

Dirichlet-Multinomials, Dirichlet-Categoricals and Hierarchical Bayesian models for learning linguistic structure

Computational Cognitive Science, Lecture 15
Stella Frank, stella.frank@ed.ac.uk
November 6, 2018

Today

1. Dirichlet Multinomials, again



2. The Learnability of Abstract Syntactic Principles





$y \sim \text{Multinomial}(\theta)$
 $\theta \sim \text{Dirichlet}(\alpha)$

$y \sim \text{Multinomial}(\theta)$

$\theta \sim \text{Dirichlet}(\alpha)$





Likelihoods: $p(y | \theta)$

Likelihoods are usually fairly determined by the question (plus domain knowledge).

- If we want to model heights:
 - ~continuous outcome (cm)
 - ~symmetric around the mean

Likelihood

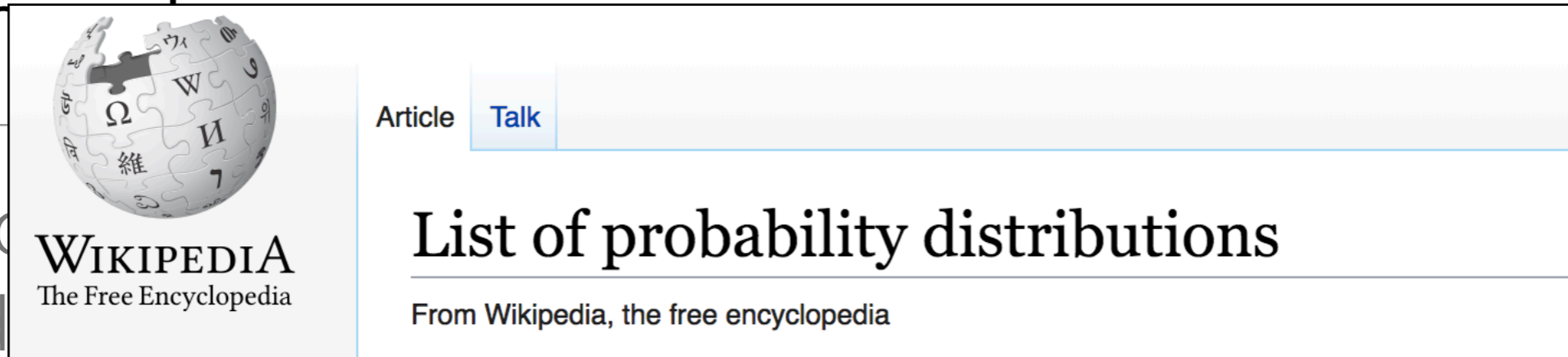
Likelihood
(plus d

• If we want

• ~continuous

• ~symmetric

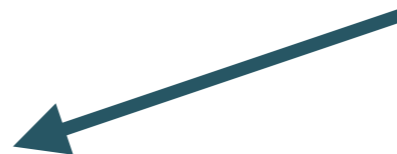
• positive



The screenshot shows the top portion of a Wikipedia article. On the left is the Wikipedia logo, a globe made of puzzle pieces with various characters, and the text 'WIKIPEDIA The Free Encyclopedia'. To the right of the logo are two tabs: 'Article' and 'Talk'. The main title of the article is 'List of probability distributions', and below it is the subtitle 'From Wikipedia, the free encyclopedia'.

- The [Johnson SU distribution](#)
- The [Landau distribution](#)
- The [Laplace distribution](#)
- The [Lévy skew alpha-stable distribution](#) or [stable distribution](#) is a family of data and critical behavior; the [Cauchy distribution](#), [Holtsmark distribution](#), [Lévy distribution](#) are special cases.
- The [Linnik distribution](#)
- The [logistic distribution](#)
- The [map-Airy distribution](#)
- The [normal distribution](#), also called the Gaussian or the bell curve. It is ubiquitous. [Central limit theorem](#): every variable that can be modelled as a sum of many small independent variables with finite [mean](#) and [variance](#) is approximately normal.
- The [Normal-exponential-gamma distribution](#)
- The [Normal-inverse Gaussian distribution](#)
- The [Pearson Type IV distribution](#) (see [Pearson distributions](#))
- The [skew normal distribution](#)
- [Student's t-distribution](#), useful for estimating unknown means of Gaussian variables.
 - The [noncentral t-distribution](#)
 - The [skew t distribution](#)

We've heard of this one



Like

Stan Modeling Language

Likelihood

User's Guide and Reference Manual

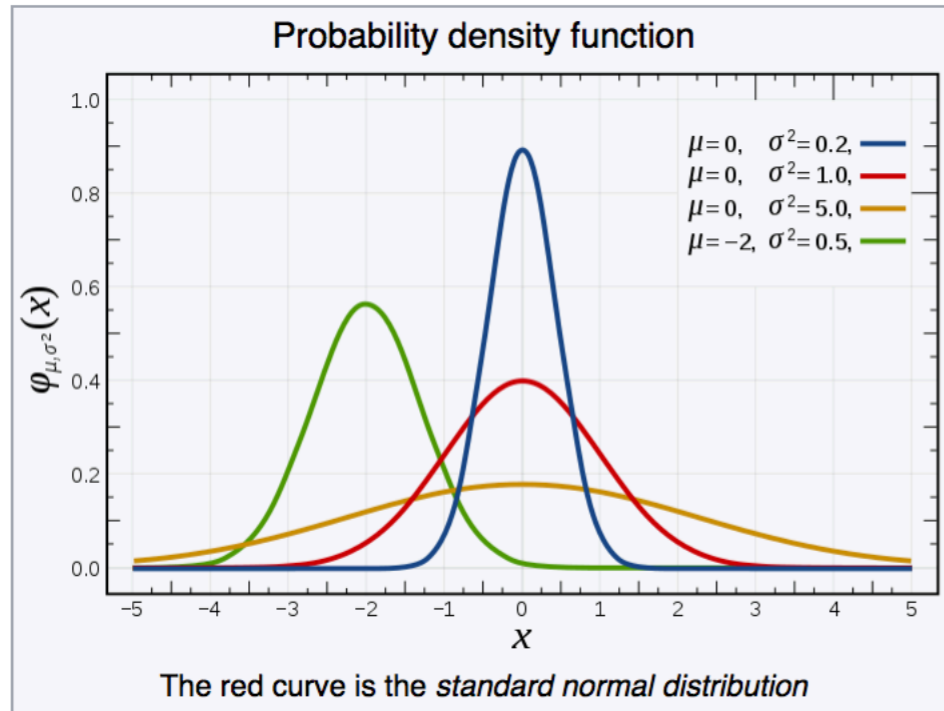
(plus domain knowledge).

Question

- If we want to model heights:
 - ~continuous outcome (cm)
 - ~symmetric around the mean

▼ VIII Discrete Distributions
Conventions for Probability Functions
Binary Distributions
Bounded Discrete Distributions
Unbounded Discrete Distributions
Multivariate Discrete Distributions
▼ IX Continuous Distributions
Unbounded Continuous Distributions
Positive Continuous Distributions
Non-negative Continuous Distributions
Positive Lower-Bounded Probabilities
Continuous Distributions on [0, 1]
Circular Distributions
Bounded Continuous Probabilities
Distributions over Unbounded Vectors
Simplex Distributions
Correlation Matrix Distributions
Covariance Matrix Distributions

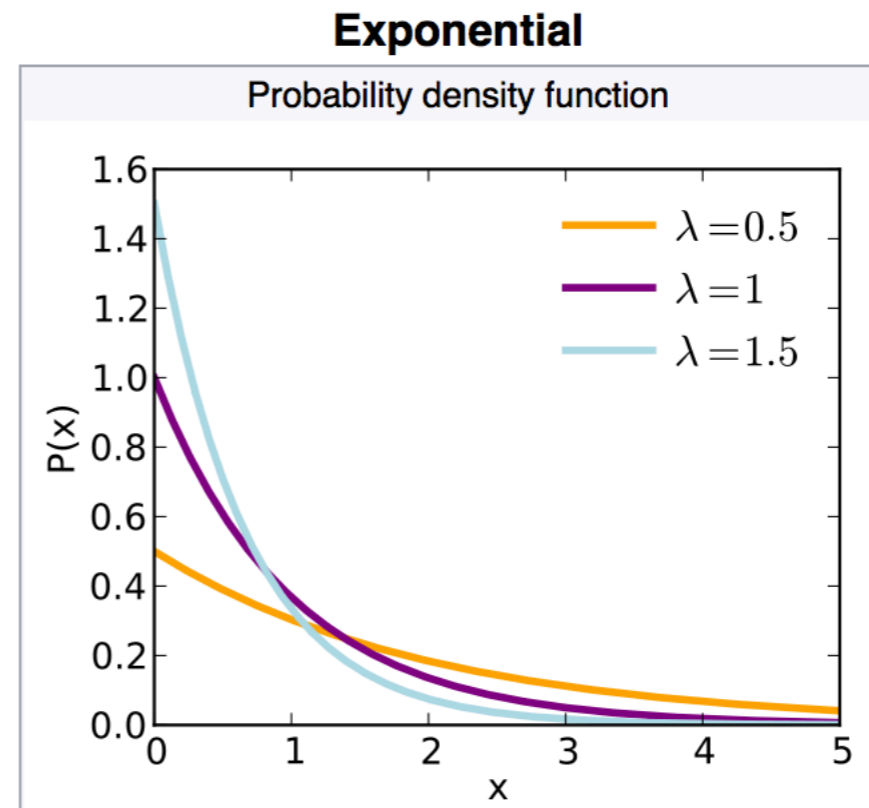
Normal Distribution



Looks reasonable...

height \sim Normal(165,30)

Not so much:





Likelihoods: $p(y | \theta)$

Likelihoods are usually mostly determined by the question (plus domain knowledge).

- Word Learning (Frank et al 2009 model):
 - need to relate words and objects with a lexicon
 - maybe other ways of doing this!

Likelihoods: $p(y | \theta)$

Likelihoods are usually mostly determined by the question (plus domain knowledge).

- Marbles world:  
 - discrete outcomes (colours)
 - multiple outcomes (sets of draws)

Likelihoods: $p(y | \theta)$

Likelihoods are usually mostly determined by the question (plus domain knowledge).

- Marbles world:
 - discrete outcomes (colours)
 - multiple outcomes (sets of draws)

53. Multivariate Discrete Distributions

The multivariate discrete distributions are over multiple integer values, which are expressed in Stan as arrays.

53.1. Multinomial Distribution

Other uses for the multivariate

- What's the probability of a bunch of words in a text?
(ignoring syntax: "bag of words" model)
- Birthday paradox: what's the probability of two of you having the same birthday?
- Distribution of students to tutorial groups

What's the probability of a bunch of marbles?

- Assume we know the distribution of marbles in the bag






$$\theta = [\text{red} = 0.1, \text{green} = 0.4, \text{blue} = 0.5]$$

- and we see  - what's $p(\text{blue} \mid \theta)$?

What's the probability of a bunch of marbles?

- Assume we know the distribution of marbles in the bag

$$\theta = [\text{red} = 0.1, \text{green} = 0.4, \text{blue} = 0.5]$$

- and we see  - what's $p(\text{blue} \mid \theta)$?
- What about   ?
- What about   ?

What's the probability of a bunch of marbles?

53. Multivariate Discrete Distributions

The multivariate discrete distributions are over multiple integer values, which are expressed in Stan as arrays.

53.1. Multinomial Distribution

Probability Mass Function

If $K \in \mathbb{N}$, $N \in \mathbb{N}$, and $\theta \in K$ -simplex, then for $y \in \mathbb{N}^K$ such that $\sum_{k=1}^K y_k = N$,

$$\text{Multinomial}(y|\theta) = \binom{N}{y_1, \dots, y_K} \prod_{k=1}^K \theta_k^{y_k},$$

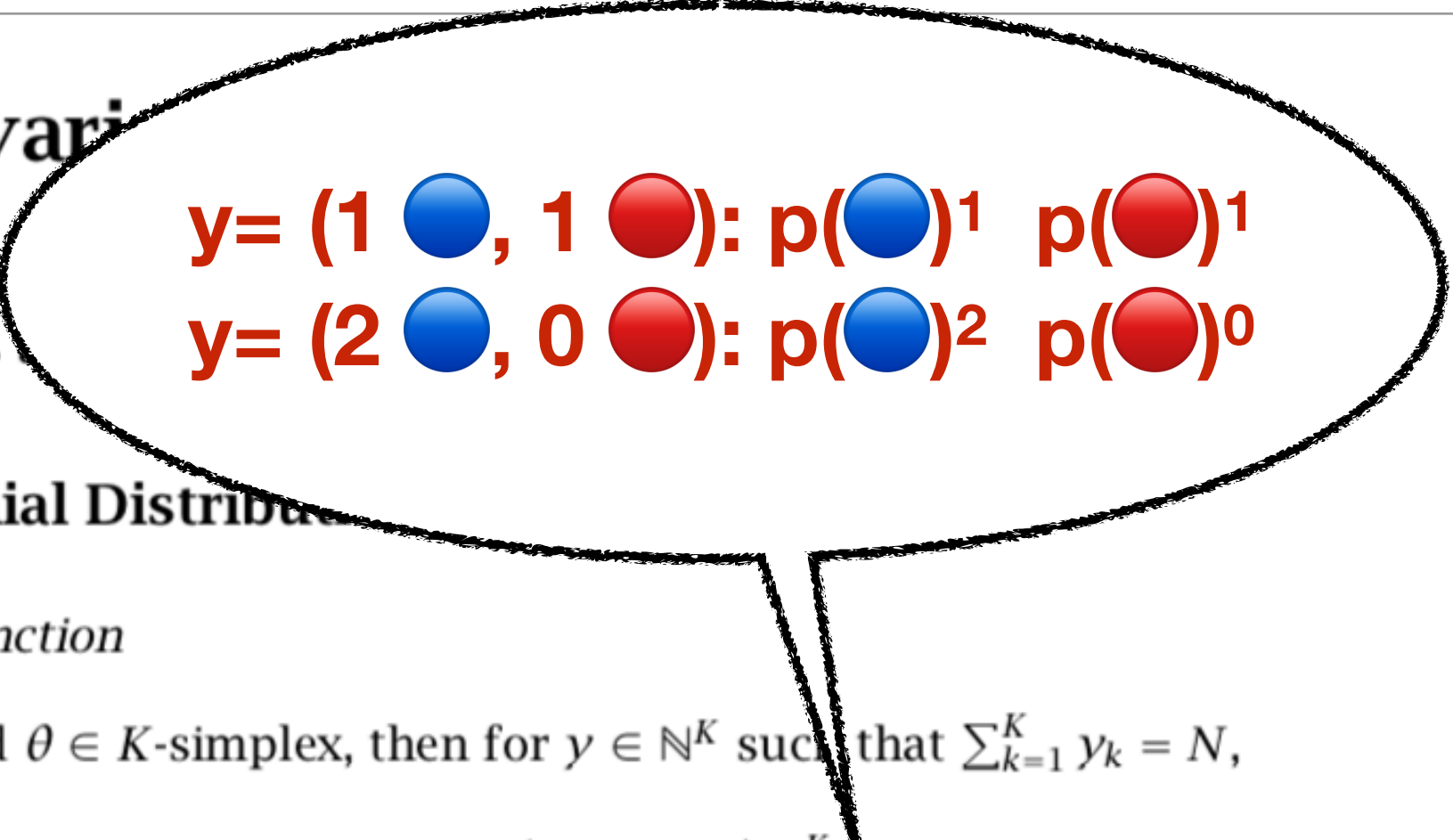
where the multinomial coefficient is defined by

$$\binom{N}{y_1, \dots, y_K} = \frac{N!}{\prod_{k=1}^K y_k!}.$$

What's the probability of a bunch of marbles?

53. Multivariate

The multivariate distribution is expressed in Stan as


$$y = (1 \text{ blue}, 1 \text{ red}): p(\text{blue})^1 p(\text{red})^1$$
$$y = (2 \text{ blue}, 0 \text{ red}): p(\text{blue})^2 p(\text{red})^0$$

53.1. Multinomial Distribution

Probability Mass Function

If $K \in \mathbb{N}$, $N \in \mathbb{N}$, and $\theta \in K$ -simplex, then for $y \in \mathbb{N}^K$ such that $\sum_{k=1}^K y_k = N$,

$$\text{Multinomial}(y|\theta) = \binom{N}{y_1, \dots, y_K} \prod_{k=1}^K \theta_k^{y_k},$$

where the multinomial coefficient is defined by

$$\binom{N}{y_1, \dots, y_K} = \frac{N!}{\prod_{k=1}^K y_k!}.$$

What's the number of sequences of a bunch of marbles?

**Multinomial Coefficient:
number of sequences giving rise
to counts y :**

$$y = (1 \text{ blue}, 1 \text{ red}) \rightarrow 2!/1! = 2$$



53.

Probability Mass Function

If $K \in \mathbb{N}$, $N \in \mathbb{N}$, and $\theta \in K$ -simplex, then for $y \in \mathbb{N}^K$ such that $\sum_{k=1}^K y_k = N$,

$$\text{Multinomial}(y|\theta) = \binom{N}{y_1, \dots, y_K} \prod_{k=1}^K \theta_k^{y_k},$$

where the multinomial coefficient is defined by

$$\binom{N}{y_1, \dots, y_K} = \frac{N!}{\prod_{k=1}^K y_k!}.$$

distributions

the integer values, which are

What's the probability of a *sequence* of marbles?

- Assume we know the distribution of marbles in the bag

$$\theta = [\text{red} = 0.1, \text{green} = 0.4, \text{blue} = 0.5]$$

- and we see  and then :

$$p(\text{blue}, \text{red} \mid \theta) = \theta_{\text{blue}} \times \theta_{\text{red}} = 0.05$$

Note: no multinomial coefficient here!

51.5. Categorical Distribution

Probability Mass Functions

If $N \in \mathbb{N}$, $N > 0$, and if $\theta \in \mathbb{R}^N$ forms an N -simplex (i.e., has nonnegative entries summing to one), then for $y \in \{1, \dots, N\}$,

$$\text{Categorical}(y \mid \theta) = \theta_y.$$

Ok, but what if we don't know θ ?

- Now we know how to calculate $P(y \mid \theta)$, if we have θ
- But we generally don't:
 - Can't inspect inside the bag of marbles
 - More realistically: we can only see a finite amount of text, but we want to estimate the distribution over words for the language as a whole

One method: search for a good θ

$y = (\text{blue} \text{ red})$ [with Multinomial Likelihood]

- $\theta = [R=0.3, G=0.3, B=0.3]: P(y | \theta) = 0.22222$
- $\theta = [R=0.2, G=0.7, B=0.1]: P(y | \theta) = 0.04$
- $\theta = [R=0.4, G=0.1, B=0.5]: P(y | \theta) = 0.4$
- $\theta = [R=0.5, G=0.0, B=0.5]: P(y | \theta) = 0.5$

One method: search for a good θ

$$y = (\text{●} \text{●})$$

- $\theta = [R=0.3, G=0.3, B=0.3]: P(y | \theta) = 0.22222$
- $\theta = [R=0.2, G=0.7, B=0.1]: P(y | \theta) = 0.04$
- $\theta = [R=0.4, G=0.1, B=0.5]: P(y | \theta) = 0.4$
- $\theta = [R=0.5, G=0.0, B=0.5]: P(y | \theta) = 0.5$

**This is MLE, Maximum Likelihood Estimation:
 $\operatorname{argmax}_{\theta} P(y | \theta)$**

One method: search for a good θ

$$y = (\text{blue circle} \text{ red circle})$$

- $\theta = [R=0.3, G=0.3, B=0.3]: P(y | \theta) = 0.22222$
- $\theta = [R=0.2, G=0.7, B=0.1]: P(y | \theta) = 0.04$
- $\theta = [R=0.4, G=0.1, B=0.5]: P(y | \theta) = 0.4$
- $\theta = [R=0.5, G=0.0, B=0.5]: P(y | \theta) = 0.5$

**do we really
want to rule out**



Frequentists vs. Bayesians

- Frequentists believe that there is a single true θ , and our goal is to find it (e.g. using MLE), by formulating hypotheses and testing them (p-values)
- Bayesians think it's more useful to consider a distribution over possible θ , i.e. $P(\theta)$.
Begin with a prior notion of $P(\theta)$ and update it based on data, using Bayes' rule:

$$P(\theta | y) = \frac{P(y | \theta)P(\theta)}{\int_{\theta'} P(y | \theta')P(\theta')}$$

Bayes needs a prior!

- In the marbles example, what is θ ? What's its type?

$$P(\theta | y) = \frac{P(y | \theta)P(\theta)}{\int_{\theta'} P(y | \theta')P(\theta')}$$



Bayes needs a prior!

- In the marbles example, what is θ ? What's its type?
 - a probability distribution (over colours)
- What's $P(\theta)$?

$$P(\theta | y) = \frac{P(y | \theta)P(\theta)}{\int_{\theta'} P(y | \theta')P(\theta')}$$



Bayes needs a prior!

- In the marbles example, what is θ ? What's its type?
 - a probability distribution (over colours)
- What's $P(\theta)$?
 - a probability distribution over probability distributions
- This prior encodes what kind of distributions we think are likely a priori

$$P(\theta | y) = \frac{P(y | \theta)P(\theta)}{\int_{\theta'} P(y | \theta')P(\theta')}$$



Hierarchical Bayesian Model, again

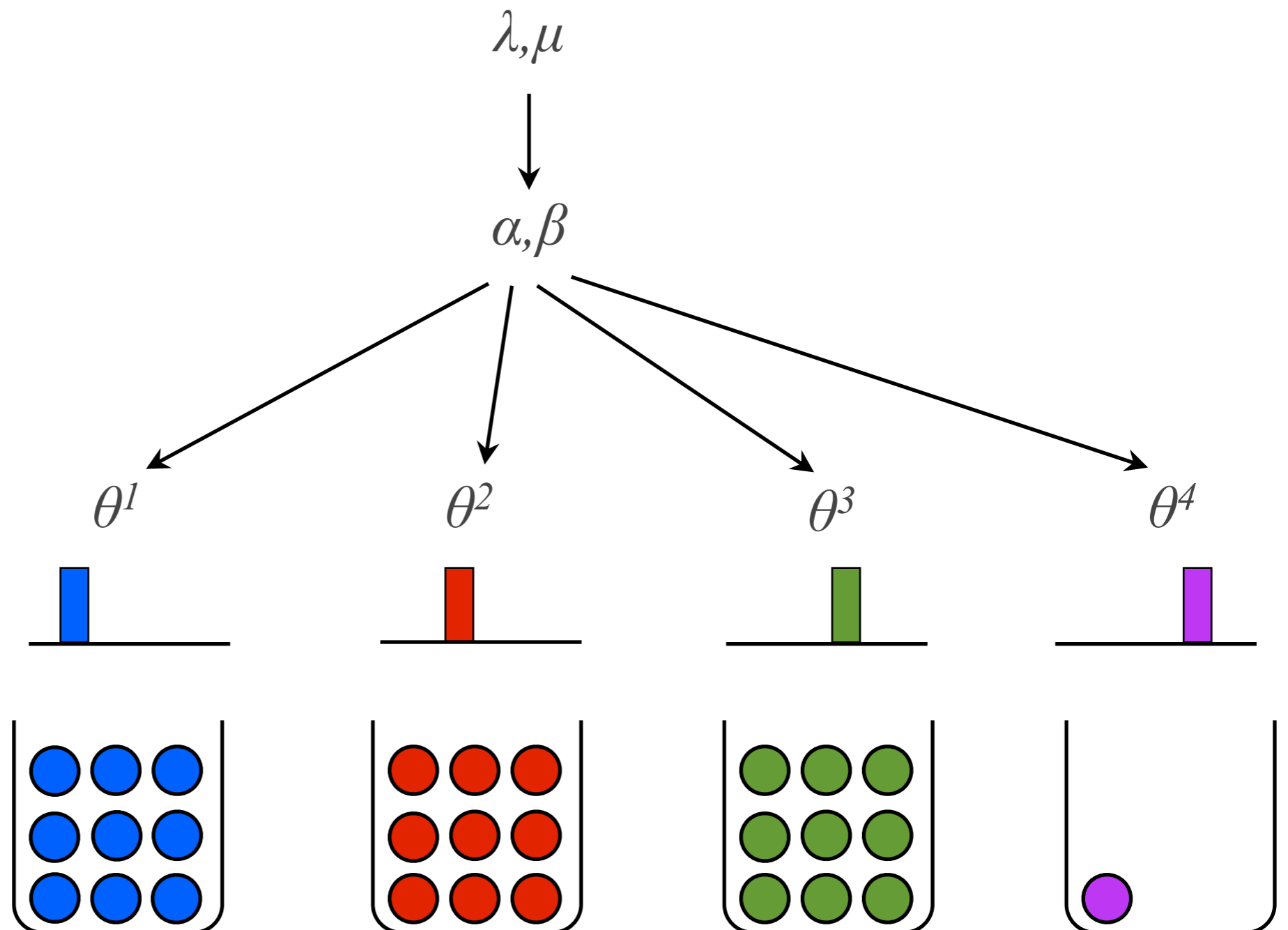
Today we are only considering levels 1 and 2.

Level 3: Prior about bags in general

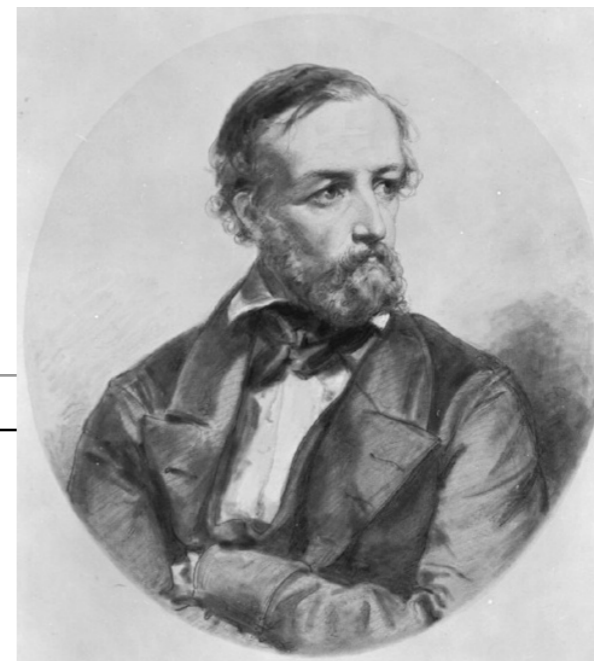
Level 2: Bags in general

Level 1: Bag proportions

Data



Bayes needs a prior!



62. Simplex Distributions

The simplex probabilities have support on the unit K -simplex for a specified K . A K -dimensional vector θ is a unit K -simplex if $\theta_k \geq 0$ for $k \in \{1, \dots, K\}$ and $\sum_{k=1}^K \theta_k = 1$.

62.1. Dirichlet Distribution

Probability Density Function

If $K \in \mathbb{N}$ and $\alpha \in (\mathbb{R}^+)^K$, then for $\theta \in K$ -simplex,

$$\text{Dirichlet}(\theta | \alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}.$$

Warning: If any of the components of θ satisfies $\theta_i = 0$ or $\theta_i = 1$, then the probability is 0 and the log probability is $-\infty$. Similarly, the distribution requires strictly positive parameters, with $\alpha_i > 0$ for each i .

Bayes needs a prior!

62. Simplex Distributions

The simplex probabilities have support on the unit K -simplex. A K -dimensional vector θ is a unit K -simplex if $\theta_k \geq 0$ and $\sum_{k=1}^K \theta_k = 1$.

62.1. Dirichlet Distribution

Probability Density Function

If $K \in \mathbb{N}$ and $\alpha \in (\mathbb{R}^+)^K$, then for $\theta \in K$ -simplex

$$\text{Dirichlet}(\theta | \alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}.$$

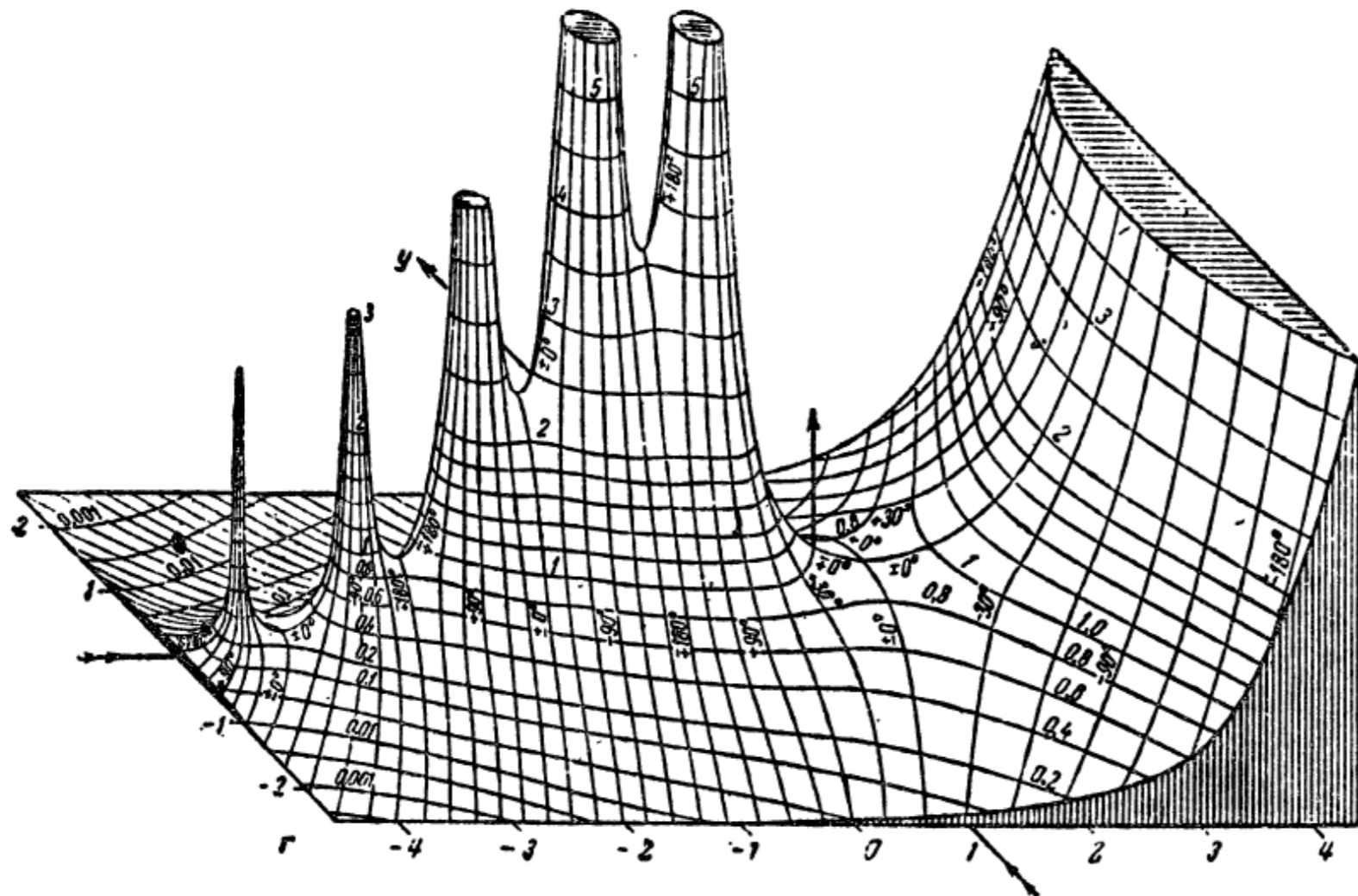
Warning: If any of the components of θ satisfies $\theta_i = 0$ or $\theta_i = 1$, then the probability is 0 and the log probability is $-\infty$. Similarly, the distribution requires strictly positive parameters, with $\alpha_i > 0$ for each i .

Gamma function:
 $\Gamma(n) = (n-1)!$
(if n is an integer)

Bayes needs a prior!

62. Simplex Distributions

The simplex probabilities have support on the n -dimensional vector θ is a unit K -simplex if $\theta_k \geq 0$



Gamma function:

$$\Gamma(n) = (n-1)! \\ \text{(if } n \text{ is an integer)}$$

$$\theta_k^{\alpha_k - 1}$$

$\theta_i = 0$ or $\theta_i = 1$, then the probability distribution requires strictly

Bayes needs a prior!

62. Simplex Distributions

The simplex probabilities have support on the unit K -simplex. A K -dimensional vector θ is a unit K -simplex if $\theta_k \geq 0$ and $\sum_{k=1}^K \theta_k = 1$.

62.1. Dirichlet Distribution

Probability Density Function

If $K \in \mathbb{N}$ and $\alpha \in (\mathbb{R}^+)^K$, then for $\theta \in K$ -simplex,

$$\text{Dirichlet}(\theta | \alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}.$$

Warning: If any of the components of θ satisfies $\theta_i = 0$ or $\theta_i = 1$, then the probability is 0 and the log probability is $-\infty$. Similarly, the distribution requires strictly positive parameters, with $\alpha_i > 0$ for each i .

This is the normalising constant

$$\int_{\theta} \text{Dir}(\theta' | \alpha) d\theta'$$

Posterior distribution of Dirichlet-Categorical

$$\begin{aligned} P(\theta|y, \alpha) &\propto P(y|\theta)P(\theta|\alpha) \\ &= \text{Cat}(y|\theta)\text{Dir}(\theta|\alpha) \\ &= \prod_{k=1}^K \theta_k^{N_k} \prod_{k=1}^K \theta_k^{\alpha_k-1} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i)} \\ &\propto \prod_{k=1}^K \theta_k^{N_k} \theta_k^{\alpha_k-1} \\ &= \prod_{k=1}^K \theta_k^{N_k+\alpha_k-1} \end{aligned}$$

where N_k is the number of items of category k in y .

Posterior distribution of Dirichlet-Categorical

$$\begin{aligned} P(\theta|y, \alpha) &\propto P(y|\theta)P(\theta|\alpha) \\ &= \text{Cat}(y|\theta)\text{Dir}(\theta|\alpha) \\ &= \prod_{k=1}^K \theta_k^{N_k} \prod_{k=1}^K \theta_k^{\alpha_k-1} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i)} \\ &\propto \prod_{k=1}^K \theta_k^{N_k} \theta_k^{\alpha_k-1} \\ &= \prod_{k=1}^K \theta_k^{N_k+\alpha_k-1} \end{aligned}$$

where N_k is the number of items of category k in y .

Posterior is also a Dirichlet: $P(\theta|y, \alpha) = \text{Dir}(\alpha')$,

where $\alpha' = \alpha + y$; i.e. $\alpha'_k = \alpha_k + y_k$.

Posterior distribution of Dirichlet-Categorical

Posterior is also a Dirichlet: $P(\theta|y, \alpha) = \text{Dir}(\alpha')$,

where $\alpha' = \alpha + y$; i.e. $\alpha'_k = \alpha_k + y_k$.

This is called *conjugacy*: Prior and posterior are same type of distribution, given a certain type of likelihood.

(Dirichlet is the conjugate prior for the categorical and the multinomial.)

Posterior Dirichlet is parameterised by counts in data y plus “pseudocounts” α .

Large values of α (or $\alpha\beta$) can thus overwhelm the data; these are “stronger” priors.

What if we care about prediction, not θ ?

What's the colour of the *next* marble going to be?

We can calculate the *predictive posterior*, while marginalising over θ : this takes all possible θ into account

$$P(y = \text{blue} \mid D, \alpha)$$

$$p(y = k \mid D, \alpha) = \int_{\theta} P(y = k \mid \theta) P(\theta \mid D, \alpha) d\theta$$

Predictive Posterior Probability

$$\begin{aligned} p(y = k|D, \alpha) &= \int_{\theta} P(y = k|\theta)P(\theta|D, \alpha)d\theta \\ &= \int_{\theta} \theta_k \prod_k \theta^{N_k + \alpha_k - 1} \frac{\Gamma(\sum_{k=1}^K N_k + \alpha_k)}{\prod_{k=1}^K \Gamma(N_k + \alpha_k)} d\theta \\ &\dots \\ &= \frac{N_k + \alpha_k}{\sum_{k' \in K} N_{k'} + \alpha_{k'}} \end{aligned}$$

where N_k is the number (count) of items in category k in D .

If α prior is symmetric, such that $\alpha_k = \alpha$ for all k ,

$$p(y = k|D, \alpha) = \frac{N_k + \alpha}{N + K\alpha}$$

Summary

- Dirichlet-Multinomial and Dirichlet-Categorical distributions may look complex
- But certain properties are simple:

- Posterior is a Dirichlet updated by counts in data:

$$P(\theta | y, \alpha) \propto P(y | \theta)P(\theta | \alpha) \sim \text{Dir}(\alpha + y)$$

- Predictive posterior is normalised counts + prior

$$P(k | y, \alpha) = \frac{y_k + \alpha_k}{\sum_{k'} y_{k'} + \alpha_{k'}}$$

- The effect of prior α : "pseudocounts" of data expected a priori

Extra: Non-parametric distributions

- Parametric distributions (e.g. Dirichlet, Multinomial) require specifying the number of categories K .
- If we don't know, or don't want to specify K , we can use *non*-parametric models
- e.g. *Dirichlet Process*(α, H):
 - α is concentration parameter, H is base distribution
 - Predictive posterior:

$$p(y_{new} = k_{new}) = \frac{\alpha}{\alpha + n - 1} \quad p(y_{new} = k_{exists}) = \frac{y_k}{\alpha + n + 1}$$

The Learnability of Abstract Syntactic Principles, Perfors, Tenenbaum & Regier (2009)

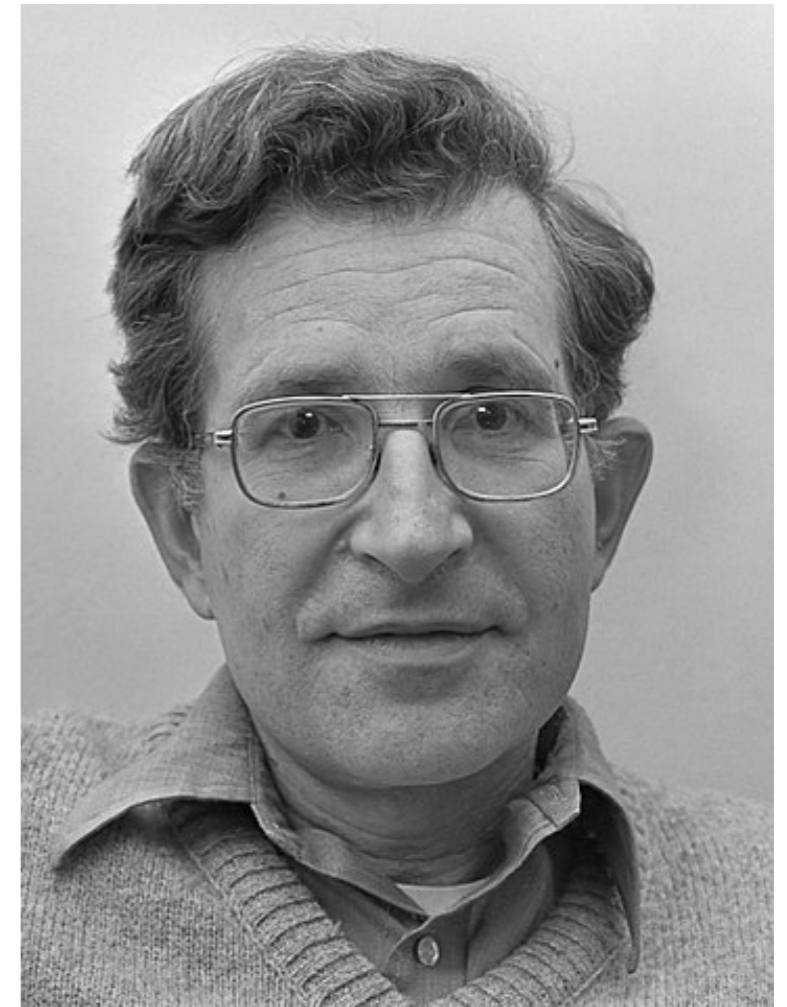


By Welleschik - http://commons.wikimedia.org/wiki/File:Lille_Meert2.JPG, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=19506073>

(There will be zero Dirichlets in this part!)

"Poverty of Stimulus" argument (Chomsky et alia)

1. Language is complex (see example)
2. Children do not hear enough instances of [complex example] to possibly learn it
3. Ergo children must have innate linguistic knowledge about [complex example]



Complex example: Aux-raising in English questions

"The bear is eating the fish's breakfast"

"Is the bear  eating the fish's breakfast?"

Complex example: Aux-raising in English questions

"The students who are in the classroom are still awake"

"Are the students who in the classroom are still awake?"



"Are the students who are in the classroom still awake?"



Complex example: Aux-raising in English questions

"The students who are in the classroom are still awake"

Linear rule: move the first auxiliary verb

"Are the students who in the classroom are still awake?"



"Are the students who are in the classroom still awake?"



Hierarchical rule: move the aux in main clause

Aux raising in other languages

German:

"Sind die Studenten, die in der Vorlesung sind, wach?"



Finnish (thanks, mom):

Poika, joka on onnellinen, on leikkimässä

(The boy who is happy is playing)

"Onko poika, joka on onnellinen, leikkimässä?"



Language "acquisition device"

- Even if this was an English-only phenomena (which it isn't):
- Any infant who can learn any other language can also learn English, so:
- (Learning) the aux-raising rule has to be part of the linguistic capabilities of all infants.
- But: English child-directed speech contains only ~0.05% complex interrogatives - can this be enough to learn?

Counter "Poverty of Stimulus" (Perfors et al)

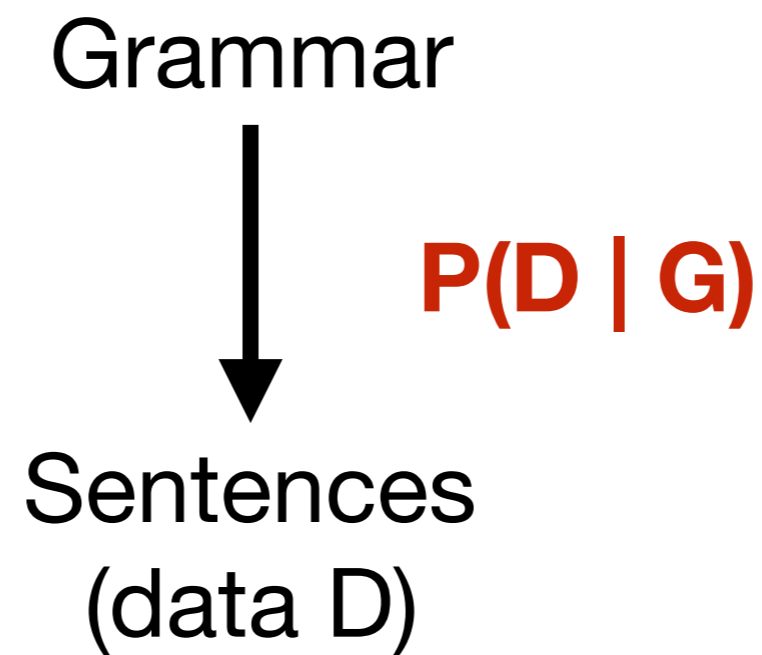
1. Language is complex
2. Sure - and children learn it as a system
3. Arguing from a single (type of) example is silly
4. Overhypotheses!



Counter "Poverty of Stimulus" (Perfors et al)

To put this point another way, while it may be sensible to ask what a rational learner can infer about language as a whole without any language-specific biases, it is less sensible to ask what a rational learner can infer about any single specific linguistic rule (such as auxiliary-fronting). The need to acquire a whole system of linguistic rules together imposes constraints among the rules, so that an *a priori* unbiased learner may acquire constraints that are based on the other linguistic rules it must learn at the same time.

Hierarchical model of linguistic structure



Hierarchical model of linguistic structure

Grammar $P(\mathbf{G} | \mathbf{D}) \propto P(\mathbf{D} | \mathbf{G}) P(\mathbf{G})$

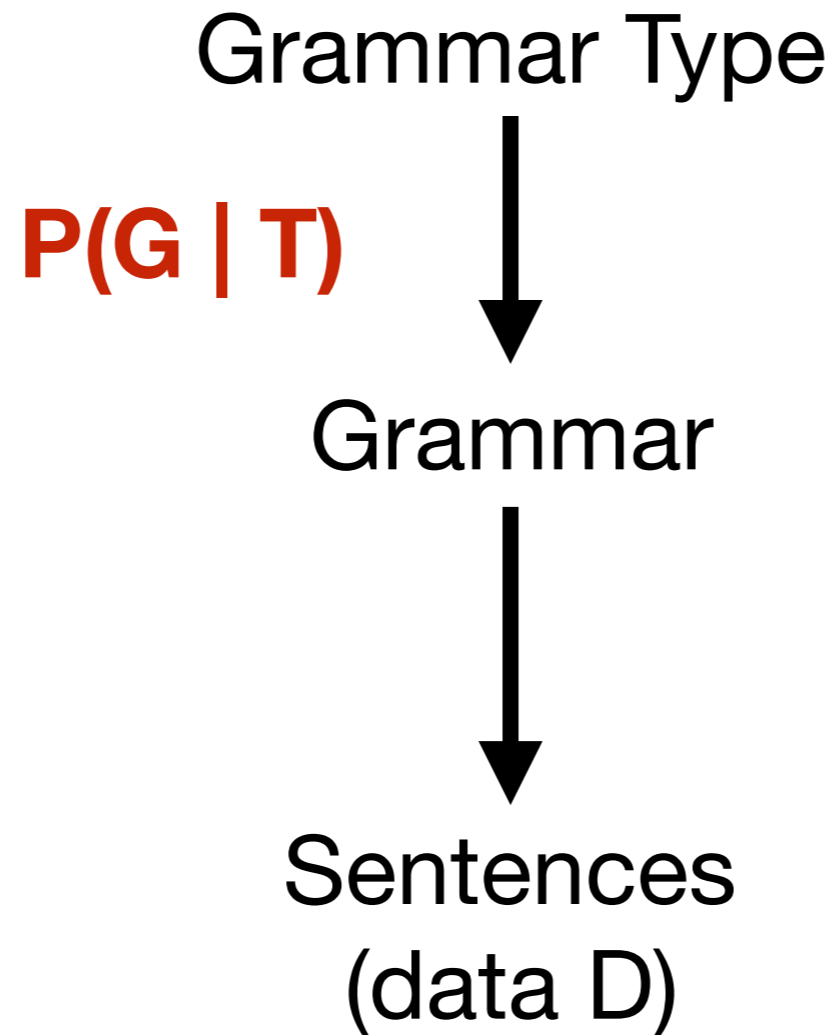


Sentences
(data \mathbf{D})

Need to define:

- prior over \mathbf{G}
- space of grammars

Hierarchical model of linguistic structure



$$P(T, G | D) \propto P(D | G) P(G | T) P(T)$$

Space of Grammar Types

- no grammar: one state that allows all possible sentences
- flat grammar: memorises all sentences in the corpus
- regular grammars: represent sentences linearly
(different numbers of non-terminals, rules)
- context-free grammars: represent sentences hierarchically

All sentences represented as sequences of syntactic categories.

All grammars are *probabilistic*: assign a probability to a sentence.

Hierarchical model of linguistic structure

$P(T) \sim \text{uniform}$

Grammar Type



Grammar



Sentences
(data D)

**$P(G|T)$ prefers
simpler grammars
of a given type**

$P(T, G|D) \propto P(D|G) P(G|T) P(T)$

Approach: find best grammar of each type, and evaluate its posterior probability, given some plausible data D.

Results by sentence frequency

Table 2

Log prior, likelihood, and posterior probabilities of each hand-designed grammar for each level of evidence. Because numbers are negative, smaller absolute values correspond to higher probability. If two grammars have log probabilities that differ by n , their actual probabilities differ by e^n ; thus, the best hierarchical phrase-structure grammar CFG-L is e^{101} ($\sim 10^{43}$) times more probable than the best linear grammar REG-M. Bold values indicate the highest posterior score at each level.

Corpus	Probability	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Level 1	Prior	-99	-148	-124	-117	-94	-155	-192
	Likelihood	-17	-20	-19	-21	-36	-27	-27
	Posterior	-116	-168	-143	-138	-130	-182	-219
Level 2	Prior	-630	-456	-442	-411	-201	-357	-440
	Likelihood	-134	-147	-157	-162	-275	-194	-177
	Posterior	-764	-603	-599	-573	-476	-551	-617
Level 3	Prior	-1198	-663	-614	-529	-211	-454	-593
	Likelihood	-282	-323	-333	-346	-553	-402	-377
	Posterior	-1480	-986	-947	-875	-764	856	-970
Level 4	Prior	-5839	-1550	-1134	-850	-234	-652	-1011
	Likelihood	-1498	-1761	-1918	-2042	-3104	-2078	-1956
	Posterior	-7337	-3311	-3052	-2892	-3338	-2730	-2967
Level 5	Prior	-10,610	-1962	-1321	-956	-244	-732	-1228
	Likelihood	-2856	-3376	-3584	-3816	-5790	-3917	-3703
	Posterior	-13,466	-5338	-4905	-4772	-6034	-4649	-4931
Level 6	Prior	-67,612	-5231	-2083	-1390	-257	-827	-1567
	Likelihood	-18,118	-24,454	-25,696	-27,123	-40,108	-27,312	-26,111
	Posterior	-85,730	-29,685	-27,779	-28,513	-40,365	-28,139	-27,678

Data from higher levels include more infrequent sentence types

Complex Aux-questions

Table 7

Ability of each grammar to parse specific sentences. The complex declarative sentence “Eagles that are alive can fly” occurs in the Adam corpus. Only the context-free grammars can parse the corresponding complex interrogative sentence.

Type	In input?	Example	Can parse?						
			FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Decl Simple	Y	Eagles can fly. (n aux vi)	Y	Y	Y	Y	Y	Y	Y
Int Simple	Y	Can eagles fly? (aux n vi)	Y	Y	Y	Y	Y	Y	Y
Decl Complex	Y	Eagles that are alive can fly. (n comp aux adj aux vi)	Y	Y	Y	Y	Y	Y	Y
Int Complex	N	Can eagles that are alive fly? (aux n comp aux adj vi)	N	N	N	N	Y	Y	Y
Int Complex	N	*Are eagles that alive can fly? (aux n comp adj aux vi)	N	N	N	N	Y	N	N

- 1-state can parse everything (by construction)
- Only CFGs parse the correct form of the question and fail to parse the incorrect form

Summary

- A learner with the representational capacity for both flat (regular) and hierarchical (context-free) grammars can infer, from child-directed speech data, that hierarchical structures capture the data better.
- Such a grammar can also correctly generalise to new structures, such as complex questions.
- No initial bias towards hierarchy or particular linguistic structures is necessary.

Next:

Words as high-dimensional objects (not discrete atomic categories), capturing semantics, syntax, phonology, etc.

- Is this representation cognitively realistic?
- How can we discover these representations?

