Deterministic Approximation Methods in Bayesian Inference

Tobias Plötz
Department of Computer Science
Technical University of Darmstadt
64289 Darmstadt
t_ploetz@rbg.informatik.tu-darmstadt.de

Abstract

In this seminar paper we give an introduction to the field of deterministic approximate inference. We cast the problem of approximating a posterior distribution over hidden variables as a variational minimization problem. In this framework we describe three algorithms: Variational Factorization, Variational Bounds and Expectation Propagation. We analyze the approximations obtained by the three algorithms in terms of convergence and accuracy. This gives a hint on which method is preferable in a specific problem at hand.

1 Introduction

Bayesian inference methods provide a very powerful and theoretically well-founded toolset to solve various machine learning problems. In a standard scenario we will have observations or data, represented by $\mathcal{D}$, and a probabilistic model that describes the dependencies between the data and unobserved, hidden variables $\mathcal{H}$[1] . Usually, we would like to infer two quantities from this model. The first is the posterior distribution over the hidden variables given the current observations, i.e. $p(\mathcal{H}|\mathcal{D})$, that may, for example, subsequently be used to get a predictive distribution for a new partially observed data point. The second quantity of interest is the model evidence or marginal data likelihood, i.e. $p(\mathcal{D})$, that measures, how well the underlying model explains the observed data and that is usually used for model comparison and averaging. The relation between these quantities can be expressed as follows, using Bayes' theorem:

$$p(\mathcal{H}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{H})p(\mathcal{H})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H})p(\mathcal{H})}{\int p(\mathcal{D}|\mathcal{H})p(\mathcal{H})d\mathcal{H}}$$

So we see that calculating the posterior distribution over the parameters requires a normalization by the model evidence that in turn requires an integration over the hidden variables. In many practical problems it turns out, that this integration might not be feasible. The reason for this might either be the analytical intractability of the problem, i.e. that no closed form solution of the integral exists, or the computational intractability that could arise, for example, if the number of terms in the integrand grows exponentially with the number of observed data points. In these cases approximate inference methods may help.

There are several ways to carry out the approximation. One group of methods replace the posterior distribution by a point estimate of the hidden variables[2]. This group includes the Maximum Likelihood and Maximum A-Posteriori estimates of $\mathcal{H}$, where the point estimate is one of the modes of $p(\mathcal{D}|\mathcal{H})$ and $p(\mathcal{D}|\mathcal{H})p(\mathcal{H})$ respectively. These methods capture just a single point of the true posterior. So they might give a very poor approximation and in case of Maximum Likelihood they

[1] With the term "hidden" variables we denote parameters as well as latent variables.

[2] So the posterior distribution is given in terms of the Dirac measure for this point estimate

are prone to overfitting. Sampling methods provide a stochastic approximation, where the true distribution is approximated by repeatedly drawing independent samples from it. A nice property of sampling is, that it is guaranteed to give exact results after an infinite amount of time. Under practical considerations, however, they might suffer from high computational costs and as the results are stochastic it is hard to tell, when the approximation is good enough. For details on sampling we refer to [Bis06], chapter 11.

This seminar paper deals with a third kind of approximate inference methods that are deterministic, i.e., the results of these methods applied to the same problem is not affected by randomness. These deterministic methods can be much faster than sampling. The general principle is as follows: We want to find a distribution q that is as close as possible to the true posterior. Because any sensible notion of "closeness" would lead to $p(\mathcal{H}|\mathcal{D})$ being the best choice for q and we assume that working with the true posterior is intractable, we have to restrict q in some way. These restriction will have two consequences. The first is, that evaluating q is now tractable and the second is, that the optimal q under this constraint will not be equal to $p(\mathcal{H}|\mathcal{D})$. Ideally, we want to find restrictions on q that render the problem tractable and at the same time lead to an approximation that is as close as possible to the true posterior.

The three methods that are described in this paper differ in the restrictions they impose on q, and in the notion of "closeness", i.e. the optimization objective for finding q. The first method, called Variational Factorization (VF)[3], requires q to factor in a certain way over the hidden variables and then minimizes the Kullback-Leibler (KL) divergence $KL(q\|p)$ from q to the exact posterior p. The second method, called Variational Bounds (VB), gives the restriction on q implicitly by defining a family of lower or upper bounds of the joint distribution $p(\mathcal{H}, \mathcal{D})$ and setting q by normalizing this bound. The integral over the bound with respect to the hidden variables represents a lower or upper bound on the marginal data likelihood and the optimization is done by maximizing and minimizing this bound respectively. The third method, called Expectation Propagation (EP), requires q to factorize like the true posterior and each factor is approximated by a exponential family distribution. In contrast to variational factorization, in EP the reverse KL-divergence $KL(p\|q)$ is minimized. In section 5 we compare the properties of the different approximations obtained by the three methods.

This seminar paper is thought to give a basic understanding of the concepts behind deterministic approximate inference, such that the reader is able to understand more detailed treatments of this topic. Throughout the paper, we will supply further reading pointers as well as references to examples. The line of argumentation is strongly aligned to the one in [Bis06], but we fill in some gaps in the derivations.

2 Variational Factorization

In this section, we seek a variational[4] distribution $q(\mathcal{H})$ that minimizes the Kullback Leibler divergence $KL(q\|p)$ between q and the exact posterior $p(\mathcal{H}|\mathcal{D})$, i.e., we want to solve the variational optimization problem:

$$\min_q(KL(q\|p)) \tag{1}$$

where the Kullback Leibler divergence is given by

$$KL(q\|p) = -\int q(x) \ln\left(\frac{p(x)}{q(x)}\right) dx$$

If we place no constraints on q, the solution to (1) would be the exact posterior as $KL(q\|p) = 0$ if and only if $q = p$. This solution would not help, since we assume that evaluating the exact posterior is intractable. Instead, we restrict the form of q. In this section we require q to factorize in n different distributions $q_1, ..., q_n$, i.e.

$$q(\mathcal{H}) = q_1(\mathcal{H}_1)q_2(\mathcal{H}_2)...q_n(\mathcal{H}_n) \tag{2}$$

for a partition $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_n$ of the hidden variables $\mathcal{H}$. A common choice is, to require a factorisation between the latent variables and the parameters of a model. This factorization introduces

[3]This method is also often referred to as "Variational Bayes".

[4]The term *variational* comes from the field of *calculus of variations*. A variational function can be seen as a placeholder for functions just as a variable is a placeholder for scalars.

artificial independence constraints to the model as in general the latent variables and the parameters of a model are not independent given the data. We could now try to directly solve this constrained variational optimization problem, but instead we have a look at how minimizing the KL divergence $KL(q\|p)$ can be interpreted as maximizing a lower bound on the log marginal data likelihood $\ln(p(\mathcal{D}))$. We first note, that for an arbitrary probability distribution q the following decomposition of $\ln(p(\mathcal{D}))$ holds:

$$\ln(p(\mathcal{D})) = \ln(p(\mathcal{D})) \int q(\mathcal{H}) d\mathcal{H} = \int q(\mathcal{H}) \ln(p(\mathcal{D})) d\mathcal{H} = \int q(\mathcal{H}) \ln\left(\frac{p(\mathcal{D},\mathcal{H}) q(\mathcal{H})}{p(\mathcal{H}|\mathcal{D}) q(\mathcal{H})}\right) d\mathcal{H}$$

$$= \int q(\mathcal{H}) \ln\left(\frac{p(\mathcal{D},\mathcal{H})}{q(\mathcal{H})}\right) d\mathcal{H} - \int q(\mathcal{H}) \ln\left(\frac{p(\mathcal{H}|\mathcal{D})}{q(\mathcal{H})}\right) d\mathcal{H} = \mathcal{L}(q) + KL(q\|p)$$

where we have defined

$$\mathcal{L}(q) = \int q(\mathcal{H}) \ln\left(\frac{p(\mathcal{D},\mathcal{H})}{q(\mathcal{H})}\right) d\mathcal{H}$$

In the above decomposition, the second term $KL(q\|p)$ is the KL divergence from q to the exact posterior $p(\mathcal{H}|\mathcal{D})$, i.e. the optimization objective of this section. As the KL divergence is always non-negative, the first term $\mathcal{L}(q)$ is a lower bound for $\ln(p(\mathcal{D}))$. To get the tightest lower bound, we can maximize the first term $\mathcal{L}(q)$ which is also equivalent to minimizing $KL(q\|p)$. The lower bound will become an exact bound if we choose the exact posterior for our variational distribution q, as the KL divergence will vanish in this case. Again, this would not lead to a simplified problem. We now constrain q to factorize like in (2). Then we can write $\mathcal{L}(q)$ in the following way (where we leave out the parameters of the q_i for readability):

$$\mathcal{L}(q) = \int \left(q_j \prod_{i \neq j}^{n} q_i\right) \left(\ln(p(\mathcal{D},\mathcal{H})) - \ln(q_j) - \sum_{i \neq j}^{n} \ln(q_i)\right) d\mathcal{H} \tag{3}$$

There is no known solution of how to maximize this expression in terms of all q_i at the same time. So instead we pick out one particular q_j and analyze, how $\mathcal{L}(q)$ can be maximized with respect to q_j when we assume, that the other $q_i, i \neq j$ are kept fixed. Therefore, we single out the terms that depend on q_j, and equation (3) will look like this:

$$\int \left(q_j \prod_{i \neq j}^{n} q_i\right) \ln(p(\mathcal{D},\mathcal{H})) d\mathcal{H} - \int \left(q_j \prod_{i \neq j}^{n} q_i\right) \ln(q_j) d\mathcal{H} - \int \left(q_j \prod_{i \neq j}^{n} q_i\right) \sum_{i \neq j}^{n} \ln(q_i) d\mathcal{H} \tag{4}$$

The second and the third integral can now be simplified by noting that the integrands consist of densities over the $\mathcal{H}_i, i \neq j$ and $\mathcal{H}_j$ respectively, multiplied with a factor that does not depend on all of the hidden variables. So we can split the integrals over $\mathcal{H}$ into several integrals and make use of the fact that integrals over the corresponding normalized densities q_i equate to one:

$$\int \left(q_j \prod_{i \neq j}^{n} q_i\right) \ln(q_j) d\mathcal{H} = \int \left\{\int \prod_{i \neq j}^{n} q_i d\mathcal{H}_i\right\} q_j \ln(q_j) d\mathcal{H}_j = \int q_j \ln(q_j) d\mathcal{H}_j$$

$$\int \left(q_j \prod_{i \neq j}^{n} q_i\right) \sum_{i \neq j}^{n} \ln(q_i) d\mathcal{H} = \int \left\{\int q_j d\mathcal{H}_j\right\} \prod_{i \neq j}^{n} q_i \sum_{i \neq j}^{n} \ln(q_i) d\mathcal{H}_i = \int \prod_{i \neq j}^{n} q_i \sum_{i \neq j}^{n} \ln(q_i) d\mathcal{H}_i$$

Note that the last equation is independent of q_j so we can absorb it into a constant term. We also rewrite the first integral of equation (4) by splitting up the integration over $\mathcal{H}$.

$$\int \left(q_j \prod_{i \neq j}^{n} q_i\right) \ln(p(\mathcal{D},\mathcal{H})) d\mathcal{H} = \int q_j \left\{\int \prod_{i \neq j}^{n} q_i \ln(p(\mathcal{D},\mathcal{H})) d\mathcal{H}_i\right\} d\mathcal{H}_j$$

$$= \int q_j \mathbb{E}_{i \neq j}[\ln(p(\mathcal{D},\mathcal{H}))] d\mathcal{H}_j$$

Here $\mathbb{E}_{i \neq j}$ denotes an expectation taken over all $\mathcal{H}_{i \neq j}$. Now we put the results back into equation (4) and obtain

$$\mathcal{L}(q) = \int q_j \mathbb{E}_{i \neq j}[\ln(p(\mathcal{D},\mathcal{H}))] d\mathcal{H}_j - \int q_j \ln(q_j) d\mathcal{H}_j + \text{const} \tag{5}$$

The first two terms in the above expression for the lower bound look very similar to a negative KL divergence for which we would know what the optimal q_j should look like. But as $\mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))]$ is not necessarily a probability distribution, we first have to normalize it in order to reformulate (5) in terms of a negative KL divergence. We therefore introduce a new distribution $\tilde{p}(\mathcal{D},\mathcal{H}_j)$ that is just the normalized $\mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))]$, i.e. the log of $\tilde{p}(\mathcal{D},\mathcal{H}_j)$ satisfies the following property[5]:

$$\ln(\tilde{p}(\mathcal{D},\mathcal{H}_j)) = \mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))] - \ln \int \mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))]d\mathcal{H}_j = \mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))] + \text{const} \tag{6}$$

If we plug this into equation (5) we get:

$$\mathcal{L}(q) = \int q_j \ln \frac{\tilde{p}(\mathcal{D},\mathcal{H}_j)}{q_j} d\mathcal{H}_j + \text{const} = -\text{KL}(q_j, \tilde{p}(\mathcal{D},\mathcal{H}_j)) + \text{const} \tag{7}$$

So we see that maximizing $\mathcal{L}(q)$ is equal to minimizing $\text{KL}(q_j, \tilde{p}(\mathcal{D},\mathcal{H}_j))$ and hence the optimal q_j^* is equal to $\tilde{p}(\mathcal{D},\mathcal{H}_j)$. Thus we get two expressions for the optimal q_j^* and the log of q_j^*:

$$\ln(q_j^*) = \ln(\tilde{p}(\mathcal{D},\mathcal{H}_j)) = \mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))] + \text{const} \tag{8}$$

$$q_j^* = \tilde{p}(\mathcal{D},\mathcal{H}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))])}{\int \exp(\mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))])d\mathcal{H}_j} \tag{9}$$

Note that the optimal solution q_j^* still depends on the other $q_{i\neq j}$ through their expectations. A general insight is, that if for some j,k the hidden variables $\mathcal{H}_j$ and $\mathcal{H}_k$ are independent in the original model $p(\mathcal{H},\mathcal{D})$ then this independence is carried over to the solution for q_j^* as the expectation with respect to q_k can be absorbed into the constant term and thus q_j^* will not depend on q_k.

We can apply the result (8) to all $q_j, 1 \leq j \leq n$ to get a set of n equations that must be fulfilled at a local maximum of $\mathcal{L}(q)$. We can now start with some initial distributions for the q_i and update each q_i in turn while holding the others fixed. Each update to a specific q_i will therefore increase the lower bound $\mathcal{L}(q)$ unless the approximation is already at a local maximum. As $\mathcal{L}(q)$ is bounded from above by the constant $\ln p(\mathcal{D})$ the algorithm is guaranteed to converge.

In order to take the expectations in the update equations we need the functional form of the factor approximations. If we would just choose them arbitrarily it can well happen that it gets intractable to calculate the normalization constants after a few iterations. So for the success of this variational method it is *crucial* that the functional form of the approximating distribution are invariant under the update equations. Sometimes it is possible to derive the functional form of the q_j by simplification of the update equations (8), but this may not always be the case. Even if the functional form of the approximations can be derived, it may still be intractable to take the required expectations. In this case the method of variational lower bounds, that is described in the next section, might help.

We now sum up the Variational Factorization method. In the first step, you have to choose a factorization over the hidden variables for the approximation q. As a second step, the equation (8) for each q_i has to be analyzed in order to derive the functional form of q_i. To actually optimize the q_i, you first initialize them. Afterwards, all factors are refined iteratively until some convergence criterion is met, e.g. the lower bound $\mathcal{L}(q)$ does not get increased any further.

Examples and further readings The Bishop book [Bis06] exemplifies the application of the Variational Factorization method in the scenario of Gaussian Mixture Models and Linear Regression. For Gaussian Mixture Models a factorization between latent variables and parameters is assumed and from equation (8) the functional form of both factors is derived. It is then shown, that evaluating the predictive density with the approximated posterior is tractable and that the lower bound on the marginal data likelihood can be used to determine the number of mixture components, something that is not possible with an EM treatment of the Gaussian mixture model.

In [Bea03] it is shown, that for a wide class of models, called the Conjugate Exponential Models, by assuming a factorization between the parameters, the functional form of the approximation will again be conjugate exponential, thus leading to simple update equations for the factors. In [Win04]

[5] It is not obvious, why this normalization should in general be possible, i.e., why $\mathbb{E}_{i\neq j}[\ln(p(\mathcal{D},\mathcal{H}))]$ should be bounded. Neither in [Bis06] nor in other references that make use of the factorized approximation a proof could be found to that claim.

the Variational Message Passing algorithm is presented as an application of the Variational Factorization framework to conjugate exponential graphical models. The assumed factorization between all the nodes of the model leads to a local message passing algorithm.

3 Variational bounds

The method in the previous section was based on minimizing the KL divergence between the approximated distribution q and the true posterior p. This leads to the lower bound $\mathcal{L}(q)$ on the log marginal data likelihood $\ln p(\mathcal{D})$. In this section we start with generally considering bounds on $p(\mathcal{D})$. Our goal is to seek a function $g(\mathcal{H}, \mathcal{D})$ that can tractably be integrated and that is a bound for the joint distribution, when holding the data fixed. Thus the integral over g will be a bound for the integral of the joint distribution over the hidden variables, i.e. the marginal data likelihood. This technique is especially useful when the joint distribution factors in a certain way, as we can replace individual factors, one by one, with either only lower or only upper bounds until we get a bound that can conveniently be integrated. By normalizing the bound, we also get an approximation to the posterior. In general, the bounds are associated with free parameters that are to be optimized in order to maximize the lower bound of the marginal likelihood or to minimize the upper bound, thus getting the tightest possible bound.

Convex duality One method to find bounds for factors or functions in general is called *convex duality*. To illustrate the general concept of convex duality, we will have a look at the log function, which is non-linear in its input parameter. Because $\ln(x)$ is concave, for every x we can find a line that is an upper bound of $\ln(x)$ and touches it at the point x, i.e., we can find a tangent. Simple math shows us, that the tangent line for a point x is given by

$$y(t) = \frac{1}{x}t + \ln(x) - 1 \tag{10}$$

$$y(t) = \eta t - \ln(\eta) - 1 \tag{11}$$

where, in the last line, we substituted with $\eta = x^{-1}$. Equation (10) denotes a tangent line for a specific x and as the log function is concave, the tangent line is an upper bound for the whole log function. Hence, by letting η vary in equation (11) we get a family of upper bounds for the log function as each η defines a tangent. Because the reverse also holds, i.e. every tangent line can be represented like (11) for some η, we see that for every x there is an η for which the bound is exact at x, i.e. $\ln(x) = y(x)$. Therefore, we can define the log function as a pointwise minimum over all tangent lines:

$$\ln(x) = \min_{\eta}(\eta x - \ln(\eta) - 1) \tag{12}$$

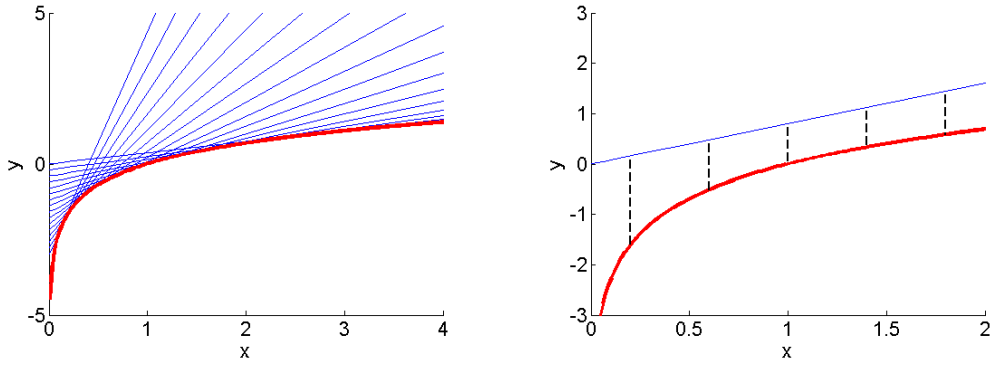
So we can upper bound the non-linear log function by a linear function, at the cost of introducing an additional free parameter η that can be optimized for every point x to give an exact bound. Figure 1a shows the log function as well as the upper bound for some choices of η. Note that every bound is exact at one point x^* and that the tightness of that bound within a small region of x^* depends on the magnitude of the derivative of $\ln(x)$ at that point. In the literature the intercept, written as a function $g(\eta)$ of η, is called the *conjugate* or *dual* function and in fact, any concave function $f(x)$ can be written as

$$f(x) = \min_{\eta}(\eta x - g(\eta)) \tag{13}$$

The dual function g is itself a concave function and can thus be written in terms of f:

$$g(\eta) = \min_{x}(\eta x - f(x)) \tag{14}$$

Figure 1b illustrates why the last relation holds. If we want the line ηx to become a tangent, we have to move it by the minimal difference from the log function to the line. As the sum of the convex function ηx and the convex function $-f(x)$ is again convex, the term $\eta x - f(x)$ has a unique minimum. Equation (14) also implies that we can determine g by setting the derivative of f to η in case that f is differentiable. In case of $f(x)$ being convex instead of concave the above relations will again hold when taking maxima instead of minima. Additionally, these relations generalize to vector valued x. In that case η will also be vector valued.



(a) (b)

Figure 1: (a): The log function as well as several upper bounds for different settings for η. (b): The intercept of the tangent line of a specific slope is given by the minimum distance from the log function to the line through the origin.

One powerful application of the concept of convex duality is bounding the sigmoid function, which plays an important part in machine learning, e.g. in logistic regression. The sigmoid function is given by

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Taking products of sigmoid terms is inconvenient, because the number of terms in the result will grow exponentially. Our goal is to derive lower and upper bounds of the sigmoid function that can nicely be multiplied together and then form a lower and upper bound on the product respectively. First, we notice, that $\sigma(x)$ is neither convex nor concave, but log concave, i.e. $\ln(\sigma(x))$ is a concave function[6]. We get the dual function $g(\eta)$ by setting the derivative of $\ln(\sigma(x))$ to η and some lines of math give the following result

$$g(\eta) = -\eta \ln(\eta) - (1 - \eta) \ln(1 - \eta) = H(\eta)$$

where $H(\eta)$ is the entropy of η. So for all $\eta \in (0, 1)$ we have[7]

$$\ln(\sigma(x)) \leq \eta x - H(\eta) \Rightarrow \sigma(x) \leq \exp(\eta x - H(\eta))$$

Figure 2a shows the sigmoid function and some variational upper bounds. As in the case with the log function, each upper bound touches the sigmoid function in exactly one point. The nice property of this upper bound is, that it is an exponential of a linear function and thus several upper bounds can nicely be multiplied together to give again an exponential of a linear function. It is also possible to derive a lower bound for the sigmoid function and a derivation can be found in [Bis06]. Note that in general, it might not be easy to find a suitable transformation of the input function such that, on the one hand, the transformed function will be convex or concave and, on the other hand, the bound will have the desired algebraic properties that simplify the integration.

To put the previous discussion into the bigger picture: Convex bounds are one way to approximate individual factors of the joint distribution by lower or upper bounds. This leads to a bound for the whole joint distribution and it should have the property that integration is now tractable. Assume that we bounded the joint distribution by a lower bound g that depends on free parameters $\boldsymbol{\eta}$

$$\forall \boldsymbol{\eta} : p(\mathcal{H}, \mathcal{D}) \geq g(\boldsymbol{\eta}, \mathcal{H}, \mathcal{D}) \tag{15}$$

Integrating out the hidden variables will therefore give a lower bound on the log marginal data likelihood

$$\ln p(\mathcal{D}) = \ln \int p(\mathcal{H}, \mathcal{D}) d\mathcal{H} \geq \ln \int g(\boldsymbol{\eta}, \mathcal{H}, \mathcal{D}) d\mathcal{H} = \mathcal{L}(\boldsymbol{\eta}) \tag{16}$$

[6]That can be seen by observing that the second derivative $-\frac{\exp(x)}{(1+\exp(x))^2}$ is always negative.

[7]Here is a small caveat: For the relation (13) to hold, it is necessary that we restrict η to "nice" values, i.e. to values where η could in fact represent the slope of a tangent line of f.

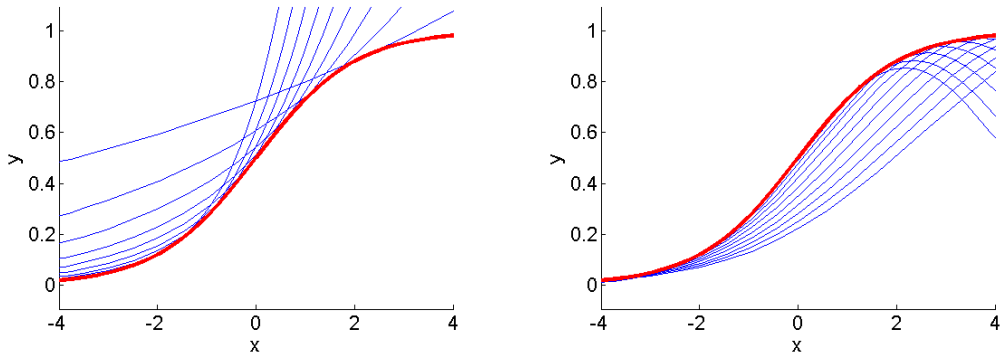


Figure 2: The sigmoid function and variational upper bounds (a) as well as lower bounds (b)

We can maximize $\mathcal{L}(\boldsymbol{\eta})$ with respect to the free parameters $\boldsymbol{\eta}$ to get optimal parameters $\boldsymbol{\eta}^*$. The parameters $\boldsymbol{\eta}^*$ can be used to evaluate $\mathcal{L}(\boldsymbol{\eta}^*)$ as an approximation to $p(\mathcal{D})$. Furthermore we can get an approximation q to the true posterior by normalizing $g(\boldsymbol{\eta}^*)$ with respect to the hidden variables:

$$q(\mathcal{H}) = \frac{g(\boldsymbol{\eta}^*, \mathcal{D}, \mathcal{H})}{\int g(\boldsymbol{\eta}^*, \mathcal{D}, \mathcal{H}) d\mathcal{H}} = \frac{g(\boldsymbol{\eta}^*)}{\mathcal{L}(\boldsymbol{\eta}^*)} \tag{17}$$

To see why this is a good approximation, consider the case where the marginal data likelihood is bounded exactly, i.e. $p(\mathcal{D}) = \mathcal{L}(\boldsymbol{\eta}^*)$. From basic properties of the integral we know, that for two functions f and h with $f \leq h$ it holds that $\int f = \int g$ if and only if $f = g$ almost everywhere. Therefore we see that $p(\mathcal{H}, \mathcal{D}) = g(\boldsymbol{\eta}^*, \mathcal{H}, \mathcal{D})$ almost everywhere and thus q would equal $p(\mathcal{H}, \mathcal{D})$ almost everywhere. Note that in the previous section about variational factorization we also maximized a lower bound on the log marginal data likelihood. But with general lower bounds we cannot interpret the difference between the bound and the exact value as a KL divergence.

For the optimization of the parameters $\boldsymbol{\eta}$ there are basically two different ways. The first is to solve the integral for $\mathcal{L}(\boldsymbol{\eta})$ analytically to get a closed form expression. Now a suitable optimization algorithm can be used to maximize $\mathcal{L}(\boldsymbol{\eta})$, e.g. conjugate gradient descent, Newton-Raphson, etc. The second approach uses EM as a general optimization technique for monotone functions of integrals to iteratively maximize the lower bound. In our case we can apply Jensen's inequality to get a lower bound on the lower bound $\mathcal{L}(\boldsymbol{\eta})$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\eta}) &= \ln \int g(\boldsymbol{\eta}, \mathcal{H}, \mathcal{D}) d\mathcal{H} = \ln \int \frac{g(\boldsymbol{\eta}, \mathcal{H}, \mathcal{D})}{q(\mathcal{H})} q(\mathcal{H}) d\mathcal{H} \\
&\geq \int \ln \left(\frac{g(\boldsymbol{\eta}, \mathcal{H}, \mathcal{D})}{q(\mathcal{H})} \right) q(\mathcal{H}) d\mathcal{H} = \int \ln \left(g(\boldsymbol{\eta}, \mathcal{H}, \mathcal{D}) \right) q(\mathcal{H}) d\mathcal{H} - \text{const}
\end{aligned}$$

where q depends on an old estimate of the parameters $\boldsymbol{\eta}^{old}$ and is defined by equation (17) for these parameters. Here, $g(\boldsymbol{\eta}, \mathcal{H}, \mathcal{D})$ takes the role of the complete-data likelihood in classical EM and thus we alternate between the E step, where we compute the expectation of the complete-data log likelihood with respect to $q(\mathcal{H})$, and the M step, where we maximize this expectation with respect to the parameters $\boldsymbol{\eta}$ that then become the new initial setting $\boldsymbol{\eta}^{old}$ for the next iteration. As in classical EM, this technique is espacially useful when pulling the logarithm inside the integral simplifies the integration, for instance because the bound is an exponential of a simple function. Every step of the EM algorithm will increase the lower bound, thus convergence to a local maximum is guaranteed.

Using lower bounds over upper bounds has the advantage that it is possible to combine lower bounds with the Variational Factorization methodology of the previous section. In VF, we get the lower bound $\mathcal{L}(q)$ on $\ln p(\mathcal{D})$. Evaluating the update equations for a single q_j requires taking expectations with respect to the other $q_i, i \neq j$. That may still be intractable and variational lower bounds can be used to get a lower bound on $\mathcal{L}(q)$ such that maximization is now tractable.

Readings and examples Jordan et. al give a great introduction to Variational Bounds [JGJS99]. They show how the factors of the joint distribution can be bounded incrementally to avoid bounds that are too loose. Also the connection between Variational Bounds and Variational Factorization is analyzed in more depth than we have made it here. They apply these bounds to different graphical models and give many references for further reading. In [Bis06] variational bounds are combined with the factorization approximation of the previous section to evaluate an approximate predictive distribution for Bayesian Logistic Regression.

In [WG07] variational upper bounds are used for approximate inference in undirected graphical models. They do not use the principle of convex duality but apply the weighted power means inequality and the arithmetic-geometric means inequality to replace the intractable product over potential functions with a sum of tractable products over approximate potential functions.

4 Expectation Propagation

In section 2 we started our discussion by considering variational distributions $q(\mathcal{H})$ that minimize the KL divergence $KL(q\|p)$ from q to the exact posterior $p(\mathcal{H}|\mathcal{D})$. In some sense, this is just the wrong direction of the KL divergence, as $KL(q\|p)$ intuitively measures, how surprised you are when you get a sample from distribution q, given that you assume p to be the true distribution. In this light it seems more natural to seek a distribution q that minimizes the other KL divergence, i.e. $KL(p\|q)$, as we actually do not want to be surprised to get a sample from the true distribution p, if we assume our approximation q to be correct.

Minimizing $KL(p\|q)$ by moment matching We now show that minimizing $KL(q\|p)$, under the constraint that q lies in the exponential family, can be done by matching the expectations of the sufficient statistic of q. Therefore, let q be of the form

$$q(\mathcal{H}) = g(\boldsymbol{\eta})h(\mathcal{H})\exp(\boldsymbol{\eta}^T u(\mathcal{H})) \tag{18}$$

where $\boldsymbol{\eta}$ denote the natural parameters of q and $u(\mathcal{H})$ denote the sufficient statistic of q (e.g. for Gaussian q, we have $u(\mathcal{H}) = [\mathcal{H}, \mathcal{H}\mathcal{H}^T]^T$). If we express $KL(p\|q)$ as a function of $\boldsymbol{\eta}$ we get the following

$$\begin{aligned} KL(p\|q) &= -\int p(\mathcal{H}|\mathcal{D}) \ln\left(\frac{q(\mathcal{H})}{p(\mathcal{H}|\mathcal{D})}\right) d\mathcal{H} = -\int p(\mathcal{H}|\mathcal{D}) \ln q(\mathcal{H}) d\mathcal{H} + \text{const} \\ &= -\mathbb{E}_{p(\mathcal{H}|\mathcal{D})}(\ln g(\boldsymbol{\eta})) - \mathbb{E}_{p(\mathcal{H}|\mathcal{D})}(\ln h(\mathcal{H})) - \mathbb{E}_{p(\mathcal{H}|\mathcal{D})}(\boldsymbol{\eta}^T u(\mathcal{H})) + \text{const} \\ &= -\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T \mathbb{E}_{p(\mathcal{H}|\mathcal{D})}(u(\mathcal{H})) + \text{const} \end{aligned}$$

Taking the derivative with respect to $\boldsymbol{\eta}$ and setting it to zero results in the following equation

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}_{p(\mathcal{H}|\mathcal{D})}(u(\mathcal{H}))$$

It can be shown, that the negative gradient of $\ln g(\boldsymbol{\eta})$ is given by the expectation of the sufficient statistic under the distribution $q(\mathcal{H})$ (cf. [Bis06] chapter 2.4.1). So if we restrict q to be from the exponential family we get the following condition for the q that minimizes $KL(p\|q)$:

$$\mathbb{E}_{q(\mathcal{H})}(u(\mathcal{H})) = \mathbb{E}_{p(\mathcal{H}|\mathcal{D})}(u(\mathcal{H})) \tag{19}$$

This procedure of equating the expectations of the sufficient statistics is called *moment matching* and for a Gaussian q this boils down to setting the mean parameter to the mean of $p(\mathcal{H}|\mathcal{D})$ and setting the variance parameter to the variance of $p(\mathcal{H}|\mathcal{D})$.

Expectation Propagation We shall assume that calculating the moments for the whole true posterior is not tractable. We now write the posterior as a product over N factors:

$$p(\mathcal{H}|\mathcal{D}) = \frac{1}{p(\mathcal{D})}\prod_{i=1}^{N} f_i(\mathcal{H}_i)\ ,\ \mathcal{H}_i \subseteq \mathcal{H} \tag{20}$$

Note that the individual factors are not necessarily probability distributions and that the $\mathcal{H}_i$ need not to be disjoint. In fact the above representation might also not be unique[8]. For notational convenience

[8]This factor model can be derived from so called factor graphs, that are a generalization of Bayesian Networks and undirected graphical models.

we drop the index from the $\mathcal{H}_i$ from now on. In Expectation Propagation we want to approximate each factor f_i by a factor approximation $\tilde{f}_i$, where we restrict each $\tilde{f}_i$ to be from the exponential family. As this family is closed under multiplication, the resulting approximation q is also from the exponential family and can be written as

$$q(\mathcal{H}) = \frac{1}{Z} \prod_{i=1}^{N} \tilde{f}_i(\mathcal{H}) \tag{21}$$

Here, Z is a normalization constant that ensures that the product of the factor approximations is a valid probability distribution. So how do we find a good approximation for the factors f_i? In EP we try to find those factor approximations by refining each $\tilde{f}_j$ in the context of all the other current $\tilde{f}_i, i \neq j$. We first show, how this works in detail and then give a short intuition, why this is sensible.

As a first step we initialize each $\tilde{f}_i$ and set the initial q to the normalized product of these factors. Then we iteratively pick a factor $\tilde{f}_j$ that we wish to refine and define an unnormalized distribution $q^{\setminus j}$ that we get by dividing our current approximation q by $\tilde{f}_j$:

$$q^{\setminus j}(\mathcal{H}) = \frac{q(\mathcal{H})}{\tilde{f}_j(\mathcal{H})} \tag{22}$$

Now we minimize the KL divergence between f_j and $\tilde{f}_j$, both multiplied with the current estimate of the other factor approximations $q^{\setminus j}$. We get a new approximation q^{new} of the whole posterior by minimizing:

$$\mathrm{KL}\left(\frac{1}{Z_j} f_j q^{\setminus j} \,\middle\|\, q^{new}\right) = \mathrm{KL}\left(\frac{1}{Z_j} f_j q^{\setminus j} \,\middle\|\, \tilde{f}_j q^{\setminus j}\right) \tag{23}$$

for the appropriate normalization constant Z_j. The solution to that minimization problem is given by moment matching as explained above. We assume that calculating the moments of the left-hand side of the above KL divergence is a tractable operation, but in practice this operation is the hard part in implementing Expectation Propagation for a specific problem[9]. Now we calculate the refined version of $\tilde{f}_j$. From q^{new} we get the new $\tilde{f}_j$ by simply dividing out the other factors $q^{\setminus j}$ and scaling by the above normalization constant Z_j:

$$\tilde{f}_j = Z_j \frac{q^{new}}{q^{\setminus j}} \tag{24}$$

Now q is set to q^{new} for the next iteration where a new factor gets refined. These updates are iterated and several passes through all factors can be made, until some convergence criterion is met, e.g. that the changes of the parameters of the single $\tilde{f}_i$ during a whole pass through the factors has not exceeded some threshold. At the end, the last q^{new} gives an approximation to the posterior and we can evaluate an approximation to the model evidence by calculating the normalization constant. To sum up the algorithm, in EP you do the following:

1. Initialize each $\tilde{f}_i$ and set $q \propto \prod \tilde{f}_i$
2. Repeat until convergence
   (a) Pick a $\tilde{f}_j$ to refine and calculate $q^{\setminus j} = \frac{q}{\tilde{f}_i}$ as well as the normalization constant $Z_j = \int f_j q^{\setminus j} d\mathcal{H}$
   (b) Calculate q^{new} by matching the expected sufficient statistic of $f_j q^{\setminus j}$
   (c) Set $\tilde{f}_j = Z_j \frac{q^{new}}{q^{\setminus j}}$ and $q = q^{new}$

We now give an intuition, why it is sensible to approximate each factor f_j by $\tilde{f}_j$ in the context of all other factor approximations $q^{\setminus j}$. Figure 3 shows two plots of factors that occur in the so called "clutter" problem [Bis06], as well as the factor approximation and the context. The context, i.e. the product of all other factor approximations, controls, which parts of the true factor f_j must

[9] In Tom Minka's video lecture [Min09] about Expectation Propagation this moment matching operation is called the "proj" operator

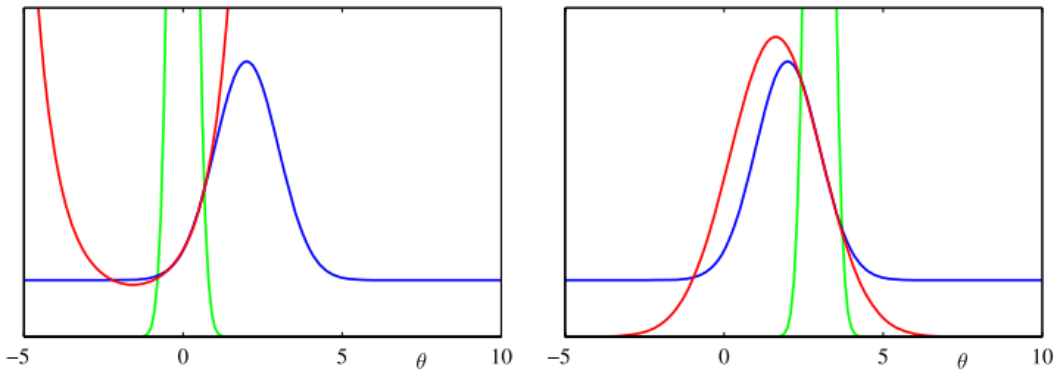


Figure 3: Both plots show a factor f_i of the clutter problem in blue and the context $q^{\setminus j}$ in green. The updated approximation $\tilde{f}_j$ (red) for the factor will be accurate where the context has considerable mass. Figure taken from [Bis06]

be approximated accurately by the factor approximation $\tilde{f}_j$. Ideally we would like to get factor approximations that lead to an overall approximation q that is as close as possible to the best q^* given by matching moments of the true posterior. Now assume that the context is already close to q^* divided by the current factor under consideration. Multiplying the true factor to the context will give a very similar result as multiplying the refined factor approximation to the context as the factor approximation is accurate in the region where the context (and therfore $q^* \setminus \tilde{f}_j$) has mass. Hence the overall approximation will be very close to q^*. The downside is, that there are no guarantees that the algorithm will converge in such a way. It is even not guaranteed that EP will converge at all. In practice, however, EP with unimodal approximations is found to converge very fast to very good solutions if the true posterior has a unimodal shape. Otherwise minimizing $KL(p\|q)$ will cause q to average over all modes, leading to a poor approximation.

Readings and examples A detailed introduction to EP can be found in [Min01b] and in [Bis06]. In the former it is shown that EP generalizes other well known approximate inference algorithms like Assumed Density Filtering, the Kalman Smoother and Loopy Belief Propagation, and it is explained, how EP can be applied to generic factor graphs. Another, nicely comprehensible introduction to EP can be found at [Min09]. As a general hub for information on the diverse applications of EP we refer to [Min07]. In [SN11] an extension to EP is presented that is both provably convergent and very fast.

5 Properties of the Three Approximation Methods

In this section we will compare the different variational methods in terms of convergence, properties of the resulting distribution that arise from the choice of the minimization objective and in terms of the accuracy of the approximation.

Convergence We have seen, that in Variational Factorization as well as Variational Bounds a lower bound on the log marginal data likelihood is maximized. In VF each refinement of the factors will cause this lower bound to increase and thus the algorithm will converge. The same holds for VB, where we can either directly optimize the lower bound or employ an EM algorithm, in which the bound is consistently increased. For Expectation Propagation it can be shown, that if the factor approximations belong to the exponential family then there are fix points for the EP iterations [Min01a], but the algorithm does not necessarily converge to one of these fix points. Furthermore, it can be shown, that fix points of EP are associated with local minima of a specific error function. It is possible to minimize this error function directly. By doing so, however, one would lose the fast convergence properties of EP for well suited problems. In practice, EP with unimodal approximations will converge very fast to a good solution, if the true posterior is unimodal or very close to unimodal. According to Minka [Min01a] the opposite also holds, i.e. if EP does not converge, then

the true posterior has a multimodal shape and hence approximating it with EP might lead to a poor solution anyway, but this claim is disputed [HZ02].

Minimization Objective In VF we seek a distribution q that minimizes the KL-divergence $KL(q\|p)$ to the true posterior while in Expectation Propagation the opposite KL-divergence $KL(p\|q)$ is the minimization objective. The choice of the minimization objective naturally has implications for the resulting approximation.

Here we reproduce the formula for the KL-divergence to take a closer look at it:

$$KL(q\|p) = -\int q(x) \ln\left(\frac{p(x)}{q(x)}\right) dx$$

Assume that in some region R the true distribution p is almost zero. If q would have any considerable mass in R this would lead to a high value of the integrand as the logarithm will evaluate to a high value which is weighted by some $q(x)$ that is not almost zero. In this case, the KL divergence would be big. In order to minimize the KL divergence, q must be chosen such that it does not put mass in regions where p does not have mass. In the extreme case this means, that if p is zero in a certain region, minimizing $KL(q\|p)$ will require q to also be zero in this region. This property is known as the *zero forcing* property of $KL(q\|p)$ [Bis06].

Now we can put it the other way around to see the implications of minimizing $KL(p\|q)$. In this case it is not so important that q has low mass in regions where p has low mass. Even if q has high mass in this region, leading to a high value of the logarithmic term, this high value is weighted by a small p. But in regions in which p does have considerable mass, q should also have mass. Otherwise the integrand would have a high value again. So minimizing $KL(p\|q)$ will enforce that q is not zero in regions where p is not zero. This is known as the *zero avoiding* property of $KL(p\|q)$.

Accuracy Now consider a multimodal posterior distribution that should be approximated. From the previous discussion it is clear that the VF approximation will seek one of the modes whereas EP, used with unimodal approximations, will lead to an approximation that averages over all the modes, given EP converges at all. In this case, using VF might be the best choice, especially if the multimodality of the posterior is due to symmetry, e.g. in mixture models the posterior distribution of the mixing coefficients is invariant under permutations of the mixture component indices. This kind of symmetry can be exploited [Bis06].

On the other hand, many posterior distributions will become more and more Gaussian shaped as the number of observations increases. This asymptotic normality of the posterior distribution is a consequence of the (Bayesian) Central Limit Theorem, but this assertion needs some conditions to hold [GCSR03]. The most important conditions require an asymmetry of the parameters (e.g. it should not be possible to simply exchange parameter indices to get an equivalent solution) and that the number of parameters should not grow with the number of observations. For problems in which these conditions do not hold, EP with Gaussian factor approximations should be a bad choice. However, if these conditions do hold, EP can produce a Gaussian approximation whose mean and variance are very close to the mean and variance of the true posterior due to the direction of the minimized KL divergence. VF will produce approximations that consistently underestimate the variance because of the zero forcing property of $KL(q\|p)$. Note that in EP we are free to choose that we want to approximate the posterior with a Gaussian, whereas in VF the functional form must be determined by the imposed factorization.

Unfortunately we could not find corresponding claims on the accuracy of VB approximations. We assume that the properties of the VB approximations depend heavily on the way, the bound is chosen.

6 Conclusion

In this seminar thesis, we gave a brief introduction to three methods for deterministically approximating a probability distribution p. We have fit each method in a general framework, where we define restrictions on the approximating family and a minimization objective. The first method, Variational Factorization, assumes a certain factorization of the approximating distribution q and minimizes the KL divergence $KL(q\|p)$. We derived expressions for the best approximating factors

(equation 8). It is important that the functional forms of the factors can be derived out of these expressions in order to turn the variational optimization into an ordinary optimization problem. Then we can optimize the parameters of the factors. The hard part of this method is choosing the factorization and deriving the functional form of the approximating factors.

The second method, Variational Bounds, starts by bounding an unnormalized true distribution. If p factors in a certain way, each factor can be bounded seperately. By doing so, the obtained bound depends on some parameters, that can be optimized to give the tightest bound. The hard part of this method is finding suitable bounds for the factors. Using convex duality, this is done by searching for a transformation of the factor and/or its input parameters to obtain a convex or concave function, which can then be bounded in terms of the dual function. Applying the inverse transformations must yield an expression, that renders the original problem tractable. We have exemplified this with the sigmoid function that plays an important role in Bayesian inference.

Expectation Propagation is the third method. It approximates each factor of the true distribution by an approximation that has to be from the exponential family. The optimization objective in EP is the KL divergence $KL(p\|q)$. That implies, that for EP, the best approximation can be found by moment matching. EP now provides us with an iterative update scheme for the single factors. In each update step we have to compute the moments from the product of the corresponding true factor and the other approximated factors (the context) and this is the hard part for EP.

References

[Bea03] Matthew J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[GCSR03] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2003.

[HZ02] Tom Heskes and Onno Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, pages 216–233, San Francisco, CA, 2002. Morgan Kaufmann Publishers.

[JGJS99] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999.

[Min01a] Thomas P. Minka. Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[Min01b] Thomas P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[Min07] Thomas P. Minka. A roadmap to research on EP. http://research.microsoft.com/en-us/um/people/minka/papers/ep/roadmap.html, 2007. [Online; accessed 24-June-2012].

[Min09] Thomas P. Minka. Approximate inference. http://videolectures.net/mlss09uk_minka_ai/, 2009. [Online; accessed 24-June-2012].

[SN11] Matthias W. Seeger and Hannes Nickisch. Fast convergent algorithms for expectation propagation approximate bayesian inference. *Journal of Machine Learning Research - Proceedings Track*, 15:652–660, 2011.

[WG07] Ydo Wexler and Dan Geiger. Variational upper bounds for probabilistic phylogenetic models. In *Proceedings of the 11th annual international conference on Research in computational molecular biology*, RECOMB'07, pages 226–237, Berlin, Heidelberg, 2007. Springer-Verlag.

[Win04] John M. Winn. *Variational Message Passing and its Applications*. PhD thesis, University of Cambridge, 2004.