# 8. Beyond MLE: Using Prior Information

## ECE 830 & CS 761, Spring 2016

# The Bayesian Paradigm

Given a parameter $\theta$, we assume observations are generated according to $p(y|\theta)$. In our work so far, we have treated the parameter $\theta$ like a fixed, deterministic, but unknown quantity while the observation $y$ is the realization of a random process.

We will now consider **probabilistic models** for $\theta$ in addition to our data.

- This allows us to incorporate prior information we have about $\theta$ (i.e. information about likely values of $\theta$ we have *before collecting any data*).
- It also allows us to make statements about our confidence in different estimates of $\theta$.

## Example: Unfair coin

Suppose you toss a single coin $6$ times and each time it comes up "heads." It might be reasonable to say that we are $98\%$ sure that the coin is unfair, biased towards heads.


Formally, we can think about this in a hypothesis testing framework using a binomial probabilistic model. Let $k := \sum_{i=1}^{6} y_i$.

$$H_0 : \text{prob heads } \equiv \theta > 0.5$$

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

$$p(\theta > 0.5|y) = ?$$

## Example: (cont.)

The problem with this is that

$$p(\theta \in H_0 | x)$$

implies that $\theta$ is a random, not deterministic, quantity.

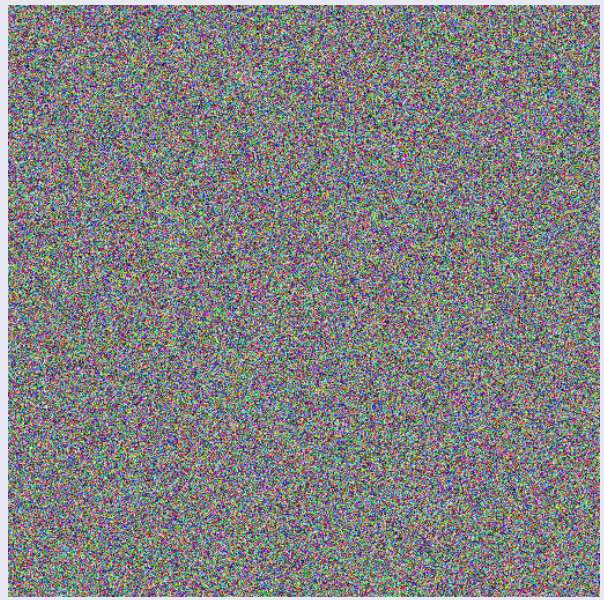
So, while "confidence" statements are very reasonable and in fact a normal part of "everyday thinking," this idea can not be supported from the classical perspective.

All of these "deficiencies" can be circumvented by a change in how we view the parameter $\theta$.

## Example: Image processing

In many imaging problems, we have a good sense of what "natural" images should look like.

Likely



Unlikely

This prior information can be exploited to improve image denoising, deblurring, reconstruction, and analysis.

# Bayes Rule

If we view $\theta$ as the realization of a random variable with density $p(\theta)$, then we can work with the generative (or forward) model

$$\underbrace{p(\theta)}_{\text{prior}} \rightarrow \theta^* \rightarrow \underbrace{p(y|\theta^*)}_{\text{likelihood}} \rightarrow y.$$

We are interested in the inverse problem

$$y \rightarrow p(\theta|y) \rightarrow \hat{\theta}.$$

Bayes Rule (Bayes, 1763) shows that

$$p(\theta|y) = \frac{p(y|\theta)\ p(\theta)}{p(y)} = \frac{p(y|\theta)\ p(\theta)}{\int p(y|\widetilde{\theta})\ p(\widetilde{\theta})\ d\widetilde{\theta}}$$

Once we can compute this posterior distribution, confidence measures such as $p(\theta \in H_0|y)$ are perfectly legitimate quantities to ask for.

## Example: Coin toss

Suppose you toss a single coin $6$ times and each time it comes up "heads." Mathematically, we can model the problem as follows. Let $\theta = \mathbb{P}(\text{Heads})$. The data (the number of heads $y$ in $n = 6$ tosses) follows a binomial distribution $p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$. The mathematical equivalent of the question "is the coin probably biased" is the probability $\mathbb{P}(\theta > 0.5|y = 6)$.

Suppose we assume $p(\theta) = \text{Unif}(0, 1)$ (all values of $\theta$ are equally probable before we begin to flip the coin, and $\mathbb{P}(\theta > \frac{1}{2}) = \frac{1}{2}$). Now compute

$$p(\theta|y) \;=\; \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \;=\; \frac{\theta^6}{\int \theta^6 d\theta} = \frac{\theta^6}{\frac{1}{7}\theta^7|_0^1} \;=\; 7\,\theta^6 \;.$$

Then

$$\mathbb{P}\left(\theta > \frac{1}{2}\,|\,y = 6\right) \;=\; \int_{\frac{1}{2}}^{1} 7\theta^6 d\theta \;=\; \theta^7|_{\frac{1}{2}}^1 \;=\; 1 - 2^{-7} \;=\; 0.984 \;.$$

(If we chose a different prior we would get a different answer!)

# Bayesian statistical models

**Definition: Bayesian statistical model**

A Bayesian statistical model is composed of a *data generation model*, $p(y|\theta)$, and a *prior* distribution on the parameters, $p(\theta)$.

The prior distribution (or "prior" for short) models the uncertainty in the parameter. More specifically, $p(\theta)$ models our knowledge - or a lack thereof - prior to collecting data.

Notice that

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)} \propto p(y|\theta)\, p(\theta)$$

Hence, $p(\theta|y)$ is proportional to the likelihood function multiplied by the prior.

## Example: DC level in AWGN

$$
\begin{aligned}
y_i &= A + \nu_i \ , \quad n = 1, \cdots, N \\
\nu_i &\sim \mathcal{N}(0, \sigma^2) \ \ iid \\
\widehat{A} &= \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{MLE estimate}
\end{aligned}
$$

Now suppose that we have prior knowledge that $-A_0 \leq A \leq A_0$. We might incorporate this by forming a new estimator

$$
\widetilde{A} = \begin{cases} -A_0 & , \ \widehat{A} < -A_0 \\ \widehat{A} & , \ -A_0 \leq \widehat{A} \leq A_0 \\ A_0 & , \ \widehat{A} > A_0 \end{cases}
$$

This is called a truncated sample mean estimator of $A$.

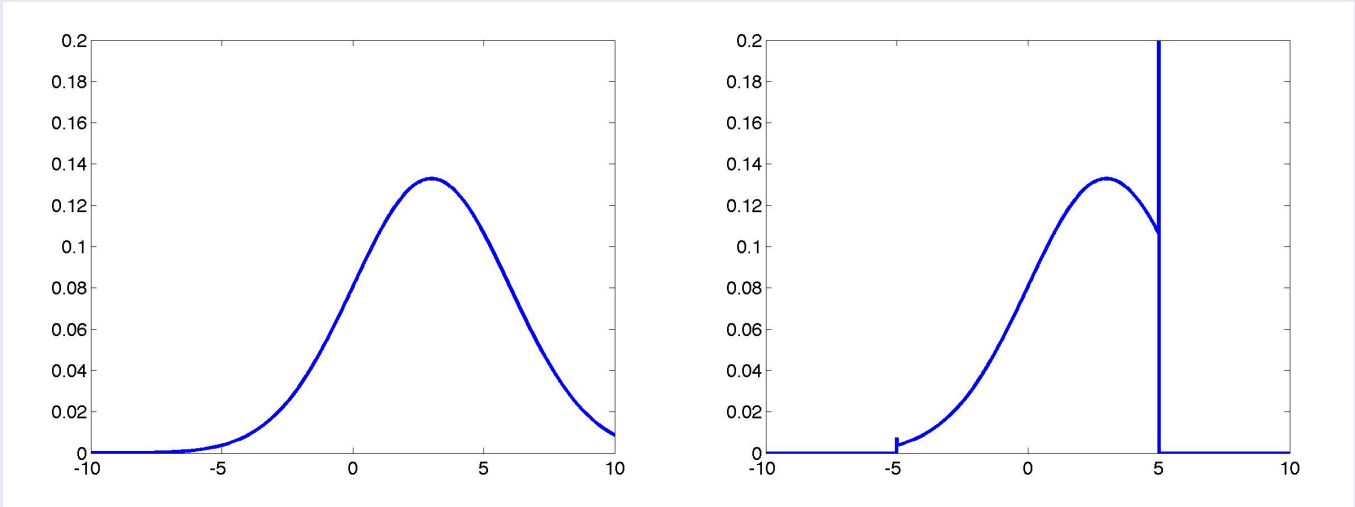
## Example: (cont.)

Is $\widetilde{A}$ a better estimator of $A$ than the sample mean $\widehat{A}$? Let $p_{\mathrm{MLE}}$ denote the density of $\widehat{A}$. Since $\widehat{A} = \frac{1}{n}\sum y_i$,

$$p_{\mathrm{MLE}} = \mathcal{N}(A, \sigma^2/n).$$

The density of the truncated sample mean (TSM) $\widetilde{A}$ is given by

$$\begin{aligned} p_{\mathrm{TSM}} =& \mathbb{P}(\widehat{A} \le -A_0)\ \delta(a + A_0) + p_{\mathrm{MLE}} I_{\{-A_0 \le a \le A_0\}} \\ &+ \mathbb{P}(\widehat{A} \ge A_0)\ \delta(a - A_0) \end{aligned}$$




$p_{\mathrm{MLE}}$ $p_{\mathrm{TSM}}$

Now consider the MSE of the sample mean $\widehat{A}$:

$$
\begin{aligned}
\mathsf{MSE}(\widehat{A}) &= \int_{-\infty}^{\infty} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&= \int_{-\infty}^{-A_0} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da + \int_{-A_0}^{A_0} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&\quad + \int_{A_0}^{\infty} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&> \int_{-\infty}^{-A_0} (-A_0-A)^2 \, p_{\mathrm{MLE}}(a) \, da + \int_{-A_0}^{A_0} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&\quad + \int_{A_0}^{\infty} (A_0-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&= (-A_0-A)^2 \, \mathbb{P}(\widehat{A} \le -A_0) + \int_{-A_0}^{A_0} (a-A)^2 \, p_{\mathrm{MLE}}(a) \, da \\
&\quad + (A-A_0)^2 \, \mathbb{P}(\widehat{A} \ge A_0) \\
&= \mathsf{MSE}(\widetilde{A})
\end{aligned}
$$

$$\mathsf{MSE}(\widehat{A}) > \mathsf{MSE}(\widetilde{A})$$

**Note**

1. $\widetilde{A}$ is biased
2. Although $\widehat{A}$ is the MLE, $\widetilde{A}$ is better in the MSE sense
3. Prior information is aptly described by regarding $A$ as a random variable with a prior distribution.

$$\text{uniform}(-A_0, A_0)$$

$\Rightarrow$ We know $-A_0 \leq A \leq A_0$, but otherwise $A$ is arbitrary.

# Elements of Bayesian Analysis

**(a)** Joint distribution

$$p(y, \theta) = p(y|\theta)p(\theta)$$

**(b)** Marginal distributions

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

$$p(\theta) = \int p(y|\theta)p(\theta)dy \text{(“prior”)}$$

**(c)** Posterior distribution

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)dy}$$

## Example: Binomial + Beta

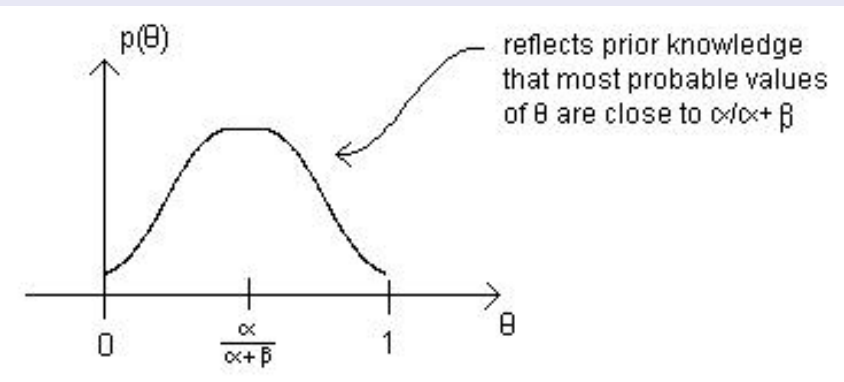
$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, 0 \leq \theta \leq 1$$

$$= \text{binomial likelihood}$$

$$p(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$= \text{Beta prior distribution}$$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$ is the Gamma function



## Example: (cont.)

- Joint Density

$$p(y, \theta) = \left[ \frac{\binom{n}{y}}{B(\alpha, \beta)} \right] \theta^{\alpha+y-1}(1-\theta)^{n-y+\beta-1}$$

- Marginal Density

$$p(y) = \left[ \binom{n}{y} \frac{1}{B(\alpha, \beta)} \right] B(\alpha + y, \beta + n - y)$$

- Posterior Density

$$p(\theta|y) = \underbrace{\frac{\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}}{B(\alpha + y, \beta + n - y)}}_{\text{beta density with parameters}}$$

$$\alpha' = \alpha + y$$
$$\beta' = \beta + n - y$$

# Bayesian Estimation

We are interested in estimating $\theta$ given the observation $y$ within a Bayesian framework. Naturally, then, any estimation strategy will be based on the posterior distribution $p(\theta|y)$.

However, we need a criterion for assessing the quality of potential estimators.

# Loss

## Definition: Loss

The quality of an *estimate* $\widehat{\theta}$ is measured by a real-valued *loss (or cost) function*

$$L(\theta, \widehat{\theta}).$$

For example, *squared error* or quadratic loss is simply

$$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta}(y))^\top (\theta - \widehat{\theta}(y)) = \|\theta - \widehat{\theta}(y)\|^2.$$

## Definition: Bayes Risk

The quality of an estimator is measured by the expected loss, known as the Bayes risk:

$$R(\widehat{\theta}) := \mathbb{E}_{y,\theta}\left[L(\theta, \widehat{\theta})\right].$$

**Note that the expectation is with respect to both $y$ and $\theta$.**

For example, if $y$ and $\theta$ are jointly continuous, then

$$\begin{aligned} R(\widehat{\theta}) &= \iint L(\theta, \widehat{\theta}(y)) p(\theta, y) dy d\theta \\ &= \iint L(\theta, \widehat{\theta}(y)) p(y|\theta) p(\theta) dy d\theta \\ &= \mathbb{E}_y \left[ \mathbb{E}_{\theta|y} \left[ L(\theta, \widehat{\theta}(y)) | y = y \right] \right] \end{aligned}$$

In general, Bayesian estimation seeks the estimator

$$\begin{aligned}
\widehat{\theta} &= \arg\min_{\widetilde{\theta}} R(\widetilde{\theta}) \\
&= \arg\min_{\widetilde{\theta}} \mathbb{E}_{y,\theta} \left\{ L\left(\theta, \widetilde{\theta}(y)\right) \right\} \\
&= \arg\min_{\widetilde{\theta}} \mathbb{E}_{y} \left\{ \mathbb{E}_{\theta|y} \left\{ L\left(\theta, \widetilde{\theta}(y)\right) | y = y \right\} \right\}
\end{aligned}$$

minimizing the Bayes risk. Thus, given the data $y$, the "best" or optimal estimator under a given loss function is given by

$$\widehat{\theta}(y) = \arg\min_{\widetilde{\theta}} \mathbb{E}\left[ L\left(\theta, \widetilde{\theta}\right) | y \right].$$

This is called the "posterior expected loss"; it depends only on the loss function and the posterior distribution.

# Bayesian estimation with squared error loss

Measure the loss as $L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|^2$. Now note

$$\begin{aligned}
&\mathbb{E}_{\theta|y=y}[(\theta - \widehat{\theta}(y))^\top(\theta - \widehat{\theta}(y))] \\
&= \mathbb{E}_{\theta|y}[(\theta - \mathbb{E}[\theta|y] + \mathbb{E}[\theta|y] - \widehat{\theta}(y))^\top(\theta - \mathbb{E}[\theta|y] + \mathbb{E}[\theta|y] - \widehat{\theta}(y))] \\
&= \mathbb{E}_{\theta|y}[(\theta - \mathbb{E}[\theta|y])^\top(\theta - \mathbb{E}[\theta|y])] + 2\mathbb{E}_{\theta|y}[(\theta - \mathbb{E}[\theta|y])^\top(\mathbb{E}[\theta|y] - \widehat{\theta}(y))] \\
&\quad + \mathbb{E}_{\theta|y}[(\mathbb{E}[\theta|y] - \widehat{\theta}(y))^\top(\mathbb{E}[\theta|y] - \widehat{\theta}(y))]
\end{aligned}$$

The first term is independent of $\widehat{\theta}(y)$ and the second term is $0$. The third term can be minimized by taking

$$\widehat{\theta}_{\mathrm{PM}}(y) = \mathbb{E}[\theta|y] = \int \theta p(\theta|y) d\theta$$

which is the posterior mean. (In signal processing this is also called the "Bayesian minimim mean squared error estimator".)

## Example: DC Level in AWGN

$$y_i = A + \nu_i$$

$i = 1, \cdots, n$, $\nu_i \sim \mathcal{N}(0, \sigma^2)$. Prior for unknown parameter $A$:

$$p(A) = \text{Unif}[-A_0, A_0]$$

$$p(y|A) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - A)^2 \right\}$$

$$p(A|y) = \begin{cases} \dfrac{\frac{1}{2A_0(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - A)^2\right\}}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a)^2\right\} da} & \text{if } |A| \le A_0 \\ 0 & \text{if } |A| > A_0 \end{cases}$$

## Example: (cont.)

Bayes Minimum MSE Estimator:

$$\begin{aligned}\widehat{A} &= \mathbb{E}[A|y] = \int_{-\infty}^{\infty} A p(A|y) dA \\ &= \frac{\int_{-A_0}^{A_0} A \cdot \frac{1}{2A_0(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - A)^2\right\} dA}{\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a)^2\right\} da}\end{aligned}$$

Notes:

1. No closed-form estimator
2. As $A_0 \to \infty$, $\widehat{A} \to \frac{1}{n}\sum_{i=1}^{n} y_i$
3. For smaller $A_0$, truncated integral produces an $\widehat{A}$ that is a function of $y$, $\sigma^2$, and $A_0$
4. As $n$ increases $\sigma^2/n$ decreases and posterior $p(A|y)$ becomes tightly clustered about $\frac{1}{n}\sum_i y_i$
   $\Rightarrow \widehat{A} \to \frac{1}{n}\sum y_i$ as $n \to \infty$
   (the data "swamps out" the prior)

# Other Common Loss Functions

## Absolute Error Loss (Laplace, 1773):

$$L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_1 := \sum_{i=1}^{p} |\theta_i - \widehat{\theta}_i|$$

Scalar case:

$$\begin{aligned} \mathbb{E}\left[L(\theta, \widehat{\theta})|y\right] &= \int_{-\infty}^{\infty} |\theta - \widehat{\theta}| p(\theta|y) d\theta \\ &= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) p(\theta|y) d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) p(\theta|y) d\theta \end{aligned}$$

The optimal estimator under this loss is referred to the "minimum mean absolute error" (MMAE) estimator.

To see what estimator minimises this loss, we differentiate $\mathbb{E}\left[L(\theta, \widehat{\theta})|y\right]$ with respect to $\widehat{\theta}$ (using Leibnitz's rule) to get

$$\frac{\partial}{\partial \widehat{\theta}} \mathbb{E}\left[L(\theta, \widehat{\theta})|y\right] = P(\widehat{\theta}(y)|y) - (1 - P(\widehat{\theta}(y)|y)),$$

where $P(\theta|y)$ is the posterior cumulative distribution function of $\theta$ given $y$. Setting this equal to zero, this implies $P(\widehat{\theta}(y)|y) = 1/2$ or

$$\mathbb{P}(\theta < \widehat{\theta}|y) = \mathbb{P}(\theta > \widehat{\theta}|y).$$

The optimal $\widehat{\theta}$ under absolute error loss is the posterior median.

# Uniform Loss:

$$L(\theta, \widehat{\theta}) = I_{\left\{\|\widehat{\theta}-\theta\|>\epsilon\right\}} = \begin{cases} 1 & \text{if } \|\theta - \widehat{\theta}\| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

where $\epsilon > 0$ is small. The posterior expected loss is

$$\mathbb{E}\left[L(\theta, \widehat{\theta})|y\right] = \mathbb{E}\left[I_{\left\{\|\widehat{\theta}-\theta\|>\epsilon\right\}}|y\right] = \mathbb{P}(\|\widehat{\theta} - \theta\| > \epsilon|y)$$

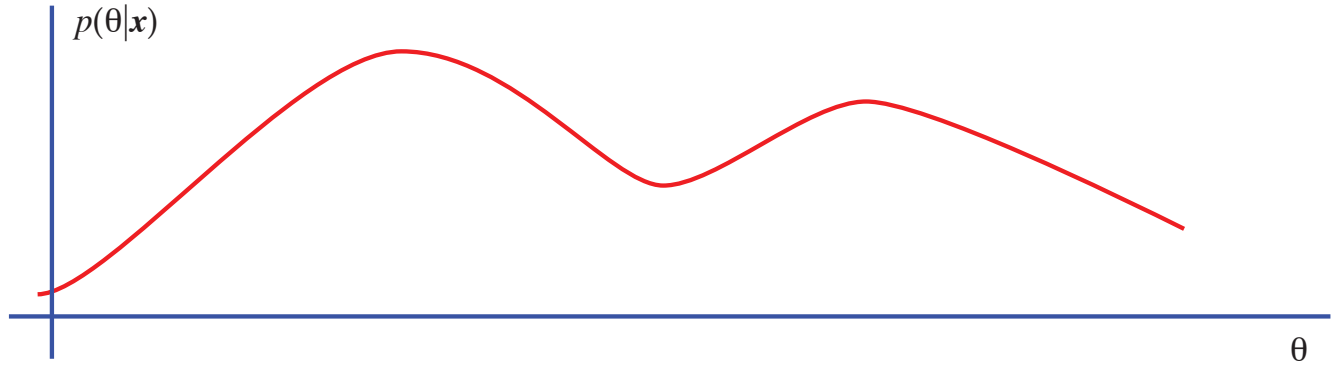
which is the posterior probability that $\theta$ deviates from $\widehat{\theta}(y)$ by more then $\epsilon$. To minimize this uniform loss we must choose $\widehat{\theta}$ to be the value of $\theta$ with highest posterior probability. Taking the limit as $\epsilon \longrightarrow 0$ gives:

The optimal estimator $\widehat{\theta}$ under uniform loss is the posterior mode.

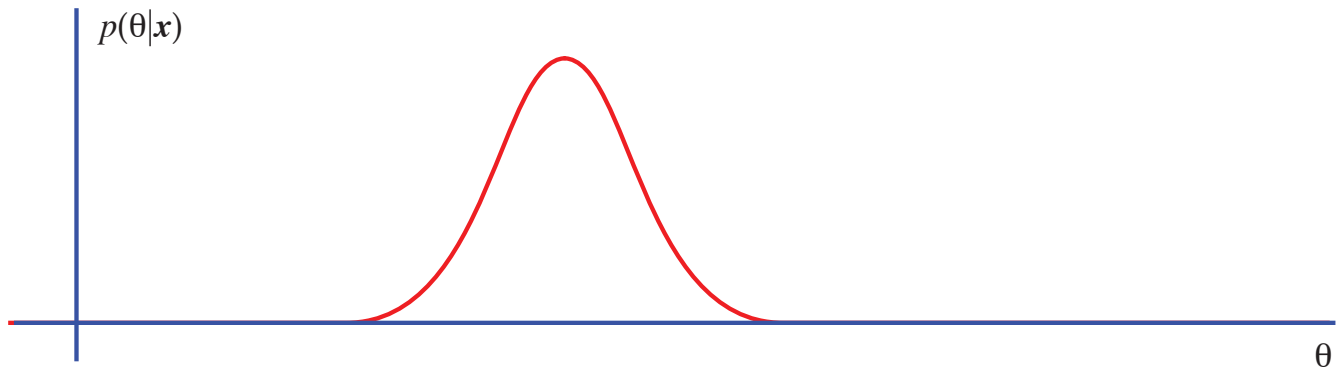
## Definition

*Maximum A Posteriori (MAP)* estimator - the value of $\theta$ where $p(\theta|y)$ is maximized:

$$\widehat{\theta}_{\mathrm{MAP}}(y) = \arg\max_{\widetilde{\theta}} p(\widetilde{\theta}|y) = \arg\max_{\widetilde{\theta}} p(y|\widetilde{\theta})p(\widetilde{\theta})$$




(a) General posterior PDF




(b) Gaussian posterior PDF

If the posterior is symmetric and unimodal, then

$$\widehat{\theta}_{\mathrm{MMSE}} = \widehat{\theta}_{\mathrm{MMAE}} = \widehat{\theta}_{\mathrm{MAP}}$$

# Computation

Both $\widehat{\theta}_{\mathrm{PM}}$ and $\widehat{\theta}_{\mathrm{MMAE}}$ require integrating with respect to $p(\theta|y)$. Often this calculation will be intractable. How can we approximate these estimators numerically?
One common approach: if we can simulate $\theta_1, \ldots, \theta_M$ from $p(\theta|y)$, then we can apply the following Monte Carlo estimates:

$$\begin{aligned} \widehat{\theta}_{\mathrm{PM}}(y) &\approx \frac{1}{M}\sum_{i=1}^{M}\theta_i \\ \widehat{\theta}_{\mathrm{MMAE}}(y) &\approx \operatorname{median}\{\theta_1, \ldots, \theta_{\mathrm{M}}\} \end{aligned}$$

If the posterior mode cannot be determined analytically, then numerical approaches can be applied.
Which of the three loss functions is used is often dictated by computational considerations.

# Choosing a Prior

Two approaches:

1. Informative (or "subjective") priors:
   - design/choose priors that are compatible with prior knowledge of unknown parameters
   - can be impractical in complicated problems with many parameters
   - injecting subjective opinion into analysis contrary to making scientific analysis as objective as possible.
2. Non-informative priors:
   - attempt to remove subjectiveness from Bayesian procedures
   - designs are often based on invariance arguments

# Selecting an Informative Prior

Clearly, the most important objective is to choose the prior $p(\theta)$ that best reflects the prior knowledge available to us.

In general, however, our prior knowledge is imprecise and any number of prior densities may aptly capture this information.

Moreover, usually the optimal estimator can't be obtained in closed-form.

Therefore, sometimes it is desirable to choose a prior density that models prior knowledge and is nicely matched in functional form to $p(y|\theta)$ so that the optimal estimator (and posterior density) can be expressed in a simple fashion.

# Conjugate Priors

Idea: Given $p(y|\theta)$, choose $p(\theta)$ so that $p(\theta|y) \propto p(y|\theta)\, p(\theta)$ has a simple functional form.

Conjugate priors: choose $p(\theta) \in \mathcal{P}$, where $\mathcal{P}$ is a family of densities (e.g., Gaussian family) so that the posterior density also belongs to that family.

## Definition

$p(\theta)$ is a *conjugate prior* for $p(y|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

## Example: Conjugate priors for exponential random variables

$$y_1, \ldots, y_n \overset{\text{iid}}{\sim} \text{exponential}(\theta)$$

$$p(y|\theta) = \prod_{i=1}^{n} \theta e^{-\theta y_i} = \theta^n e^{-\theta t}$$

where $t := \sum y_i$.

## Example: (cont.)

Let $\theta \sim \text{Gamma}(\alpha, \beta)$, so that $p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$ for $\theta \in [0, \infty)$. Then

$$
\begin{aligned}
p(y, \theta) &= \theta^n e^{-\theta t} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \\
p(y) &= \int p(y, \theta) d\theta \; = \; \int_0^\infty \theta^n e^{-\theta t} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^{N+\alpha-1} e^{-\theta(\beta+t)} d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(N+\alpha)}{(\beta+t)^{N+\alpha}} \\
p(\theta|y) &= \frac{p(y, \theta)}{p(y)} \\
&= \frac{(\beta+t)^{N+\alpha}}{\Gamma(N+\alpha)} \theta^{N+\alpha-1} e^{-\theta(\beta+t)} \\
&= \text{Gamma}(N+\alpha, \beta+t)
\end{aligned}
$$

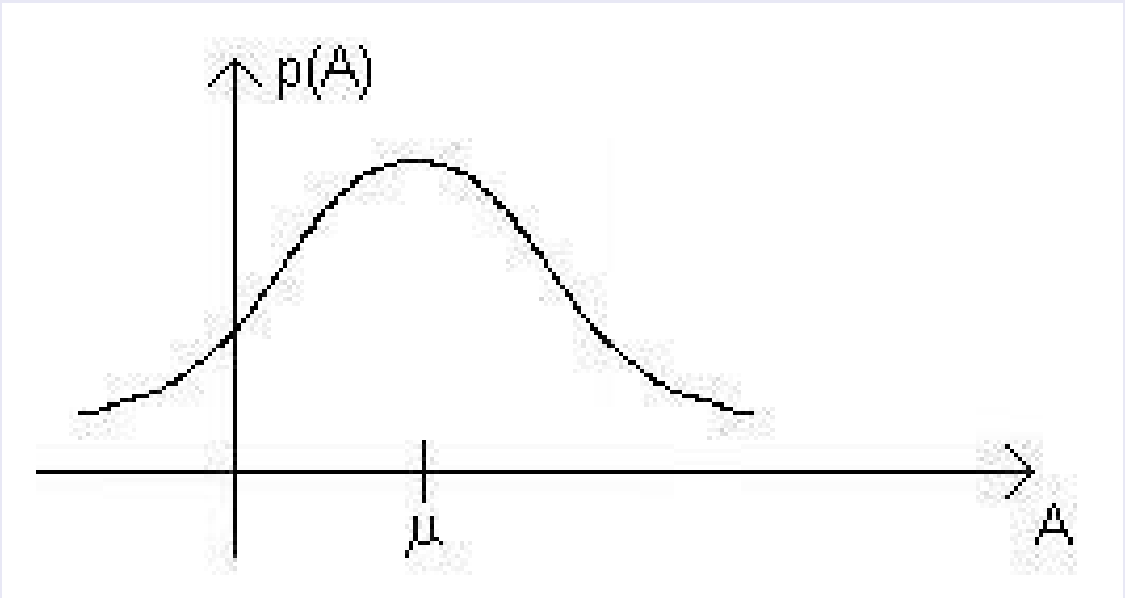
Thus the Gamma prior is conjugate for the exponential distribution!

## Example: Constant in AWGN

$$y_i = A + \nu_i \ , \ i = 1, \cdots, n; \qquad \nu_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Rather than modeling $A \sim \text{Uniform}(-A_0, A_0)$ (which did not yield a closed-form estimator) consider

$$p(A) = \frac{1}{\sqrt{2\pi {\sigma_A}^2}} \exp\left\{ -\frac{1}{2{\sigma_A}^2}(A - \mu)^2 \right\}$$



## Example: (cont.)

With $\mu = 0$ and $\sigma_A = \frac{1}{3}A_0$ this Gaussian prior also reflects prior knowledge that it is unlikely for $|A| \geq A_0$.
The Gaussian prior is also conjugate to the Gaussian likelihood

$$p(y|A) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - A)^2\right]$$

so that the resulting posterior density is also a simple Gaussian, as shown next. First note that

$$p(y|A) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2\right] \cdot \exp\left[-\frac{1}{2\sigma^2}\left(nA^2 - 2nA\bar{y}\right)\right]$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

## Example: (cont.)

$$
\begin{aligned}
p(A|y) &= \frac{p(y|A)\, p(A)}{\int p(y|a)\, p(a)\, da} \\
&= \frac{\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(nA^2 - 2nA\bar{y}) + \frac{1}{{\sigma_A}^2}(A-\mu)^2\right)\right]}{\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(na^2 - 2na\bar{y}) + \frac{1}{{\sigma_A}^2}(a-\mu)^2\right)\right] da} \\
&= \frac{e^{-\frac{1}{2}Q(A)}}{\int_{-\infty}^{\infty} e^{-\frac{1}{2}Q(a)} da}
\end{aligned}
$$

where

$$
Q(A) = \frac{n}{\sigma^2}A^2 - \frac{2nA\bar{y}}{\sigma^2} + \frac{A^2}{{\sigma_A}^2} - \frac{2\mu A}{{\sigma_A}^2} + \frac{\mu^2}{{\sigma_A}^2}
$$

Now let

$$
\begin{aligned}
\sigma^2_{A|y} &:= \frac{1}{\frac{n}{\sigma^2} + \frac{1}{{\sigma_A}^2}} \\
\mu^2_{A|y} &:= \left(\frac{n}{\sigma^2}\bar{y} + \frac{\mu}{{\sigma_A}^2}\right)\sigma^2_{A|y}
\end{aligned}
$$

## Example: (cont.)

Then by "completing the square" we have

$$\begin{aligned} Q(A) &= \frac{1}{\sigma^2_{A|y}}\left(A^2 - 2\mu_{A|y}A + \mu^2_{A|y}\right) - \frac{\mu^2_{A|y}}{\sigma^2_{A|y}} + \frac{\mu^2}{{\sigma_A}^2} \\ &= \frac{1}{\sigma^2_{A|y}}\left(A^2 - \mu_{A|y}\right)^2 - \frac{\mu^2_{A|y}}{\sigma^2_{A|y}} + \frac{\mu^2}{{\sigma_A}^2} \end{aligned}$$

Hence

$$p(A|y) = \frac{\exp\left[-\frac{1}{2\sigma^2_{A|y}}(A-\mu_{A|y})^2\right]\exp\left[-\frac{1}{2}\left(\frac{\mu^2}{{\sigma_A}^2} - \frac{\mu^2_{A|y}}{\sigma^2_{A|y}}\right)\right]}{\int_{-\infty}^{\infty} \underbrace{\exp\left[-\frac{1}{2\sigma^2_{A|y}}(a-\mu_{A|y})^2\right]}_{\text{"unnormalized" Gaussian density}} \underbrace{\exp\left[-\frac{1}{2}\left(\frac{\mu^2}{{\sigma_A}^2} - \frac{\mu^2_{A|y}}{\sigma^2_{A|y}}\right)\right]}_{\text{Constant, indep. of } a} da}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2_{A|y}}}\exp\left[-\frac{1}{2\sigma^2_{A|y}}(A-\mu_{A|y})^2\right]$$

$$A|y \sim \mathcal{N}(\mu_{A|y}, \sigma^2_{A|y})$$

## Example: (cont.)

Now

$$\begin{aligned}\widehat{A} &= \mathbb{E}[A|y] = \mu_{A|y} = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{\mu}{{\sigma_A}^2}}{\frac{n}{\sigma^2} + \frac{1}{{\sigma_A}^2}} \\ &= \left(\frac{{\sigma_A}^2}{{\sigma_A}^2 + \sigma^2/n}\right)\bar{y} + \left(\frac{\sigma^2/n}{{\sigma_A}^2 + \sigma^2/n}\right)\mu \\ &= \alpha\bar{y} + (1-\alpha)\mu\end{aligned}$$

where

$$0 < \alpha = \frac{{\sigma_A}^2}{{\sigma_A}^2 + \sigma^2/n} < 1$$

**Interpretation**

1. When there is little data ($\sigma_A^2 \ll \frac{\sigma^2}{n}$), $\alpha$ is small and $\widehat{A} \approx \mu$.

2. When there is a lot of data ($\sigma_A^2 \gg \frac{\sigma^2}{n}$), $\alpha \approx 1$ and $\widehat{A} \approx \bar{y}$.

Interplay between data and prior knowledge
small $n \longrightarrow \widehat{A}$ favors prior, large $n \longrightarrow \widehat{A}$ favors data