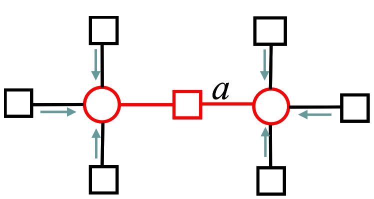
School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Variational (Bayesian) Inference and Mean Field Approximations

Willie Neiswanger

Lecture 13, February 25, 2015



Goals of Inference

Recall, the goals of inference in graphical models include:

- Computing the likelihood of observed data (in models with latent variables).
- Computing the marginal distribution over a given subset of nodes in the model.
- Computing the conditional distribution over a subsets of nodes given a disjoint subset of nodes.
- Computing a mode of the density (for the above distributions).

Approaches to Inference

Recall, approaches to inference include:

- Exact inference algorithms:
  - Brute force.
  - The elimination algorithm.
  - Message passing (sum-product algorithm, belief propagation).
  - Junction tree algorithm.

- Approximate inference algorithms:
  - Loopy belief propagation (← Last Class)
  - Variational (Bayesian) inference + mean field approximations (← Today)
  - Stochastic simulation / sampling / MCMC (← Future Classes)

From Last Class: Loopy Belief Propagation

Recall, from last class:

- We introduced message passing ("belief propagation") on loopy graphs (non-trees).
  - Messages *may* circulate indefinitely.
  - However, it often seems to work empirically.

- But what is happening, theoretically, when it works?

- We can view it as a case of "**variational inference**".

From Last Class: Loopy Belief Propagation

Viewing Loopy Belief Propagation as variational inference:

- We wrote down the KL-divergence between an approximate distribution Q and the distribution P we want to infer.
- We defined a similar value: the (Gibbs) “Free Energy”.
  - This Free Energy consists of an entropy term and an expected log marginal term.
- Computing the Free Energy is hard, in general, so we instead use approximations, such as the Bethe approximation.
- We then minimize the Bethe Free Energy (i.e. the Free Energy with Bethe approximation).
- We also described another approximation in “generalized belief propagation”.
  - Allows for a more general variational approximation.

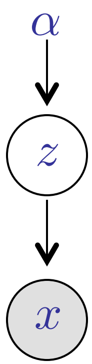
Variational (Bayesian) Inference and Mean Field Approximations

(Notation and examples from David Blei's tutorial on Variational Inference)

Notation

We use the following notation for the rest of the lecture:

- n observations: $x = x_{1:n}$
- m latent variables: $z = z_{1:m}$
- fixed parameters: α
  - These parameters could be for the distribution over the observations or over the hidden variables.
- This notation can describe (just about) any graphical model.
  - (i.e. any Bayes net or Markov random field).
- Example graphical model ---------------->



Problem Setup

- In modern machine learning, variational (Bayesian) inference, which we will refer to here as **variational Bayes**, is most often used to infer the conditional distribution over the latent variables given the observations (and parameters).
- This is also known as the **posterior distribution** over the latent variables.
- With our notation, the posterior is written:

$$p(z|x,\alpha) = \frac{p(z,x|\alpha)}{\int_z p(z,x|\alpha)}$$

Motivating Example

- Why do we often need to use an approximate inference methods (such as variational Bayes) to compute the posterior distribution over nodes in our graphical model?

- It's because we cannot directly compute the posterior distribution for many interesting models.
  - I.e. the posterior density is in an intractable form (often involving integrals) which cannot be easily analytically solved.

- As a motivating example, we will try to compute the posterior for a (Bayesian) mixture of Gaussians.

Motivating Example

Bayesian mixture of Gaussians

- The likelihood (i.e. the generative process):

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$ for $k = 1, \ldots, K$.
2. For $i = 1, \ldots, n$
   (a) Draw $z_i \sim \text{Cat}(\pi)$.
   (b) Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2)$.

- Note that we have observed variables, $x_{1:n}$, latent variables $\mu_{1:k}$ and $z_{1:n}$, and parameters $\{\tau^2, \pi, \sigma^2\}$.

Motivating Example

- We can write the posterior distribution as:

$$p(\mu_{1:K}, z_{1:n} | x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})}$$

- Where we have suppressed writing the parameters for ease of notation.

Motivating Example

- Can we compute this density?
- The numerator can be computed for any choice of the latent variables.
- The problem is the denominator (the marginal probability of the observations):

$$p(x_{1:n}) = \int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})$$

$$= \int_{\mu_{1:K}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} \sum_{z_{1:n}} p(z_i) p(x_i | z_i, \mu_{1:K})$$

- This integral cannot easily be computed analytically.

Variational Bayes

The main idea behind variational Bayes:

- Choose a family of distributions over the latent variables $z_{1:m}$ with its own set of variational parameters ν, i.e.

$$q(z_{1:m}|\nu)$$

- Then, we find the setting of the parameters that makes our approximation q closest to the posterior distribution.
  - This is where optimization algorithms come in.
- Then we can use q with the fitted parameters in place of the posterior.
  - E.g. to form predictions about future data, or to investigate the posterior distribution over the hidden variables, find modes, etc.

Kullback-Leibler Divergence

- We measure the closeness of the two distributions with the Kullback-Leibler (KL) divergence, defined to be

$$\mathrm{KL}(q||p) = \int_z q(z) \log \frac{q(z)}{p(z|x)} = \mathbb{E}_q \left[\log \frac{q(z)}{p(z|x)} \right]$$

- Intuitively, there are three “cases” of importance:
  - If q is high and p is high, then we are happy (i.e. low KL divergence).
  - If q is high and p is low then we pay a price (i.e. high KL divergence).
  - If q is low then we don’t care (i.e. also low KL divergence, regardless of p).

- Intuitively, it might make more sense to consider $\mathrm{KL}(p||q)$
  - however, we do not do this for computational reasons (which we will explain).

The Evidence Lower Bound

- So: to do variational Bayes, we want to minimize the KL divergence between our approximation q and our posterior p.
- However, we can't actually minimize this quantity (we will show why later), but we can minimize a function that is equal to it up to a constant.
- This function is known as the **evidence lower bound (ELBO)**.
- Recall that the "evidence" is a term used for the marginal likelihood of observations (or the log of that).

Deriving the Evidence Lower Bound

- First recall Jensen's inequality (applied to random variables X):
  When f is concave, $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$.
- We apply Jensen's inequality to the log (marginal) probability of the observations to get the ELBO.

$$
\begin{aligned}
\log p(x) &= \log \int_z p(x,z) \\
&= \log \int_z p(x,z) \frac{q(z)}{q(z)} \\
&= \log \left(\mathbb{E}_q \left[\frac{p(x,z)}{q(z)} \right] \right) \\
&\geq \mathbb{E}_q [\log p(x,z)] - \mathbb{E}_q [\log q(z)]
\end{aligned}
$$

This final line is the ELBO! It is a lower bound for the evidence.

The Evidence Lower Bound

All together, the Evidence Lower Bound (ELBO) for a probability model $p(x,z)$ and approximation $q(z)$ to the posterior is :

$$\mathbb{E}_q\left[\log p(x,z)\right] - \mathbb{E}_q\left[\log q(z)\right]$$

- This quantity is less than or equal to the evidence (log marginal probability of the observations).
- We optimize this quantity (over densities $q(z)$) in Variational Bayes to find an "optimal approximation".

The Evidence Lower Bound

Notes:

- We choose a family of variational distributions (i.e. a family of approximations) such that these two expectations can be computed.
- The second expectation is the “entropy”, another quantity from information theory.
- In variational inference, we find settings of the variational parameters ν that maximize the ELBO, which is equivalent to minimizing the KL divergence.
  - Why is this? On next slide.

The Evidence Lower Bound

Why do we maximize the ELBO?

- First recall that

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

- Next, we can write the KL divergence as:

$$\begin{aligned}
\mathrm{KL}(q||p) &= \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x)}\right] \\
&= \mathbb{E}_q\left[\log q(z)\right] - \mathbb{E}_q\left[\log p(z|x)\right] \\
&= \mathbb{E}_q\left[\log q(z)\right] - \mathbb{E}_q\left[\log p(z,x)\right] + \log p(x) \\
&= -\left(\mathbb{E}_q\left[\log p(z,x)\right] - \mathbb{E}_q\left[\log q(z)\right]\right) + \log p(x)
\end{aligned}$$

This final line is the negative ELBO plus a constant (that does not depend on q).

The Evidence Lower Bound

Hence…

- Therefore, finding an approximation q that maximizes the ELBO is equivalent to finding the q that minimizes the KL divergence to the posterior!

- Note: the difference between the ELBO and the KL divergence is the log normalizer (i.e. the evidence), which is the quantity that the ELBO bounds.

Quick Recap

Quick recap on what we've covered so far:

- We often cannot compute posteriors, and so we need to approximate them, using (for e.g.) variational methods.
- In variational Bayes, we'd like to find an approximation within some family that minimizes the KL divergence to the posterior, but we can't directly minimize this.
- Therefore, we defined the ELBO, which we can maximize, and this is equivalent to minimizing the KL divergence.
- Next, we will discuss a specific family of approximations.

Mean Field Variational Inference

- We now describe a popular family of variational approximations called **mean field approximations**.

- In this type of variational inference, we assume the variational distribution over the latent variables factorizes as

$$q(z_1, \ldots, z_m) = \prod_{j=1}^{m} q(z_j)$$

  (where we omit variational parameters for ease of notation).

  - We refer to $q(z_j)$, the variational approximation for a single latent variable, as a "local variational approximation".

- In the above expression, the variational approximation $q(z_j)$ over each latent variable z_j is independent.

Mean Field Variational Inference

- Note that this is a fairly general setup; we can also partition the latent variables $z_1, \ldots, z_m$ into R groups $z_{G_1}, \ldots, z_{G_R}$, and use the approximation:

$$q(z_1, \ldots, z_m) = q(z_{G_1}, \ldots, z_{G_R}) = \prod_{r=1}^{R} q(z_{G_r})$$

  - Often called "generalized mean field" versus (the above) "naïve mean field".
  - More on this later, applied to Markov random fields.

- Typically, this approximation does not contain the true posterior (because the latent variables are dependent).
  - E.g.: in the (Bayesian) mixture of Gaussians model, all of the cluster assignments z_i for $i = 1, \ldots, n$ are dependent on each other and on the cluster locations $\mu_{1:K}$, given data $x_{1:n}$.

Optimizing the ELBO in Mean Field Variational Inference

How do we optimize the ELBO in mean field variational inference?

- Typically, we use coordinate ascent optimization.
- I.e. we optimize each latent variable's variational approximation $q(z_j)$ in turn while holding the others fixed.
  - At each iteration we get an updated "local" variational approximation.
  - And we iterate through each latent variable until convergence.
- Note: this is not the only way to optimize the ELBO in mean field approximations (e.g. one can do gradient ascent, using the "natural gradient"), however it is a very popular method.

Optimizing the ELBO in Mean Field Variational Inference

- First, recall that the (probability) chain rule gives:

$$p(z_{1:m}, x_{1:n}) = p(x_{1:n}) \prod_{j=1}^{m} p(z_j | z_{1:(j-1)}, x_{1:n})$$

Note that the latent variables in this product can occur in any order (i.e. the indexing from 1 to m is arbitrary)---this will be important later.

- Second, note that we can decompose the entropy term of the ELBO (using the mean field variational approximation) as

$$\mathbb{E}_q \left[\log q(z_{1:m}) \right] = \sum_{j=1}^{m} \mathbb{E}_{q_j} \left[\log q(z_j) \right]$$

Optimizing the ELBO in Mean Field Variational Inference

- Third, using the previous two facts, we can decompose the ELBO $\mathcal{L}$ for the mean field variational approximation into a nice form.

- Recall that the ELBO is defined as:

$$\mathbb{E}_q\left[\log p(x, z)\right] - \mathbb{E}_q\left[\log q(z)\right]$$

- Therefore, under the mean field approximation, the ELBO can be written:

$$\mathcal{L} = \log p(x_{1:n}) + \sum_{j=1}^{m}(\mathbb{E}_q\left[\log p(z_j | z_{1:(j-1)}, x_{1:n})\right] - \mathbb{E}_{q_j}\left[\log q(z_j)\right])$$

Optimizing the ELBO in Mean Field Variational Inference

Before we can continue, we need to introduce some terminology:

- "**The conditional**" for latent variable z_j is:

$$p(z_j|z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_m, x) = p(z_j|z_{-j}, x)$$

- Where the $-j$ notation denotes all indices other than the j^{th}.

- This is actually the "posterior conditional" of z_j, given all other latent variables and observations.

- This posterior conditional is very important in mean field variational Bayes, and will be important in future inference algorithms used in this class, such as Gibbs sampling.

Optimizing the ELBO in Mean Field Variational Inference

- Again, we wrote the ELBO $\mathcal{L}$ for the mean field variational approximation as:

$$\mathcal{L} = \log p(x_{1:n}) + \sum_{j=1}^{m}\left(\mathbb{E}_q\left[\log p(z_j|z_{1:(j-1)}, x_{1:n})\right] - \mathbb{E}_{q_j}\left[\log q(z_j)\right]\right)$$

- Next, we want to derive the coordinate ascent update for a latent variable z_j, keeping all other latent variables fixed.
  - i.e. we want the $\mathrm{argmax}_{q_j}\mathcal{L}$.

- Removing the parts that do not depend on $q(z_j)$, we can write:

$$\begin{aligned}\mathrm{argmax}_{q_j}\mathcal{L} &= \mathrm{argmax}_{q_j}\left(\mathbb{E}_q\left[\log p(z_j|z_{-j}, x)\right] - \mathbb{E}_{q_j}\left[\log q(z_j)\right]\right) \\ &= \mathrm{argmax}_{q_j}\left(\int q(z_j)\mathbb{E}_{q_{-j}}\left[\log p(z_j|z_{-j}, x)\right] dz_j - \int q(z_j)\log q(z_j) dz_j\right)\end{aligned}$$

Optimizing the ELBO in Mean Field Variational Inference

Notes:

- In the previous expression, to get the term $\mathbb{E}_q\left[\log p(z_j|z_{-j},x)\right]$, we have re-ordered the latent variables in our sum so that the j^{th} latent variable comes last.
- The notation $\mathbb{E}_{q_{-j}}$ is the expectation over all "other" latent variables (except for the j^{th}).
- We define the term inside the argmax on the last line to be called $\mathcal{L}_j$, i.e.

$$\mathcal{L}_j = \int q(z_j)\mathbb{E}_{q_{-j}}\left[\log p(z_j|z_{-j},x)\right]dz_j - \int q(z_j)\log q(z_j)dz_j$$

- Note here that we have decomposed the expectation over q as an integral over z_j of an expectation over $q(z_{-j})$.

Optimizing the ELBO in Mean Field Variational Inference

- To find this argmax, we take the derivative of $\mathcal{L}_j$ with respect to $q(z_j)$, use Lagrange multipliers, and set the derivative to zero:

$$\frac{d\mathcal{L}_j}{dq(z_j)} = \mathbb{E}_{q_{-j}}\left[\log p(z_j|z_{-j}, x)\right] - \log q(z_j) - 1 = 0$$

- From this, we arrive at the coordinate ascent update:

$$q^*(z_j) \propto \exp\left\{\mathbb{E}_{q_{-j}}\left[\log p(z_j|z_{-j}, x)\right]\right\}$$

- However, since the denominator of the conditional does not depend on z_j, we can equivalently write:

$$q^*(z_j) \propto \exp\left\{\mathbb{E}_{q_{-j}}\left[\log p(z_j, z_{-j}, x)\right]\right\}$$

Optimizing the ELBO in Mean Field Variational Inference

Notes:

- This coordinate ascent procedure convergences to a ***local* maximum**.
- The coordinate ascent update for $q(z_j)$ only depends on the other, fixed approximations $q(z_k)$, $k \neq j$.
- While this determines the optimal $q(z_j)$, we haven't yet specified the form (i.e. what specific distribution family) of q we aim to use, only the factorization.
- Depending on what form we use, the coordinate update $q^*(z_j)$ might not be easy to work with (and might not be in the same form as $q(z_j)$ …).
    - But in many cases it is!
    - And we will specify what forms yield good coordinate updates.

Optimizing the ELBO in Mean Field Variational Inference

Simple Example: multinomial conditionals

- Suppose we have chosen a model whose conditional distribution is a multinomial, i.e.

$$p(z_j|z_{-j}, x) = \pi(z_{-j}, x)$$

- Then the optimal (coordinate update for) $q(z_j)$ is:

$$q^*(z_j) \propto \exp\{\mathbb{E}[\log \pi(z_{-j}, x)]\}$$

- Which is also a multinomial, and is easy to compute. So choosing a multinomial family of approximations for each latent variable gives closed form coordinate ascent updates.

Quick Recap

Quick recap on what we've covered:

- We defined a family of approximations called "mean field" approximations, in which there are no dependencies between latent variables (and also a generalized version of this).
- We decomposed the ELBO into a nice form under mean field assumptions.
- We derived coordinate ascent updates to iteratively optimize each local variational approximation under mean field assumptions.
- Next, we will discuss specific forms for the local variational approximations in which we can easily compute (closed-form) coordinate ascent updates.

Exponential Family Conditionals

- Is there a general form for models in which the coordinate updates in mean field variational inference are easy to compute and lead to closed-form updates?

- Yes: the answer is exponential family conditionals.

- I.e. models with conditional densities that are in an exponential family, i.e. of the form:

$$p(z_j|z_{-j},x) = h(z_j)\exp\left\{\eta(z_{-j},x)^\top t(z_j) - a(\eta(z_{-j},x))\right\}$$

where h, η, t, and a are functions that parameterize the exponential family.

- Different choices of these parameters lead to many popular densities (normal, gamma, exponential, Bernouilli, Dirichlet, categorical, beta, Poisson, geometric, etc.).

Exponential Family Conditionals

- We call these "exponential-family-conditional" models.
  - Also known as "conditionally conjugate models".
- Many popular models fall into this category, including:
  - Bayesian mixtures of exponential family models with conjugate priors.
  - Hierarchical hidden Markov models.
  - Kalman filter models and switching Kalman filters.
  - Mixed-membership models of exponential families.
  - Factorial mixtures / hidden Markov models of exponential families.
  - Bayesian linear regression.
  - Any model containing only conjugate pairs and multinomials.
- Some popular models do not fall into this category, including:
  - Bayesian logistic regression and other nonconjugate Bayesian generalized linear models.
  - Correlated topic model, dynamic topic model.
  - Discrete choice models.
  - Nonlinear matrix factorization models.

Exponential Family Conditionals

- We can derive a general formula for the coordinate ascent update for all exponential-family-conditional models.
- First, we will choose the form of our local variational approximation $q(z_j)$ to be the same as the conditional distribution (i.e. in an exponential family).
- When we perform our coordinate ascent update, we will see that the update yields an optimal $q(z_j)$ in the same family.
- Recall from above that we derived the coordinate ascent updates for optimizing the ELBO (under the mean field assumption) as:

$$q^*(z_j) \propto \exp\left\{\mathbb{E}_{q_{-j}}\left[\log p(z_j|z_{-j}, x)\right]\right\}$$

Exponential Family Conditionals

Coordinate ascent updates for exponential-family-conditional models (under the mean field approximation):

- The log of the conditional:

$$\log p(z_j | z_{-j}, x) = \log h(z_j) + \eta(z_{-j}, x)^\top t(z_j) - a(\eta(z_{-j}, x))$$

- The expectation of this with respect to $q(z_{-j})$ is:

$$\mathbb{E}_{q_{-j}}\left[\log p(z_j | z_{-j}, x)\right] = \log h(z_j) + \mathbb{E}_{q_{-j}}\left[\eta(z_{-j}, x)\right]^\top t(z_j) - \mathbb{E}_{q_{-j}}\left[a(\eta(z_{-j}, x))\right]$$

- The last term does not depend on $q(z_j)$, so we have the update:

$$q^*(z_j) \propto h(z_j) \exp\left\{\mathbb{E}_{q_{-j}}\left[\eta(z_{-j}, x)\right]^\top t(z_j)\right\}$$

- So the optimal $q(z_j)$ is in the same exponential family as the conditional.

Exponential Family Conditionals

Writing this update in terms of variational parameters ν.

- Give each latent variable a variational parameter ν_j. Under the mean field assumption, we can write the full approximation as :

$$q(z_{1:m}|\nu) = \prod_{j=1}^{m} q(z_j|\nu_j)$$

where each local variational approximation has an exponential family form.

- Then the coordinate ascent algorithm updates each variational parameter, in turn, as:

$$\nu_j^* = \mathbb{E}_{q_{-j}}\left[\eta(z_{-j}, x)\right]$$

Quick Recap

Quick recap on what we've covered:

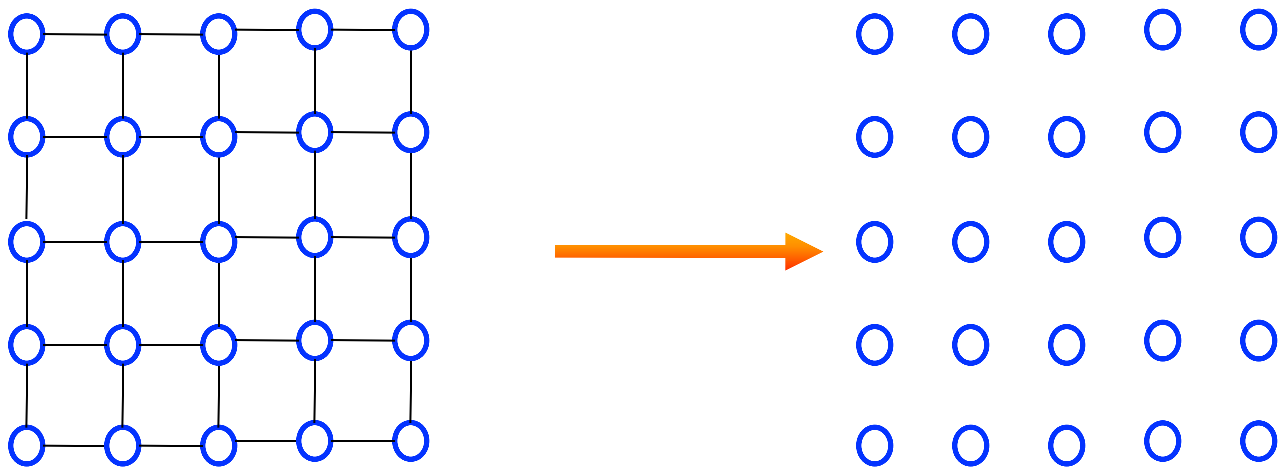
- We found a family of models (exponential-family-conditional models) in which we have closed form coordinate ascent updates to optimize the ELBO.
  - And we gave a number of examples (and non-examples) of these models.

- We gave an explicit form for the coordinate ascent update for these exponential-family-conditional models.
  - And also looked at the update in terms of the local variational parameters.

Mean Field for Markov Random Fields

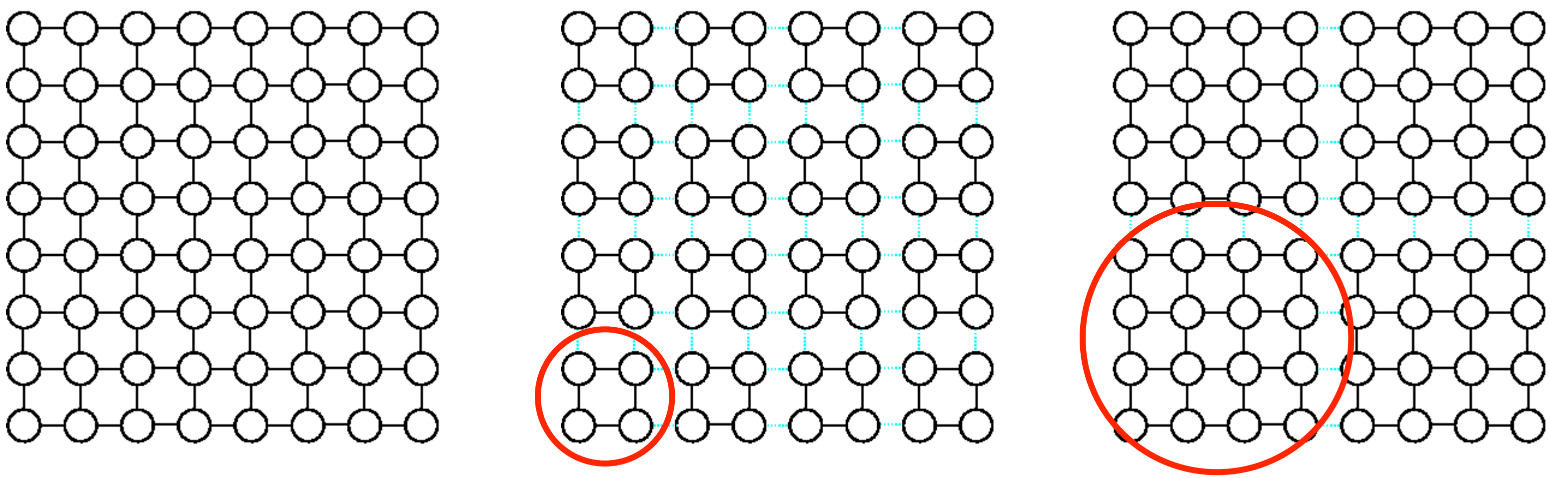
- We can also apply similar mean field approximations for Markov random fields (such as the Ising model):

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



Mean Field for Markov Random Fields

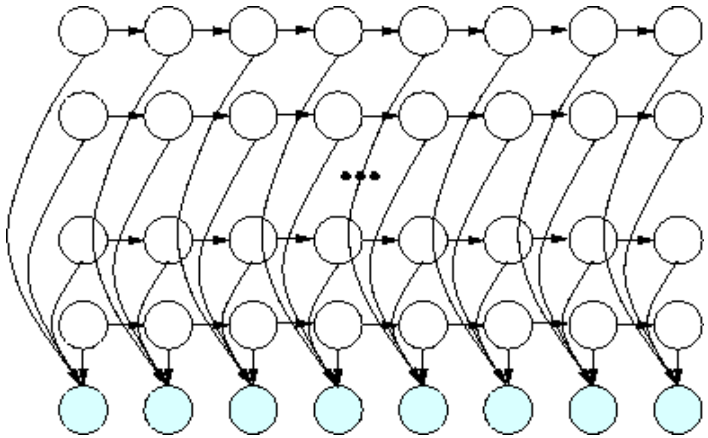
- We can also apply more general forms of mean field approximations (involving clusters) to the Ising model:
- Instead of making all latent variables independent (i.e. naïve mean field, previous figure), clusters of (disjoint) latent variables are independent.

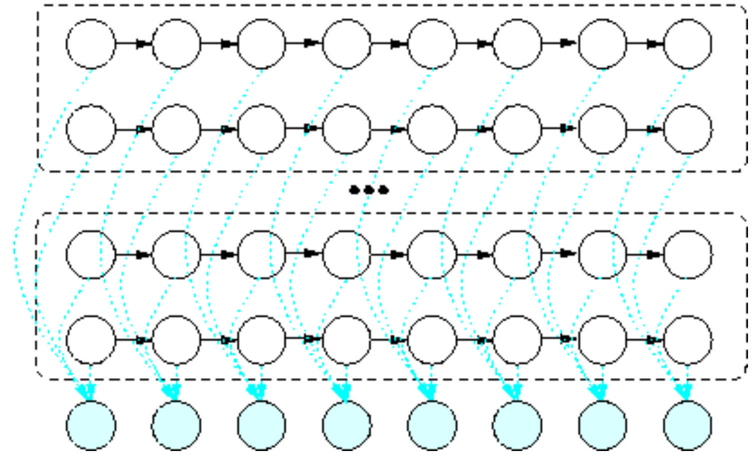


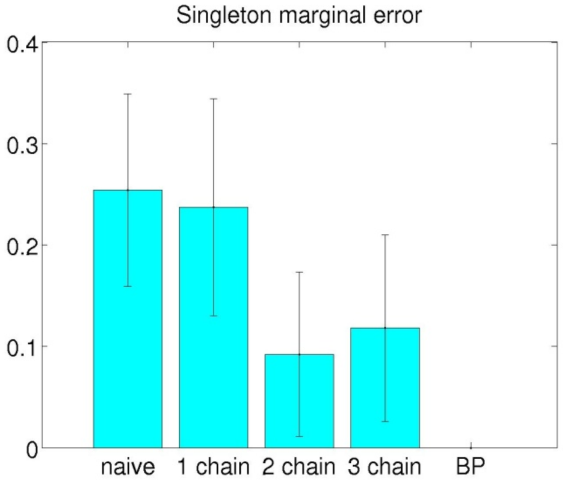
Generalized (Cluster-based) Mean Field for MRFs

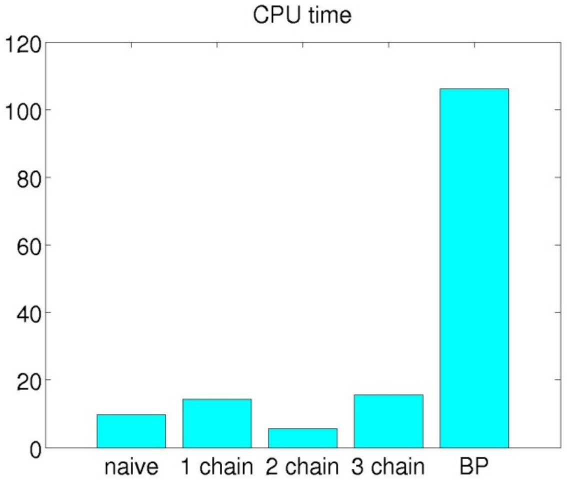
- For these MRFs there exist a general, iterative message passing algorithm for inference (similar to the loopy-BP algorithm learned in the previous class).
- Clustering completely defines the approximation.
  - Preserves dependencies.
  - Allows for a flexible performance/cost trade-off.
  - Clustering can be done in an automated fashion.
- Generalizes model-specific structured VI algorithms, including:
  - fHMM, LDA.
  - Variational Bayesian learning algorithms
- Provides new structured VI approximations to complex models

Some Results: Factorial HMMs

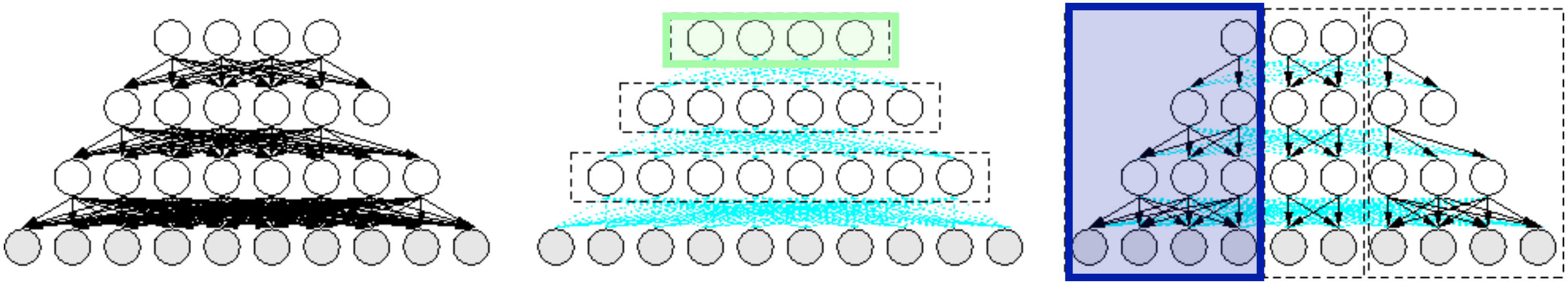


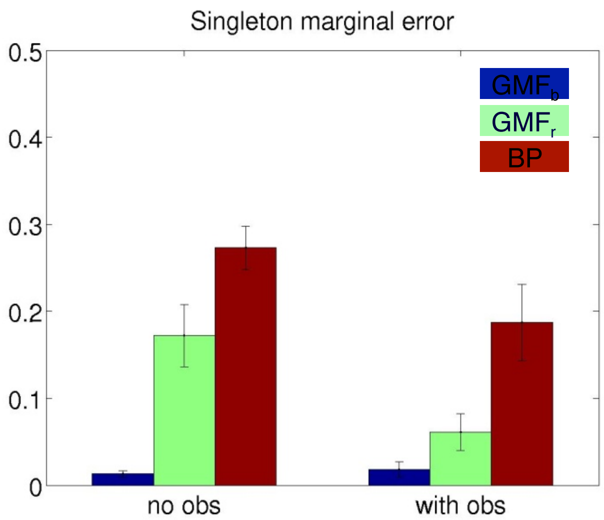


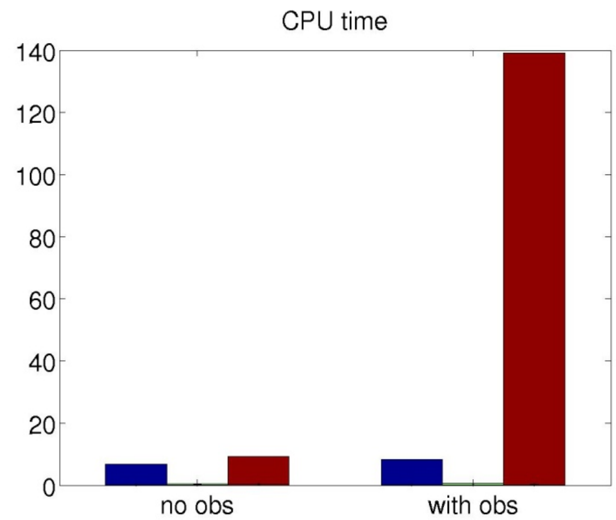




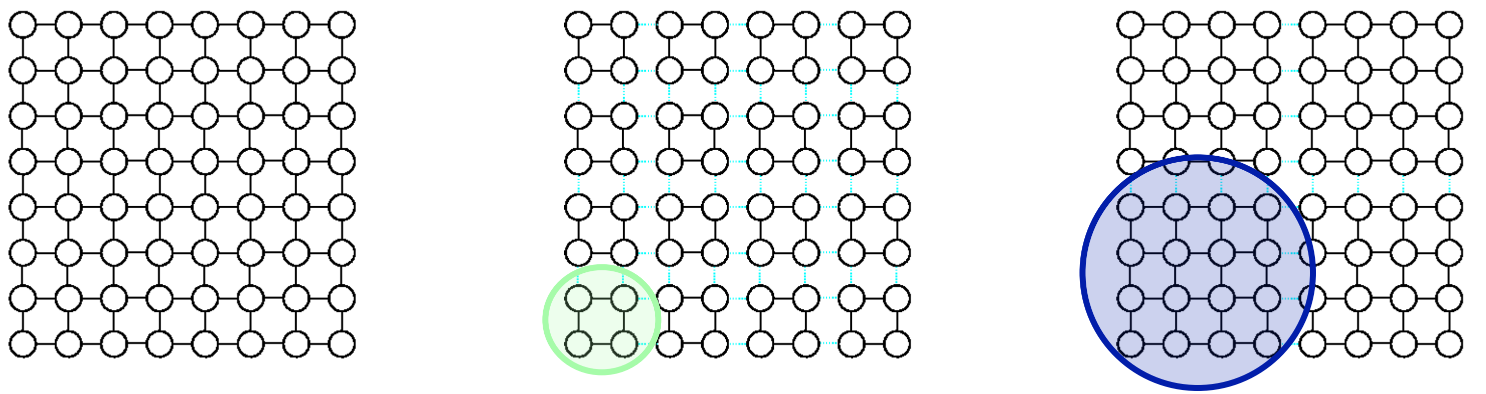
Some Results: Sigmoid Belief Networks

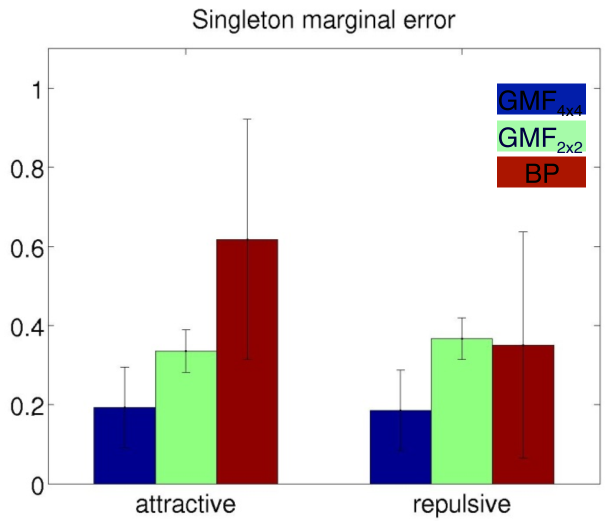


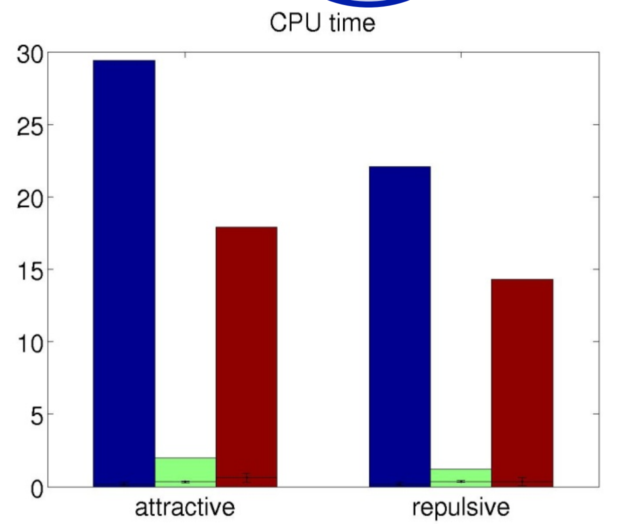




Some Results: Ising Models







Attractive coupling: positively weighted
Repulsive coupling: negatively weighted