A Brief Introduction to Bayesian Inference

Mark Johnson

CG168 notes

A brief review of *discrete* probability theory

- Ω is the set of all *elementary events* (c.f. interpretations in logic)
- If $\omega \in \Omega$, then $\mathrm{P}(\omega)$ is the probability of event ω
  - $\mathrm{P}(\omega) \geq 0$
  - $\sum_{\omega \in \Omega} \mathrm{P}(\omega) = 1$
- A *random variable* X is a function from Ω 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

$$\mathrm{P}(X = x) = \sum_{\substack{\omega \in \Omega \\ X(\omega) = x}} \mathrm{P}(\omega)$$

Independence and conditional distributions

- Two RVs X and Y are *independent* iff $\mathrm{P}(X, Y) = \mathrm{P}(X)\mathrm{P}(Y)$
- The *conditional distribution* of Y given X is:

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

  so X and Y are independent iff $\mathrm{P}(Y|X) = \mathrm{P}(Y)$ (here and below I assume strictly positive distributions)
- We can decompose the joint distribution of a sequence of RVs into a product of conditionals:

$$\begin{aligned} & \mathrm{P}(X_1, \ldots, X_n) \\ = \;& \mathrm{P}(X_1)\mathrm{P}(X_2|X_1)\mathrm{P}(X_3|X_2, X_1)\ldots\mathrm{P}(X_n|X_{n-1}, \ldots, X_1) \end{aligned}$$

  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 $\boldsymbol{X} = (X_1, \ldots, X_n)$ into a product of conditionals

$$\mathrm{P}(\boldsymbol{X}) \;=\; \prod_{i=1}^{n} \mathrm{P}(X_i | X_1, \ldots, X_{i-1})$$

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

$$\mathrm{P}(\boldsymbol{X}) \;=\; \prod_{i} \mathrm{P}(X_i | \boldsymbol{X}_{\mathsf{Pa}(i)})$$

  where $\mathsf{Pa}(i) \subseteq \{1, \ldots, i-1\}$ and $\boldsymbol{X}_S = \{X_j : j \in S\}$
- X and Y are *conditionally independent* given Z iff $\mathrm{P}(X, Y | Z) = \mathrm{P}(X|Z)\,\mathrm{P}(Y|Z)$ or equivalently, $\mathrm{P}(X|Y, Z) = \mathrm{P}(X|Z)$
- Note: the "parents" $\mathsf{Pa}(i)$ 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

$$\mathrm{P}(\boldsymbol{X}) = \prod_i \mathrm{P}(X_i | \boldsymbol{X}_{\mathsf{Pa}(i)})$$

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

Bayes rule

- Bayes theorem:

$$\mathrm{P}(Y|X) \;=\; \frac{\mathrm{P}(X|Y)\,\mathrm{P}(Y)}{\mathrm{P}(X)}$$

- Bayes inversion: swap direction of arcs in Bayes net
- Interpreted as a recipe for "belief updating":

$$\underbrace{\mathrm{P}(\mathsf{Hypothesis}|\mathsf{Data})}_{\text{Posterior}} \;\propto\; \underbrace{\mathrm{P}(\mathsf{Data}|\mathsf{Hypothesis})}_{\text{Likelihood}}\;\underbrace{\mathrm{P}(\mathsf{Hypothesis})}_{\text{Prior}}$$

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

$$\mathrm{P}(\mathsf{Data}) \;=\; \sum_{\mathsf{Hypothesis}'} \mathrm{P}(\mathsf{Data}|\mathsf{Hypothesis}')\,\mathrm{P}(\mathsf{Hypothesis}')$$

which is the probability of generating the data under *any* hypothesis

Iterated Bayesian belief updating

- Suppose the data consists of 2 components $\boldsymbol{D} = (D_1, D_2)$, and $\mathrm{P}(H)$ is our prior over hypotheses H

$$\begin{aligned} \mathrm{P}(H|D_1, D_2) &\propto \mathrm{P}(D_1, D_2|H)\,\mathrm{P}(H) \\ &\propto \mathrm{P}(D_2|H, D_1)\,\mathrm{P}(H|D_1) \end{aligned}$$

- This means the following are equivalent:
  - update the prior $\mathrm{P}(H)$ treating (D_1, D_2) as a single observation
  - update the prior $\mathrm{P}(H)$ wrt the first observation D_1 producing posterior $\mathrm{P}(H|D_1) \propto \mathrm{P}(D_1|H)\,\mathrm{P}(H)$, 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 $\mathrm{P}(H)$

$$\mathrm{P}(H|d_1, d_2) \propto \mathrm{P}(d_1, d_2|H)\,\mathrm{P}(H) = \mathrm{P}(d_1, d_2, H)$$

- Incremental Bayesian belief updating
  - Bayesian belief updating on d_1 using prior $\mathrm{P}(H)$

$$\mathrm{P}(H|d_1) \propto \mathrm{P}(d_1|H)\,\mathrm{P}(H) = \mathrm{P}(d_1, H)$$

  - Bayesian belief updating on d_2 using prior $\mathrm{P}(H|d_1)$

$$\begin{aligned}\mathrm{P}(H|d_1, d_2) &\propto \mathrm{P}(d_2|H, d_1)\,\mathrm{P}(H|d_1) \\ &\propto \mathrm{P}(d_2|H, d_1)\,\mathrm{P}(H, d_1) \\ &= \mathrm{P}(d_2, d_1, H)\end{aligned}$$

"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:

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

  (This is for discrete RVs).
- You'll sometimes see the notion

$$X \,|\, Y \;\sim\; F$$

  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 $\mathrm{P}(H)$ and $\mathrm{P}(H|D)$ are *continuous functions* of H

$$\underbrace{\mathrm{P}(H|D)}_{\text{Posterior}} \quad \propto \quad \underbrace{\mathrm{P}(D|H)}_{\text{Likelihood}} \; \underbrace{\mathrm{P}(H)}_{\text{Prior}}$$

- The normalizing constant is:

$$\mathrm{P}(D) \quad = \quad \int \mathrm{P}(D|H')\, \mathrm{P}(H')\, dH'$$

- Some of the approaches you can take:
  - Monte Carlo sampling procedures (which we'll talk about later)
  - Choose $\mathrm{P}(H)$ so that $\mathrm{P}(H|D)$ is easy to calculate
    $\Rightarrow$ 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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, where $\mathrm{P}(X = j|\boldsymbol{\theta}) = \theta_j$ (so $\sum_{j=1}^{m} \theta_j = 1$)
  - Example: An m-sided die, where $\theta_j =$ prob. of face j
- Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ and each $X_i|\boldsymbol{\theta} \sim \textsc{Categorical}(\boldsymbol{\theta})$. Then:

$$\mathrm{P}(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \textsc{Categorical}(X_i; \boldsymbol{\theta}) = \prod_{j=1}^{m} \theta_j^{N_j}$$

  where N_j is the number of times j occurs in $\boldsymbol{X}$.
- Goal of next few slides: compute $\mathrm{P}(\boldsymbol{\theta}|\boldsymbol{X})$

Multinomial distributions

- Suppose $X_i \sim \textsc{Categorical}(\boldsymbol{\theta})$ for $i = 1, \ldots, n$, and N_j is the number of times j occurs in $\boldsymbol{X}$
- Then $\boldsymbol{N}|n, \boldsymbol{\theta} \sim \textsc{Multi}(\boldsymbol{\theta}, n)$, and

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

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

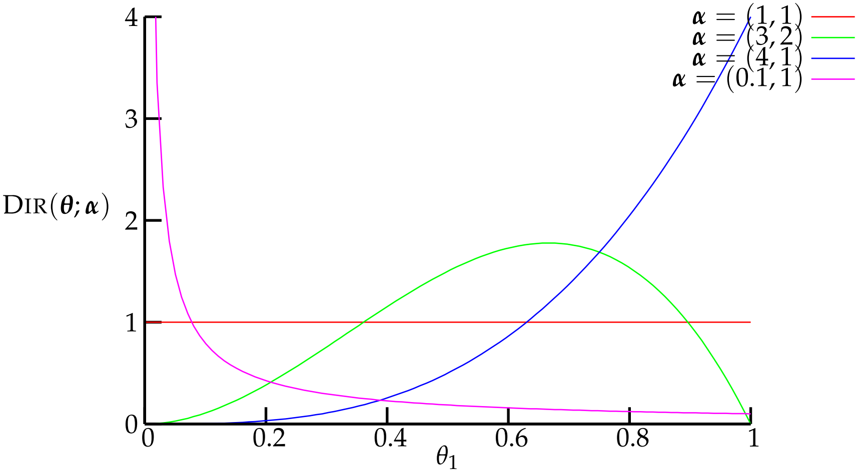
Dirichlet distributions

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

$$
\begin{aligned}
\textsc{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha}) &= \frac{1}{C(\boldsymbol{\alpha})} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \\
C(\boldsymbol{\alpha}) &= \int \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \, d\boldsymbol{\theta} = \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{m} \alpha_j)}
\end{aligned}
$$

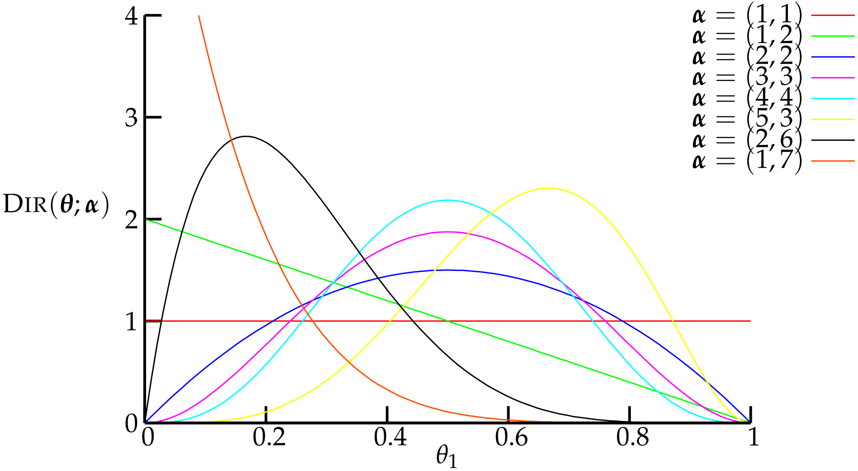
- Γ is a generalization of the factorial function
- $\Gamma(k) = (k-1)!$ for positive integer k
- $\Gamma(x) = (x-1)\Gamma(x-1)$ for all x

Plots of the Dirichlet distribution



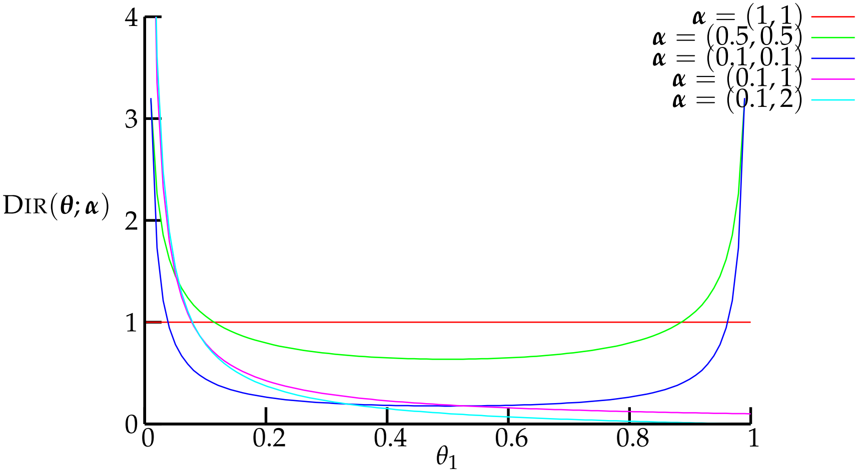
$$\mathrm{DIR}(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{m}\alpha_j)}{\prod_{j=1}^{m}\Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j-1}$$

Plots of the Dirichlet distribution (2)



$$\textsc{Dir}(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{m}\alpha_j)}{\prod_{j=1}^{m}\Gamma(\alpha_j)}\prod_{j=1}^{m}\theta_j^{\alpha_j-1}$$

Plots of the Dirichlet distribution (3)



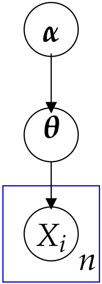
$$\textsc{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1}$$

Dirichlet distributions as priors for $\boldsymbol{\theta}$

- Generative model:

$$\begin{array}{rcl} \boldsymbol{\theta} \mid \boldsymbol{\alpha} & \sim & \text{DIR}(\boldsymbol{\alpha}) \\ X_i \mid \boldsymbol{\theta} & \sim & \text{CATEGORICAL}(\boldsymbol{\theta}), \quad i = 1, \ldots, n \end{array}$$

- We can depict this as a Bayes net using *plates*, which indicate *replication*



Inference for $\boldsymbol{\theta}$ with Dirichlet priors

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

$$\begin{aligned} \mathrm{P}(\boldsymbol{\theta}|\boldsymbol{X}) &\propto \mathrm{P}(\boldsymbol{X}|\boldsymbol{\theta})\ \mathrm{P}(\boldsymbol{\theta}) \\ &\propto \left(\prod_{j=1}^{m} \theta_j^{N_j}\right) \left(\prod_{j=1}^{m} \theta_j^{\alpha_j - 1}\right) \\ &= \prod_{j=1}^{m} \theta_j^{N_j + \alpha_j - 1}, \text{ so} \\ \mathrm{P}(\boldsymbol{\theta}|\boldsymbol{X}) &= \textsc{Dir}(\boldsymbol{N} + \boldsymbol{\alpha}) \end{aligned}$$

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

$\Rightarrow$ can regard Dirichlet parameters $\boldsymbol{\alpha}$ as "pseudo-counts" from "pseudo-data"

Point estimates from Bayesian posteriors

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

$$\widehat{H} \;=\; \underset{H}{argmax}\,\mathrm{P}(H|D) \;=\; \underset{H}{argmax}\,\mathrm{P}(D|H)\,\mathrm{P}(H)$$

- The *expected value* $\mathrm{E}_{\mathrm{P}}[X]$ of X under distribution P is:

$$\mathrm{E}_{\mathrm{P}}[X] \;=\; \int x\,\mathrm{P}(X=x)\,dx$$

The expected value is a kind of average, weighted by $\mathrm{P}(X)$. The *expected value* $\mathrm{E}[\boldsymbol{\theta}]$ of $\boldsymbol{\theta}$ is an estimate of $\boldsymbol{\theta}$.

The posterior mode of a Dirichlet

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

$$\widehat{H} \;=\; \underset{H}{argmax}\, \mathrm{P}(H|D) \;=\; \underset{H}{argmax}\, \mathrm{P}(D|H)\, \mathrm{P}(H)$$

- For Dirichlets with parameters $\boldsymbol{\alpha}$, the MAP estimate is:

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

so if the posterior is $\textsc{Dir}(\boldsymbol{N} + \boldsymbol{\alpha})$, the MAP estimate for $\boldsymbol{\theta}$ is:

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

- If $\boldsymbol{\alpha} = \mathbf{1}$ then $\hat{\theta}_j = N_j/n$, which is also the *maximum likelihood estimate* (MLE) for $\boldsymbol{\theta}$

The expected value of $\boldsymbol{\theta}$ for a Dirichlet

- The *expected value* $\mathrm{E}_{\mathrm{P}}[X]$ of X under distribution P is:

$$\mathrm{E}_{\mathrm{P}}[X] \;=\; \int x\, \mathrm{P}(X = x)\, dx$$

- For Dirichlets with parameters $\boldsymbol{\alpha}$, the expected value of θ_j is:

$$\mathrm{E}_{\mathrm{DIR}(\boldsymbol{\alpha})}[\theta_j] \;=\; \frac{\alpha_j}{\sum_{j'=1}^{m} \alpha_{j'}}$$

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

$$\mathrm{E}_{\mathrm{DIR}(\boldsymbol{N}+\boldsymbol{\alpha})}[\theta_j] \;=\; \frac{N_j + \alpha_j}{n + \sum_{j'=1}^{m} \alpha_{j'}}$$

- $\mathrm{E}[\boldsymbol{\theta}]$ *smooths* or *regularizes* the MLE by adding pseudo-counts $\boldsymbol{\alpha}$ to $\boldsymbol{N}$

Sampling from a Dirichlet

$$\boldsymbol{\theta} \,|\, \boldsymbol{\alpha} \;\sim\; \textsc{Dir}(\boldsymbol{\alpha}) \quad \text{iff} \quad \mathrm{P}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \;=\; \frac{1}{\mathrm{C}(\boldsymbol{\alpha})} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1}, \text{ where:}$$

$$\mathrm{C}(\boldsymbol{\alpha}) \;=\; \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{m} \alpha_j)}$$

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

Conjugate priors

- If prior is $\mathrm{DIR}(\boldsymbol{\alpha})$ and likelihood is i.i.d. $\mathrm{CATEGORICAL}(\boldsymbol{\theta})$, then posterior is $\mathrm{DIR}(\boldsymbol{N} + \boldsymbol{\alpha})$
  $\Rightarrow$ prior parameters $\boldsymbol{\alpha}$ specify "pseudo-observations"
- A class $\mathcal{C}$ of prior distributions $\mathrm{P}(H)$ is *conjugate* to a class of likelihood functions $\mathrm{P}(D|H)$ iff the posterior $\mathrm{P}(H|D)$ is also a member of $\mathcal{C}$
- In general, conjugate priors encode "pseudo-observations"
  - the difference between prior $\mathrm{P}(H)$ and posterior $\mathrm{P}(H|D)$ are the observations in D
  - but $\mathrm{P}(H|D)$ belongs to same family as $\mathrm{P}(H)$, 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 *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 $\boldsymbol{X} = (X_1, \ldots, X_n)$
- Hypothesis h_1: $\boldsymbol{X}$ is generated from a fair coin, i.e., $\theta_H = 0.5$
- Hypothesis h_2: $\boldsymbol{X}$ is generated from a biased coin with unknown bias, i.e., $\theta_H \sim \textsc{Dir}(\boldsymbol{\alpha})$

$$\mathrm{P}(H|\boldsymbol{X}) \;=\; \mathrm{P}(\boldsymbol{X}|H)\,\mathrm{P}(H)$$

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

$$\mathrm{P}(\boldsymbol{X}|\boldsymbol{\alpha}, h_2) \;=\; \int \mathrm{P}(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{\alpha})\, d\boldsymbol{\theta}$$

Posterior with Dirichlet priors

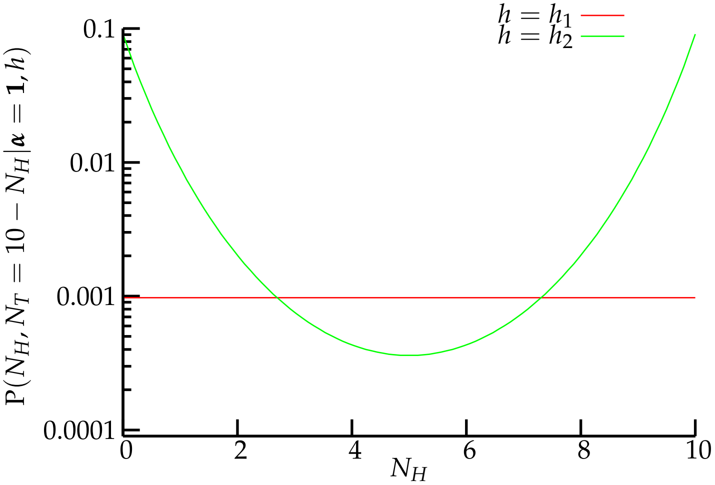
$$\begin{array}{rcll} \boldsymbol{\theta} & \mid & \boldsymbol{\alpha} & \sim \ \mathrm{DIR}(\boldsymbol{\alpha}) \\ X_i & \mid & \boldsymbol{\theta} & \sim \ \mathrm{CATEGORICAL}(\boldsymbol{\theta}), \quad i = 1, \ldots, n \end{array}$$

- *Integrate out* $\boldsymbol{\theta}$ to calculate posterior probability of $\boldsymbol{X}$

$$\begin{aligned} \mathrm{P}(\boldsymbol{X}|\boldsymbol{\alpha}) &= \int \mathrm{P}(\boldsymbol{X}, \boldsymbol{\theta}|\alpha)\, d\boldsymbol{\theta} \;=\; \int \mathrm{P}(\boldsymbol{X}|\boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta}|\boldsymbol{\alpha})\, d\boldsymbol{\theta} \\ &= \int \left(\prod_{j=1}^{m} \theta_j^{N_j} \right) \left(\frac{1}{\mathrm{C}(\boldsymbol{\alpha})} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \right) d\boldsymbol{\theta} \\ &= \frac{1}{\mathrm{C}(\boldsymbol{\alpha})} \int \prod_{j=1}^{m} \theta_j^{N_j + \alpha_j - 1}\, d\boldsymbol{\theta} \\ &= \frac{\mathrm{C}(\boldsymbol{N} + \boldsymbol{\alpha})}{\mathrm{C}(\boldsymbol{\alpha})}, \text{ where } \mathrm{C}(\boldsymbol{\alpha}) \;=\; \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{m} \alpha_j)} \end{aligned}$$

- *Collapsed Gibbs samplers* and the *Chinese Restaurant Process* rely on this result

Posteriors under h_1 and h_2



Understanding the posterior

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

$$\begin{aligned}
P(\boldsymbol{X}|\boldsymbol{\alpha}) &= \left(\frac{\prod_{j=1}^{m}\Gamma(N_j+\alpha_j)}{\Gamma(n+\alpha_\bullet)}\right)\left(\frac{\Gamma(\alpha_\bullet)}{\prod_{j=1}^{m}\Gamma(\alpha_j)}\right) \\
&= \left(\prod_{j=1}^{m}\frac{\Gamma(N_j+\alpha_j)}{\Gamma(\alpha_j)}\right)\left(\frac{\Gamma(\alpha_\bullet)}{\Gamma(n+\alpha_\bullet)}\right) \\
&= \frac{\alpha_1}{\alpha_\bullet} \times \frac{\alpha_1+1}{\alpha_\bullet+1} \times \ldots \times \frac{\alpha_1+N_1-1}{\alpha_\bullet+N_1-1} \\
&\quad \times \frac{\alpha_2}{\alpha_\bullet+N_1} \times \frac{\alpha_2+1}{\alpha_\bullet+N_1+1} \times \ldots \times \frac{\alpha_2+N_2-1}{\alpha_\bullet+N_1+N_2-1} \\
&\quad \times \ldots \\
&\quad \times \frac{\alpha_m}{\alpha_\bullet+n-N_m-1} \times \frac{\alpha_m+1}{\alpha_\bullet+n-N_m} \times \ldots \times \frac{\alpha_m+N_m-1}{\alpha_\bullet+n-1}
\end{aligned}$$

Exchangability

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

$$\begin{aligned}\mathrm{P}(X_n = k | X_1, \ldots, X_{n-1}, \boldsymbol{\alpha}) &= \frac{\mathrm{P}(X_1, \ldots, X_n | \boldsymbol{\alpha})}{\mathrm{P}(X_1, \ldots, X_{n-1} | \boldsymbol{\alpha})} \\ &= \frac{\alpha_k + N_k(X_1, \ldots, X_{n-1})}{\alpha_\bullet + n - 1}\end{aligned}$$

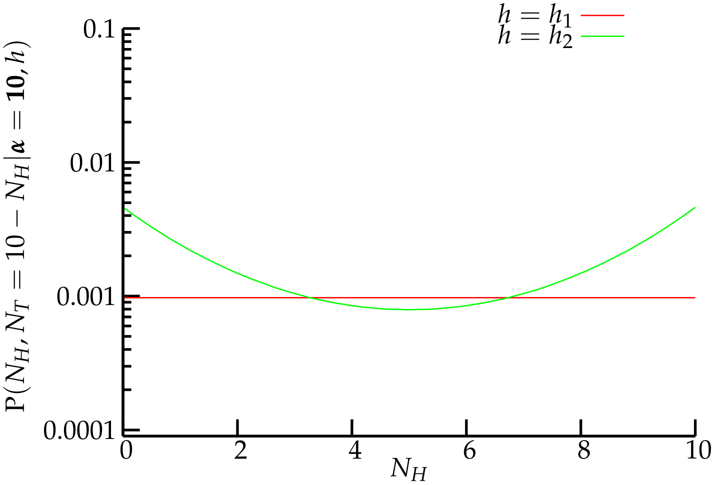
- but $X_1, \ldots, X_n$ are *exchangable*
  - $\mathrm{P}(\boldsymbol{X}|\boldsymbol{\alpha})$ depends only on $\boldsymbol{N}$

  $\Rightarrow$ doesn't depend on *the order* in which the $\boldsymbol{X}$ occur
- A distribution over a sequence of random variables is *exchangable* iff *the probability of all permutations of the random variables are equal*

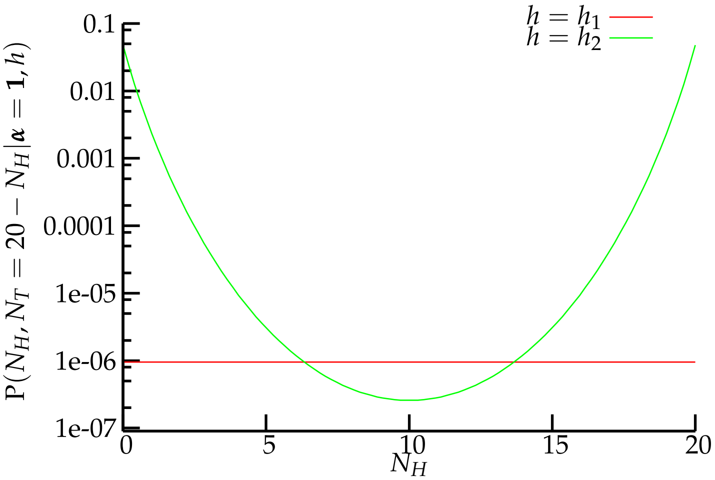
Summary so far

- Bayesian inference can compare models of *differing complexity* (assuming we can *calculate posterior probability*)
  - Hypothesis h_1 has no free parameters
  - Hypothesis h_2 has one free parameter θ_H
- *Bayesian Occam's Razor:* "A more complex hypothesis is only prefered 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* $\theta_H \approxeq 0.5$
  What's going on here?

Posteriors with $n = 10, \boldsymbol{\alpha} = \mathbf{10}$



Posteriors with $n = 20, \boldsymbol{\alpha} = \mathbf{1}$



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 $\boldsymbol{N}$ outcomes out of n trials is:

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

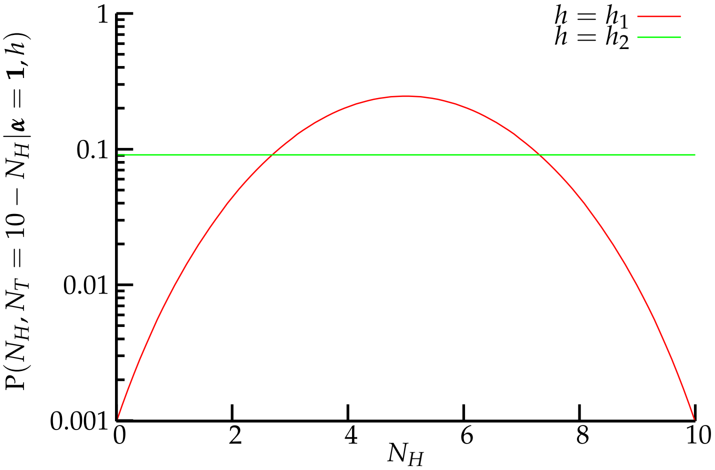
- The probability of observing $\boldsymbol{N}$ given $\boldsymbol{\theta}$ is:

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

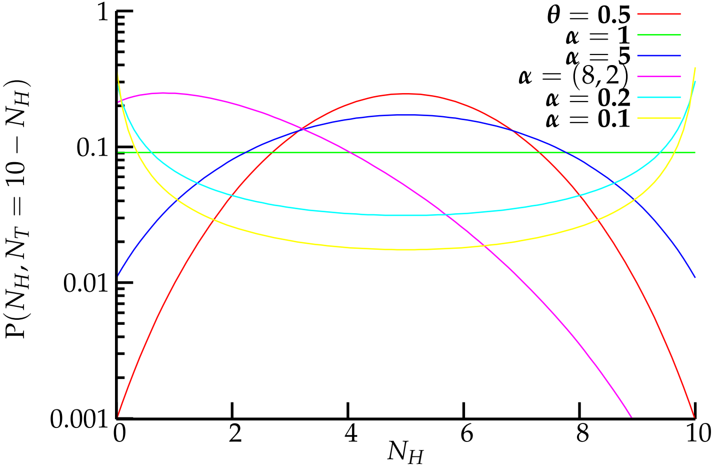
- The probability of observing $\boldsymbol{N}$ given $\boldsymbol{\alpha}$ is:

$$\mathrm{P}(\boldsymbol{N}|\boldsymbol{\alpha}) = \frac{n}{\prod_{j=1}^{m} N_j!} \frac{\mathrm{C}(\boldsymbol{N}+\boldsymbol{\alpha})}{\mathrm{C}(\boldsymbol{\alpha})}$$

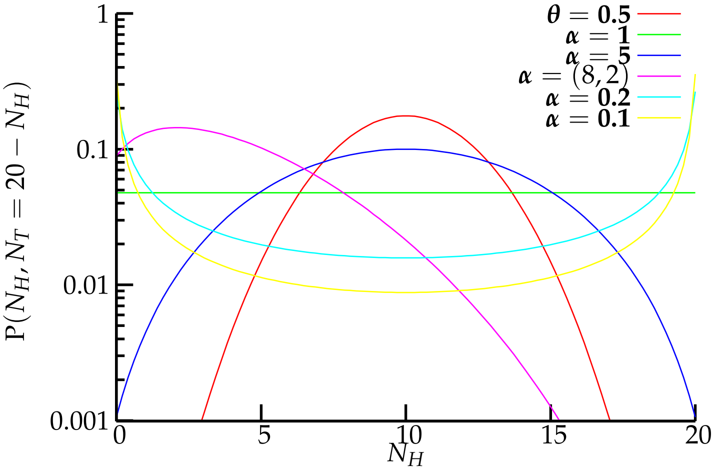
Dirichlet-**multinomial** posteriors with $n = 10, \boldsymbol{\alpha} = \mathbf{1}$



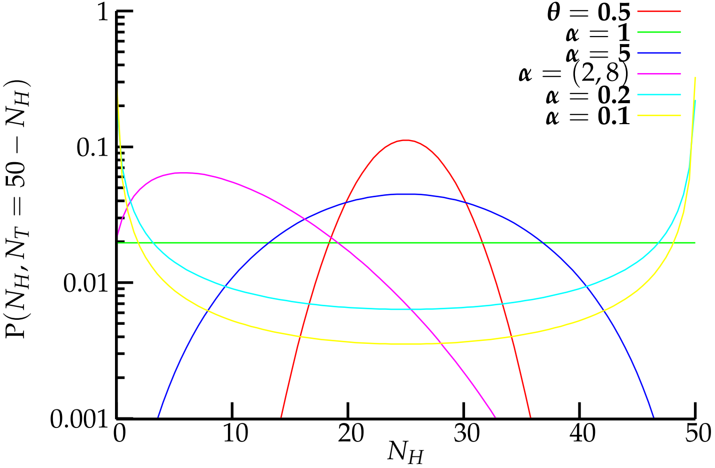
Dirichlet-**multinomial** posteriors with $n = 10$, varying $\boldsymbol{\alpha}$



Dirichlet-**multinomial** posteriors with $n = 20$, varying $\boldsymbol{\alpha}$



Dirichlet-**multinomial** posteriors with $n = 50$, varying $\boldsymbol{\alpha}$



Entropy vs. "rich get richer"

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

$$\mathrm{P}(X_n = k | \boldsymbol{\alpha}, \boldsymbol{X}_{-n}) = \frac{N_k(\boldsymbol{X}_{-n}) + \alpha_k}{\alpha_\bullet + n - 1}$$

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