STOCHASTIC VARIATIONAL INFERENCE

based on the paper by Hoffman, Blei, Wang and Paisley

Gregor Gebhardt[1] & Tobias Plötz[2]

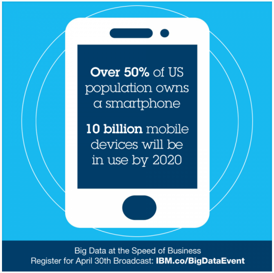
June 17th, 2015

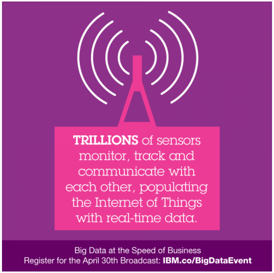
[1] Computational Learning for Autonomous Systems (CLAS) and Intelligent Autonomous Systems (IAS)
[2] Visual Inference Group

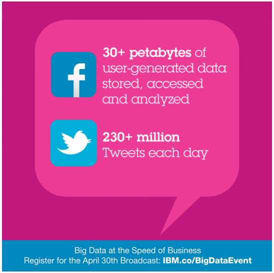
World of Big Data

http://www.strongautomotive.com/

WORLD OF BIG DATA







http://www.ibmbigdatahub.com/

Analyzing Big Data

Nowadays, huge amounts of data are available.

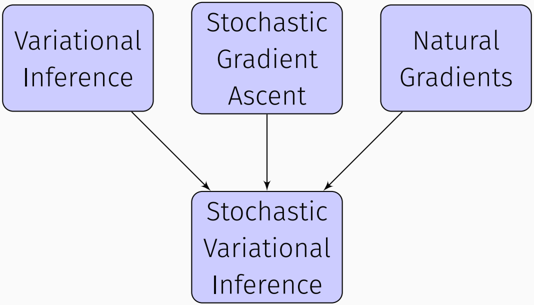
We want to analyze them using Bayesian methods.

Variational inference is a powerful method for running inference in complex probabilistic models, but it does not scale to large datasets.

Analyzing Big Data

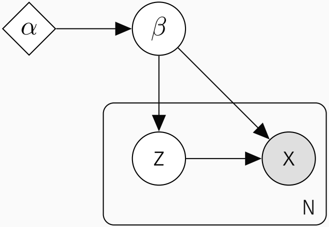
Stochastic Variational Inference helps to make Variational Inference scale to large datasets.

It makes use of three concepts



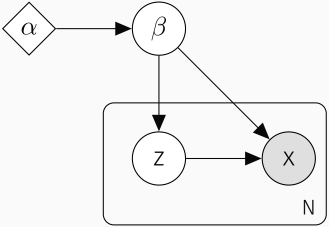
The high level story

We consider following model class:



- N observations $x = x_{1:N}$
- N local hidden variables $z = z_{1:N}$
- global hidden variables β
- fixed hyper-parameters α

The high level story



Our goal is to calculate the posterior of the hidden variables

$$p(\beta, z|x) \quad \text{This is intractable ☹}$$

Meanfield Variational Inference allows to find approximation

$$p(\beta, z|x) \approx q(\beta, z) = q(\beta|\lambda)q(z|\phi)$$

The high level story

$$p(\beta, z|x) \approx q(\beta, z) = q(\beta|\lambda)q(z|\phi)$$

This results in optimization problem over λ and ϕ

Usually solved with a coordinate-ascent algorithm:

1. Update λ leaving $q(z|\phi)$ fixed
2. Update ϕ leaving $q(\beta|\lambda)$ fixed
3. Repeat until convergence

The high level story

Update of q($\beta|\lambda$) involves all data points x $\Rightarrow$ **not scalable**

Stochastic gradient ascent: follow a noisy, but unbiased estimate of the gradient instead of exact gradient.

Noisy gradient shall be obtained by using small sample of all data points to solve scalability issue.

However, the (noisy) gradient requires complex computations ☹

The **natural gradient** is an alternative, more “sensible” gradient.

In this case, it is also easier to compute.

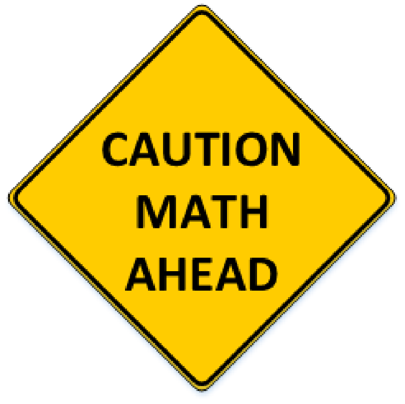
The high level story

Stochastic Variational Inference

=

Variational inference with stochastic updates of global parameter λ along natural gradient.

Before we proceed ...

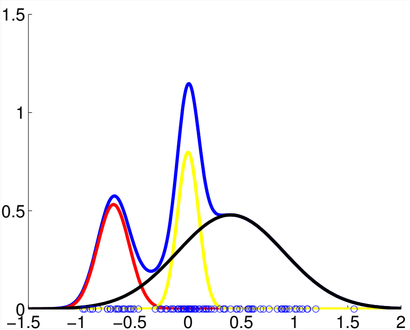


VARIATIONAL INFERENCE

Motivation - Gaussian Mixture Models

Simple example

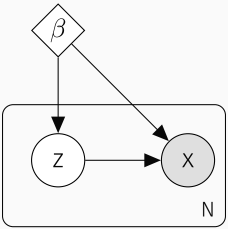
- We are given some data points $x = \{x_1, ..., x_n\}$
- We want to fit a Gaussian mixture model to this data
- $p(x|\beta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)$



Maximum likelihood learning

How can we learn model parameters $\beta = \{\mu_k, \Sigma_k\}_{k=1}^{K}$?

Rewrite $p(x|\beta) = \sum_z p(x, z|\beta) = \prod_{i=1}^{n} \sum_{z_i} \prod_{k=1}^{K} \pi_k^{z_{ik}} \mathcal{N}(x_i|\mu_k, \Sigma_k)^{z_{ik}}$



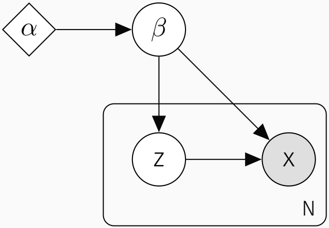
Now use Expectation Maximization to learn
$\beta^* = \arg\max_\beta \sum_z p(x, z; \beta)$

Variational Inference

Shortcomings of maximum likelihood learning?

- No assessment of uncertainty in estimate β^*
- How many mixture components do we need?

The Bayesian alternative: Treat β as random variable!



Variational inference

Instead of learning, compute posterior over latent variables given observed data

$$p(\beta, z|x) = \frac{p(\beta, z, x)}{\sum_{\beta,z} p(\beta, z, x)}$$

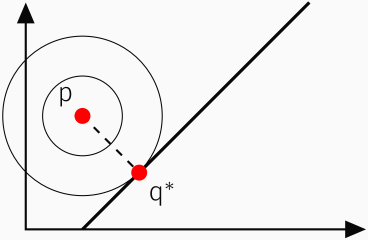
Computing $p(\beta, z|x)$ is usually intractable because of normalization term.

What can we do now?

- Sampling
- Approximate Variational Inference

Variational Inference

- We have intractable posterior $p(\beta, z|x)$
- Choose family of tractable distributions $q(\beta, z) \in Q$
- Choose some kind of distance measure Δ
- Find $q^* = \arg\min_{q \in Q} \Delta(q, p)$



We call Q the family of **Variational Distributions**.

Depending on choice of Q and Δ, we get different algorithms

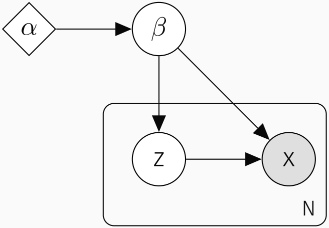
- Q product of exponential family distributions, $\Delta = \mathrm{KL}(p||q)$ leads to **Expectation Propagation**[1]
- $Q = q(\beta)\prod q(z_i)$, $\Delta = \mathrm{KL}(q||p)$ leads to **mean-field**[2]

Stochastic Variational Inference works in the context of mean-field inference in a special model class.

[1] Minka, "Expectation Propagation for Approximate Bayesian Inference".
[2] Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics).

CONJUGATE EXPONENTIAL MODELS

Model Structure



- N observations $x = x_{1:N}$
- N local hidden variables $z = z_{1:N}$, where each $z_n = z_{n,1:J}$
- global hidden variables β
- fixed hyper-parameters α

Conjugate Exponential Models

The joint distribution factorizes into a global term and a product of local terms

$$p(x, z, \beta|\alpha) = p(\beta|\alpha) \prod_{n+1}^{N} p(x_n, z_n|\beta)$$

The **complete conditionals** have to be in the exponential family

$$p(\beta|x, z, \alpha) = h(\beta) \exp\{\eta_g(x, z, \alpha)^\intercal t(\beta) - a_g(\eta_g(x, z, \alpha))\}$$

$$p(z_{nj}|x_n, z_{n,-j}, \beta) = h(z_{nj}) \exp\{\eta_l(x_n, z_{n,-j}, \beta)^\intercal t(z_{nj}) - a_l(\eta_l(x_n, z_{n,-j}, \beta))\}$$

base measure $h(\cdot)$

natural parameter $\eta(\cdot)$

log-normalizer $a(\cdot)$

sufficient statistics $t(\cdot)$

Conjugate Exponential Models

The assumptions on the complete conditionals

$$p(\beta|x, z, \alpha) = h(\beta) \exp\{\eta_g(x, z, \alpha)^\intercal t(\beta) - a_g(\eta_g(x, z, \alpha))\}$$
$$p(z_{nj}|x_n, z_{n,-j}, \beta) = h(z_{nj}) \exp\{\eta_l(x_n, z_{n,-j}, \beta)^\intercal t(z_{nj}) - a_l(\eta_l(x_n, z_{n,-j}, \beta))\}$$

imply an **exponential family local context**

$$p(x_n, z_n|\beta) = h(x_n, z_n) \exp\{\beta^\intercal t(x_n, z_n) - a_l(\beta)\},$$

and a **conjugate exponential family prior** on the global parameters

$$p(\beta|\alpha) = h(\beta) \exp\{\alpha^\intercal t(\beta) - a_g(\alpha)\}$$

Conjugate Exponential Models

The **conjugate exponential model** with prior

$$p(\beta|\alpha) = h(\beta)\exp\{\alpha^\intercal t(\beta) - a_g(\alpha)\}$$

implies following form of the sufficient statistics and natural parameters

$$t(\beta) = [\beta, -a_l(\beta)]$$
$$\alpha = [\alpha_1, \alpha_2]$$

The natural parameters of the posterior are given by

$$\eta_g(x, z, \alpha) = [\alpha_1 + \sum_{n=1}^{N} t(z_n, x_n), \alpha_2 + N]$$

MEAN-FIELD VARIATIONAL INFERENCE

Mean-Field Variational Inference

- Maximize a lower bound on the logarithm of the **marginal probability of the observations** $\log p(x)$.
  - Equivalent to minimizing the KL divergence from the variational distribution to the posterior.
- Assume that each hidden variable is **independent** and governed by its **own parameter**.
- The variational distributions are in the **same family** as the complete conditional distributions.
  - As a result of the conjugacy of complete conditionals and prior distributions.

Evidence Lower Bound (ELBO)

Lower bound on the logarithm of the marginal probability of the observations

$$
\begin{aligned}
\log p(x) &= \log \int p(x, z, \beta) dz d\beta \\
&= \log \int p(x, z, \beta) \frac{q(z, \beta)}{q(z, \beta)} dz d\beta \\
&= \log \left(\mathbb{E}_q \left[\frac{p(x, z, \beta)}{q(z, \beta)} \right] \right) \\
&\geq \mathbb{E}_q \left[\log \frac{p(x, z, \beta)}{q(z, \beta)} \right] \\
&= \mathbb{E}_q \left[\log p(x, z, \beta) \right] - \mathbb{E}_q \left[\log q(z, \beta) \right] \\
&=: \mathcal{L}(q)
\end{aligned}
\tag{1}
$$

[1] Jensen's inequality

Evidence Lower Bound (ELBO)

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

The ELBO consists of two terms:

- the **expected log joint likelihood** $\mathbb{E}_q[\log p(x, z, \beta)]$,
- the **entropy of the variational distribution** $-\mathbb{E}_q[\log q(z, \beta)]$.

Maximizing the ELBO is equivalent to finding the member of the exponential family that is closest (in terms of KL) to the true posterior

$$\begin{aligned} KL(q(z, \beta)||p(z, \beta|x)) &= \mathbb{E}_q[\log q(z, \beta)] - \mathbb{E}_q[\log p(z, \beta|x)] \\ &= \mathbb{E}_q[\log q(z, \beta)] - \mathbb{E}_q[\log p(x, z, \beta)] + \log p(x) \\ &= -\mathcal{L}(q) + \text{const.} \end{aligned}$$

The Mean-Field Variational Family

Each hidden variable is independent and governed by its own parameter

$$q(z, \beta) = q(\beta|\lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj}|\phi_{nj}).$$

Thus, the entropy term decomposes into

$$-\mathbb{E}_q [\log q(z, \beta)] = -\mathbb{E}_\lambda [\log q(\beta)] - \sum_{n+1}^{N} \sum_{j=1}^{J} \mathbb{E}_{\phi_{nj}} \left[\log q(z_{nj})\right].$$

The expected log joint likelihood can be separated by applying the chain rule

$$\mathbb{E}_q [\log p(x, z, \beta)] = \mathbb{E}_q [\log p(x, z)] + \mathbb{E}_q [\log p(\beta|x, z)].$$

The Mean-Field Variational Family

The variational distributions are in the same exponential family as the complete conditional distributions

$$q(\beta|\lambda) = h(\beta)\exp\{\lambda^\intercal t(\beta) - a_g(\lambda)\},$$
$$q(z_{nj}|\phi_{nj}) = h(z_{nj})\exp\{\phi_{nj}^\intercal t(z_{nj}) - a_l(\phi_{nj})\}$$

the gradients of the ELBO w.r.t the parameters λ and ϕ_{nj} can be obtained as

$$\nabla_\lambda \mathcal{L} = \nabla_\lambda^2 a_g(\lambda)(\mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda)$$
$$\nabla_{\phi_{nj}} \mathcal{L} = \nabla_{\phi_{nj}}^2 a_l(\phi_{nj})(\mathbb{E}_q[\eta_l(x_n, z_{n,-j}, \beta)] - \phi_{nj})$$

This leads to **coordinate ascent variational inference**

$$\lambda = \mathbb{E}_q\left[\eta_g(x, z, \alpha)\right]$$
$$\phi_{nj} = \mathbb{E}_q\left[\eta_l(x_n, z_{n,-j}, \beta)\right]$$

How does it work in the end?

```
Initialize λ^(0) randomly.
repeat
    for all local variational parameters φ_nj do
        Update φ_nj with φ_nj^(t) = E_{q^(t-1)}[η_{l,j}(x_n, z_{n,-j}, β)].
    end for
    Update the global variational parameters,
    λ^(t) = E_{q^(t)}[η_g(z_{1:N}, x_{1:N}, α)]
until the ELBO converges.
```

From Mean-Field to Stochastic Variational Inference

- The global parameter λ is initialized randomly
  - all local updates are based on this initial random guess.
  - one could already learn about the structure of the data from a subset.
- Better: update the global parameters after each local update using **stochastic optimization**.
  - Sample one data point from the data set.
  - Compute the optimal local variational parameters.
  - Form intermediate global parameters
    - by repeating the sampled data point occured N times
    - and performing classical coordinate ascent
  - Set the global parameters to a weighted average of the old estimate and the intermediate global parameters.
- The Gradient is based on a Euclidean metric on the parameters.
  - The **natural gradient** accounts for the information geometry in the parameter space.

NATURAL GRADIENT

Classical Gradient Ascent

Maximize a function by taking small steps in the direction of the gradient

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho \nabla_{\lambda} f(\lambda^{(t)})$$

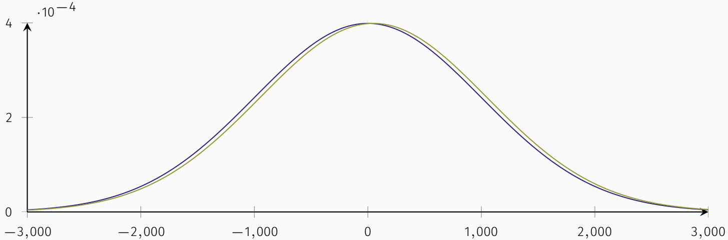
The classical gradient points in the direction of the steepest ascent constrained by the **Euclidean metric** in the **parameter space**.

$$\nabla_{\lambda} = \arg\max_{d\lambda} f(\lambda + d\lambda) \qquad \text{s.t.} \quad ||d\lambda||^2 < \epsilon^2 \quad \text{with } \epsilon \to 0$$

This might not be the best option for probability distributions...

Classical Gradient Ascent

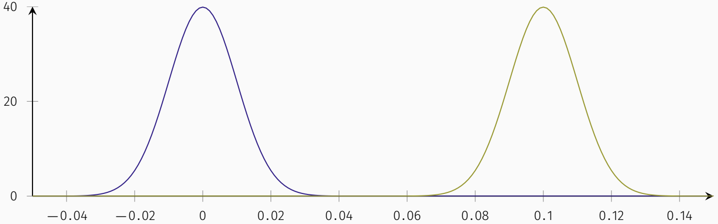
Consider the two Gaussian distributions $\mathcal{N}(0, 1000)$ and $\mathcal{N}(50, 1000)$.



The distributions are nearly identical, but the Euclidean distance between the parameter vectors is 50.

Classical Gradient Ascent

Now consider the two Gaussians $\mathcal{N}(0, 0.01)$ and $\mathcal{N}(0.1, 0.01)$.



The distributions barely overlap, however, the Euclidean distance of their parameter vector is only 0.1.

Natural Gradient Ascent

Apply a different measure: the symmetrized Kullback-Leibler divergence

$$D_{KL}^{sym}(\lambda, \lambda') = \mathbb{E}_\lambda \left[\log \frac{q(\beta|\lambda)}{q(\beta|\lambda')} \right] + \mathbb{E}_{\lambda'} \left[\log \frac{q(\beta|\lambda')}{q(\beta|\lambda)} \right]$$

We want a Riemannian metric $G(\lambda)$ that transforms the squared Euclidean distance to the symmetric KL divergence

$$d\lambda^\intercal G(\lambda) d\lambda = D_{KL}^{sym}(\lambda, \lambda + d\lambda)$$

The natural gradient is the the gradient premultiplied by the inverse Riemannian metric[3]

$$\hat{\nabla}_\lambda f(\lambda) = G(\lambda)^{-1} \nabla_\lambda f(\lambda)$$

[3]Amari, "Natural gradient works efficiently in learning".

Natural Gradient Ascent

We can derive the matrix $G(\lambda)$ by plugging the first-order Taylor approximations

$$\log q(\beta|\lambda + d\lambda) = \log q(\beta|\lambda) + d\lambda^{\intercal}\nabla_{\lambda} \log q(\beta|\lambda) + O(d\lambda^2)$$
$$q(\beta|\lambda + d\lambda) = q(\beta|\lambda) + q(\beta|\lambda)d\lambda^{\intercal}\nabla_{\lambda} \log q(\beta|\lambda) + O(d\lambda^2)$$

into the symmetric Kulback-Leibler divergence ($\lambda' = \lambda + d\lambda$)

$$\begin{aligned} D_{KL}^{sym}(\lambda, \lambda') &= \mathbb{E}_{\lambda}\left[\log \frac{q(\beta|\lambda)}{q(\beta|\lambda')}\right] + \mathbb{E}_{\lambda'}\left[\log \frac{q(\beta|\lambda')}{q(\beta|\lambda)}\right] \\ &= \int_{\beta} [q(\beta|\lambda') - q(\beta|\lambda)] [\log q(\beta|\lambda') - \log q(\beta|\lambda)] \, d\beta \\ &= \int_{\beta} \left[q(\beta|\lambda)d\lambda^{\intercal}\nabla_{\lambda} \log q(\beta|\lambda) + O(d\lambda^2)\right] \\ &\qquad \left[d\lambda^{\intercal}\nabla_{\lambda} \log q(\beta|\lambda) + O(d\lambda^2)\right] d\beta \end{aligned}$$

Natural Gradient Ascent

$$
\begin{aligned}
D_{KL}^{sym}(\lambda, \lambda') &= \int_\beta \left[q(\beta|\lambda) d\lambda^\intercal \nabla_\lambda \log q(\beta|\lambda) + O(d\lambda^2)\right] \\
&\qquad\qquad \left[d\lambda^\intercal \nabla_\lambda \log q(\beta|\lambda) + O(d\lambda^2)\right] d\beta \\
&= O(d\lambda^3) + \int_\beta q(\beta|\lambda) \left[d\lambda^\intercal \nabla_\lambda \log q(\beta|\lambda)\right]^2 d\beta \\
&\approx \mathbb{E}_\lambda \left[(d\lambda^\intercal \nabla_\lambda \log q(\beta|\lambda))^2 \right] = d\lambda^\intercal \mathbb{E}_\lambda \left[(\nabla_\lambda \log q(\beta|\lambda))^2 \right] d\lambda
\end{aligned}
$$

$G(\lambda) = \mathbb{E}_\lambda \left[(\nabla_\lambda \log q(\beta|\lambda))^2 \right]$ is the **Fisher information matrix**.

For the exponential family, the Fisher information matrix is the **second derivative of the *log-normalizer*** $\nabla_\lambda^2 a(\lambda)$.

STOCHASTIC GRADIENT ASCENT

Stochastic Gradient Ascent

Follow noisy estimates of the gradient.

- Noisy estimates are often cheaper to compute.
- Allow to escape from shallow local optima.

Assuming we have

- an objective function $f(\lambda)$ and
- a random function $B(\lambda)$, where $\mathbb{E}_q[B(\lambda)] = \nabla_\lambda f(\lambda)$,

we update the parameters by

$$\lambda^{(t)} = \lambda^{(t-1)} + \rho_t b_t\left(\lambda^{(t-1)}\right),$$

where b_t is a sample of the random function $B(\lambda^{(t)})$.

Stochastic Gradient Ascent

If the step size ρ satisfies

$$\sum \rho_t = \infty, \qquad \sum \rho_t^2 < \infty,$$

then $\lambda^{(t)}$ will converge to the optimum λ^* or a local optimum of f.[4]

The same applies if the gradient is premultiplied by a sequence of positive-definite matrices G_t^{-1}:

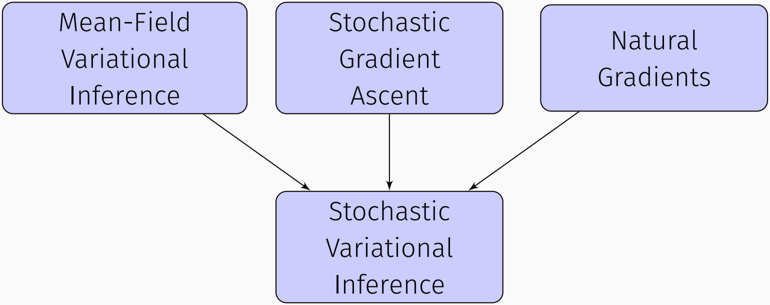
$$\lambda^{(t)} = \lambda^{(t-1)} + \rho_t G_t^{-1} b_t \left(\lambda^{(t-1)}\right)$$

E.g., the Fisher information matrix $G(\lambda)$.

[4] Robbins and Monro, "A stochastic approximation method".

STOCHASTIC VARIATIONAL INFERENCE

Stochastic Variational Inference



Stochastic Variational Inference

1. Sample a data point from the data set.
   - Optimize the local variational parameters.
2. Form intermediate global parameters.
   - Classical coordinate ascent.
3. Update the global variational parameters.
   - Weighted average of the intermediate and the old global parameters.

This algorithm is stochastic natural gradient ascent on the global variational parameters.

The Noisy Natural Gradient of ELBO

Recall the evidence lower bound (ELBO)

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

Let $\phi(\lambda)$ be a function that returns a local optimum of the local variational parameters

$$\nabla_\phi \mathcal{L}(\lambda, \phi(\lambda)) = 0$$

Locally maximized ELBO, with fixed λ and locally optimal $\phi(\lambda)$

$$\mathcal{L}(\lambda) := \mathcal{L}(\lambda, \phi(\lambda))$$

The gradient is the same as for the ELBO

$$\begin{aligned} \nabla_\lambda \mathcal{L}(\lambda) &= \nabla_\lambda \mathcal{L}(\lambda, \phi(\lambda)) + (\nabla_\lambda \phi(\lambda))^\intercal \nabla_\phi \mathcal{L}(\lambda, \phi(\lambda)) \\ &= \nabla_\lambda \mathcal{L}(\lambda, \phi(\lambda)) \end{aligned}$$

The Noisy Natural Gradient of ELBO

Recall the evidence lower bound (ELBO)

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

$\mathcal{L}(\lambda)$ can be decomposed into a global term and a local term

$$\begin{aligned}\mathcal{L}(\lambda) = \mathbb{E}_q [\log p(\beta)] - \mathbb{E}_q [\log q(\beta)] \\ + \sum_{n=1}^{N} \max_{\phi_n} (\mathbb{E}_q [\log p(x_n, z_n | \beta)] - \mathbb{E}_q [\log q(z_n)])\end{aligned}$$

Define the random function $\mathcal{L}_i(\lambda)$ of the variational parameters with a uniformly drawn index $i \sim \text{Unif}(1, \ldots, N)$

$$\begin{aligned}\mathcal{L}_i(\lambda) = \mathbb{E}_q [\log p(\beta)] - \mathbb{E}_q [\log q(\beta)] \\ + N \max_{\phi_i} (\mathbb{E}_q [\log p(x_i, z_i | \beta)] - \mathbb{E}_q [\log q(z_i)])\end{aligned}$$

$\mathbb{E}_{\text{Unif}}[\mathcal{L}(\lambda)] = \mathbb{E}_{\text{Unif}}[\mathcal{L}_i(\lambda)]$, so the noisy natural gradient is unbiased.

The Noisy Natural Gradient of ELBO

Because

$$\nabla_\lambda \mathcal{L}(\lambda, \phi(\lambda)) = \nabla_\lambda \mathcal{L}(\lambda) \quad \text{and} \quad \mathcal{L}(\lambda) = \mathcal{L}_i(\lambda),$$

supposed that the data set $\left\{x_i^{(N)}, z_i^{(N)}\right\}$ is formed by N replicates of the sampled data point (x_i, z_i), the noisy natural gradient is

$$\begin{aligned}\hat{\nabla}_\lambda \mathcal{L}_i &= G(\lambda)^{-1} \nabla^2_\lambda a_g(\lambda) \mathbb{E}_q \left[\eta_g(x_i^{(N)}, z_i^{(N)}, \alpha)\right] - \lambda \\ &= \mathbb{E}_q \left[\eta_g \left(x_i^{(N)}, z_i^{(N)}, \alpha\right)\right] - \lambda\end{aligned}$$

Exploiting the assumptions on the prior $p(\beta|\alpha)$ and the distribution of the local context $p(x_i, z_i|\beta)$

$$\eta_g \left(x_i^{(N)}, z_i^{(N)}, \alpha\right) = \alpha + N \cdot (t(x_i, z_i), 1)$$

The Noisy Natural Gradient of ELBO

The **noisy natural gradient** becomes

$$\hat{\nabla}_\lambda \mathcal{L}_i = \alpha + N \cdot \left(\mathbb{E}_{\phi_i(\lambda)}\left[t(x_i, z_i)\right], 1\right) - \lambda$$

The **intermediate global parameters** are

$$\hat{\lambda}_t = \alpha + N \cdot \left(\mathbb{E}_{\phi_i(\lambda)}\left[t(x_i, z_i)\right], 1\right)$$

The **global variational parameters** are updated as

$$\begin{aligned}\lambda^{(t)} &= \lambda^{(t-1)} + \rho_t \left(\hat{\lambda}_t - \lambda^{(t-1)}\right) \\ &= (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}_t\end{aligned}$$

Stochastic Variational Inference

Initialize $\lambda^{(0)}$ randomly.
Choose an appropriate step-size schedule ρ_t
repeat
    Sample a data point x_i uniformly from the data set.
    Compute its local variational parameter,

$$\phi_{ij} = \mathbb{E}_{\lambda^{(t-1)}} \left[\eta_{lj} \left(x_i, z_{i,-j}, \beta \right) \right]$$

    Compute intermediate global parameters,

$$\hat{\lambda}_t = \mathbb{E}_{\phi_i} \left[\eta_g \left(x_i^{(N)}, z_i^{(N)} \right) \right].$$

    Update the global variational parameters,

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

until convergence.

applications & extensions

Topic Models

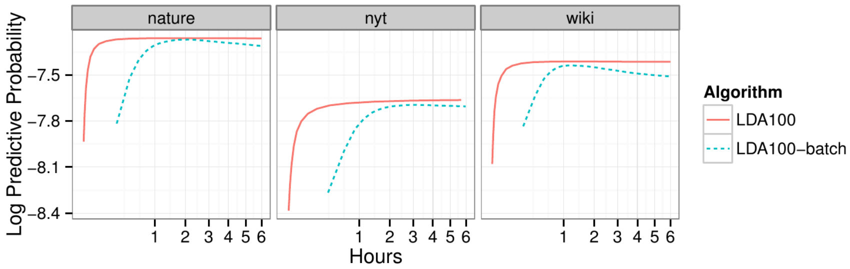
SVI applied to

- Latent Dirichlet Allocation David M. Blei, Ng, and Jordan, “Latent Dirichlet Allocation”
- Hierarchical Dirichlet Processes Teh et al., “Hierarchical Dirichlet processes”

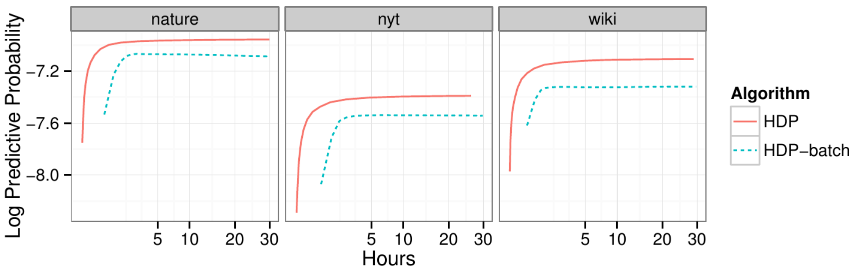
Evaluated on corpora:

	# documents	# words	vocabulary size
Nature	350k	58M	4200
New York Times	1.8M	461M	8000
Wikipedia	3.8M	482M	7700

Results for LDA



Results for HDP



Structured Stochastic Variational Inference[5]

Relax fully-factorized mean-field assumption.

$$q(\beta, z) = q(\beta|\lambda) \prod_n q(z_n|\phi_n(\beta))$$

- $q(z_n|\phi_n(\beta))$ does not need to factorize into $q(z_{n,i}|\dots)$
- $q(z_n|\phi_n(\beta))$ may depend on β

Several ways of updating $\phi_n(\beta)$ with fixed $q(\beta|\lambda)$

Update of λ still stochastic

[5] Hoffman and David M Blei, "Structured Stochastic Variational Inference".

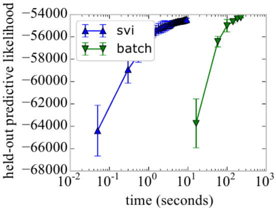
Time Series[6]

Application of SVI

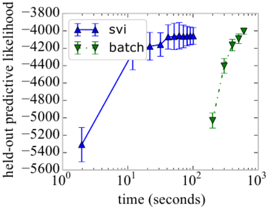
- Bayesian Hidden (Semi-)Markov Models
- Their non-parametric extensions with HDPs

Local parameters: hidden states of each observations.

Global parameters: Parameters governing state transition and observation distribution.



(a) SVI vs Batch HDP-HMM



(b) SVI vs Batch HDP-HSMM

[6] Johnson and Willsky, "Stochastic Variational Inference for Bayesian Time Series Models".