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:

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

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

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

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

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

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

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

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

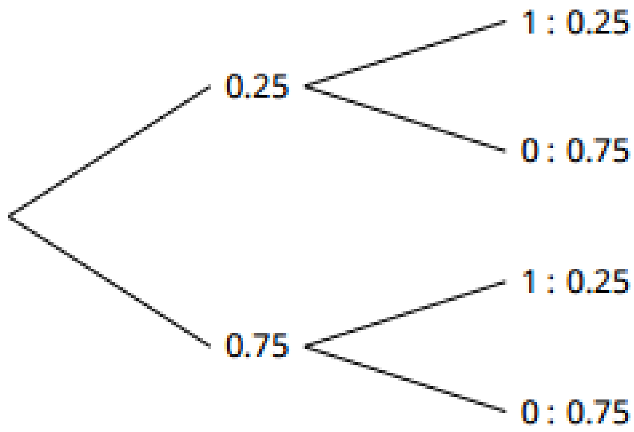
Proof.

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

$\square$

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 $X^*|X_1, X_2, X_3 \sim \text{Bern}(0.625)$. 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.

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

$\square$

In general,

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

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

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? $\text{Supp}(\theta) =$ 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

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

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.