Scalable machine learning

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS424/524, SML302

Lecture 23

Scaling algorithms for fitting models to data

Last class, we discussed Monte Carlo methods and Markov chain Monte Carlo methods.

Today, we will discuss a class of method to scale ML to larger-scale models and data: Variational methods.

We will focus on *mean field variational methods*, and discuss their parallels to Gibbs sampling.

We will end by summarizing approaches to fitting models.

Monte Carlo methods in practice

Last class, we discussed Monte Carlo methods:

- Rejection sampling: reject samples from proposal distribution
- Importance sampling: weight samples from proposal distribution

and Markov chain Monte Carlo methods:

- Metropolis-Hastings: add a Markov chain to the proposal distribution
- Gibbs sampling: conditionally sample each parameter

Review: Gibbs sampling

Gibbs sampling accepts all proposal distributions

Gibbs sampling does this by selecting a proposal distribution that, by definition, produces an acceptance probability equal to one in the Metropolis-Hastings approach

Gibbs sampling is useful for sampling in a high dimensional space (as we are in with Gaussian mixture models)

Recall that Gibbs sampling requires sampling from the complete conditional distribution for each variable

Recall the problem setup

Inference in graphical models

Let $z \in \Theta$ be the set of unknown variables, including parameters and latent variables.

Let $y \in \mathcal{D}$ be data observations.

Posterior distribution: Estimate the posterior distribution $p(z \mid y)$

Parameter estimates: Find a value of z that maximizes some objective function; here, data likelihood or posterior probability

We cannot optimize directly in a computationally feasible way. Often the space is non-convex and hard to search.

MCMC: Gibbs sampling and auxiliary variables

Write $z = \{z_1, z_2, \ldots, z_q\}$ for the q-dimensional space we are sampling

For GMMs with $K = 3$, $q = 8 + n$ (means, variances, proportions, and cluster assignments)

Define $p(z_j \mid z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_q)$, or, more succinctly, $p(z_j \mid z_{-j})$.

Gibbs sampling: MCMC in high dimensions

Often, adding additional *auxiliary* variables leads to *faster mixing*.

In a mixture model, auxiliary variables are hard cluster assignments c_i.

Gibbs sampling updates have the coordinate-ascent flavor of EM, except both E-steps and M-steps are samples from a conditional distribution.

Gibbs sampling: algorithm

Gibbs sampling

- Initialize z^0
- For $s = 1 : S$
  - sample $z_1^{s+1} \sim p(z_1 \mid z_2^s, z_3^s, \ldots, z_q^s)$
  - sample $z_2^{s+1} \sim p(z_2 \mid z_1^{s+1}, z_3^s, \ldots, z_q^s)$
  - sample $z_3^{s+1} \sim p(z_3 \mid z_1^{s+1}, z_2^{s+1}, \ldots, z_q^s)$
  - sample $z_q^{s+1} \sim p(z_q \mid z_1^{s+1}, z_2^{s+1}, \ldots, z_{q-1}^{s+1})$

Gibbs sampling in practice: Gaussian mixture models

Gibbs sampling on a Gaussian mixture model:

- initialize: assign each sample to one of K mixture components in c
- iterate S times:
  - for each parameter z_j, draw a sample from $p(z_j \mid z_{-j}, c^s, \mathbf{y})$.
  - given new parameter values, draw a sample of c_i

How do we compute these conditional probabilities?

Gibbs sampling in practice: Gaussian mixture models

Gibbs sampling on a Gaussian mixture model

- assign each sample to one of K mixture components in c
- iterate S times:
  - for each parameter z_j, draw a sample from $p(z_j \mid z_{-j}, c^s, \mathbf{y})$:

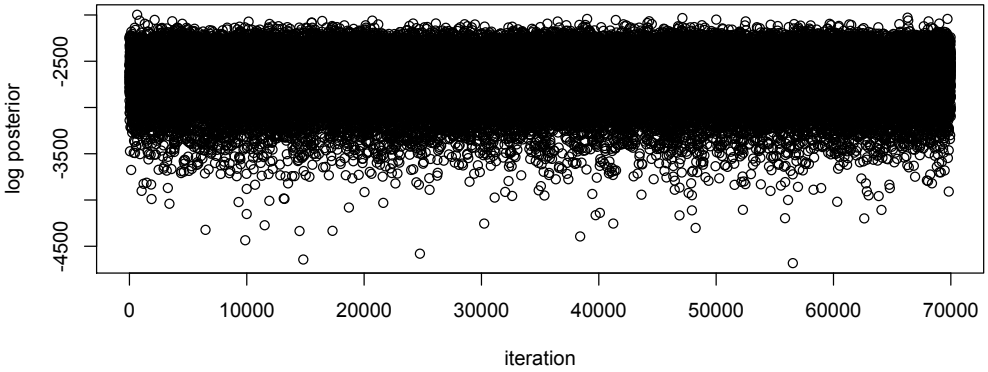
$$\mu_k \sim \mathcal{N}\left(\frac{\sigma_0^2}{\frac{\sigma^2}{|k|}+\sigma_0^2}\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_k(c_i^s)}\sum_{i=1}^n \mathbb{1}_k(c_i^s) y_i\right) + \frac{\sigma^2}{\frac{\sigma^2}{|k|}+\sigma_0^2}\mu_0, \left(\frac{1}{\sigma_0^2}+\frac{|k|}{\sigma^2}\right)^{-1}\right)$$

$$\pi \sim Dir\left(\left[\left(\alpha_1 + \sum_{i=1}^n \mathbb{1}_1(c_i^s)\right), \ldots, \left(\alpha_K + \sum_{i=1}^n \mathbb{1}_K(c_i^s)\right)\right]\right)$$

  - given new parameter values, draw a sample of c_i:

$$c_i \sim Mult\left(\frac{\pi_k^{s+1}\mathcal{N}\left(\mu_k^{s+1}, \sigma_k^{2(s+1)}\right)}{\sum_{\ell=1}^K \pi_\ell^{s+1}\mathcal{N}\left(\mu_\ell^{s+1}, \sigma_\ell^{2(s+1)}\right)}\right).$$

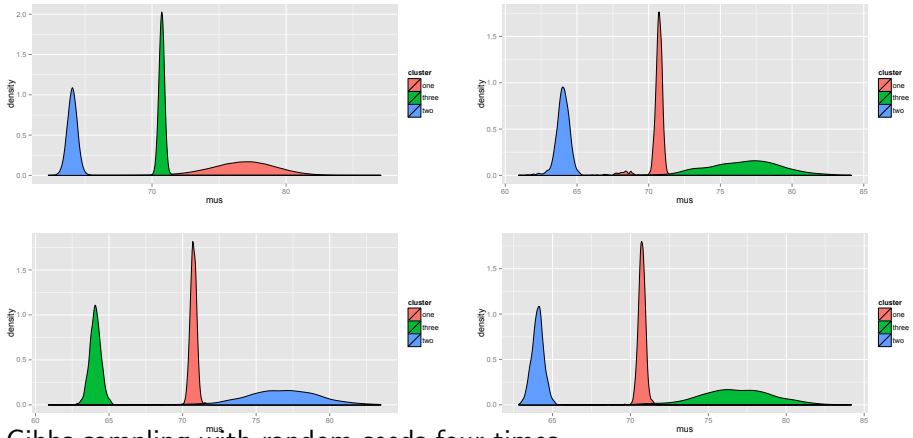
Gibbs sampling in practice



Ran 70,000 iterations of Gibbs sampling on a Gaussian mixture model.

- added auxiliary variables to assign each component to one cluster

Gibbs sampling in practice: repeated restarts



- Run Gibbs sampling with random seeds four times
- plot marginal posterior distributions of each cluster mean
- What works? What does not work?

Is there something faster for high-dimensional inference?

Gibbs sampling draws new samples conditionally on other parameters.

The slow part of this process is the convergence of the Markov chain to the stationary distribution

What if I treated this as an optimization problem?

Instead of sampling parameters, set the new parameter value equal to the expected value of the conditional distribution.

From sampling to optimization

Instead of sampling parameters, set the new parameter value equal to the expected value of the conditional distribution.

- could I justify this in terms of optimizing an objective function?
- does it converge to a local optima?
- what assumptions does it make about the probability model?
- when will it not work?

Is there something faster for high-dimensional inference?

Treated as an optimization problem: set new parameter value equal to expected value of conditional distribution.

- Gibbs sampling: samples from conditional probability will be biased toward high posterior probability samples
- Expected value of (approximate) posterior minimizes (under certain conditions) mean squared error of parameter estimates

Approximate inference for Gaussian mixture model

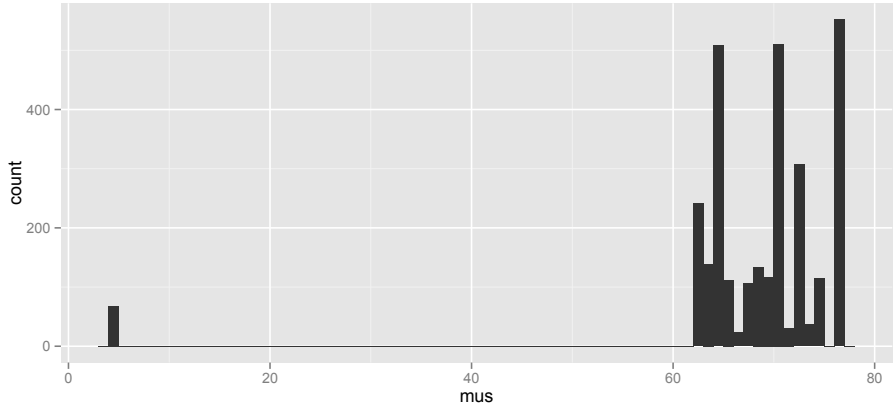
- assign each sample to one of K mixture components in c
- Until convergence:
  - for each parameter z_j, set equal to $\mathrm{E}(z_j \mid z_{-j}, c^s, \mathbf{y})$:

$$\begin{aligned}
\mu_k^{(s+1)} &= \frac{\sigma_0^2}{\frac{\sigma^2}{n} + \sigma_0^2}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + \frac{\sigma^2}{\frac{\sigma^2}{n} + \sigma_0^2}\mu_0 \\
\pi^{(s+1)} &= \left[\left(\alpha_1 + \sum_{i=1}^{n} \mathbb{1}_1(c_i^s)\right), \ldots, \left(\alpha_K + \sum_{i=1}^{n} \mathbb{1}_K(c_i^s)\right)\right]
\end{aligned}$$

  - given new parameter values, set $c_i^{(s+1)}$:

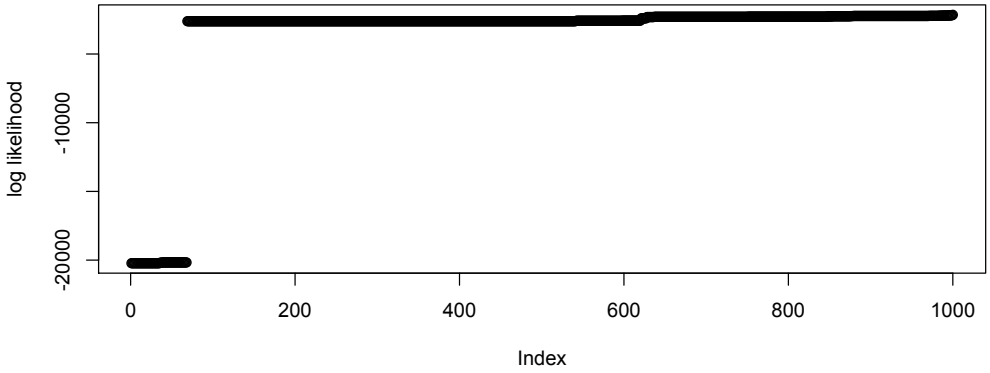
$$c_i^{(s+1)} = \left(\frac{\pi_k^{s+1}\mathcal{N}\left(\mu_k^{s+1}, \sigma_k^{2(s+1)}\right)}{\sum_{\ell=1}^{K} \pi_\ell^{s+1}\mathcal{N}\left(\mu_\ell^{s+1}, \sigma_\ell^{2(s+1)}\right)}\right).$$

Approximate inference in practice



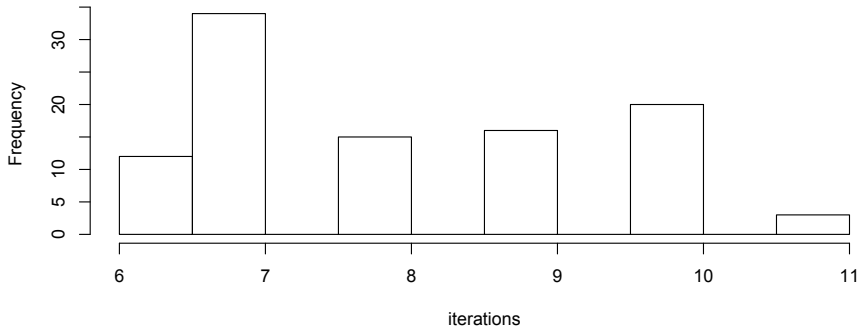
- histogram of cluster means across 100 runs to convergence
- some runs are off the mark, e.g., one cluster is (near) empty
- but: half of the runs give consistent and good results.
- can we tell the difference between good and bad runs?

Approximate inference in practice



- bad runs have low log likelihoods
- idea: run many times from random starting points, remove bad runs.
- have we improved upon Gibbs sampling?

Approximate inference in practice



- takes two orders of magnitude fewer iterations to converge
- even if each iteration took similar time, this would be a big win
- often these iterations are faster (but not always)
- always need to perform random restarts
- What is this algorithm? Why does it work?

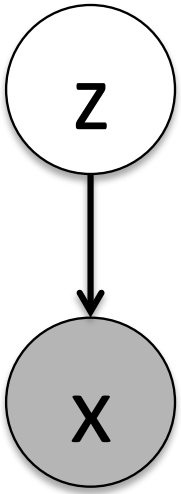
Recall: Generative model for a mixture model

Generative process for each sample

1. Choose $z_i \sim \mathrm{Mult}(\pi)$
2. Choose $x_i \mid z_i \sim p(x_i \mid \theta_{z_i})$

Recall that this generative graphical model factorizes the joint probability of x_i, z_i as:

$$p(x_i, z_i \mid \theta_{1:K}) = p(z_i \mid \pi) \prod_{k=1}^{K} p(x_i \mid \theta_k)^{z_i^k}$$

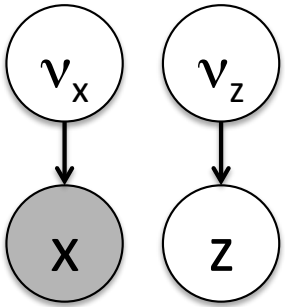


Fully factorized mixture models

This fully factorized model is written as:

$$p(x_i, z_i \mid \theta_{1:K}) = p(z_i \mid \nu_z) \prod_{k=1}^{K} p(x_i \mid \nu_x)$$

where we add *variational parameters* ν to parameterize factorized distributions



Variational methods: general idea

- Idea behind *variational methods* is to select a family of distributions over latent variables with their own *variational parameters*:

$$p(z_{1:p} \mid x) \approx q(z_1 \mid \nu_1) \dots q(z_p \mid \nu_p)$$

- Then, find values of variational parameters that makes q as close to the true posterior p as possible.
- May use q with fitted parameters as an estimate of true posterior:
  - predict future data
  - study posterior distributions of latent variables
  - interpret point estimates of latent variables

Mean field variational inference: fully factorized models

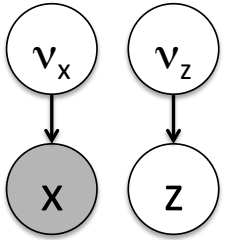
One form of variational inference is *mean field variational inference*.

Here, we abandon our complex model in favor of a fully factorized model.

This model is factorized with respect to the latent variables, and is mathematically tractable.

$$p(x, z \mid \nu, \alpha) = p(\pi \mid \alpha) \prod_{i=1}^{n} p(z_i \mid \nu_{z_i}) \prod_{k=1}^{K} p(x_i \mid \nu_{x_i})$$

Objective: find parameters that minimize distance between distributions $p(z)$ and $q(z)$.



Kullback-Leibler (KL) Divergence

Measure the proximity of $p(z)$ and $q(z)$ with KL divergence:

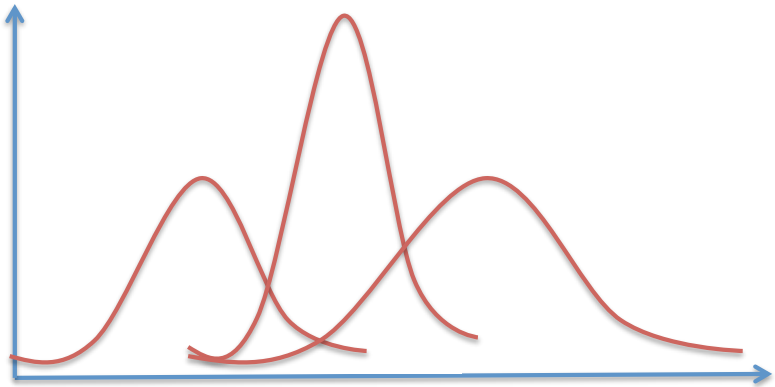
- The KL divergence for variational inference is:

$$KL(q \,||\, p) = \mathrm{E}_q \left[\log \frac{q(z)}{p(z)} \right] = \int_{\mathcal{Z}} q(z) \log \frac{q(z)}{p(z)} dz$$

- when $q(z)$ is large, if $p(z)$ is large, divergence is small
- when $q(z)$ is large, if $p(z)$ is small, divergence is large
- when $q(z)$ is small, $p(z)$ does not matter

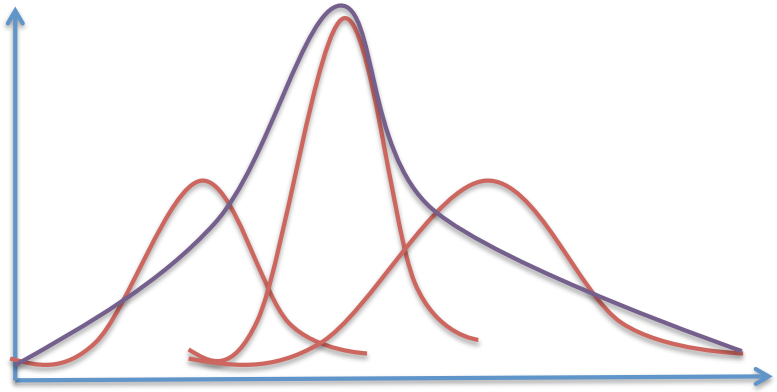
Isn't this formulation backwards?

What approximation are we making here?



Let's consider the Gaussian mixture model.

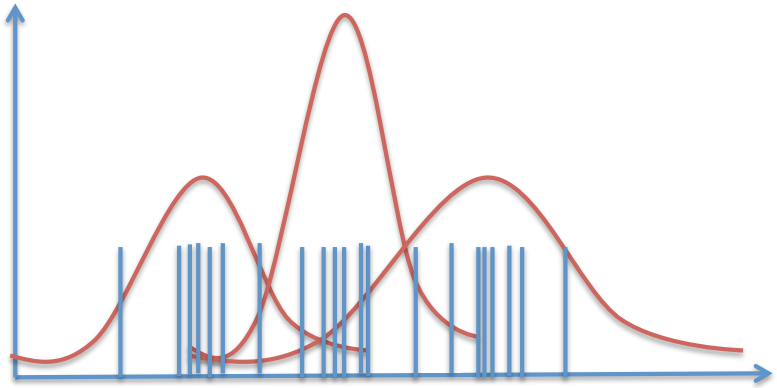
Variational approximation of GMM



Variational approach to fitting a Gaussian mixture model:

- construct a *tractable* distribution to approximate the true posterior
- fit variational parameters to correspond well to the true posterior

Gibbs sampling approximation of GMM



Gibbs sampling approach to fitting a Gaussian mixture model.

Kullback-Leibler (KL) Divergence

Why not consider $KL(p \,||\, q)$?

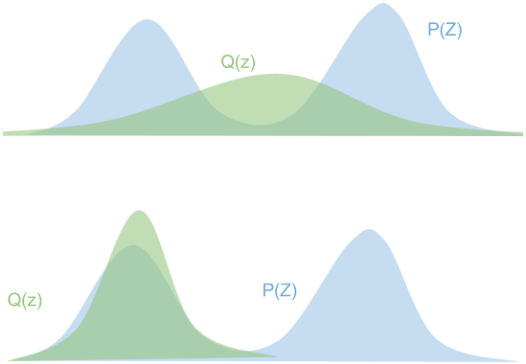
This is more intuitive (expectation under true posterior)

We choose a simple q to take an expectation with respect to a simple distribution.

We can reverse the arguments; this leads to a type of variational inference called *expectation propagation*, which is often more computationally expensive than mean field methods.

KL divergence, both ways, for Gaussian mixture model

Preferred when $KL(q \,||\, p)$:



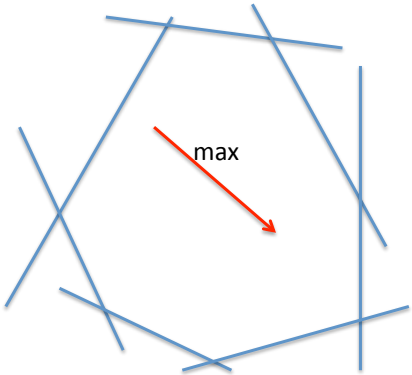
preferred when $KL(p \,||\, q)$

[Figure from Eric Jang's blog]

Variational approximation of optimization problem

Consider parameter estimation as an optimization problem:

- find the parameters that maximize the likelihood of the data
- parameters are subject to constraints in the form of conditional dependencies encoded in the conditional distributions of the model.



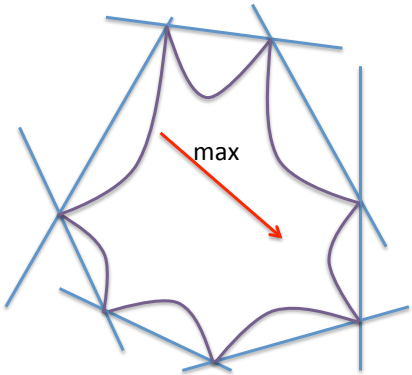
Variational approximation affects constraints in optimization problem.

- find the parameters that maximize the likelihood of the data
- parameters are subject to constraints encoded in the fully factorized version of the model

Variational approximation of optimization problem

Mean field variational approximations impose additional constraints, despite factorizing the distribution

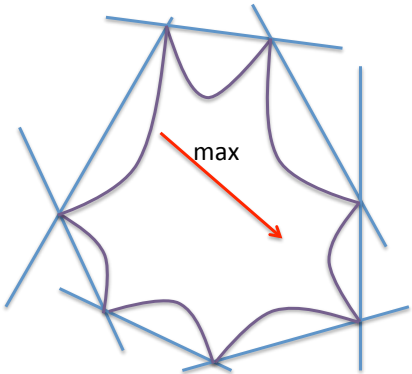
- MFV methods constrain the higher order expectations
- full conditional distributions explicitly model these higher order expectations
- E.g., multivariate Gaussian: covariance matrix off diagonals are 0.



Variational approximation of optimization problem

What do these new constraints say about the method?

- the optimum may not be in the search space
- local optima are likely
- despite having a non-convex structure, coordinate ascent is computationally tractable with a factorized distribution.



Gaussian mixture model

- In the GMM, the posterior is intractable because
  - the cluster assignments are conditionally dependent
  - the cluster means are conditionally dependent
- these dependencies make this posterior intractable
- For the GMM, we factorize all cluster assignment variables

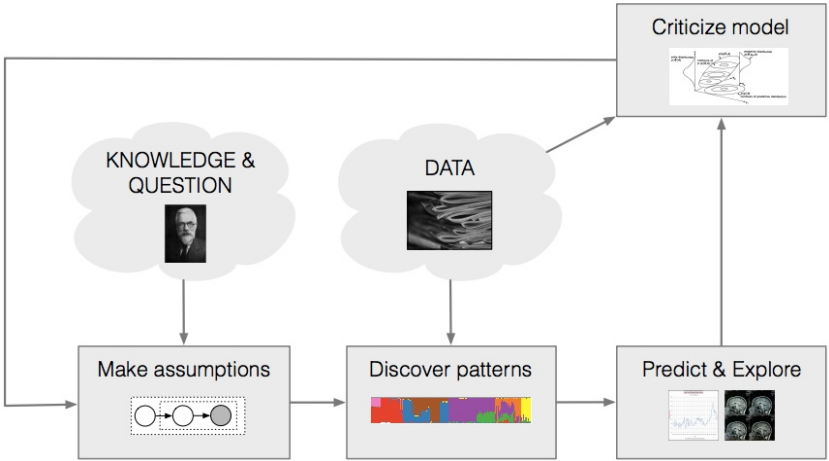
Choice of how to optimize in this space

- using coordinate ascent, optimize each factorized distribution holding the others fixed, conditional on data
- using coordinate ascent variational inference, we replicate the intuitive algorithm based on Gibbs sampling.
- In Gibbs sampling, we sample from $p(z_j \mid z_{-j}, y)$
- In coordinate ascent MFV inference, we set $z_j = \mathrm{E}_q[p(z_j \mid z_{-j}, y)]$

MFV methods in practice

- There are many details I am sweeping under the rug.
- That said, coordinate ascent MFV methods are often as easy as Gibbs sampling to derive and implement
- No choices to be made*; distributions are by definition simple to work with for most conjugate models
- With large data, for intractable Gibbs/EM, this is a good first pass.
- Always run multiple times with different initializations; compare results based on log likelihood

Cycle of data analysis *[Blei 2016]*



MFV methods in practice: warnings

- many types of models where these methods are known to work poorly (binary, sparse, etc.)
- mapping optima back to original constraint set is not always reasonable (cluster assignment c_i)
- convergence to local optima
- point estimate instead of full posterior estimate; stability, uncertainty not straightforward to estimate.

More advanced variational approaches

Structured mean field methods

- do not factorize distribution entirely; retain some structure encoded in tractable distributions

Loopy belief propagation

- run inference algorithms as if model were a tree, not a directed acyclic graph

Variational EM

- E-step: find the values of the variational parameters that are optimal with respect to KL divergence
- M-step: maximize variational bound on log likelihood with respect to true model parameters (same as before)

Note: use variational approximation in place of any difficult optimization

More advanced variational approaches

Stochastic variational inference

- use natural gradient for optimization instead of second derivative
- select one (or a subset) of samples for each update

Variational Bayes

- variational EM, except M-step uses MAP estimates instead of MLE
- point estimates of approximate posterior hyperparameters, instead of point estimates of parameters themselves

Expectation propagation

- Identical framework as MFV methods, but KL divergence is $KL(p(z) \,||\, q(z))$
- practically: iteratively propagates expected values (first moments) instead of full conditional probabilities

Gibbs sampling versus mean field variational methods

- Choosing a method to fit a model is part of the data analysis process
- Both are simple to describe for hierarchical models for which distributions are conjugate

Gibbs sampling

- Gibbs sampling samples from the exact posterior after mixing
- mixing might take a while; no way to know when a chain has mixed

Mean field variational inference

- Variational methods have an unknown bias
- Variational methods are deterministic and convergence is clear
- Variational methods are typically faster than Gibbs sampling.

What models can we now scale to real data?

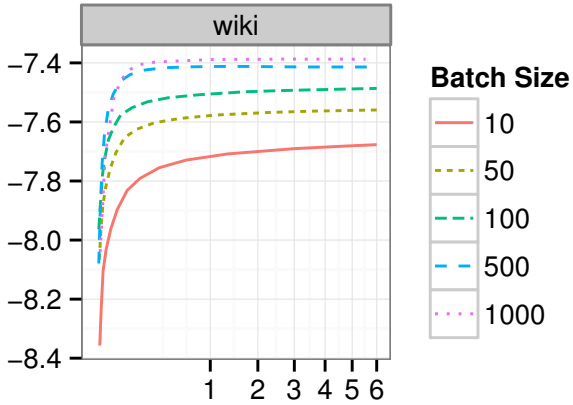
- latent Dirichlet allocation: very large document sets (stochastic variational inference), genomic data
- Matrix factorization methods: non-negative matrix factorization, factor analysis, Poisson matrix factorization (SVI) to Netflix data
- Factorial HMMs: genotype data (MFV methods)
- Sparse regression: spike and slab priors
- Stochastic block model for community detection in networks

What data can we now tackle?

- The Web
- Facebook
- Thousands of genomes
- All archived books
- Music repositories
- Netflix users and movies
- fMRI (neuroscience)
- Images on the web

Example: Topic model for all of Wikipedia (3.5M articles)

Using stochastic variational inference:

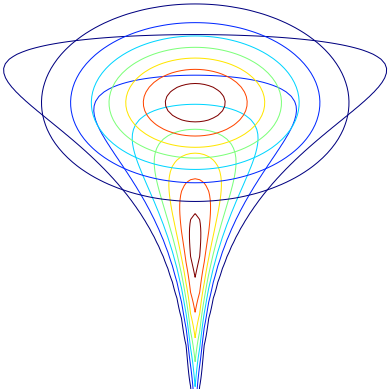


[From Wang et al. 2013]

Newer ideas: Black box variational inference

What do these new constraints say about the method?

- Black box SVI using Autograd to compute gradients automatically *[Duvenaud & Adams]*
- Stan: automatic differentiation VI *[Kucukelbir et al. 2017]*
- Edward: specify model, and VI is automated *[Blei group]*



Summary of variational inference

- When prototyping models, often need fast inference methods
- When point estimates of parameters are sufficient, consider variational inference
- Mean field methods have similar structure as Gibbs sampling, but converge much faster
- Many additional ways to speed up variational inference (stochasticity, in particular)

Additional Resources

- MLAPA: Chapter 24, Chapter 22
- Andrieu et al. 2003 *An Introduction to MCMC for Machine Learning*
- MH Papers: [Hastings 1970] and [Metropolis et al. 1953]
- Gelfand & Smith 1992 *Sampling-Based Approaches to Calculating Marginal Densities*
- (video) Nando de Freitas *Monte Carlo simulation for statistical inference, model selection and decision making*
- (video) David Sontag *Approximate inference in graphical models using LP relaxations*
- Metacademy – Metropolis-Hastings
- Metacademy – Gibbs Sampling
- Metacademy – Slice Sampling
- Metacademy – Hamiltonian Monte Carlo
- Metacademy – Loopy Belief Propagation