Variational Inference

Material adapted from David Blei

University of Maryland

Introduction

Variational Inference

- Inferring hidden variables

Variational Inference

- Inferring hidden variables
- Unlike MCMC:
  - Deterministic
  - Easy to gauge convergence
  - Requires dozens of iterations
- Doesn't require conjugacy
- Slightly hairier math

Conjugacy: the joint distribution has special properties—we'll talk more about what that means in a second if you haven't heard that term before.

Setup

- $\vec{x} = x_{1:n}$ observations
- $\vec{z} = z_{1:m}$ hidden variables
- α fixed parameters
- Want the posterior distribution

$$p(z|x,\alpha) = \frac{p(z,x|\alpha)}{\int_z p(z,x|\alpha)} \tag{1}$$

Motivation

- Can't compute posterior for many interesting models

GMM (finite)

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$
2. For each observation $i = 1 \dots n$:
    2.1 Draw $z_i \sim \text{Mult}(\pi)$
    2.2 Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_0^2)$

Motivation

- Can't compute posterior for many interesting models

GMM (finite)

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$
2. For each observation $i = 1 \ldots n$:
    2.1 Draw $z_i \sim \text{Mult}(\pi)$
    2.2 Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_0^2)$

- Posterior is intractable for large n, and we might want to add priors

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})} \tag{2}$$

Consider all means

Motivation

- Can't compute posterior for many interesting models

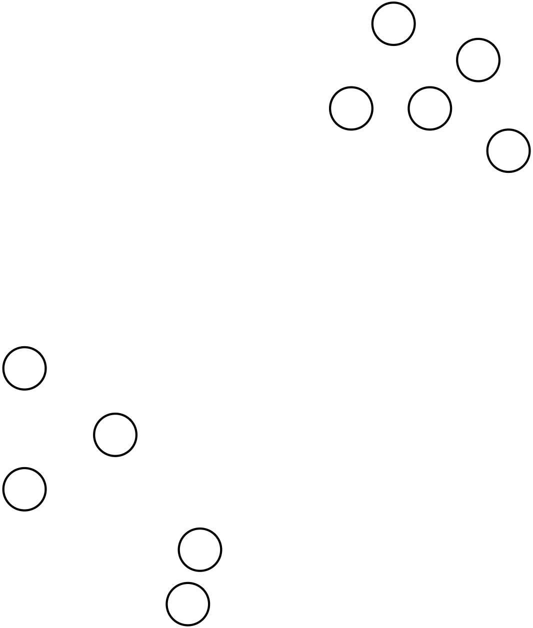
GMM (finite)

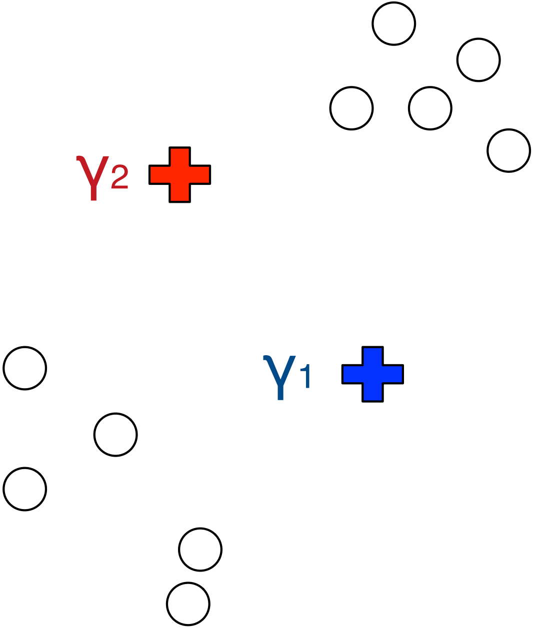
1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$
2. For each observation $i = 1 \ldots n$:
   2.1 Draw $z_i \sim \text{Mult}(\pi)$
   2.2 Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_0^2)$

- Posterior is intractable for large n, and we might want to add priors

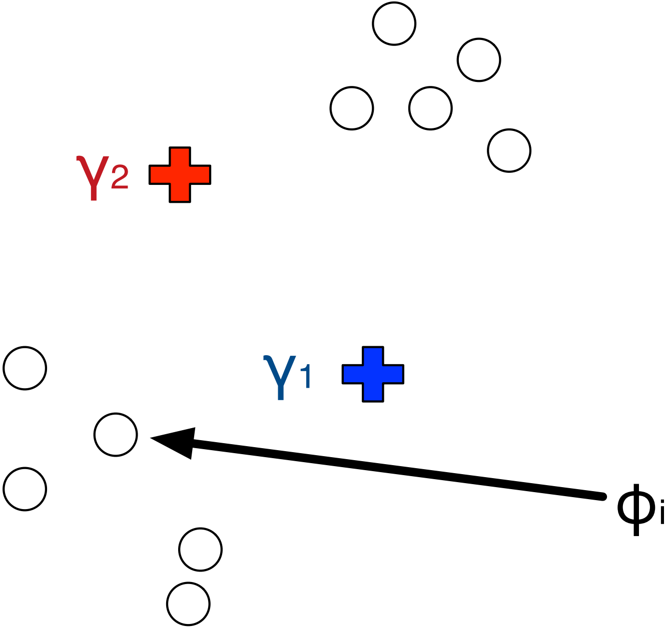
$$p(\mu_{1:K}, z_{1:n} | x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})} \tag{2}$$

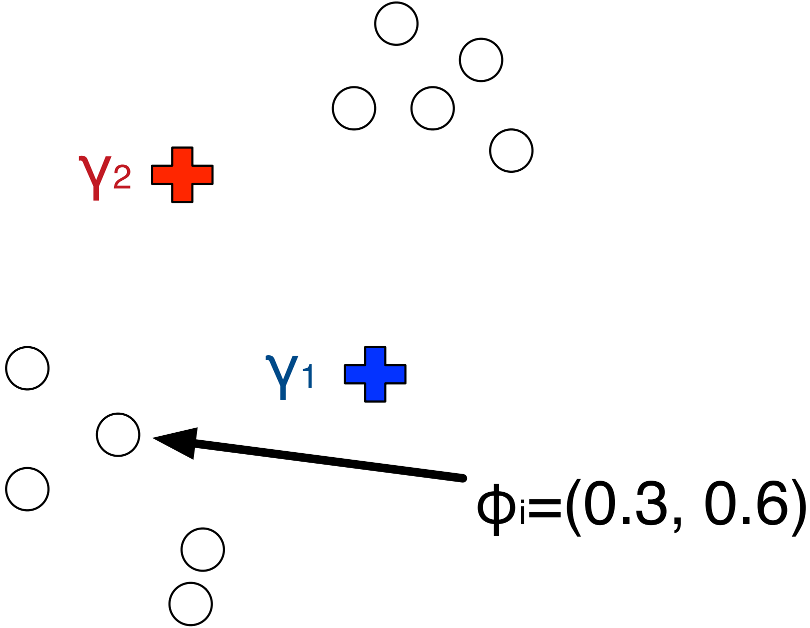
Consider all assignments



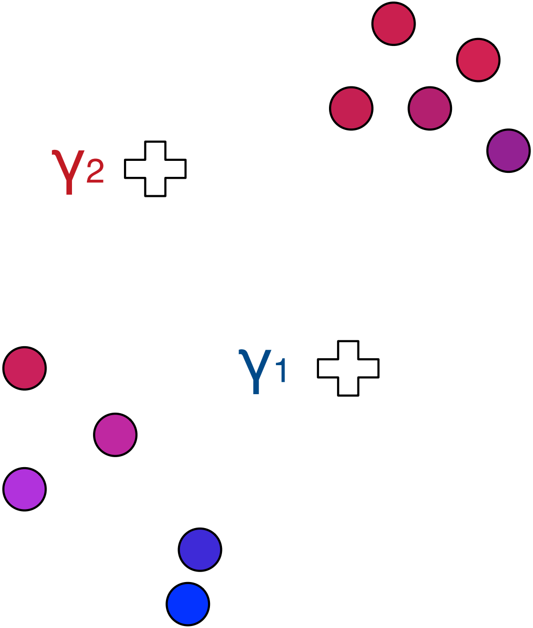


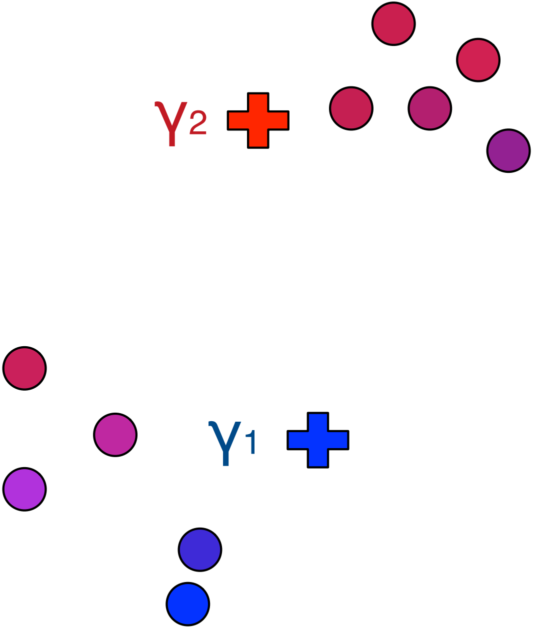
γ2
γ1


γ2
γ1
φi



γ2
γ1
ϕi=(0.3, 0.6)


γ2
γ1


γ2
γ1

Main Idea

- We create a **variational distribution** over the latent variables

$$q(z_{1:m} \mid \nu) \tag{3}$$

- Find the settings of ν so that q is close to the posterior
- If $q == p$, then this is vanilla EM

What does it mean for distributions to be close?

- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\mathrm{KL}(q||p) \equiv \mathbb{E}_q\left[\log \frac{q(Z)}{p(Z|x)}\right] \tag{4}$$

What does it mean for distributions to be close?

- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\mathrm{KL}(q||p) \equiv \mathbb{E}_q\left[\log \frac{q(Z)}{p(Z|x)}\right] \tag{4}$$

- Characterizing KL divergence
  - If q and p are high, we're happy
  - If q is high but p isn't, we pay a price
  - If q is low, we don't care
  - If $\mathrm{KL} = 0$, then distribution are equal

What does it mean for distributions to be close?

- We measure the closeness of distributions using Kullback-Leibler Divergence

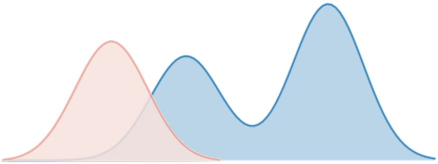
$$\mathrm{KL}(q\,||\,p) \equiv \mathbb{E}_q\left[\log \frac{q(Z)}{p(Z\,|\,x)}\right] \tag{4}$$

- Characterizing KL divergence
  - If q and p are high, we're happy
  - If q is high but p isn't, we pay a price
  - If q is low, we don't care
  - If $\mathrm{KL} = 0$, then distribution are equal

This behavior is often called "mode splitting": we want **a** good solution, not every solution.

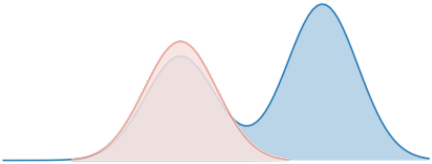
KL(Q‖P)

Bad



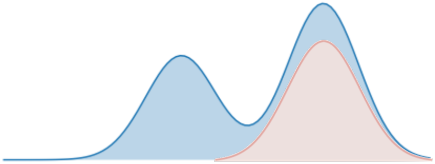
KL(Q‖P)

Good

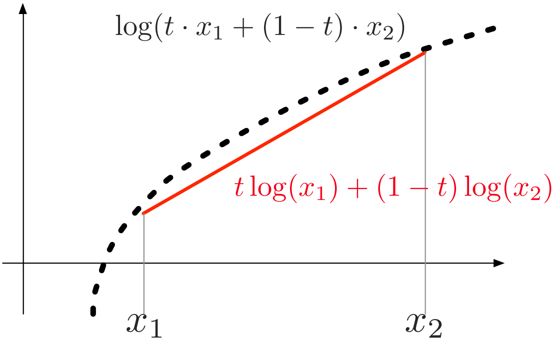


KL(Q‖P)

Equally Good



Jensen's Inequality: Concave Functions and Expectations



When f is concave

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

If you haven't seen this before, spend fifteen minutes to convince yourself that it's true

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$\log p(x) = \log\left[\int_z p(x,z)\right]$$

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log\left[\int_z p(x,z)\right] \\
&= \log\left[\int_z p(x,z)\frac{q(z)}{q(z)}\right]
\end{aligned}
$$

Add a term that is equal to one

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log\left[\int_z p(x,z)\right] \\
&= \log\left[\int_z p(x,z)\frac{q(z)}{q(z)}\right] \\
&= \log\left[\mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]\right]
\end{aligned}
$$

Take the numerator to create an expectation

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log\left[\int_z p(x,z)\right] \\
&= \log\left[\int_z p(x,z)\frac{q(z)}{q(z)}\right] \\
&= \log\left[\mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]\right] \\
&\geq \mathbb{E}_q\left[\log p(x,z)\right] - \mathbb{E}_q\left[\log q(z)\right]
\end{aligned}
$$

Apply Jensen's equality and use log difference

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log\left[\int_z p(x,z)\right] \\
&= \log\left[\int_z p(x,z)\frac{q(z)}{q(z)}\right] \\
&= \log\left[\mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]\right] \\
&\geq \mathbb{E}_q\left[\log p(x,z)\right] - \mathbb{E}_q\left[\log q(z)\right]
\end{aligned}
$$

- Fun side effect: Entropy
- Maximizing the ELBO gives as tight a bound on on log probability

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log\left[\int_z p(x,z)\right] \\
&= \log\left[\int_z p(x,z)\frac{q(z)}{q(z)}\right] \\
&= \log\left[\mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]\right] \\
&\geq \mathbb{E}_q\left[\log p(x,z)\right] - \mathbb{E}_q\left[\log q(z)\right]
\end{aligned}
$$

- Fun side effect: Entropy
- Maximizing the ELBO gives as tight a bound on on log probability

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log \left[\int_z p(x,z) \right] \\
&= \log \left[\int_z p(x,z) \frac{q(z)}{q(z)} \right] \\
&= \log \left[\mathbb{E}_q \left[\frac{p(x,z)}{q(z)} \right] \right] \\
&\geq \mathbb{E}_q \left[\log p(x,z) \right] - \mathbb{E}_q \left[\log q(z) \right]
\end{aligned}
$$

- Fun side effect: Entropy
- Maximizing the ELBO gives as tight a bound on on log probability

Relation to KL Divergence

- Conditional probability definition

$$p(z|x) = \frac{p(z,x)}{p(x)} \tag{5}$$

Relation to KL Divergence

- Conditional probability definition

$$p(z|x) = \frac{p(z,x)}{p(x)} \tag{5}$$

- Plug into KL divergence

$$\mathrm{KL}(q(z)\,||\,p(z|x)) = \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x)}\right]$$

Relation to KL Divergence

- Conditional probability definition

$$p(z|x) = \frac{p(z,x)}{p(x)} \tag{5}$$

- Plug into KL divergence

$$\mathrm{KL}(q(z)||p(z|x)) = \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x)}\right]$$

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{5}$$

- Plug into KL divergence

$$\begin{aligned} \mathrm{KL}(q(z) \,||\, p(z \mid x)) &= \mathbb{E}_q \left[\log \frac{q(z)}{p(z \mid x)} \right] \\ &= \mathbb{E}_q \left[\log q(z) \right] - \mathbb{E}_q \left[\log p(z \mid x) \right] \end{aligned}$$

Break quotient into difference

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z,x)}{p(x)} \tag{5}$$

- Plug into KL divergence

$$\begin{aligned}
\mathrm{KL}(q(z) \,||\, p(z \mid x)) =& \mathbb{E}_q \left[\log \frac{q(z)}{p(z \mid x)} \right] \\
=& \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z \mid x)] \\
=& \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z,x)] + \mathbb{E}_q [\log p(x)]
\end{aligned}$$

Apply definition of conditional probability

Relation to KL Divergence

- Conditional probability definition

$$p(z|x) = \frac{p(z,x)}{p(x)} \tag{5}$$

- Plug into KL divergence

$$\begin{aligned}
\mathrm{KL}(q(z)\,||\,p(z|x)) &= \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x)}\right] \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z|x)] \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z,x)] + \mathbb{E}_q[\log p(x)] \\
&= -\big(\mathbb{E}_q[\log p(z,x)] - \mathbb{E}_q[\log q(z)]\big) + \log p(x)
\end{aligned}$$

Reorganize terms

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z,x)}{p(x)} \tag{5}$$

- Plug into KL divergence

$$\begin{aligned}
\mathrm{KL}(q(z) \,||\, p(z \mid x)) =& \mathbb{E}_q \left[\log \frac{q(z)}{p(z \mid x)} \right] \\
=& \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z \mid x)] \\
=& \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z,x)] + \mathbb{E}_q [\log p(x)] \\
=& -\big(\mathbb{E}_q [\log p(z,x)] - \mathbb{E}_q [\log q(z)]\big) + \log p(x)
\end{aligned}$$

- Negative of ELBO (plus constant); minimizing KL divergence is the same as maximizing ELBO

colourized by pastincolour.com

SOLVAY CONFERENCE 1927

A. PICARD E. HENRIOT P. EHRENFEST Ed. HERSEN Th. DE DONDER E. SCHRÖDINGER E. VERSCHAFFELT W.PAULI W. HEISENBERG R.H FOWLER L. BRILLOUIN

P. DEBYE M. KNUDSEN W.L. BRAGG H.A. KRAMERS P.A.M. DIRAC A.H. COMPTON L. de BROGLIE M. BORN N. BOHR

I. LANGMUIR M. PLANCK Mme CURIE H.A.LORENTZ A. EINSTEIN P. LANGEVIN Ch.E. GUYE C.T.R. WILSON O.W. RICHARDSON

Absents : Sir W.H. BRAGG, H. DESLANDRES et E. VAN AUBEL

210 LE MAGNÉTISME.

parallèlement ou en sens opposés. On peut dire que l'énergie, ou la fonction d'Hamilton, est dans ce cas

$$\mathcal{H} = -A \sum_k (\sigma_k, \sigma_{k+1}),$$

où σ_k représente le vecteur de pivotement du $k^{\text{ième}}$ électron. On a $(\sigma, \sigma_{k+1}) = \pm 1$ suivant que les pivotements sont parallèles ou opposés. Ising n'a rien trouvé, dans les propriétés de ce modèle, qui corresponde au ferromagnétisme, mais une dépendance entre

colourized by pastincolour.com

SOLVAY CONFERENCE 1927

A. PICARD E. HENRIOT P. EHRENFEST Ed. HERSEN Th. DE DONDER E. SCHRÖDINGER E. VERSCHAFFELT W.PAULI W. HEISENBERG R.H FOWLER L. BRILLOUIN

P. DEBYE M. KNUDSEN W.L. BRAGG H.A. KRAMERS P.A.M. DIRAC A.H. COMPTON L. de BROGLIE M. BORN N. BOHR

I. LANGMUIR M. PLANCK Mme CURIE H.A.LORENTZ A. EINSTEIN P. LANGEVIN Ch.E. GUYE C.T.R. WILSON O W. RICHARDSON

Absents : Sir W.H. BRAGG, H. DESLANDRES et E. VAN AUBEL

Mean field variational inference

- Assume that your variational distribution factorizes

$$q(z_1, \dots, z_m) = \prod_{j=1}^{m} q(z_j) \tag{6}$$

- You may want to group some hidden variables together
- Does not contain the true posterior because hidden variables are dependent

Gates

Tom Minka
Microsoft Research Ltd.
Cambridge, UK

John Winn
Microsoft Research Ltd.
Cambridge, UK

Abstract

Gates are a new notation for representing mixture models and context-sensitive independence in factor graphs. Factor graphs provide a natural representation for message-passing algorithms, such as expectation propagation. However, message passing in mixture models is not well captured by factor graphs unless the entire mixture is represented by one factor, because the message equations have a containment structure. Gates capture this containment structure graphically, allowing both the independences and the message-passing equations for a model to be readily visualized. Different variational approximations for mixture models can be understood as different ways of drawing the gates in a model. We present general equations for expectation propagation and variational message passing in the presence of gates.

General Blueprint

- Choose q
- Derive ELBO
- Coordinate ascent of each q_i
- Repeat until convergence

General Blueprint

- Choose q
- Derive ELBO
- Coordinate ascent of each q_i
- Repeat until convergence

Variational Inference

Material adapted from David Blei

University of Maryland

LDA Derivation

Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

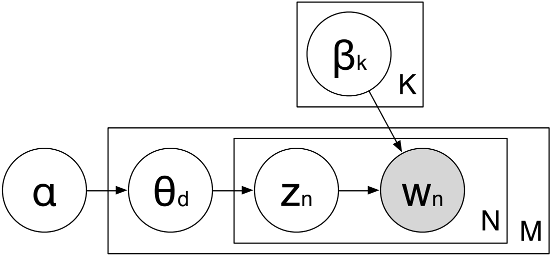
Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

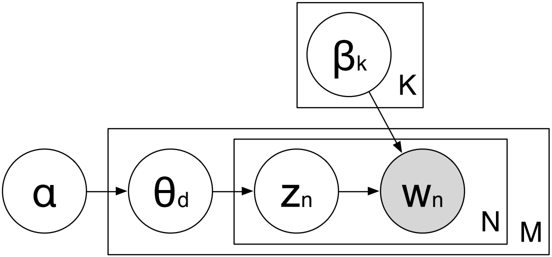
Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

LDA Generative Model



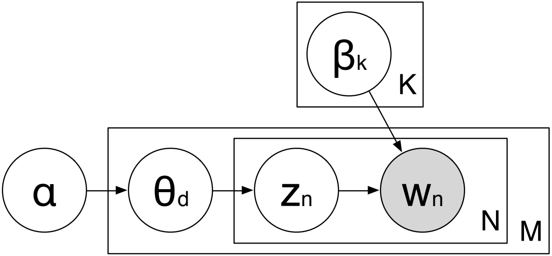
- For each topic $k \in \{1, \ldots, K\}$, a multinomial distribution β_k

LDA Generative Model



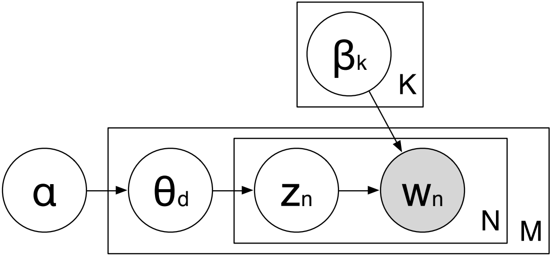
- For each topic $k \in \{1, \ldots, K\}$, a multinomial distribution β_k
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

LDA Generative Model



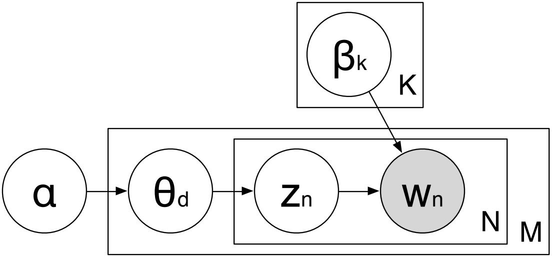
- For each topic $k \in \{1, \ldots, K\}$, a multinomial distribution β_k
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.

LDA Generative Model



- For each topic $k \in \{1, \ldots, K\}$, a multinomial distribution β_k
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

LDA Generative Model



- For each topic $k \in \{1, \dots, K\}$, a multinomial distribution β_k
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

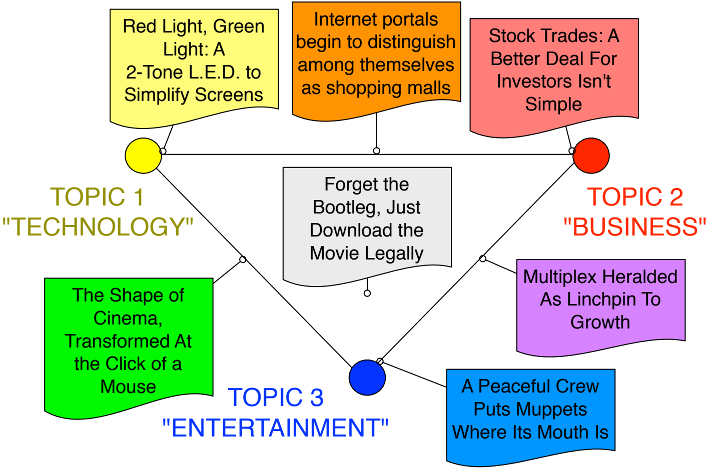
Statistical inference uncovers most unobserved variables given data.

Example: Latent Dirichlet Allocation

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

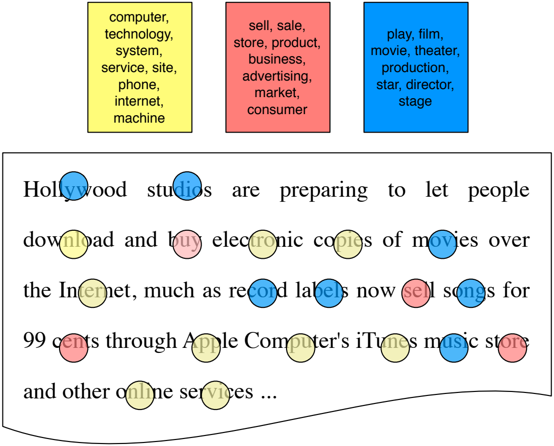
Distribution over words given a topic i: β_i

Example: Latent Dirichlet Allocation



Distribution over topics given a document d: θ_d

Example: Latent Dirichlet Allocation



Assignment of token to topic: $z_{d,n}$

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}) \quad (7)$$

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}) \quad (7)$$

- $p(\theta_d \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_k \theta_{d,k}^{\alpha_k - 1}$ (Dirichlet)

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}) \quad (7)$$

- $p(\theta_d \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_k \theta_{d,k}^{\alpha_k - 1}$ (Dirichlet)
- $p(z_{d,n} \mid \theta_d) = \theta_{d,z_{d,n}}$ (Draw from Multinomial)

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}) \quad (7)$$

- $p(\theta_d \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_k \theta_{d,k}^{\alpha_k - 1}$ (Dirichlet)
- $p(z_{d,n} \mid \theta_d) = \theta_{d, z_{d,n}}$ (Draw from Multinomial)
- $p(w_{d,n} \mid \beta, z_{d,n}) = \beta_{z_{d,n}, w_{d,n}}$ (Draw from Multinomial)

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}) \tag{7}$$

Variational distribution:

$$q(\Theta, Z) = \prod_d q(\theta_d \mid \gamma_d) \prod_n q(z_{d,n} \mid \phi_{d,n}) \tag{8}$$

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}) \tag{7}$$

Variational distribution:

$$q(\Theta, Z) = \prod_d q(\theta_d \mid \gamma_d) \prod_n q(z_{d,n} \mid \phi_{d,n}) \tag{8}$$

ELBO:

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) = & \mathbb{E}_q [\log p(\theta \mid \alpha)] + \mathbb{E}_q [\log p(z \mid \theta)] + \mathbb{E}_q [\log p(w \mid z, \beta)] \\ & - \mathbb{E}_q [\log q(\theta)] - \mathbb{E}_q [\log q(z)] \end{aligned} \tag{9}$$

What is the variational distribution?

$$q(\vec{\theta}, \vec{z}) = \prod_d q(\theta_d \,|\, \gamma_d) \prod_n q(z_{d,n} \,|\, \phi_{d,n}) \tag{10}$$

- Variational document distribution over topics γ_d
  - Vector of length K for each document
  - Non-negative
  - Doesn't sum to 1.0
- Variational token distribution over topic assignments $\phi_{d,n}$
  - Vector of length K for every token
  - Non-negative, sums to 1.0

Expectation of log Dirichlet

- Most expectations are straightforward to compute
- Dirichlet is harder

$$\mathbb{E}_{\text{dir}}\left[\log p(\theta_i \mid \alpha)\right] = \Psi(\alpha_i) - \Psi\left(\sum_j \alpha_j\right) \tag{11}$$

Reminders

- Gamma function

$$\Gamma(n) = (n-1)! = \int_0^\infty x^{n-1} e^{-x}\, dx \tag{12}$$

- Digamma function

$$\psi(x) = \frac{d}{dx} \ln\left(\Gamma(x)\right) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{13}$$

Expectation 1

$$\mathbb{E}_q[\log p(\theta \,|\, \alpha)] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{14}$$

$$\tag{15}$$

Expectation 1

$$\mathbb{E}_q\left[\log p(\theta \mid \alpha)\right] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{14}$$

$$= \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\right\} + \sum_i \log \theta_i^{\alpha_i - 1}\right] \tag{15}$$

Log of products becomes sum of logs.

Expectation 1

$$\mathbb{E}_q[\log p(\theta \mid \alpha)] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{14}$$

$$= \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\right\} + \sum_i \log \theta_i^{\alpha_i - 1}\right]$$

$$= \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) + \mathbb{E}_q\left[\sum_i (\alpha_i - 1) \log \theta_i\right] \tag{15}$$

Log of exponent becomes product, expectation of constant is constant

Expectation 1

$$
\begin{aligned}
\mathbb{E}_q[\log p(\theta \mid \alpha)] = & \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}\right\}\right] \\
= & \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\right\} + \sum_i \log \theta_i^{\alpha_i - 1}\right] \\
= & \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) + \mathbb{E}_q\left[\sum_i (\alpha_i - 1) \log \theta_i\right] \\
= & \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) \\
& + \sum_i (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right)\right)
\end{aligned} \tag{14}
$$

Expectation of log Dirichlet

Reminder: Indicator Function

$$\mathbb{1}[x] \equiv \begin{cases} 1 & \text{if } x \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

Expectation 2

$$\mathbb{E}_q[\log p(z \mid \theta)] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n == i]}\right] \tag{16}$$

$$\tag{17}$$

Expectation 2

$$\mathbb{E}_q\left[\log p(z \mid \theta)\right] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n == i]}\right] \quad (16)$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n == i]}\right] \quad (17)$$

$$(18)$$

Products to sums

Expectation 2

$$\mathbb{E}_q[\log p(z \mid \theta)] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{16}$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{17}$$

$$= \sum_n \sum_i \mathbb{E}_q\left[\log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{18}$$

$$\tag{19}$$

Linearity of expectation

Expectation 2

$$\mathbb{E}_q[\log p(z \mid \theta)] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n == i]}\right] \quad (16)$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n == i]}\right] \quad (17)$$

$$= \sum_n \sum_i \mathbb{E}_q\left[\log \theta_i^{\mathbb{1}[z_n == i]}\right] \quad (18)$$

$$= \sum_n \sum_i \phi_{ni} \mathbb{E}_q[\log \theta_i] \quad (19)$$

$$(20)$$

Independence of variational distribution, exponents become products

Expectation 2

$$
\begin{aligned}
\mathbb{E}_q\left[\log p(z \mid \theta)\right] &= \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n == i]}\right] && (16)\\
&= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n == i]}\right] && (17)\\
&= \sum_n \sum_i \mathbb{E}_q\left[\log \theta_i^{\mathbb{1}[z_n == i]}\right] && (18)\\
&= \sum_n \sum_i \phi_{ni} \mathbb{E}_q\left[\log \theta_i\right] && (19)\\
&= \sum_n \sum_i \phi_{ni} \left(\Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right)\right) && (20)
\end{aligned}
$$

Expectation of log Dirichlet

Expectation 3

$$\mathbb{E}_q[\log p(w \mid z, \beta)] = \mathbb{E}_q\left[\log \beta_{z_{d,n}, w_{d,n}}\right] \tag{21}$$

(22)

Expectation 3

$$\mathbb{E}_q\left[\log p(w \mid z, \beta)\right] = \mathbb{E}_q\left[\log \beta_{z_{d,n}, w_{d,n}}\right] \tag{21}$$

$$= \mathbb{E}_q\left[\log \prod_{v}^{V} \prod_{i}^{K} \beta_{i,v}^{\mathbb{1}[v = w_{d,n}, z_{d,n} = i]}\right] \tag{22}$$

$$\tag{23}$$

Expectation 3

$$\mathbb{E}_q\left[\log p(w \mid z, \beta)\right] = \mathbb{E}_q\left[\log \beta_{z_{d,n}, w_{d,n}}\right] \tag{21}$$

$$= \mathbb{E}_q\left[\log \prod_v^V \prod_i^K \beta_{i,v}^{\mathbb{1}[v = w_{d,n}, z_{d,n} = i]}\right] \tag{22}$$

$$= \sum_v^V \sum_i^K \mathbb{E}_q\left[\mathbb{1}\left[v = w_{d,n}, z_{d,n} = i\right]\right] \log \beta_{i,v} \tag{23}$$

$$\tag{24}$$

Expectation 3

$$
\begin{aligned}
\mathbb{E}_q[\log p(w \mid z, \beta)] &= \mathbb{E}_q\left[\log \beta_{z_{d,n}, w_{d,n}}\right] && (21)\\
&= \mathbb{E}_q\left[\log \prod_v^V \prod_i^K \beta_{i,v}^{\mathbb{1}[v = w_{d,n}, z_{d,n} = i]}\right] && (22)\\
&= \sum_v^V \sum_i^K \mathbb{E}_q\left[\mathbb{1}\left[v = w_{d,n}, z_{d,n} = i\right]\right] \log \beta_{i,v} && (23)\\
&= \sum_v^V \sum_i^K \phi_{n,i} w_{d,n}^v \log \beta_{i,v} && (24)
\end{aligned}
$$

Entropies

Entropy of Dirichlet

$$\mathbb{H}_q[\gamma] = -\log\Gamma\left(\sum_j \gamma_j\right) + \sum_i \log\Gamma(\gamma_i) - \sum_i (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right)$$

Entropies

Entropy of Dirichlet

$$\mathbb{H}_q[\gamma] = -\log\Gamma\left(\sum_j \gamma_j\right) + \sum_i \log\Gamma(\gamma_i) - \sum_i (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)$$

Entropy of Multinomial

$$\mathbb{H}_q[\phi_{d,n}] = -\sum_i \phi_{d,n,i}\log\phi_{d,n,i} \tag{25}$$

Complete objective function

$$
\begin{aligned}
L(\gamma,\phi;\alpha,\beta) &= \log\Gamma\left(\sum_{j=1}^{k}\alpha_j\right) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i-1)\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V}\phi_{ni}w_n^j\log\beta_{ij} \\
&- \log\Gamma\left(\sum_{j=1}^{k}\gamma_j\right) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i-1)\left(\Psi(\gamma_i)-\Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) \\
&- \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni},
\end{aligned}
$$

Note the entropy terms at the end (negative sign)

Deriving the algorithm

- Compute partial wrt to variable of interest
- Set equal to zero
- Solve for variable

Update for ϕ

Derivative of ELBO:

$$\frac{\partial \mathcal{L}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right) + \log \beta_{i,v} - \log \phi_{ni} - 1 + \lambda \tag{26}$$

Solution:

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{27}$$

Update for ϕ

Derivative of ELBO:

$$\frac{\partial \mathscr{L}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right) + \log \beta_{i,v} - \log \phi_{ni} - 1 + \lambda \tag{26}$$

Solution:

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{27}$$

Remind you of Gibbs?

How much the topic likes word v

Update for ϕ

Derivative of ELBO:

$$\frac{\partial \mathcal{L}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right) + \log \beta_{i,v} - \log \phi_{ni} - 1 + \lambda \tag{26}$$

Solution:

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{27}$$

Remind you of Gibbs?

How much document *d* likes topic *i*

Update for γ

Derivative of ELBO:

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \gamma_i} =& \Psi'(\gamma_i)(\alpha_i + \phi_{n,i} - \gamma_i) \\ &- \Psi'\left(\sum_j \gamma_j\right) \sum_j \left(\alpha_j + \sum_n \phi_{nj} - \gamma_j\right)\end{aligned}$$

Update for γ

Derivative of ELBO:

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \gamma_i} =& \Psi'(\gamma_i)\left(\alpha_i + \phi_{n,i} - \gamma_i\right) \\ & - \Psi'\left(\sum_j \gamma_j\right) \sum_j \left(\alpha_j + \sum_n \phi_{nj} - \gamma_j\right)\end{aligned}$$

Update for γ

Derivative of ELBO:

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'(\gamma_i)(\alpha_i + \phi_{n,i} - \gamma_i) - \Psi'\left(\sum_j \gamma_j\right) \sum_j \left(\alpha_j + \sum_n \phi_{nj} - \gamma_j\right)$$

Solution:

$$\gamma_i = \alpha_i + \sum_n \phi_{ni} \tag{28}$$

Update for β

Slightly more complicated (requires Lagrange parameter), but solution is obvious:

$$\beta_{ij} \propto \sum_d \sum_n \phi_{dni} w_{dn}^j \tag{29}$$

Overall Algorithm

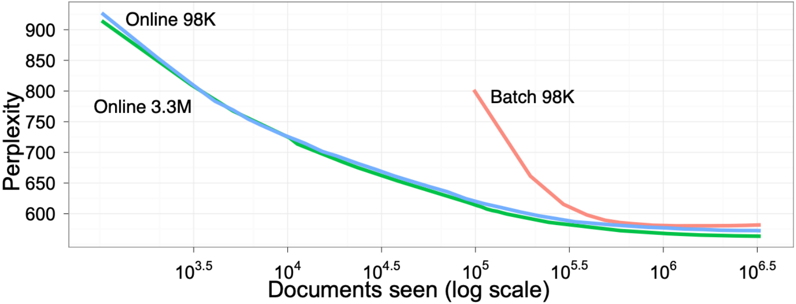
1. Randomly initialize variational parameters (can't be uniform)
2. For each iteration:
    2.1 For each document, update γ and ϕ
    2.2 For corpus, update β
    2.3 Compute $\mathscr{L}$ for diagnostics
3. Return expectation of variational parameters for solution to latent variables

Online Learning for Latent Dirichlet Allocation

Matthew D. Hoffman
Department of Computer Science
Princeton University
Princeton, NJ
mdhoffma@cs.princeton.edu

David M. Blei
Department of Computer Science
Princeton University
Princeton, NJ
blei@cs.princeton.edu

Francis Bach
INRIA—Ecole Normale Supérieure
Paris, France
francis.bach@ens.fr



Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first

Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first . . . then match implementation
- Check with synthetic data

Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first ... then match implementation
- Check with synthetic data
- Try to match variables with paper
- Write unit tests for each atomic update
- Monitor variational bound (with asserts)

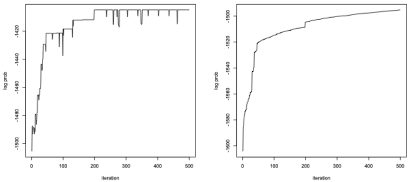
Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first . . . then match implementation
- Check with synthetic data
- Try to match variables with paper
- Write unit tests for each atomic update
- Monitor variational bound (with asserts)
- Write the state (checkpointing and debugging)
- Visualize variational parameters

Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first . . . then match implementation
- Check with synthetic data
- Try to match variables with paper
- Write unit tests for each atomic update
- Monitor variational bound (with asserts)
- Write the state (checkpointing and debugging)
- Visualize variational parameters
- Cache / memoize gamma / digamma functions

Relationship with Gibbs Sampling



- Gibbs objective can jump around (left), variational always increases (right)
- Batch is sometimes a good fit for smaller datasets
- Gibbs sampling: sample from the conditional distribution of all other variables
- Variational inference: each factor is set to the exponentiated log of the conditional
- Variational is easier to parallelize, Gibbs faster per step
- Gibbs typically easier to implement