Bayesian inference

Arash Ali Amini

EE126, UC Berkeley

February 26, 2011

Bayesian inference

- Bayes rule can in general be used to do inference:

Bayesian inference

- Bayes rule can in general be used to do inference: You
  - have "*prior*" knowledge on an event A or a random variable X,

Bayesian inference

- Bayes rule can in general be used to do inference: You
  - have "*prior*" knowledge on an event A or a random variable X,
  - observe something: an event B or another random variable Y,

Bayesian inference

- Bayes rule can in general be used to do inference: You
  - have "*prior*" knowledge on an event A or a random variable X,
  - observe something: an event B or another random variable Y,
  - update your knowledge on A or X, based on B or Y, which is then called "*posterior*".

Bayesian inference

- Bayes rule can in general be used to do inference: You
  - have "*prior*" knowledge on an event A or a random variable X,
  - observe something: an event B or another random variable Y,
  - update your knowledge on A or X, based on B or Y, which is then called "*posterior*".
- For this to work, A (or X) and B (or Y) should be somehow related.

Bayesian inference

- Bayes rule can in general be used to do inference: You
  - have "*prior*" knowledge on an event A or a random variable X,
  - observe something: an event B or another random variable Y,
  - update your knowledge on A or X, based on B or Y, which is then called "*posterior*".
- For this to work, A (or X) and B (or Y) should be somehow related.
- Have already seen it for events

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

Bayesian inference

- Bayes rule can in general be used to do inference: You
  - have "*prior*" knowledge on an event A or a random variable X,
  - observe something: an event B or another random variable Y,
  - update your knowledge on A or X, based on B or Y, which is then called "*posterior*".
- For this to work, A (or X) and B (or Y) should be somehow related.
- Have already seen it for events

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

Can get (three) other versions by letting

$$A = \{X \in (x, x+\delta_x)\} \quad \text{and/or} \quad B = \{Y \in (y, y+\delta_y)\}$$

for $\delta_x, \delta_y > 0$ small.

Simple example

- Usually X and Y do NOT play symmetric roles
- Y is usually a "parameter" in a "generative model" of X.

Simple example

- Usually X and Y do NOT play symmetric roles
- Y is usually a "parameter" in a "generative model" of X.
- That makes Bayes rule very useful:
- Have model for X given $Y = y$ and want to turn the table around.

Simple example

- Usually X and Y do NOT play symmetric roles
- Y is usually a “parameter” in a “generative model” of X.
- That makes Bayes rule very useful:
- Have model for X given $Y = y$ and want to turn the table around.

Consider the example:

- Light bulb produced by Company G has exponentially distributed “life time” X with rate λ.

Simple example

- Usually X and Y do NOT play symmetric roles
- Y is usually a "parameter" in a "generative model" of X.
- That makes Bayes rule very useful:
- Have model for X given $Y = y$ and want to turn the table around.

Consider the example:

  - Light bulb produced by Company G has exponentially distributed "life time" X with rate λ.
  - Turns out parameter λ is itself random, uniformly distributed on $[0, 4]$.

Simple example

- Usually X and Y do NOT play symmetric roles
- Y is usually a "parameter" in a "generative model" of X.
- That makes Bayes rule very useful:
- Have model for X given $Y = y$ and want to turn the table around.

Consider the example:

- Light bulb produced by Company G has exponentially distributed "life time" X with rate λ.
- Turns out parameter λ is itself random, uniformly distributed on $[0, 4]$.
- We test a lightbulb and it has a lifetime of x.

Simple example

- Usually X and Y do NOT play symmetric roles
- Y is usually a "parameter" in a "generative model" of X.
- That makes Bayes rule very useful:
- Have model for X given $Y = y$ and want to turn the table around.

Consider the example:

- Light bulb produced by Company G has exponentially distributed "life time" X with rate λ.
- Turns out parameter λ is itself random, uniformly distributed on $[0, 4]$.
- We test a lightbulb and it has a lifetime of x.
- What can be said about the underlying parameter λ?

From verbal description to model

Light bulb produced by Company G has exponentially distributed "life time" X with rate λ. Turns out the parameter λ is itself random, uniformly distributed on $[0, 4]$. We test a lightbulb and it has a lifetime of x. What can be said about the underlying parameter λ?

From verbal description to model

Light bulb produced by Company G has exponentially distributed "life time" X with rate λ. Turns out the parameter λ is itself random, uniformly distributed on $[0, 4]$. We test a lightbulb and it has a lifetime of x. What can be said about the underlying parameter λ?

Here is a probabilistic model for above problem:

- Let Λ be the random rate. We have

From verbal description to model

Light bulb produced by Company G has exponentially distributed "life time" X with rate λ. Turns out the parameter λ is itself random, uniformly distributed on $[0, 4]$. We test a lightbulb and it has a lifetime of x. What can be said about the underlying parameter λ?

Here is a probabilistic model for above problem:

- Let Λ be the random rate. We have

$$\begin{aligned} X \mid \Lambda = \lambda \; &\sim \; \text{Expo}(\lambda), \\ \Lambda \; &\sim \; \text{Unif}(0, 4). \end{aligned}$$

From verbal description to model

Light bulb produced by Company G has exponentially distributed "life time" X with rate λ. Turns out the parameter λ is itself random, uniformly distributed on $[0, 4]$. We test a lightbulb and it has a lifetime of x. What can be said about the underlying parameter λ?

Here is a probabilistic model for above problem:

- Let Λ be the random rate. We have

$$\begin{aligned} X \mid \Lambda = \lambda \ &\sim \ \text{Expo}(\lambda), \\ \Lambda \ &\sim \ \text{Unif}(0, 4). \end{aligned}$$

- More detailed (in terms of PDFs)

$$\begin{aligned} f_{X|\Lambda}(x|\lambda) &= \lambda e^{-\lambda x} \, 1\{x > 0\} \\ f_\Lambda(\lambda) &= \frac{1}{4} 1\{0 < \lambda < 4\}. \end{aligned}$$

From verbal description to model

Light bulb produced by Company G has exponentially distributed "life time" X with rate λ. Turns out the parameter λ is itself random, uniformly distributed on $[0, 4]$. We test a lightbulb and it has a lifetime of x. What can be said about the underlying parameter λ?

Here is a probabilistic model for above problem:

- Let Λ be the random rate. We have

$$X \mid \Lambda = \lambda \sim \text{Expo}(\lambda),$$
$$\Lambda \sim \text{Unif}(0, 4).$$

- More detailed (in terms of PDFs)

$$f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x} \, 1\{x > 0\}$$
$$f_{\Lambda}(\lambda) = \frac{1}{4} 1\{0 < \lambda < 4\}.$$

- From this we have the joint distribution of (X, Λ), i.e.,
- We have a fully-specified model for (X, Λ).

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.
- Note the $\mathbb{P}(A) \approx f_\Lambda(\lambda)\,\delta_\lambda$.

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.
- Note the $\mathbb{P}(A) \approx f_\Lambda(\lambda)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B) \approx f_X(x)\,\delta_x$.

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.
- Note the $\mathbb{P}(A) \approx f_\Lambda(\lambda)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B) \approx f_X(x)\,\delta_x$.
- Similarly $\mathbb{P}(A|B) \approx f_{\Lambda|B}(\lambda|B)\,\delta_\lambda$.

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.
- Note the $\mathbb{P}(A) \approx f_\Lambda(\lambda)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B) \approx f_X(x)\,\delta_x$.
- Similarly $\mathbb{P}(A|B) \approx f_{\Lambda|B}(\lambda|B)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B|A) \approx f_{X|A}(x|A)\,\delta_x$.

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.
- Note the $\mathbb{P}(A) \approx f_\Lambda(\lambda)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B) \approx f_X(x)\,\delta_x$.
- Similarly $\mathbb{P}(A|B) \approx f_{\Lambda|B}(\lambda|B)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B|A) \approx f_{X|A}(x|A)\,\delta_x$.

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.
- Note the $\mathbb{P}(A) \approx f_\Lambda(\lambda)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B) \approx f_X(x)\,\delta_x$.
- Similarly $\mathbb{P}(A|B) \approx f_{\Lambda|B}(\lambda|B)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B|A) \approx f_{X|A}(x|A)\,\delta_x$.

Plugging in

$$f_{\Lambda|B}(\lambda|B)\,\delta_\lambda \approx \frac{f_{X|A}(x|A)\,\delta_x\;f_\Lambda(\lambda)\,\delta_\lambda}{f_X(x)\,\delta_x}.$$

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.
- Note the $\mathbb{P}(A) \approx f_\Lambda(\lambda)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B) \approx f_X(x)\,\delta_x$.
- Similarly $\mathbb{P}(A|B) \approx f_{\Lambda|B}(\lambda|B)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B|A) \approx f_{X|A}(x|A)\,\delta_x$.

Plugging in

$$f_{\Lambda|B}(\lambda|B)\,\delta_\lambda \approx \frac{f_{X|A}(x|A)\,\delta_x \;\, f_\Lambda(\lambda)\,\delta_\lambda}{f_X(x)\,\delta_x}.$$

Letting $\delta_x, \delta_\lambda \to 0$, we have $B \to \{X = x\}$ and $A \to \{\Lambda = \lambda\}$. Hence

$$f_{\Lambda|X}(\lambda|x) = \frac{f_{X|\Lambda}(x|\lambda)\;\, f_\Lambda(\lambda)}{f_X(x)}.$$

Derivation of relevant Bayes rule

- Recall $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{\Lambda \in (\lambda, \lambda + \delta_\lambda)\}$ and $B = \{X \in (x, x + \delta_x)\}$.
- Note the $\mathbb{P}(A) \approx f_\Lambda(\lambda)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B) \approx f_X(x)\,\delta_x$.
- Similarly $\mathbb{P}(A|B) \approx f_{\Lambda|B}(\lambda|B)\,\delta_\lambda$.
- Similarly $\mathbb{P}(B|A) \approx f_{X|A}(x|A)\,\delta_x$.

Plugging in

$$f_{\Lambda|B}(\lambda|B)\,\delta_\lambda \approx \frac{f_{X|A}(x|A)\,\delta_x\;f_\Lambda(\lambda)\,\delta_\lambda}{f_X(x)\,\delta_x}.$$

Letting $\delta_x, \delta_\lambda \to 0$, we have $B \to \{X = x\}$ and $A \to \{\Lambda = \lambda\}$. Hence

$$f_{\Lambda|X}(\lambda|x) = \frac{f_{X|\Lambda}(x|\lambda)\;f_\Lambda(\lambda)}{f_X(x)}.$$

You might have guessed this from the start!

Back to our simple example

- We had

$$f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}\, 1\{x > 0\}$$
$$f_{\Lambda}(\lambda) = \frac{1}{4} 1\{0 < \lambda < 4\}.$$

Back to our simple example

- We had

$$f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}\, 1\{x > 0\}$$
$$f_\Lambda(\lambda) = \frac{1}{4} 1\{0 < \lambda < 4\}.$$

- Plugging in (assuming $x > 0$) for $\lambda \in (0, 4)$

$$f_{\Lambda|X}(\lambda|x) = \frac{f_{X|\Lambda}(x|\lambda)\; f_\Lambda(\lambda)}{f_X(x)}$$
$$= \frac{\lambda e^{-\lambda x} \frac{1}{4}}{f_X(x)}$$

Back to our simple example

- We had

$$f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}\ 1\{x > 0\}$$
$$f_\Lambda(\lambda) = \frac{1}{4}1\{0 < \lambda < 4\}.$$

- Plugging in (assuming $x > 0$) for $\lambda \in (0, 4)$

$$f_{\Lambda|X}(\lambda|x) = \frac{f_{X|\Lambda}(x|\lambda)\ f_\Lambda(\lambda)}{f_X(x)}$$
$$= \frac{\lambda e^{-\lambda x}\frac{1}{4}}{f_X(x)}$$

- We can try to find $f_X(x)$ by "total probability law"

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|\Lambda}(x|\lambda)\ f_\Lambda(\lambda) d\lambda.$$

Back to our simple example

- We had

$$f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}\ 1\{x > 0\}$$
$$f_{\Lambda}(\lambda) = \frac{1}{4} 1\{0 < \lambda < 4\}.$$

- Plugging in (assuming $x > 0$) for $\lambda \in (0, 4)$

$$f_{\Lambda|X}(\lambda|x) = \frac{f_{X|\Lambda}(x|\lambda)\ f_{\Lambda}(\lambda)}{f_X(x)} = \frac{\lambda e^{-\lambda x} \frac{1}{4}}{f_X(x)}$$

- We can try to find $f_X(x)$ by "total probability law"

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|\Lambda}(x|\lambda)\ f_{\Lambda}(\lambda) d\lambda.$$

- But this is not always needed!

- We note that we essentially have the "shape" of posterior $f_\Lambda(\cdot|x)$,

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda e^{-\lambda x}\frac{1}{4}}{f_X(x)}\, 1\{\lambda \in (0,4)\}$$
$$\propto_\lambda \;\lambda e^{-\lambda x}\, 1\{\lambda \in (0,4)\}$$

- We note that we essentially have the "shape" of posterior $f_\Lambda(\cdot|x)$,

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda e^{-\lambda x}\frac{1}{4}}{f_X(x)}\,1\{\lambda \in (0,4)\}$$
$$\propto_\lambda \; \lambda e^{-\lambda x}\,1\{\lambda \in (0,4)\}$$

- What we lack is just a "normalization constant" C, i.e.,

$$f_{\Lambda|X}(\lambda|x) = C\lambda e^{-\lambda x}\,1\{\lambda \in (0,4)\}$$

- We note that we essentially have the "shape" of posterior $f_\Lambda(\cdot|x)$,

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda e^{-\lambda x}\frac{1}{4}}{f_X(x)}\, 1\{\lambda \in (0,4)\}$$
$$\propto_\lambda \; \lambda e^{-\lambda x}\, 1\{\lambda \in (0,4)\}$$

- What we lack is just a "normalization constant" C, i.e.,

$$f_{\Lambda|X}(\lambda|x) = C\lambda e^{-\lambda x}\, 1\{\lambda \in (0,4)\}$$

- which can be obtained by integration

$$1 = f_{\Lambda|X}(\lambda|x)\, d\lambda = C\int_0^4 \lambda e^{-\lambda x} d\lambda.$$

- We note that we essentially have the "shape" of posterior $f_\Lambda(\cdot|x)$,

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda e^{-\lambda x}\frac{1}{4}}{f_X(x)}\,1\{\lambda \in (0,4)\}$$
$$\propto_\lambda \ \lambda e^{-\lambda x}\,1\{\lambda \in (0,4)\}$$

- What we lack is just a "normalization constant" C, i.e.,

$$f_{\Lambda|X}(\lambda|x) = C\lambda e^{-\lambda x}\,1\{\lambda \in (0,4)\}$$

- which can be obtained by integration

$$1 = f_{\Lambda|X}(\lambda|x)\,d\lambda = C\int_0^4 \lambda e^{-\lambda x} d\lambda.$$

- Exercise: find the value of C. (It will depend on x)

- In our lightbulb problem, computing the integral for $f_X(x)$ is essentially equivalent to computing the normalization integral.

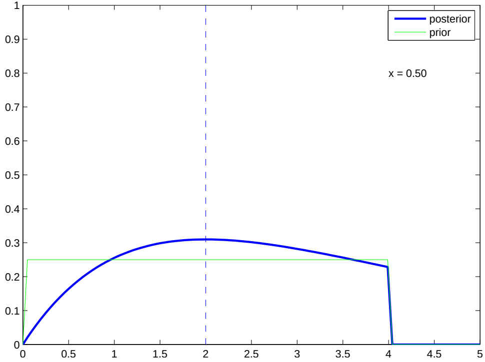
- In our lightbulb problem, computing the integral for $f_X(x)$ is essentially equivalent to computing the normalization integral.
- Sometimes, however, you recognize the posterior by its shape, and someone else has already figured out the normalization constant!

- In our lightbulb problem, computing the integral for $f_X(x)$ is essentially equivalent to computing the normalization integral.
- Sometimes, however, you recognize the posterior by its shape, and someone else has already figured out the normalization constant!
- This is particularly the case when posterior has the same shape as the prior only with updated parameters. (Catch word: conjugate priors.)

Some more observations

In our lightbulb example

- We know that for Expo(λ), the mean is $\frac{1}{\lambda}$.
- So, once we observe $X = x$, we might guess $\Lambda \approx \frac{1}{x}$.

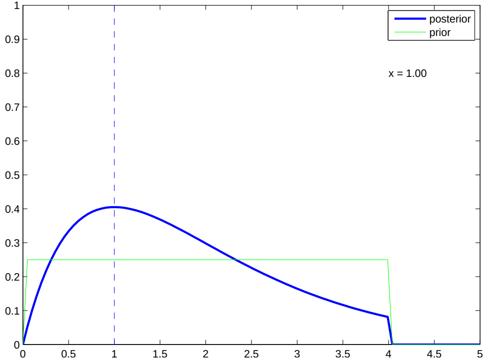


Note how the posterior concentrates roughly around $\frac{1}{x}$.

Some more observations

In our lightbulb example

- We know that for Expo(λ), the mean is $\frac{1}{\lambda}$.
- So, once we observe $X = x$, we might guess $\Lambda \approx \frac{1}{x}$.

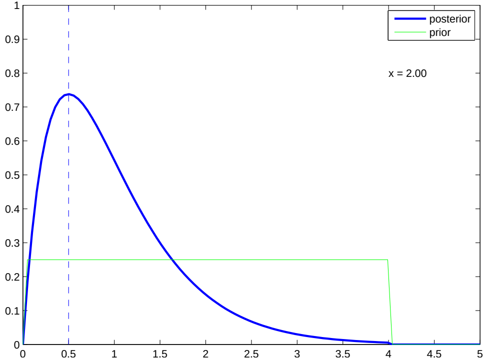


Note how the posterior concentrates roughly around $\frac{1}{x}$.

Some more observations

In our lightbulb example

- We know that for Expo(λ), the mean is $\frac{1}{\lambda}$.
- So, once we observe $X = x$, we might guess $\Lambda \approx \frac{1}{x}$.

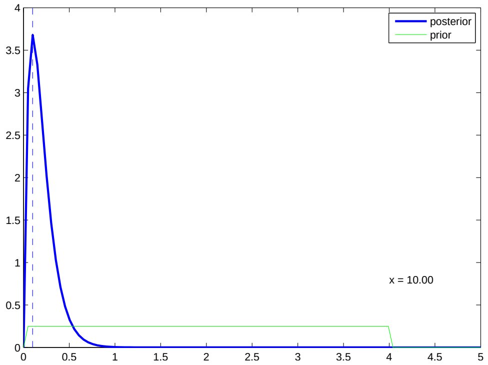


Note how the posterior concentrates roughly around $\frac{1}{x}$.

Some more observations

In our lightbulb example

- We know that for Expo(λ), the mean is $\frac{1}{\lambda}$.
- So, once we observe $X = x$, we might guess $\Lambda \approx \frac{1}{x}$.

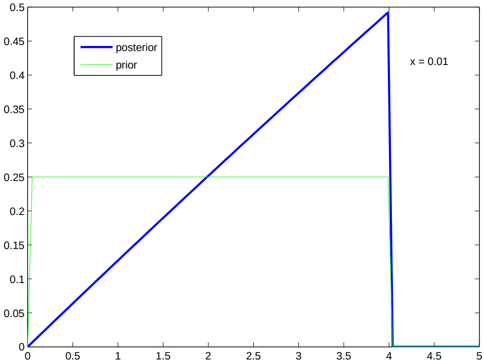


Note how the posterior concentrates roughly around $\frac{1}{x}$.

Some more observations

In our lightbulb example

- We know that for Expo(λ), the mean is $\frac{1}{\lambda}$.
- So, once we observe $X = x$, we might guess $\Lambda \approx \frac{1}{x}$.



Note how the posterior concentrates roughly around $\frac{1}{x}$.
Explain what is going on in the last two plots.

Signal detection example

- Binary signal S transmitted,

$$S = \begin{cases} +1 & \text{w.p.} \quad p \\ -1 & \text{w.p.} \quad 1-p \end{cases}$$

Signal detection example

- Binary signal S transmitted,

$$S = \begin{cases} +1 & \text{w.p.} \quad p \\ -1 & \text{w.p.} \quad 1-p \end{cases}$$

- Received signal is $Y = S + \sigma N$, where N is standard normal,

Signal detection example

- Binary signal S transmitted,

$$S = \begin{cases} +1 & \text{w.p.} \quad p \\ -1 & \text{w.p.} \quad 1-p \end{cases}$$

- Received signal is $Y = S + \sigma N$, where N is standard normal,
- N and S are independent.

Signal detection example

- Binary signal S transmitted,

$$S = \begin{cases} +1 & \text{w.p.} \quad p \\ -1 & \text{w.p.} \quad 1-p \end{cases}$$

- Received signal is $Y = S + \sigma N$, where N is standard normal,
- N and S are independent.
- Equivalently

$$Y \mid S = s \;\sim\; \mathcal{N}(s, \sigma^2)$$

Signal detection example

- Binary signal S transmitted,

$$S = \begin{cases} +1 & \text{w.p.} \quad p \\ -1 & \text{w.p.} \quad 1-p \end{cases}$$

- Received signal is $Y = S + \sigma N$, where N is standard normal,
- N and S are independent.
- Equivalently

$$Y \mid S = s \;\sim\; \mathcal{N}(s, \sigma^2)$$

- We observe value y for Y. What can be said about S?

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.
- Recall Bayes rule: $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.
- Recall Bayes rule: $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{S = +1\}$ and $B = \{Y \in (y, y + \delta_y)\}$.

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.
- Recall Bayes rule: $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{S = +1\}$ and $B = \{Y \in (y, y + \delta_y)\}$.
- Recall $\mathbb{P}(B) \approx f_Y(y)\,\delta_y$.

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.
- Recall Bayes rule: $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{S = +1\}$ and $B = \{Y \in (y, y + \delta_y)\}$.
- Recall $\mathbb{P}(B) \approx f_Y(y)\,\delta_y$.
- Similarly $\mathbb{P}(B|A) \approx f_{Y|A}(y)\,\delta_y = f_{Y|S}(y|1)\,\delta_y$.

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.
- Recall Bayes rule: $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{S = +1\}$ and $B = \{Y \in (y, y + \delta_y)\}$.
- Recall $\mathbb{P}(B) \approx f_Y(y)\,\delta_y$.
- Similarly $\mathbb{P}(B|A) \approx f_{Y|A}(y)\,\delta_y = f_{Y|S}(y|1)\,\delta_y$.

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.
- Recall Bayes rule: $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{S = +1\}$ and $B = \{Y \in (y, y + \delta_y)\}$.
- Recall $\mathbb{P}(B) \approx f_Y(y)\,\delta_y$.
- Similarly $\mathbb{P}(B|A) \approx f_{Y|A}(y)\,\delta_y = f_{Y|S}(y|1)\,\delta_y$.

Plugging in

$$\mathbb{P}(S = 1 \mid B) = \frac{f_{Y|S}(y|1)\,\delta_y\;\mathbb{P}(S = 1)}{f_Y(y)\,\delta_y}$$

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.
- Recall Bayes rule: $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{S = +1\}$ and $B = \{Y \in (y, y + \delta_y)\}$.
- Recall $\mathbb{P}(B) \approx f_Y(y)\,\delta_y$.
- Similarly $\mathbb{P}(B|A) \approx f_{Y|A}(y)\,\delta_y = f_{Y|S}(y|1)\,\delta_y$.

Plugging in

$$\mathbb{P}(S = 1 \mid B) = \frac{f_{Y|S}(y|1)\,\delta_y\;\mathbb{P}(S = 1)}{f_Y(y)\,\delta_y}$$

Letting $\delta_y \to 0$, we get $B \to \{Y = y\}$, hence

$$\mathbb{P}(S = 1 \mid Y = y) = \frac{f_{Y|S}(y|1)\;\mathbb{P}(S = 1)}{f_Y(y)}$$

Relevant Bayes formula

- Posterior distribution of S will be discrete (supported on $\{-1, +1\}$)
- It is enough to figure out $\mathbb{P}(S = +1 \mid Y = y)$.
- Recall Bayes rule: $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.
- Take $A = \{S = +1\}$ and $B = \{Y \in (y, y + \delta_y)\}$.
- Recall $\mathbb{P}(B) \approx f_Y(y)\,\delta_y$.
- Similarly $\mathbb{P}(B|A) \approx f_{Y|A}(y)\,\delta_y = f_{Y|S}(y|1)\,\delta_y$.

Plugging in

$$\mathbb{P}(S = 1 \mid B) = \frac{f_{Y|S}(y|1)\,\delta_y\ \mathbb{P}(S = 1)}{f_Y(y)\,\delta_y}$$

Letting $\delta_y \to 0$, we get $B \to \{Y = y\}$, hence

$$\mathbb{P}(S = 1 \mid Y = y) = \frac{f_{Y|S}(y|1)\ \ \mathbb{P}(S = 1)}{f_Y(y)}$$

Again not hard to guess.

Applying to signal detection

- We had $Y \mid S = 1 \sim \mathcal{N}(1, \sigma^2)$, i.e.,

$$f_{Y|S}(y|1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-1)^2/2\sigma^2}$$

Applying to signal detection

- We had $Y \mid S = 1 \sim \mathcal{N}(1, \sigma^2)$, i.e.,

$$f_{Y|S}(y|1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-1)^2/2\sigma^2}$$

- For also need $f_Y(y)$. (This actually important in this case! That is, we know the shape of posterior, it is all about the normalization.)

Applying to signal detection

- We had $Y \mid S = 1 \sim \mathcal{N}(1, \sigma^2)$, i.e.,

$$f_{Y|S}(y|1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-1)^2/2\sigma^2}$$

- For also need $f_Y(y)$. (This actually important in this case! That is, we know the shape of posterior, it is all about the normalization.)
- Use total probability

$$\begin{aligned} f_Y(y) &= \sum_{i \in \{+1,-1\}} f_{Y|S}(y|i)\, \mathbb{P}(S = i) \\ &= p\, f_{Y|S}(y|1) \;+\; (1-p)\, f_{Y|S}(y|-1) \\ &= \frac{p}{\sqrt{2\pi}\sigma} e^{-(y-1)^2/2\sigma^2} + \frac{1-p}{\sqrt{2\pi}\sigma} e^{-(y+1)^2/2\sigma^2} \end{aligned}$$

Can plug these into formula and get our result. But, we notice a possible simplification.

Can plug these into formula and get our result. But, we notice a possible simplification.

$$
\begin{aligned}
f_{Y|S}(y|i) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-i)^2/2\sigma^2} \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-(y^2-2iy+i^2)/2\sigma^2} \\
&= \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-(y^2+1)/2\sigma^2}}_{=:A_y} \; e^{iy/\sigma^2}
\end{aligned}
$$

since $i \in \{+1, 1\}$ implies $i^2 = 1$.

Can plug these into formula and get our result. But, we notice a possible simplification.

$$\begin{aligned} f_{Y|S}(y|i) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-i)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(y^2-2iy+i^2)/2\sigma^2} \\ &= \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-(y^2+1)/2\sigma^2}}_{=:A_y} \; e^{iy/\sigma^2} \end{aligned}$$

since $i \in \{+1, 1\}$ implies $i^2 = 1$. It follows that

$$f_Y(y) = A_y \Big\{ p e^{y/\sigma^2} + (1-p) e^{-y/\sigma^2} \Big\}$$

Now we plug-in

$$\begin{aligned}\mathbb{P}(S = 1 \mid Y = y) &= \frac{f_{Y|S}(y|1)\ \mathbb{P}(S = 1)}{f_Y(y)} \\ &= \frac{A_y e^{y/\sigma^2} p}{A_y \left\{ p e^{y/\sigma^2} + (1-p) e^{-y/\sigma^2} \right\}} \\ &= \frac{p e^{y/\sigma^2}}{p e^{y/\sigma^2} + (1-p) e^{-y/\sigma^2}}.\end{aligned}$$

Now we plug-in

$$\begin{aligned}\mathbb{P}(S=1 \mid Y=y) &= \frac{f_{Y|S}(y|1)\ \mathbb{P}(S=1)}{f_Y(y)} \\ &= \frac{A_y e^{y/\sigma^2} p}{A_y\left\{pe^{y/\sigma^2} + (1-p)e^{-y/\sigma^2}\right\}} \\ &= \frac{pe^{y/\sigma^2}}{pe^{y/\sigma^2} + (1-p)e^{-y/\sigma^2}}.\end{aligned}$$

or more cleanly

$$\mathbb{P}(S=1 \mid Y=y) = \frac{1}{1+\gamma e^{-2y/\sigma^2}}$$

where $\gamma = \frac{1-p}{p}$ is the odds ratio.

Now we plug-in

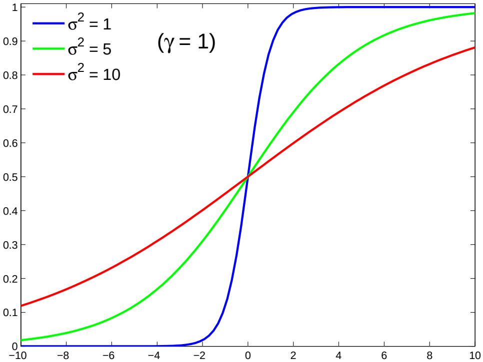
$$\begin{aligned}\mathbb{P}(S=1 \mid Y=y) &= \frac{f_{Y|S}(y|1)\ \mathbb{P}(S=1)}{f_Y(y)} \\ &= \frac{A_y e^{y/\sigma^2} p}{A_y\left\{pe^{y/\sigma^2} + (1-p)e^{-y/\sigma^2}\right\}} \\ &= \frac{pe^{y/\sigma^2}}{pe^{y/\sigma^2} + (1-p)e^{-y/\sigma^2}}.\end{aligned}$$

or more cleanly

$$\mathbb{P}(S=1 \mid Y=y) = \frac{1}{1+\gamma e^{-2y/\sigma^2}}$$

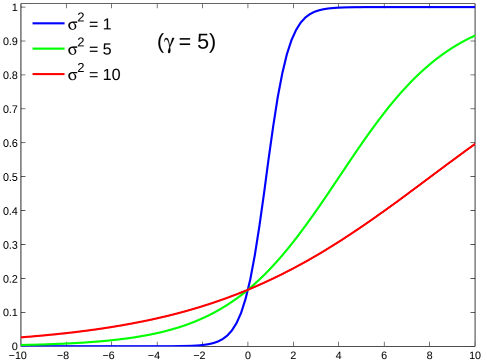
where $\gamma = \frac{1-p}{p}$ is the odds ratio. Note that this is a sigmoid as a function of y.

Here is the plot of $\mathbb{P}(S = 1 \mid Y = y)$.



Conforms with intuition.

Here is the plot of $\mathbb{P}(S = 1 \mid Y = y)$.



Conforms with intuition.

Hierarchical Bayesian Modeling

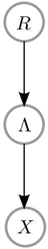
- Consider the lightbulb example. Suppose you only know that Λ is uniform in some interval $[0, r]$.
- That is, you have a prior which itself depends on a parameter "r"; this is sometimes called a *hyperparameter*.
- A true Bayesian might decide to treat the now unknown hyperparameter "r" as a random variable and put a prior on it.
- For example, we might build the hierarchical model

$$\begin{aligned} X \mid \Lambda = \lambda, R = r \; &\sim \text{Expo}(\lambda) \\ \Lambda \mid R = r \; &\sim \text{Unif}(0, r), \\ R &\sim \text{Expo}(12) \end{aligned}$$

- The prior on R seems rather arbitrary here. (There is a great deal of discussion among Bayesians on choosing effective priors.)
- The higher one goes up in the hierarchy, the lesser the effect of prior becomes on the outcome of inference.

- Note that in this new model, the conditional distribution of X given $\Lambda = \lambda$ and $R = r$, does not actually depend on r. That is, given $\Lambda = \lambda$, X and R are conditionally independent.
- This means that we could have expressed the model as

$$\begin{aligned} X \mid \Lambda = \lambda \; &\sim \text{Expo}(\lambda) \\ \Lambda \mid R = r \; &\sim \text{Unif}(0, r), \\ R &\sim \text{Expo}(12) \end{aligned}$$



together with the graphical representation to the right emphasizing the conditional independence of X and R given Λ.

- A typical question in this model is to ask for posterior (joint) distribution of (Λ, R) given $X = x$.
- Or we might only ask for (posterior) distribution of Λ given $X = x$, which can be obtained by marginalization of the joint.

Extension to multiple samples

- Back to our original lightbulb problem, we might extend it in a different direction.
- Assume that instead of just one lightbulb, we measure the lifetimes of "n" lightbulbs produced by the same process in the same company.
- This problem may be modeled as

$$X_i \mid \Lambda = \lambda \overset{i.i.d.}{\sim} \text{Expo}(\lambda), \quad i = 1, \dots, n$$
$$\Lambda \sim \text{Unif}(0, 4).$$

- We have (dropping subscripts on PDFs for brevity)

$$f(x_1, \dots, x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \lambda^n e^{-\lambda \sum_i x_i}, \quad \lambda > 0$$

- Hence, letting $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, we get

$$f(\lambda|x_1, \dots, x_n) \propto_\lambda f(x_1, \dots,_{x_n}|\lambda) f(\lambda) \propto_\lambda \lambda^n e^{-n\bar{x}\lambda}\, 1\{\lambda \in (0, 4)\}$$

(which is a truncated Gamma distribution)