CSC696H: Probabilistic Methods in ML

Variational Inference

Prof. Jason Pacheco

Material adapted from: David Blei, NeurIPS 2016 Tutorial

Outline

- Variational Inference
- Stochastic Variational

Outline

- Variational Inference
- Stochastic Variational

Posterior Inference Review

Posterior on latent variable x given data $\mathcal{Y}$ by Bayes' rule:

$$p(x \mid \mathcal{Y}) = \frac{p(x)p(\mathcal{Y} \mid x)}{p(\mathcal{Y})}$$

Marginal likelihood given by,

$$p(\mathcal{Y}) = \int p(x)p(\mathcal{Y} \mid x)dx$$

- Posterior: belief over unknowns, given observed data (knowns)
- Marginal Likelihood: quality of model fit to the observed data

Variational Inference Preview

- Formulate statistical inference as an optimization problem
- Maximize variational lower bound on marginal likelihood

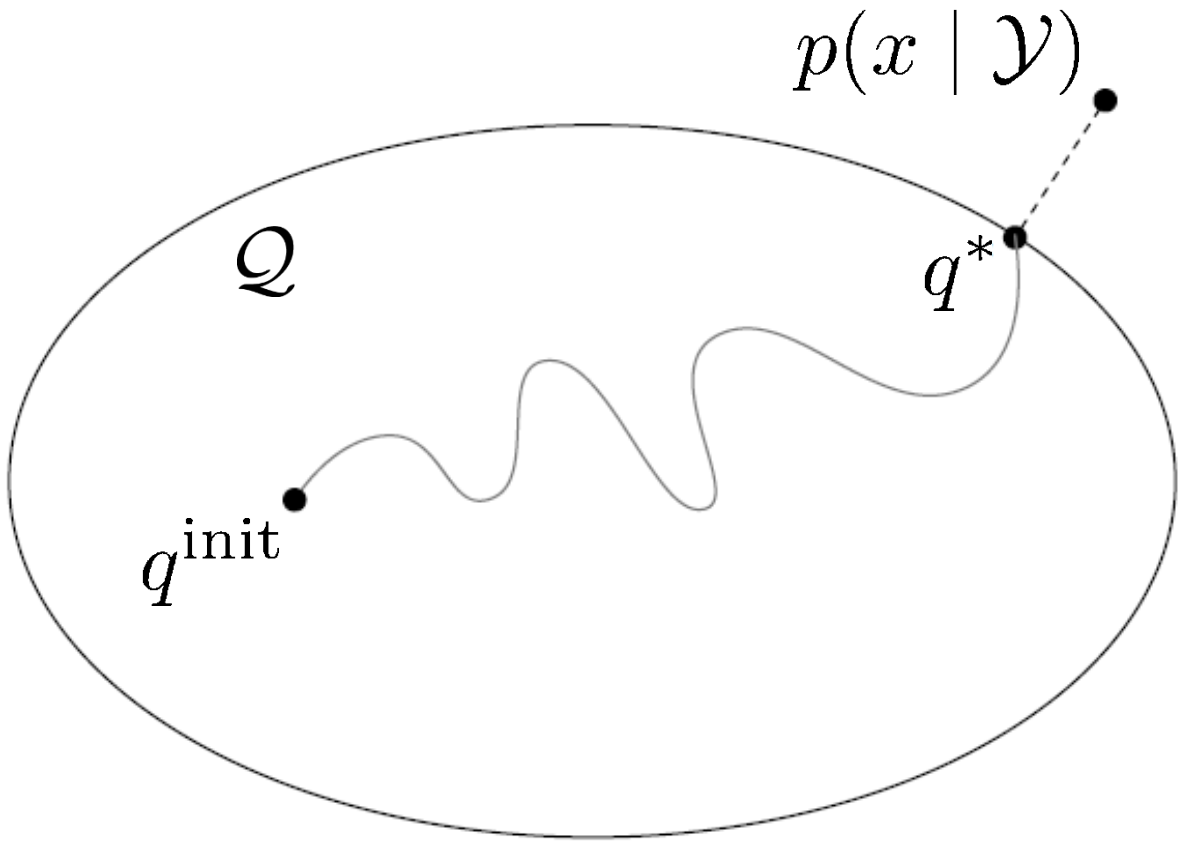
$$\log p(\mathcal{Y}) \geq \max_{q \in \mathcal{Q}} \mathcal{L}(q)$$

- Solution to RHS yields posterior approximation

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q) \approx p(x \mid \mathcal{Y})$$

- Constraint set $\mathcal{Q}$ defines tractable family of approximating distributions
- Very often $\mathcal{Q}$ is an *exponential family*

Variational Inference



[Source: David Blei]

Expectation Maximization (EM) Lower Bound

Recall EM lower bound of marginal likelihood

$$\log p(\mathcal{Y}) = \log \int p(x)p(\mathcal{Y} \mid x)\, dx$$

(Multiply by q(x)/q(x)=1)

$$= \log \int p(x)p(\mathcal{Y} \mid x) \left(\frac{q(x)}{q(x)} \right) dx$$

(Definition of Expected Value)

$$= \log \mathbf{E}_q \left[\frac{p(x)p(\mathcal{Y} \mid x)}{q(x)} \right]$$

(Jensen's Inequality)

$$\geq \mathbf{E}_q \left[\log \frac{p(x)p(\mathcal{Y} \mid x)}{q(x)} \right]$$

A Little Information Theory

- The *entropy* is a natural measure of the inherent uncertainty:

$$H(p) = -\int p(x) \log p(x)\, dx = \mathbb{E}_p[-\log p(x)]$$

- **Interpretation** Difficulty of compression of some random variable

- The *relative entropy* or *Kullback-Leibler (KL) divergence* is a non-negative, but asymmetric, "distance" between a given pair of probability distributions:

$$KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx \qquad KL(p\|q) \geq 0 \qquad KL(p\|q) \neq KL(q\|p)$$

- The KL divergence equals zero if and only if $p(x) = q(x)$ for all *x*.
- **Interpretation** The cost of compressing data from distribution *p(x)* with a code optimized for distribution *q(x)*

EM Lower Bound

$$\mathbf{E}_q\left[\log \frac{p(x)p(\mathcal{Y} \mid x)}{q(x)}\right] = \mathbf{E}_q\left[\log \frac{p(x)p(\mathcal{Y} \mid x)}{q(x)} \frac{p(\mathcal{Y})}{p(\mathcal{Y})}\right]$$

(Multiply by 1)

$$= \log p(\mathcal{Y}) - \mathrm{KL}(q(x) \| p(x \mid \mathcal{Y}))$$

(Definition of KL)

Bound gap is the Kullback-Leibler divergence KL(q||p),

$$\mathrm{KL}(q(x) \| p(x \mid \mathcal{Y})) = \int q(x) \log \frac{q(x)}{p(x \mid \mathcal{Y})}$$

Solution to **E-step** is,

$$q^* = \arg\min_q \mathrm{KL}(q(x) \| p(x \mid \mathcal{Y})) = p(x \mid \mathcal{Y})$$

This doesn't help us if $p(x \mid \mathcal{Y})$ is intractable

Variational Lower Bound

Idea Restrict optimization to a set $\mathcal{Q}$ of analytic distributions

$$\log p(\mathcal{Y}) \geq \max_{q \in \mathcal{Q}} \mathcal{L}(q) \equiv \mathbf{E}_q \left[\log \frac{p(x)p(\mathcal{Y} \mid x)}{q(x)} \right]$$

- If posterior is in set $p(x \mid \mathcal{Y}) \in \mathcal{Q}$ then exact inference $q(x) = p(x \mid \mathcal{Y})$
- Otherwise, if $p(x \mid \mathcal{Y}) \notin \mathcal{Q}$ posterior is closest approximation in KL

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\big(q(x) \| p(x \mid \mathcal{Y})\big)$$

… and we recover strict lower bound on marginal likelihood with gap

$$\log p(\mathcal{Y}) - \mathcal{L}(q^*) = \mathrm{KL}\big(q^*(x) \| p(x \mid \mathcal{Y})\big)$$

Variational Lower Bound

Two competing terms in variational bound…

$$\begin{aligned}\mathcal{L}(q) &\equiv \mathbb{E}_q\left[\log \frac{p(x)p(\mathcal{Y} \mid x)}{q(x)}\right] \\ &= \mathbb{E}_q[\log p(x,\mathcal{Y})] - \mathbb{E}_q[\log q(x)] \\ &= \mathbb{E}_q[\log p(x,\mathcal{Y})] + H(q)\end{aligned}$$

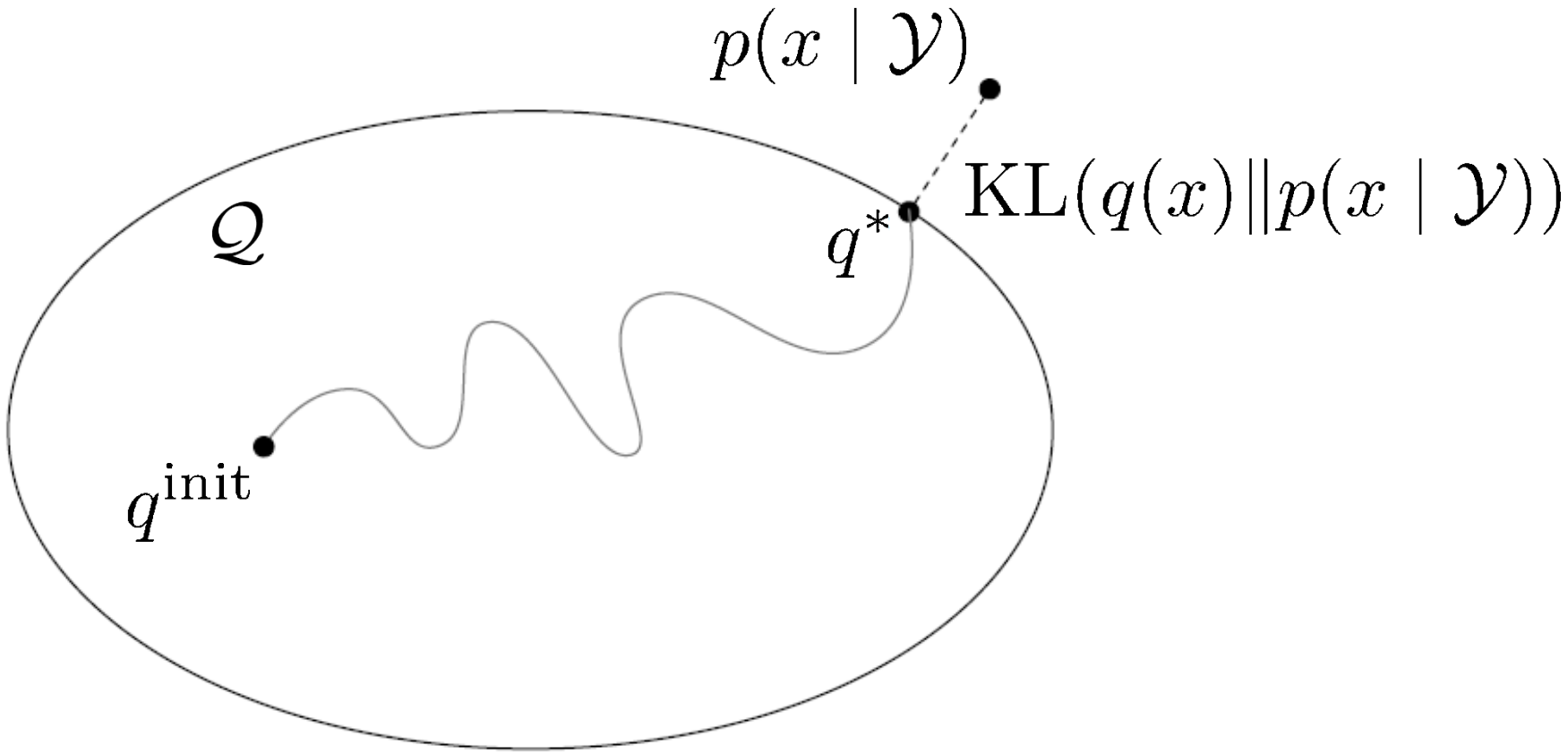
Average (negative) Energy

Encourages q(x) to “agree” with model p(x,y)

Entropy

Encourages q(x) to have large uncertainty (good for generalization)

Variational Approximation



Minimize KL between $q(x)$ and posterior $p(x \mid \mathcal{Y})$.

[Source: David Blei]

Relation to EM

- EM is means for approximate *learning,* but we are using it to motivate approximate *inference*

- EM lower bound takes same form as VI lower bound, but with different constraint sets

- Connection with variational inference (VI) is in E-step, which performs inference with fixed parameters

Variational Inference

$$\log p(\mathcal{Y}) \geq \max_{q \in \mathcal{Q}} \mathcal{L}(q) \equiv \mathbb{E}_q[\log p(x, \mathcal{Y})] + H(q)$$

Different sets $\mathcal{Q}$ yield different VI algorithms to optimize bound:

- **Mean Field** Ignore posterior dependencies among variables
- **Loopy BP** *Locally consistent* marginals (exact for tree-structured models)
- **Expectation Propagation (EP)** *Locally consistent moments* (equivalent to Loopy BP for tree-structure exponential families)

Why is it called “variational”?

Differential Calculus

- Typically, we optimize a function $\max_x f(x)$ w.r.t. a **variable** X
- Use standard derivatives/gradients $\nabla_x f(x)$
- Extrema given by zero-gradient conditions $\|\nabla_x f(x)\| = 0$

Calculus of Variations

- Optimize a *functional* (function of a function): $\max_{q(x)} f(q(x))$
- *Functional derivative* characterizes change w.r.t. function q(x)
- Extrema given by Euler-Lagrange equation; analogous to zero-gradient condition

In practice, we typically parameterize $q_\mu(x)$ and take standard gradients w.r.t. parameters μ

Summary: Variational Inference

1) Begin with intractable model posterior:

$$p(x \mid \mathcal{Y}) = \frac{p(x)p(\mathcal{Y} \mid x)}{p(\mathcal{Y})}$$

Marginal Likelihood

2) Choose a family of approximating distributions $\mathcal{Q}$ that is tractable

3) Maximize variational lower bound on marginal likelihood:

$$\log p(\mathcal{Y}) \geq \max_{q \in \mathcal{Q}} \mathcal{L}(q) \equiv \mathbb{E}_q[\log p(x, \mathcal{Y})] + H(q)$$

4) Maximizer is posterior approximation (in KL divergence)

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(x) \| p(x \mid \mathcal{Y}))$$

Still need to show…

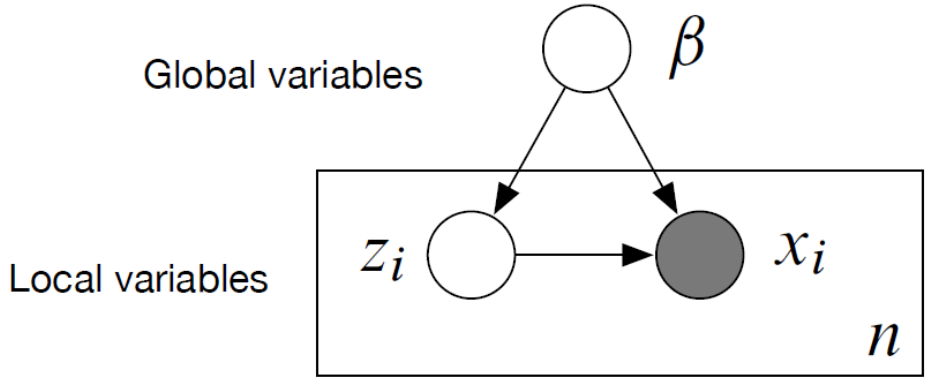
a) How to define approximating variational family $\mathcal{Q}$

b) How to optimize lower bound

Outline

- Variational Inference
- Stochastic Variational

A Generic Class of Directed Models

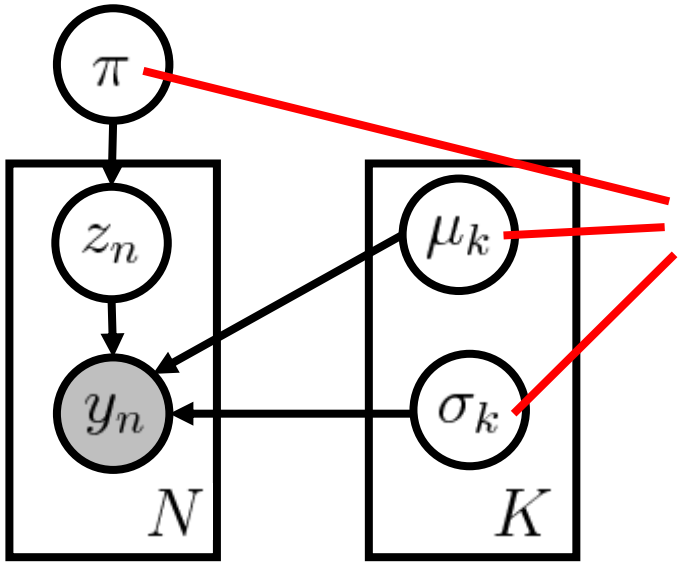


$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- Bayesian mixture models
- Time series & sequence models (HMMs, Linear dynamical systems)
- Matrix factorization (factor analysis, PCA, CCA)
- Multilevel regression (linear, probit, Poisson)
- Stochastic block models
- Mixed-membership models (Linear discriminant analysis)

[Source: David Blei]

Example: Gaussian Mixture Model

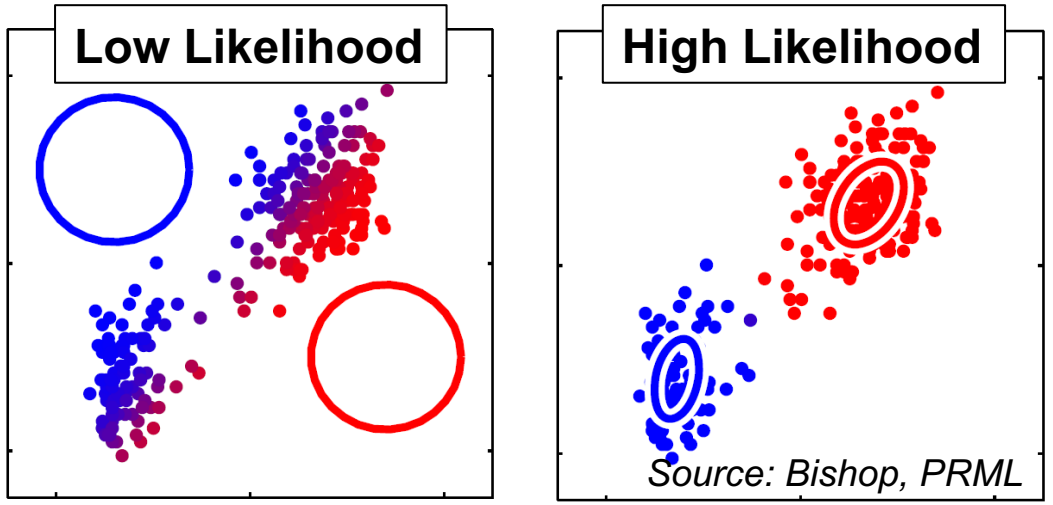


Global variables:

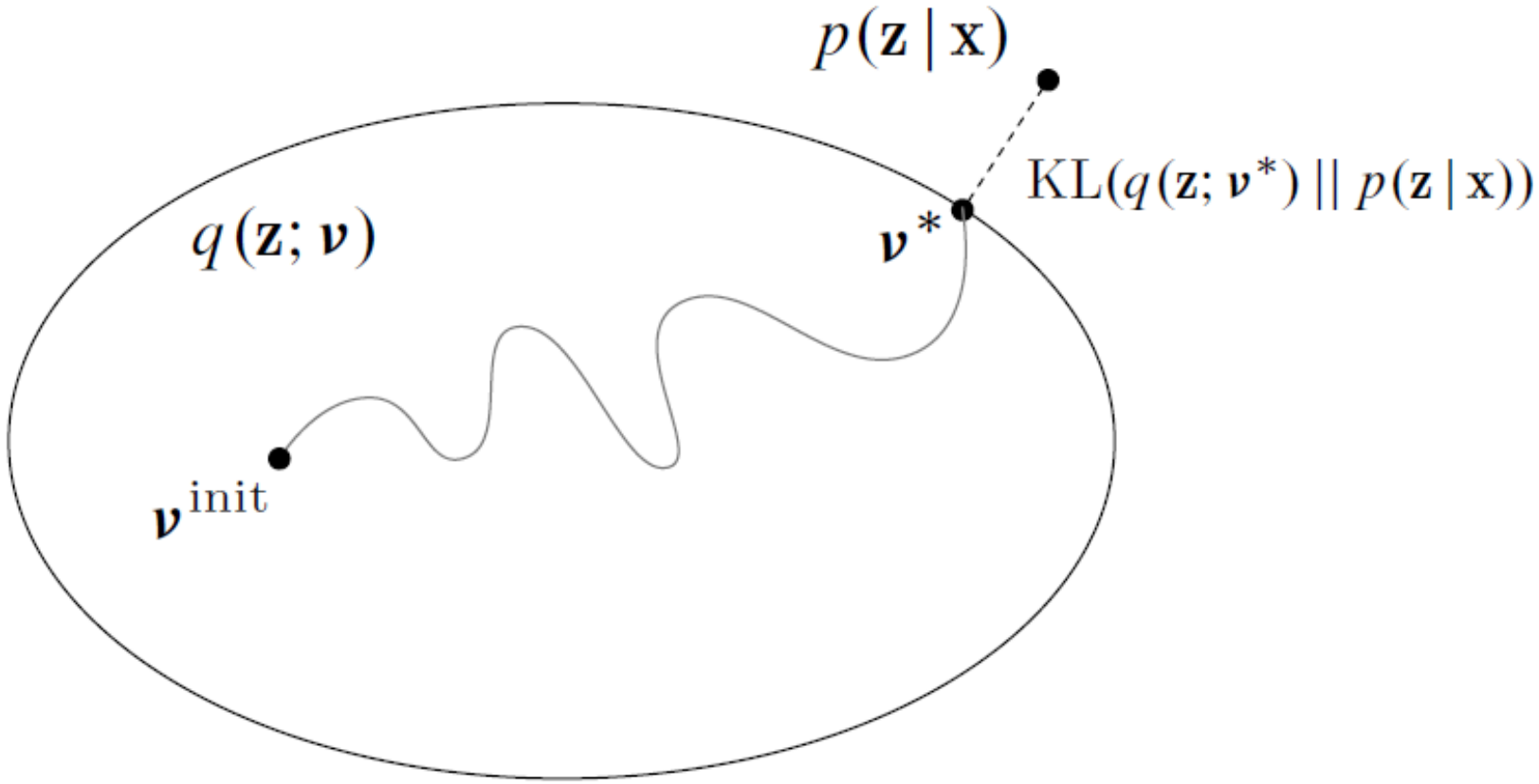
$$\beta = \{\pi, \mu_1, \sigma_1, \ldots, \mu_K, \sigma_K\} \qquad \mathcal{Y} = \{y_1, \ldots, y_N\}$$

Local variables Z control component assignments

GMM



Variational Approximation



Minimize KL between $q(\beta, \mathbf{z}; \nu)$ and posterior $p(\beta, \mathbf{z} \mid \mathbf{x})$.

[Source: David Blei]

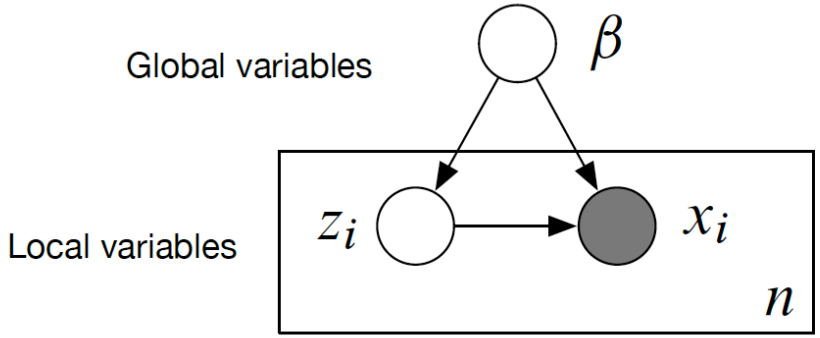
Variational Lower Bound – ELBO

$$\mathcal{L}(\nu) = \mathbb{E}_{q_\nu}[\log p(\beta, \mathbf{z}, \mathbf{x})] - \mathbb{E}_{q_\nu}[\log q(\beta, \mathbf{z}; \nu)]$$

- KL is intractable; VI optimizes **evidence lower bound** (ELBO)
  - Lower bounds log p(x) – marginal likelihood, or *evidence*
  - Maximizing ELBO is equivalent to minimizing KL w.r.t. posterior
- The ELBO trades off two terms
  - The first term prefers q(.) to place mass on the MAP estimate
  - Second term encourages q(.) to be *diffuse* (maximize entropy)
- The ELBO is non-convex

[Source: David Blei]

Conditionally conjugate models



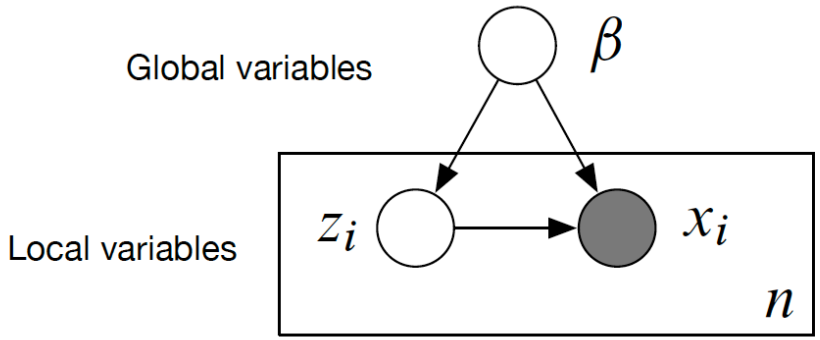
$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- The observations are $\mathbf{x} = x_{1:n}$.
- The **local** variables are $\mathbf{z} = z_{1:n}$.
- The **global** variables are β.
- The ith data point x_i only depends on z_i and β.

Compute $p(\beta, \mathbf{z} \mid \mathbf{x})$.

[Source: David Blei]

Conditionally conjugate models



$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- A **complete conditional** is the conditional of a latent variable given the observations and other latent variables.
- Assume each complete conditional is in the exponential family,

$$p(z_i \mid \beta, x_i) = \text{expfam}(z_i\,;\, \eta_\ell(\beta, x_i))$$
$$p(\beta \mid \mathbf{z}, \mathbf{x}) = \text{expfam}(\beta\,;\, \eta_g(\mathbf{z}, \mathbf{x})),$$

where $\text{expfam}(z\,;\,\eta) = h(z)\exp\{\eta^\top z - a(\eta)\}$.

[Source: David Blei]

Aside: The exponential family

$$p(x) = h(x) \exp\{\eta^\top t(x) - a(\eta)\}$$

Terminology:

η	the natural parameter
$t(x)$	the sufficient statistics
$a(\eta)$	the log normalizer
$h(x)$	the base density

[Source: David Blei]

Aside: The exponential family

$$p(x) = h(x)\exp\{\eta^\top t(x) - a(\eta)\}$$

- The log normalizer is

$$a(\eta) = \log \int \exp\{\eta^\top t(x)\} dx$$

- It ensures the density integrates to one.
- Its gradient calculates the expected sufficient statistics

$$\mathbb{E}[t(X)] = \nabla_\eta a(\eta).$$

[Source: David Blei]

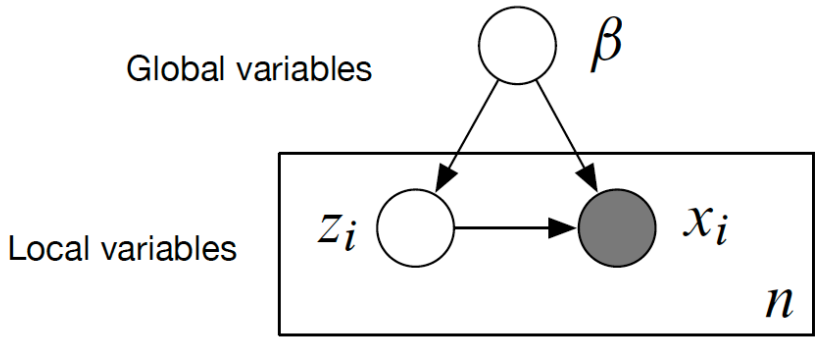
Aside: The exponential family

$$p(x) = h(x) \exp\{\eta^\top t(x) - a(\eta)\}$$

- Many common distributions are in the exponential family
  - Bernoulli, categorical, Gaussian, Poisson, Beta, Dirichlet, Gamma, etc.
- Outlines the theory around conjugate priors and corresponding posteriors
- Connects closely to variational inference [Wainwright and Jordan, 2008]

[Source: David Blei]

Conditionally conjugate models



$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

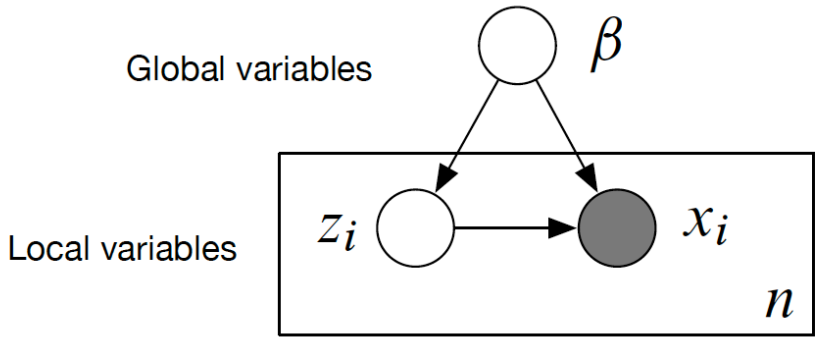
- Each **complete conditional** is in the exponential family.
- The global parameter comes from conjugacy [Bernardo and Smith, 1994]

$$\eta_g(\mathbf{z}, \mathbf{x}) = \alpha + \sum_{i=1}^{n} t(z_i, x_i),$$

where α is a hyperparameter and $t(\cdot)$ are sufficient statistics for $[z_i, x_i]$.

[Source: David Blei]

Conditionally conjugate models

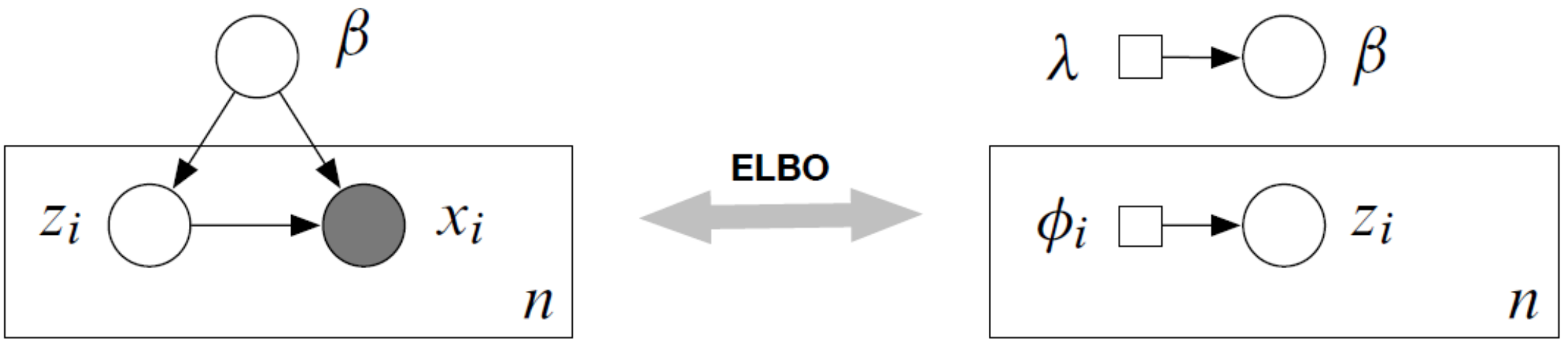


$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- Bayesian mixture models
- Time series models (HMMs, linear dynamic systems)
- Factorial models
- Matrix factorization (factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression (linear, probit, Poisson)
- Stochastic block models
- Mixed-membership models (LDA and some variants)

[Source: David Blei]

Mean Field for Generic Directed Model



PGM of Mean Field Approximation

Recall: mean field family is *fully factorized*

$$q(\beta, \mathbf{z}; \lambda, \phi) = q(\beta; \lambda) \prod_{i=1}^{n} q(z_i; \phi_i)$$

Variational Parameters

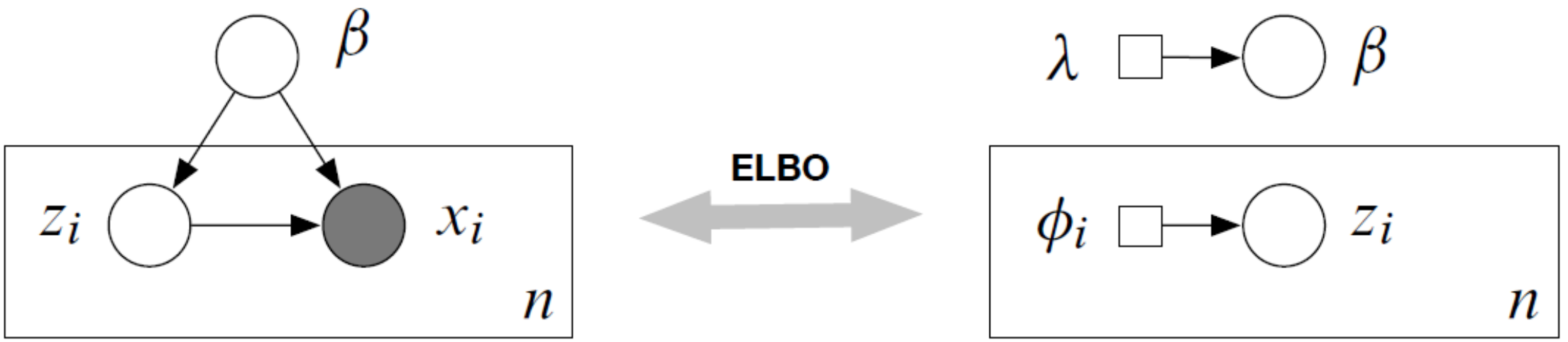
Conditional conjugacy: Each factor is the same expfam as complete conditional

$$p(\beta \mid \mathbf{z}, \mathbf{x}) = h(\beta) \exp\{\eta_g(\mathbf{z}, \mathbf{x})^\top \beta - a(\eta_g(\mathbf{z}, \mathbf{x}))\}$$

$$q(\beta; \lambda) = h(\beta) \exp\{\lambda^\top \beta - a(\lambda)\}.$$

[Source: David Blei]

Mean Field for Generic Directed Model



PGM of Mean Field Approximation

Recall: mean field family is *fully factorized*

$$q(\beta, \mathbf{z}; \lambda, \phi) = q(\beta; \lambda) \prod_{i=1}^{n} q(z_i; \phi_i)$$

Variational Parameters

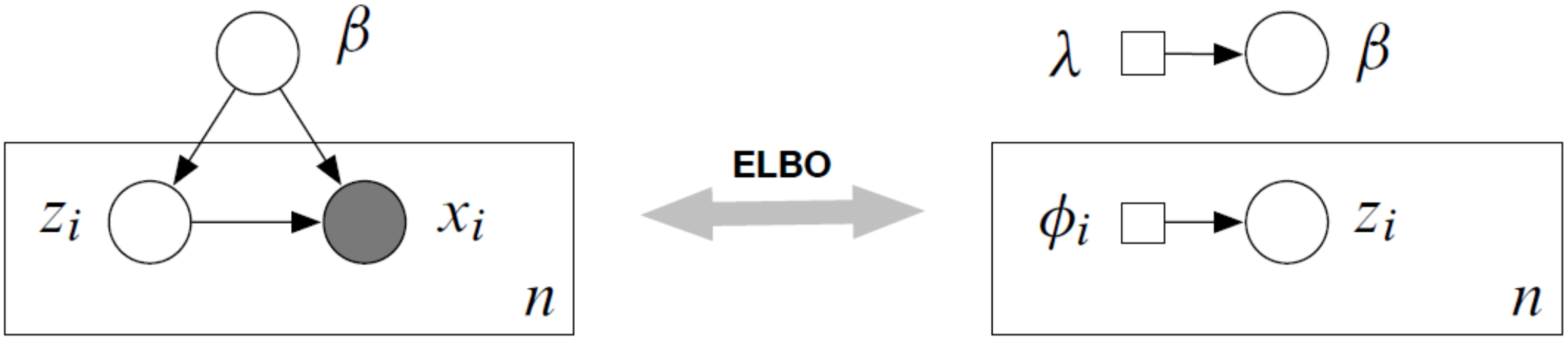
Global parameter ensure conjugacy to (z,x):

$$\eta_g(\mathbf{z}, \mathbf{x}) = \alpha + \sum_{i=1}^{n} t(z_i, x_i),$$

where α is prior hyperparameter and t(.) are sufficient statistics for $[z_i, x_i]$

[Source: David Blei]

Mean Field for Generic Directed Model



PGM of Mean Field Approximation

Optimize ELBO,

$$\mathcal{L}(\lambda, \phi) = \mathbb{E}_q[\log p(\beta, \mathbf{z}, \mathbf{x})] - \mathbb{E}_q[\log q(\beta, \mathbf{z})]$$

Don't forget… entropy decomposes as sum over individual entropies

Traditional VI uses coordinate ascent,

$$\lambda^* = \mathbb{E}_\phi\left[\eta_g(\mathbf{z}, \mathbf{x})\right];\ \phi_i^* = \mathbb{E}_\lambda\left[\eta_\ell(\beta, x_i)\right]$$

Iteratively update each parameter, holding others fixed

- Obvious relationship with Gibbs sampling
- Remember, ELBO is not convex

[Source: David Blei]

Coordinate Ascent Mean Field for Generic Model

Input: data $\mathbf{x}$, model $p(\beta, \mathbf{z}, \mathbf{x})$.

Initialize λ randomly.

repeat

  for *each data point* i **do**

    Set local parameter $\phi_i \leftarrow \mathbb{E}_\lambda[\eta_\ell(\beta, x_i)]$.

  end

  Set global parameter

$$\lambda \leftarrow \alpha + \sum_{i=1}^{n} \mathbb{E}_{\phi_i}[t(Z_i, x_i)].$$

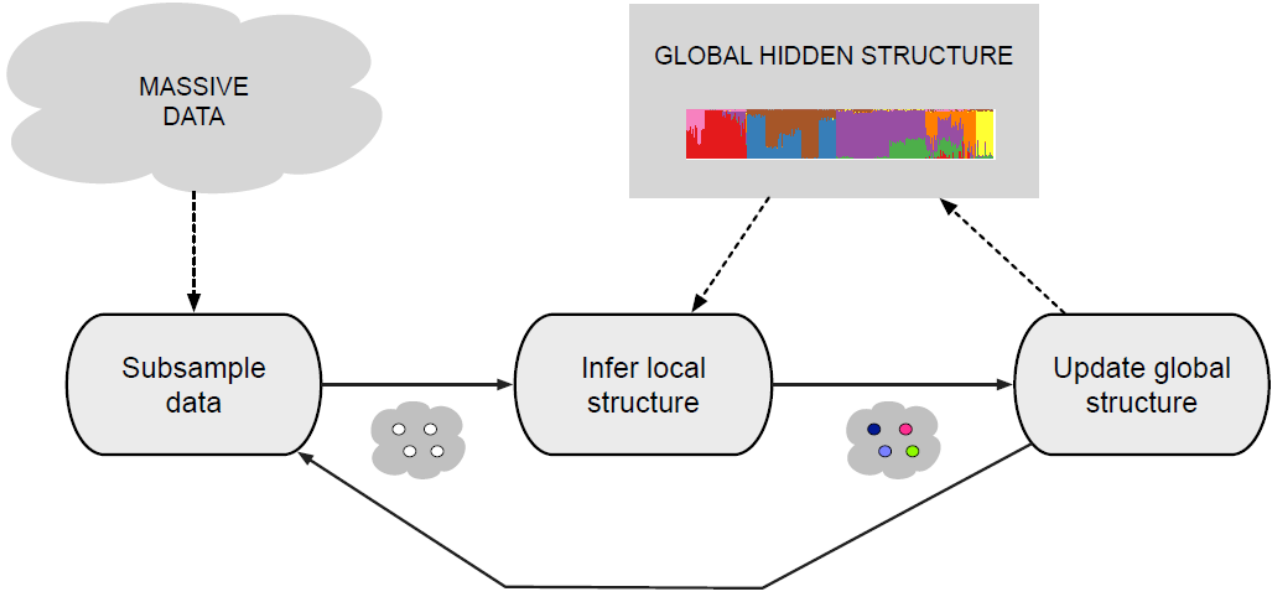
until *the ELBO has converged*

Need to visit *every* data point

Need to sum *every* data point

[Source: David Blei]

Stochastic (Mean Field) Variational Inference



Classical mean field VI is inefficient for large data

- Do some local computation *for each data point*
- Aggregate computations to re-estimate global structure
- Repeat

Idea visit random subsets of data to estimate gradient updates on full dataset

[Source: David Blei]

Stochastic Gradient Ascent/Descent

A STOCHASTIC APPROXIMATION METHOD[1]

By Herbert Robbins and Sutton Monro

University of North Carolina

1. Summary. Let $M(x)$ denote the expected value at level x of the response to a certain experiment. $M(x)$ is assumed to be a monotone function of x but is unknown to the experimenter, and it is desired to find the solution $x = \theta$ of the equation $M(x) = \alpha$, where α is a given constant. We give a method for making successive experiments at levels $x_1, x_2, \cdots$ in such a way that x_n will tend to θ in probability.

- Use cheaper noisy gradient estimates [Robbins and Monro, 1951]
- Guaranteed to converge to local optimum [Bottou, 1996]
- Popular in modern machine learning (e.g. learning deep neural nets)

[Source: David Blei]

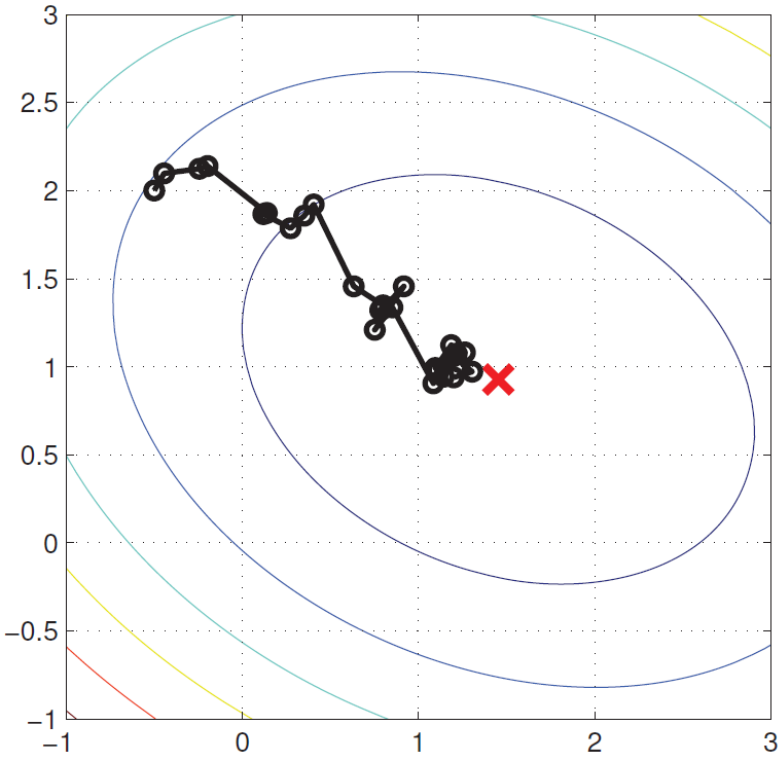
Stochastic Gradient Ascent/Descent

- Stochastic gradients update:

$$\nu_{t+1} = \nu_t + \rho_t \hat{\nabla}_\nu \mathcal{L}(\nu_t)$$

- Gradient estimator must be *unbiased*

$$\mathbb{E}[\hat{\nabla}_\nu \mathcal{L}(\nu)] = \nabla_\nu \mathcal{L}(\nu)$$



- Sequence of step sizes ρ_t must follow Robbins-Monro conditions

$$\sum_{t=0}^{\infty} \rho_t = \infty, \qquad \sum_{t=0}^{\infty} \rho_t^2 < \infty$$

[Source: David Blei]

Stochastic Variational Inference

- The **natural gradient** of the ELBO [Amari, 1998; Sato, 2001]

$$\nabla_\lambda^{\text{nat}} \mathcal{L}(\lambda) = \left(\alpha + \sum_{i=1}^n \mathbb{E}_{\phi_i^*}[t(Z_i, x_i)]\right) - \lambda.$$

- Construct a **noisy natural gradient**,

$$j \sim \text{Uniform}(1, \ldots, n)$$
$$\hat{\nabla}_\lambda^{\text{nat}} \mathcal{L}(\lambda) = \alpha + n\mathbb{E}_{\phi_j^*}[t(Z_j, x_j)] - \lambda.$$

- This is a good noisy gradient.
  - Its expectation is the exact gradient (*unbiased*).
  - It only depends on optimized parameters of one data point (*cheap*).

[Source: David Blei]

Stochastic Variational Inference

Input: data $\mathbf{x}$, model $p(\beta, \mathbf{z}, \mathbf{x})$.

Initialize λ randomly. Set ρ_t appropriately.

repeat

Sample $j \sim \text{Unif}(1, \ldots, n)$.

Set local parameter $\phi \leftarrow \mathbb{E}_\lambda\left[\eta_\ell(\beta, x_j)\right]$.

Set intermediate global parameter

$$\hat{\lambda} = \alpha + n\mathbb{E}_\phi[t(Z_j, x_j)].$$

Set global parameter

$$\lambda = (1 - \rho_t)\lambda + \rho_t\hat{\lambda}.$$

until *forever*

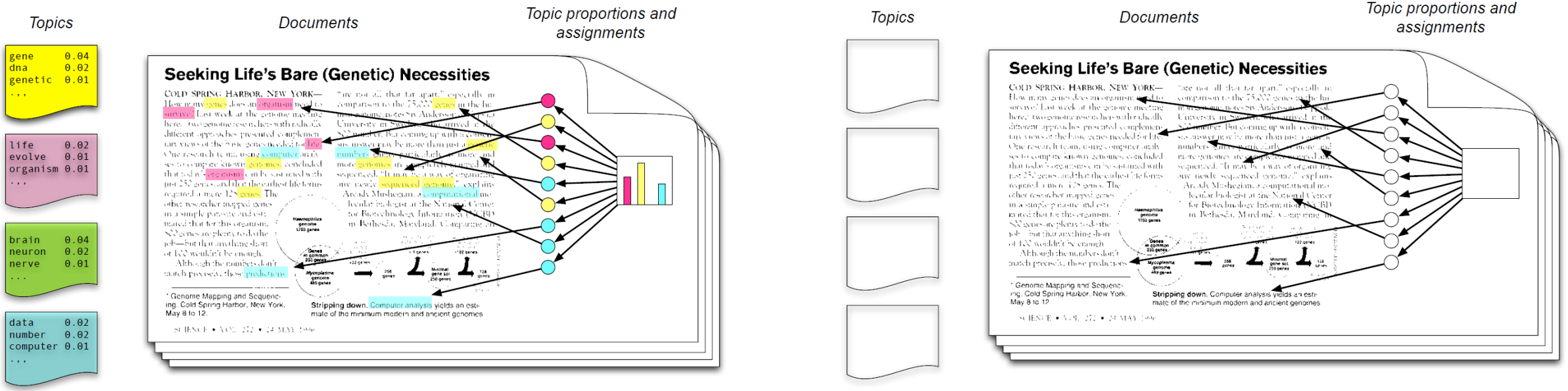
[Source: David Blei]

Topic Models

Topic models discover hidden thematic structure in large collections of documents

[Source: David Blei]

Topic Models



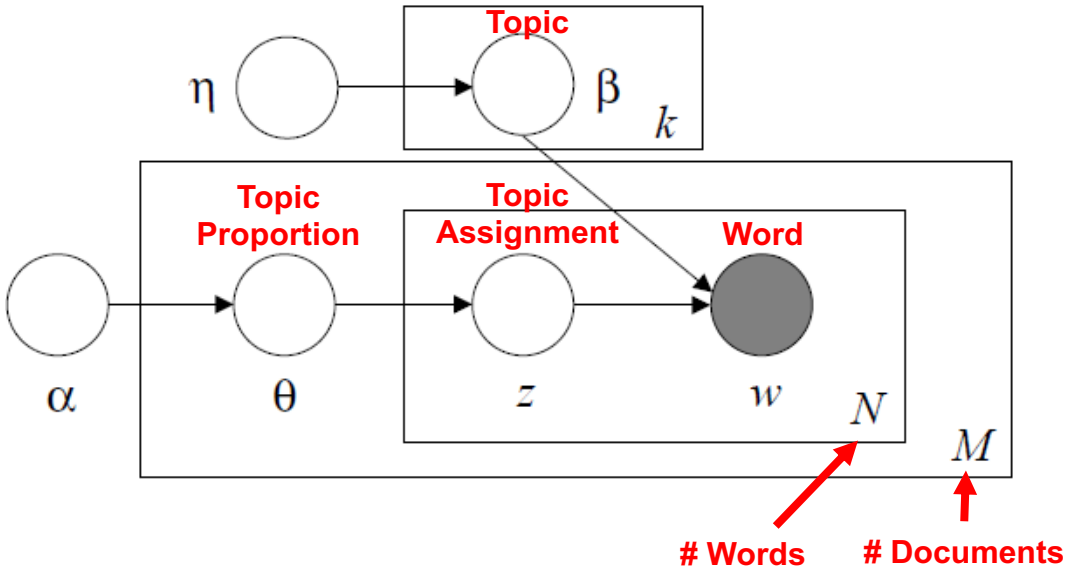
- Each *topic* is a distribution over words (vocabulary)
- Each *document* is a mixture of corpus-wide topics
- Each *word* is drawn from one of the topics (they are distributions)
- But we only observe documents; everything else is hidden (unsupervised learning problem)
- Need to calculate posterior (for millions of documents; billions of latent variables):

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

[Source: David Blei]

Topic Models

Latent Dirichlet Allocation (LDA)



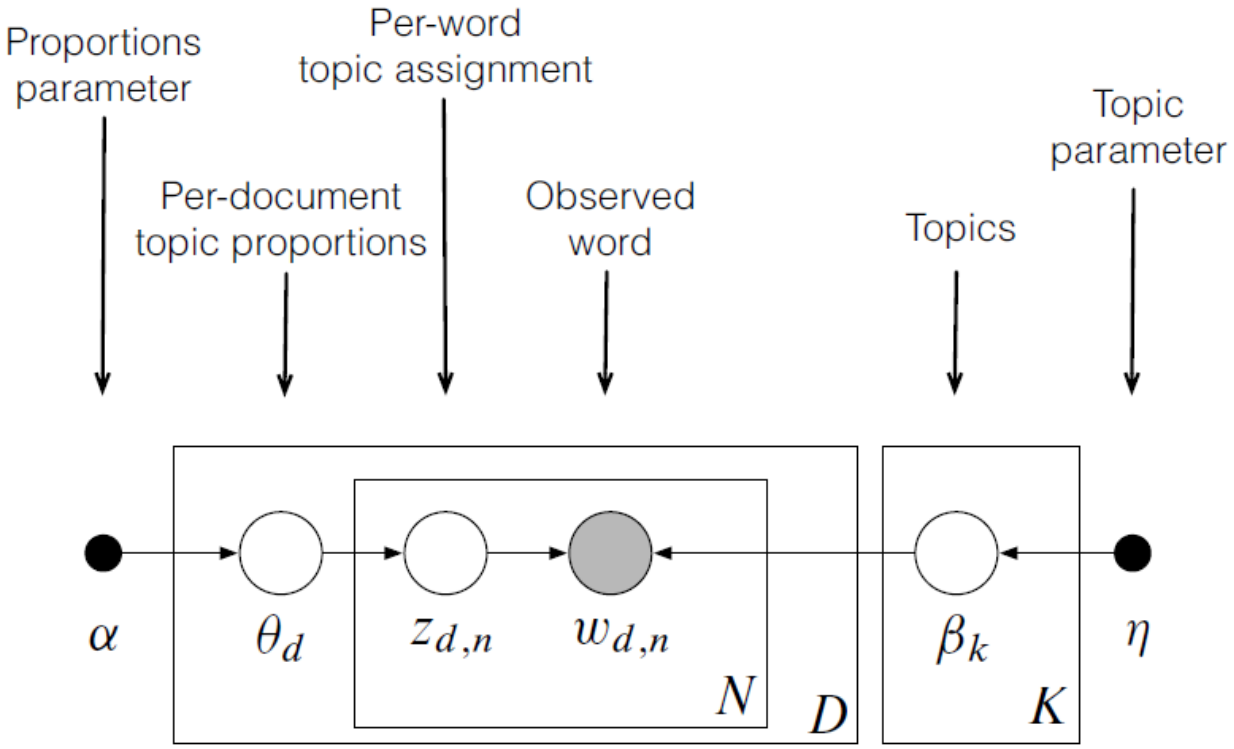
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Allows *unsupervised learning* of document corpus via mixture modeling

Blei, D. et al., JMLR 2003

Example: Latent Dirichlet Allocation



Latent Dirichlet Allocation (LDA):

$$\beta_k \sim \text{Dirichlet}(\eta)$$
$$\theta_d \sim \text{Dirichlet}(\alpha)$$
$$z_{d,n} \mid \theta_d \sim \text{Cat}(\theta_d)$$
$$w_{d,n} \mid z_{d,n}, \beta \sim \text{Cat}(\beta_{z_{d,n}})$$

- Assumes words are *exchangeable* ("bag-of-words" model)
- Reduces parameters while still yielding useful insights
- Complete conditionals are closed-form (we can do mean field)

[Source: David Blei]

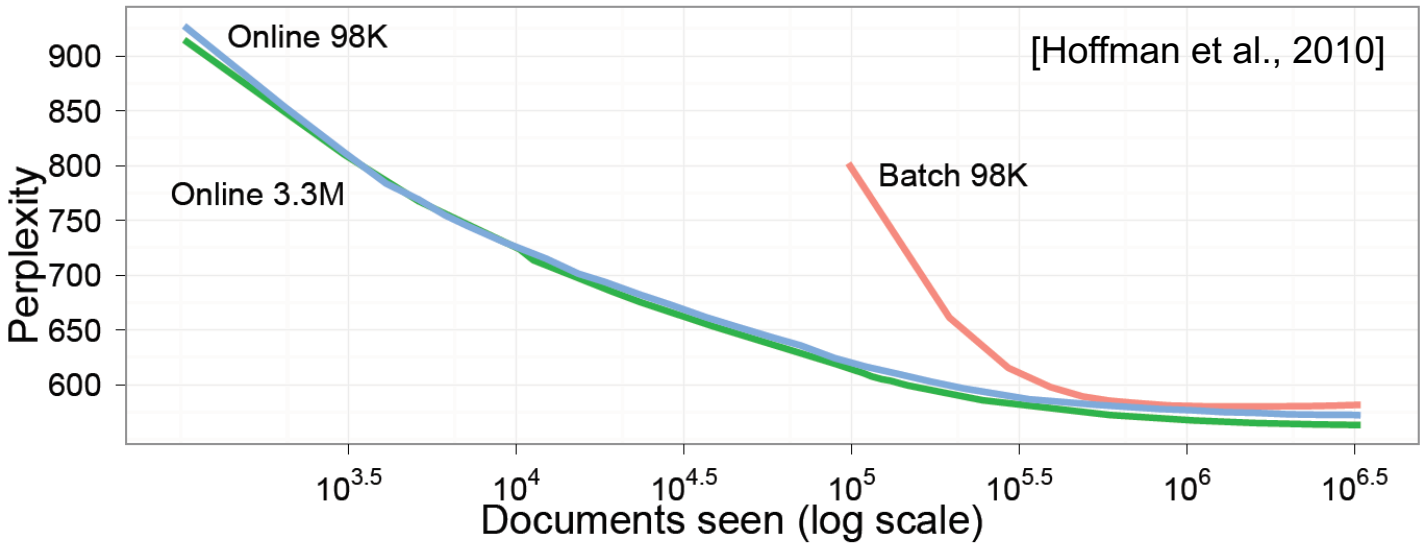
Example: Latent Dirichlet Allocation

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



- Stochastic VI (online) shows faster learning as compared to standard (batch) updates
- Similar learning rate when dataset increased from 98K to 3.3M documents
- Perplexity measures posterior uncertainty (lower is better)

$$\text{Perplexity} = 2^{H(p)} = 2^{-\sum_x p(x) \log p(x)}$$

[Source: David Blei]

Summary: Variational Inference

1) Begin with intractable model posterior:

$$p(x \mid \mathcal{Y}) = \frac{p(x)p(\mathcal{Y} \mid x)}{p(\mathcal{Y})}$$

Marginal Likelihood

2) Choose a family of approximating distributions $\mathcal{Q}$ that is tractable

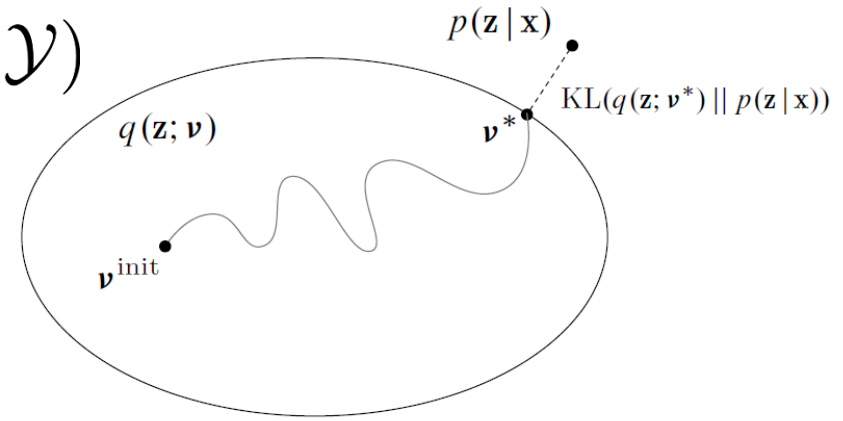
3) Maximize variational lower bound on marginal likelihood:

$$\log p(\mathcal{Y}) \geq \max_{q \in \mathcal{Q}} \mathcal{L}(q) \equiv \mathbb{E}_q[\log p(x, \mathcal{Y})] + H(q)$$

4) Maximizer is posterior approximation (in KL divergence)

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(x) \| p(x \mid \mathcal{Y}))$$

Different approximating families $\mathcal{Q}$ lead to different forms of optimizing variational bound



Summary: Mean Field VI

- Mean field family assumes fully factorized approximating distribution

$$q(x) = \prod_{s \in \mathcal{V}} q_s(x_s)$$

- Mean field algorithm performs coordinate ascent on lower bound

$$q_s(x_s) \propto \exp\left\{\mathbb{E}_{q_{\setminus s}}[\log p(x, \mathcal{Y})]\right\}$$

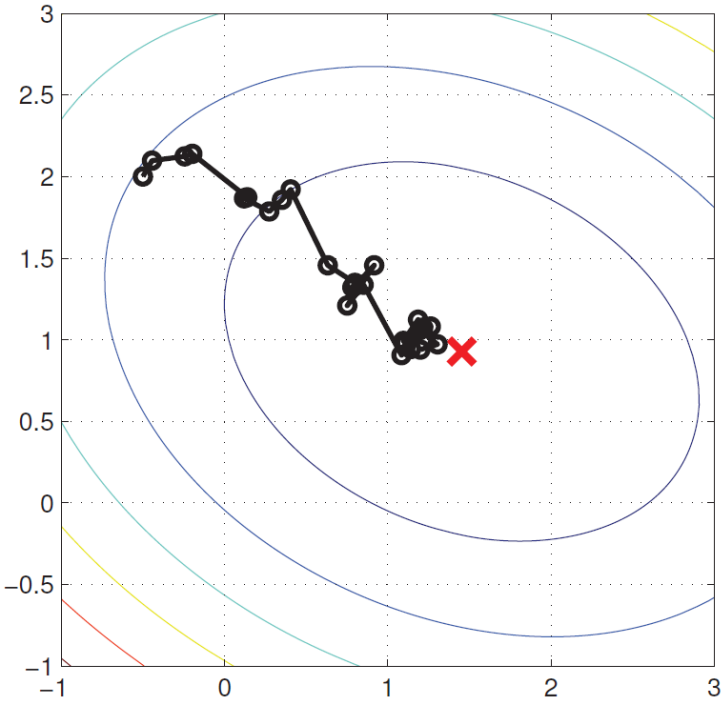
- Coordinate ascent updates require complete conditionals to be conjugate
  - Similar, but stricter, assumption to Gibbs sampling
- MF update takes specific form depending on model $p(.)$, e.g. pairwise MRF:

$$\mu_{sk}^{(i)} \propto \psi_s(k) \exp\left\{\sum_{t \in \Gamma(s)} \mathbb{E}_{\mu_t^{(i-1)}}[\phi_{st}(k, x_t)]\right\}$$

Summary: Stochastic (Mean Field) VI

- MF coordinate ascent updates require visiting *all data*
  - Doesn't scale to large datasets
- Stochastic VI updates using stochastic gradient ascent
  - Randomly subsample dataset
  - Compute stochastic estimate of full gradient based on subsample
  - Stochastic gradient step on variational parameters (ν here):

$$\nu_{t+1} = \nu_t + \rho_t \hat{\nabla}_\nu \mathcal{L}(\nu_t)$$



- Step sizes must decrease over time while satisfying Robbins-Monro conditions

$$\sum_{t=0}^{\infty} \rho_t = \infty, \qquad \sum_{t=0}^{\infty} \rho_t^2 < \infty$$

- Often call standard MF "batch" since updates based on full data