10-708: Probabilistic Graphical Models 10-708, Spring 2017

Lecture 13 : Variational Inference: Mean Field Approximation

Lecturer: Willie Neiswanger *Scribes: Xupeng Tong, Minxing Liu*

1 Problem Setup

1.1 Recap on Inference

- Goals of Inference
  - 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
  - Exact inference algorithms
    - Brute force
    - The elimination algorithm.
    - Message passing (sum-product algorithm, belief propagation).
    - Junction tree algorithm.
  - Approximate inference algorithms
    - Loopy belief propagation
    - Variational (Bayesian) inference + mean field approximations
    - Stochastic simulation / sampling / MCMC

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.

The posterior can be written as,

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

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.)

1.2 Motivating Example

Take **Bayesian mixture of Gaussians** for example,

- The likelihood (i.e. the generative process):
  - Draw $\mu_k \sim N(0, \tau^2)$ for $k = 1, \cdots, K$
  - For $i = 1, \cdots, n$
    - Draw $z_i \sim Cat(\pi)$
    - Draw $x_i \sim N(\mu_{z_i}, \sigma^2)$
- $p(\mu_{1:n}, z_{1:n}|x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i)p(x_i|z_i,\mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i)p(x_i|z_i,\mu_{1:K})}$

Since the integral cannot be easily computed analytically, we will use variational bayes. The main idea behind this is to 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 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.)

1.3 Kullback-Leibler Divergence

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

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

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 dont care (i.e. also low KL divergence, regardless of p).

Intuitively, it might make more sense to consider $KL(p\|q)$, however, we do not do this for computational reasons.

1.4 Evidence Lower Bound

To do variational Bayes, we want to minimize the KL divergence between our approximation q and our posterior p. However, we cant 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).

To derive the Evidence Lower Bound, we introduce Jensens inequality (applied to random variables X) here:

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

We apply Jensens inequality to the log (marginal) probability of the observations to get the ELBO.

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

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

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

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

Notes that,

- 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.

By rewriting the KL divergence as,

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

We observe that this final line is the negative ELBO plus a constant (that does not depend on q). Therefore, we conclude that finding an approximation q that maximizes the ELBO is equivalent to finding the q that minimizes the KL divergence to the posterior.

We often cannot compute posteriors, and so we need to approximate them, using (for e.g.) variational methods. In variational Bayes, wed like to find an approximation within some family that minimizes the KL divergence to the posterior, but we cant directly minimize this. Therefore, we defined the ELBO, which we can maximize, and this is equivalent to minimizing the KL divergence. 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.

2 Mean Field Variational Inference

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

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

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

This setup is a fairly general, however we can also partition the latent variables $z_1, \cdots, z_m$ into R groups $z_{G_1}, \cdots, z_{G_R}$, and use the approximation:

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

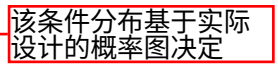
This is often called "generalized mean field" as compared to "naive mean field".

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 $i = 1 \cdots n$ are dependent on each other and on the cluster locations $\mu_{1:K}$, given data $x_{1:n}$.

We now want to optimize the ELBO in mean field variational inference. Typically, coordinate ascent optimization is used (i.e. we optimize each latent variables variational approximation q_{z_j} in turn while holding the others fixed.). We have to note that 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.

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

  $$p(z_{1:m}, x_{1:n}) = p(x_{1n}) \prod_{j=1}^{m} p(z_j | z_{1(j1)}, x_{1n})$$

  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

  $$\mathbb{E}_q[\log(q_{1:m})] = \sum_{j}^{m} \mathbb{E}_{q_j}[\log(q_j)]$$

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

  The ELBO is defined as,

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

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

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

Introducing some terminologies,

- "The conditional" for latent variable is:

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

  where the notation $-j$ 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.

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

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

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

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

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

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

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

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

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

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

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 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 $q(z_j)$.

To wrap up, we first 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). Then we decomposed the ELBO into a nice form under mean field assumptions. Later on, we derived a coordinate ascent updates to iteratively optimize each local variational approximation under mean field assumptions. There are many specific forms for the local variational approximations in which we can easily compute (closed-form) coordinate ascent updates.

3 Probabilistic Topic Models

Humans cannot afford to deal with a huge number of text documents (e.g., search, browse, or measure similarity). We then turn to computational tools to help solve these tasks. Generally, we want to find a map from each document to a vector space representation, $D \to R^d$ (see Figure 1). This problem is also called document embedding. Document embedding enables us to perform several tasks relevant to documents, like calculating the similarity between documents, content classification, document clustering, distilling semantics and perspectives from documents, etc.

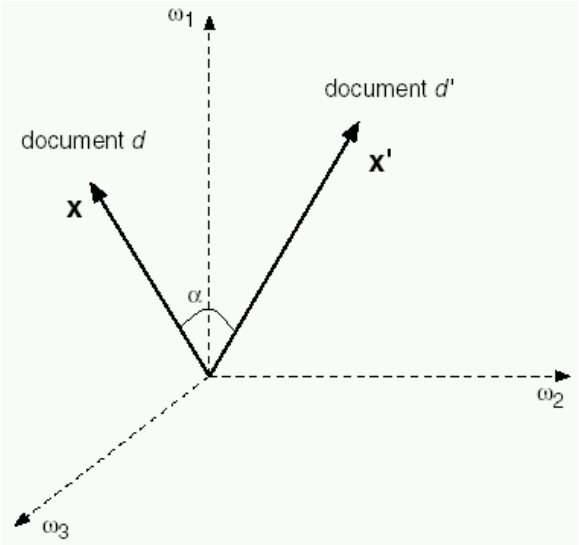


Figure 1: Document Embedding

3.1 Bags of Words Representation

A common representation of documents is to represent each document as a bag of words. Bag of words representation ignores the order of words in a document, and only the frequency of the word matters. Finally each document will be mapped to a high-dimensional sparse vector. Bags of words representation is simple, but has a few drawbacks. The high dimensionality and sparsity makes it inefficient for text processing tasks, e.g., document classification, document similarity, etc. It also fails to consider the semantics embedded in the document. For example, two sentences may talk about the same topic even if they don't share any common words. In bags of words representation, they will not at all be considered similar. Last but not least, the bags of words representation is also not effective for browsing.

3.2 Topic Model

Probabilistic Topic Modeling is developed to organize the documents according to the topics. Now it's like that we map the document to the vector space defined by topics, which is relatively a lower-dimensional space compared to directly using the bags of words representation. In specific, we view each document as generated by a certain topic model, with the steps below:

Draw θ from the prior;
for *each word* n **do**
  Draw topic distribution z_n from $multinomial(\theta)$;
  Draw word $w_n|z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$;
end

Algorithm 1: Generating a document

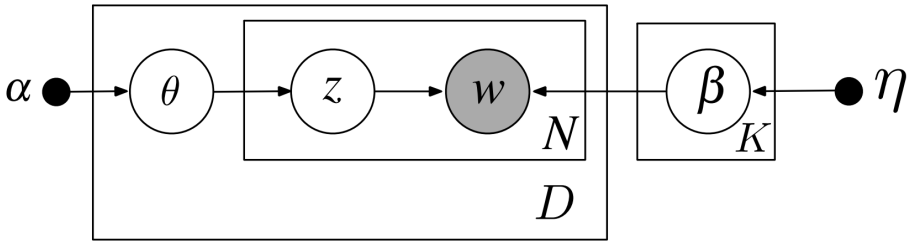


Figure 2: Graphical Model of LDA

In detail, we first generate a topic distribution for a document. In this way, a document can have multiple topics. Then for each word, we select a topic from the topic distribution, and then select a word for the document according to the probability of words given on the topic. The prior θ and β are the distribution of the topic distribution and word|topic distribution. If the prior is a Dirichlet distribution, the model is called Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003). The graphical model representation for the LDA model is in Figure 2.

Another prior used in topic models is logistic normal (Blei *et al.*, 2005). Different from Dirichlet prior, which only captures variations in each topic's intensity independently, logistic normal prior captures the intuition that some topics are highly correlated and can rise up in intensity together. For example, a document about genetics is more likely to also be about disease than X-ray astronomy. However, logistic normal distribution is not conjugate so it's hard to make inference for the topic model.

3.3 Approximate Inference for LDA

We care about the posterior probability of hidden variables $p(\beta, \theta, z|w)$ in LDA model. Directly calculating the probability is intractable, so we would instead solve $q(\beta, \theta, z)$, a close approximate estimation to the true posterior, which is achieved by approximate inference. The methods used are:

- Variational Inference
  - Mean field approximation (Blei *et al.*)

  - Expectation propagation (Minka *et al.*)
  - Variational 2nd-order Taylor approximation (Xing)

- Markov Chain Monte Carlo
  - Gibbs sampling (Griffiths *et al.*)
  - Stochastic gradient MCMC methods

Here we focus on using **mean field algorithm** for approximate inference.

When we are doing the mean field approximation, we assume the variational approximation q over β, θ, d are independent. Thus we can use the fully factorized distribution:

$$q(\beta, \theta, z) = \prod_k q(\beta_k) \prod_d q(\theta_d) \prod_n q(z_{dn})$$

Note that mean field family usually does NOT include the true posterior.

According to the theorems above, we can develop a coordinate ascent algorithm to find the optimal q. For example, the general update rule for $q(\theta_d)$ is:

$$q(\theta_d) \propto \exp\{\mathbb{E}_{\prod_n q(z_{dn})}[\log(p(\theta_d|\alpha)) + \Sigma_n \log[p(z_{dn}|\theta_d)]\}$$

where d is a document, and n is a term in the document.

In LDA, $p(\theta_d|\alpha)$ is a Dirichlet distribution, and $p(z_{dn}|\theta_d)$ is a Multinomial distribution:

$$p(\theta_d|\alpha) \propto \exp\{\Sigma_{k=1}^K (\alpha_k - 1) \log(\theta_{dk})\}$$

$$p(z_{dn}|\theta_d) \propto \exp\{\Sigma_{k=1}^K 1[z_{dn} = k] \log \theta_{dk}\}$$

where K is the number of topics.

We can now write the coordinate ascent for $q(\theta_d)$ as

$$q(\theta_d) \propto \exp\{\Sigma_{k=1}^K (\Sigma_{n=1}^N q(z_{dn} = k) + \alpha_k - 1) \log \theta_{dk}\}$$

The coordinate ascent algorithm for LDA is as follows:

Initialize variational topics $q(\beta_k)$, k=1,2,...,K;
repeat
  for *each document* $d \in \{1, 2, ..., D\}$ **do**
    Initialize variational topic assignments $q(z_{dn})$, n=1,2,...,N;
    repeat
      Update variational topic proportions $q(\theta_d)$;
      Update variational topic assignments $q(z_{dn})$, n=1,2,...,N;
    until *Change of* $q(\theta_d)$ *is small enough*;
  end
  Update variational topics $q(\beta_k)$, k=1,2,...K;
until *Lower bound* $L(q)$ *converges*;

Algorithm 2: Coordinate Ascent Algorithm for LDA

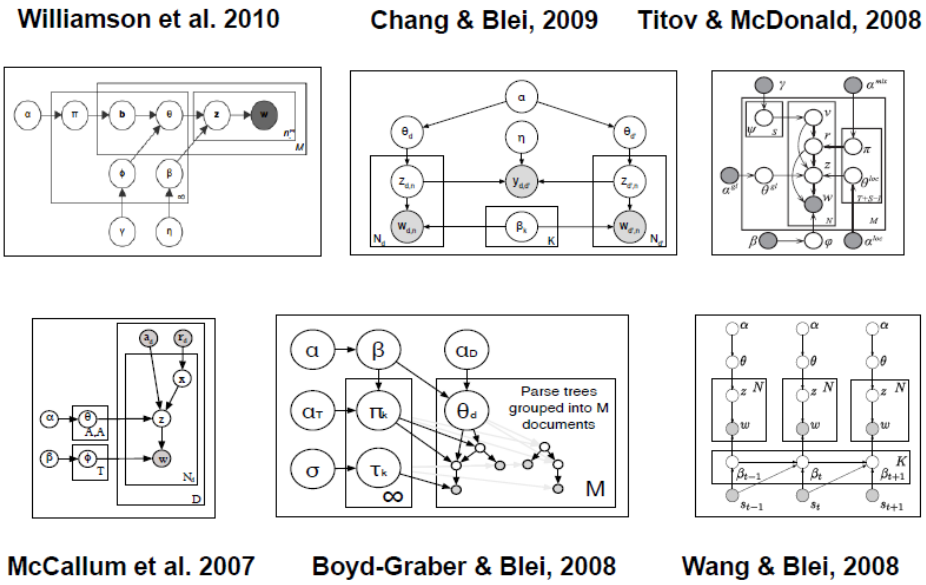


Figure 3: Topic Model zoo

3.4 Conclusion

In this section, we mainly focused on using mean field algorithm to conduct approximate inference on a popular topic model, Latent Dirichlet Allocation. Probabilistic topic models are one of the most active models in the cutting edge research, with several improved models (see Figure 3 for an overview). However, similar approximate inference methods can be applied on these models.

4 Exponential Family Conditionals

In the beginning of the lecture, we asked a question that are there specific forms for the local variational approximations in which we can easily compute closed-form conditional ascent updates? The answer is yes.

Consider a 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.

So, answering the question, 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:

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

where η, h, a, and t are functions that parameterize the exponential family. And different choices of these parameters lead to many popular densities (normal, gamma, exponential, Bernouilli, Dirichlet, categorical, beta, Poisson, geometric, etc.).

We call these **exponential-family-conditional models**, a.k.a **conditionally conjugate models**

There are 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.

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.

5 Mean Field for Markov Random Fields

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

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

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

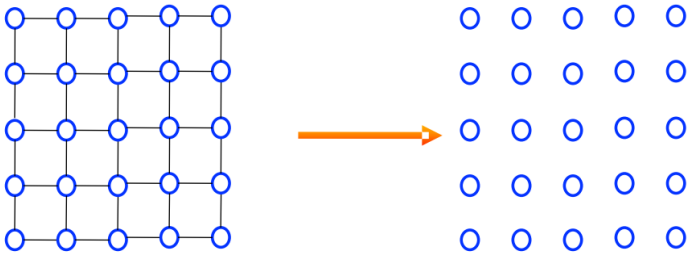


Figure 4: Mean Field for Markov Random Fields

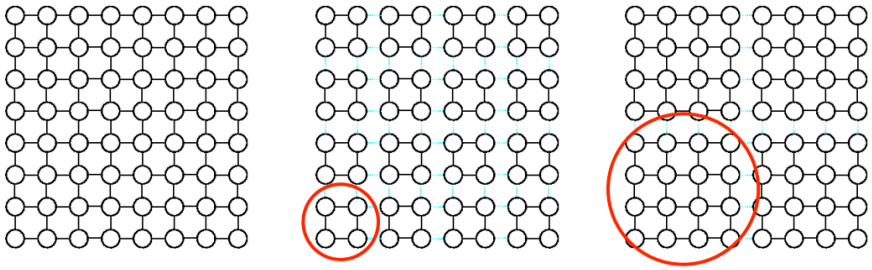


Figure 5: Generalized (Cluster-based) Mean Field for MRFs

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 and Variational Bayesian learning algorithms. It also provides new structured VI approximations to complex models.

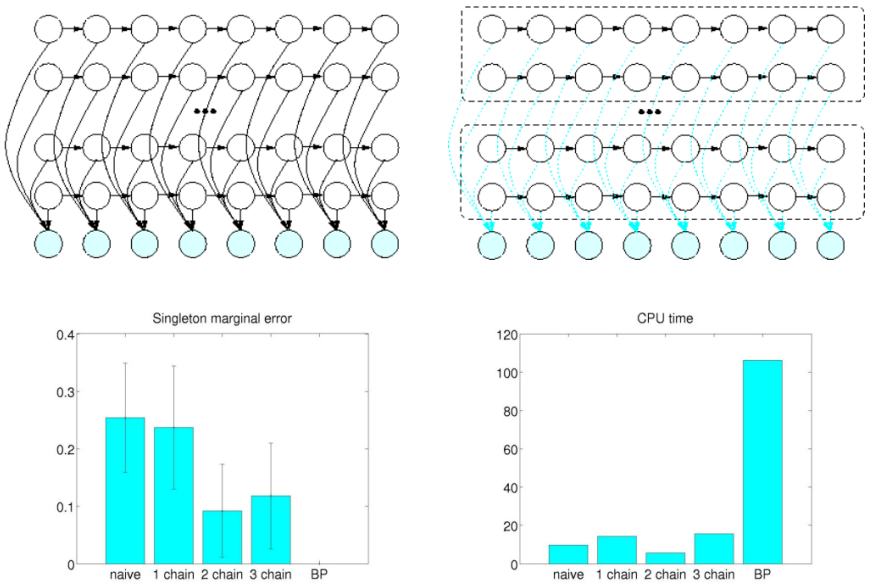


Figure 6: Some Results: Factorial HMMs

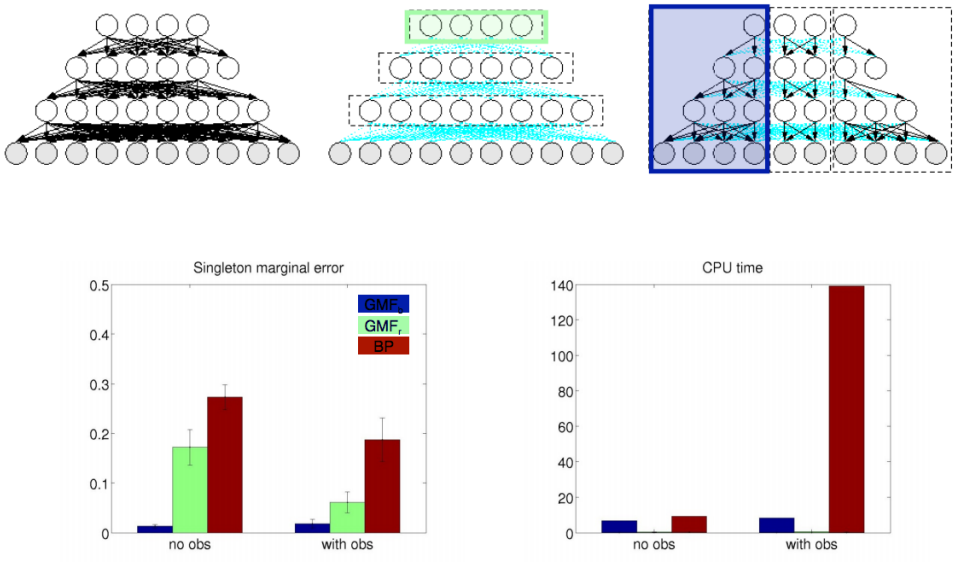


Figure 7: Some Results: Sigmoid Belief Networks

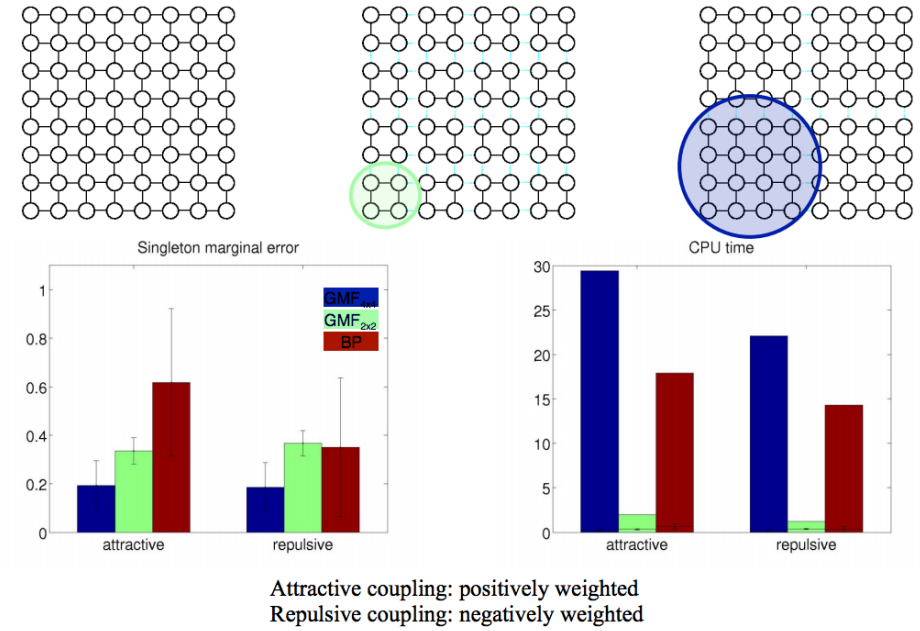


Figure 8: Some Results: Sigmoid Belief Networks