Today’s content

- Quick review of probability: axioms, joint & conditional probabilities, Bayes’ Rule, conditional independence
- Bayes Nets (a.k.a. directed acyclic graphical models, DAGs)
- The Gaussian distribution
  - Example: human phoneme categorization
- Maximum likelihood estimation
- Bayesian parameter estimation
- Frequentist hypothesis testing
- Bayesian hypothesis testing
Probability spaces

Traditionally, probability spaces are defined in terms of sets. An event $E$ is a subset of a sample space $\Omega$: $E \subseteq \Omega$. 
Probability spaces

Traditionally, probability spaces are defined in terms of sets. An event $E$ is a subset of a sample space $\Omega$: $E \subseteq \Omega$.

A probability space $P$ on a sample space $\Omega$ is a function from events $E$ in $\Omega$ to real numbers such that the following three axioms hold:

1. $P(E) \geq 0$ for all $E \subset \Omega$ (non-negativity).
2. If $E_1$ and $E_2$ are disjoint, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (disjoint union).
3. $P(\Omega) = 1$ (properness).
Joint, conditional, and marginal probabilities

Given the joint distribution $P(X, Y)$ over two random variables $X$ and $Y$, the conditional distribution $P(Y|X)$ is defined as

$$P(Y|X) \equiv \frac{P(X, Y)}{P(X)}$$
Joint, conditional, and marginal probabilities

Given the joint distribution $P(X, Y)$ over two random variables $X$ and $Y$, the conditional distribution $P(Y|X)$ is defined as

$$P(Y|X) \equiv \frac{P(X, Y)}{P(X)}$$

The marginal probability distribution $P(X)$ is

$$P(X = x) = \sum_y P(X = x, Y = y)$$

These concepts can be extended to arbitrary numbers of random variables.
The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]
The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]

And this generalizes to more than two variables:
The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]

And this generalizes to more than two variables:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]
The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]

And this generalizes to more than two variables:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]
\[ P(E_1, E_2, E_3) = P(E_3|E_1, E_2)P(E_2|E_1)P(E_1) \]
The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

\[ P(E_1, E_2) = P(E_2 | E_1)P(E_1) \]

And this generalizes to more than two variables:

\[ P(E_1, E_2) = P(E_2 | E_1)P(E_1) \]
\[ P(E_1, E_2, E_3) = P(E_3 | E_1, E_2)P(E_2 | E_1)P(E_1) \]
\[ \vdots \]
The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]

And this generalizes to more than two variables:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]
\[ P(E_1, E_2, E_3) = P(E_3|E_1, E_2)P(E_2|E_1)P(E_1) \]
\[ \vdots \]
The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

\[
P(E_1, E_2) = P(E_2|E_1)P(E_1)
\]

And this generalizes to more than two variables:

\[
P(E_1, E_2) = P(E_2|E_1)P(E_1)
\]
\[
P(E_1, E_2, E_3) = P(E_3|E_1, E_2)P(E_2|E_1)P(E_1)
\]
\[\vdots\]
\[
P(E_1, E_2, \ldots, E_n) = P(E_n|E_1, E_2, \ldots, E_{n-1}) \ldots P(E_2|E_1)P(E_1)
\]
The chain rule

A joint probability can be rewritten as the product of marginal and conditional probabilities:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]

And this generalizes to more than two variables:

\[ P(E_1, E_2) = P(E_2|E_1)P(E_1) \]
\[ P(E_1, E_2, E_3) = P(E_3|E_1, E_2)P(E_2|E_1)P(E_1) \]
\[ \vdots \]
\[ P(E_1, E_2, \ldots, E_n) = P(E_n|E_1, E_2, \ldots, E_{n-1}) \ldots P(E_2|E_1)P(E_1) \]

Breaking a joint probability down into the product a marginal probability and several joint probabilities this way is called **chain rule decomposition**.
Bayes’ Rule (Bayes’ Theorem)

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]
Bayes’ Rule (Bayes’ Theorem)

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

With extra "background" random variables \( I \):

\[ P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \]
Bayes’ Rule (Bayes’ Theorem)

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

With extra "background" random variables \( I \):

\[ P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \]

This “theorem” follows directly from def’n of conditional probability:

\[ P(A, B) = P(B|A)P(A) \]
Bayes’ Rule (Bayes’ Theorem)

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

With extra "background" random variables \( I \):

\[ P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \]

This “theorem” follows directly from def’n of conditional probability:

\[ P(A, B) = P(B|A)P(A) \]
\[ P(A, B) = P(A|B)P(B) \]
Bayes’ Rule (Bayes’ Theorem)

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

With extra "background" random variables \( I \):

\[ P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \]

This “theorem” follows directly from def’n of conditional probability:

\[ P(A, B) = P(B|A)P(A) \]
\[ P(A, B) = P(A|B)P(B) \]

So

\[ P(A|B)P(B) = P(B|A)P(A) \]
Bayes’ Rule (Bayes’ Theorem)

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

With extra "background" random variables \( I \):

\[ P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \]

This “theorem” follows directly from def’n of conditional probability:

\[ P(A, B) = P(B|A)P(A) \]
\[ P(A, B) = P(A|B)P(B) \]

So

\[ \frac{P(A|B)P(B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \]
Bayes’ Rule (Bayes’ Theorem)

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

With extra "background" random variables \( I \):

\[ P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \]

This “theorem” follows directly from def’n of conditional probability:

\[ P(A, B) = P(B|A)P(A) \]
\[ P(A, B) = P(A|B)P(B) \]

So

\[ \frac{P(A|B)P(B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \]
Other ways of writing Bayes’ Rule

$$P(A|B) = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Normalizing constant}}$$

- The hardest part of using Bayes’ Rule was calculating the normalizing constant (a.k.a. the \textit{partition function})
Other ways of writing Bayes’ Rule

\[ P(A|B) = \frac{\text{Likelihood Prior}}{\text{Normalizing constant}} \]

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

- The hardest part of using Bayes’ Rule was calculating the normalizing constant (a.k.a. the partition function)
- Hence there are often two other ways we write Bayes’ Rule:
Other ways of writing Bayes’ Rule

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

- The hardest part of using Bayes’ Rule was calculating the normalizing constant (a.k.a. the partition function)
- Hence there are often two other ways we write Bayes’ Rule:
  1. Emphasizing explicit marginalization:

\[ P(A|B) = \frac{P(B|A)P(A)}{\sum_a P(A = a, B)} \]
Other ways of writing Bayes’ Rule

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

- The hardest part of using Bayes’ Rule was calculating the normalizing constant (a.k.a. the \textit{partition function})
- Hence there are often two other ways we write Bayes’ Rule:
  1. Emphasizing explicit marginalization:
     \[ P(A|B) = \frac{\sum_a P(A = a, B)}{\sum_a P(A = a, B)} \]
  2. Ignoring the partition function:
     \[ P(A|B) \propto P(B|A)P(A) \]
(Conditional) Independence

Events $A$ and $B$ are said to be Conditionally Independent given information $C$ if

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

Conditional independence of $A$ and $B$ given $C$ is often expressed as

$$A \perp B|C$$
Directed graphical models

- A lot of the interesting joint probability distributions in the study of language involve *conditional independencies* among the variables.
Directed graphical models

- A lot of the interesting joint probability distributions in the study of language involve *conditional independencies* among the variables
- So next we’ll introduce you to a general framework for specifying conditional independencies among collections of random variables
Directed graphical models

- A lot of the interesting joint probability distributions in the study of language involve *conditional independencies* among the variables.
- So next we’ll introduce you to a general framework for specifying conditional independencies among collections of random variables.
- It won’t allow us to express *all possible* independencies that may hold, but it goes a long way.
Directed graphical models

- A lot of the interesting joint probability distributions in the study of language involve *conditional independencies* among the variables.
- So next we’ll introduce you to a general framework for specifying conditional independencies among collections of random variables.
- It won’t allow us to express *all possible* independencies that may hold, but it goes a long way.
- And I hope that you’ll agree that the framework is intuitive too!
A non-linguistic example

- Imagine a factory that produces three types of coins in equal volumes:
A non-linguistic example

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
A non-linguistic example

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
A non-linguistic example

Imagine a factory that produces three types of coins in equal volumes:
- Fair coins;
- 2-headed coins;
- 2-tailed coins.
A non-linguistic example

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.
- Generative process:
A non-linguistic example

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.
- Generative process:
  - The factory produces a coin of type $X$ and sends it to you;
A non-linguistic example

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.
- Generative process:
  - The factory produces a coin of type $X$ and sends it to you;
  - You receive the coin and flip it twice, with H(eads)/T(ails) outcomes $Y_1$ and $Y_2$
A non-linguistic example

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.

- Generative process:
  - The factory produces a coin of type $X$ and sends it to you;
  - You receive the coin and flip it twice, with H(eads)/T(ails) outcomes $Y_1$ and $Y_2$.

- Receiving a coin from the factory and flipping it twice is sampling (or taking a sample) from the joint distribution $P(X, Y_1, Y_2)$.
This generative process a Bayes Net

The directed acyclic graphical model (DAG), or Bayes net:

Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable given only its parents
This generative process a Bayes Net

The directed acyclic graphical model (DAG), or Bayes net:

Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable \textbf{given only its parents}

In this DAG, \( P(X, Y_1, Y_2) = P(X)P(Y_1|X)P(Y_2|X) \)
This generative process a Bayes Net

The directed acyclic graphical model (DAG), or Bayes net:

Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable \textbf{given only its parents}.

In this DAG, \( P(X, Y_1, Y_2) = P(X)P(Y_1|X)P(Y_2|X) \)

\begin{tabular}{l|l}
X & \(P(X)\) \\
Fair & \(\frac{1}{3}\) \\
2-H & \(\frac{1}{3}\) \\
2-T & \(\frac{1}{3}\) \\
\end{tabular}
This generative process a Bayes Net

The directed acyclic graphical model (DAG), or Bayes net:

Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable **given only its parents**

In this DAG, $P(X, Y_1, Y_2) = P(X)P(Y_1|X)P(Y_2|X)$

|   | $P(X)$ |   | $P(Y_1 = H|X)$ | $P(Y_1 = T|X)$ |
|---|--------|---|----------------|----------------|
| Fair | $\frac{1}{3}$ | Fair | $\frac{1}{2}$ | $\frac{1}{2}$ |
| 2-H  | $\frac{1}{3}$ | 2-H  | 1              | 0              |
| 2-T  | $\frac{1}{3}$ | 2-T  | 0              | 1              |
This generative process a Bayes Net

The directed acyclic graphical model (DAG), or Bayes net:

Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable \textbf{given only its parents}

In this DAG, \( P(X, Y_1, Y_2) = P(X)P(Y_1|X)P(Y_2|X) \)
Conditional independence in Bayes nets

| X    | P(X) | X    | P(Y_1 = H | X) | P(Y_1 = T | X) | X    | P(Y_2 = H | X) | P(Y_2 = T | X) |
|------|------|------|--------|--------|--------|------|--------|--------|
| Fair | 1/3  | Fair | 1/2    | 1/2    | Fair  | 1/2    | 1/2    |
| 2-H  | 1/3  | 2-H  | 1      | 0      | 2-H   | 1      | 0      |
| 2-T  | 1/3  | 2-T  | 0      | 1      | 2-T   | 0      | 1      |

Question:

- **Conditioned on not having any further information, are the two coin flips Y_1 and Y_2 in this generative process independent?**
Conditional independence in Bayes nets

<table>
<thead>
<tr>
<th></th>
<th>(X)</th>
<th>(P(X))</th>
<th>(X)</th>
<th>(P(Y_1 = H \mid X))</th>
<th>(P(Y_1 = T \mid X))</th>
<th>(X)</th>
<th>(P(Y_2 = H \mid X))</th>
<th>(P(Y_2 = T \mid X))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fair</td>
<td>(\frac{1}{3})</td>
<td>Fair</td>
<td>(\frac{1}{2})</td>
<td>(\frac{1}{2})</td>
<td>Fair</td>
<td>(\frac{1}{2})</td>
<td>(\frac{1}{2})</td>
</tr>
<tr>
<td></td>
<td>2-H</td>
<td>(\frac{1}{3})</td>
<td>2-H</td>
<td>1</td>
<td>0</td>
<td>2-H</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2-T</td>
<td>(\frac{1}{3})</td>
<td>2-T</td>
<td>0</td>
<td>1</td>
<td>2-T</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Question:**

- *Conditioned on not having any further information, are the two coin flips \(Y_1\) and \(Y_2\) in this generative process independent?*

- That is, if \(C = \emptyset\), is it the case that \(A \perp B \mid C\)?
Conditional independence in Bayes nets

| X  | P(X) | X  | P(Y_1 = H|X) | P(Y_1 = T|X) | X  | P(Y_2 = H|X) | P(Y_2 = T|X) |
|----|------|----|-------------|-------------|----|-------------|-------------|
| Fair | \(\frac{1}{3}\) | Fair | \(\frac{1}{2}\) | \(\frac{1}{2}\) | Fair | \(\frac{1}{2}\) | \(\frac{1}{2}\) |
| 2-H | \(\frac{1}{3}\) | 2-H | 1 | 0 | 2-H | 1 | 0 |
| 2-T | \(\frac{1}{3}\) | 2-T | 0 | 1 | 2-T | 0 | 1 |

Question:

- *Conditioned on not having any further information, are the two coin flips \(Y_1\) and \(Y_2\) in this generative process independent?*
- That is, if \(C = \{\}\), is it the case that \(A \perp B|C\)?
- *No!*
Conditional independence in Bayes nets

| X   | P(X) | X   | P(Y_1 = H|X) | P(Y_1 = T|X) | X   | P(Y_2 = H|X) | P(Y_2 = T|X) |
|-----|------|-----|-------------|-------------|-----|-------------|-------------|
| Fair| \(\frac{1}{3}\) | Fair| \(\frac{1}{2}\) | \(\frac{1}{2}\) | Fair| \(\frac{1}{2}\) | \(\frac{1}{2}\) |
| 2-H | \(\frac{1}{3}\) | 2-H | 1            | 0           | 2-H | 1            | 0           |
| 2-T | \(\frac{1}{3}\) | 2-T | 0            | 1           | 2-T | 0            | 1           |

Question:

- **Conditioned on not having any further information, are the two coin flips \(Y_1\) and \(Y_2\) in this generative process independent?**
- That is, if \(C = \{\}\), is it the case that \(A \perp B|C\)?
- **No!**
  - \(P(Y_2 = H) = \frac{1}{2}\) (you can see this by symmetry)
**Conditional independence in Bayes nets**

|       | $P(X)$ | $X$ | $P(Y_1 = H|X)$ | $P(Y_1 = T|X)$ | $X$ | $P(Y_2 = H|X)$ | $P(Y_2 = T|X)$ |
|-------|--------|-----|---------------|---------------|-----|---------------|---------------|
| Fair  | $\frac{1}{3}$ | Fair | $\frac{1}{2}$ | $\frac{1}{2}$ | Fair | $\frac{1}{2}$ | $\frac{1}{2}$ |
| 2-H   | $\frac{1}{3}$ | 2-H | 1             | 0             | 2-H | 1             | 0             |
| 2-T   | $\frac{1}{3}$ | 2-T | 0             | 1             | 2-T | 0             | 1             |

**Question:**

▶ **Conditioned on not having any further information, are the two coin flips $Y_1$ and $Y_2$ in this generative process independent?**

▶ That is, if $C = \emptyset$, is it the case that $A \perp B|C$?

▶ **No!**

▶ $P(Y_2 = H) = \frac{1}{2}$ (you can see this by symmetry)

▶ But $P(Y_2 = H|Y_1 = H) = \frac{1}{3} \times \frac{1}{2} + \frac{2}{3} \times 1 = \frac{5}{6}$
Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.
Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.
- A path between two disjoint node sets $A$ and $B$ is a sequence of edges connecting some node in $A$ with some node in $B$. 
The comprehensive criterion for assessing conditional independence is known as D-separation.

A path between two disjoint node sets $A$ and $B$ is a sequence of edges connecting some node in $A$ with some node in $B$.

Any node on a given path has converging arrows if two edges on the path connect to it and point to it.
Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.
- A path between two disjoint node sets $A$ and $B$ is a sequence of edges connecting some node in $A$ with some node in $B$.
- Any node on a given path has converging arrows if two edges on the path connect to it and point to it.
- A node on the path has non-converging arrows if two edges on the path connect to it, but at least one does not point to it.
Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.
- A path between two disjoint node sets $A$ and $B$ is a sequence of edges connecting some node in $A$ with some node in $B$.
- Any node on a given path has converging arrows if two edges on the path connect to it and point to it.
- A node on the path has non-converging arrows if two edges on the path connect to it, but at least one does not point to it.
- A third disjoint node set $C$ d-separates $A$ and $B$ if for every path between $A$ and $B$, either:
  1. there is some node on the path with converging arrows which is not in $C$; or
  2. there is some node on the path whose arrows do not converge and which is in $C$. 
Major types of d-separation

A node set $C$ d-separates $A$ and $B$ if for every path between $A$ and $B$, either:

1. there is some node on the path with converging arrows which is not in $C$; or
2. there is some node on the path whose arrows do not converge and which is in $C$.

Common-cause d-separation (from knowing $Z$)

Intervening d-separation (from knowing $Y$)

Explaining away: knowing $Z$ prevents d-separation

D-separation in the absence of knowledge of $Z$
Back to our example
Back to our example

Without looking at the coin before flipping it, the outcome $Y_1$ of the first flip gives me information about the type of coin, and affects my beliefs about the outcome of $Y_2$. 

$X$ 

$Y_1$ 

$Y_2$
Back to our example

- *Without looking at the coin before flipping it*, the outcome $Y_1$ of the first flip gives me information about the type of coin, and affects my beliefs about the outcome of $Y_2$.

- But if I *look* at the coin before flipping it, $Y_1$ and $Y_2$ are rendered independent.
An example of explaining away

I saw an exhibition about the, uh...
An example of explaining away

I saw an exhibition about the, uh...

There are several causes of disfluency, including:
An example of explaining away

*I saw an exhibition about the, uh...*

There are several causes of disfluency, including:

- An upcoming word is difficult to produce (e.g., low frequency, *astrolabe*)
An example of explaining away

*I saw an exhibition about the, uh…*

There are several causes of disfluency, including:

- An upcoming word is difficult to produce (e.g., low frequency, *astrolabe*)
- The speaker’s attention was distracted by something in the non-linguistic environment
An example of explaining away

*I saw an exhibition about the, uh...*

There are several causes of disfluency, including:

- An upcoming word is difficult to produce (e.g., low frequency, *astrolabe*)
- The speaker’s attention was distracted by something in the non-linguistic environment
An example of explaining away

*I saw an exhibition about the, uh...*

There are several causes of disfluency, including:

- An upcoming word is difficult to produce (e.g., low frequency, *astrolabe*)
- The speaker’s attention was distracted by something in the non-linguistic environment

A reasonable graphical model:
An example of explaining away

- Without knowledge of $D$, there’s no reason to expect that $W$ and $A$ are correlated
An example of explaining away

- Without knowledge of $D$, there's no reason to expect that $W$ and $A$ are correlated
- But hearing a disfluency demands a cause
An example of explaining away

- Without knowledge of $D$, there's no reason to expect that $W$ and $A$ are correlated.
- But hearing a disfluency demands a cause.
- Knowing that there was a distraction explains away the disfluency, reducing the probability that the speaker was planning to utter a hard word.
An example of the disfluency model

- Let’s suppose that both hard words and distractions are unusual, the latter more so

\[
P(W = \text{hard}) = 0.25 \quad P(A = \text{distracted}) = 0.15
\]
An example of the disfluency model

Let’s suppose that both hard words and distractions are unusual, the latter more so

\[
P(W = \text{hard}) = 0.25 \\
P(A = \text{distracted}) = 0.15
\]

Hard words and distractions both induce disfluencies; having both makes a disfluency really likely

<table>
<thead>
<tr>
<th>$W$</th>
<th>$A$</th>
<th>$D=$no disfluency</th>
<th>$D=$disfluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td>undistracted</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>easy</td>
<td>distracted</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>hard</td>
<td>undistracted</td>
<td>0.85</td>
<td>0.15</td>
</tr>
<tr>
<td>hard</td>
<td>distracted</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>
An example of the disfluency model

\[ P(W = \text{hard}) = 0.25 \]
\[ P(A = \text{distracted}) = 0.15 \]

Suppose that we observe the speaker uttering a disfluency. What is \( P(W = \text{hard}|D = \text{disfluent}) \)?
An example of the disfluency model

\[ P(W = \text{hard}) = 0.25 \]
\[ P(A = \text{distracted}) = 0.15 \]

<table>
<thead>
<tr>
<th>( W )</th>
<th>( A )</th>
<th>( D = \text{no disfluency} )</th>
<th>( D = \text{disfluency} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td>undistracted</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>easy</td>
<td>distracted</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>hard</td>
<td>undistracted</td>
<td>0.85</td>
<td>0.15</td>
</tr>
<tr>
<td>hard</td>
<td>distracted</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

- Suppose that we observe the speaker uttering a disfluency. What is \( P(W = \text{hard} | D = \text{disfluent}) \)?
- Now suppose we also learn that her attention is distracted. What does that do to our beliefs about \( W \)
An example of the disfluency model

\[ P(W = \text{hard}) = 0.25 \]
\[ P(A = \text{distracted}) = 0.15 \]

<table>
<thead>
<tr>
<th>$W$</th>
<th>$A$</th>
<th>$D=\text{no disfluency}$</th>
<th>$D=\text{disfluency}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy</td>
<td>undistracted</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>easy</td>
<td>distracted</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>hard</td>
<td>undistracted</td>
<td>0.85</td>
<td>0.15</td>
</tr>
<tr>
<td>hard</td>
<td>distracted</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

- Suppose that we observe the speaker uttering a disfluency. What is $P(W = \text{hard} | D = \text{disfluent})$?
- Now suppose we also learn that her attention is distracted. What does that do to our beliefs about $W$?
- That is, what is $P(W = \text{hard} | D = \text{disfluent}, A = \text{distracted})$?
An example of the disfluency model

Fortunately, there is automated machinery to “turn the Bayesian crank”:

\[ P(W = \text{hard}) = 0.25 \]
An example of the disfluency model

Fortunately, there is automated machinery to “turn the Bayesian crank”:

\[
P(W = \text{hard}) = 0.25
\]
\[
P(W = \text{hard}|D = \text{disfluent}) = 0.57
\]
An example of the disfluency model

Fortunately, there is automated machinery to “turn the Bayesian crank”:

\[
\begin{align*}
P(W = \text{hard}) &= 0.25 \\
P(W = \text{hard}|D = \text{disfluent}) &= 0.57 \\
P(W = \text{hard}|D = \text{disfluent}, A = \text{distracted}) &= 0.40
\end{align*}
\]
An example of the disfluency model

Fortunately, there is automated machinery to “turn the Bayesian crank”:

\[
P(W = \text{hard}) = 0.25
\]
\[
P(W = \text{hard}|D = \text{disfluent}) = 0.57
\]
\[
P(W = \text{hard}|D = \text{disfluent}, A = \text{distracted}) = 0.40
\]

- Knowing that the speaker was distracted (A) decreased the probability that the speaker was about to utter a hard word (W)—A explained D away.
An example of the disfluency model

Fortunately, there is automated machinery to “turn the Bayesian crank”:

\[
P(W = \text{hard}) = 0.25 \\
P(W = \text{hard}|D = \text{disfluent}) = 0.57 \\
P(W = \text{hard}|D = \text{disfluent}, A = \text{distracted}) = 0.40
\]

- Knowing that the speaker was distracted (A) decreased the probability that the speaker was about to utter a hard word (W)—A explained D away.
- A caveat: the type of relationship among A, W, and D will depend on the values one finds in the probability table!
Summary thus far

Key points:
- Bayes’ Rule is a compelling framework for modeling inference under uncertainty
- DAGs/Bayes Nets are a broad class of models for specifying joint probability distributions with conditional independencies
- Classic Bayes Net references: ??; ?; ?, Chapter 14; ?, Chapter 8.
An example of the disfluency model

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>hard</td>
<td>W=hard</td>
<td></td>
</tr>
<tr>
<td>easy</td>
<td>W=easy</td>
<td></td>
</tr>
<tr>
<td>disfl</td>
<td>D=disfluent</td>
<td></td>
</tr>
<tr>
<td>distr</td>
<td>A=distracted</td>
<td></td>
</tr>
<tr>
<td>undistr</td>
<td>A=undistracted</td>
<td></td>
</tr>
</tbody>
</table>

\[
P(W = \text{hard} | D = \text{disfluent}, A = \text{distracted})
\]

\[
P(\text{hard} | \text{disfl}, \text{distr}) = \frac{P(\text{disfl} | \text{hard}, \text{distr})P(\text{hard} | \text{distr})}{P(\text{disfl} | \text{distr})} \quad \text{(Bayes' Rule)}
\]

\[
= \frac{P(\text{disfl} | \text{hard}, \text{distr})P(\text{hard})}{P(\text{disfl} | \text{distr})}
\]

\[
P(\text{disfl} | \text{distr}) = \sum_{W} P(\text{disfl} | W = w')P(W = w') \quad \text{(Marginalization)}
\]

\[
= P(\text{disfl} | \text{hard})P(\text{hard}) + P(\text{disfl} | \text{easy})P(\text{easy})
\]

\[
= 0.6 \times 0.25 + 0.3 \times 0.75
\]

\[
= 0.375
\]

\[
P(\text{hard} | \text{disfl}, \text{distr}) = \frac{0.6 \times 0.25}{0.375}
\]

\[
= 0.4
\]
An example of the disfluency model

\[ P(W = \text{hard}|D = \text{disfluent}) \]

\[
P(\text{hard}|\text{disfl}) = \frac{P(\text{disfl}|\text{hard})P(\text{hard})}{P(\text{disfl})} \quad \text{(Bayes’ Rule)}
\]

\[
P(\text{disfl}|\text{hard}) = \sum_{a'} P(\text{disfl}|A = a', \text{hard})P(A = a'|\text{hard})
\]

\[
= P(\text{disfl}|A = \text{distr, hard})P(A = \text{distr}|\text{hard}) + P(\text{disfl}|\text{undistr, hard})P(\text{undistr}|\text{hard})
\]

\[
= 0.6 \times 0.15 + 0.15 \times 0.85
\]

\[
= 0.2175
\]

\[
P(\text{disfl}) = \sum_{w'} P(\text{disfl}|W = w')P(W = w')
\]

\[
= P(\text{disfl}|\text{hard})P(\text{hard}) + P(\text{disfl}|\text{easy})P(\text{easy})
\]

\[
P(\text{disfl}|\text{easy}) = \sum_{a'} P(\text{disfl}|A = a', \text{easy})P(A = a'|\text{easy})
\]

\[
= P(\text{disfl}|A = \text{distr, easy})P(A = \text{distr}|\text{easy}) + P(\text{disfl}|\text{undistr, easy})P(\text{undistr}|\text{easy})
\]

\[
= 0.3 \times 0.15 + 0.01 \times 0.85
\]

\[
= 0.0535
\]

\[
P(\text{disfl}) = 0.2175 \times 0.25 + 0.0535 \times 0.75
\]

\[
= 0.0945
\]

\[
P(\text{hard}|\text{disfl}) = \frac{0.2175 \times 0.25}{0.0945}
\]

\[
= 0.575396825396825
\]
Bayesian parameter estimation

The scenario: you are a native English speaker in whose experience passivizable constructions are passivized with frequency $q$.

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

You encounter a new dialect of English and hear data $y$ consisting of $n$ passivizable utterances, $m$ of which were passivized:

\[ X \sim Bern(\pi) \]

**Goal:**

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- **Or** place a probability distribution on the number of passives in the next $N$ passivizable utterances.
Anatomy of Bayesian inference

Simplest possible scenario:

\[ I \rightarrow \theta \rightarrow Y \]
Anatomy of Bayesian inference

Simplest possible scenario:

The corresponding Bayesian inference:

\[ P(\theta|y, I) = \frac{P(y|\theta, I)P(\theta|I)}{P(y|I)} \]
Anatomy of Bayesian inference

Simplest possible scenario:

\[ I \xrightarrow{\theta} Y \]

The corresponding Bayesian inference:

\[
P(\theta|y, I) = \frac{P(y|\theta, I)P(\theta|I)}{P(y|I)}
\]

- Likelihood for \( \theta \)
- Prior over \( \theta \)

\[
= \frac{P(y|\theta)P(\theta|I)}{P(y|I)}
\]

Likelihood marginalized over \( \theta \)

(because \( y \perp I \mid \theta \))
Anatomy of Bayesian inference

Simplest possible scenario:

\[ I \rightarrow \theta \rightarrow Y \]

The corresponding Bayesian inference:

\[
P(\theta|y, I) = \frac{P(y|\theta, I)P(\theta|I)}{P(y|I)}
\]

\[ \text{Likelihood for } \theta \quad \text{Prior over } \theta \]

\[
= \frac{P(y|\theta)}{P(y|I)} \cdot \frac{P(\theta|I)}{P(y|I)}
\]

\[ \text{Likelihood marginalized over } \theta \] (because \( y \perp I \mid \theta \))

- At the “bottom” of the graph, our model is the binomial distribution:

\[
P(y|\theta) \sim Binom(n, \theta)
\]

- But to get things going we have to set the prior \( P(\theta|I) \).
Priors for the binomial distribution

For a model with parameters $\theta$, a prior distribution is just some joint probability distribution $P(\theta)$

Because the prior is often supposed to account for "knowledge we bring to the table", we often write $P(\theta|I)$ to be explicit
Priors for the binomial distribution

- For a model with parameters $\theta$, a prior distribution is just some joint probability distribution $P(\theta)$
  - Because the prior is often supposed to account for “knowledge we bring to the table”, we often write $P(\theta|I)$ to be explicit
- Model parameters are nearly always real-valued, so $P(\theta)$ is generally a multivariate continuous distribution
Priors for the binomial distribution

- For a model with parameters $\theta$, a prior distribution is just some joint probability distribution $P(\theta)$
  - Because the prior is often supposed to account for “knowledge we bring to the table”, we often write $P(\theta|I)$ to be explicit
- Model parameters are nearly always real-valued, so $P(\theta)$ is generally a multivariate continuous distribution
- In general, the sky is the limit as to what you choose for $P(\theta)$
Priors for the binomial distribution

- For a model with parameters $\theta$, a prior distribution is just some joint probability distribution $P(\theta)$
  - Because the prior is often supposed to account for “knowledge we bring to the table”, we often write $P(\theta|I)$ to be explicit
- Model parameters are nearly always real-valued, so $P(\theta)$ is generally a multivariate continuous distribution
- In general, the sky is the limit as to what you choose for $P(\theta)$
- But in many cases there are useful priors that will make your life easier
The beta distribution

The beta distribution has two parameters $\alpha_1, \alpha_2 > 0$ and is defined as:

$$P(\pi | \alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1}$$

$$(0 \leq \pi \leq 1, \alpha_1 > 0, \alpha_2 > 0)$$

where the beta function $B(\alpha_1, \alpha_2)$ serves as a normalizing constant:

$$B(\alpha_1, \alpha_2) = \int_0^1 \pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1} d\pi$$
Some beta distributions

If $X \sim B(\alpha_1, \alpha_2)$:

1. $E[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2}$

2. If $\alpha_1, \alpha_2 > 1$, then $X$ has a mode at $\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence $I = \langle \alpha_1, \alpha_2 \rangle$. 
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence \( I = \langle \alpha_1, \alpha_2 \rangle \).

\[
P(\pi \mid y, \alpha_1, \alpha_2) = \frac{P(y \mid \pi) P(\pi \mid \alpha_1, \alpha_2)}{P(y \mid \alpha_1, \alpha_2)}
\] (1)
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence \( I = \langle \alpha_1, \alpha_2 \rangle \).

\[
P(\pi | y, \alpha_1, \alpha_2) = \frac{P(y | \pi)P(\pi | \alpha_1, \alpha_2)}{P(y | \alpha_1, \alpha_2)} \quad (1)
\]

Since the denominator is not a function of \( \pi \), it is a normalizing constant. Ignore it and work in terms of proportionality:

\[
P(\pi | y, \alpha_1, \alpha_2) \propto P(y | \pi)P(\pi | \alpha_1, \alpha_2)
\]
Let us use a beta distribution as a prior for our problem—hence $I = \langle \alpha_1, \alpha_2 \rangle$.

\[
P(\pi | y, \alpha_1, \alpha_2) = \frac{P(y | \pi)P(\pi | \alpha_1, \alpha_2)}{P(y | \alpha_1, \alpha_2)}
\] (1)

Since the denominator is not a function of $\pi$, it is a normalizing constant. Ignore it and work in terms of proportionality:

\[
P(\pi | y, \alpha_1, \alpha_2) \propto P(y | \pi)P(\pi | \alpha_1, \alpha_2)
\]

Likelihood for the binomial distribution is

\[
P(y | \pi) = \binom{n}{m} \pi^m (1 - \pi)^{n-m}
\]
Using the beta distribution as a prior

1. The ball hit the window. (Active)
2. The window was hit by the ball. (Passive)

Let us use a beta distribution as a prior for our problem—hence
\[ l = \langle \alpha_1, \alpha_2 \rangle. \]

\[
P(\pi | y, \alpha_1, \alpha_2) = \frac{P(y | \pi) P(\pi | \alpha_1, \alpha_2)}{P(y | \alpha_1, \alpha_2)}
\]  

(1)

Since the denominator is not a function of \( \pi \), it is a normalizing constant. Ignore it and work in terms of proportionality:

\[ P(\pi | y, \alpha_1, \alpha_2) \propto P(y | \pi) P(\pi | \alpha_1, \alpha_2) \]

Likelihood for the binomial distribution is

\[
P(y | \pi) = \binom{n}{m} \pi^m (1 - \pi)^{n-m}
\]

Beta prior is

\[
P(\pi | \alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}
\]
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
P(\pi | y, \alpha_1, \alpha_2) \propto \pi^m (1 - \pi)^{n-m} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}
\]
Using the beta distribution as a prior

Ignore $\binom{n}{m}$ and $B(\alpha_1, \alpha_2)$ (both constant in $\pi$):

$$P(\pi|y, \alpha_1, \alpha_2) \propto \underbrace{\pi^m(1-\pi)^{n-m}}_{\text{Likelihood}} \underbrace{\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}}_{\text{Prior}}$$

$$\propto \pi^{m+\alpha_1-1}(1-\pi)^{n-m+\alpha_2-1}$$
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
P(\pi|y, \alpha_1, \alpha_2) \propto \pi^m (1 - \pi)^{n-m} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}
\]

\[
\propto \pi^{m+\alpha_1-1} (1 - \pi)^{n-m+\alpha_2-1}
\]

**Crucial trick:** this is itself a beta distribution! Recall that if \( \theta \sim Beta(\alpha_1, \alpha_2) \) then

\[
P(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}
\]
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
P(\pi | y, \alpha_1, \alpha_2) \propto \underbrace{\pi^m (1 - \pi)^{n-m}}_{\text{Likelihood}} \underbrace{\pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}}_{\text{Prior}}
\]

\[
\propto \pi^{m+\alpha_1-1} (1 - \pi)^{n-m+\alpha_2-1}
\]

Crucial trick: this is itself a beta distribution! Recall that if \( \theta \sim \text{Beta} (\alpha_1, \alpha_2) \) then

\[
P(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}
\]

Hence \( P(\theta | y, \alpha_1, \alpha_2) \) is distributed as \( \text{Beta}(\alpha_1 + m, \alpha_2 + n - m) \).
Using the beta distribution as a prior

Ignore \( \binom{n}{m} \) and \( B(\alpha_1, \alpha_2) \) (both constant in \( \pi \)):

\[
P(\pi|y, \alpha_1, \alpha_2) \propto \pi^m (1 - \pi)^{n - m} \pi^{\alpha_1 - 1} (1 - \pi)^{\alpha_2 - 1}
\]

\[
\propto \pi^{m + \alpha_1 - 1} (1 - \pi)^{n - m + \alpha_2 - 1}
\]

**Crucial trick:** this is itself a beta distribution! Recall that if \( \theta \sim Beta(\alpha_1, \alpha_2) \) then

\[
P(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1 - 1} (1 - \pi)^{\alpha_2 - 1}
\]

Hence \( P(\theta|y, \alpha_1, \alpha_2) \) is distributed as \( Beta(\alpha_1 + m, \alpha_2 + n - m) \).

- With a beta prior and a binomial likelihood, the posterior is still beta-distributed. This is called conjugacy.
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivable utterances.

To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;

- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

- To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability

- $P(\text{passive}|\text{passivizable clause}) \approx 0.08$ (Roland et al., 2007)
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- **Or** place a probability distribution on the number of passives in the next $N$ passivizable utterances.

- To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability
- $P(\text{passive}|\text{passivizable clause}) \approx 0.08$ (Roland et al., 2007)
- The mode of a beta distribution is $\frac{\alpha_1-1}{\alpha_1+\alpha_2-2}$
Using our beta-binomial model

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability

- $P(\text{passive}|\text{passivizable clause}) \approx 0.08$ (Roland et al., 2007)
- The mode of a beta distribution is $\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$
- Hence we might use $\alpha_1 = 3$, $\alpha_2 = 24$ (note that $\frac{2}{25} = 0.08$)
Using our beta-binomial model

Goal:
- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

To estimate $\pi$ it is common to use Maximum a-posteriori (MAP) estimation: choose the value of $\pi$ with highest posterior probability

- $P(\text{passive}|\text{passivizable clause}) \approx 0.08$ (Roland et al., 2007)
- The mode of a beta distribution is $\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$
- Hence we might use $\alpha_1 = 3$, $\alpha_2 = 24$ (note that $\frac{2}{25} = 0.08$)
- Suppose that $n = 7$, $m = 2$: our posterior will be $Beta(5, 29)$, hence $\hat{\pi} = \frac{4}{32} = 0.125$
Beta-binomial posterior distributions
Fully Bayesian density estimation

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.
Fully Bayesian density estimation

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivable utterances.

In the fully Bayesian view, we don’t summarize our posterior beliefs into a point estimate; rather, we marginalize over them in predicting the future:
Fully Bayesian density estimation

Goal:

- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- Or place a probability distribution on the number of passives in the next $N$ passivizable utterances.

In the fully Bayesian view, we don’t summarize our posterior beliefs into a point estimate; rather, we marginalize over them in predicting the future:

$$P(y_{new} | y, I) = \int_{\theta} P(y_{new} | \theta) P(\theta | y, I) \, d\theta$$
Fully Bayesian density estimation

**Goal:**
- Estimate the success parameter $\pi$ associated with passivization in the new English dialect;
- **Or** place a probability distribution on the number of passives in the next $N$ passivizable utterances.

In the fully Bayesian view, we don’t summarize our posterior beliefs into a point estimate; rather, we marginalize over them in predicting the future:

$$P(y_{new}|y, I) = \int_\theta P(y_{new}|\theta)P(\theta|y, I)\,d\theta$$

This leads to the beta-binomial predictive model:

$$P(r|k, I, y) = \binom{k}{r} \frac{B(\alpha_1 + m + r, \alpha_2 + n - m + k - r)}{B(\alpha_1 + m, \alpha_2 + n - m)}$$
Fully Bayesian density estimation

\[ P(k \text{ passives out of 50 trials} \mid y, I) \]

- Binomial
- Beta–Binomial
In this case (as in many others), marginalizing over the model parameters allows for greater dispersion in the model’s predictions.
Fully Bayesian density estimation

- In this case (as in many others), marginalizing over the model parameters allows for greater dispersion in the model’s predictions.
- This is because the new observations are only conditionally independent given $\theta$—with uncertainty about $\theta$, they are linked!
Fully Bayesian density estimation

- In this case (as in many others), marginalizing over the model parameters allows for greater dispersion in the model’s predictions.
- This is because the new observations are only conditionally independent given $\theta$—with uncertainty about $\theta$, they are linked!