Approximate Bayesian Inference

Evgeny Burnaev

Skoltech, Moscow, Russia

Skoltech
Skolkovo Institute of Science and Technology

Outline

1. Introduction
2. Variational Inference
3. ELBO
4. Mean-field Variational Bayes

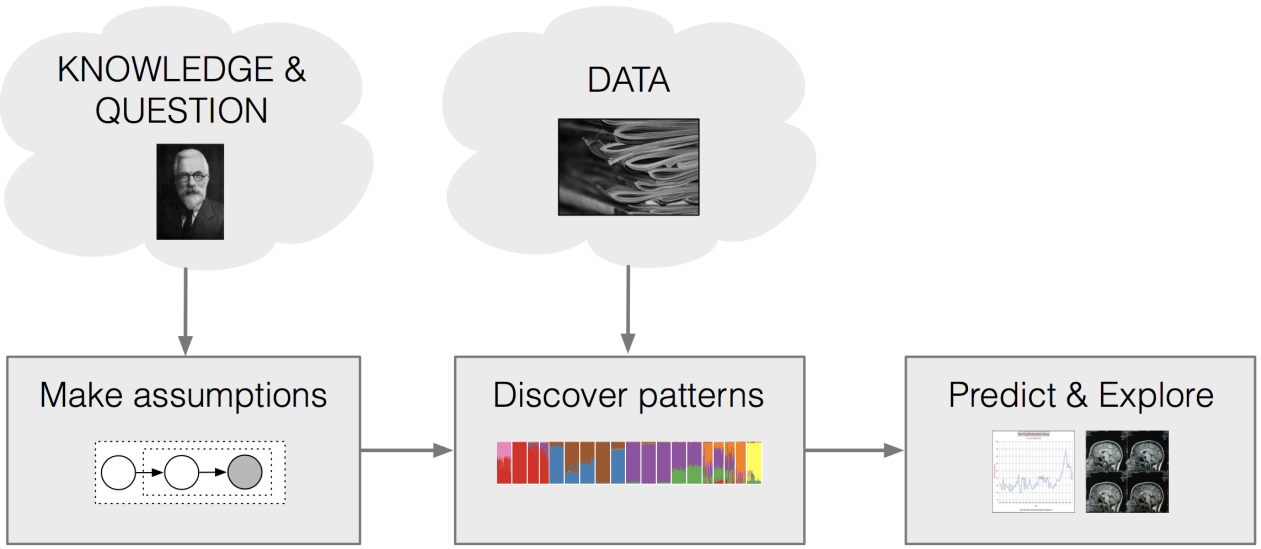
1 Introduction

2 Variational Inference

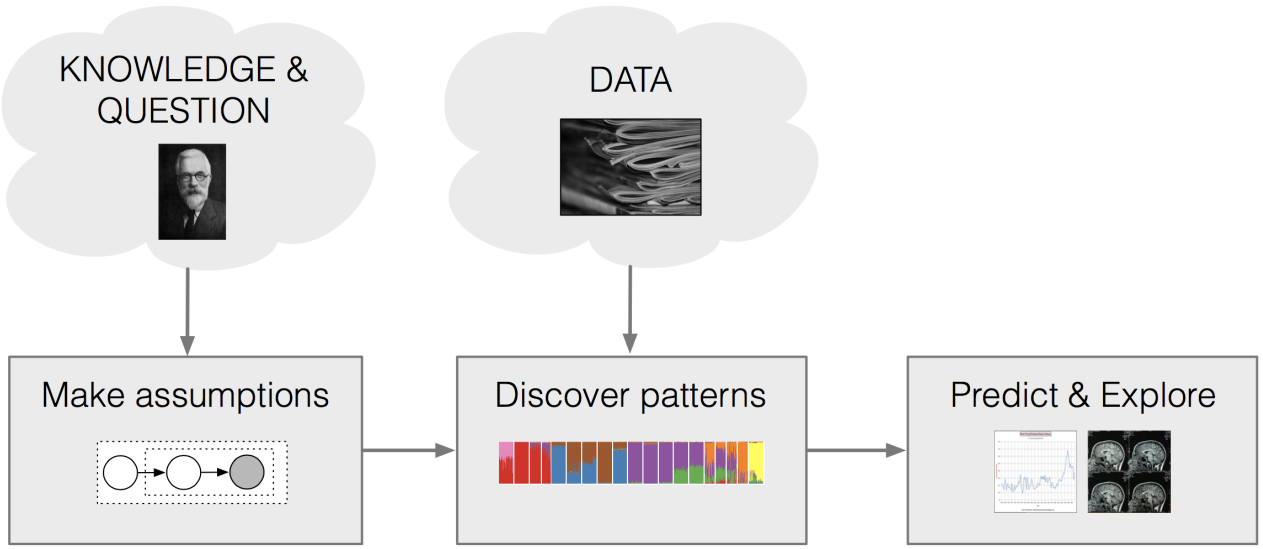
3 ELBO

4 Mean-field Variational Bayes

The probabilistic pipeline

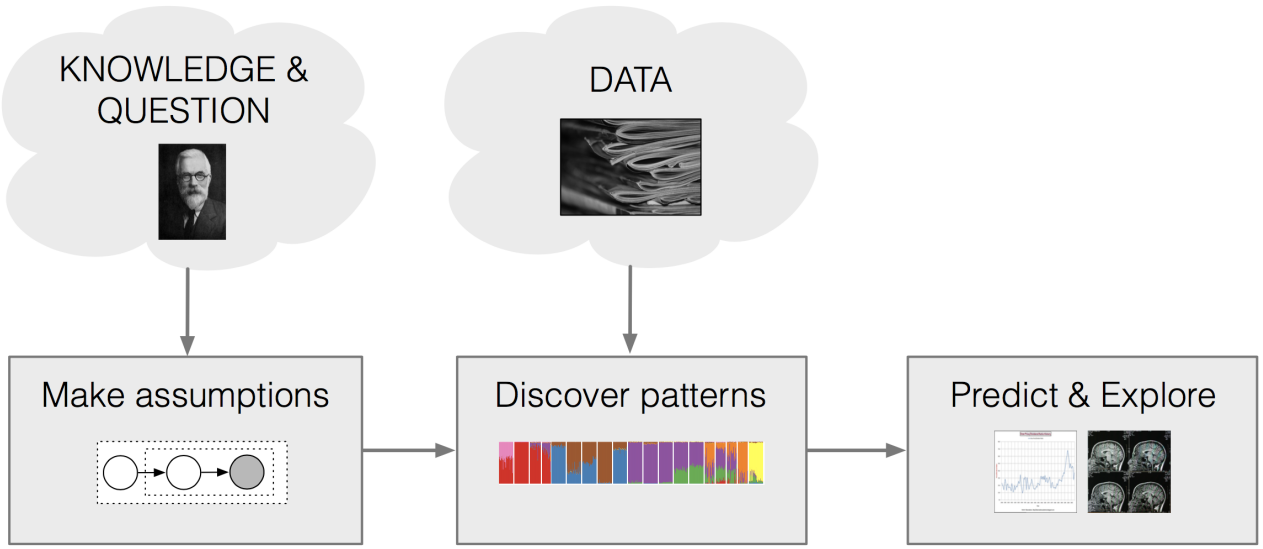


The probabilistic pipeline



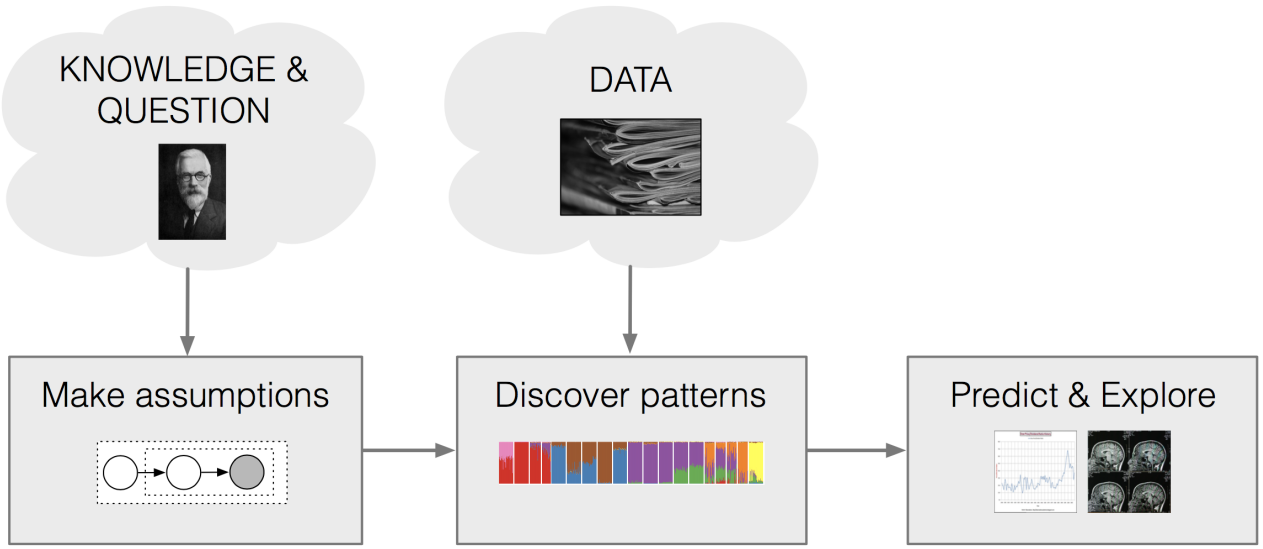
- Customized data analysis is important to many fields

The probabilistic pipeline



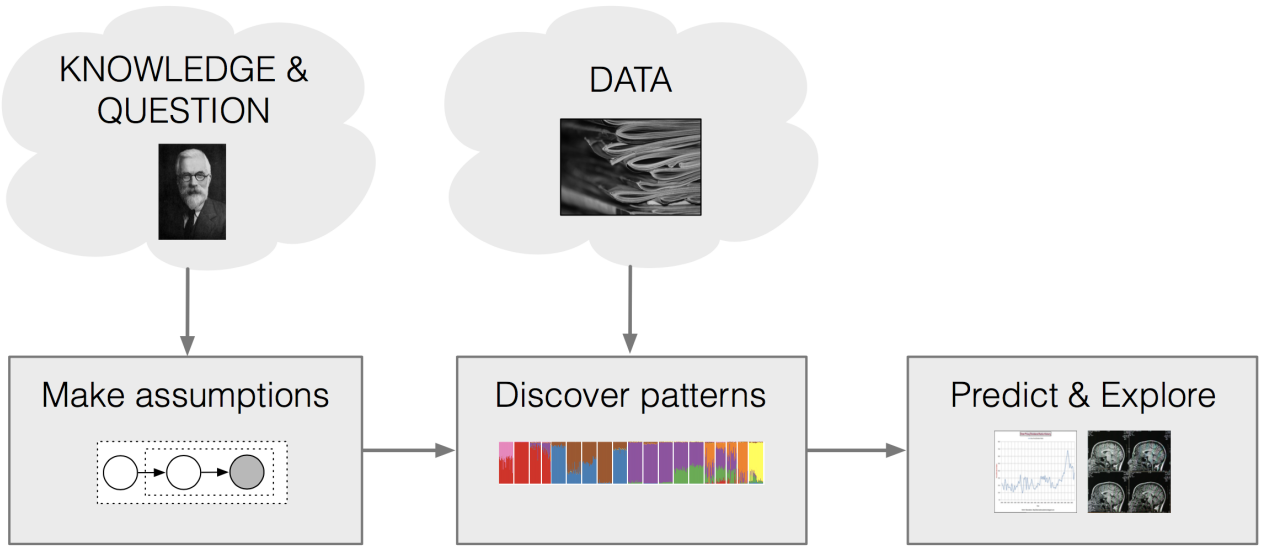
- Customized data analysis is important to many fields
- Pipeline separates assumptions, computation, application

The probabilistic pipeline



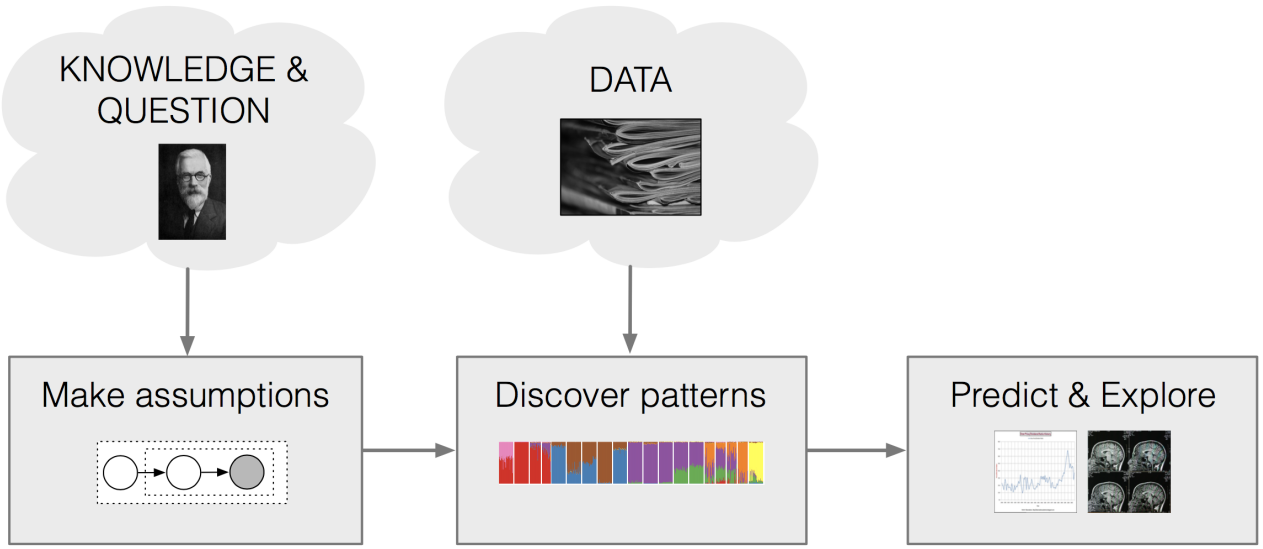
- Customized data analysis is important to many fields
- Pipeline separates assumptions, computation, application
- Eases collaborative solutions to statistics problems

The probabilistic pipeline



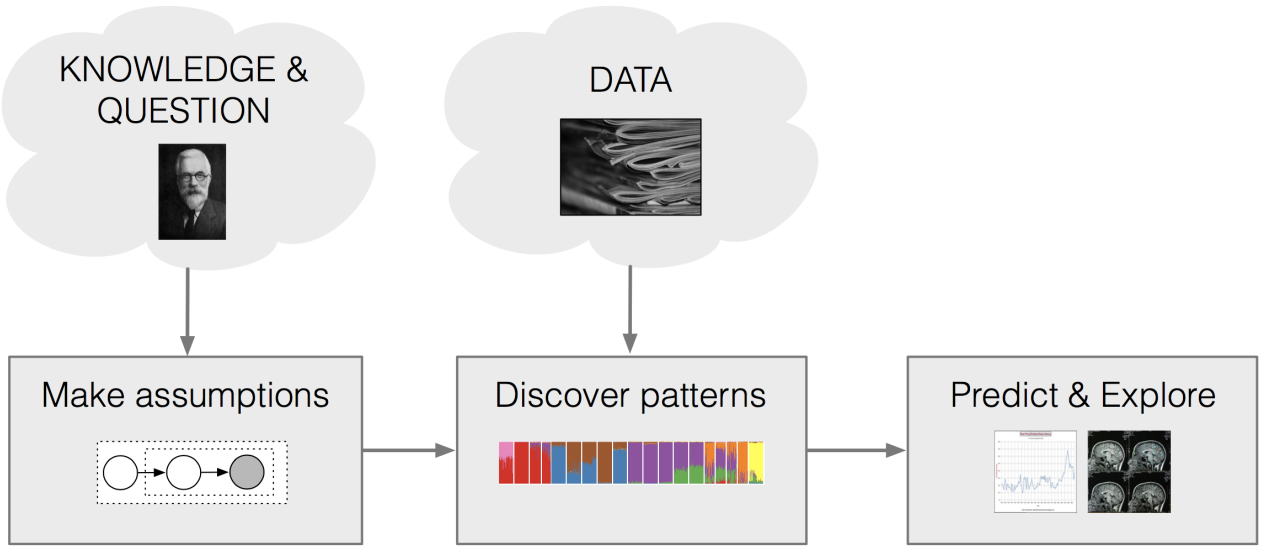
- Inference is the key algorithmic problem

The probabilistic pipeline



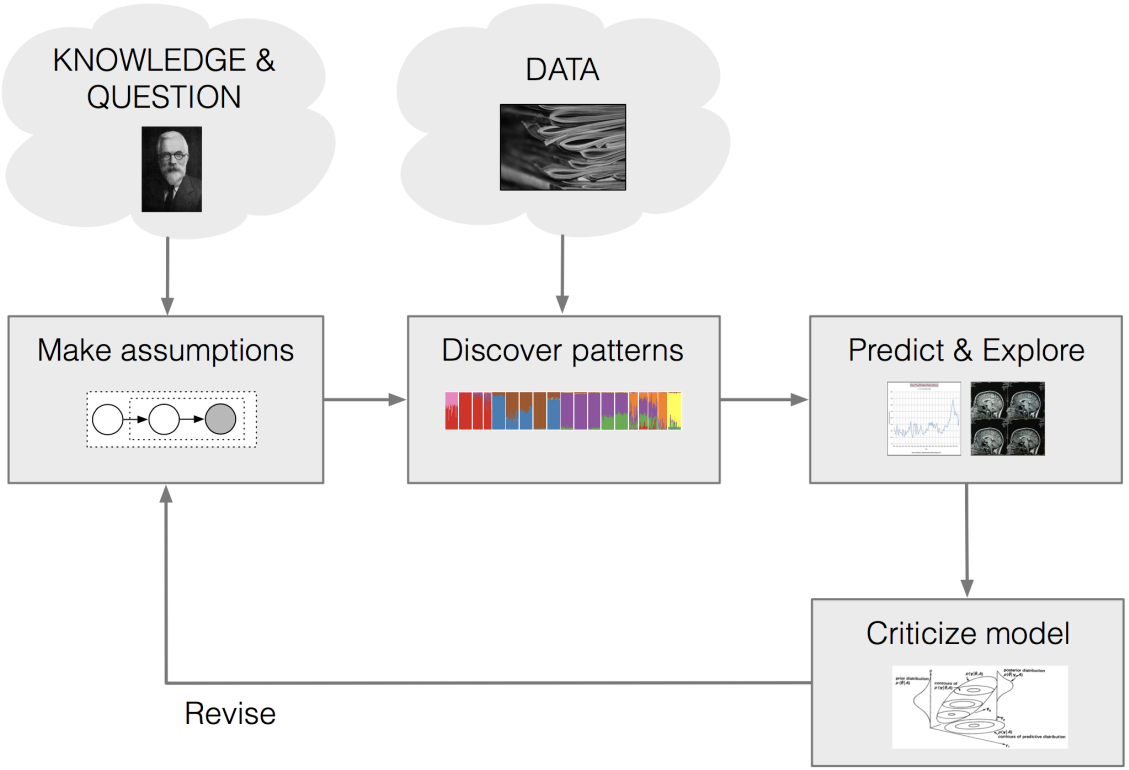
- Inference is the key algorithmic problem
- Answers the question: What does this model say about this data?

The probabilistic pipeline



- Inference is the key algorithmic problem
- Answers the question: What does this model say about this data?
- Our goal: General and scalable approaches to inference

The probabilistic pipeline



Basic Notation

- $p(\mathbf{x})$ — probability of $\mathbf{x}$
- $p(\mathbf{x}|\boldsymbol{\theta})$ — conditional probability of $\mathbf{x}$ given $\boldsymbol{\theta}$
- Bayes Rule

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})},$$

where

$$p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta}$$

Basic Notation

- $p(\mathbf{x})$ — probability of $\mathbf{x}$
- $p(\mathbf{x}|\boldsymbol{\theta})$ — conditional probability of $\mathbf{x}$ given $\boldsymbol{\theta}$
- Bayes Rule

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})},$$

where

$$p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta}$$

Basic Notation

- $p(\mathbf{x})$ — probability of $\mathbf{x}$
- $p(\mathbf{x}|\boldsymbol{\theta})$ — conditional probability of $\mathbf{x}$ given $\boldsymbol{\theta}$
- Bayes Rule

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})},$$

where

$$p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta}$$

Inference Problem

- Given a dataset $\mathcal{D}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$
- Bayes Rule

$$p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_m)}$$

  where
  - $p(\mathcal{D}_m|\boldsymbol{\theta})$ — likelihood function of $\boldsymbol{\theta}$
  - $p(\boldsymbol{\theta})$ — prior probability of $\boldsymbol{\theta}$
  - $p(\boldsymbol{\theta}|\mathcal{D}_m)$ — posterior distribution over $\boldsymbol{\theta}$
- Computing posterior distribution is known as the **inference** problem: the integral

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m, \boldsymbol{\theta}) d\boldsymbol{\theta}$$

  can be very high-dimensional and difficult to compute

Inference Problem

- Given a dataset $\mathcal{D}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$
- Bayes Rule

$$p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_m)}$$

  where
  - $p(\mathcal{D}_m|\boldsymbol{\theta})$ — likelihood function of $\boldsymbol{\theta}$
  - $p(\boldsymbol{\theta})$ — prior probability of $\boldsymbol{\theta}$
  - $p(\boldsymbol{\theta}|\mathcal{D}_m)$ — posterior distribution over $\boldsymbol{\theta}$
- Computing posterior distribution is known as the **inference** problem: the integral

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m, \boldsymbol{\theta}) d\boldsymbol{\theta}$$

  can be very high-dimensional and difficult to compute

Inference Problem

- Given a dataset $\mathcal{D}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$
- Bayes Rule

$$p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_m)}$$

  where
  - $p(\mathcal{D}_m|\boldsymbol{\theta})$ — likelihood function of $\boldsymbol{\theta}$
  - $p(\boldsymbol{\theta})$ — prior probability of $\boldsymbol{\theta}$
  - $p(\boldsymbol{\theta}|\mathcal{D}_m)$ — posterior distribution over $\boldsymbol{\theta}$
- Computing posterior distribution is known as the **inference** problem: the integral

$$p(\mathcal{D}_m) = \int p(\mathcal{D}_m, \boldsymbol{\theta}) d\boldsymbol{\theta}$$

  can be very high-dimensional and difficult to compute

Prediction

$$p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_m)}$$

where

- $p(\mathcal{D}_m|\boldsymbol{\theta})$ — likelihood function of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta})$ — prior probability of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta}|\mathcal{D}_m)$ — posterior distribution over $\boldsymbol{\theta}$

Prediction

$$p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_m)}$$

where

- $p(\mathcal{D}_m|\boldsymbol{\theta})$ — likelihood function of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta})$ — prior probability of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta}|\mathcal{D}_m)$ — posterior distribution over $\boldsymbol{\theta}$

Prediction: given $\mathcal{D}_m$, compute conditional probability of $\mathbf{x}^*$ (predictive distribution)

Prediction

$$p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_m)}$$

where

- $p(\mathcal{D}_m|\boldsymbol{\theta})$ — likelihood function of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta})$ — prior probability of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta}|\mathcal{D}_m)$ — posterior distribution over $\boldsymbol{\theta}$

Prediction: given $\mathcal{D}_m$, compute conditional probability of $\mathbf{x}^*$ (predictive distribution)

$$p(\mathbf{x}^*|\mathcal{D}_m) = \int p(\mathbf{x}^*|\boldsymbol{\theta}, \mathcal{D}_m)p(\boldsymbol{\theta}|\mathcal{D}_m)d\boldsymbol{\theta}$$

Prediction

$$p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_m)}$$

where

- $p(\mathcal{D}_m|\boldsymbol{\theta})$ — likelihood function of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta})$ — prior probability of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta}|\mathcal{D}_m)$ — posterior distribution over $\boldsymbol{\theta}$

Prediction: given $\mathcal{D}_m$, compute conditional probability of $\mathbf{x}^*$ (predictive distribution)

$$\begin{aligned} p(\mathbf{x}^*|\mathcal{D}_m) &= \int p(\mathbf{x}^*|\boldsymbol{\theta}, \mathcal{D}_m)p(\boldsymbol{\theta}|\mathcal{D}_m)d\boldsymbol{\theta} \\ &= \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_m)}[p(\mathbf{x}^*|\boldsymbol{\theta}, \mathcal{D}_m)] \end{aligned}$$

Prediction

$$p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_m)}$$

where

- $p(\mathcal{D}_m|\boldsymbol{\theta})$ — likelihood function of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta})$ — prior probability of $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta}|\mathcal{D}_m)$ — posterior distribution over $\boldsymbol{\theta}$

Prediction: given $\mathcal{D}_m$, compute conditional probability of $\mathbf{x}^*$ (predictive distribution)

$$\begin{aligned} p(\mathbf{x}^*|\mathcal{D}_m) &= \int p(\mathbf{x}^*|\boldsymbol{\theta}, \mathcal{D}_m)p(\boldsymbol{\theta}|\mathcal{D}_m)d\boldsymbol{\theta} \\ &= \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_m)}[p(\mathbf{x}^*|\boldsymbol{\theta}, \mathcal{D}_m)] \end{aligned}$$

Computing predictive distribution require posterior $p(\boldsymbol{\theta}|\mathcal{D}_m)$

Model Selection

- Compare model classes, e.g. $\mathcal{M}_1$ and $\mathcal{M}_2$. Need to compute posterior probabilities given $\mathcal{D}$

$$p(\mathcal{M}|\mathcal{D}_m) = \frac{p(\mathcal{D}_m|\mathcal{M})p(\mathcal{M})}{p(\mathcal{D}_m)},$$

where

$$p(\mathcal{D}_m|\mathcal{M}) = \int p(\mathcal{D}_m|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}$$

is known as the marginal likelihood or evidence

Computational Challenges

- Computing marginal likelihoods often requires computing very high-dimensional integrals
- Computing posterior distributions (and hence predictive distributions) is often analytically intractable
- Let us look at some specific examples:
  - Bayesian Probabilistic Matrix Factorization
  - Bayesian Neural Networks
  - Probabilistic Topic Modeling

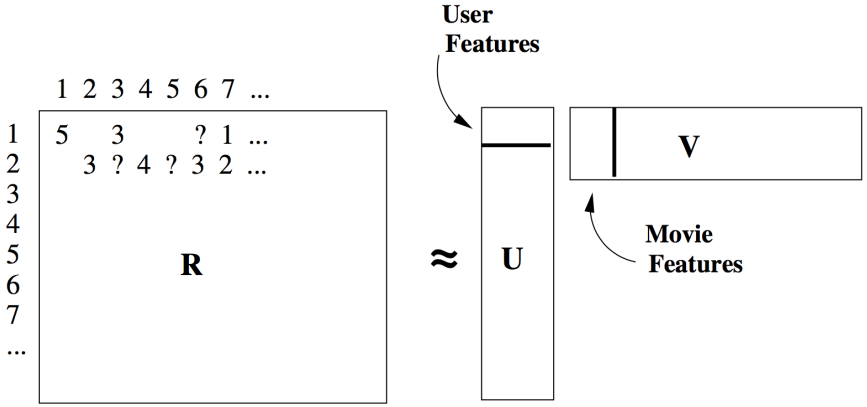
Computational Challenges

- Computing marginal likelihoods often requires computing very high-dimensional integrals
- Computing posterior distributions (and hence predictive distributions) is often analytically intractable
- Let us look at some specific examples:
  - Bayesian Probabilistic Matrix Factorization
  - Bayesian Neural Networks
  - Probabilistic Topic Modeling

Computational Challenges

- Computing marginal likelihoods often requires computing very high-dimensional integrals
- Computing posterior distributions (and hence predictive distributions) is often analytically intractable
- Let us look at some specific examples:
  - Bayesian Probabilistic Matrix Factorization
  - Bayesian Neural Networks
  - Probabilistic Topic Modeling

Bayesian PMF

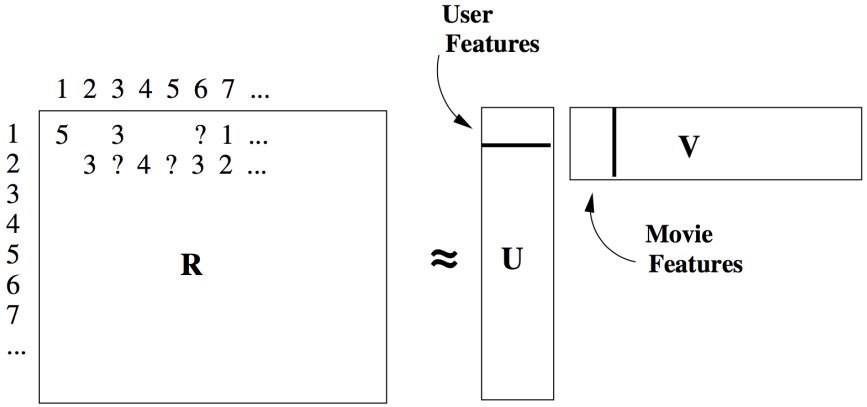


- We have N users, M movies, and integer rating values from 1 to K
- Let r_{ij} be the rating of user i for movie j,and $U \in \mathbb{R}^{D \times N}$, $V \in \mathbb{R}^{D \times M}$ be latent user and movie feature matrices:

$$R \approx U^{\mathrm{T}} V$$

- Goal: Predict missing ratings

Bayesian PMF

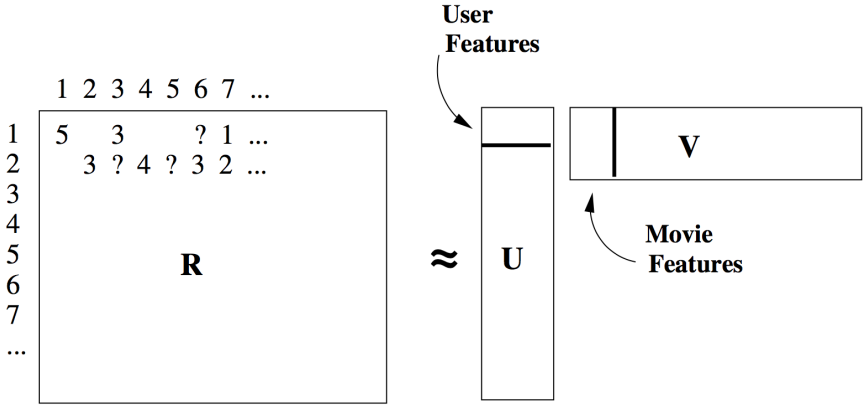


- We have N users, M movies, and integer rating values from 1 to K
- Let r_{ij} be the rating of user i for movie j,and $U \in \mathbb{R}^{D \times N}$, $V \in \mathbb{R}^{D \times M}$ be latent user and movie feature matrices:

$$R \approx U^{\mathrm{T}} V$$

- Goal: Predict missing ratings

Bayesian PMF



- We have N users, M movies, and integer rating values from 1 to K
- Let r_{ij} be the rating of user i for movie j,and $U \in \mathbb{R}^{D \times N}$, $V \in \mathbb{R}^{D \times M}$ be latent user and movie feature matrices:

$$R \approx U^{\mathrm{T}} V$$

- Goal: Predict missing ratings

Bayesian PMF

- Probabilistic linear model with Gaussian observation noise. Likelihood:

$$p(r_{ij}|u_i, v_j, \sigma^2) = \mathcal{N}(r_{ij}|u_i^{\mathrm{T}} v_j, \sigma^2)$$

- Gaussian priors over parameters:

$$p(U|\mu_U, \Lambda_U) = \prod_{i=1}^{N} \mathcal{N}(u_i|\mu_u, \Sigma_u),$$

$$p(V|\mu_V, \Lambda_V) = \prod_{i=1}^{M} \mathcal{N}(v_i|\mu_v, \Sigma_v)$$

- Conjugate Gaussian-inverse-Wishart priors on the user and movie hyperparameters $\boldsymbol{\theta}_U = \{\mu_u, \Sigma_u\}$ and $\boldsymbol{\theta}_V = \{\mu_v, \Sigma_v\}$
- **Hierarchical Prior**

Bayesian PMF

- Probabilistic linear model with Gaussian observation noise. Likelihood:

$$p(r_{ij}|u_i, v_j, \sigma^2) = \mathcal{N}(r_{ij}|u_i^{\mathrm{T}} v_j, \sigma^2)$$

- Gaussian priors over parameters:

$$p(U|\mu_U, \Lambda_U) = \prod_{i=1}^{N} \mathcal{N}(u_i|\mu_u, \Sigma_u),$$

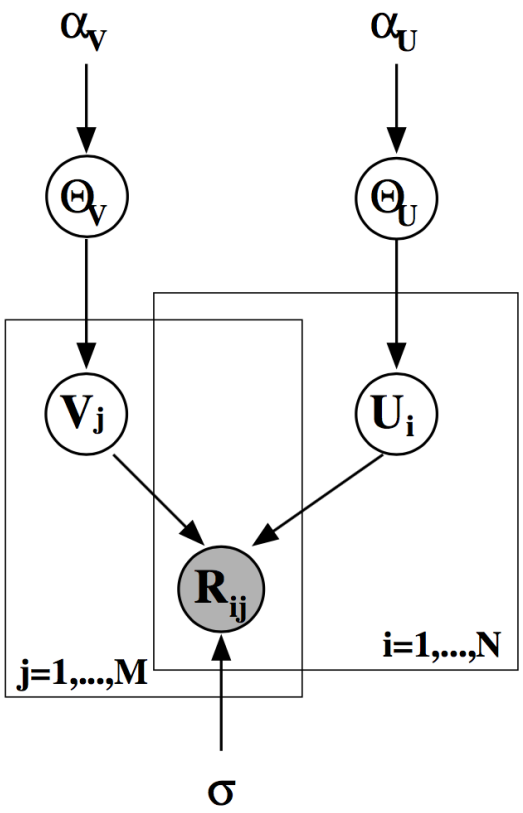
$$p(V|\mu_V, \Lambda_V) = \prod_{i=1}^{M} \mathcal{N}(v_i|\mu_v, \Sigma_v)$$

- Conjugate Gaussian-inverse-Wishart priors on the user and movie hyperparameters $\boldsymbol{\theta}_U = \{\mu_u, \Sigma_u\}$ and $\boldsymbol{\theta}_V = \{\mu_v, \Sigma_v\}$
- **Hierarchical Prior**

Bayesian PMF

- Probabilistic linear model with Gaussian observation noise. Likelihood:
$$p(r_{ij}|u_i, v_j, \sigma^2) = \mathcal{N}(r_{ij}|u_i^{\mathrm{T}} v_j, \sigma^2)$$
- Gaussian priors over parameters:
$$p(U|\mu_U, \Lambda_U) = \prod_{i=1}^{N} \mathcal{N}(u_i|\mu_u, \Sigma_u),$$
$$p(V|\mu_V, \Lambda_V) = \prod_{i=1}^{M} \mathcal{N}(v_i|\mu_v, \Sigma_v)$$
- Conjugate Gaussian-inverse-Wishart priors on the user and movie hyperparameters $\boldsymbol{\theta}_U = \{\mu_u, \Sigma_u\}$ and $\boldsymbol{\theta}_V = \{\mu_v, \Sigma_v\}$
- **Hierarchical Prior**

Bayesian PMF



Bayesian PMF

- Predictive distribution: consider predicting a rating r_{ij}^* for user i and query movie j:

$$p(r_{ij}^*|R) = \int p(r_{ij}^*|u_i, v_j)p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)d\{U, V\}d\{\boldsymbol{\theta}_U, \boldsymbol{\theta}_V\},$$

  where $p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)$ is the posterior over parameters and hyperparameters
- Exact evaluation of this predictive distribution is analytically intractable
- Posterior distribution $p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)$ is complicated and does not have a closed form expression
- Need to approximate

Bayesian PMF

- Predictive distribution: consider predicting a rating r_{ij}^* for user i and query movie j:

$$p(r_{ij}^*|R) = \int p(r_{ij}^*|u_i, v_j)p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)d\{U, V\}d\{\boldsymbol{\theta}_U, \boldsymbol{\theta}_V\},$$

  where $p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)$ is the posterior over parameters and hyperparameters
- Exact evaluation of this predictive distribution is analytically intractable
- Posterior distribution $p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)$ is complicated and does not have a closed form expression
- Need to approximate

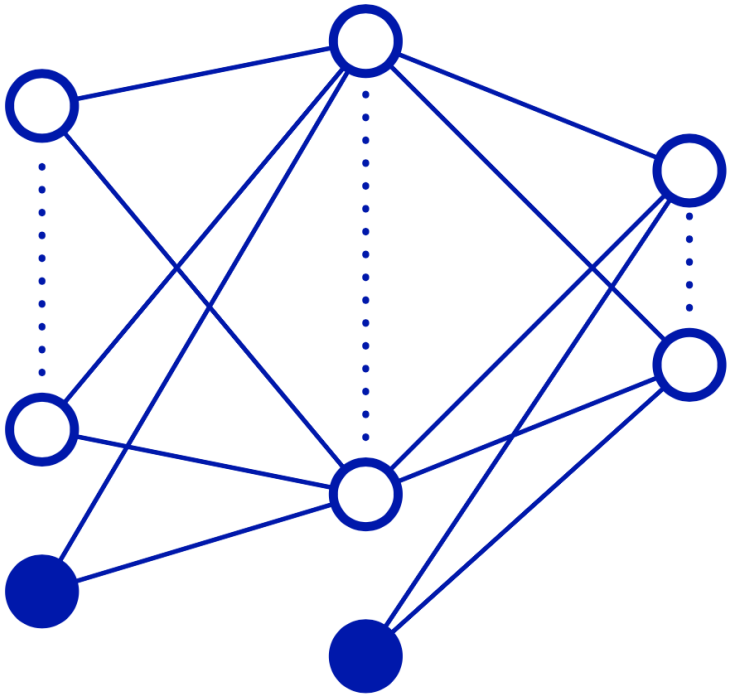
Bayesian PMF

- Predictive distribution: consider predicting a rating r^*_{ij} for user i and query movie j:

$$p(r^*_{ij}|R) = \int p(r^*_{ij}|u_i, v_j)p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)d\{U, V\}d\{\boldsymbol{\theta}_U, \boldsymbol{\theta}_V\},$$

  where $p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)$ is the posterior over parameters and hyperparameters
- Exact evaluation of this predictive distribution is analytically intractable
- Posterior distribution $p(U, V, \boldsymbol{\theta}_U, \boldsymbol{\theta}_V|R)$ is complicated and does not have a closed form expression
- Need to approximate

Bayesian Neural Nets



Bayesian Neural Nets

- Regression problem: given a set of i.i.d observations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^m$ with corresponding targets $\mathbf{Y} = \{y_i\}_{i=1}^m$
- Likelihood

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{m} \mathcal{N}(y_i | f(\mathbf{x}_i, \mathbf{w}), \beta^{-1}),$$

where the mean is given by the output of the neural network

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M} w_{kj}^2 \sigma(\sum_{i=0}^{d} w_{ji}^1 x_i), \; \sigma(z) = (1 + e^{-z})^{-1}$$

- Gaussian prior over the network parameters: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathrm{I})$. Posterior is analytically intractable

$$p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})}{\int p(\mathbf{Y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})d\mathbf{w}}$$

Bayesian Neural Nets

- Regression problem: given a set of i.i.d observations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^m$ with corresponding targets $\mathbf{Y} = \{y_i\}_{i=1}^m$
- Likelihood

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{m} \mathcal{N}(y_i|f(\mathbf{x}_i, \mathbf{w}), \beta^{-1}),$$

where the mean is given by the output of the neural network

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M} w_{kj}^2 \sigma(\sum_{i=0}^{d} w_{ji}^1 x_i),\ \sigma(z) = (1 + e^{-z})^{-1}$$

- Gaussian prior over the network parameters: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathrm{I})$. Posterior is analytically intractable

$$p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})}{\int p(\mathbf{Y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})d\mathbf{w}}$$

Bayesian Neural Nets

- Regression problem: given a set of i.i.d observations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^m$ with corresponding targets $\mathbf{Y} = \{y_i\}_{i=1}^m$
- Likelihood

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{m} \mathcal{N}(y_i|f(\mathbf{x}_i, \mathbf{w}), \beta^{-1}),$$

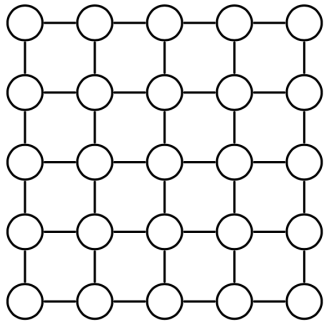
where the mean is given by the output of the neural network

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M} w_{kj}^2 \sigma(\sum_{i=0}^{d} w_{ji}^1 x_i),\ \sigma(z) = (1 + e^{-z})^{-1}$$

- Gaussian prior over the network parameters: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathrm{I})$. Posterior is analytically intractable

$$p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})}{\int p(\mathbf{Y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})d\mathbf{w}}$$

Undirected Models



- $\mathbf{x}$ is a binary random vector with $x_i \in \{+1, -1\}$

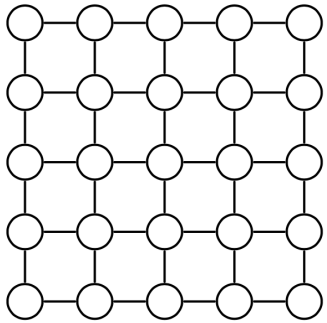
$$p(\mathbf{x}) = \mathcal{Z}^{-1} \exp\Big(\sum_{(i,j)\in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \Big),$$

where $\mathcal{Z}$ is known as partition function

$$\mathcal{Z} = \sum_{\mathbf{x}} \exp\Big(\sum_{(i,j)\in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \Big)$$

- If $\mathbf{x}$ is 100-dimensional, need to sum over 2^{100} terms. The sum might decompose. Otherwise we need to approximate

Undirected Models



- $\mathbf{x}$ is a binary random vector with $x_i \in \{+1, -1\}$

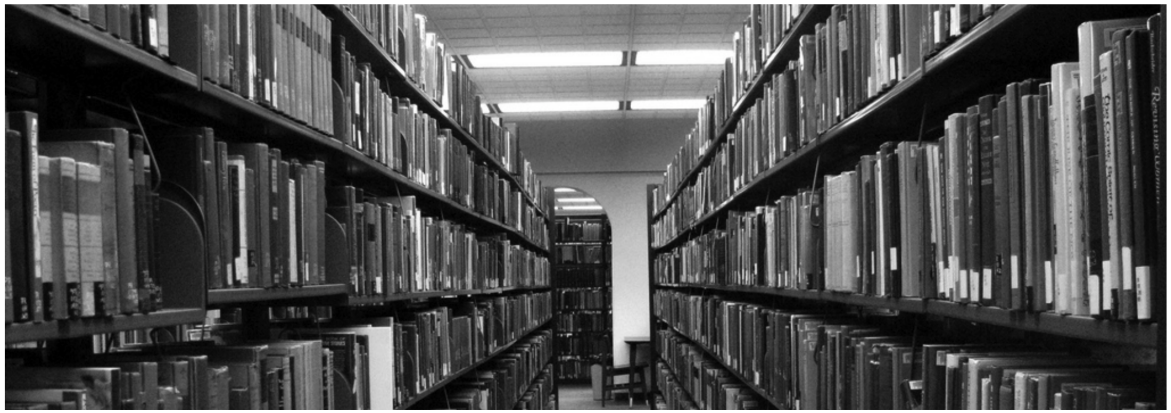
$$p(\mathbf{x}) = \mathcal{Z}^{-1} \exp\Big(\sum_{(i,j)\in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \Big),$$

where $\mathcal{Z}$ is known as partition function

$$\mathcal{Z} = \sum_{\mathbf{x}} \exp\Big(\sum_{(i,j)\in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \Big)$$

- If $\mathbf{x}$ is 100-dimensional, need to sum over 2^{100} terms. The sum might decompose. Otherwise we need to approximate

Topic Modeling



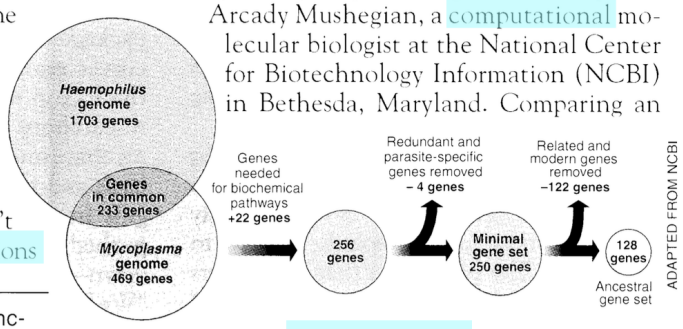
Topic models use posterior inference to discover the hidden thematic structure in a large collection of documents

Latent Dirichlet Allocation (LDA)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



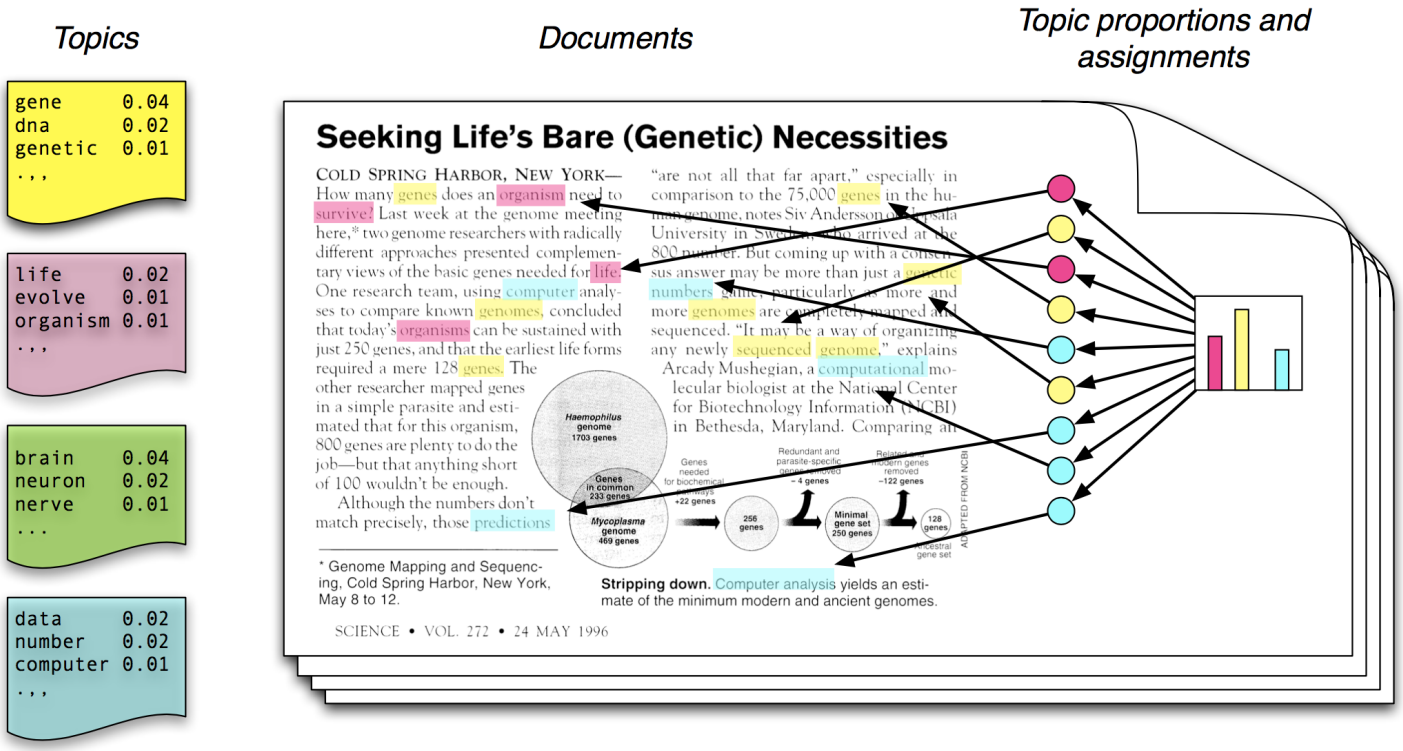
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

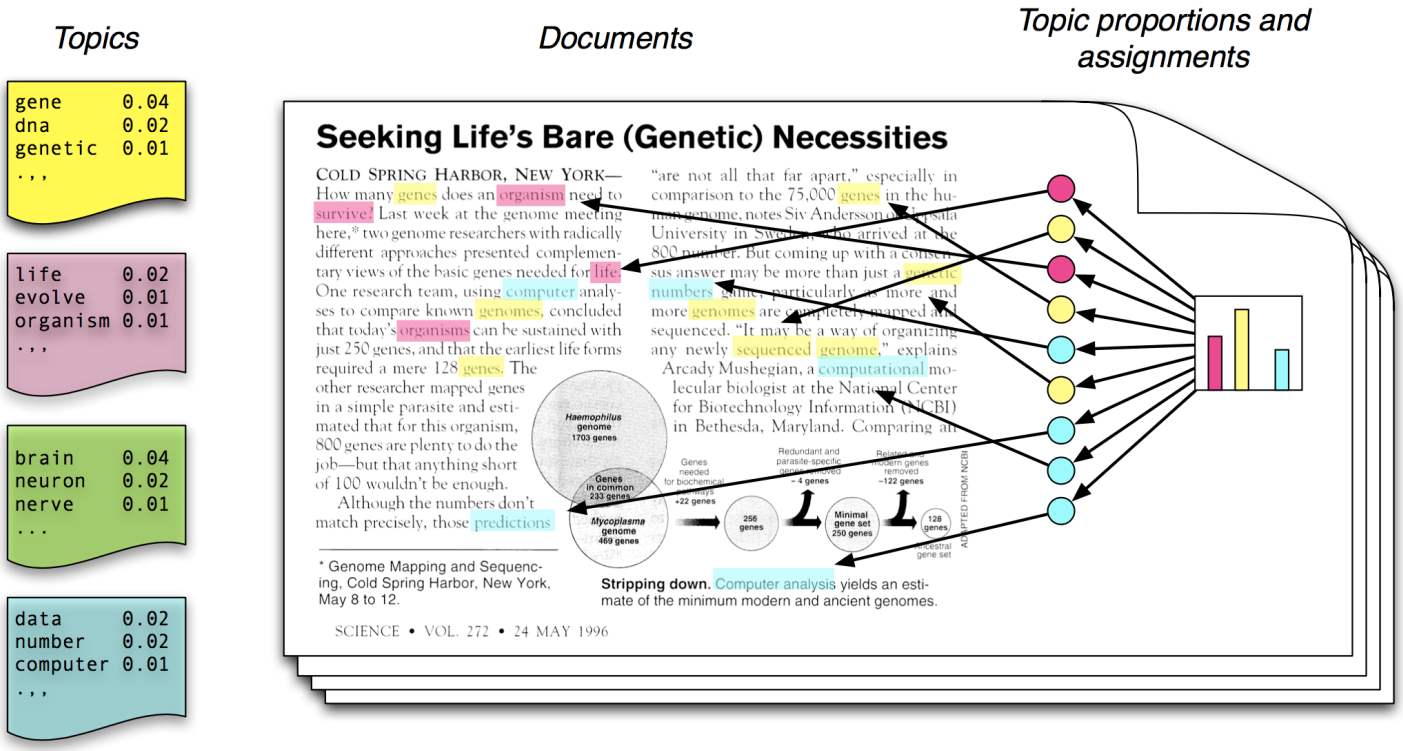
Documents exhibit multiple topics

Latent Dirichlet Allocation (LDA)



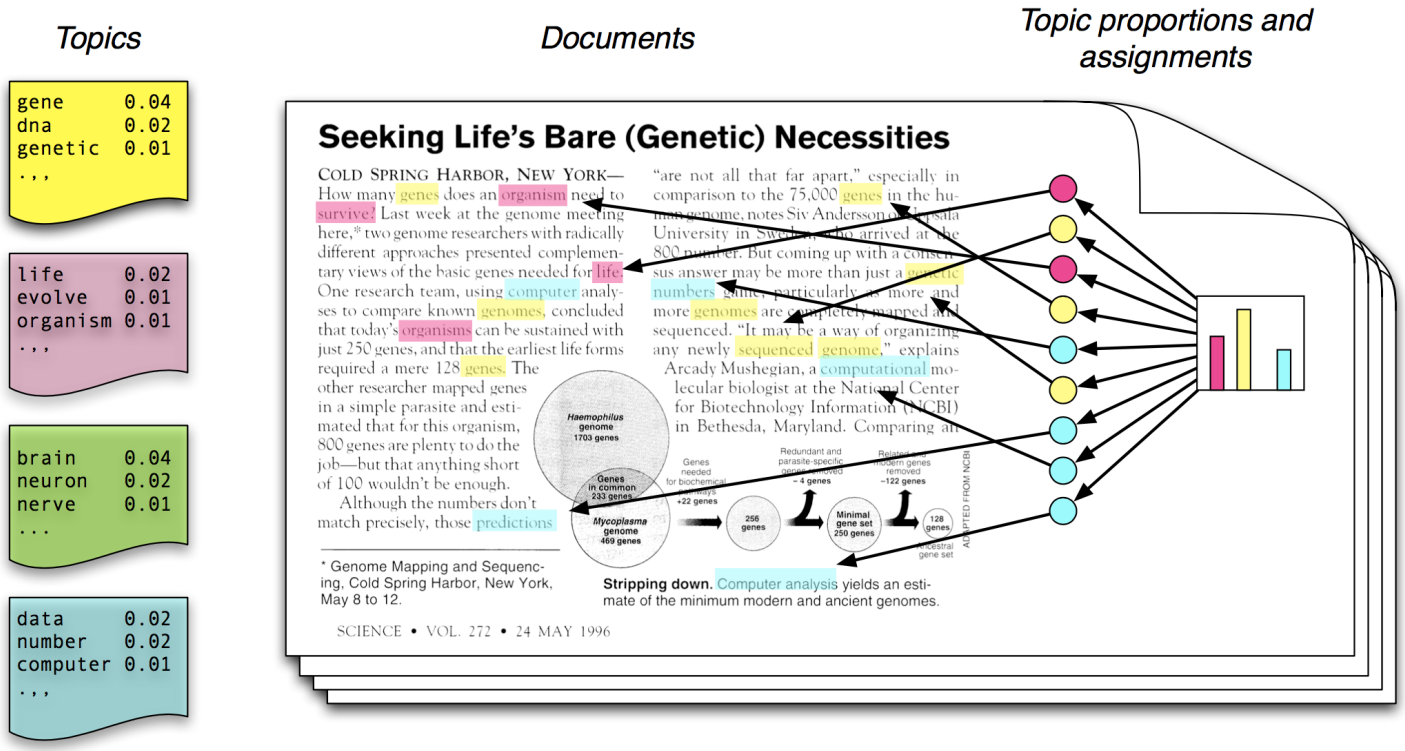
- Each topic is a distribution over words
- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics

Latent Dirichlet Allocation (LDA)



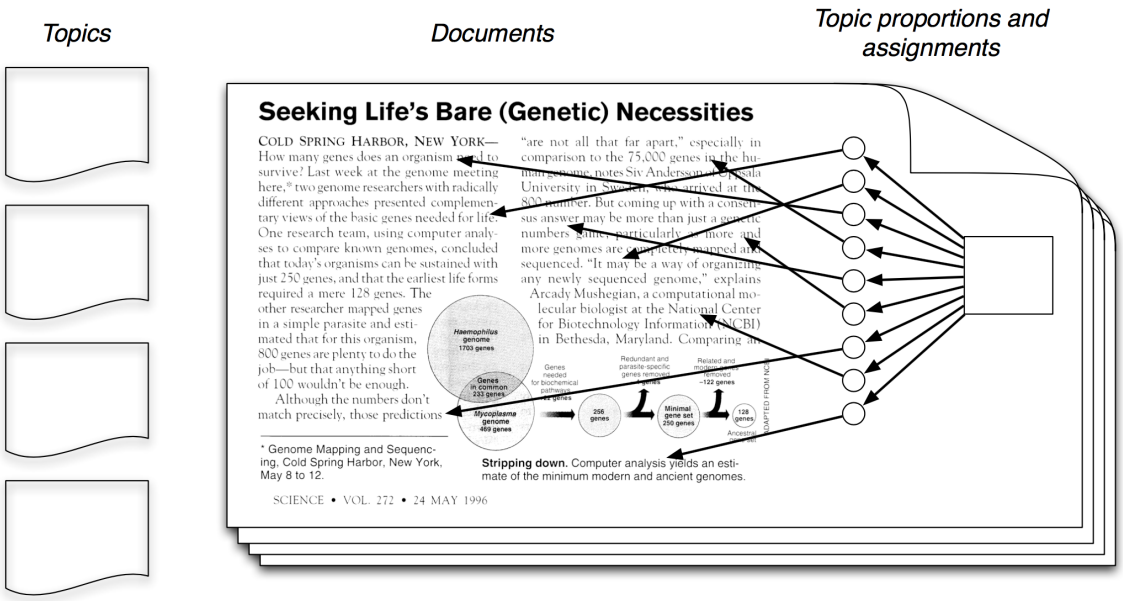
- Each topic is a distribution over words
- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics

Latent Dirichlet Allocation (LDA)



- Each topic is a distribution over words
- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics

Latent Dirichlet Allocation (LDA)

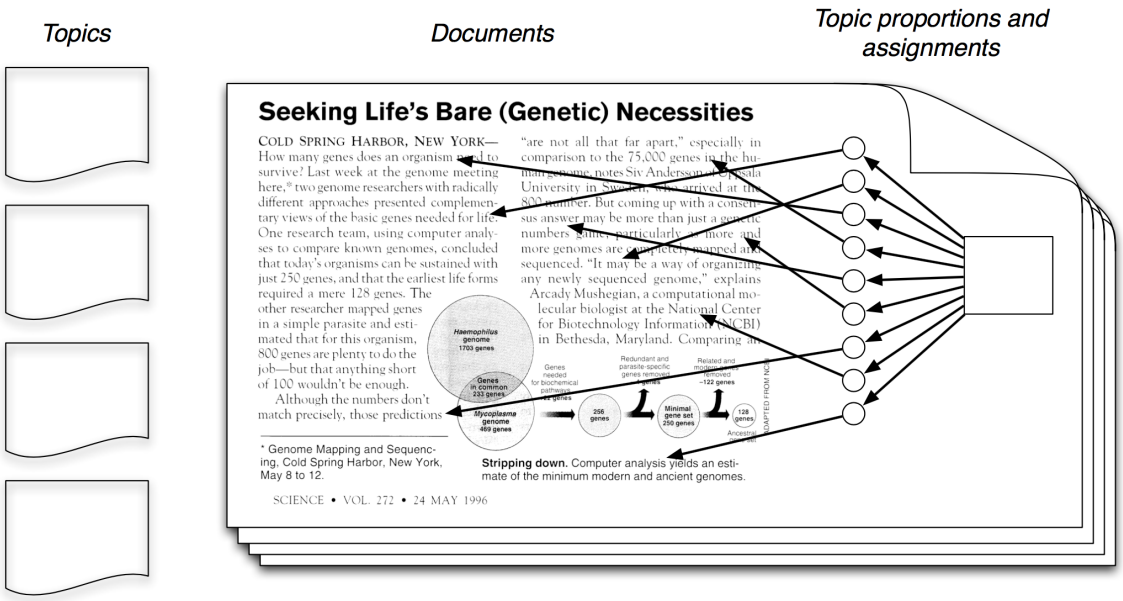


- But we only observe the documents; everything else is hidden
- So we want to calculate the posterior

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

(Note: millions of documents; billions of latent variables)

Latent Dirichlet Allocation (LDA)

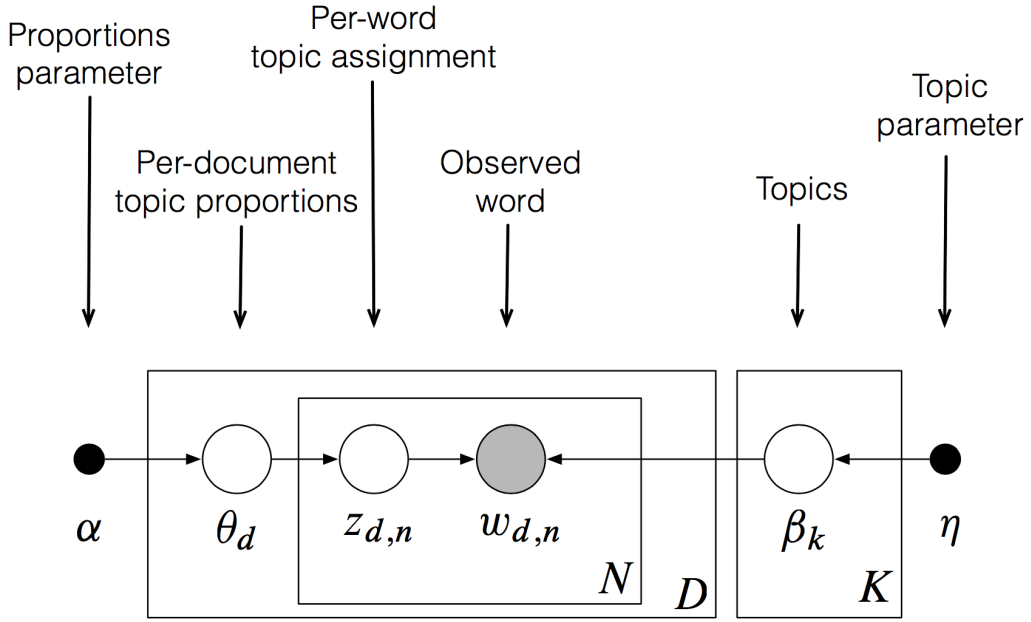


- But we only observe the documents; everything else is hidden
- So we want to calculate the posterior

$$p(\text{topics, proportions, assignments} \,|\, \text{documents})$$

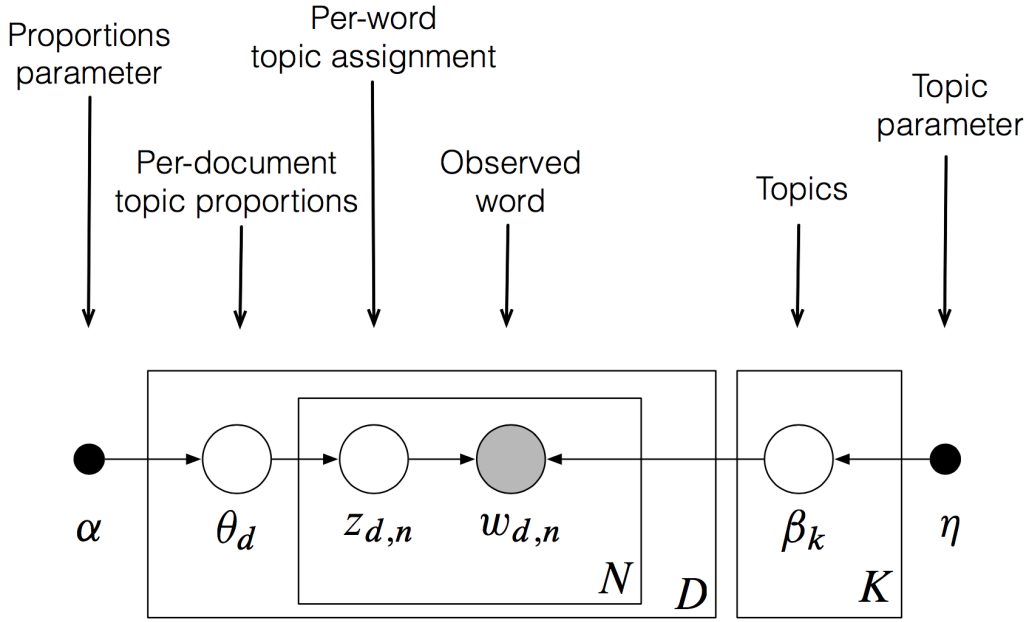
(Note: millions of documents; billions of latent variables)

LDA as a Graphical Model



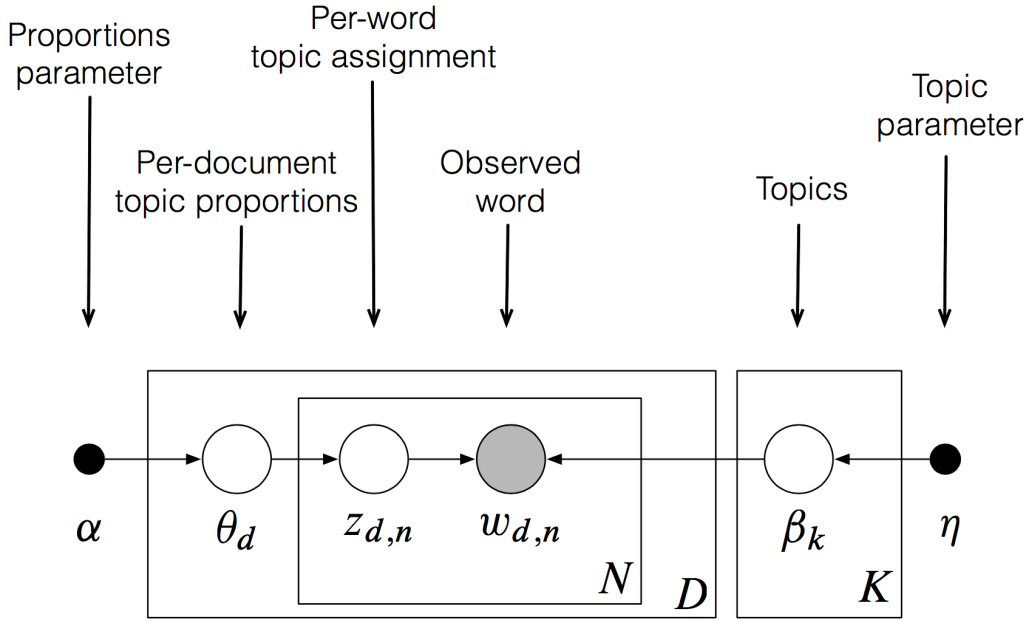
- Encodes assumptions about data with a factorization of the joint
- Connects assumptions to algorithms for computing with data
- Defines the posterior (through the joint)

LDA as a Graphical Model



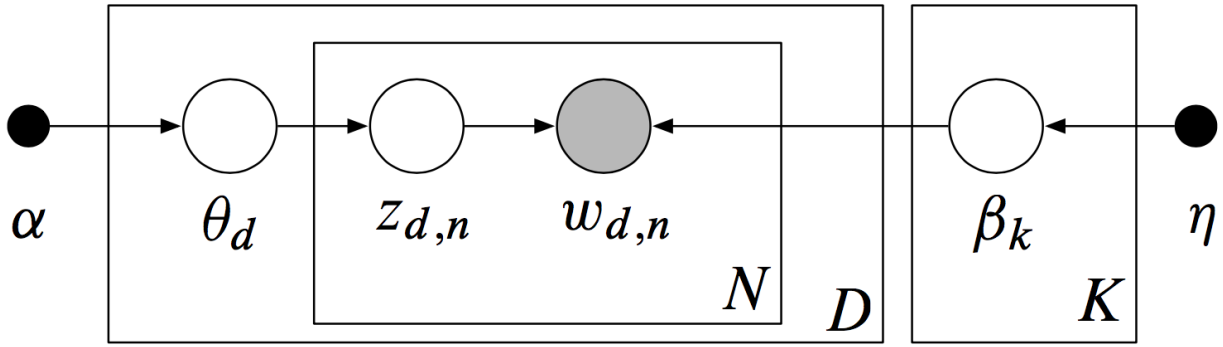
- Encodes assumptions about data with a factorization of the joint
- Connects assumptions to algorithms for computing with data
- Defines the posterior (through the joint)

LDA as a Graphical Model



- Encodes assumptions about data with a factorization of the joint
- Connects assumptions to algorithms for computing with data
- Defines the posterior (through the joint)

LDA as a Graphical Model

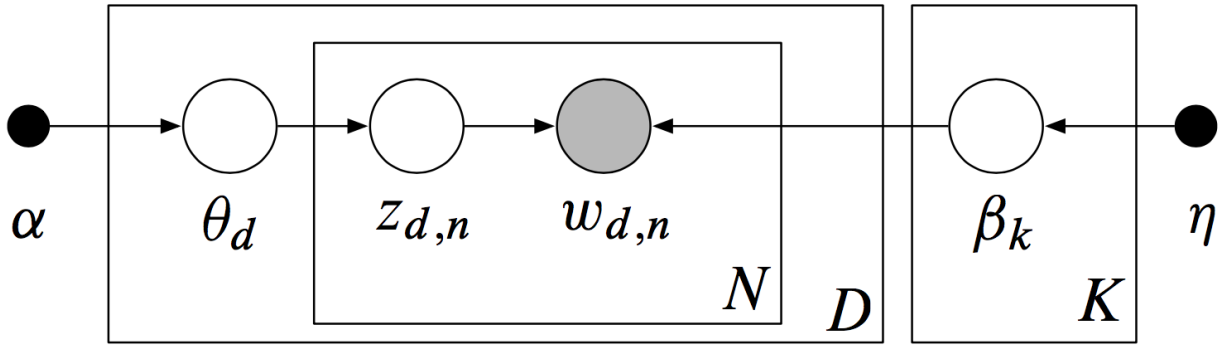


- The posterior of the latent variables given the documents is

$$p(\beta, \boldsymbol{\theta}, \mathbf{z}|\mathbf{w}) = \frac{p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{\int_\beta \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) d\beta d\boldsymbol{\theta}}$$

- We can't compute the denominator, the marginal $p(\mathbf{w})$
- We use approximate inference

LDA as a Graphical Model

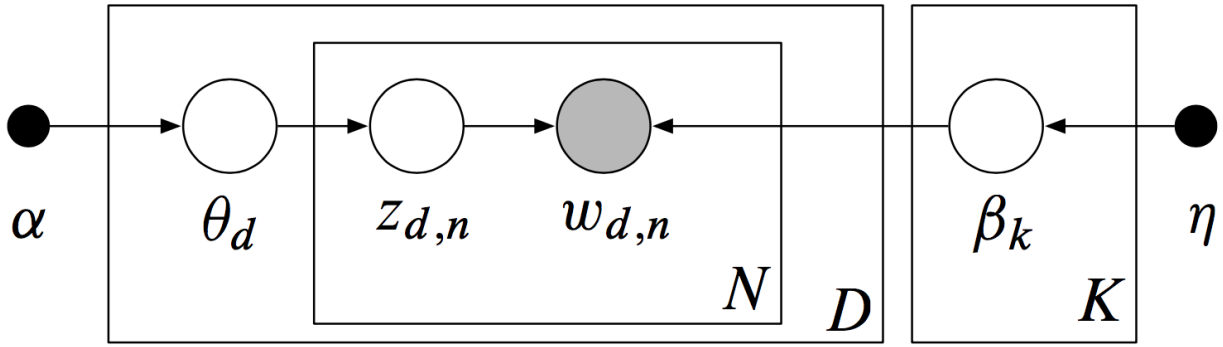


- The posterior of the latent variables given the documents is

$$p(\beta, \boldsymbol{\theta}, \mathbf{z}|\mathbf{w}) = \frac{p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{\int_\beta \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) d\beta d\boldsymbol{\theta}}$$

- We can't compute the denominator, the marginal $p(\mathbf{w})$
- We use approximate inference

LDA as a Graphical Model



- The posterior of the latent variables given the documents is

$$p(\beta, \boldsymbol{\theta}, \mathbf{z}|\mathbf{w}) = \frac{p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{\int_\beta \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) d\beta d\boldsymbol{\theta}}$$

- We can't compute the denominator, the marginal $p(\mathbf{w})$
- We use approximate inference

1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

Topics found in 1.8M articles from the New York Times

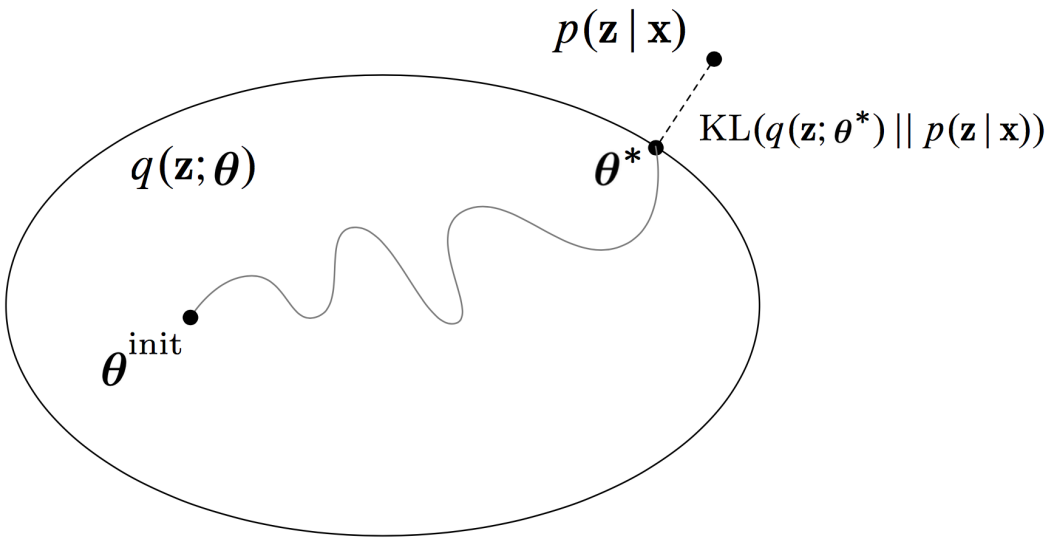
1 Introduction

2 Variational Inference

3 ELBO

4 Mean-field Variational Bayes

Variational Inference

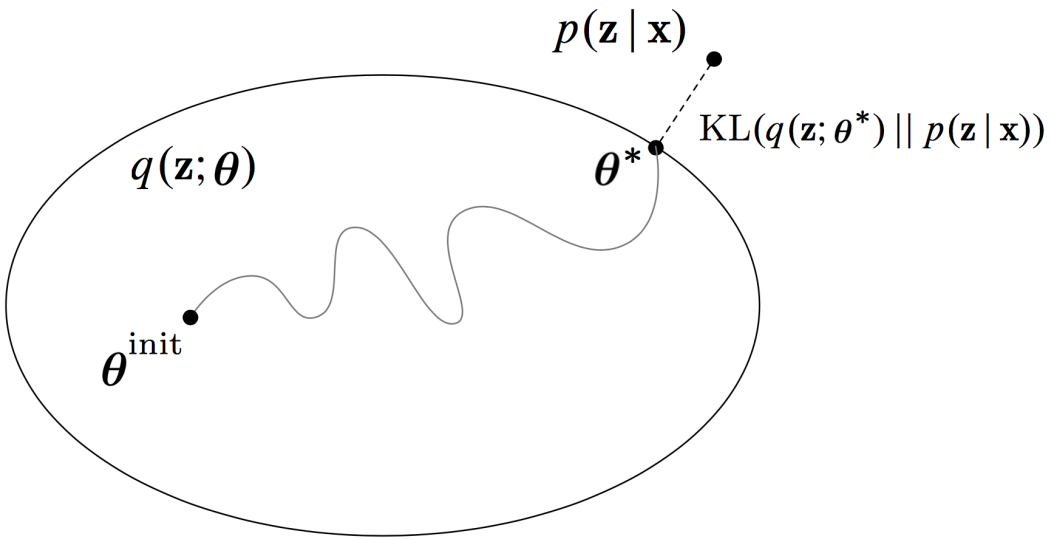


- VI turns inference into optimization.
- Posit a variational family of distributions over the latent variables

$$q(\mathbf{z}; \boldsymbol{\theta})$$

- Fit the variational parameters $\boldsymbol{\theta}$ to be close (in KL) to the exact posterior. (There are alternative divergences, which connect to algorithms like EP, BP, and others.)

Variational Inference

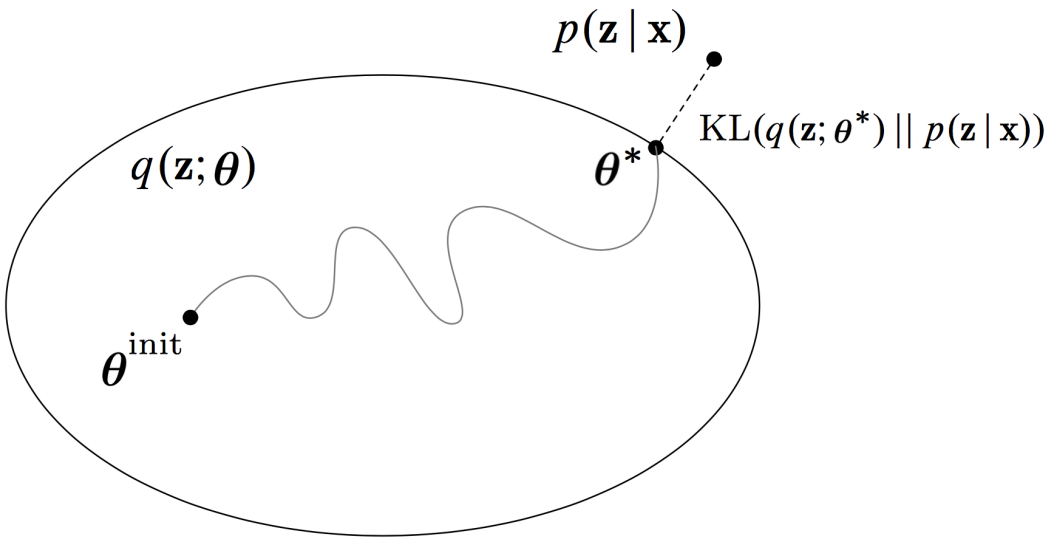


- VI turns inference into optimization.
- Posit a variational family of distributions over the latent variables

$$q(\mathbf{z}; \boldsymbol{\theta})$$

- Fit the variational parameters $\boldsymbol{\theta}$ to be close (in KL) to the exact posterior. (There are alternative divergences, which connect to algorithms like EP, BP, and others.)

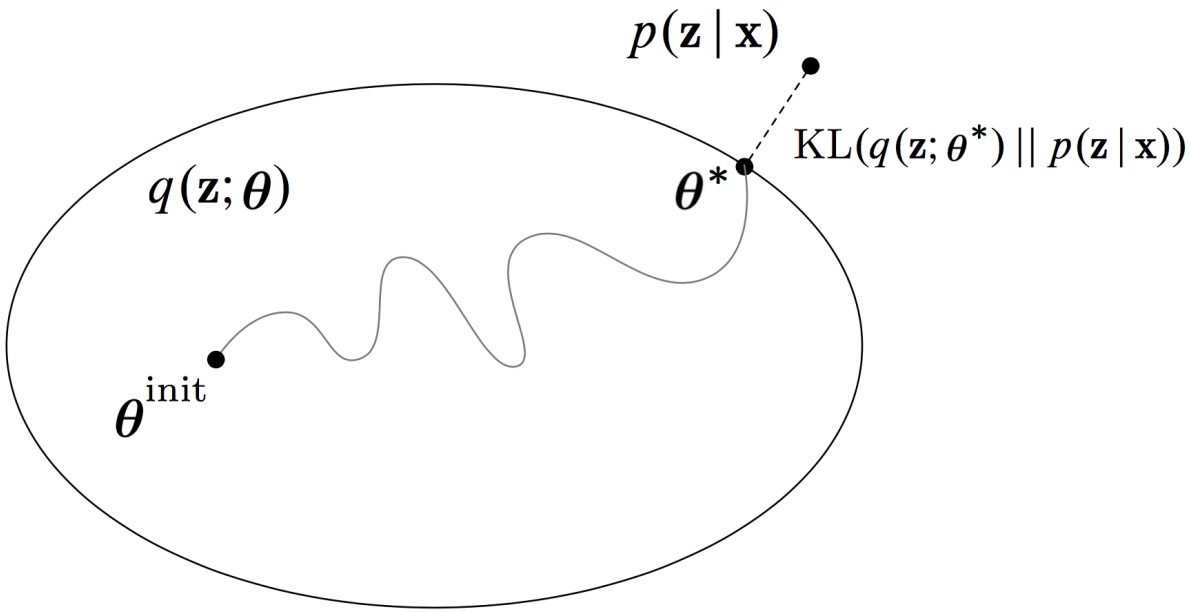
Variational Inference



- VI turns inference into optimization.
- Posit a variational family of distributions over the latent variables

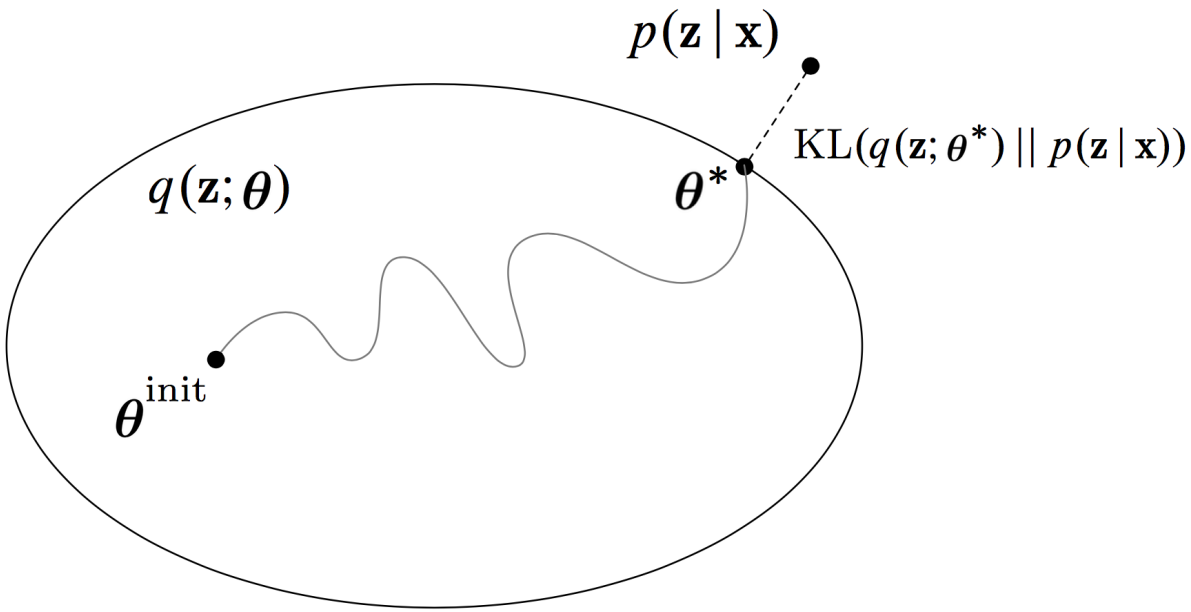
$$q(\mathbf{z}; \boldsymbol{\theta})$$

- Fit the variational parameters $\boldsymbol{\theta}$ to be close (in KL) to the exact posterior. (There are alternative divergences, which connect to algorithms like EP, BP, and others.)



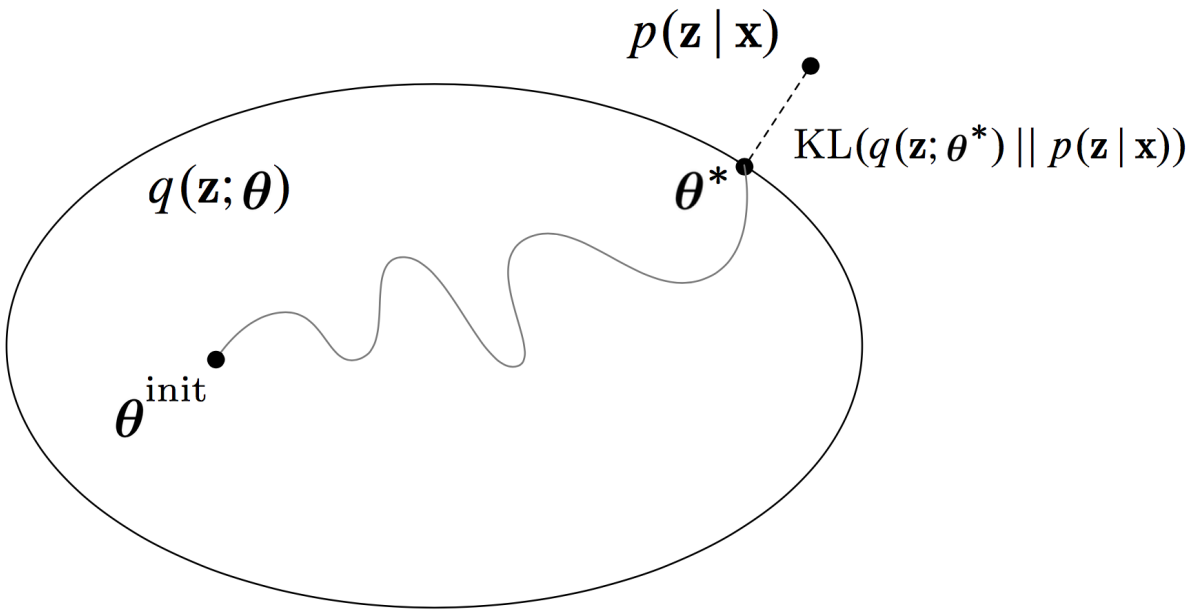
VI approximates difficult quantities from complex models.
With stochastic optimization we can

- scale up VI to massive data
- enable VI on a wide class of difficult models
- enable VI with elaborate and flexible families of approximations



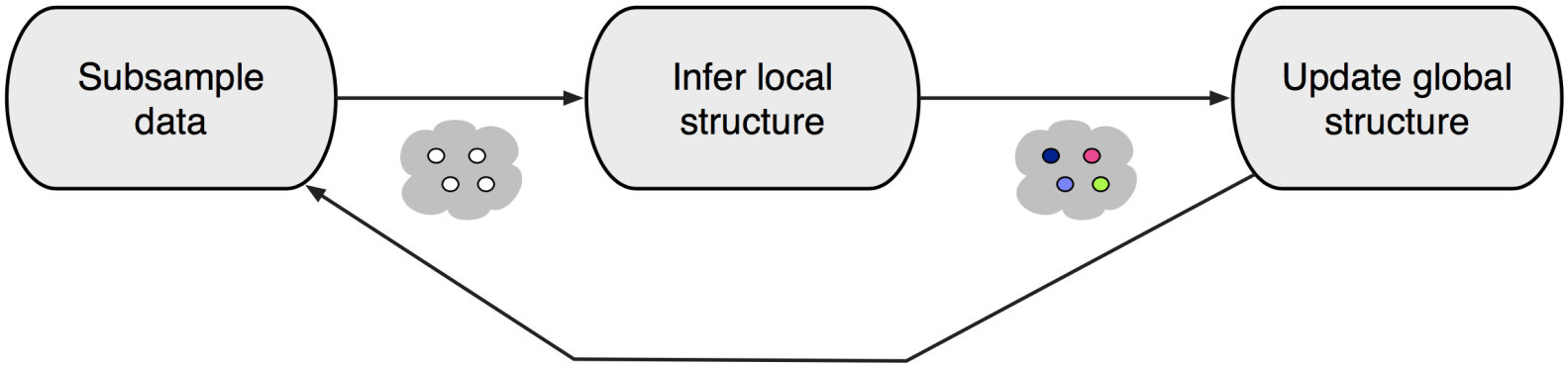
VI approximates difficult quantities from complex models.
With stochastic optimization we can

- scale up VI to massive data
- enable VI on a wide class of difficult models
- enable VI with elaborate and flexible families of approximations



VI approximates difficult quantities from complex models.
With stochastic optimization we can

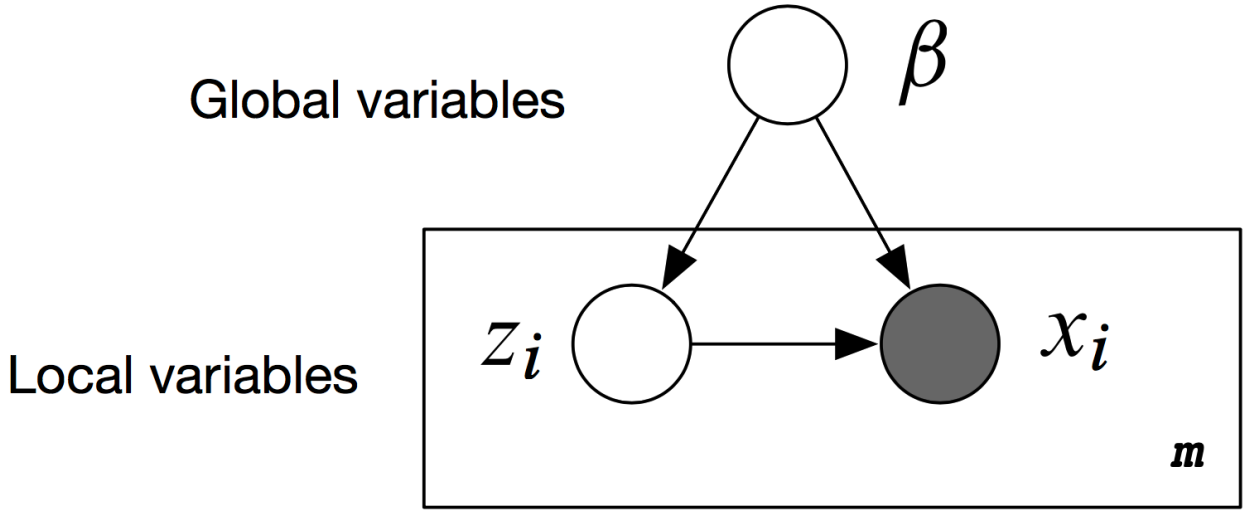
- scale up VI to massive data
- enable VI on a wide class of difficult models
- enable VI with elaborate and flexible families of approximations



Road map:

- Consider variational lower bound criterion
- Derive classical mean-field VI
- Consider VI based SGD approaches
- Consider Variatinal Autoencoder for image modeling
- Consider Application for Multidimensional Time-Series Prediction

A Generic Class of Models



$$p(\beta, \mathbf{Z}, \mathbf{X}) = p(\beta) \prod_{i=1}^{m} p(\mathbf{z}_i, \mathbf{x}_i | \beta)$$

- The observations are $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$
- The local variables are $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$
- The global variables are β
- The i-th data point $\mathbf{x}_i$ only depends on $\mathbf{z}_i$ and β

Compute $p(\beta, \mathbf{Z} | \mathbf{X})$

1 Introduction

2 Variational Inference

3 ELBO

4 Mean-field Variational Bayes

Why KL?

- Variational Bayes for a data sample $\mathcal{D}_m$

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|\mathcal{D}_m))$$

- Evidence Variational Lower Bound (ELBO)

- $KL \geq 0 \Rightarrow$

$$\log p(\mathcal{D}_m) - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq 0$$

- Thus $\log p(\mathcal{D}_m) \geq \mathcal{L}(q)$, and so we define

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

Why KL?

- Variational Bayes for a data sample $\mathcal{D}_m$

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|\mathcal{D}_m))$$

- Evidence Variational Lower Bound (ELBO)

- $KL \geq 0 \Rightarrow$

$$\log p(\mathcal{D}_m) - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq 0$$

- Thus $\log p(\mathcal{D}_m) \geq \mathcal{L}(q)$, and so we define

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

Why KL?

- Variational Bayes for a data sample $\mathcal{D}_m$

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|\mathcal{D}_m))$$

- Evidence Variational Lower Bound (ELBO)

$$KL(q(\cdot)||p(\cdot|\mathcal{D}_m)) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D}_m)} d\boldsymbol{\theta}$$

- $KL \geq 0 \Rightarrow$

$$\log p(\mathcal{D}_m) - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq 0$$

- Thus $\log p(\mathcal{D}_m) \geq \mathcal{L}(q)$, and so we define

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

Why KL?

- Variational Bayes for a data sample $\mathcal{D}_m$

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|\mathcal{D}_m))$$

- Evidence Variational Lower Bound (ELBO)

$$KL(q(\cdot)||p(\cdot|\mathcal{D}_m)) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D}_m)} d\boldsymbol{\theta}$$

$$= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta}) p(\mathcal{D}_m)}{p(\boldsymbol{\theta}, \mathcal{D}_m)} d\boldsymbol{\theta}$$

- $KL \geq 0 \Rightarrow$

$$\log p(\mathcal{D}_m) - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq 0$$

- Thus $\log p(\mathcal{D}_m) \geq \mathcal{L}(q)$, and so we define

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

Why KL?

- Variational Bayes for a data sample $\mathcal{D}_m$

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|\mathcal{D}_m))$$

- Evidence Variational Lower Bound (ELBO)

$$KL(q(\cdot)||p(\cdot|\mathcal{D}_m)) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D}_m)} d\boldsymbol{\theta}$$

$$= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta}) p(\mathcal{D}_m)}{p(\boldsymbol{\theta}, \mathcal{D}_m)} d\boldsymbol{\theta}$$

$$= \log p(\mathcal{D}_m) - \underbrace{\int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}}_{\text{ELBO } \mathcal{L}(q)}$$

- $KL \geq 0 \Rightarrow$

$$\log p(\mathcal{D}_m) - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq 0$$

- Thus $\log p(\mathcal{D}_m) \geq \mathcal{L}(q)$, and so we define

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

Why KL?

- Variational Bayes for a data sample $\mathcal{D}_m$

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|\mathcal{D}_m))$$

- Evidence Variational Lower Bound (ELBO)

$$KL(q(\cdot)||p(\cdot|\mathcal{D}_m)) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D}_m)} d\boldsymbol{\theta}$$

$$= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta}) p(\mathcal{D}_m)}{p(\boldsymbol{\theta}, \mathcal{D}_m)} d\boldsymbol{\theta}$$

$$= \log p(\mathcal{D}_m) - \underbrace{\int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}}_{\text{ELBO } \mathcal{L}(q)}$$

- $KL \geq 0 \Rightarrow$

$$\log p(\mathcal{D}_m) - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq 0$$

- Thus $\log p(\mathcal{D}_m) \geq \mathcal{L}(q)$, and so we define

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

Why KL?

- Variational Bayes for a data sample $\mathcal{D}_m$

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|\mathcal{D}_m))$$

- Evidence Variational Lower Bound (ELBO)

$$KL(q(\cdot)||p(\cdot|\mathcal{D}_m)) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D}_m)} d\boldsymbol{\theta}$$

$$= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta}) p(\mathcal{D}_m)}{p(\boldsymbol{\theta}, \mathcal{D}_m)} d\boldsymbol{\theta}$$

$$= \log p(\mathcal{D}_m) - \underbrace{\int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}}_{\text{ELBO } \mathcal{L}(q)}$$

- $KL \geq 0 \Rightarrow$

$$\log p(\mathcal{D}_m) - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, \mathcal{D}_m)}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \geq 0$$

- Thus $\log p(\mathcal{D}_m) \geq \mathcal{L}(q)$, and so we define

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

Variational Bayes

- As Q we should choose some "nice" distributions to get tractable optimization

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$$

- $q_j(\theta_j)$ is often also from the exponential family

Variational Bayes

- As Q we should choose some "nice" distributions to get tractable optimization

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$$

- $q_j(\theta_j)$ is often also from the exponential family

Variational Bayes

- As Q we should choose some "nice" distributions to get tractable optimization

$$q^* = \arg\max_{q \in Q} \mathcal{L}(q)$$

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$$

- $q_j(\theta_j)$ is often also from the exponential family

Variational Bayes

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$$

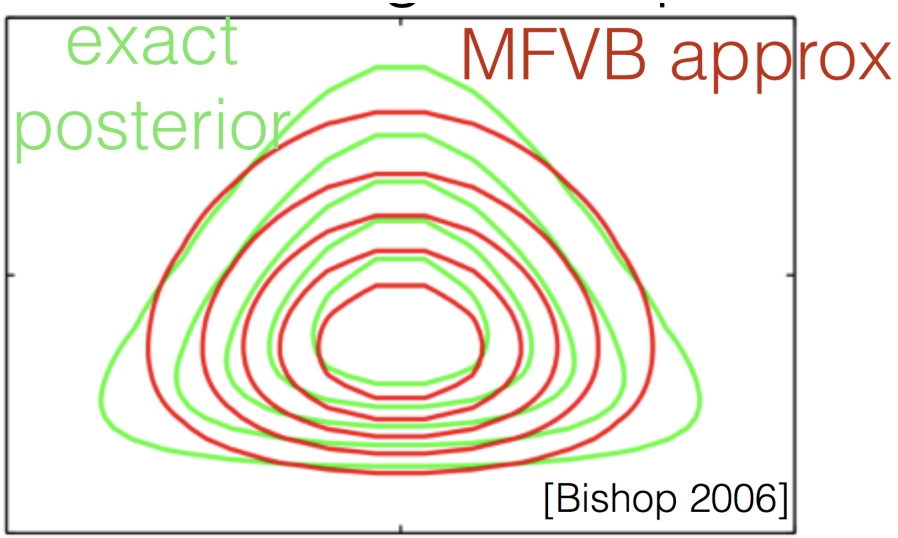
- **Not a modeling assumption!**

- Often coordinate descent in $q_1, \ldots, q_p$ to optimize $\mathcal{L}(q)$ w.r.t. $q \in Q$

- Mean-field variational Bayes (MFVB)

$$Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$$

- **Not a modeling assumption!**

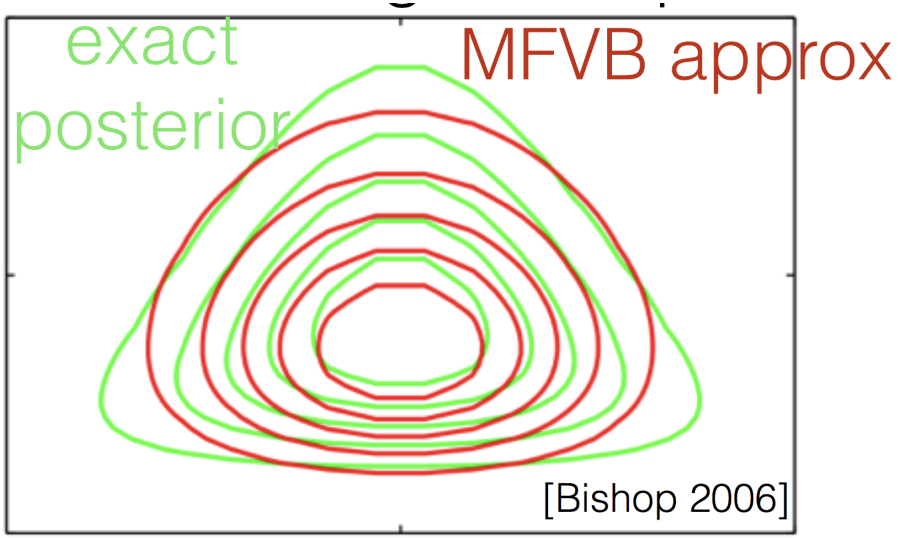


- Often coordinate descent in $q_1, \ldots, q_p$ to optimize $\mathcal{L}(q)$ w.r.t. $q \in Q$

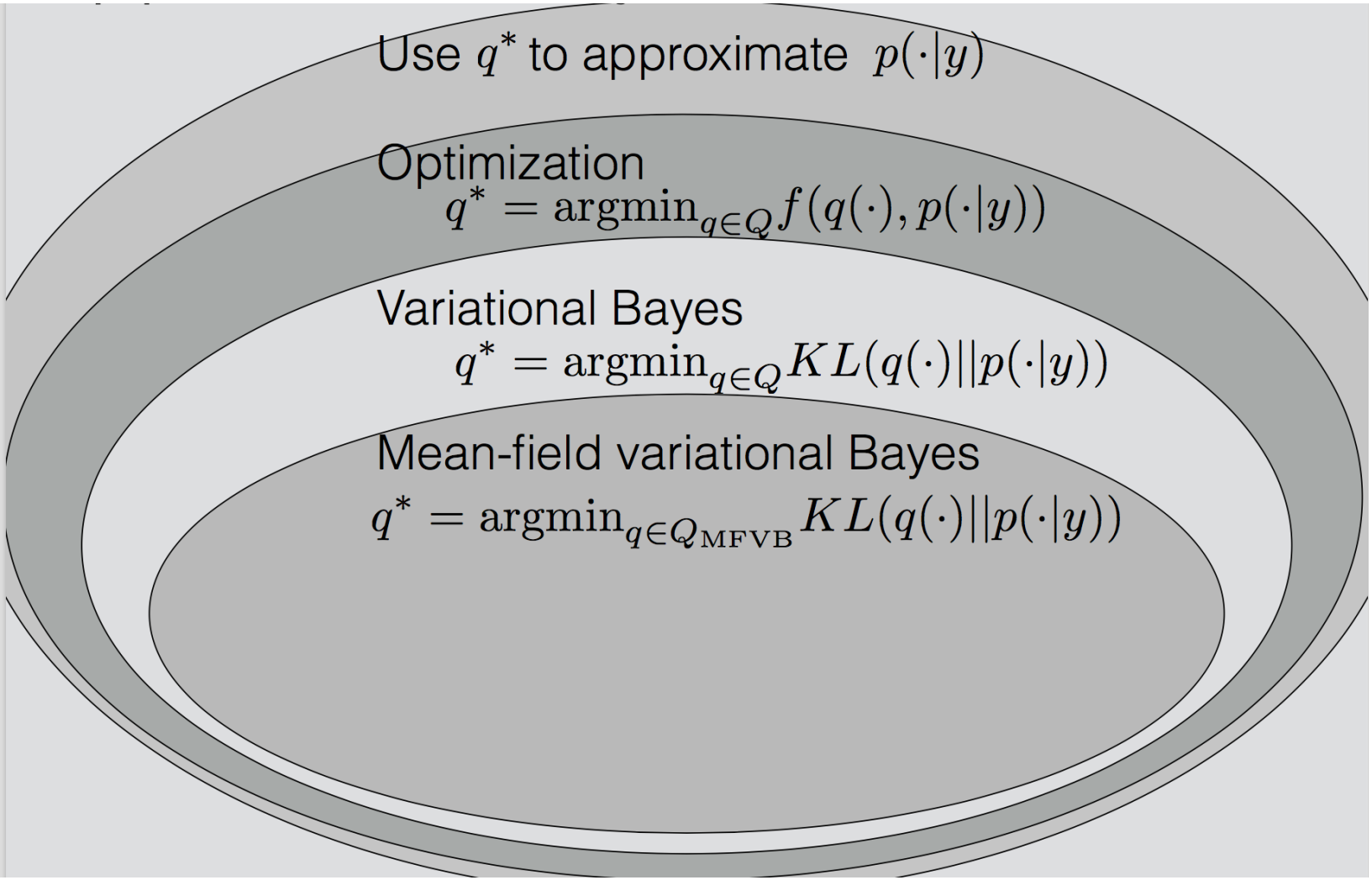
- Mean-field variational Bayes (MFVB)

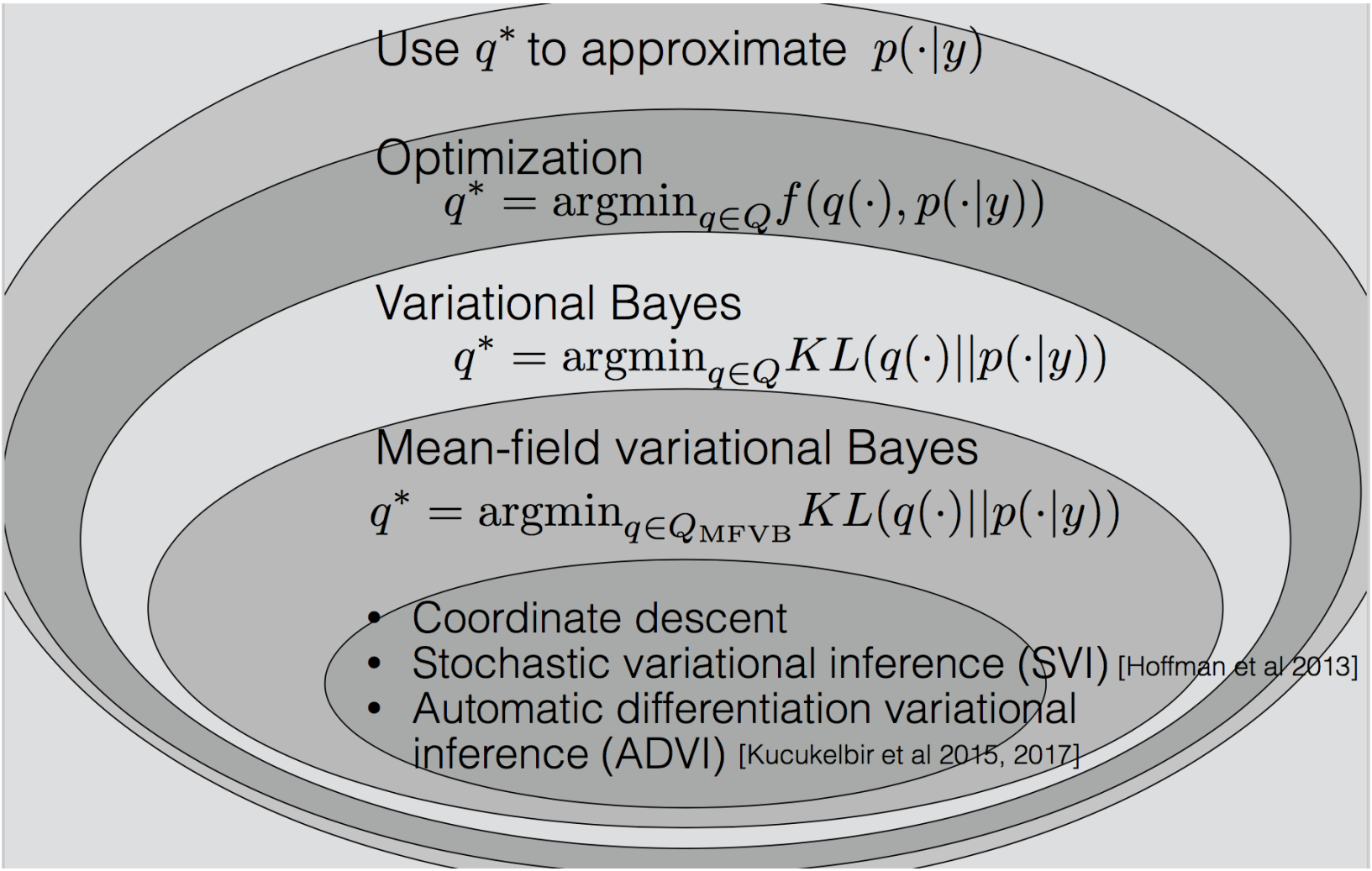
$$Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$$

- **Not a modeling assumption!**



- Often coordinate descent in $q_1, \ldots, q_p$ to optimize $\mathcal{L}(q)$ w.r.t. $q \in Q$


Use q^* to approximate $p(\cdot|y)$
Optimization
$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$
Variational Bayes
$q^* = \mathrm{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$
Mean-field variational Bayes
$q^* = \mathrm{argmin}_{q \in Q_{\mathrm{MFVB}}} KL(q(\cdot)||p(\cdot|y))$


Use q^* to approximate $p(\cdot|y)$
Optimization
$q^* = \mathrm{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$
Variational Bayes
$q^* = \mathrm{argmin}_{q \in Q} KL(q(\cdot)||p(\cdot|y))$
Mean-field variational Bayes
$q^* = \mathrm{argmin}_{q \in Q_{\mathrm{MFVB}}} KL(q(\cdot)||p(\cdot|y))$
• Coordinate descent
• Stochastic variational inference (SVI) [Hoffman et al 2013]
• Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]

1 Introduction

2 Variational Inference

3 ELBO

4 Mean-field Variational Bayes

Mean-field Variational Bayes

- Mean-field variational Bayes (MFVB)
  $Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$
- The variational lower bound takes the form:

- Here we denote by

$$\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] = \int \log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i$$

Mean-field Variational Bayes

- Mean-field variational Bayes (MFVB)
  $Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$
- The variational lower bound takes the form:

- Here we denote by

$$\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] = \int \log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i$$

Mean-field Variational Bayes

- Mean-field variational Bayes (MFVB)
  $Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$
- The variational lower bound takes the form:

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \log p(\mathcal{D}_m, \boldsymbol{\theta}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{1}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

- Here we denote by

$$\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] = \int \log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i$$

Mean-field Variational Bayes

- Mean-field variational Bayes (MFVB)
  $Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$
- The variational lower bound takes the form:

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \log p(\mathcal{D}_m, \boldsymbol{\theta}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{1}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

$$= \int q_j(\theta_j) \underbrace{\left[\log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i\right]}_{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]} d\theta_j +$$

- Here we denote by

$$\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] = \int \log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i$$

Mean-field Variational Bayes

- Mean-field variational Bayes (MFVB)
  $Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$
- The variational lower bound takes the form:

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \log p(\mathcal{D}_m, \boldsymbol{\theta}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{1}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

$$= \int q_j(\theta_j) \underbrace{\left[\log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i\right]}_{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]} d\theta_j +$$

$$+ \int q_j(\theta_j) \log \frac{1}{q(\theta_j)} d\theta_j + \sum_{i \neq j} \int q_i(\theta_i) \log \frac{1}{q(\theta_i)} d\theta_i$$

- Here we denote by

$$\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] = \int \log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i$$

Mean-field Variational Bayes

- Mean-field variational Bayes (MFVB)
  $Q_{MFVB} = \{q : q(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_j(\theta_j)\}$
- The variational lower bound takes the form:

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \log p(\mathcal{D}_m, \boldsymbol{\theta}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{1}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

$$= \int q_j(\theta_j) \underbrace{\left[\log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i\right]}_{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]} d\theta_j +$$

$$+ \int q_j(\theta_j) \log \frac{1}{q(\theta_j)} d\theta_j + \sum_{i \neq j} \int q_i(\theta_i) \log \frac{1}{q(\theta_i)} d\theta_i$$

- Here we denote by

$$\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] = \int \log p(\mathcal{D}_m, \boldsymbol{\theta}) \prod_{i \neq j} q_i(\theta_i) d\theta_i$$

Mean-field Variational Bayes

- Suppose we keep $\{\prod_{i \neq j} q_i(\theta_i)\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$
- Let us introduce a new distribution $\widetilde{q}(\theta_j) = \widetilde{q}(\mathcal{D}_m, \theta_j)$ by the relation
- Optimum of

$$\mathcal{L}(q) = \int q_j(\theta_j) \log \widetilde{q}(\theta_j) d\theta_j + \int q_j(\theta_j) \log \frac{1}{q_j(\theta_j)} d\theta_j + \text{const} \to \max_{q_j(\cdot)}$$

Mean-field Variational Bayes

- Suppose we keep $\{\prod_{i \neq j} q_i(\theta_i)\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$
- Let us introduce a new distribution $\widetilde{q}(\theta_j) = \widetilde{q}(\mathcal{D}_m, \theta_j)$ by the relation

- Optimum of

$$\mathcal{L}(q) = \int q_j(\theta_j) \log \widetilde{q}(\theta_j) d\theta_j + \int q_j(\theta_j) \log \frac{1}{q_j(\theta_j)} d\theta_j + \text{const} \to \max_{q_j(\cdot)}$$

Mean-field Variational Bayes

- Suppose we keep $\{\prod_{i \neq j} q_i(\theta_i)\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$
- Let us introduce a new distribution $\widetilde{q}(\theta_j) = \widetilde{q}(\mathcal{D}_m, \theta_j)$ by the relation

$$\log \widetilde{q}(\theta_j) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] + \text{const}, \quad \text{or}$$

- Optimum of

$$\mathcal{L}(q) = \int q_j(\theta_j) \log \widetilde{q}(\theta_j) d\theta_j + \int q_j(\theta_j) \log \frac{1}{q_j(\theta_j)} d\theta_j + \text{const} \to \max_{q_j(\cdot)}$$

Mean-field Variational Bayes

- Suppose we keep $\{\prod_{i \neq j} q_i(\theta_i)\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$
- Let us introduce a new distribution $\widetilde{q}(\theta_j) = \widetilde{q}(\mathcal{D}_m, \theta_j)$ by the relation

$$\log \widetilde{q}(\theta_j) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] + \text{const}, \quad \text{or}$$
$$\widetilde{q}(\theta_j) \sim \exp\{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]\}$$

Here const is a normalizing constant

- Optimum of

$$\mathcal{L}(q) = \int q_j(\theta_j) \log \widetilde{q}(\theta_j) d\theta_j + \int q_j(\theta_j) \log \frac{1}{q_j(\theta_j)} d\theta_j + \text{const} \to \max_{q_j(\cdot)}$$

Mean-field Variational Bayes

- Suppose we keep $\{\prod_{i \neq j} q_i(\theta_i)\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$
- Let us introduce a new distribution $\widetilde{q}(\theta_j) = \widetilde{q}(\mathcal{D}_m, \theta_j)$ by the relation

$$\log \widetilde{q}(\theta_j) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] + \text{const}, \quad \text{or}$$
$$\widetilde{q}(\theta_j) \sim \exp\{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]\}$$

  Here const is a normalizing constant

- Optimum of

$$\mathcal{L}(q) = \int q_j(\theta_j) \log \widetilde{q}(\theta_j) d\theta_j + \int q_j(\theta_j) \log \frac{1}{q_j(\theta_j)} d\theta_j + \text{const} \to \max_{q_j(\cdot)}$$

Mean-field Variational Bayes

- Suppose we keep $\{\prod_{i \neq j} q_i(\theta_i)\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$
- Let us introduce a new distribution $\widetilde{q}(\theta_j) = \widetilde{q}(\mathcal{D}_m, \theta_j)$ by the relation

$$\log \widetilde{q}(\theta_j) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] + \text{const}, \quad \text{or}$$
$$\widetilde{q}(\theta_j) \sim \exp\left\{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]\right\}$$

  Here const is a normalizing constant

- Optimum of

$$\mathcal{L}(q) = \int q_j(\theta_j) \log \widetilde{q}(\theta_j) d\theta_j + \int q_j(\theta_j) \log \frac{1}{q_j(\theta_j)} d\theta_j + \text{const} \to \max_{q_j(\cdot)}$$

  is obtained on

$$q_j^*(\theta_j) = \widetilde{q}(\theta_j)$$

Mean-field Variational Bayes

- Thus, if all $q_i(\theta_i)$ are fixed except $q_j(\theta_j)$, then the optimal $q_j^*(\theta_j)$ has the form

- First we initialize all factors $q_i(\theta_i)$
- Cycle through the factors and replace each in turn with a revised estimate

Mean-field Variational Bayes

- Thus, if all $q_i(\theta_i)$ are fixed except $q_j(\theta_j)$, then the optimal $q_j^*(\theta_j)$ has the form

$$\log q_j^*(\theta_j) = \log \widetilde{q}_j(\theta_j) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] + \text{const},$$

- First we initialize all factors $q_i(\theta_i)$
- Cycle through the factors and replace each in turn with a revised estimate

Mean-field Variational Bayes

- Thus, if all $q_i(\theta_i)$ are fixed except $q_j(\theta_j)$, then the optimal $q_j^*(\theta_j)$ has the form

$$\log q_j^*(\theta_j) = \log \widetilde{q}_j(\theta_j) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] + \text{const},$$
$$q_j^*(\theta_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] d\theta_j\right)}$$

- First we initialize all factors $q_i(\theta_i)$
- Cycle through the factors and replace each in turn with a revised estimate

Mean-field Variational Bayes

- Thus, if all $q_i(\theta_i)$ are fixed except $q_j(\theta_j)$, then the optimal $q_j^*(\theta_j)$ has the form

$$\log q_j^*(\theta_j) = \log \widetilde{q}_j(\theta_j) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] + \text{const},$$
$$q_j^*(\theta_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] d\theta_j\right)}$$

- First we initialize all factors $q_i(\theta_i)$
- Cycle through the factors and replace each in turn with a revised estimate

- Thus, if all $q_i(\theta_i)$ are fixed except $q_j(\theta_j)$, then the optimal $q_j^*(\theta_j)$ has the form

$$\log q_j^*(\theta_j) = \log \widetilde{q}_j(\theta_j) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] + \text{const},$$
$$q_j^*(\theta_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}_m, \boldsymbol{\theta})] d\theta_j\right)}$$

- First we initialize all factors $q_i(\theta_i)$
- Cycle through the factors and replace each in turn with a revised estimate

Example: 1d Gaussian model for Midge wing length

[CSIRO 2004]

- Catalogued midge wing lengths (mm), $\mathcal{D}_m = \{x_1, \ldots, x_m\}$
- Parameters: population mean and variance $\boldsymbol{\theta} = (\mu, \sigma^2)$. For convenience we define $\tau = (\sigma^2)^{-1}$
- Model:

$$p(\mathcal{D}_m|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{m/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{m}(x_i - \mu)^2\right\}$$

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}),\ p(\tau) = \Gamma(\tau|a_0, b_0)$$

Example: 1d Gaussian model for Midge wing length

[CSIRO 2004]

- Catalogued midge wing lengths (mm), $\mathcal{D}_m = \{x_1, \ldots, x_m\}$
- Parameters: population mean and variance $\boldsymbol{\theta} = (\mu, \sigma^2)$. For convenience we define $\tau = (\sigma^2)^{-1}$
- Model:

$$p(\mathcal{D}_m|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{m/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{m}(x_i - \mu)^2\right\}$$

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}),\ p(\tau) = \Gamma(\tau|a_0, b_0)$$

Example: 1d Gaussian model for Midge wing length

[CSIRO 2004]

- Catalogued midge wing lengths (mm), $\mathcal{D}_m = \{x_1, \ldots, x_m\}$
- Parameters: population mean and variance $\boldsymbol{\theta} = (\mu, \sigma^2)$. For convenience we define $\tau = (\sigma^2)^{-1}$
- Model:

$$p(\mathcal{D}_m|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{m/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{m}(x_i - \mu)^2\right\}$$

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}),\ p(\tau) = \Gamma(\tau|a_0, b_0)$$

Example: 1d Gaussian model for Midge wing length

- Variational apprximation: $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$
- We get
- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Variational apprximation: $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$
- We get

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Variational apprximation: $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$
- We get

$$\log q_\mu^*(\mu) = \mathbb{E}_\tau [\log p(\mathcal{D}_m | \mu, \tau) + \log p(\mu | \tau)] + \text{const}$$

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Variational apprximation: $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$
- We get

$$\log q_\mu^*(\mu) = \mathbb{E}_\tau[\log p(\mathcal{D}_m|\mu, \tau) + \log p(\mu|\tau)] + \text{const}$$
$$= -\frac{\mathbb{E}[\tau]}{2}\left[\sum_{i=1}^{m}(x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] + \text{const}$$

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Variational apprximation: $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$
- We get

$$\log q_\mu^*(\mu) = \mathbb{E}_\tau[\log p(\mathcal{D}_m|\mu, \tau) + \log p(\mu|\tau)] + \text{const}$$
$$= -\frac{\mathbb{E}[\tau]}{2}\left[\sum_{i=1}^{m}(x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] + \text{const}$$

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Variational apprximation: $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$
- We get

$$\log q_\mu^*(\mu) = \mathbb{E}_\tau[\log p(\mathcal{D}_m|\mu, \tau) + \log p(\mu|\tau)] + \text{const}$$
$$= -\frac{\mathbb{E}[\tau]}{2}\left[\sum_{i=1}^{m}(x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] + \text{const}$$

- Thus the solution is

$$q_\mu(\mu) = \mathcal{N}(\mu|\mu_m, \lambda_m^{-1})$$

Example: 1d Gaussian model for Midge wing length

- Variational apprximation: $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$
- We get

$$\log q_\mu^*(\mu) = \mathbb{E}_\tau[\log p(\mathcal{D}_m|\mu, \tau) + \log p(\mu|\tau)] + \text{const}$$
$$= -\frac{\mathbb{E}[\tau]}{2}\left[\sum_{i=1}^m (x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] + \text{const}$$

- Thus the solution is

$$q_\mu(\mu) = \mathcal{N}(\mu|\mu_m, \lambda_m^{-1})$$
$$\mu_m = \frac{\lambda_0\mu_0 + m\overline{x}_m}{\lambda_0 + m}, \;\; \lambda_m = (\lambda_0 + m)\mathbb{E}[\tau]$$

Example: 1d Gaussian model for Midge wing length

- Variational apprximation: $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$
- We get

$$\log q_\mu^*(\mu) = \mathbb{E}_\tau[\log p(\mathcal{D}_m|\mu, \tau) + \log p(\mu|\tau)] + \text{const}$$
$$= -\frac{\mathbb{E}[\tau]}{2}\left[\sum_{i=1}^m (x_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] + \text{const}$$

- Thus the solution is

$$q_\mu(\mu) = \mathcal{N}(\mu|\mu_m, \lambda_m^{-1})$$
$$\mu_m = \frac{\lambda_0\mu_0 + m\overline{x}_m}{\lambda_0 + m}, \;\; \lambda_m = (\lambda_0 + m)\mathbb{E}[\tau]$$
$$\overline{x}_m = \frac{1}{m}\sum_{i=1}^m x_i$$

Example: 1d Gaussian model for Midge wing length

- Analogously

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Analogously

$$\log q_\tau^*(\tau) = \mathbb{E}_\mu[\log p(\mathcal{D}_m|\mu,\tau) + \log p(\mu,\tau)] + \log p(\tau) + \text{const}$$

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Analogously

$$\log q_\tau^*(\tau) = \mathbb{E}_\mu[\log p(\mathcal{D}_m|\mu,\tau) + \log p(\mu,\tau)] + \log p(\tau) + \text{const}$$
$$= -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{i=1}^m (x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] +$$

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Analogously

$$\log q_\tau^*(\tau) = \mathbb{E}_\mu[\log p(\mathcal{D}_m|\mu,\tau) + \log p(\mu,\tau)] + \log p(\tau) + \text{const}$$
$$= -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{i=1}^m (x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] +$$
$$+ (a_0-1)\log\tau - b_0\tau + \frac{m}{2}\log\tau + \text{const}$$

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Analogously

$$\log q_\tau^*(\tau) = \mathbb{E}_\mu[\log p(\mathcal{D}_m|\mu,\tau) + \log p(\mu,\tau)] + \log p(\tau) + \text{const}$$
$$= -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{i=1}^m (x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] +$$
$$+ (a_0-1)\log\tau - b_0\tau + \frac{m}{2}\log\tau + \text{const}$$

- Thus the solution is

Example: 1d Gaussian model for Midge wing length

- Analogously

$$\log q_\tau^*(\tau) = \mathbb{E}_\mu[\log p(\mathcal{D}_m|\mu,\tau) + \log p(\mu,\tau)] + \log p(\tau) + \text{const}$$
$$= -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{i=1}^m (x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] +$$
$$+ (a_0-1)\log\tau - b_0\tau + \frac{m}{2}\log\tau + \text{const}$$

- Thus the solution is

$$q_\tau(\tau) = \Gamma(\tau|a_m, b_m),$$

Example: 1d Gaussian model for Midge wing length

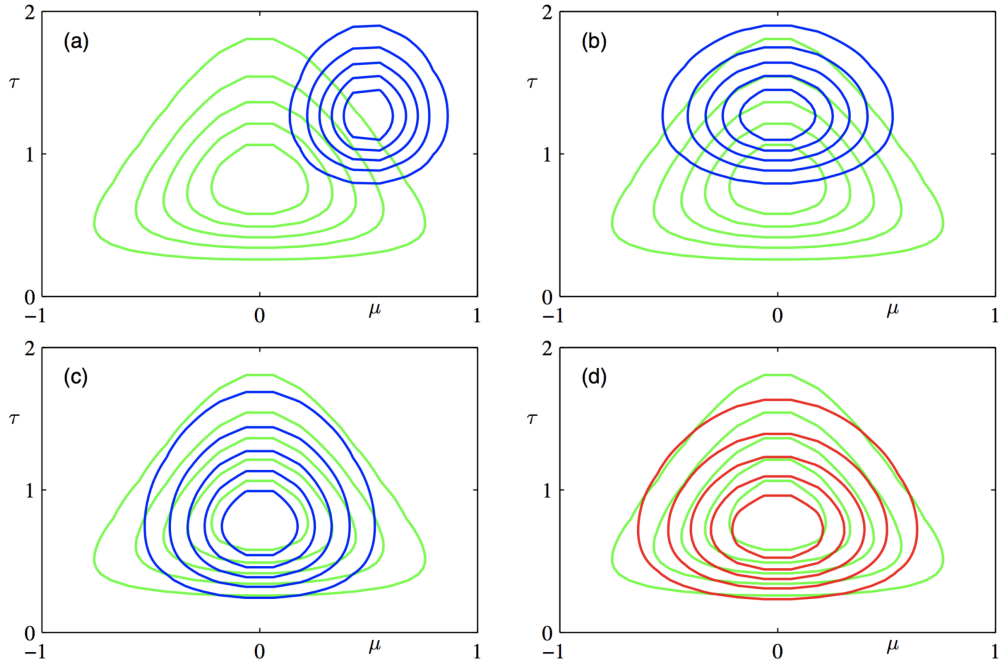
- Analogously

$$\log q_\tau^*(\tau) = \mathbb{E}_\mu[\log p(\mathcal{D}_m|\mu,\tau) + \log p(\mu,\tau)] + \log p(\tau) + \text{const}$$
$$= -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{i=1}^m (x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] +$$
$$+ (a_0-1)\log\tau - b_0\tau + \frac{m}{2}\log\tau + \text{const}$$

- Thus the solution is

$$q_\tau(\tau) = \Gamma(\tau|a_m, b_m),$$
$$a_m = a_0 + \frac{m}{2},\ b_m = b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_{i=1}^m (x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

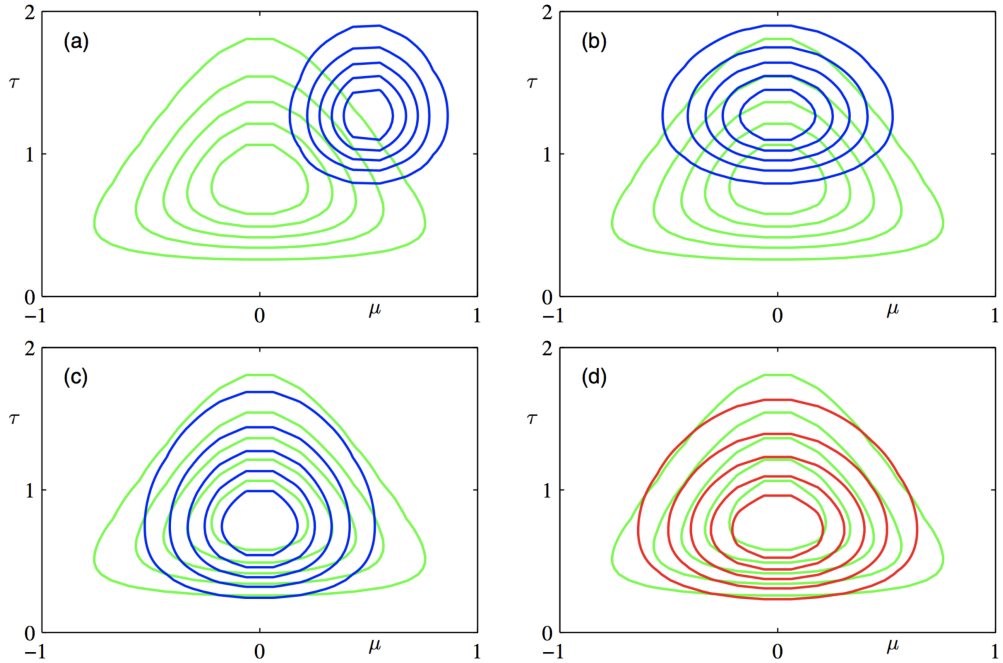
Example: 1d Gaussian model for Midge wing length



Variational inference for μ and τ. Contours of the true posterior $p(\mu, \tau|\mathcal{D}_m)$ are shown in green

- (a) Contours of the initial approximation $q_\mu(\mu)q_\tau(\tau)$ are in blue
- (b) After re-estimating $q_\mu(\mu)$
- (c) After re-estimating $q_\tau(\tau)$
- (d) Contours of the optimal factorized approximation are in red

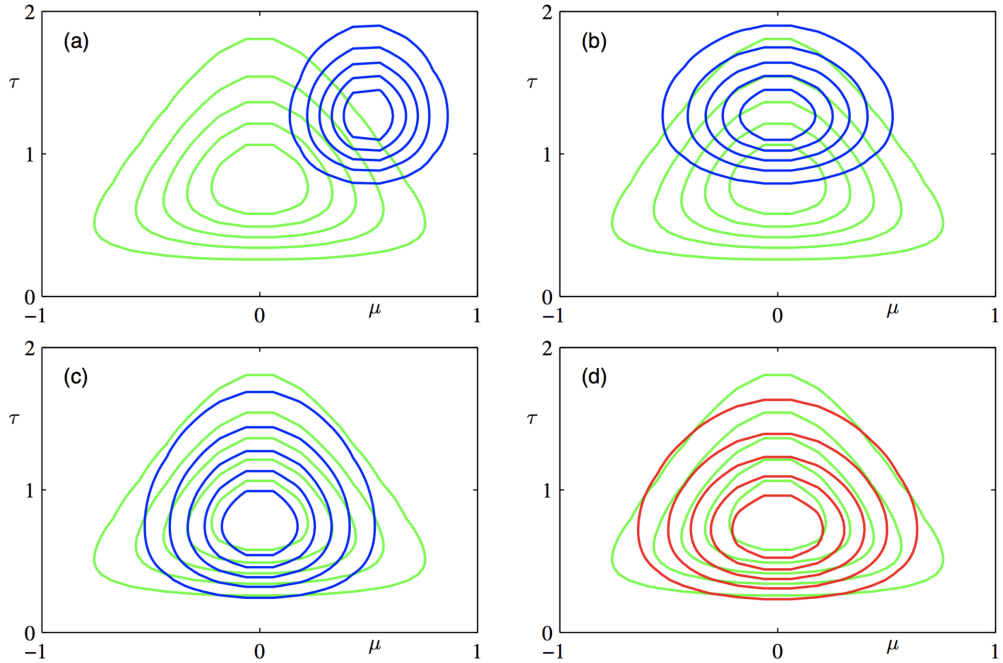
Example: 1d Gaussian model for Midge wing length



Variational inference for μ and τ. Contours of the true posterior $p(\mu, \tau|\mathcal{D}_m)$ are shown in green

- (a) Contours of the initial approximation $q_\mu(\mu)q_\tau(\tau)$ are in blue
- (b) After re-estimating $q_\mu(\mu)$
- (c) After re-estimating $q_\tau(\tau)$
- (d) Contours of the optimal factorized approximation are in red

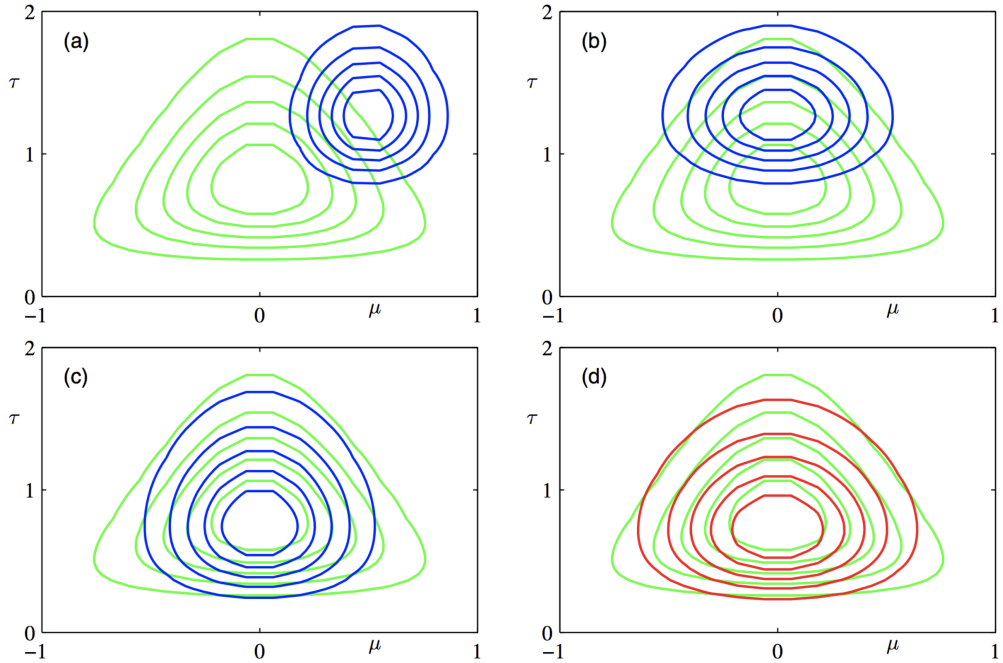
Example: 1d Gaussian model for Midge wing length



Variational inference for μ and τ. Contours of the true posterior $p(\mu, \tau|\mathcal{D}_m)$ are shown in green

- (a) Contours of the initial approximation $q_\mu(\mu)q_\tau(\tau)$ are in blue
- (b) After re-estimating $q_\mu(\mu)$
- (c) After re-estimating $q_\tau(\tau)$
- (d) Contours of the optimal factorized approximation are in red

Example: 1d Gaussian model for Midge wing length



Variational inference for μ and τ. Contours of the true posterior $p(\mu, \tau | \mathcal{D}_m)$ are shown in green

- (a) Contours of the initial approximation $q_\mu(\mu) q_\tau(\tau)$ are in blue
- (b) After re-estimating $q_\mu(\mu)$
- (c) After re-estimating $q_\tau(\tau)$
- (d) Contours of the optimal factorized approximation are in red