Computer Science

CSC535: Probabilistic Graphical Models

Variational Inference

Prof. Jason Pacheco

Material adapted from: David Blei, NeurIPS 2016 Tutorial

Outline

- Variational Inference
- Mean Field Variational
- Stochastic Variational

Outline

- Variational Inference
- Mean Field Variational
- 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

Posterior Inference Review

- ➢ Tree-structured discrete / Gaussian models can use **sum-product BP**
- ➢ Posterior & marginal likelihood intractable in many practical cases

Monte Carlo methods and MCMC

- **PROs** Asymptotic guarantees, easy to implement for most models, more computation = higher accuracy
- **CON**s Difficult to diagnose convergence, few non-asymptotic guarantees, slow

Loopy (sum-product) BP

- **PROs** Often yields good solutions quickly, easy to diagnose convergence
- **CONs** No computation/accuracy tradeoff, restricted to discrete/Gaussian models

Loopy BP is an instance of a wider class of *variational methods*

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*

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:

$$H(p) = -\int p(x) \log p(x)\, dx$$

- **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:

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

- 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…

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

Average (negative) Energy

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

Entropy

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

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…

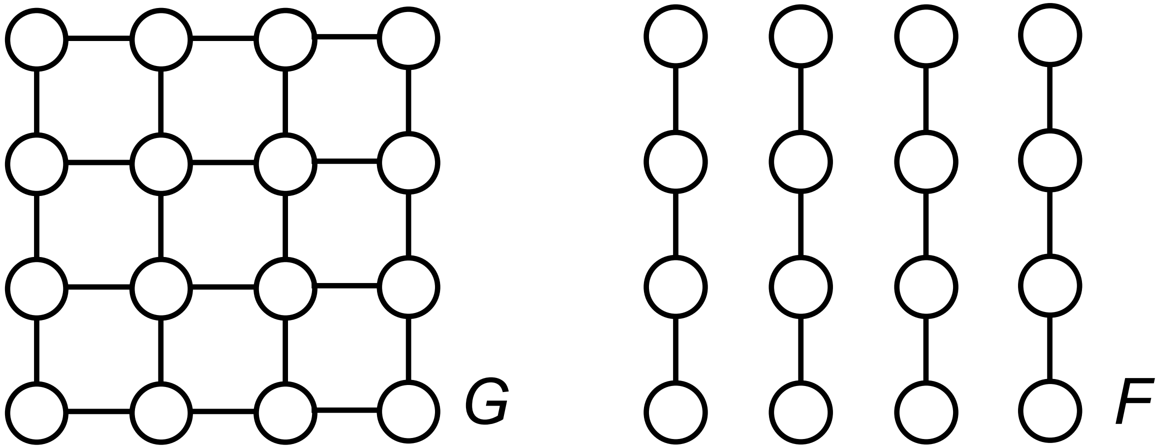
a) How to define approximating variational family $\mathcal{Q}$

b) How to optimize lower bound

Outline

- Variational Inference
- Mean Field Variational
- Stochastic Variational

Mean Field Variational Methods



Mean field assumes Markov with respect to sub-graph *F* of original graph *G*:

- Sub-graph picked so that entropy is “simple”, and thus optimization tractable

Mean field provides lower bound on true log-normalizer:

- Optimize over smaller set where true objective can be evaluated

Mean field optimization has local optima:

- Constraint set of distributions Markov w.r.t. subgraph F is non-convex

Naïve Mean Field

Assume discrete pairwise MRF model in *exponential family* form:

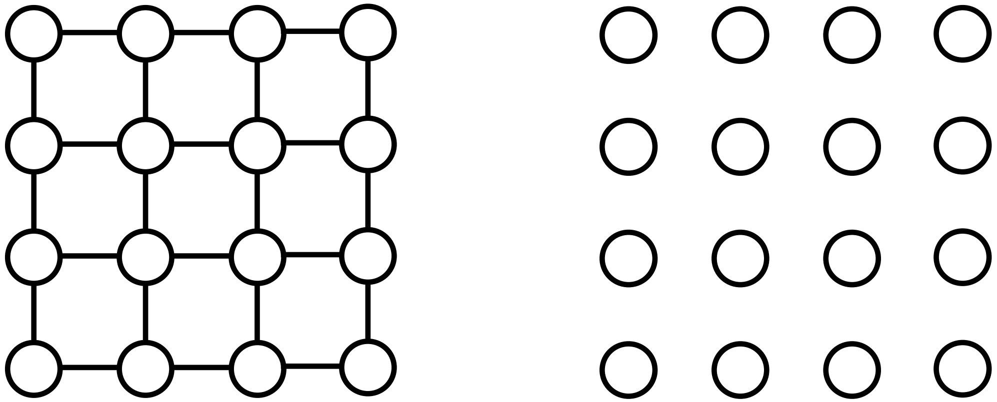
Absorbed observations into potential functions

$$p(x \mid \mathcal{Y}) \propto \exp\left\{ \sum_{(s,t)\in\mathcal{E}} \phi_{st}(x_s, x_t) + \sum_{s\in\mathcal{V}} \phi_s(x_s) \right\}$$

A *naïve mean field method* approximates distribution as fully factorized:

Free parameters to be optimized:

$$q(x) = \prod_{s\in\mathcal{V}} q_s(x_s) \qquad q_s(x_s = k) = \mu_{sk} \geq 0, \quad \textstyle\sum_{k=1}^{K_s} \mu_{sk} = 1.$$

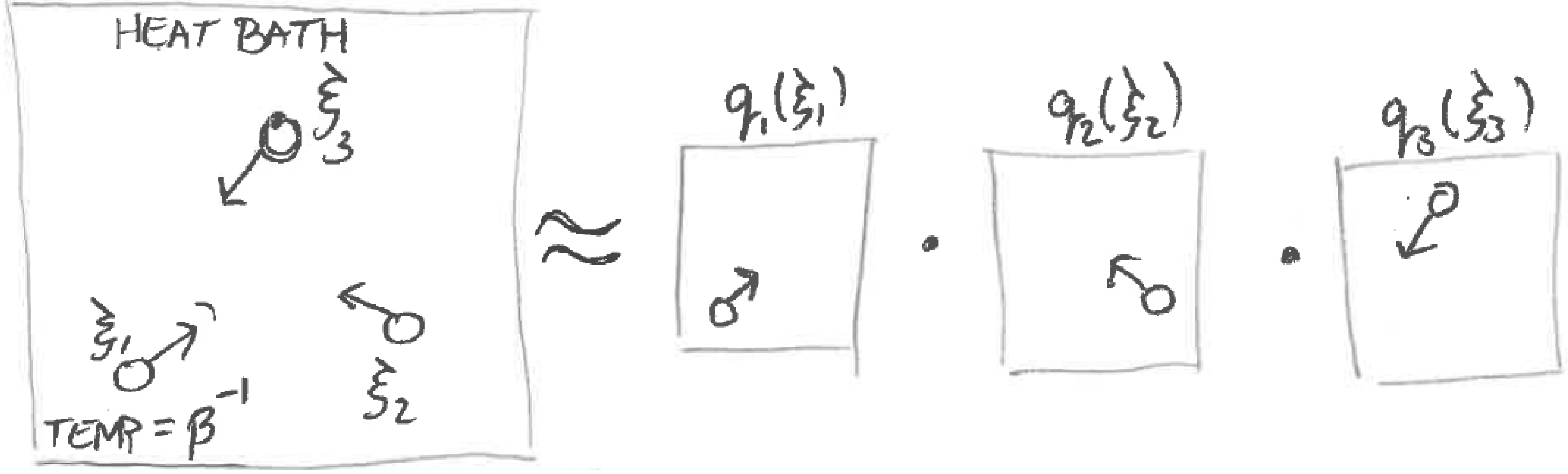


Why "Mean Field"?

Originates from the *many body problem* in statistical mechanics…

$\xi_i =$ "Microstates" e.g. spin, velocity, position, …

$p(\xi)$



Gibbs' distribution:

Hamiltonian

$$p(\xi) = \frac{1}{Z} e^{-\beta H(\xi)} \approx \prod_i \frac{1}{Z_i} e^{-\beta h_i(\xi_i)} \equiv \prod_i q_i(\xi_i)$$

Mean Field Lower Bound

Write optimization in terms of parameters μ:

$$\max_{\mu \geq 0} \mathcal{L}(\mu) \equiv \mathbb{E}_\mu[\log p(x, \mathcal{Y})] + H(\mu)$$

$$\text{subject to } \sum_{k=1}^{K_s} \mu_{sk} = 1 \;\; \forall s \in \mathcal{V}$$

For discrete pairwise MRF terms expand to:

$$H(\mu) = -\sum_{s \in \mathcal{V}} \sum_k \mu_{sk} \log \mu_{sk}$$

$$E(\mu) = \sum_{(s,t) \in \mathcal{E}} \sum_{k,\ell} \mu_{sk} \mu_{t\ell} \phi_{st}(k, \ell) + \sum_{s \in \mathcal{V}} \sum_k \mu_{sk} \phi_s(k)$$

Mean Field Algorithm : Pairwise MRF

1: Initialize parameters $\mu^{(0)}$, set i=0

2: While NOT converged

3: | i ← i+1

4: | For each node $s \in \mathcal{V}$ and value $k = 1, \ldots, K_s$

5: | | Update parameter μ_{sk} holding all others fixed

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

6: | Check if converged

Where we define: $\psi_s = \exp(\phi_s)$

Mean Field Updates : Pairwise MRF

$$\mathcal{L}(\mu) = \mathbb{E}_\mu[p(x)] + H(\mu) = \sum_{(s,t)\in\mathcal{E}} \sum_{k=1}^{K_s} \sum_{\ell=1}^{K_t} \mu_{sk}\mu_{t\ell}\phi(k,\ell) - \sum_{s\in\mathcal{V}} \sum_{k=1}^{K_s} \mu_{sk} \log \mu_{sk}$$

Updates via coordinate ascent on each parameter,

$$0 = \frac{\partial \mathcal{L}}{\partial \mu_{sk}} = \sum_{t\in\Gamma(s)} \sum_{\ell=1}^{K_t} \mu_{t\ell}\phi(k,\ell) + \phi_s(k) - \log \mu_{sk} - 1$$

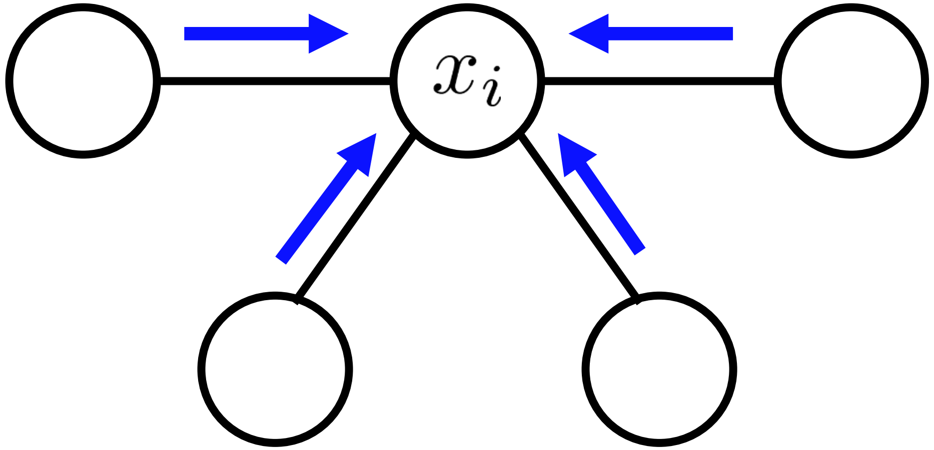
$$\log \mu_{sk} = \sum_{t\in\Gamma(s)} \sum_{\ell=1}^{K_t} \mu_{t\ell}\phi(k,\ell) + \phi_s(k) - 1$$

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

Normalization enforced via Lagrange multiplier (I glossed over this)

Pairwise MRF Mean Field as Message Passing

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s) \qquad \phi_{st}(x_s, x_t) = \log \psi_{st}(x_s, x_t)$$



$$q_i(x_i) \propto \psi_i(x_i) \prod_{j\in\Gamma(i)} m_{ji}(x_i) \qquad m_{ji}(x_i) \propto \exp\left\{\mathbb{E}_{q_j}[\phi_{ij}(x_i, x_j)]\right\}$$

- Compared to *belief propagation*, has identical formula for estimating marginals from messages, but a different message update equation
- If neighboring marginals degenerate to single state, recover *Gibbs sampling* message

General Mean Field Updates

1: Initialize mean field distributions $q_s(x_s)$

2: While NOT converged

3: | For each node $s \in \mathcal{V}$

4: | | Update marginal $q_s(x_s)$ holding all others fixed

| |

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

| |

5: | Check if converged

- Here $\mathbb{E}_{q_{\setminus s}}[\cdot]$ is expectation w.r.t. all marginals besides $q_s(x_s)$
- Expectation only depends on variables in Markov blanket

Derivation of General Mean Field Updates

Mean field variational lower bound,

$$\log p(\mathcal{Y}) \geq L(q) \equiv \mathbb{E}_q[\log \widetilde{p}(x)] + \sum_i H(q_i)$$

where we use shorthand $\widetilde{p}(x) \equiv p(x, \mathcal{Y})$

Notice joint entropy decomposes to sum of marginal entropies

$$H(q) = -\sum_x \prod_i q_i(x_i) \sum_k \log q_k(x_k) = \sum_i H(q_i)$$

To update q_j view bound as function of q_j and do coordinate ascent…

Derivation of General Mean Field Updates

$$
\begin{aligned}
L(q_j) &= \sum_{\mathbf{x}} \prod_i q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right] \\
&= \sum_{\mathbf{x}_j} \sum_{\mathbf{x}_{-j}} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right]
\end{aligned}
$$

Derivation of General Mean Field Updates

$$L(q_j) = \sum_{\mathbf{x}} \prod_i q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right]$$

$$= \sum_{\mathbf{x}_j} \sum_{\mathbf{x}_{-j}} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right]$$

Linearity of expectation $= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x})$

$$- \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\sum_{k \neq j} \log q_k(\mathbf{x}_k) + q_j(\mathbf{x}_j) \right]$$

Derivation of General Mean Field Updates

$$L(q_j) = \sum_{\mathbf{x}} \prod_i q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right]$$

$$= \sum_{\mathbf{x}_j} \sum_{\mathbf{x}_{-j}} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right]$$

Linearity of expectation

$$= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x})$$

$$- \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\sum_{k \neq j} \log q_k(\mathbf{x}_k) + q_j(\mathbf{x}_j) \right]$$

Group terms not Involving q_j to const.

$$= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) + \text{const}$$

Where, $\log f_j(\mathbf{x}_j) \triangleq \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x}) = \mathbb{E}_{-q_j} [\log \tilde{p}(\mathbf{x})]$

Derivation of General Mean Field Updates

Thus we have,

$$L(q_j) = \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) + \text{const}$$

Where,

$$\log f_j(\mathbf{x}_j) \triangleq \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x}) = \mathbb{E}_{-q_j} [\log \tilde{p}(\mathbf{x})]$$

Observing that by definition of the Kullback-Leibler divergence we have,

$$L(q_j) = -\mathbb{KL}(q_j || f_j)$$

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

Which we maximize by setting $q_j = f_j$ as,

$$q_j(\mathbf{x}_j) = \frac{1}{Z_j} \exp\left(\mathbb{E}_{-q_j}[\log \tilde{p}(\mathbf{x})]\right)$$

Conditionally Conjugate Models

The coordinate update does not have a closed form for all models…

$$q_j(\mathbf{x}_j) \quad = \quad \frac{1}{Z_j} \exp\left(\mathbb{E}_{-q_j}\left[\log \tilde{p}(\mathbf{x})\right]\right)$$

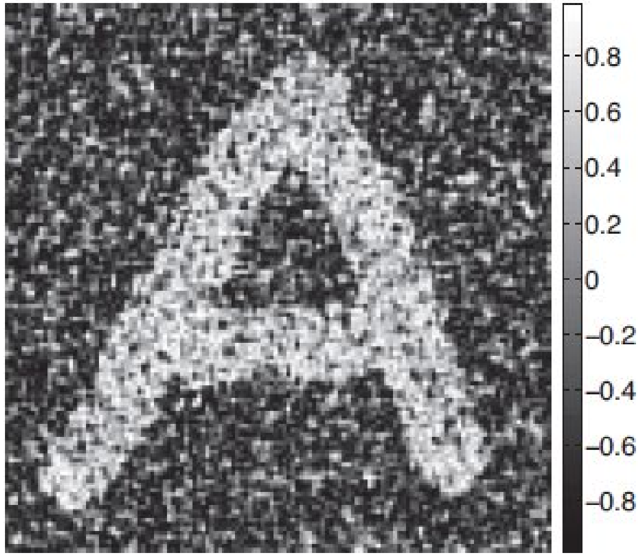
One case where things work out nice is *conditionally conjugate* models

$$\widetilde{p}(x) = \widetilde{p}_j(x_j)\widetilde{p}_{-j}(x_{-j} \mid x_j) \propto \widetilde{p}_j(x_j \mid x_{-j})$$

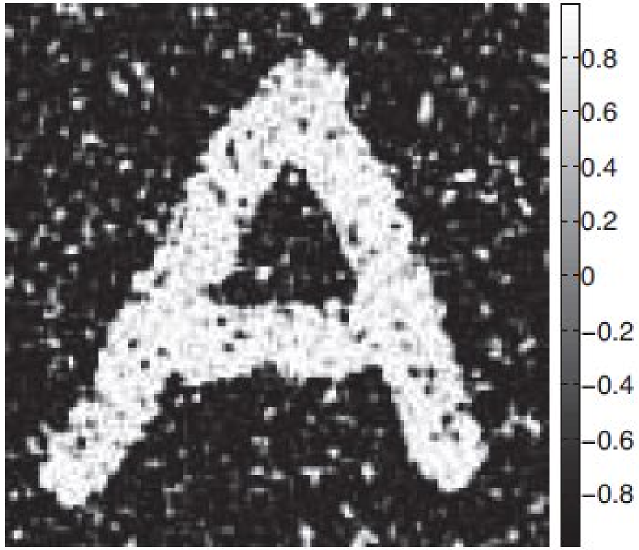
- In conditionally conjugate models $\widetilde{p}_j(x_j)$ is the *same distribution family* as the *complete conditional* $\widetilde{p}_j(x_j \mid x_{-j})$
- Similar, but stronger, condition to Gibbs sampler
- In Gibbs sampler the complete conditionals must be easy to sample, not necessarily conjugate

Example: Image Denoising

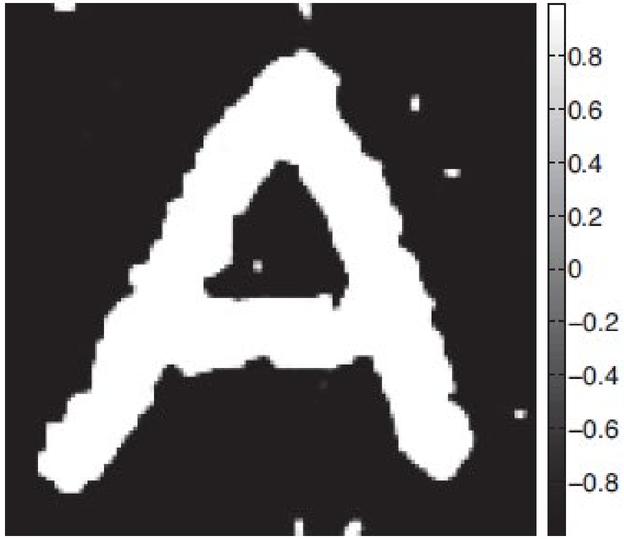
Noisy Image



3 Iterations of MF



15 Iterations of MF



Model is pairwise MRF on binary variables $x_i \in \{0, 1\}$ (a.k.a. "Ising" model)

$$p(\mathbf{x}) = \frac{1}{Z_0} \exp(-E_0(\mathbf{x})) \qquad p(\mathbf{y}|\mathbf{x}) = \prod_i p(\mathbf{y}_i|x_i) = \sum_i \exp(-L_i(x_i))$$

Where, $$E_0(\mathbf{x}) = -\sum_{i=1}^{D} \sum_{j \in \mathrm{nbr}_i} W_{ij} x_i x_j$$

Source: K. Murphy

Example: Image Denoising

Naïve mean field assumption—fully factorized variational approximation,

$$q(\mathbf{x}) = \prod_i q(x_i, \mu_i)$$

MF probability param for node i

Write out unnormalized log-joint probability,

$$\log \tilde{p}(\mathbf{x}) = x_i \sum_{j \in \mathrm{nbr}_i} W_{ij} x_j + L_i(x_i) + \mathrm{const}$$

Expectation w.r.t. neighbors of x_i (e.g. Markov blanket),

$$\mathbb{E}_{q_{-i}} \left[\log \widetilde{p}(x) \right] = x_i \sum_{j \in \mathrm{nbr}_i} W_{ij} \mu_j + L_i(x_i)$$

Update for q_i is exponentiated expectation w.r.t. Markov blanket,

$$q_i(x_i) \propto \exp\left(x_i \boxed{\sum_{j \in \mathrm{nbr}_i} W_{ij} \mu_j} + L_i(x_i) \right)$$

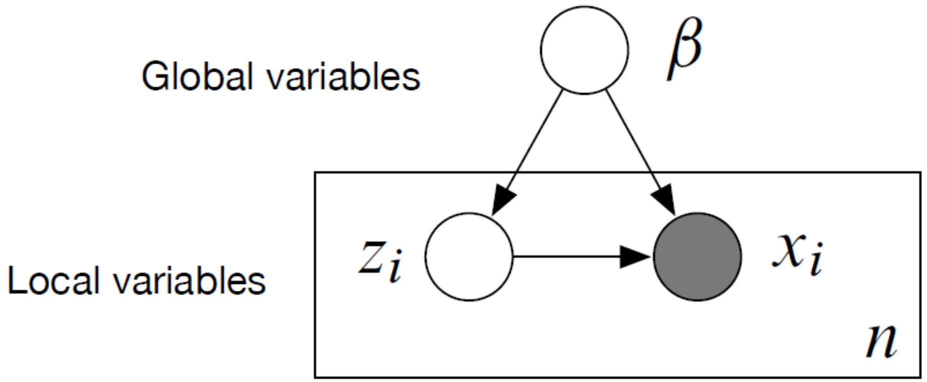
Average of
neighboring states

Source: K. Murphy

Outline

- Variational Inference
- Mean Field Variational
- Stochastic Variational

A Generic Class of Directed Models

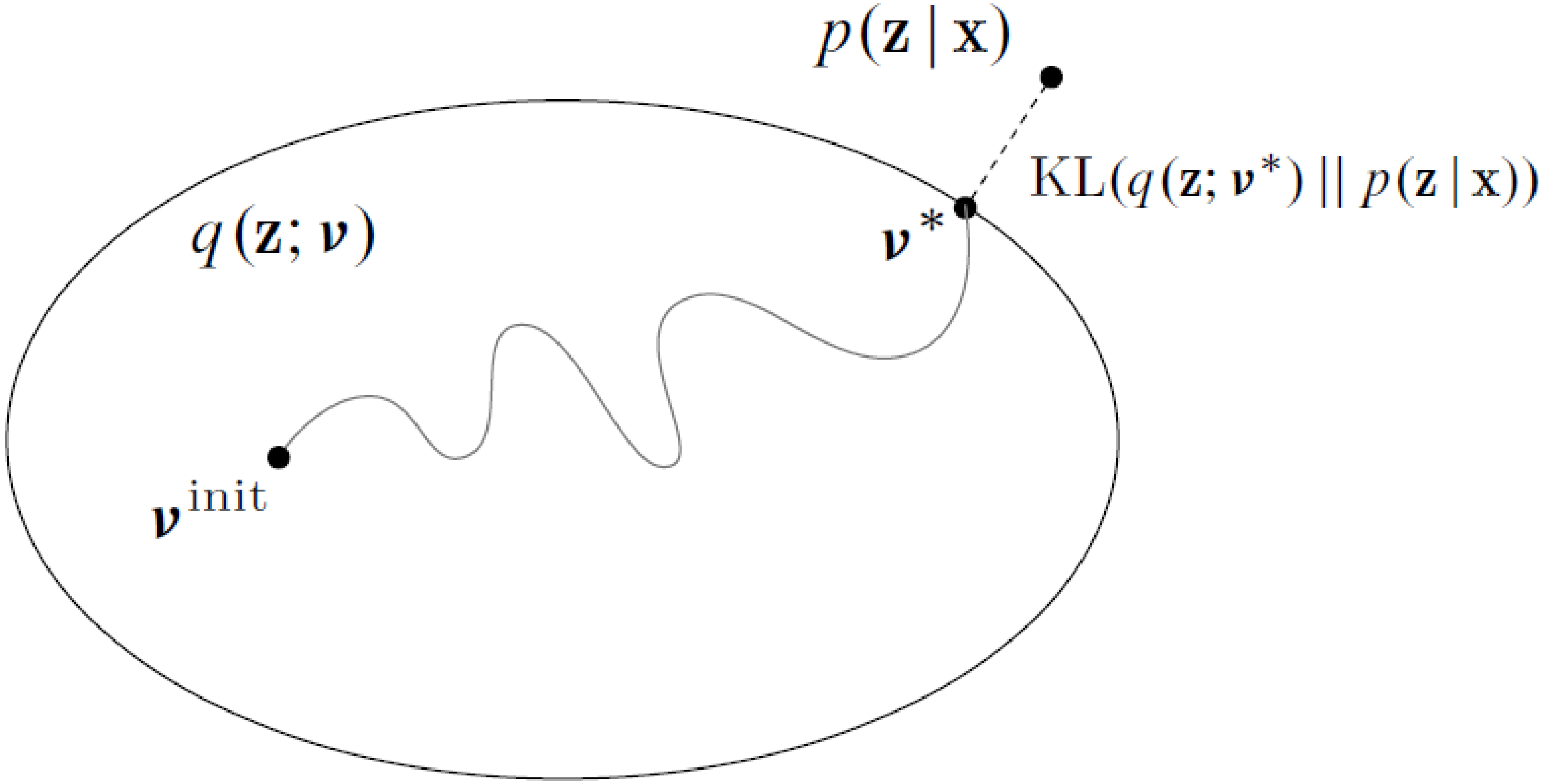


$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta)\prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- 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]

Variational Approximation



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

[Source: David Blei]

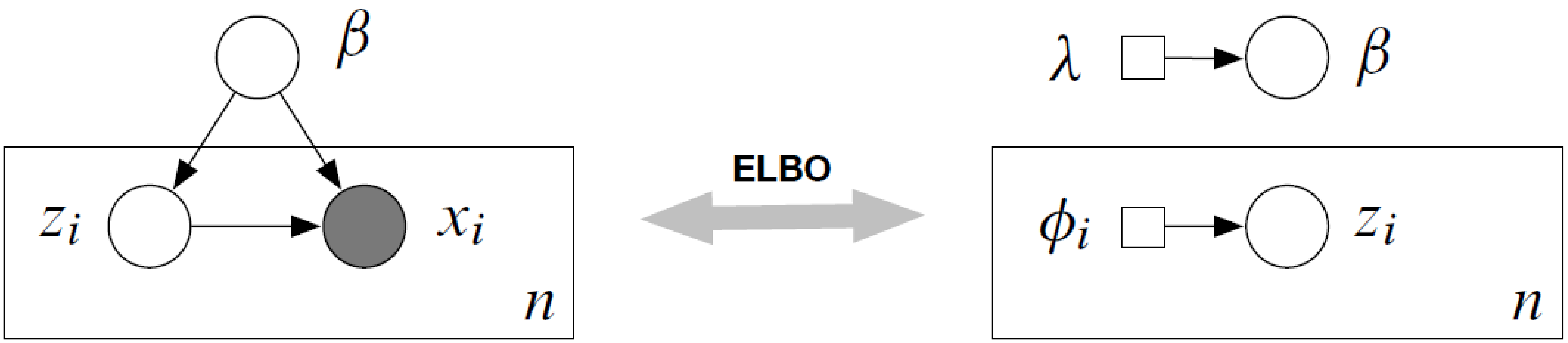
Variational Lower Bound – ELBO

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

- 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]

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

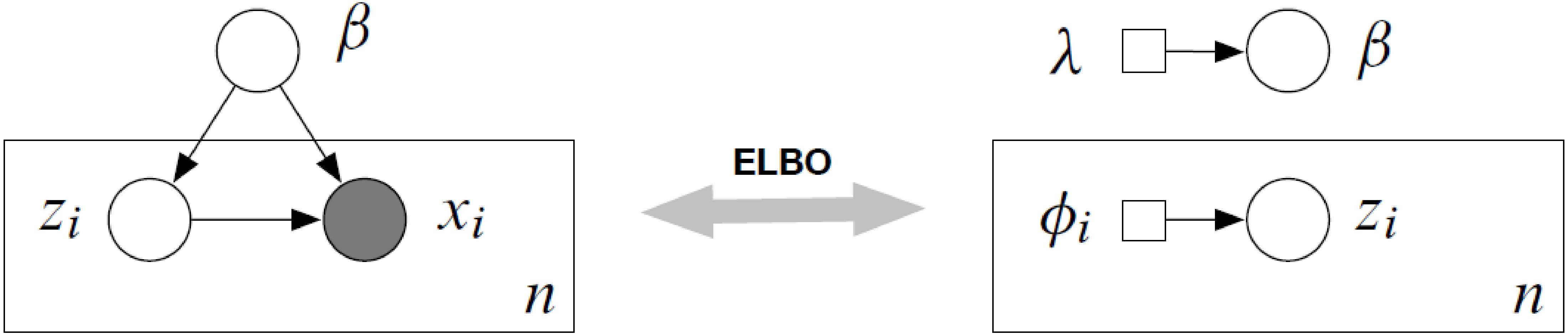
Conditional conjugacy: Each factor is the same expfam as complete conditional

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

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

[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

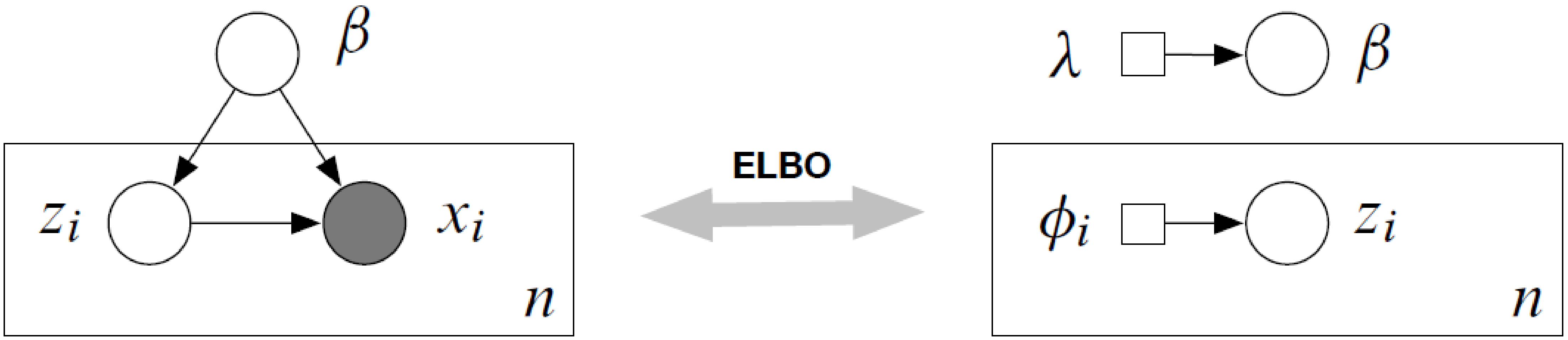
Global parameter ensure conjugacy to (z,x):

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

where α is prior hyperparameter and t(.) are sufficient stats 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

By gradient 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

Need to visit *every* data point

  Set global parameter

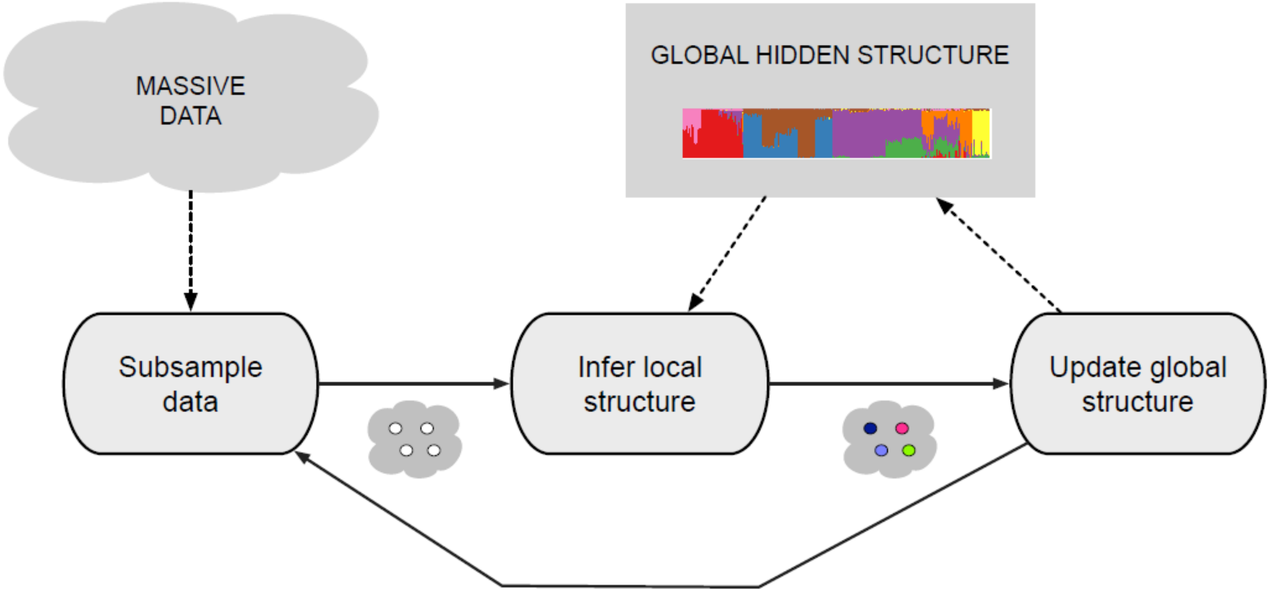
$$\lambda \leftarrow \alpha + \textstyle\sum_{i=1}^n \mathbb{E}_{\phi_i}[t(Z_i, x_i)]\,.$$

Need to sum *every* data point

until *the ELBO has converged*

[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. DNN learning)

[Source: David Blei]

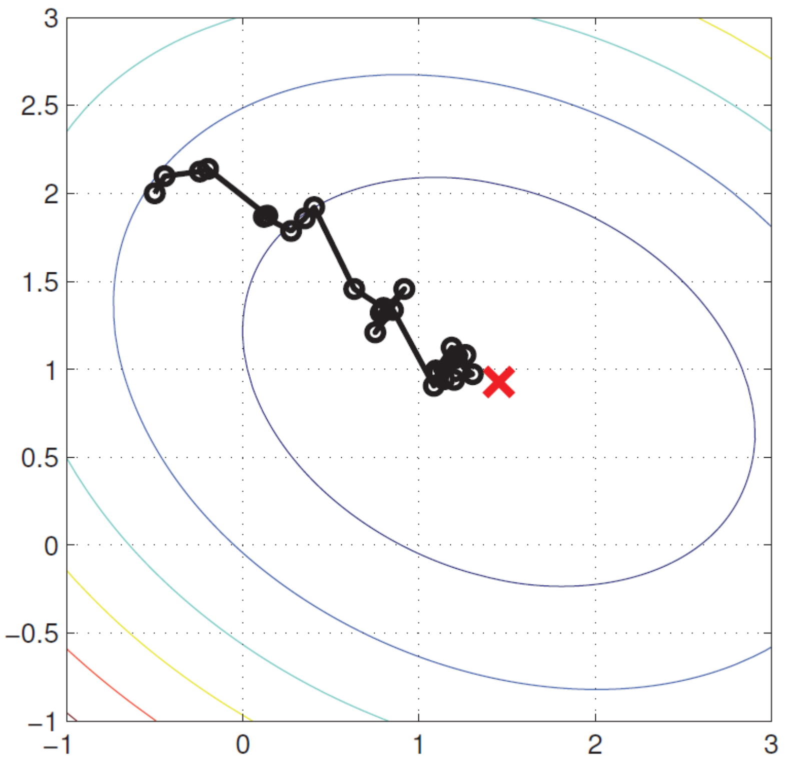
Stochastic Gradient Ascent/Descent

- 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*

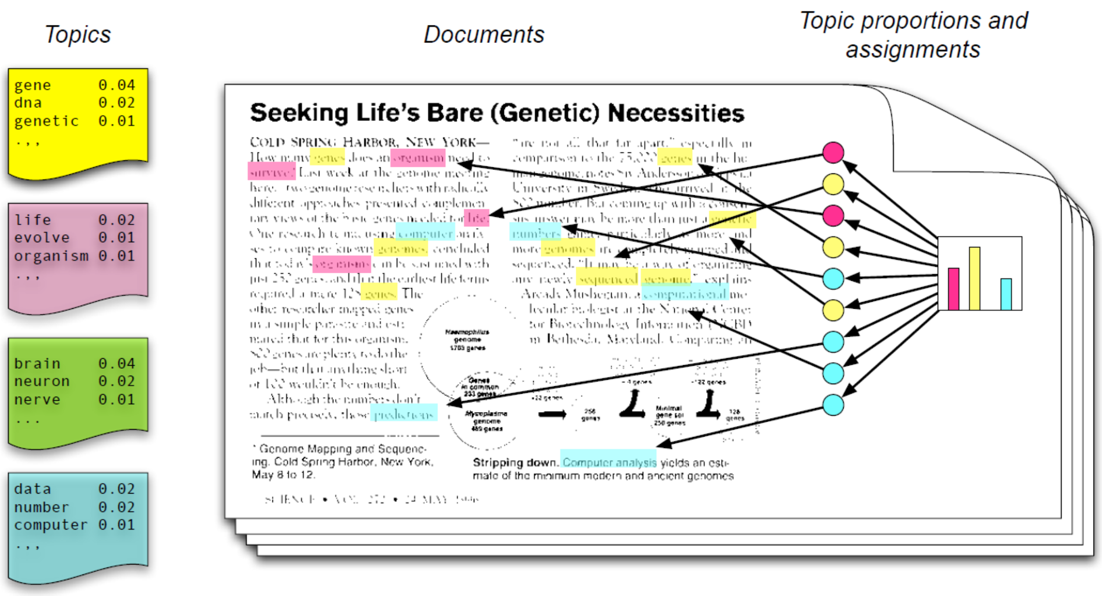
[Source: David Blei]

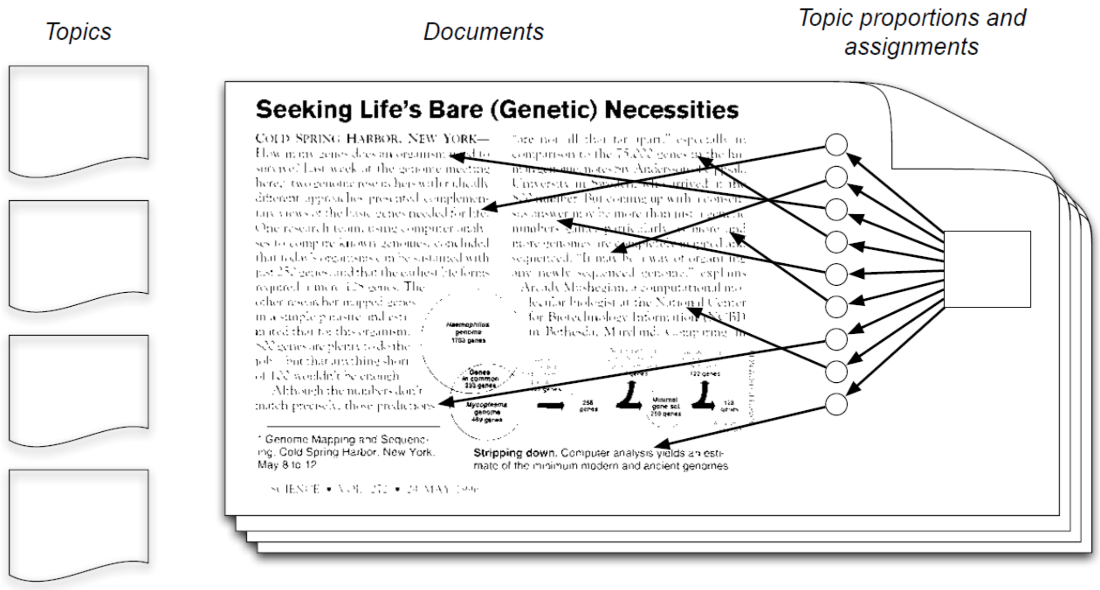
Topic Models

Topic models discover hidden thematic structure in large collections of documents

[Source: David Blei]

Topic Models



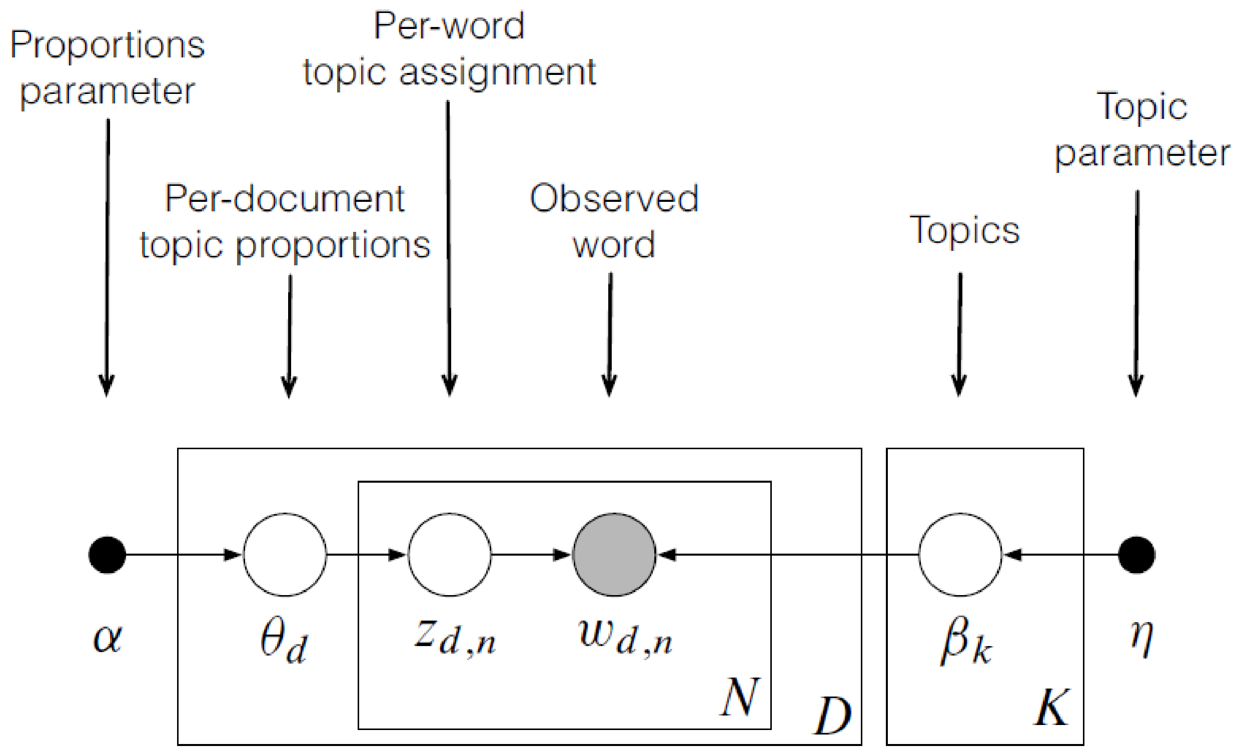


- Each *topic* is a distribution over words (vocabulary)
- Each *document* is a mixture of corpus-wide topics
- Each *word* is drawn from one of the topics (they are distributions)
- But we only observe documents; everything else is hidden (unsupervised learning problem)
- Need to calculate posterior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

P(topics, proportions, assignments | documents)

[Source: David Blei]

Example: Latent Dirichlet Allocation



Latent Dirichlet Allocation (LDA):

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

- 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]

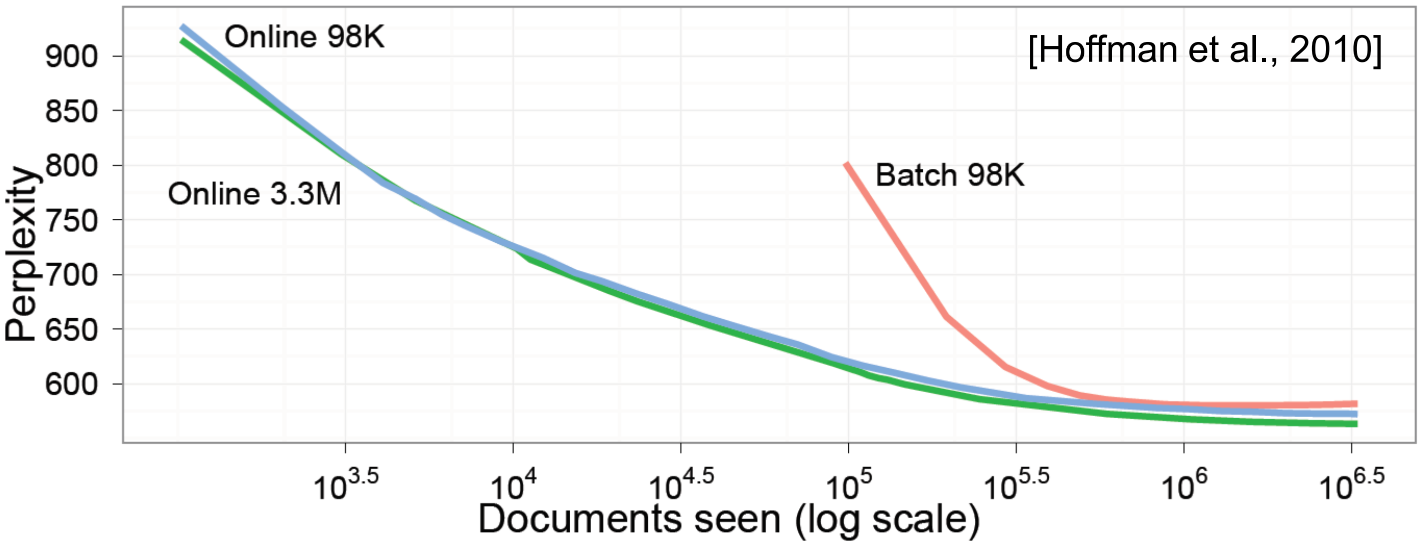
Example: Latent Dirichlet Allocation

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

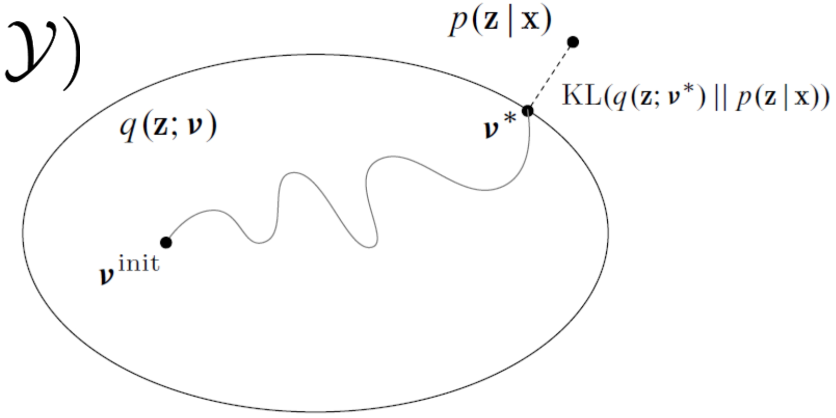
2) Choose a family of approximating distributions $\mathcal{Q}$ that is tractable

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

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

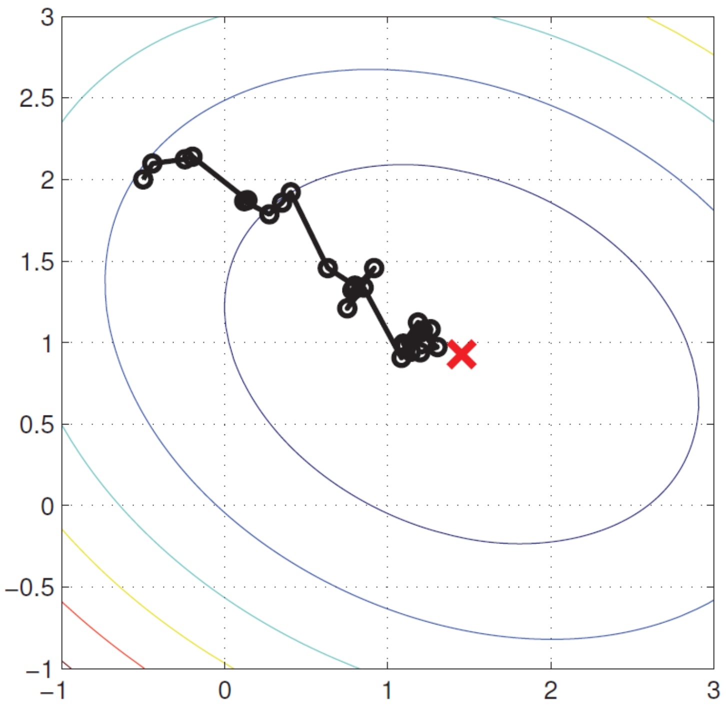
- Coordinate ascent updates require complete conditionals to be conjugate
  - Similar, but stricter, assumption to Gibbs sampling
- MF update takes specific form depending on model $p(.)$, e.g. pairwise MRF:

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

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