Recap:

$$\underbrace{f(\theta \mid \text{data})}_{\text{posterior}} \propto \underbrace{f(\text{data} \mid \theta)}_{\text{likelihood}} \underbrace{f(\theta)}_{\text{prior}}.$$

This is the picture you should have in mind:

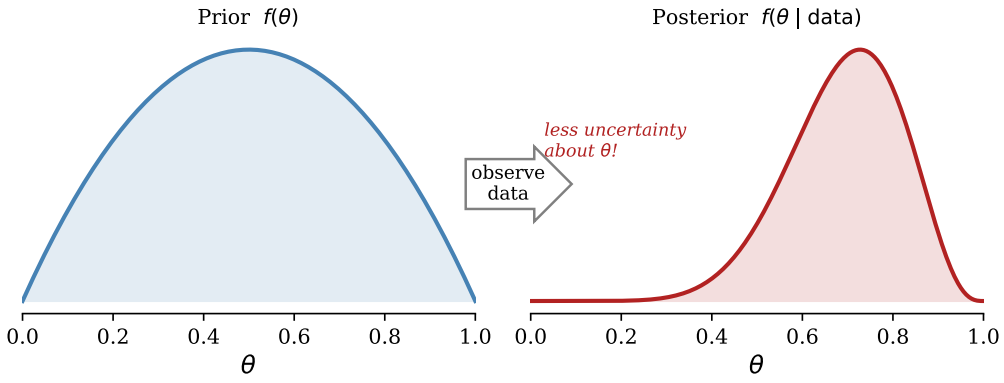



Figure 35: Observing data transforms a wide prior into a narrower posterior. The posterior carries less uncertainty about $\theta$ than the prior did, because the data gave us information about $\theta$.

## 4.5 Recognizing posteriors by their structure

A key computational trick: since the posterior is determined up to a normalizing constant, we can often *identify* which family it belongs to just by inspecting the functional form of $f(\theta \mid \text{data}) \propto L(\theta)\, f(\theta)$, without ever computing the normalizing constant explicitly.
The general strategy is: compute $L(\theta)\, f(\theta)$, drop all factors that do not depend on $\theta$, then *squint* at what remains and match it to a known density family. Once the family is identified, the normalizing constant comes for free.

**Example 4.3.** Suppose after computing $L(\theta)\, f(\theta)$ we find

$$f(\theta \mid X_1, \ldots, X_n) \propto e^{-\frac{\theta^2}{2\bar{X}_n}}.$$

This matches the form of a Gaussian density. Indeed, using $\theta$ as a dummy variable, we have

$$\text{pdf of } \mathcal{N}(\mu, \sigma^2) \;\propto e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}.$$

Comparing the above two equations, we find that $\mu = 0$ and $\sigma^2 = \bar{X}_n$. We conclude that

$$\theta \mid X_1, \ldots, X_n \sim \mathcal{N}(0,\, \bar{X}_n).$$

If we cared about the normalizing constant, we could fill it in, using the density of the normal:

$$f(\theta \mid X_1, \ldots, X_n) = \frac{1}{\sqrt{2\pi \bar{X}_n}}\, e^{-\frac{\theta^2}{2\bar{X}_n}}.$$

**Example 4.4.** Suppose after computing $L(\theta)\, f(\theta)$ we find

$$f(\theta \mid X_1, \ldots, X_n) \propto \theta^t e^{-\lambda\theta}.$$

Here, $t$ and $\lambda$ should depend on the data in some way. We're not focusing on any specific model for now, so it doesn't matter what $t$ and $\lambda$ are exactly. We recognize that the above density matches the form of a gamma density. Indeed, using $\theta$ as a dummy variable, we have

$$\text{pdf of Gamma}(\alpha, \beta) \;\propto \theta^{\alpha-1} e^{-\beta\theta}.$$

Comparing the above two equations, we find that $\alpha = t + 1$ and $\beta = \lambda$. We conclude that

$$\theta \mid X_1, \ldots, X_n \sim \text{Gamma}(t + 1,\ \lambda).$$

If we cared about the normalizing constant, we could fill it in, using the known density of the gamma:

$$f(\theta \mid X_1, \ldots, X_n) = \frac{\lambda^{t+1}}{\Gamma(t+1)}\, \theta^{(t+1)-1} e^{-\lambda\theta}.$$

## 4.6 Conjugate priors

**Definition 4.1** (Conjugate prior)**.** A prior $f(\theta)$ is called **conjugate** (for a given likelihood) if the posterior $f(\theta \mid \text{data})$ belongs to the same parametric family as the prior.

Conjugate priors are computationally convenient: the "Bayesian update" simply changes the parameters of the distribution rather than changing the distributional family entirely. We now work through three canonical examples.

### 4.6.1 Gamma–Poisson conjugacy

- **Prior:** $\theta \sim \text{Gamma}(a_0,\ b_0)$, so $f(\theta) = \dfrac{b_0^{a_0}}{\Gamma(a_0)}\, \theta^{a_0-1} e^{-b_0\theta}$.
- **Data model:** $X_1, \ldots, X_n \mid \theta \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$, so $f(X_i \mid \theta) = \dfrac{e^{-\theta}\theta^{X_i}}{X_i!}$.
- **Posterior:**

$$\begin{aligned} f(\theta \mid X_1, \ldots, X_n) \propto L(\theta)\, f(\theta) &= \prod_{i=1}^{n} f(X_i \mid \theta) \cdot f(\theta) \\ &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{X_i}}{X_i!} \cdot \frac{b_0^{a_0}}{\Gamma(a_0)}\, \theta^{a_0-1} e^{-b_0\theta} \\ &\propto e^{-n\theta}\, \theta^{\sum X_i} \cdot \theta^{a_0-1} e^{-b_0\theta} \\ &= \theta^{(\sum X_i + a_0)-1}\, e^{-(n+b_0)\theta}. \end{aligned}$$

This is the kernel of a Gamma$(\sum X_i + a_0,\ n + b_0)$ density. Therefore:

$$\boxed{\theta \mid X_1, \dots, X_n \ \sim\ \text{Gamma}\big(a_0 + \textstyle\sum_{i=1}^{n} X_i,\ \ b_0 + n\big)\,.}$$

The Gamma prior is conjugate for Poisson data. Observing data increments the shape parameter by the total count $\sum X_i$ and increments the rate parameter by $n$.

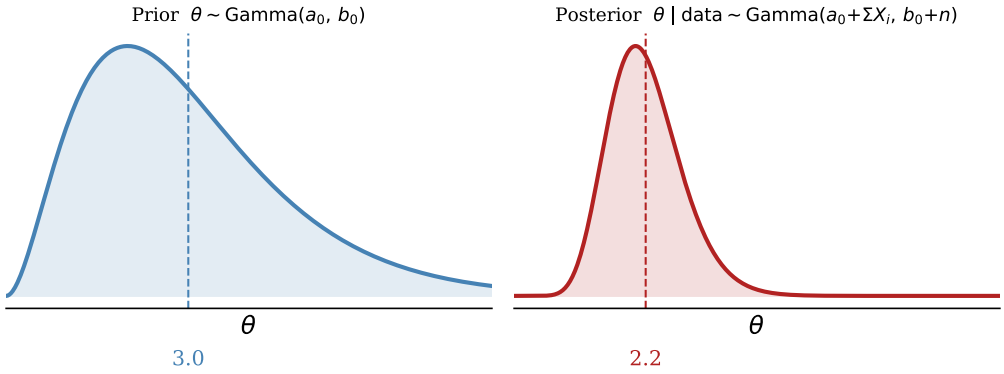



Figure 36: Gamma–Poisson conjugacy with $a_0 = 3$, $b_0 = 1$ and $n = 5$ observations summing to 10. Thus the prior is Gamma$(3, 1)$ and the posterior is Gamma$(13, 6)$. The prior mean is 3.0 with variance 3.0; the posterior mean is 2.2 with variance $13/36 \approx 0.36$. The variance was reduced by a factor of about 8.3. The vertical dashed lines show the means.

### 4.6.2 Beta–Bernoulli conjugacy

The Beta distribution, which we introduce now, is a conjugate prior for Bernoulli (coin-flip) data.

**Definition 4.2** (Beta distribution)**.** We say $X \sim \text{Beta}(a, b)$ if $\text{Range}(X) = (0, 1)$ and

$$f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1.$$

The parameters $a$ and $b$ are both positive.

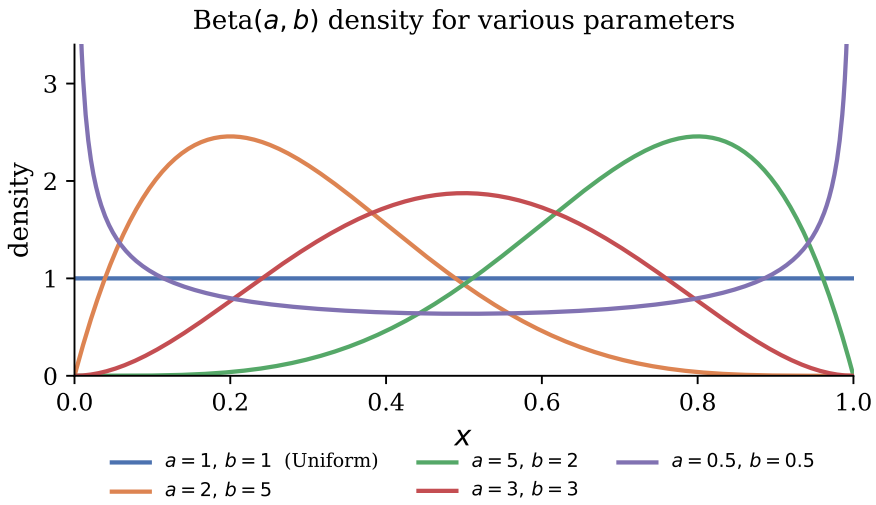



Figure 37: The Beta$(a, b)$ density for several parameter choices. The shape ranges from U-shaped ($a = b = \frac{1}{2}$) to uniform ($a = b = 1$) to unimodal and skewed or symmetric.

**Remark 4.1.** When $a = b = 1$:

$$f_X(x) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} x^0(1-x)^0 = \frac{1!}{1 \cdot 1} = 1, \qquad 0 < x < 1.$$

So $\text{Beta}(1, 1) = \text{Unif}(0, 1)$.

**Beta integral identity.** Since $f_X$ integrates to 1, we immediately obtain the useful identity

$$\int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$$

This identity is invaluable for computing normalizing constants and expectations involving Beta-shaped kernels.

**Mean of the Beta distribution.**

$$\begin{aligned}\mathbb{E}[X] &= \int_0^1 x\, f_X(x)\,\mathrm{d}x = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^a(1-x)^{b-1}\,\mathrm{d}x \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\,\Gamma(b)}{\Gamma(a+1+b)} = \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{\Gamma(a+1)}{\Gamma(a+b+1)}.\end{aligned}$$

Using $\Gamma(a+1) = a\,\Gamma(a)$ and $\Gamma(a+b+1) = (a+b)\,\Gamma(a+b)$:

$$\mathbb{E}[X] = \frac{\Gamma(a+b)}{\Gamma(a)} \cdot \frac{a\,\Gamma(a)}{(a+b)\,\Gamma(a+b)} = \boxed{\frac{a}{a+b}.}$$

**Variance of the Beta distribution** (do it yourself!):

$$\text{var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

**Beta–Bernoulli conjugacy.**

- **Prior:** $\theta \sim \text{Beta}(a_0,\, b_0)$.
- **Data model:** $X_1, \ldots, X_n \mid \theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, so $f(X_i \mid \theta) = \theta^{X_i}(1-\theta)^{1-X_i}$.
- **Posterior:**

$$f(\theta \mid X_1, \ldots, X_n) \propto \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} \cdot \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)}\,\theta^{a_0-1}(1-\theta)^{b_0-1}$$
$$\propto \theta^{\sum X_i + a_0 - 1}\,(1-\theta)^{n-\sum X_i + b_0 - 1}.$$

This is the kernel of a Beta density, so:

$$\boxed{\theta \mid X_1, \ldots, X_n \;\sim\; \text{Beta}\Big(a_0 + \textstyle\sum_{i=1}^n X_i,\;\; b_0 + n - \sum_{i=1}^n X_i\Big)\,.}$$

Thus we have shown the Beta prior is conjugate for Bernoulli data. Observing heads increments the first shape parameter; observing tails increments the second.

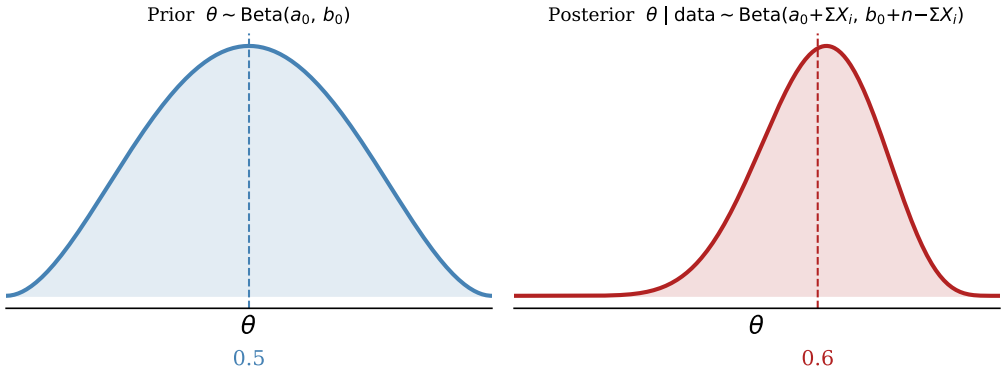



Figure 38: Beta–Bernoulli conjugacy with $a_0 = b_0 = 3$ and $n = 10$ trials yielding 7 heads. Thus the prior is Beta(3, 3) and the posterior is Beta(10, 6). The prior mean is 0.5 with variance $1/28 \approx 0.036$; the posterior mean is 0.6 with variance $60/4352 \approx 0.014$. The variance was reduced by a factor of about 2.6. The vertical dashed lines show the means.

### 4.6.3 Normal–Normal conjugacy

We consider data drawn i.i.d. from $\mathcal{N}(\theta, 1)$, and want to infer $\theta$.

- **Prior:** $\theta \sim \mathcal{N}(m_0,\, \tau_0^2)$.
- **Data model:** $X_1, \ldots, X_n \mid \theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta,\, 1)$. The variance $\sigma^2 = 1$ is *known* to us.

  You will work this problem out in the study guide, so I am skipping the algebra. In the end, we obtain

- **Posterior:**

$$\boxed{\theta \mid X_1, \dots, X_n \sim \mathcal{N}(m_n,\, \tau_n^2)}, \quad \text{where} \quad m_n = \frac{n\tau_0^2\, \bar{X}_n + m_0}{n\tau_0^2 + 1}, \quad \tau_n^2 = \frac{\tau_0^2}{n\tau_0^2 + 1}.$$

Thus the normal prior is conjugate for normal data.

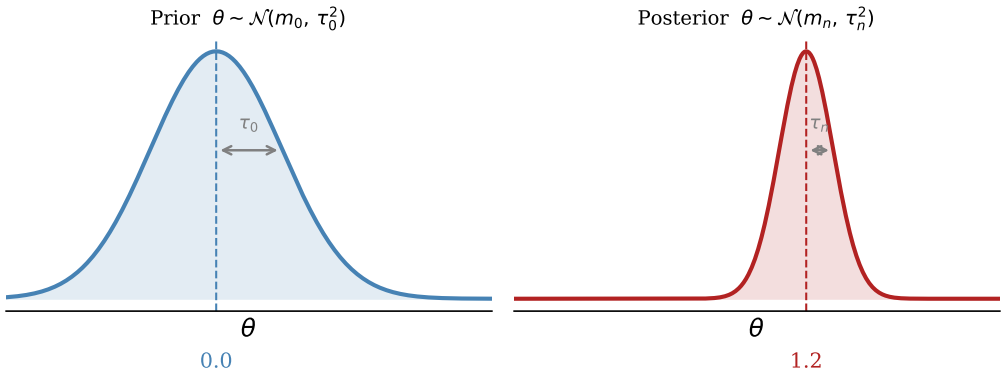


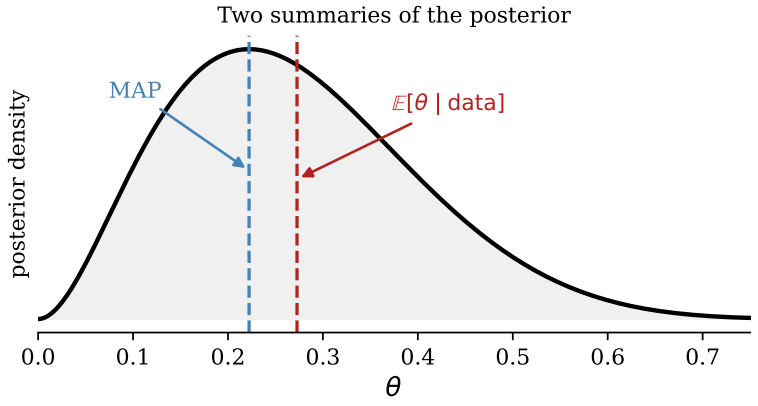

Figure 39: Normal–Normal conjugacy with $m_0 = 0$, $\tau_0 = 1$, $n = 5$ observations, and $\bar{X}_n = 1.5$. Thus the prior is $\mathcal{N}(0,\, 1)$ and the posterior is $\mathcal{N}(1.25,\, 1/6)$. The prior mean is 0.0 with variance 1; the posterior mean is 1.2 with variance $1/6 \approx 0.17$. The variance was reduced by a factor of 6. The vertical dashed lines show the means.

## 4.7 Summarizing the posterior

The full posterior $f(\theta \mid X_1, \dots, X_n)$ is a complete description of our uncertainty about $\theta$ after seeing data. In practice we often want a single-number summary. Two natural choices are:

1. **Posterior mean** (Bayes estimator): $\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta \mid X_1, \dots, X_n]$.
2. **Posterior mode** (Maximum a posteriori, or MAP):

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} f(\theta \mid X_1, \dots, X_n).$$



**MAP vs MLE.** Since $f(\theta \mid \text{data}) \propto L(\theta)\, f(\theta)$, the MAP satisfies

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} L(\theta)\, f(\theta).$$

With a *uniform* prior $f(\theta) = c$ (i.e. constant over its domain), this reduces to

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ln L(\theta) = \hat{\theta}_{\text{MLE}}.$$

Thus if we were asked to give a single "most likely" value of $\theta$ based on the posterior, and our prior assigned equal probabilities to all values of $\theta$, then we would just return the MLE.

Another useful way to summarize the posterior is with a *credible* interval.

**Definition 4.3** (Credible interval)**.** A $100(1-\alpha)\%$ credible interval for $\theta$ is any interval $(a, b)$ such that

$$\mathbb{P}(a < \theta < b \mid X_{1:n}) = 1 - \alpha.$$

For example, for a 90% credible interval $(a, b)$, we have $\mathbb{P}(a < \theta < b \mid X_{1:n}) = 0.9$.

Although any interval whose probability is 0.9 counts as a 90% credible interval, we are often most interested in a centrally-located credible interval. We can construct one by chopping off upper and lower chunks of the distribution, each of which have mass 0.05. Then the middle region will have probability $1 - 0.05 - 0.05 = 0.9$. See Figure 40.

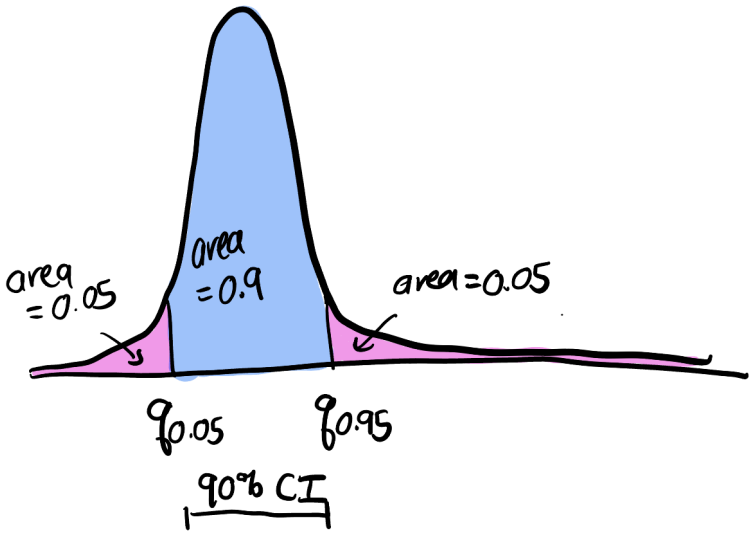



Figure 40: A centrally-located 90% credible interval for $\theta$.

Formally, this middle region is the interval between $q_{0.05}$ and $q_{0.95}$, where $q_\alpha$ is the $\alpha$th quantile of the posterior: the number such that

$$\mathbb{P}(\theta \leq q_\alpha \mid X_{1:n}) = \alpha.$$

Let us check that $\theta$ lies between $q_{0.05}$ and $q_{0.95}$ with probability 0.9:

$$\begin{aligned}\mathbb{P}(q_{0.05} \le \theta \le q_{0.95} \mid X_{1:n}) &= \mathbb{P}(\theta \le q_{0.95} \mid X_{1:n}) - \mathbb{P}(\theta \le q_{0.05} \mid X_{1:n}) \\ &= 0.95 - 0.05 = 0.9.\end{aligned}$$

## 4.8 The posterior mean as a classical estimator

The posterior mean $\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta \mid X_1, \dots, X_n]$ is a point estimator in the classical sense: it is functions of the data $X_1, \dots, X_n$ which can be used to estimate the unknown parameter $\theta$. So it makes sense to ask about its properties: what is the bias, variance, and MSE? Is it a consistent estimator? We will primarily be concerned with how $\hat{\theta}_{\text{Bayes}}$ performs relative to the other main estimator we've seen so far: the MLE.

Let us consider the three examples from Section 4.6. In each case, we will show that $\hat{\theta}_{\text{Bayes}}$ is a *convex combination* of the MLE and the prior mean. We will then use this fact to compare the MSE of $\hat{\theta}_{\text{Bayes}}$ to the MSE of $\hat{\theta}_{\text{MLE}}$.

**Definition 4.4** (Convex combination)**.** A **convex combination** of $x$ and $y$ is any number of the form $wx + (1-w)y$ where $w \in (0,1)$. Geometrically, all convex combinations lie on the line segment connecting $x$ and $y$.

- **Gamma prior, Poisson data.** Recall the setting from Section 4.6.1: we have data from a Poisson($\theta$), and we put a prior Gamma($a_0, b_0$) on $\theta$. We computed that the posterior is Gamma$\big(a_0 + \sum_{i=1}^n X_i,\ \ b_0 + n\big)$. We're going to suppose for simplicity that $b_0 = 1$. Then the prior mean is $a_0/1 = a_0$, and the posterior mean is $(a_0 + \sum X_i)/(1+n)$.

  Also, the MLE for Poisson data is $\hat{\theta}_{\text{MLE}} = \bar{X}_n$ (do it yourself!).

  We can rewrite the posterior mean as a convex combination of the prior mean $a_0$ and the MLE $\bar{X}_n$.

$$\begin{aligned}\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta \mid X_{1:n}] &= \frac{a_0 + \sum X_i}{1+n} \\ &= \frac{a_0 + n\bar{X}_n}{1+n} \\ &= \frac{1}{1+n} a_0 + \frac{n}{1+n} \bar{X}_n \\ &= (1 - w_n) a_0 + w_n \bar{X}_n,\end{aligned}$$

  where $w_n = n/(1+n)$.

- **Beta prior, Bernoulli data.** Recall the setting from Section 4.6.2: we have data from Bernoulli($\theta$), and we put a prior Beta($a_0, b_0$) on $\theta$. We computed that the posterior is Beta$(a_0 + \sum X_i,\ \ b_0 + n - \sum X_i)$. We're going to suppose for simplicity that $b_0 = 1 - a_0$. Then the prior mean is $a_0/(b_0 + a_0) = a_0$, and the posterior mean is $(a_0 + \sum X_i)/(1+n)$.

  Also, the MLE for Bernoulli data is $\hat{\theta}_{\text{MLE}} = \bar{X}_n$ (recall Example 3.3).

The prior mean, posterior mean, and MLE are exactly the same as in the Gamma-Poisson example, so we get

$$\hat{\theta}_{\text{Bayes}} = (1 - w_n)a_0 + w_n \bar{X}_n,$$

where $w_n = n/(1+n)$.

- **Normal prior, normal data.** Recall the setting from Section 4.6.3: we have data from $\mathcal{N}(\theta, 1)$, and we put a prior $\mathcal{N}(m_0, \tau_0)$ on $\theta$. We computed that the posterior is $\mathcal{N}(m_n, \tau_n)$, where $m_n = \frac{n\tau_0^2 \bar{X}_n + m_0}{n\tau_0^2 + 1}$. Let's suppose for simplicity that $\tau_0 = 1$, and use $a_0$ for the prior mean, instead of $m_0$. Then the posterior mean is

$$\hat{\theta}_{\text{Bayes}} = m_n = \frac{n\bar{X}_n + a_0}{n+1} = (1 - w_n)a_0 + w_n \bar{X}_n,$$

where $w_n = n/(1+n)$.

The MLE for normal data is, again, $\hat{\theta}_{\text{MLE}} = \bar{X}_n$ (you worked this out in Lab 10). Thus as in the previous two examples, $\hat{\theta}_{\text{Bayes}}$ is a convex combination of the prior mean and the MLE.

We see that in all three examples,

$$\hat{\theta}_{\text{Bayes}} = (1 - w_n)a_0 + w_n \bar{X}_n. \tag{30}$$

Or in plain language,

$$\text{posterior mean} = (1 - w_n)(\text{prior mean}) + w_n(\text{MLE}).$$

The weight $w_n = n/(n+1)$ converges to 1 as $n \to \infty$, so the posterior mean becomes closer and closer to the MLE. At the same time, the weight assigned to the prior mean diminishes. This makes sense: as we accrue more data, the data become more informative than the prior.

Let's use the general formula (30) to study the bias, variance, and MSE of $\hat{\theta}_{\text{Bayes}}$ as compared to that of the MLE $\bar{X}_n$. Since $\bar{X}_n$ is the only random part of (30), and $\mathbb{E}[\bar{X}_n] = \theta$, we get:

| | $\hat{\theta}_{\text{Bayes}}$ | $\bar{X}_n$ |
|---|---|---|
| bias | $(1 - w_n)a_0 + w_n\theta - \theta = (1 - w_n)(a_0 - \theta)$ | 0 |
| variance | $w_n^2 \operatorname{Var}(\bar{X}_n)$ | $\operatorname{Var}(\bar{X}_n)$ |
| MSE | $(1 - w_n)^2(a_0 - \theta)^2 + w_n^2 \operatorname{Var}(\bar{X}_n)$ | $\operatorname{Var}(\bar{X}_n)$ |

Since $w_n < 1$, the Bayes estimator has strictly smaller variance than the MLE. But it is biased (unless $a_0 = \theta$), so it is not obvious which has the smaller MSE.

**When does the Bayes estimator beat the MLE?** We have $\mathrm{MSE}(\hat{\theta}_{\mathrm{Bayes}}) < \mathrm{MSE}(\bar{X}_n)$ iff

$$(1-w_n)^2(a_0-\theta)^2 + w_n^2 \operatorname{Var}(\bar{X}_n) < \operatorname{Var}(\bar{X}_n)$$
$$(1-w_n)^2(a_0-\theta)^2 < (1-w_n^2)\operatorname{Var}(\bar{X}_n).$$

Dividing both sides by $(1-w_n)$ and using $1-w_n^2 = (1-w_n)(1+w_n)$:

$$(1-w_n)(a_0-\theta)^2 < (1+w_n)\operatorname{Var}(\bar{X}_n),$$

i.e.

$$(a_0-\theta)^2 < \frac{1+w_n}{1-w_n}\operatorname{Var}(\bar{X}_n).$$

With $w_n = n/(n+1)$ we have $1+w_n = (2n+1)/(n+1)$ and $1-w_n = 1/(n+1)$, so $\frac{1+w_n}{1-w_n} = 2n+1$. The condition becomes

$$\boxed{(a_0-\theta)^2 < (2n+1)\operatorname{Var}(\bar{X}_n) = \frac{2n+1}{n}\operatorname{Var}(X_1) \approx 2\operatorname{Var}(X_1).}$$

The Bayes estimator beats the MLE whenever the prior mean $a_0$ is within roughly $\sqrt{2\operatorname{Var}(X_1)}$ of the truth $\theta$. For each of the three models:

- **Poisson**: $\operatorname{Var}(X_1) = \theta$, so the condition is $(a_0-\theta)^2 < 2\theta$.
- **Bernoulli**: $\operatorname{Var}(X_1) = \theta(1-\theta)$, so the condition is $(a_0-\theta)^2 < 2\theta(1-\theta)$.
- **Normal**: $\operatorname{Var}(X_1) = 1$, so the condition is $(a_0-\theta)^2 < 2$, i.e. $|a_0-\theta| < \sqrt{2}$.

**Remark 4.2.** In all three cases the Bayes estimator is biased but consistent, and it beats the MLE whenever the prior mean is close enough to the truth. This is the bias–variance tradeoff in action: the prior introduces bias but reduces variance.

**Remark 4.3.** The posterior mode $\hat{\theta}_{\mathrm{MAP}}$ is also a totally valid estimator of $\theta$, and I encourage you to analyze its properties: what is its bias and variance, and is it a consistent estimator? How does it compare to the MLE? We have already shown that when the prior is uniform, MAP=MLE.