A Brief Introduction to Bayesian Inference

Mark Johnson

CG168 notes
A brief review of *discrete* probability theory

- $\Omega$ is the set of all *elementary events* (c.f. interpretations in logic)
- If $\omega \in \Omega$, then $P(\omega)$ is the probability of event $\omega$
  - $P(\omega) \geq 0$
  - $\sum_{\omega \in \Omega} P(\omega) = 1$
- A *random variable* $X$ is a function from $\Omega$ to some set of values $\mathcal{X}$
  - If $\mathcal{X}$ is countable then $X$ is a *discrete* random variable
  - If $\mathcal{X}$ is continuous then $X$ is a *continuous* random variable
- If $x$ is a possible value for $X$, then

\[
P(X = x) = \sum_{\substack{\omega \in \Omega \\ X(\omega) = x}} P(\omega)
\]
Independence and conditional distributions

• Two RVs $X$ and $Y$ are *independent* iff $P(X, Y) = P(X)P(Y)$

• The *conditional distribution* of $Y$ given $X$ is:

\[
P(Y|X) = \frac{P(Y, X)}{P(X)}
\]

so $X$ and $Y$ are independent iff $P(Y|X) = P(Y)$ (here and below I assume strictly positive distributions)

• We can decompose the joint distribution of a sequence of RVs into a product of conditionals:

\[
P(X_1, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1) \ldots P(X_n|X_{n-1}, \ldots, X_1)
\]

i.e., the probability of generating $X_1, \ldots, X_n$ “at once” is the same as generating them one at a time if each $X_i$ is conditioned on the $X_1, \ldots, X_{i-1}$ that preceded it
Conditional distributions

- It's always possible to factor any distribution over \( X = (X_1, \ldots, X_n) \) into a product of conditionals

\[
P(X) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\]

- But in many interesting cases, \( X_i \) depends only on a subset of \( X_1, \ldots, X_{i-1} \), i.e.,

\[
P(X) = \prod_{i} P(X_i | X_{\text{Pa}(i)})
\]

where \( \text{Pa}(i) \subseteq \{1, \ldots, i - 1\} \) and \( X_S = \{X_j : j \in S\} \)

- \( X \) and \( Y \) are \textit{conditionally independent} given \( Z \) iff

\[
P(X, Y | Z) = P(X | Z) \ P(Y | Z)
\]

or equivalently,

\[
P(X | Y, Z) = P(X | Z)
\]

- Note: the "parents" \( \text{Pa}(i) \) of \( X_i \) depend on the order in which the variables are enumerated!
Bayes nets

- A Bayes net is a graphical depiction of a factorization of a probability distribution into products of conditional distributions

\[ P(X) = \prod_i P(X_i|X_{Pa(i)}) \]

- A Bayes net has a node for each variable \( X_i \) and an arc from \( X_j \) to \( X_i \) iff \( j \in Pa(i) \)
Bayes rule

- Bayes theorem:

\[ P(Y|X) = \frac{P(X|Y) \ P(Y)}{P(X)} \]

- Bayes inversion: swap direction of arcs in Bayes net

- Interpreted as a recipe for “belief updating”:

\[ P(\text{Hypothesis}|\text{Data}) \propto P(\text{Data}|\text{Hypothesis}) \ P(\text{Hypothesis}) \]

- The normalizing constant (which you have to divide Likelihood times Prior by) is:

\[ P(\text{Data}) = \sum_{\text{Hypothesis'}} P(\text{Data}|\text{Hypothesis'}) \ P(\text{Hypothesis'}) \]

which is the probability of generating the data under any hypothesis
Iterated Bayesian belief updating

• Suppose the data consists of 2 components $D = (D_1, D_2)$, and $P(H)$ is our prior over hypotheses $H$

$$P(H|D_1, D_2) \propto P(D_1, D_2|H) \cdot P(H)$$
$$\propto P(D_2|H, D_1) \cdot P(H|D_1)$$

• This means the following are equivalent:
  ▷ update the prior $P(H)$ treating $(D_1, D_2)$ as a single observation
  ▷ update the prior $P(H)$ wrt the first observation $D_1$ producing posterior $P(H|D_1) \propto P(D_1|H) \cdot P(H)$, which serves as the prior for the second observation $D_2$
Incremental Bayesian belief updating

- Consider a “two-part” data set \((d_1, d_2)\). We show posterior obtained by Bayesian belief updating on \((d_1, d_2)\) together is same as posterior obtained by updating on \(d_1\) and then updating on \(d_2\).
- Bayesian belief updating on both \((d_1, d_2)\) using prior \(P(H)\)

\[
P(H|d_1, d_2) \propto P(d_1, d_2|H)P(H) = P(d_1, d_2, H)
\]

- Incremental Bayesian belief updating
  - Bayesian belief updating on \(d_1\) using prior \(P(H)\)

\[
P(H|d_1) \propto P(d_1|H)P(H) = P(d_1, H)
\]
  - Bayesian belief updating on \(d_2\) using prior \(P(H|d_1)\)

\[
P(H|d_1, d_2) \propto P(d_2|H, d_1)P(H|d_1)
\]

\[
\propto P(d_2|H, d_1)P(H, d_1)
\]

\[
= P(d_2, d_1, H)
\]
“Distributed according to” notation

• A probability distribution $F$ is a non-negative function from some set $\mathcal{X}$ whose values sum (integrate) to 1

• A random variable $X$ is distributed according to a distribution $F$, or more simply, $X$ has distribution $F$, written $X \sim F$, iff:

$$P(X = x) = F(x) \text{ for all } x$$

(This is for discrete RVs).

• You’ll sometimes see the notion

$$X \mid Y \sim F$$

which means “$X$ is generated conditional on $Y$ with distribution $F$” (where $F$ usually depends on $Y$)
Outline

Dirichlet priors for categorical and multinomial distributions

Comparing discrete and continuous hypotheses
Continuous hypothesis spaces

- Bayes rule is the same when $H$ ranges over a continuous space except that $P(H)$ and $P(H|D)$ are continuous functions of $H$

\[
\frac{P(H|D)}{P(H)} \propto \frac{P(D|H)}{P(H)}
\]

- The normalizing constant is:

\[
P(D) = \int P(D|H') P(H') \, dH'
\]

- Some of the approaches you can take:
  - Monte Carlo sampling procedures (which we’ll talk about later)
  - Choose $P(H)$ so that $P(H|D)$ is easy to calculate
    ⇒ use a prior conjugate to the likelihood
Categorical distributions

- A *categorical distribution* has a finite set of outcomes 1, \ldots, m.
- A categorical distribution is parameterized by a vector \( \theta = (\theta_1, \ldots, \theta_m) \), where \( P(X = j|\theta) = \theta_j \) (so \( \sum_{j=1}^{m} \theta_j = 1 \))
  - Example: An \( m \)-sided die, where \( \theta_j = \text{prob. of face } j \)
- Suppose \( X = (X_1, \ldots, X_n) \) and each \( X_i|\theta \sim \text{CATEGORICAL(}\theta) \). Then:

\[
P(X|\theta) = \prod_{i=1}^{n} \text{CATEGORICAL}(X_i; \theta) = \prod_{j=1}^{m} \theta_j^{N_j}
\]

where \( N_j \) is the number of times \( j \) occurs in \( X \).
- Goal of next few slides: compute \( P(\theta|X) \)
Multinomial distributions

• Suppose \( X_i \sim \text{CATEGORICAL}(\theta) \) for \( i = 1, \ldots, n \), and \( N_j \) is the number of times \( j \) occurs in \( X \)
• Then \( N|n, \theta \sim \text{MULTI}(\theta, n) \), and

\[
P(N|n, \theta) = \frac{n!}{\prod_{j=1}^{m} N_j!} \prod_{j=1}^{m} \theta_j^{N_j}
\]

where \( n! / \prod_{j=1}^{m} N_j! \) is the number of sequences of values with occurrence counts \( N \)
• The vector \( N \) is known as a \textit{sufficient statistic} for \( \theta \) because it supplies as much information about \( \theta \) as the original sequence \( X \) does.


**Dirichlet distributions**

- **Dirichlet distributions** are probability distributions over multinomial parameter vectors
  - called **Beta distributions** when \( m = 2 \)
- Parameterized by a vector \( \alpha = (\alpha_1, \ldots, \alpha_m) \) where \( \alpha_j > 0 \) that determines the shape of the distribution

\[
\text{DIR}(\theta; \alpha) = \frac{1}{C(\alpha)} \prod_{j=1}^{m} \theta_j^{\alpha_j-1}
\]

\[
C(\alpha) = \int \prod_{j=1}^{m} \theta_j^{\alpha_j-1} \, d\theta = \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{m} \alpha_j)}
\]

- \( \Gamma \) is a generalization of the factorial function
- \( \Gamma(k) = (k - 1)! \) for positive integer \( k \)
- \( \Gamma(x) = (x - 1)\Gamma(x - 1) \) for all \( x \)
Plots of the Dirichlet distribution

\[ \text{DIR}(\theta; \alpha) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \]
Plots of the Dirichlet distribution (2)

\[
\text{DIR}(\theta; \alpha) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j-1}
\]
Plots of the Dirichlet distribution (3)

\[ \text{DIR}(\theta; \alpha) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j-1} \]

\( \alpha = (1, 1) \)  
\( \alpha = (0.5, 0.5) \)  
\( \alpha = (0.1, 0.1) \)  
\( \alpha = (0.1, 1) \)  
\( \alpha = (0.1, 2) \)
Dirichlet distributions as priors for $\theta$

- Generative model:

\[
\begin{align*}
\theta & \mid \alpha \sim \text{DIR}(\alpha) \\
X_i & \mid \theta \sim \text{CATEGORICAL}(\theta), \quad i = 1, \ldots, n
\end{align*}
\]

- We can depict this as a Bayes net using *plates*, which indicate replication.
Inference for $\theta$ with Dirichlet priors

- Data $X = (X_1, \ldots, X_n)$ generated i.i.d. from $\text{CATEGORICAL} (\theta)$
- Prior is $\text{DIR}(\alpha)$. By Bayes Rule, posterior is:

$$P(\theta | X) \propto P(X | \theta) P(\theta)$$

$$\propto \left( \prod_{j=1}^{m} \theta_j^{N_j} \right) \left( \prod_{j=1}^{m} \theta_j^{\alpha_j-1} \right)$$

$$= \prod_{j=1}^{m} \theta_j^{N_j + \alpha_j - 1}, \text{ so}$$

$$P(\theta | X) = \text{DIR}(N + \alpha)$$

- So if prior is Dirichlet with parameters $\alpha$, posterior is Dirichlet with parameters $N + \alpha$

$\Rightarrow$ can regard Dirichlet parameters $\alpha$ as “pseudo-counts” from “pseudo-data”
Point estimates from Bayesian posteriors

- A “true” Bayesian prefers to use the full $P(H|D)$, but sometimes we have to choose a “best” hypothesis.
- The **Maximum a posteriori** (MAP) or *posterior mode* is

$$\hat{H} = \arg\max_H P(H|D) = \arg\max_H P(D|H)P(H)$$

- The *expected value* $E_P[X]$ of $X$ under distribution $P$ is:

$$E_P[X] = \int x P(X = x) \, dx$$

The expected value is a kind of average, weighted by $P(X)$. The *expected value* $E[\theta]$ of $\theta$ is an estimate of $\theta$. 
The posterior mode of a Dirichlet

- The *Maximum a posteriori* (MAP) or *posterior mode* is:

  \[ \hat{H} = \arg \max_H P(H|D) = \arg \max_H P(D|H) P(H) \]

- For Dirichlets with parameters \( \alpha \), the MAP estimate is:

  \[ \hat{\theta}_j = \frac{\alpha_j - 1}{\sum_{j'=1}^m (\alpha_{j'} - 1)} \]

  so if the posterior is \( \text{DIR}(N + \alpha) \), the MAP estimate for \( \theta \) is:

  \[ \hat{\theta}_j = \frac{N_j + \alpha_j - 1}{n + \sum_{j'=1}^m (\alpha_{j'} - 1)} \]

- If \( \alpha = 1 \) then \( \hat{\theta}_j = N_j / n \), which is also the *maximum likelihood estimate* (MLE) for \( \theta \)
The expected value of $\theta$ for a Dirichlet

- The *expected value* $E_P[X]$ of $X$ under distribution $P$ is:

$$E_P[X] = \int x P(X = x) \, dx$$

- For Dirichlets with parameters $\alpha$, the expected value of $\theta_j$ is:

$$E_{\text{DIR}}(\alpha)[\theta_j] = \frac{\alpha_j}{\sum_{j'=1}^{m} \alpha_{j'}}$$

- Thus if the posterior is $\text{DIR}(N + \alpha)$, the expected value of $\theta_j$ is:

$$E_{\text{DIR}}(N + \alpha)[\theta_j] = \frac{N_j + \alpha_j}{n + \sum_{j'=1}^{m} \alpha_{j'}}$$

- $E[\theta]$ *smooths* or *regularizes* the MLE by adding pseudo-counts $\alpha$ to $N$
Sampling from a Dirichlet

\[ \theta \mid \alpha \sim \text{DIR}(\alpha) \iff P(\theta \mid \alpha) = \frac{1}{C(\alpha)} \prod_{j=1}^{m} \theta_{j}^{\alpha_{j} - 1}, \text{ where:} \]

\[ C(\alpha) = \frac{\prod_{j=1}^{m} \Gamma(\alpha_{j})}{\Gamma(\sum_{j=1}^{m} \alpha_{j})} \]

- There are several algorithms for producing samples from \( \text{DIR}(\alpha) \). A simple one relies on the following result:
- If \( V_k \sim \text{GAMMA}(\alpha_k) \) and \( \theta_k = V_k / (\sum_{k'=1}^{m} V_{k'}) \), then \( \theta \sim \text{DIR}(\alpha) \)
- This leads to the following algorithm for producing a sample \( \theta \) from \( \text{DIR}(\alpha) \)
  - Sample \( v_k \) from \( \text{GAMMA}(\alpha_k) \) for \( k = 1, \ldots, m \)
  - Set \( \theta_k = v_k / (\sum_{k'=1}^{m} v_{k'}) \)
Conjugate priors

- If prior is $\text{DIR}(\alpha)$ and likelihood is i.i.d. $\text{CATEGORICAL}(\theta)$, then posterior is $\text{DIR}(N + \alpha)$
  $\Rightarrow$ prior parameters $\alpha$ specify “pseudo-observations”

- A class $\mathcal{C}$ of prior distributions $P(H)$ is \textit{conjugate} to a class of likelihood functions $P(D|H)$ iff the posterior $P(H|D)$ is also a member of $\mathcal{C}$

- In general, conjugate priors encode “pseudo-observations”
  - the difference between prior $P(H)$ and posterior $P(H|D)$ are the observations in $D$
  - but $P(H|D)$ belongs to same family as $P(H)$, and can serve as prior for inferences about more data $D'$
  $\Rightarrow$ must be possible to encode observations $D$ using parameters of prior

- In general, the likelihood functions that have conjugate priors belong to the \textit{exponential family}
Outline

Dirichlet priors for categorical and multinomial distributions

Comparing discrete and continuous hypotheses
Categorical and continuous hypotheses about coin flips

- Data: A sequence of coin flips $X = (X_1, \ldots, X_n)$
- Hypothesis $h_1$: $X$ is generated from a fair coin, i.e., $\theta_H = 0.5$
- Hypothesis $h_2$: $X$ is generated from a biased coin with unknown bias, i.e., $\theta_H \sim \text{DIR}(\alpha)$

\[ P(H|X) = P(X|H) P(H) \]

- Assume $P(h_1) = P(h_2) = 0.5$
- $P(X|h_1) = 2^{-n}$, but what is $P(X|h_2)$?
- $P(X|h_2)$ is the probability of generating $\theta$ from $\text{DIR}(\alpha)$ and then generating $X$ from $\text{CATEGORICAL}(\theta)$. But we don’t care about the value of $\theta$, so we marginalize or integrate out $\theta$

\[ P(X|\alpha, h_2) = \int P(X, \theta|\alpha) d\theta \]
Posterior with Dirichlet priors

\[ \begin{align*}
\theta & \mid \alpha \sim \text{DIR}(\alpha) \\
X_i & \mid \theta \sim \text{CATEGORICAL}(\theta), \ i = 1, \ldots, n
\end{align*} \]

- **Integrate out** \( \theta \) to calculate posterior probability of \( X \)

\[
P(X|\alpha) = \int P(X, \theta|\alpha) \, d\theta = \int P(X|\theta) P(\theta|\alpha) \, d\theta
\]

\[
= \int \left( \prod_{j=1}^{m} \theta_j^{N_j} \right) \left( \frac{1}{C(\alpha)} \prod_{j=1}^{m} \theta_j^{\alpha_j-1} \right) \, d\theta
\]

\[
= \frac{1}{C(\alpha)} \int \prod_{j=1}^{m} \theta_j^{N_j+\alpha_j-1} \, d\theta
\]

\[
= \frac{C(N + \alpha)}{C(\alpha)}, \text{ where } C(\alpha) = \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{m} \alpha_j)}
\]

- **Collapsed Gibbs samplers** and the **Chinese Restaurant Process** rely on this result
Posteriors under $h_1$ and $h_2$
Understanding the posterior

\[
P(X|\alpha) = \frac{C(N + \alpha)}{C(\alpha)} \quad \text{where} \quad C(\alpha) = \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\alpha_\cdot)} \quad \text{and} \quad \alpha_\cdot = \sum_{j=1}^{m} \alpha_j
\]

\[
P(X|\alpha) = \left( \frac{\prod_{j=1}^{m} \Gamma(N_j + \alpha_j)}{\Gamma(n + \alpha_\cdot)} \right) \left( \frac{\Gamma(\alpha_\cdot)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \right)
\]

\[
= \left( \prod_{j=1}^{m} \frac{\Gamma(N_j + \alpha_j)}{\Gamma(\alpha_j)} \right) \left( \frac{\Gamma(\alpha_\cdot)}{\Gamma(n + \alpha_\cdot)} \right)
\]

\[
= \alpha_1 \times \frac{\alpha_1 + 1}{\alpha_\cdot + 1} \times \ldots \times \frac{\alpha_1 + N_1 - 1}{\alpha_\cdot + N_1 - 1} \times \frac{\alpha_2}{\alpha_\cdot + N_1} \times \frac{\alpha_2 + 1}{\alpha_\cdot + N_1 + 1} \times \ldots \times \frac{\alpha_2 + N_2 - 1}{\alpha_\cdot + N_1 + N_2 - 1} \times \ldots \times \frac{\alpha_m}{\alpha_\cdot + n - N_m - 1} \times \frac{\alpha_m + 1}{\alpha_\cdot + n - N_m} \times \ldots \times \frac{\alpha_m + N_m - 1}{\alpha_\cdot + n - 1}
\]
Exchangability

- The individual $X_i$ in a Dirichlet-multinomial distribution $P(X|\alpha) = C(N + \alpha)/C(\alpha)$ are not independent
  - the probability of $X_i$ depends on $X_1, \ldots, X_{i-1}$

$$P(X_n = k|X_1, \ldots, X_{n-1}, \alpha) = \frac{P(X_1, \ldots, X_n|\alpha)}{P(X_1, \ldots, X_{n-1}|\alpha)} = \frac{\alpha_k + N_k(X_1, \ldots, X_{n-1})}{\alpha \cdot n - 1}$$

- but $X_1, \ldots, X_n$ are exchangable
  - $P(X|\alpha)$ depends only on $N$
    $\Rightarrow$ doesn’t depend on the order in which the $X$ occur

- A distribution over a sequence of random variables is exchangable iff the probability of all permutations of the random variables are equal
Summary so far

• Bayesian inference can compare models of different complexity (assuming we can calculate posterior probability)
  ▶ Hypothesis \( h_1 \) has no free parameters
  ▶ Hypothesis \( h_2 \) has one free parameter \( \theta_H \)

• Bayesian Occam’s Razor: “A more complex hypothesis is only preferred if its greater complexity consistently provides a better account of the data”

• But: \( h_1 \) makes every sequence equally likely. \( h_2 \) seems to dislike \( \theta_H \approx 0.5 \)

What’s going on here?
Posteriors with $n = 10, \alpha = 10$

\[ P(N_{H}, N_{T} = 10 - N_{H} | \alpha = 10, h) \]

Diagram showing the probability distribution $P(N_{H}, N_{T} = 10 - N_{H} | \alpha = 10, h)$ for different values of $N_{H}$.

Graph labels:
- $h = h_{1}$ (red line)
- $h = h_{2}$ (green line)
Posteriors with $n = 20, \alpha = 1$

\[ P(N_H, N_T = 20 - N_H | \alpha = 1, \beta) \]
Dirichlet-Multinomial distributions

- Only one sequence of 10 heads out of 10 coin flips
- but 252 different sequences of 5 heads out of 10 coin flips
- Each particular sequence of 5 heads out of 10 flips is unlikely, but there are so many of them that *the group is very likely*
- The number of ways of picking $N$ outcomes out of $n$ trials is:

$$\frac{n!}{\prod_{j=1}^{m} N_j!}$$

- The probability of observing $N$ given $\theta$ is:

$$P(N|\theta) = \frac{n}{\prod_{j=1}^{m} N_j!} \prod_{j=1}^{m} \theta_j^{N_j}$$

- The probability of observing $N$ given $\alpha$ is:

$$P(N|\alpha) = \frac{n}{\prod_{j=1}^{m} N_j!} \frac{C(N + \alpha)}{C(\alpha)}$$
Dirichlet-multinomial posteriors with $n = 10, \alpha = 1$
Dirichlet-multinomial posteriors with $n = 10$, varying $\alpha$

$$\theta = 0.5$$
$$\alpha = 1$$
$$\alpha = 5$$
$$\alpha = (8, 2)$$
$$\alpha = 0.2$$
$$\alpha = 0.1$$
Dirichlet-multinomial posteriors with \( n = 20 \), varying \( \alpha \)
Dirichlet-multinomial posteriors with $n = 50$, varying $\alpha$
Entropy vs. “rich get richer”

- Notation: If $X = (X_1, \ldots, X_n)$, then $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$

$$P(X_n = k|\alpha, X_{-n}) = \frac{N_k(X_{-n}) + \alpha_k}{\alpha + n - 1}$$

- The probability of generating an outcome is proportional to the number of times it has been seen before (including prior)

  $\Rightarrow$ Next outcome is most likely to be most frequently generated previous outcome $\Rightarrow$ **sparse outcomes**

- But there are far fewer sparse outcomes than non-sparse outcomes $\Rightarrow$ entropy “prefers” non-sparse outcomes

- If $\alpha > 1$ then most likely outcomes are not sparse i.e., entropy is stronger than prior

- If $\alpha < 1$ then most likely outcomes are sparse i.e., prior is stronger than entropy