LINKÖPING UNIVERSITY

Variational inference

Fredrik Lindsten, Linköping University

2023-10-31

Outline

Aim: Introduce variational inference (VI) and show how it can be used to solve the LDA inference problem, even in the large data regime.

Outline:

1. Variational inference: *Bayesian inference as optimization*
2. The mean-field approximation and CAVI
   (illustrated using LDA)
3. Stochastic VI for problems with large data

Bayesian inference

Task: Given a **probabilistic model** $p(\mathbf{y}, \mathbf{x})$ and **observed data** $\mathbf{y}$, compute the posterior distribution

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x})}{p(\mathbf{y})}.$$

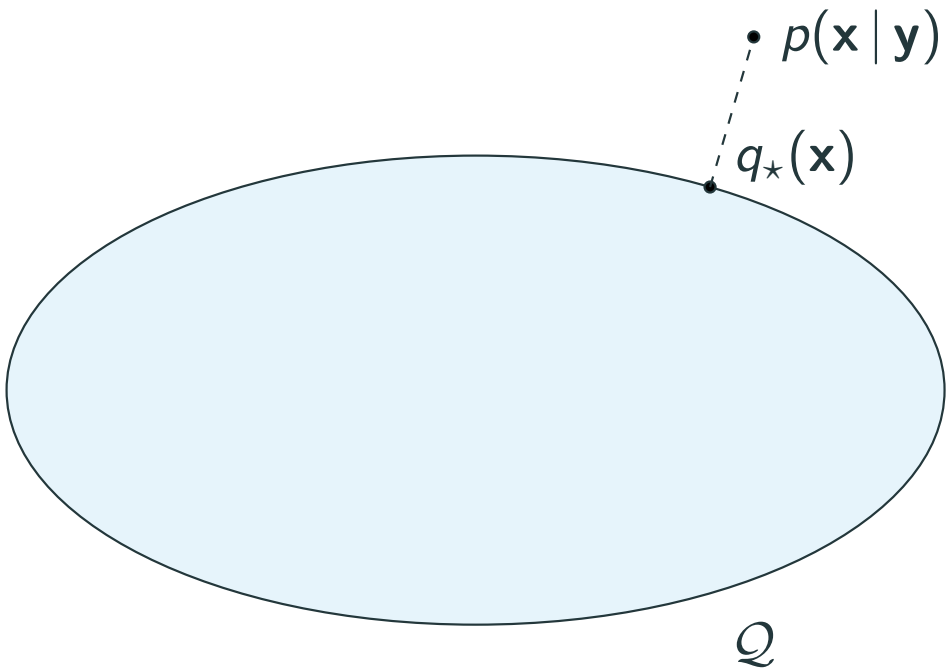
Inference as optimization

Variational inference turns probabilistic inference into optimization.

- $p(\mathbf{x} \mid \mathbf{y})$

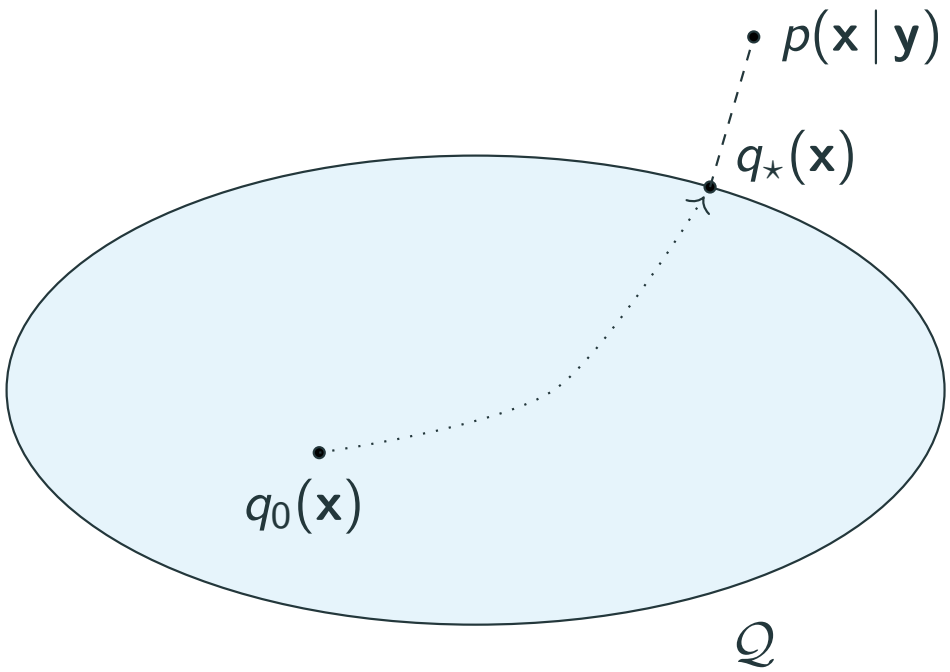
Inference as optimization

Variational inference turns probabilistic inference into optimization.



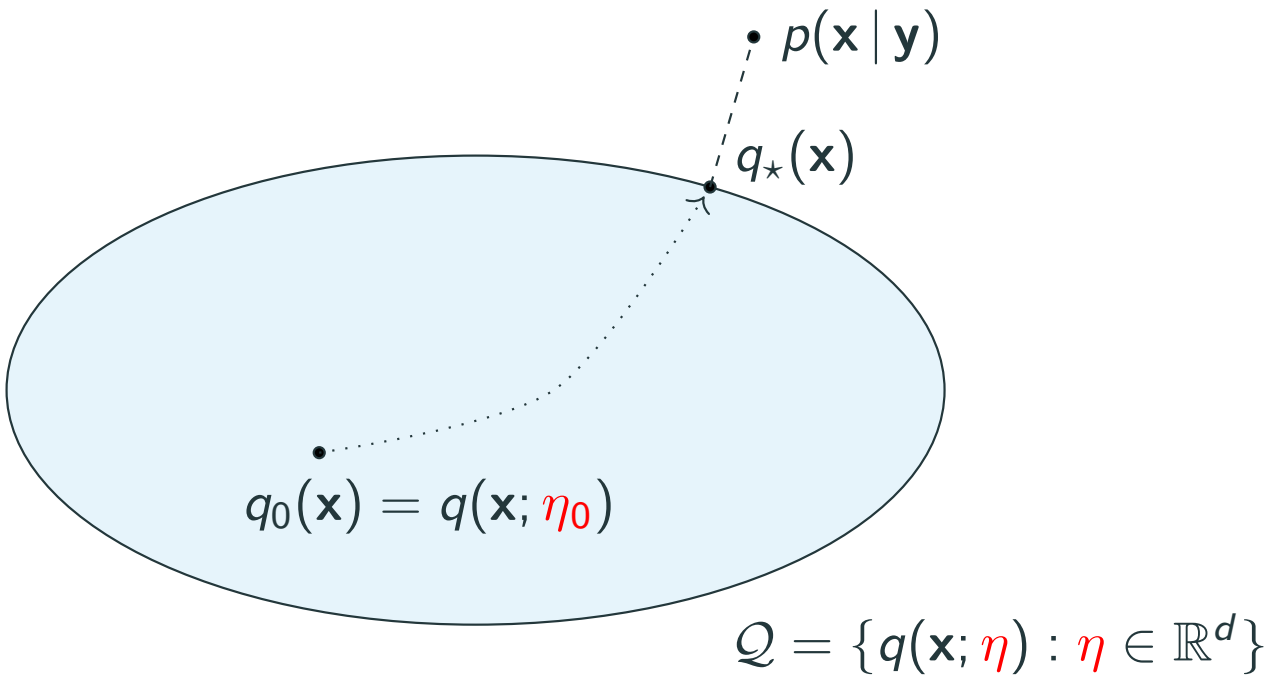
Inference as optimization

Variational inference turns probabilistic inference into optimization.



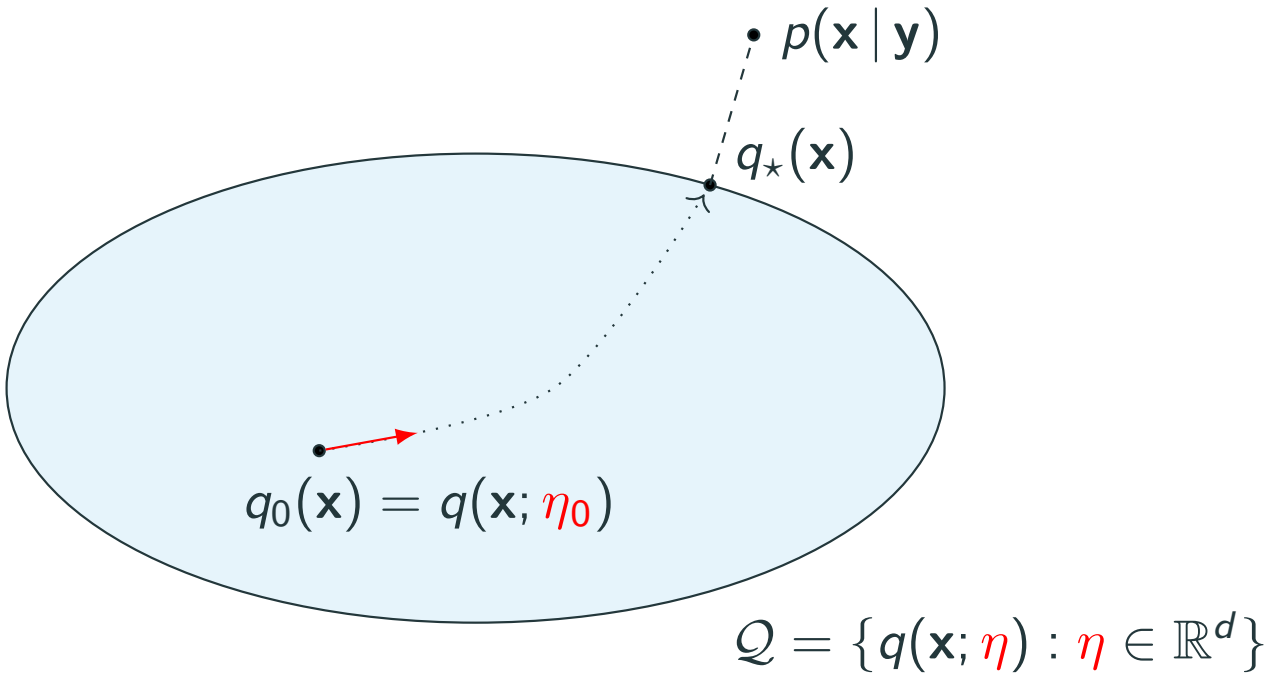
Inference as optimization

Variational inference turns probabilistic inference into optimization.



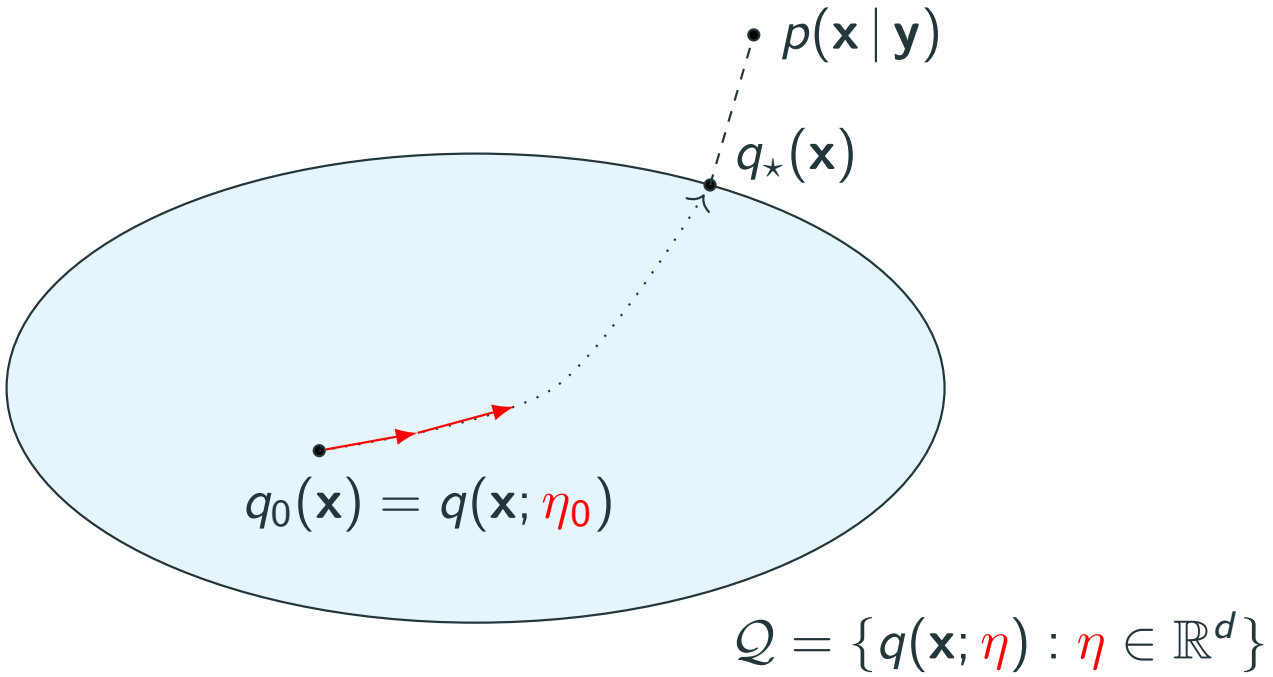
Inference as optimization

Variational inference turns probabilistic inference into optimization.



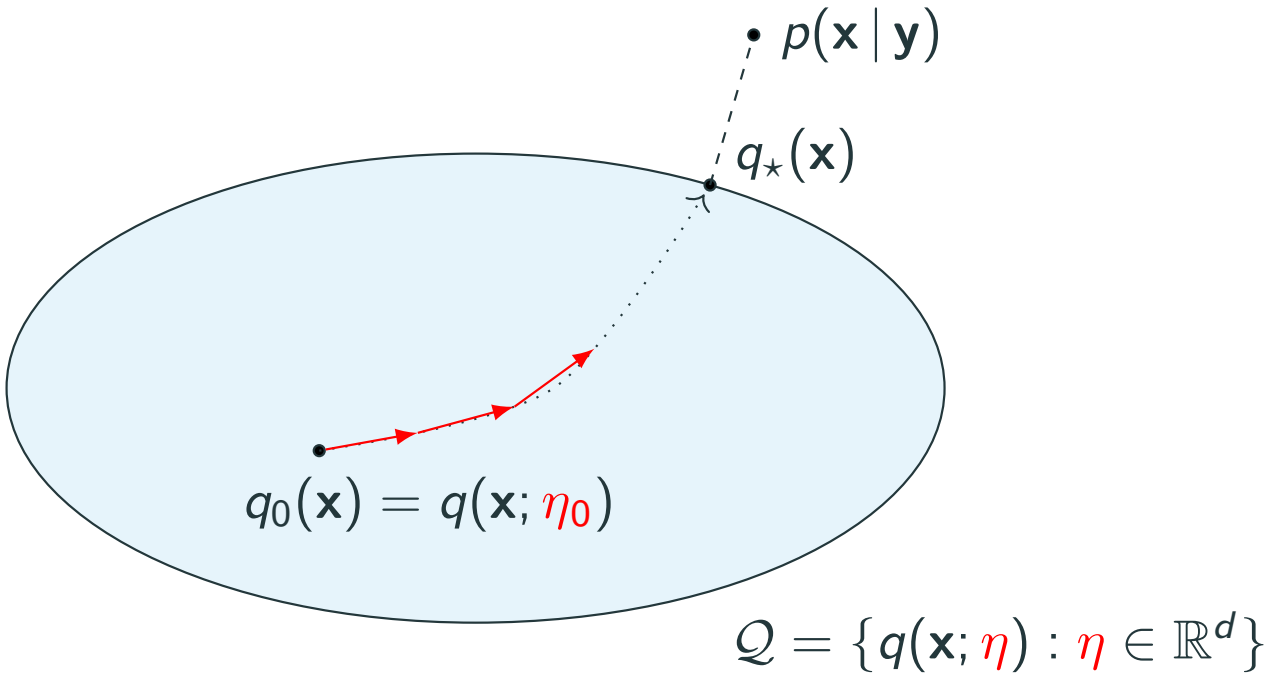
Inference as optimization

Variational inference turns probabilistic inference into optimization.



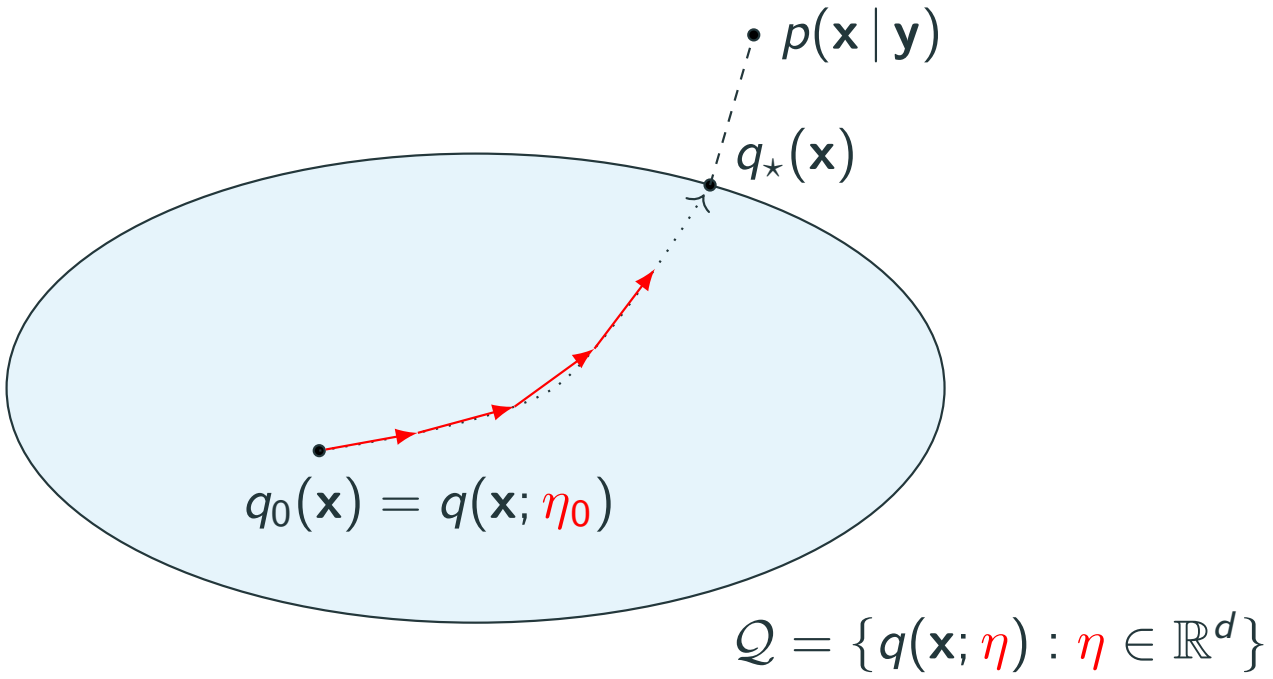
Inference as optimization

Variational inference turns probabilistic inference into optimization.



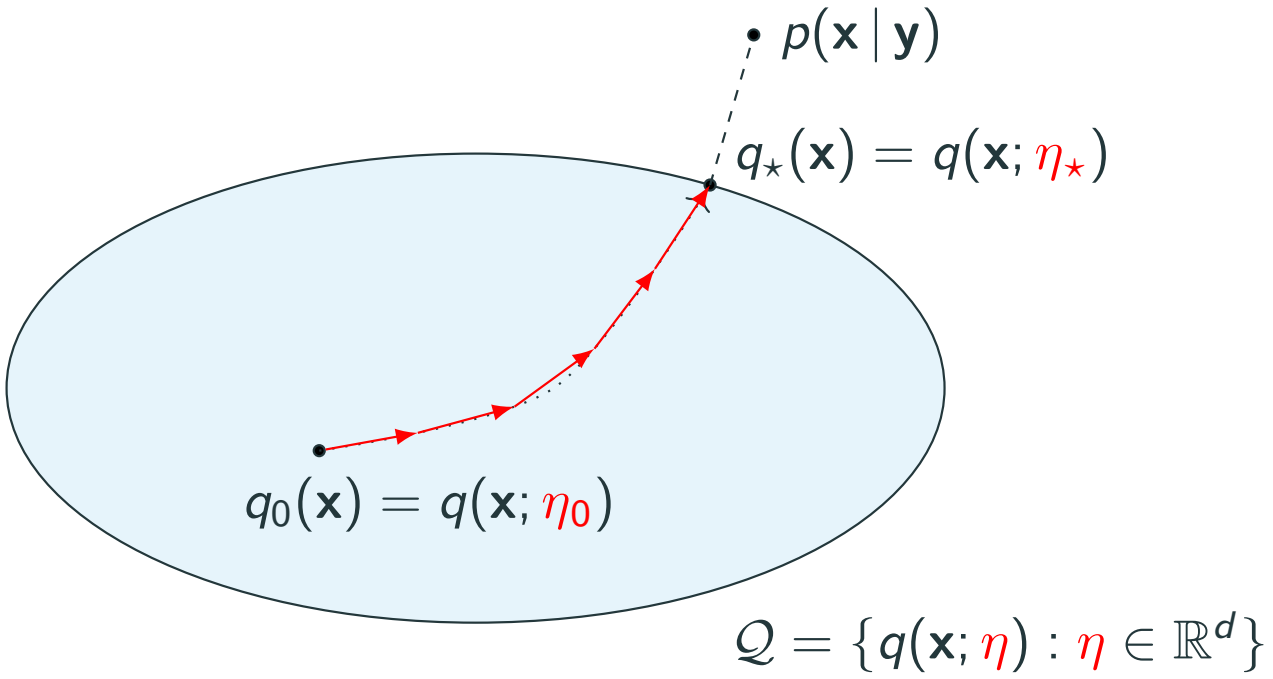
Inference as optimization

Variational inference turns probabilistic inference into optimization.



Inference as optimization

Variational inference turns probabilistic inference into optimization.



Inference as optimization

We can make use of our extensive optimization toolbox to solve inference problems! Innovations in optimization directly carry over to inference.

Inference as optimization

We can make use of our extensive optimization toolbox to solve inference problems! Innovations in optimization directly carry over to inference.

Some details still need to be sorted out...

1. How do we measure the similarity between $q(\mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{y})$?

Inference as optimization

We can make use of our extensive optimization toolbox to solve inference problems! Innovations in optimization directly carry over to inference.

Some details still need to be sorted out...

1. How do we measure the similarity between $q(\mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{y})$?
2. What is the objective function?

Inference as optimization

We can make use of our extensive optimization toolbox to solve inference problems! Innovations in optimization directly carry over to inference.

Some details still need to be sorted out...

1. How do we measure the similarity between $q(\mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{y})$?
2. What is the objective function?
3. How do we specify $\mathcal{Q}$?

Inference as optimization

We can make use of our extensive optimization toolbox to solve inference problems! Innovations in optimization directly carry over to inference.

Some details still need to be sorted out...

1. How do we measure the similarity between $q(\mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{y})$?
2. What is the objective function?
3. How do we specify $\mathcal{Q}$?
4. How do we solve the optimization problem?

Kullback–Leibler minimization and the ELBO

Measuring similarity

Detail #1: We need a way to measure the **closeness** between $q(\mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{y})$.

Measuring similarity

Detail #1: We need a way to measure the **closeness** between $q(\mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{y})$. In variational inference, the typical choice is:

Kullback–Leibler divergence:

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

Why? Main reason is that it leads to a tractable objective function.

Measuring similarity

Detail #1: We need a way to measure the **closeness** between $q(\mathbf{x})$ and $p(\mathbf{x} \mid \mathbf{y})$. In variational inference, the typical choice is:

Kullback–Leibler divergence:

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

Why? Main reason is that it leads to a tractable objective function.

Note, this is different from EP which uses the reverse Kullback–Leibler divergence, $\mathrm{KL}(p(\cdot \mid \mathbf{y}) \| q)$.

The variational inference problem

Detail #2: The VI optimization problem can thus be formulated as:

Find an approximation of the posterior $q_\star(\mathbf{x}) \approx p(\mathbf{x} \mid \mathbf{y})$ by solving,

$$q_\star(\mathbf{x}) = \arg\max_{q \in \mathcal{Q}} \text{ELBO}(q),$$

where

$$\text{ELBO}(q) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{y})] - \mathbb{E}_q[\log q(\mathbf{x})].$$

The mean field approximation

Specifying $\mathcal{Q}$

Detail #3: How do we specify $\mathcal{Q}$?

A **richer family** of distributions will **improve the accuracy at the optimum**, but it can also be **more difficult** to optimize over.

Specifying $\mathcal{Q}$

Detail #3: How do we specify $\mathcal{Q}$?

A **richer family** of distributions will **improve the accuracy at the optimum**, but it can also be **more difficult** to optimize over.

Classical choice: Mean-field factorization,

$$q(\mathbf{x}) = \prod_{k=1}^{m} q_k(x_k).$$

Not as restrictive as it may seem! Possible to fit all posterior marginals.

Specifying $\mathcal{Q}$

Detail #3: How do we specify $\mathcal{Q}$?

A **richer family** of distributions will **improve the accuracy at the optimum**, but it can also be **more difficult** to optimize over.

Classical choice: Mean-field factorization,

$$q(\mathbf{x}) = \prod_{k=1}^{m} q_k(x_k).$$

Not as restrictive as it may seem! Possible to fit all posterior marginals.

N.B. This is the same factorization that we assumed in EP to derive message passing.

ex) Mean field approximation for LDA

The LDA posterior distribution is given by,[1]

$$
\begin{aligned}
p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta} \mid \mathbf{w}) &\propto p(\boldsymbol{\beta}) p(\boldsymbol{\theta}) p(\mathbf{z} \mid \boldsymbol{\theta}) p(\mathbf{w} \mid \mathbf{z}, \boldsymbol{\beta}) \\
&= \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} \left[p(\theta_d) \prod_{n=1}^{N_d} \left[p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid z_{d,n}, \beta_{z_{d,n}}) \right] \right]
\end{aligned}
$$

[1] We drop the hyperparameters of the prior distributions, η and α, from the notation for brevity. They are considered fixed and known throughout this course module.

ex) Mean field approximation for LDA

The LDA posterior distribution is given by,[1]

$$
\begin{aligned}
p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta} \mid \mathbf{w}) &\propto p(\boldsymbol{\beta}) p(\boldsymbol{\theta}) p(\mathbf{z} \mid \boldsymbol{\theta}) p(\mathbf{w} \mid \mathbf{z}, \boldsymbol{\beta}) \\
&= \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} \left[p(\theta_d) \prod_{n=1}^{N_d} \left[p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid z_{d,n}, \beta_{z_{d,n}}) \right] \right]
\end{aligned}
$$

The mean field approximation is of the form,

$$
q(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}) = \prod_{k=1}^{K} q(\beta_k) \prod_{d=1}^{D} \left[q(\theta_d) \prod_{n=1}^{N_d} q(z_{d,n}) \right]
$$

[1] We drop the hyperparameters of the prior distributions, η and α, from the notation for brevity. They are considered fixed and known throughout this course module.

Coordinate Ascent Variational Inference

Detail #4: How do we optimize the ELBO?

The mean field assumption enables an efficient coordinate-ascent-type method.

Coordinate Ascent Variational Inference (CAVI)

Initialize $q_k(x_k)$ arbitrarily, $k = 1, \ldots, m$.

1: **while** the ELBO has not converged **do**
2: **for** $k = 1, \ldots, m$ **do**
3: Update the kth factor (keeping the other factors fixed)

$$q_k^{\text{new}}(x_k) = \arg\max_{q_k} \text{ELBO}(q_k; q_{-k}).$$

4: **end for**
5: **end while**

Computing the updates

What does $q_k^*(x_k) \propto \exp\left(\mathbb{E}_{q_{-k}}\left[\log p(x_k \mid \mathbf{x}_{-k}, \mathbf{y})\right]\right)$ mean in practice?!

Computing the updates

What does $q_k^*(x_k) \propto \exp\left(\mathbb{E}_{q_{-k}}\left[\log p(x_k \mid \mathbf{x}_{-k}, \mathbf{y})\right]\right)$ mean in practice?!

ex) CAVI for LDA. Consider the topic proportion vector θ_d for document d. From before (cf. Gibbs sampler) we have,

$$p(\theta_d \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}) = \mathrm{Dir}(\theta_d \mid \alpha + c_d)$$

where $c_{d,k} = \sum_n \mathbb{1}\{z_{d,n} = k\}$ for $k = 1, \ldots, K$.

Computing the updates

What does $q_k^*(x_k) \propto \exp\left(\mathbb{E}_{q_{-k}}\left[\log p(x_k \mid \mathbf{x}_{-k}, \mathbf{y})\right]\right)$ mean in practice?!

ex) CAVI for LDA. Consider the topic proportion vector θ_d for document d. From before (cf. Gibbs sampler) we have,

$$p(\theta_d \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}) = \mathrm{Dir}(\theta_d \mid \alpha + c_d) \propto \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k + c_{d,k}}$$

where $c_{d,k} = \sum_n \mathbb{1}\{z_{d,n} = k\}$ for $k = 1, \ldots, K$.

Computing the updates

What does $q_k^*(x_k) \propto \exp\left(\mathbb{E}_{q_{-k}}\left[\log p(x_k \mid \mathbf{x}_{-k}, \mathbf{y})\right]\right)$ mean in practice?!

ex) CAVI for LDA. Consider the topic proportion vector θ_d for document d. From before (cf. Gibbs sampler) we have,

$$p(\theta_d \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}) = \mathrm{Dir}(\theta_d \mid \alpha + c_d) \propto \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k + c_{d,k}}$$

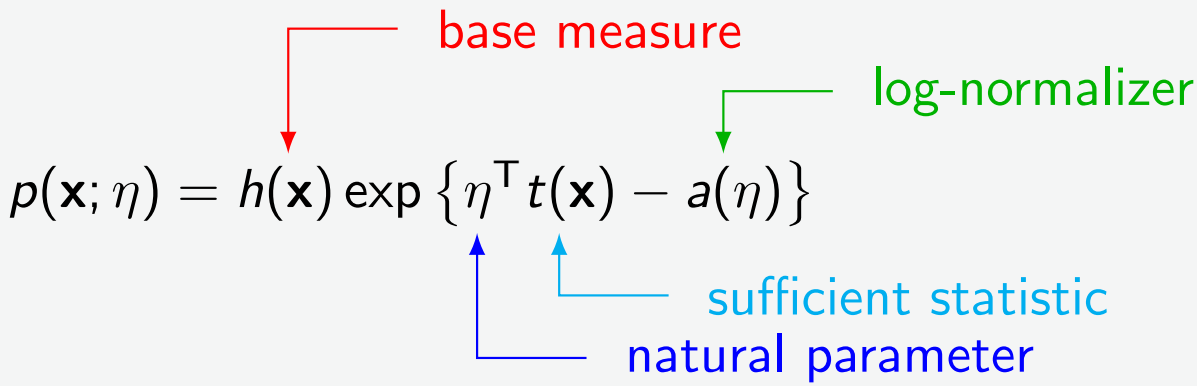
where $c_{d,k} = \sum_n \mathbb{1}\{z_{d,n} = k\}$ for $k = 1, \ldots, K$.

We get $q^*(\theta_d) = \mathrm{Dir}(\theta_d \mid \alpha + \mathbb{E}_q[c_d])$

CAVI for conditional exponential family models

This computation can be repeated for any **conditional distribution** in the **exponential family.**

Def. A probability distribution $p(\mathbf{x})$ belongs to the **exponential family** with (natural) hyper-parameter η if its PDF can be written as:



CAVI for conditional exponential family models

Assume that all **complete conditionals** of $p(\mathbf{x} \mid \mathbf{y})$ belong to the exponential family,

$$p(x_k \mid \mathbf{x}_{-k}, \mathbf{y}) = h_k(x_k) \exp\left\{\eta_k(\mathbf{x}_{-k}, \mathbf{y})^\mathsf{T} t_k(x_k) - a_k(\eta_k(\mathbf{x}_{-k}, \mathbf{y}))\right\}$$

CAVI for conditional exponential family models

Assume that all **complete conditionals** of $p(\mathbf{x} \mid \mathbf{y})$ belong to the exponential family,

$$p(x_k \mid \mathbf{x}_{-k}, \mathbf{y}) = h_k(x_k) \exp\left\{\eta_k(\mathbf{x}_{-k}, \mathbf{y})^\mathsf{T} t_k(x_k) - a_k(\eta_k(\mathbf{x}_{-k}, \mathbf{y}))\right\}$$

The coordinate updates in the CAVI algorithm,

$$q_k^*(x_k) \propto \exp\left(\mathbb{E}_{q_{-k}}\left[\log p(x_k \mid \mathbf{x}_{-k}, \mathbf{y})\right]\right)$$

results in $q_k^*(x_k) = q_k^*(x_k; \eta_k)$ being of **the same parametric form** as $p(x_k \mid \mathbf{x}_{-k}, \mathbf{y})$, with natural parameters $\eta_k = \mathbb{E}_{q_{-k}}[\eta_k(\mathbf{x}_{-k}, \mathbf{y})]$.

ex) CAVI for LDA

For LDA we have derived the complete conditionals (Gibbs sampler):

- $p(\theta_d \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}) = \mathrm{Dir}(\theta_d \mid \alpha + c_d)$
- $p(\beta_k \mid \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) = \mathrm{Dir}(\beta_k \mid \eta + \tilde{c}_k)$
- $p(z_{d,n} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{w}) = \mathrm{Cat}(z_{d,n} \mid \left\{\theta_{d,k}\beta_{k,w_{d,n}}\right\}_{k=1}^{K})$

ex) CAVI for LDA

For LDA we have derived the complete conditionals (Gibbs sampler):

- $p(\theta_d \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}) = \mathrm{Dir}(\theta_d \mid \alpha + c_d)$
- $p(\beta_k \mid \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) = \mathrm{Dir}(\beta_k \mid \eta + \tilde{c}_k)$
- $p(z_{d,n} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{w}) = \mathrm{Cat}(z_{d,n} \mid \{\theta_{d,k}\beta_{k,w_{d,n}}\}_{k=1}^{K})$

Both the **Dirichlet** and the **categorical** distributions **belong to the exponential family**.

The mean field approximation becomes,

$$
\begin{aligned}
q(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}) &= \prod_{k=1}^{K} q(\beta_k) \prod_{d=1}^{D} \left[q(\theta_d) \prod_{n=1}^{N_d} q(z_{d,n}) \right] \\
&= \prod_{k=1}^{K} \mathrm{Dir}(\beta_k \mid \lambda_k) \prod_{d=1}^{D} \left[\mathrm{Dir}(\theta_d \mid \gamma_d) \prod_{n=1}^{N_d} \mathrm{Cat}(z_{d,n} \mid \phi_{d,n}) \right]
\end{aligned}
$$

ex) CAVI for LDA

The **CAVI algorithm** for the **LDA model** becomes:

Initialize $\boldsymbol{\lambda}$, γ arbitrarily

while the ELBO has not converged **do**
    for each document d **do**
        for each word n **do**
            Update the categorical probabilities $\phi_{d,n}$.
        end for
        Update the Dirichlet parameters γ_d.
    end for
    for each topic k **do**
        Update the Dirichlet parameters λ_k.
    end for
end while

Stochastic Variational Inference

Limitations of CAVI

The CAVI algorithm is **often very efficient**, but suffers from some limitations:

- Based on $\mathrm{KL}(q\|p)$ minimization, leading to a "mode seeking".

Limitations of CAVI

The CAVI algorithm is **often very efficient**, but suffers from some limitations:

- Based on $\mathrm{KL}(q\|p)$ minimization, leading to a "mode seeking".
- Based on the mean field factorization, which limits its expressivity.

Limitations of CAVI

The CAVI algorithm is **often very efficient**, but suffers from some limitations:

- Based on $\mathrm{KL}(q\|p)$ minimization, leading to a "mode seeking".
- Based on the mean field factorization, which limits its expressivity.
- Based on coordinate ascent optimization, which means that innovations in gradient-based-optimization are not readily applicable.

Limitations of CAVI

The CAVI algorithm is **often very efficient**, but suffers from some limitations:

- Based on $\mathrm{KL}(q\|p)$ minimization, leading to a "mode seeking".
- Based on the mean field factorization, which limits its expressivity.
- Based on coordinate ascent optimization, which means that innovations in gradient-based-optimization are not readily applicable.

Active research on addressing each one of these limitations.

Limitations of CAVI

The CAVI algorithm is **often very efficient**, but suffers from some limitations:

- Based on $\mathrm{KL}(q\|p)$ minimization, leading to a "mode seeking".
- Based on the mean field factorization, which limits its expressivity.
- Based on coordinate ascent optimization, which means that innovations in gradient-based-optimization are not readily applicable.

Active research on addressing each one of these limitations.

We will focus on the last point, and in particular...

...how we can scale VI to **massive data** by using (subsampling-based) **stochastic gradient optimization.**

CAVI for large document collections

The CAVI algorithm for LDA involves,

2: **for** each document d **do**

$\vdots$

3: **end for**

8: **for** each topic k **do**

$\vdots$

9: **end for**

We need to loop over **all documents** for each update of the **global topic parameters**. For large document collections this is very wasteful.

Global and local parameters

Note that, in the LDA model, we have:

- **Global** topic variables $\boldsymbol{\beta}$, with variational parameters $\boldsymbol{\lambda}$.
- **Local, per document,** variables θ_d and $z_{d,n}$ with variational parameters γ_d and $\phi_{d,n}$, respectively.

Global and local parameters

Note that, in the LDA model, we have:

- **Global** topic variables $\boldsymbol{\beta}$, with variational parameters $\boldsymbol{\lambda}$.
- **Local, per document,** variables θ_d and $z_{d,n}$ with variational parameters γ_d and $\phi_{d,n}$, respectively.

Idea: To improve efficiency over CAVI, can we optimize the ELBO in the following way?

1. Subsample a **mini-batch of documents**, and update the local variational parameters for these documents.
2. Update the global variational parameters based on the mini-batch

Global and local parameters

Note that, in the LDA model, we have:

- **Global** topic variables $\boldsymbol{\beta}$, with variational parameters $\boldsymbol{\lambda}$.
- **Local, per document,** variables θ_d and $z_{d,n}$ with variational parameters γ_d and $\phi_{d,n}$, respectively.

Idea: To improve efficiency over CAVI, can we optimize the ELBO in the following way?

1. Subsample a **mini-batch of documents**, and update the local variational parameters for these documents.
2. Update the global variational parameters based on the mini-batch

Yes! But we need to switch to stochastic gradient optimization.

Stochastic gradient ascent

Stochastic gradient methods are key enablers of large scale machine learning.

The problem: $\text{maximize}_\eta\, f(\eta)$

The setting:

- The gradients $g(\eta) \stackrel{\text{def}}{=} \nabla_\eta f(\eta)$ is expensive (or intractable) to compute, but...

Stochastic gradient ascent

Stochastic gradient methods are key enablers of large scale machine learning.

The problem: $\text{maximize}_\eta \, f(\eta)$

The setting:

- The gradients $g(\eta) \stackrel{\text{def}}{=} \nabla_\eta f(\eta)$ is expensive (or intractable) to compute, but...
- ... $\hat{g}(\eta)$, a stochastic **unbiased** estimate of $g(\eta)$, is readily available.

Stochastic gradient ascent

Stochastic gradient methods are key enablers of large scale machine learning.

The problem: $\text{maximize}_\eta\, f(\eta)$

The setting:

- The gradients $g(\eta) \overset{\text{def}}{=} \nabla_\eta f(\eta)$ is expensive (or intractable) to compute, but...
- ... $\hat{g}(\eta)$, a stochastic **unbiased** estimate of $g(\eta)$, is readily available.

The solution: The iterates

$$\eta_{\tau+1} = \eta_\tau + \epsilon_\tau \hat{g}(\eta_\tau)$$

with $\sum_{\tau=1}^{\infty} \epsilon_\tau = \infty$ and $\sum_{\tau=1}^{\infty} \epsilon_\tau^2 < \infty$ **converges to a maxima** of f.

Reduced formulation

Mean field VI for LDA:

$$\underset{\boldsymbol{\lambda},\gamma,\phi}{\text{maximize}} \ \text{ELBO}(\boldsymbol{\lambda}, \gamma, \phi)$$

Reduced formulation

Mean field VI for LDA:

$$\underset{\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi}}{\text{maximize}}\ \text{ELBO}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi}) = \underset{\boldsymbol{\lambda}}{\text{maximize}}\ \underset{\boldsymbol{\gamma},\boldsymbol{\phi}}{\text{maximize}}\ \text{ELBO}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})$$

Assume that, for any $\boldsymbol{\lambda}$, explicit solutions exist for the local variables:

$$\begin{cases} \boldsymbol{\gamma}^* = \gamma(\boldsymbol{\lambda}) \\ \boldsymbol{\phi}^* = \phi(\boldsymbol{\lambda}) \end{cases}$$

Reduced formulation

Mean field VI for LDA:

$$\underset{\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi}}{\text{maximize}}\ \text{ELBO}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi}) = \underset{\boldsymbol{\lambda}}{\text{maximize}}\ \underset{\boldsymbol{\gamma},\boldsymbol{\phi}}{\text{maximize}}\ \text{ELBO}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})$$

Assume that, for any $\boldsymbol{\lambda}$, explicit solutions exist for the local variables:

$$\begin{cases} \boldsymbol{\gamma}^* = \gamma(\boldsymbol{\lambda}) \\ \boldsymbol{\phi}^* = \phi(\boldsymbol{\lambda}) \end{cases}$$

Equivalent formulation:

$$\underset{\boldsymbol{\lambda}}{\text{maximize}}\ \underbrace{\text{ELBO}(\boldsymbol{\lambda}, \gamma(\boldsymbol{\lambda}), \phi(\boldsymbol{\lambda}))}_{\overset{\text{def}}{=} \text{ELBO}(\boldsymbol{\lambda})}.$$

The natural gradient of the ELBO

Recall: $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^K$, i.e., we have one global parameter per topic.

It can be shown that

$$\nabla_{\lambda_k} \text{ELBO}(\boldsymbol{\lambda}) = \nabla^2_{\lambda_k} a(\lambda_k) \left(\mathbb{E}_q[\tilde{c}_k] + \eta - \lambda_k\right)$$

where $\nabla^2_{\lambda_k} a(\lambda_k)$ is the Fisher information matrix of $q(\beta_k; \lambda_k)$

The natural gradient of the ELBO

Recall: $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^K$, i.e., we have one global parameter per topic.

It can be shown that

$$\nabla_{\lambda_k} \text{ELBO}(\boldsymbol{\lambda}) = \nabla^2_{\lambda_k} a(\lambda_k) \left(\mathbb{E}_q[\tilde{c}_k] + \eta - \lambda_k\right)$$

where $\nabla^2_{\lambda_k} a(\lambda_k)$ is the Fisher information matrix of $q(\beta_k; \lambda_k)$

Pre-multiplying by the inverse Fisher matrix gives the **natural gradient**,

$$g_k(\lambda_k) = \mathbb{E}_q[\tilde{c}_k] + \eta - \lambda_k$$

The natural gradient of the ELBO

Recall: $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^K$, i.e., we have one global parameter per topic.

It can be shown that

$$\nabla_{\lambda_k} \text{ELBO}(\boldsymbol{\lambda}) = \nabla^2_{\lambda_k} a(\lambda_k) \left(\mathbb{E}_q[\tilde{c}_k] + \eta - \lambda_k\right)$$

where $\nabla^2_{\lambda_k} a(\lambda_k)$ is the Fisher information matrix of $q(\beta_k; \lambda_k)$

Pre-multiplying by the inverse Fisher matrix gives the **natural gradient**,

$$g_k(\lambda_k) = \mathbb{E}_q[\tilde{c}_k] + \eta - \lambda_k$$

Why natural gradient?

- Better adapted to the geometry of probability distributions.
- Simpler expression, **enabling data subsampling**

M. Hoffman, D. Blei, C. Wang, and J. Paisley. **Stochastic Variational Inference.** *JMLR*, 24:1303–1347, 2013.

Unbiased estimate of natural gradient

Writing out the vth element of $\mathbb{E}_q[\tilde{c}_k]$,

$$\mathbb{E}_q[\tilde{c}_{k,v}] = \sum_d \sum_n \mathbb{1}\{w_{d,n} = v\}\phi^k_{d,n}$$

it follows that we can compute a **cheap** and **unbiased** approximation of the natural gradient:

1. Pick a document d uniformly at random.
2. Compute

$$\hat{g}_k(\lambda_k) \stackrel{\text{def}}{=} D\hat{c}_k + \eta - \lambda_k, \quad \text{where}$$
$$\hat{c}_{k,v} = \sum_n \mathbb{1}\{w_{d,n} = v\}\phi^k_{d,n}.$$

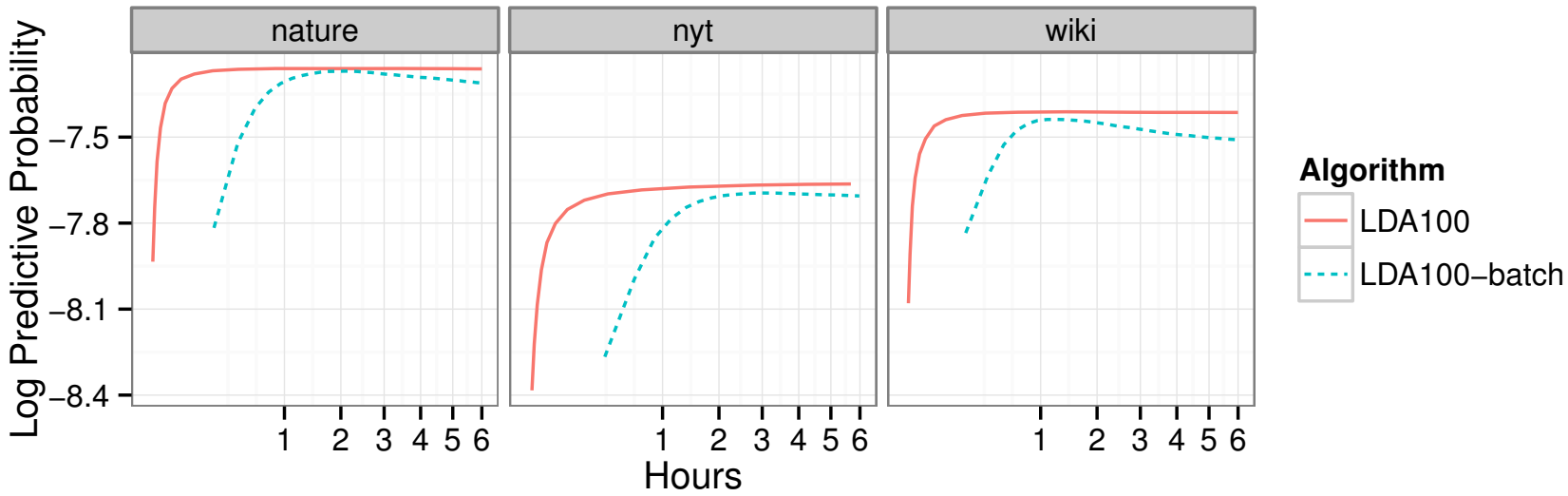
Stochastic variational inference for LDA

The **SVI algorithm** for the **LDA model** becomes:

Initialize $\boldsymbol{\lambda}$, γ arbitrarily.

while convergence criterion is not met **do**
  Sample a document d from the collection uniformly at random.
  while local parameters have not converged **do**
    for each word n **do**
      Update the categorical probabilities $\phi_{d,n}$.
    end for
    Update the Dirichlet parameters γ_d.
  end while
  for each topic k **do**
    Update the Dirichlet parameters $\lambda_k \leftarrow \lambda_k + \epsilon_\tau \hat{g}_k(\lambda_k)$.
  end for
end while

Numerical example, SVI vs CAVI



Results on three data sets: *Nature* (350k document, 58M words), *New York Times* (1.8M documents, 461M words), *Wikipedia* (3.8M documents, 482M words).

Borrowed from:

M. Hoffman, D. Blei, C. Wang, and J. Paisley. **Stochastic Variational Inference.** *JMLR*, 24:1303–1347, 2013.

Wrapping up

Types of approximation

Two types of approximate Bayesian inference methods:

1. Parametric ("deterministic") approximations:

- Expectation propagation, approximate message passing
- Variational inference
- Laplace approximations, INLA
- ...

2. Non-parametric, sampling-based approximations:

- Markov Chain Monte Carlo
- Sequential Monte Carlo
- Piecewise-deterministic Markov processes

Parametric methods

For **parametric methods** the approximation depend on:

- the approximating family of distributions $\mathcal{Q}$
- the loss function that is optimized to fit $q \approx \pi$
- the optimization procedure used
- ...

EP vs VI: Selecting $\mathcal{Q}$

Expectation propagation:

- We let $q(\mathbf{x})$ factorize in the same way as $\pi(\mathbf{x})$.
- For graphical models, common to also assume that $q(\mathbf{x})$ factorizes over the components of $\mathbf{x}$

$\Rightarrow$ we obtain an **approximate message passing** algorithm.

EP vs VI: Selecting $\mathcal{Q}$

Expectation propagation:

- We let $q(\mathbf{x})$ factorize in the same way as $\pi(\mathbf{x})$.
- For graphical models, common to also assume that $q(\mathbf{x})$ factorizes over the components of $\mathbf{x}$

$\Rightarrow$ we obtain an **approximate message passing** algorithm.

Variational inference:

- Classical choice: **mean field** factorization
- Alternative: use **flexible transform-based** distributions, $q : \mathbf{z} \sim p(\mathbf{z})$, $\mathbf{x} = f_\eta(\mathbf{z})$.

EP vs VI: The loss function

Expectation propagation:

- In each iteration i we minimize $\mathrm{KL}(f_i(\mathbf{x})q^{(-i)}(\mathbf{x})\|q(\mathbf{x}))$.
- Support-covering behavior (*zero avoiding*)

EP vs VI: The loss function

Expectation propagation:

- In each iteration i we minimize $\mathrm{KL}(f_i(\mathbf{x})q^{(-i)}(\mathbf{x})\|q(\mathbf{x}))$.
- Support-covering behavior (*zero avoiding*)

Variational inference:

- We minimize $\mathrm{KL}(q\|\pi)$.
- Mode-seeking behavior (*zero forcing*)

  – underestimating posterior variance
- Expectation is w.r.t. approximation, $\mathrm{KL}(q\|p) = \mathbb{E}_q\left[\log \frac{q}{p}\right]$!
- Recent developments: generalize VI to other divergences.

C. A. Naesseth, F. Lindsten, and D. M. Blei. **Markovian Score Climbing: Variational Inference with KL($p\|q$).** *Advances in Neural Information Processing Systems 33*, 2020.

Sampling-based approximations

Sampling-based approximations such as **Markov chain Monte Carlo** are (typically) **consistent**.

Sampling-based approximations

Sampling-based approximations such as **Markov chain Monte Carlo** are (typically) **consistent**.

- ▲ **Arbitrary accuracy** of approximation, if given enough compute time.
- ▼ "Enough" can be prohibitively long.
- ▼ Scalability to large data more difficult than for optimization-based methods

Sampling-based approximations

Sampling-based approximations such as **Markov chain Monte Carlo** are (typically) **consistent**.

- ▲ **Arbitrary accuracy** of approximation, if given enough compute time.
- ▼ "Enough" can be prohibitively long.
- ▼ Scalability to large data more difficult than for optimization-based methods

Extensions:

- Improve scalability (e.g., by data subsampling)
- Combine deterministic and sampling-based approximations – **can we get the best of both worlds?**

F. Lindsten, J. Helske, and M. Vihola. **Graphical model inference: Sequential Monte Carlo meets deterministic approximations.** *Advances in Neural Information Processing Systems 31*, 2018.