

Language and Statistics II

Lecture 21: Going Bayesian

Being Bayesian in a Nutshell

- Probabilities are for managing uncertainty.
- Being Bayesian means managing even more uncertainty.
- Example to motivate:

- Maximum likelihood estimation $\max_{\theta} p_{\theta}(\text{data})$

- Maximum *a posteriori* estimation $\max_{\theta} p_{\theta}(\text{data})p_{\alpha}(\theta)$

- More Bayesian

$$p(\theta \mid \alpha, \text{data}) = \frac{p_{\theta}(\text{data})p_{\alpha}(\theta)}{\int p_{\theta'}(\text{data})p_{\alpha}(\theta')d\theta'}$$

$$\max_{\alpha} \int p_{\theta}(\text{data})p_{\alpha}(\theta)d\theta$$

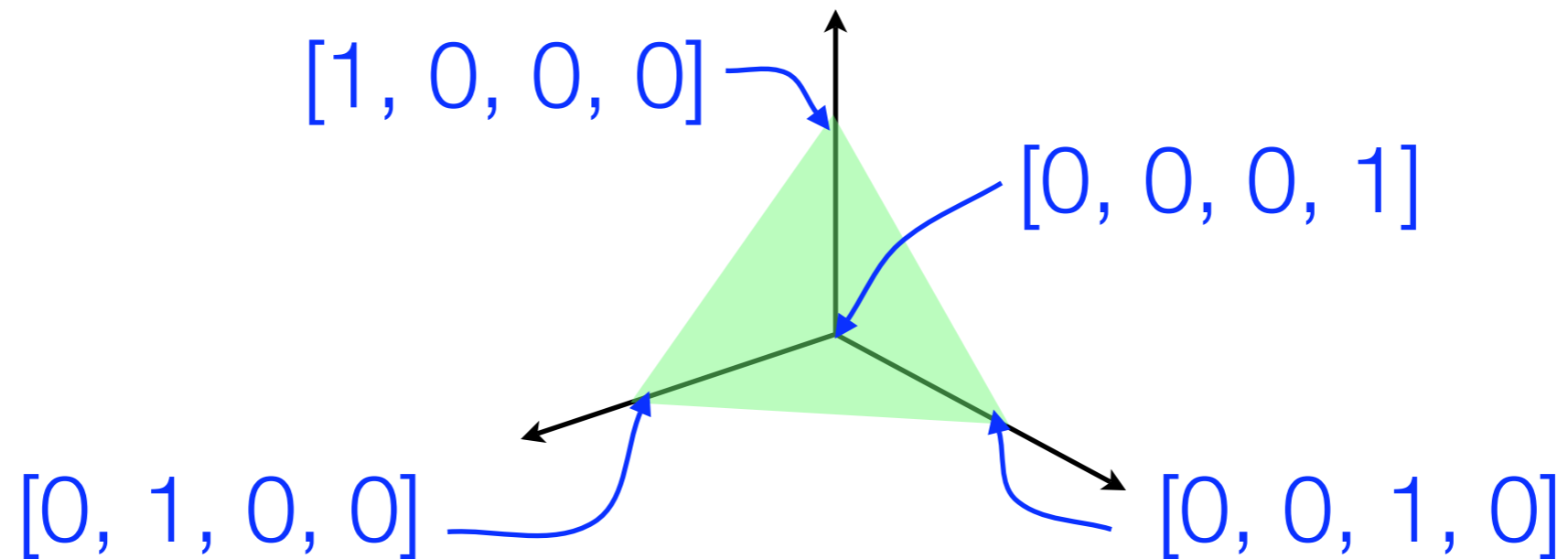
?

Multinomial Distributions

- We've seen these many times.
 - Unigram model
 - Bigram model (collection of multinomials)
 - HMM (collection)
 - PCFG
- Most Bayesian NLP research focuses on multinomial-based models.
 - Future work: “going Bayesian with log-linear models”

Distributions over Multinomials

- You can think of a multinomial distribution over d events as a point in the $(d-1)$ simplex.



- To randomly pick a point in this space, we need a **continuous** distribution over the simplex.

Dirichlet Distribution

- A distribution over the d -event probability simplex.

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^d \theta_i^{\alpha_i - 1}$$

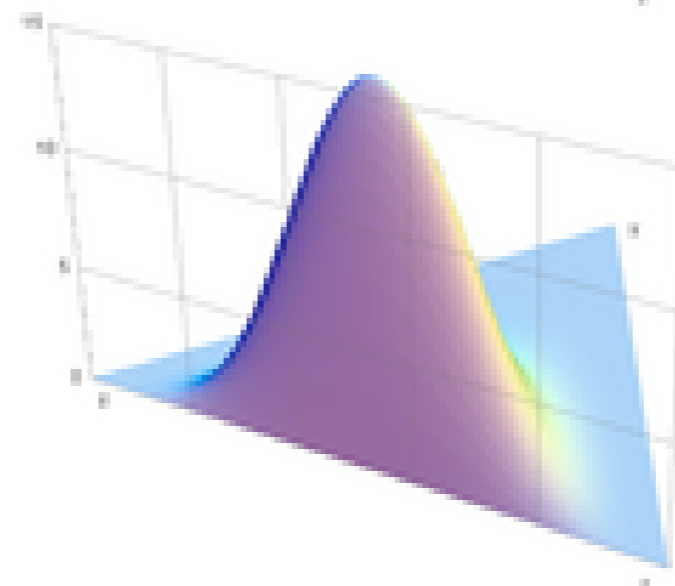
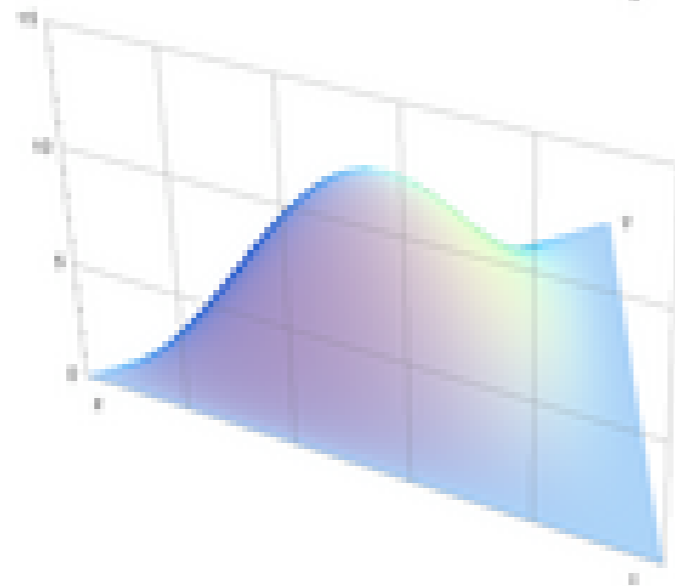
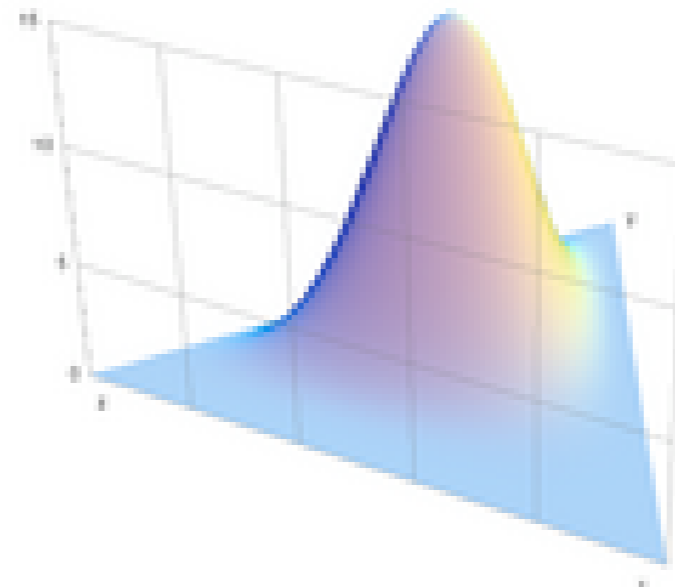
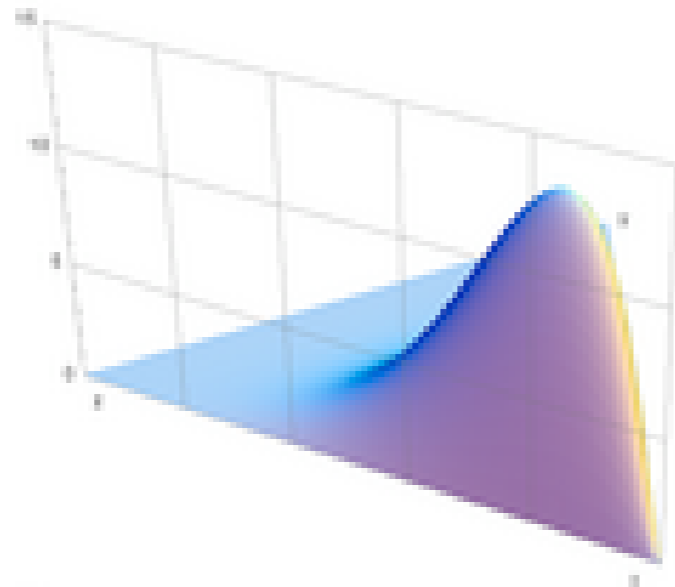
- Parameters: $\boldsymbol{\alpha}$, a vector of positive values.

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^d \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^d \alpha_i\right)}$$

- Beta function:

- Gamma function (generalized factorial):

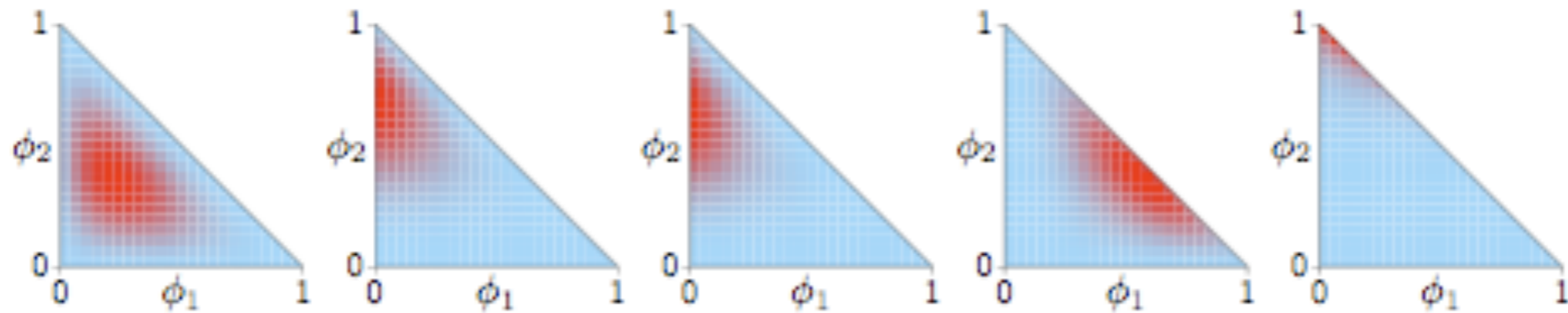
$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$



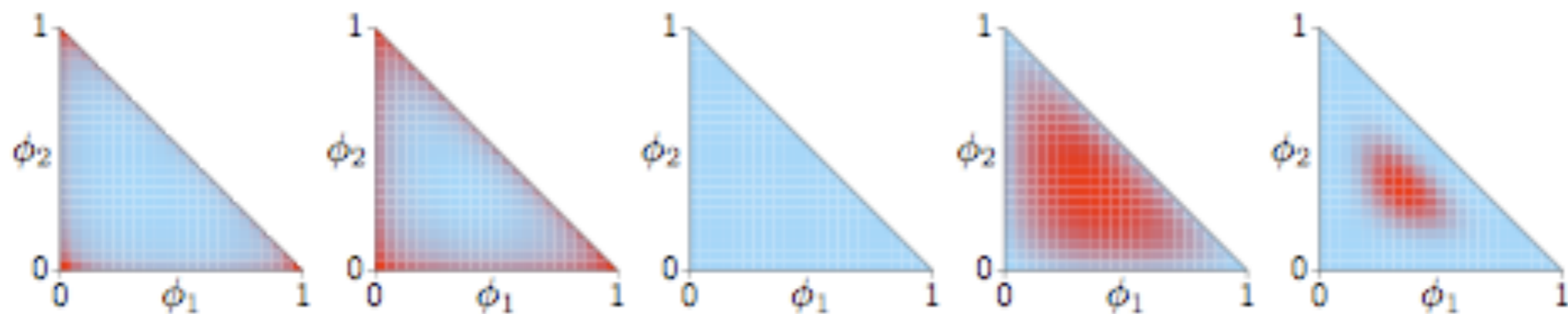
Dirichlet, $d=3$
(various parameter settings)

from answers.com

Different means:



Different variances:



Dirichlet, $d=3$
(different “means” and
“variances”)

from Liang and Klein, 2007

MAP with a Dirichlet

- Recall that we can use a prior to “smooth” an MLE estimate.

Mixture of Unigrams

- The generative story for a classical document-clustering model would be something like this (Nigam et al., 2000):
- For $i = 1 \dots M$:
 - Draw a document length N_i from some distribution.
 - Draw a topic z for the document from a multinomial.
 - For $j = 1 \dots N$:
 - Draw word w_{ij} from the multinomial θ_z .
- Nigam et al. learned this using EM.

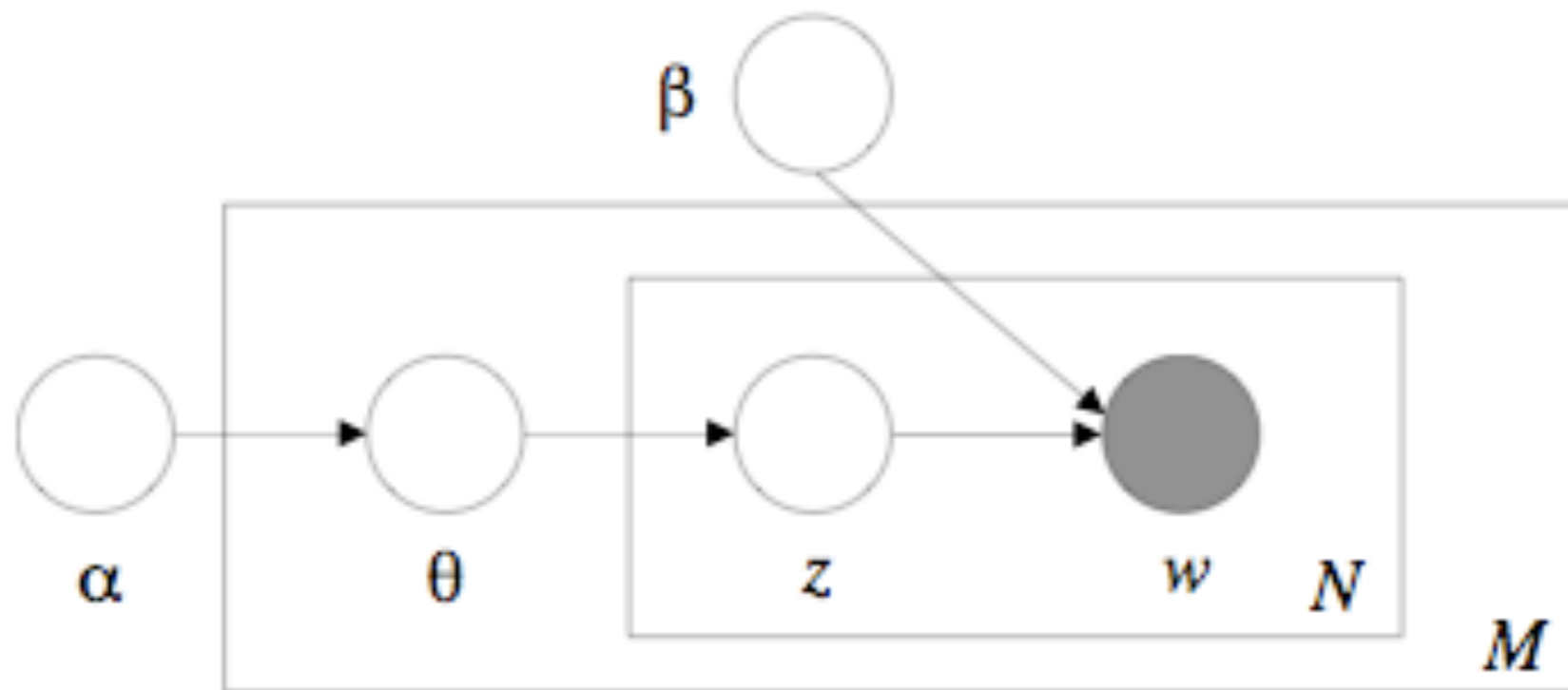
The Rest of the Lecture

- Building more interesting models
 - Topic models (Blei, Ng, and Jordan, 2003)
 - POS tagging (Goldwater and Griffiths, 2007)
- Inference
 - Markov Chain Monte Carlo example: Gibbs sampling
 - Variational methods example: Variational Bayes for structured models

Latent Dirichlet Allocation (Blei et al., 2003)

- Given: M (# documents), α (prior over topic distributions), β (per-topic unigram distributions)
- For $i = 1 \dots M$:
 - Choose N , the number of words (Poisson or whatever).
 - Choose a distribution θ_i over topics (Dirichlet(α)).
 - For $j = 1 \dots N$:
 - Choose a topic z_{ij} according to θ_i .
 - Choose a word w_{ij} according to $\beta[z_{ij}, *]$.

Graphical Model



A Bayesian HMM for POS Tagging (Goldwater and Griffiths, 2007)

- Given: α (prior over tag trigram distributions), β (prior over emission distributions)
- Pick a trigram tag distribution γ according to α
- Pick an emission distribution η according to β
- Sample from the HMM defined by (γ, η)

What might learning mean?

- Estimating the (hyper)parameters of the model.
 - In a classic HMM: EM training.
 - In LDA: estimate α and β .
 - Called “empirical Bayes,” “type 2 maximum likelihood,” “generalized maximum likelihood,” or “evidence approximation.”
- Inferring a **distribution** over the parameters of the model.
 - Often approximate, called “variational Bayes.”
- Inferring a particular hidden variable of interest on some data.
 - In a classic HMM: decoding.
 - In LDA: guess the topic z_i for each document
 - Typically used for evaluation
- Inferring a **distribution** over the hidden variables.

We've seen these ideas before!

- EM:
 - E step: infer a distribution over hidden variables
 - M step: estimate parameters
- Bayesian view: Hidden variables and parameters are the same kinds of things (Things We Don't Know), and we should infer **distributions** over both, rather than **guess**.

Computationally, this is intractable.

A Motto for Bayesian NLPers

*State mathematically what you wish you could do,
then approximate whatever's intractable.*

Two Main Tricks for Bayesian Inference

- Randomization
 - Use random sampling to approximate distributions.
 - Example: Gibbs sampling.
- Variational methods
 - Define a simpler, more factored model and fit that model to the true one.
 - Turns an intractable calculation into an optimization problem!
 - Example: mean-field approximation