lambda_{1,1}\\ \lambda_{1,2}\\ m_1\end{pmatrix}, \dots \right\rangle$$

Drawing a vertical line through the three graphs constitutes 1 data point. All have the same burn-in point and converges quickly. Discard the data points before the burn-in point. These data points dip below the significance level.

Recall the Bayeisan Protocol:

1. Pick $\mathcal{F}$, the likelihood model
2. Pick $\mathbb{P}(\theta)$, the prior
3. Collect data x
4. Obtain posterior $\mathbb{P}(\theta \mid X)$ for inference
    - do it directly in closed form
    - if only $k(\theta \mid X)$, use grid sampling if you think it'll be accurate
    - Gibbs sampling

What if 1 and 2 went wrong (the model is wrong)? How do you access the degree of departure from reality? Model Checking.
First Check (easy to pass): Recall $\mathbb{P}(X) = \int_\Theta \mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta)\, d\theta$, the prior predictive distribution. It shows you what data looks like coming from the model $\mathcal{F}$ subject to the parameters from your prior idea.
For example, if $\mathbb{P}(X \mid \theta) = \text{Binom}(100,\theta)$ and $\mathbb{P}(\theta) = U(0,1) = \text{Beta}(1,1)$, then $\mathbb{P}(X) = \text{BetaBinom}(100,1,1)$.
How to Check?

1. Sample many points from $\mathbb{P}(X)$

2. Plot the data x

3. Does the data x look plausible coming from $\mathbb{P}(X)$?

Second Check (harder to checK): Recall $\mathbb{P}(X^* \mid X) = \int_\Theta \mathbb{P}(X^* \mid \theta)\mathbb{P}(\theta \mid X)\, d\theta$, the posterior predictive distribution or the posterior replicative distribution where X^* is "replicated" data that could be observed tomorrow. In the above case, $\mathbb{P}(X^* \mid X) = \text{BetaBiinom}(100, 30, 62)$. How to Check:

1. Sample many points from $\mathbb{P}(X^* \mid X)$

2. Plot data x

3. Does the data look like other replicates of the data?

Gibbs Sampler: We want to sample from $\mathbb{P}(\theta_1, \dots, \theta_p \mid X)$, which is not easily sampled from directly. You have $\forall j, \mathbb{P}(\theta_j \mid \theta_{-j}, X)$, all the conditionals distributions that are easy to sample from.

Suppose $X_1, \dots, X_n, \mid \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho \overset{iid}{\sim} \rho N(\theta_1, \sigma_1^2) + (1-\rho)N(\theta_2, \sigma_2^2)$. Assume the following priors:

$$\begin{aligned}
\mathbb{P}(\theta_1) &\propto 1 \\
\mathbb{P}(\theta_2) &\propto 1 \\
\mathbb{P}\left(\sigma_1^2\right) &\propto \frac{1}{\sigma_1^2} \\
\mathbb{P}\left(\sigma_2^2\right) &\propto \frac{1}{\sigma_2^2} \\
\mathbb{P}(\rho) &\propto U(0,1) \propto 1
\end{aligned}$$

Use data augmentation to get $\mathbb{P}(I_1, \dots, I_n, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho \mid X)$.

$$\begin{aligned}
&\mathbb{P}\left(I_1, \dots, I_n, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho \mid X\right) \propto \mathbb{P}\left(X_1, \dots, X_n \mid I_1, \dots, I_n, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho\right) \\
&\cdot \mathbb{P}\left(I_1, \dots, I_n, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho\right) \\
&\propto \mathbb{P}\left(X_1, \dots, X_n \mid I_1, \dots, I_n, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho\right) \cdot \mathbb{P}\left(I_1, \dots, I_n \mid \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho\right) \cdot \mathbb{P}\left(\theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho\right) \\
&\propto \mathbb{P}\left(X_1, \dots, X_n \mid I_1, \dots, I_n, \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, \rho\right) \cdot \prod_{i=1}^n \rho^{I_i}(1-\rho)^{1-I_i} \cdot \frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2} \\
&\propto \left(\prod_{i=1}^n \Big(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i-\theta_1)^2}\Big)^{I_i} \Big(\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(X_i-\theta_2)^2}\Big)^{n-\sum I_i} \rho^{I_i}(1-\rho)^{1-I_i}\right)\frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2} \\
&\propto (\rho\frac{1}{\sqrt{2\pi\sigma_1^2}})^{\sum I_i} e^{-\frac{1}{2\sigma_1^2}\sum I_i(X_i-\theta_1)^2} \Big((1-\rho)\frac{1}{\sqrt{2\pi\sigma_2^2}}\Big)^{n-\sum I_i} e^{-\frac{1}{2\sigma_2^2}\sum(1-I_i)(X_i-\theta_2)^2} \frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2}
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{P}\left(\theta_1 \mid \theta_2, \sigma_1^2, \sigma_2^2, \rho, I_1, \dots, I_n, X\right) &\propto e^{-\frac{1}{2\sigma_1^2}\sum I_i(X_i^2 - 2X_i\theta_1 + \theta_1^2)} \\
&\propto e^{-\frac{1}{2\sigma_1^2}(-2\theta_1 \sum I_i X_i + \theta_1^2 \sum I_i)} \\
&= \propto e^{\frac{\sum I_i X_i}{\sigma_1^2}\theta_1 - \frac{\sum I_i}{\sigma_1^2}\theta_1^2} \\
&\propto N\Big(\frac{\sum I_i X_i}{\sum I_i}, \frac{\sigma_1^2}{\sum I_i}\Big) \\
\mathbb{P}\left(\theta_2 \mid \theta_1, \sigma_1^2, \sigma_2^2, \rho, I_1, \dots, I_n, X\right) &\propto N\Big(\frac{\sum (1 - I_I) X_I}{\sum 1 - I_i}, \frac{\sigma_2^2}{\sum 1 - I_i}\Big) \\
\mathbb{P}\left(\sigma_1^2 \mid \theta_1, \theta_2, \sigma_2^2, \rho, I_1, \dots, I_n, X\right) &\propto (\sigma_1^2)^{-\frac{\sum I_i}{2} - 1} e^{-\frac{\sum I_i (X_i - \theta_1)^2 / 2}{\sigma_1^2}} \\
&\propto \text{InvGamma}\Big(\frac{\sum I_i}{2}, \frac{\sum I_i (X_i - \theta_1)^2}{2}\Big) \\
\mathbb{P}\left(\sigma_2^2 \mid \theta_1, \theta_2, \sigma_1^2, \rho, I_1, \dots, I_n, X\right) &\propto \text{InvGamma}\Big(\frac{\sum 1 - I_i}{2}, \frac{\sum (1 - I_i)(X_i - \theta_2)^2}{2}\Big) \\
\mathbb{P}\left(\rho \mid \theta_1, \theta_2, \sigma_1^2, \sigma_2^2, I_1, \dots, I_n, X\right) &\propto \rho^{\sum I_i}(1 - \rho)^{\sum 1 - I_i} \\
&\propto \text{Beta}(1 + \sum I_i, 1 + \sum 1 - I_i) \\
\mathbb{P}\left(I_1 \mid \theta_2, \sigma_1^2, \sigma_2^2, I_2, \dots, I_n, X\right) &\propto \Big(\rho \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i - \theta_1)^2}\Big)^{I_i} \Big((1 - \rho)\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(X_i - \theta_2)^2}\Big)^{1 - I_i} \\
&\propto \text{Bern}\left(\frac{\rho \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i - \theta_1)^2}}{\rho \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i - \theta_1)^2} + (1 - \rho)\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(X_i - \theta_2)^2}}\right) \\
\mathbb{P}\left(I_2 \mid \theta_2, \sigma_1^2, \sigma_2^2, I_1, I_3, \dots, I_n, X\right) &\propto \dots \\
&\vdots \\
\mathbb{P}\left(I_n \mid \theta_2, \sigma_1^2, \sigma_2^2, I_1, \dots, I_{n-1}, X\right) &\propto \dots
\end{aligned}$$