Variational Algorithms for Approximate Bayesian Inference: Local Variational Methods

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, IN, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

April 4, 2018

Contents

- Recap of Approximation Inference
- Motivating Examples, Local Variational Methods, Convex Duality, Upper Bound for the Logistic Sigmoid Function, Lower Bound to the Logistic Sigmoid
- Variational Logistic Regression, Inference of Hyperparameters,
- Bohning's Bound to the log-sum-exp, Multinomial Logistic Regression, Bounds to the Sigmoid Function, Product of Sigmoids, Jensen's inequality, Multivariate Delta Method, Variational inference Based on Upper Bounds
- Expectation Propagation, Algorithm, Assumed Density Filtering, Convergence of EP, The clutter problem
- Expectation Propagation in Graphs, Tracking Problem

Following:

- Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 10
- Machine Learning: A Probabilistic Perspective, Kevin Murphy, Chapter 21.
- A roadmap to research EP, Msft Research (video Presentation)

Recap of Approximation Inference

Suppose we have a fully Bayesian model based around N i.i.d observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with latent parameters $\mathbf{Z} = \{z_1, \ldots, z_N\}$. Our model tells us the joint distribution $\mathrm{p}(\mathbf{X}, \mathbf{Z})$.

We want to approximate the posterior $p(\boldsymbol{Z}|\mathbf{X})$ and model evidence $\mathrm{p}(\boldsymbol{X})$ using some approximating function $\mathrm{q}(\mathbf{Z})$ (usually analytically tractable function).

Variational Inference is based on the following decomposition of the log marginal probability:

$$\ln p(x) = \mathcal{L}(q) + KL(q||p)$$

Where we have defined the lower bound and KL divergence respectively as follows:

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z}$$

$$KL(q||p) = -\int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} = \int q(\boldsymbol{Z}) \ln \left\{ \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} \right\} d\boldsymbol{Z}$$

Recap of Approximate Inference

Whilst Variational Inference is chiefly concerned with minimizing the KL divergence, we can also consider alternative forms of KL divergence.

For example, the reverse KL divergence:

$$KL(p||q) = -\int p(\boldsymbol{X}|\boldsymbol{Z}) \ln\left\{\frac{q(\boldsymbol{Z})}{p(\boldsymbol{X}|\boldsymbol{Z})}\right\} d\boldsymbol{Z} = \int p(\boldsymbol{X}|\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{X}|\boldsymbol{Z})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z}$$

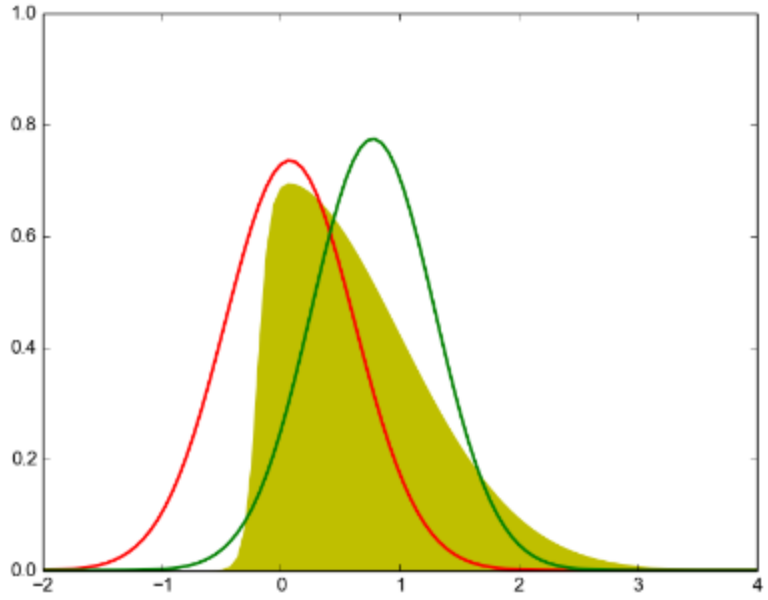
Note that in general $KL(p||q) \neq KL(q||p)$ and therefore using the reverse KL divergence would yield different results.

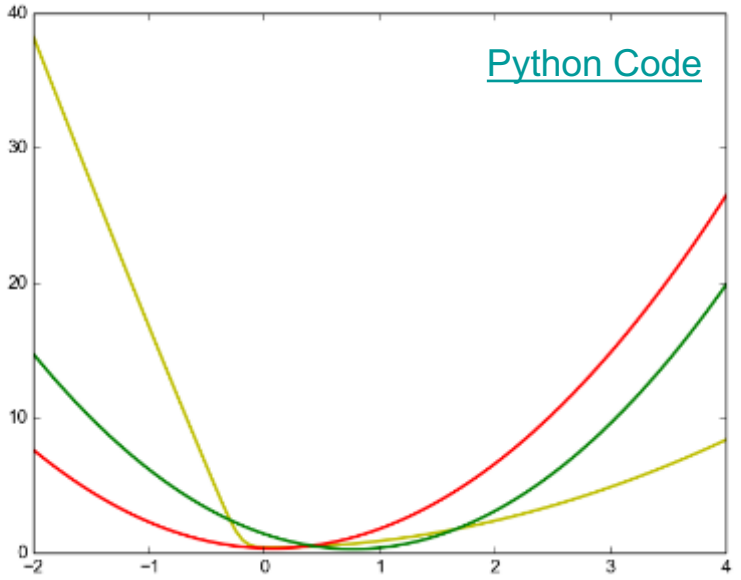
For example, reverse KL divergence gives rise to a class of methods known as Expectation Propagation.

This can sometimes outperform Variational Bayes e.g. in the clutter problem.

Outline of Variational Inference: Example

Example – approximating $\mathrm{f(Z)} = \mathrm{Aexp}\left(-\frac{\mathrm{Z}^2}{2}\right)\sigma(20Z+4)$ where $\sigma(Z)$ is the logistic sigmoid function defined by $\sigma(Z) = (1+e^{-Z})^{-1}$. Have restricted $q(Z) \sim \mathcal{N}(\mu, \sigma^2)$ and performed numerical optimization to find values of μ and σ^2 s.t. $KL(q||p)$ is minimized. The Laplace approximation also shown is centered on the mode *of* $p(Z)$.





Yellow: original function; red: Laplace approximation; green: variational approximation. Right plot shows the negative logs of the corresponding curves.

Local variational bounds

So far, we have been focusing on mean field inference, which is a form of variational inference based on minimizing $KL(q||\tilde{p})$, where q is the approximate posterior, assumed to be factorized, and $\tilde{p} = p(\boldsymbol{X}, \boldsymbol{Z})$ is the exact (but unnormalized) posterior.

$$KL(q||\tilde{p}) = -\int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} = -\int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z}$$

$$= -\int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} - ln\boldsymbol{p}(\boldsymbol{X}) = KL(q||p) - lnZ$$

where J is the normalization factor $\mathrm{Z} = \boldsymbol{p}(\boldsymbol{X})$. Of course

$$\mathcal{L}(q) = -KL(q||\tilde{p}) = \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z}.$$

However, there is another kind of variational inference, where we replace a specific term in the joint distribution with a simpler function, to simplify computation of the posterior.

Such an approach is sometimes called *a local variational approximation*, since we are only modifying one piece of the model, unlike mean field, which is a global approximation.

Motivating Example: Variational Logistic Regression

Consider the problem of approximating the parameter posterior for multiclass logistic regression model under a Gaussian prior.

One approach is to use a Gaussian (Laplace) approximation. However, a variational approach can produce a more accurate approximation to the posterior, since it has tunable parameters.

Another advantage is that the variational approach monotonically optimizes a lower bound on the likelihood of the data.

To see why we need a bound, recalling that $p(C_k|\boldsymbol{x}) = \frac{e^{\eta_k}}{\sum_j e^{\eta_j}}, \eta_j = \boldsymbol{x}^T \boldsymbol{w}_j$, note that the likelihood can be written as follows:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{i=1}^{N} exp(\boldsymbol{y}^T \boldsymbol{\eta}_i - lse(\boldsymbol{\eta}_i))$$

where $\boldsymbol{\eta}_i = [\eta_{ij}, \dots, \eta_{iM}] = \boldsymbol{X}_i \boldsymbol{w}_i = [\boldsymbol{x}_i^T \boldsymbol{w}_1, \dots, \boldsymbol{x}_i^T \boldsymbol{w}_M]$, where $M = C - 1$ (since we set $\mathbf{w}_C = \mathbf{0}$ for identifiability), and where we define the **log-sum-exp** or **lse** function as follows:

$$lse(\boldsymbol{\eta}_i) = ln\left(1 + \sum_{m=1}^{M} e^{\eta_{im}}\right)$$

Since the likelihood is not conjugate to the Gaussian prior, we need to compute "Gaussian-like" lower bounds to this likelihood, which give rise to approximate Gaussian posteriors.

Motivating Example: Multitask learning

One important application of Bayesian inference for logistic regression is fitting multiple related classifiers.

To share information between the parameters for each classifier requires that we maintain a posterior distribution over the parameters, so we have a measure of confidence as well as an estimate of the values.

We can embed the above variational method inside of a larger hierarchical model in order to perform such multi-task learning.

- Braun, M. and J. McAuliffe (2010). Variational Inference for Large-Scale Models of Discrete Choice. *J. of the Am. Stat. Assoc. 105*(489), 324–335.

Motivating Example: Discrete factor analysis

Another situation where variational bounds are useful arises when we fit a factor analysis model to discrete data (PCA for categorical data).

This model is just like multinomial logistic regression, except the input variables are hidden factors. We need to perform inference on the hidden variables as well as the regression weights.

The data has the form $y_{ij} \in \{1, \ldots, C\}$, where $j = 1 : R$ is the number of observed response variables. We assume each y_{ij} is generated from a latent variable $\boldsymbol{z_i} \in \mathbb{R}^L$, with a Gaussian prior, which is passed through the softmax function as follows:

$$p(\boldsymbol{z}_i) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

$$p(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) = \prod_{r=1}^{R} Cat(y_{ir}|S(\boldsymbol{W}_r^T \boldsymbol{z}_i + \boldsymbol{w}_{0r})), \; S(\boldsymbol{\eta})|_c = \frac{e^{\eta_c}}{\sum_j e^{\eta_j}}$$

where $\mathbf{W}_r \in \mathbb{R}^{L \times M}$ is the factor loading matrix for response j, and $\mathbf{w}_{0r} \in \mathbb{R}^M$ is the offset term for response r, and $\boldsymbol{\theta} = (\boldsymbol{W}_r, \boldsymbol{w}_{0r})_{r=1}^R$.

For simplicity, we might perform point estimation of the weights, and just integrate out the hidden variables. We can do this using variational EM, where we use the variational bound in the E step.

Machine Learning: A Probabilistic Perspective, Kevin Murphy, Chapter 9, Section 12.4.

Motivating Example: Correlated topic model

A topic model is a latent variable model for text documents and other forms of discrete data.

Often we assume the distribution over topics has a Dirichlet prior.

A more powerful model (correlated topic model), uses a Gaussian prior, which can model correlations more easily.

Unfortunately, this model involves the lse function. However, we can use variational bounds in the context of a variational EM algorithm, as we discuss in the follow up notes.

Machine Learning: A Probabilistic Perspective, Kevin Murphy, Chapter 9, Sections 27.3 and 27.4.1

Local Variational Methods

The methods discussed in the earlier two variational inference lectures can be considered as 'global' methods in the sense that they directly seek an approximation to the full posterior distribution over all random variables.

Alternative, we can use a 'local approach'. e.g we may seek bounds on a conditional distribution which is just one component of a larger probabilistic model.

A local approximation can be applied to multiple variables in-turn until we obtain a converged approximation. We will see this can be applied with success to the Logistic Regression problem.

In global variational methods, the convexity property of the log function ($\frac{d^2 lnx}{dx^2} = -\frac{1}{x^2} < 0$) in the KL divergence plays a central role in the development of the lower bound. Convexity is also important in local variational methods as it guarantees the existence of certain upper bounds and lower bounds that we will discuss.

- Rockafellar, R. (1972). *Convex Analysis*. Princeton University Press.
- Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in Graphical Models*, pp. 105– 162. MIT Press.

Local Variational Methods

Take a convex function such as $f(x) = \exp(-x)$ and suppose we want to find its tangent at a point $x = \xi$. Define $\eta = -\exp(-\xi)$ and then the tangent line is given by

$$y(x,\eta) = \eta x - \eta + \eta \ln(-\eta)$$

$$\begin{aligned} Use: y(x,\eta) &= f(\xi) + f'(\xi)(x-\xi) \\ &= -\eta + \eta(x + \ln(-\eta)) \\ where: f'(\xi) &= -e^{-\xi} \equiv \eta,\ f(x) \geq y(x,\eta) \end{aligned}$$

Different values of η correspond to different tangent lines, y(x,η) but all are lower bounds of $f(x)$. We can therefore write $f(x)$ as the supremum of the family of tangent lines. That is

$$f(x) = \max_{\eta} (\eta x - \eta + \eta \ln(-\eta))$$

We have approximated the convex function *f*(*x*) by a simpler, linear function *y*(*x*, η). The price we have paid is that we have introduced a variational parameter η, and to obtain the tightest bound we must optimize with respect to η.

Local Variational Bounds

The red curve shows the function exp(−*x*), and the blue line shows the tangent at $x = \xi$ defined by

$$y(x,\eta) = f(\xi) + f'(\xi)(x-\xi) = -\eta + \eta(x + \ln(-\eta))$$

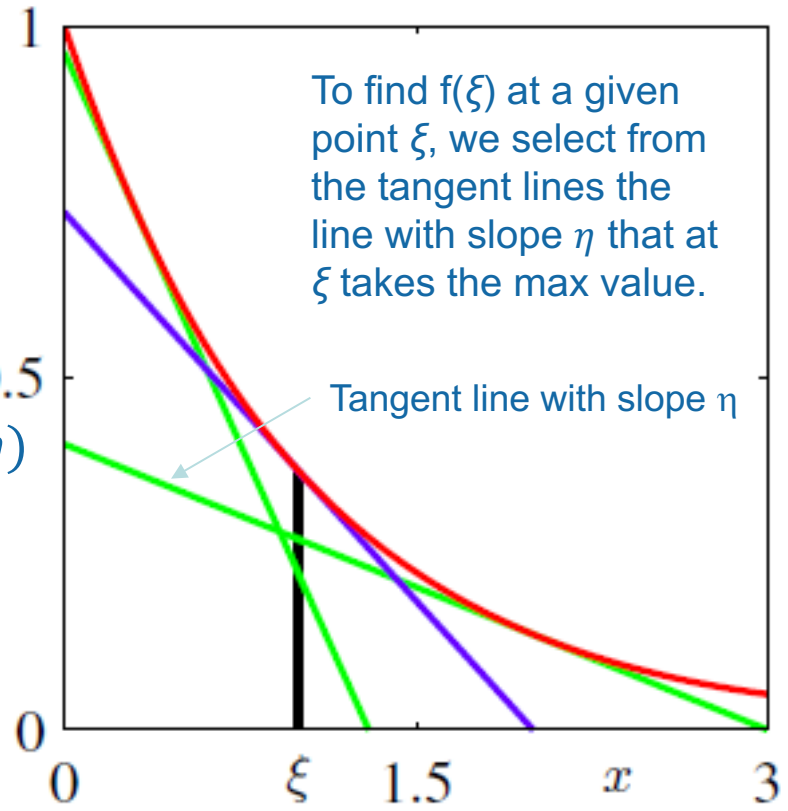
$$where: f'(\xi) = -e^{-\xi} \equiv \eta,\ f(x) \geq y(x,\eta)$$

with $\xi = 1$. This line has slope $\eta = f'(\xi) = -\exp(-\xi)$.

Note that any other tangent line (shown in green), will have a smaller value of *y* at $x = \xi$.

$f(x) = \max_{\eta}(\eta x - \eta + \eta\ln(-\eta))$

Note that $y(x,\eta) = \eta x - \eta + \eta\ln(-\eta)$ is a line tangent to $f(x) = \exp(-x)$ at some point $-\ln(-\eta)$

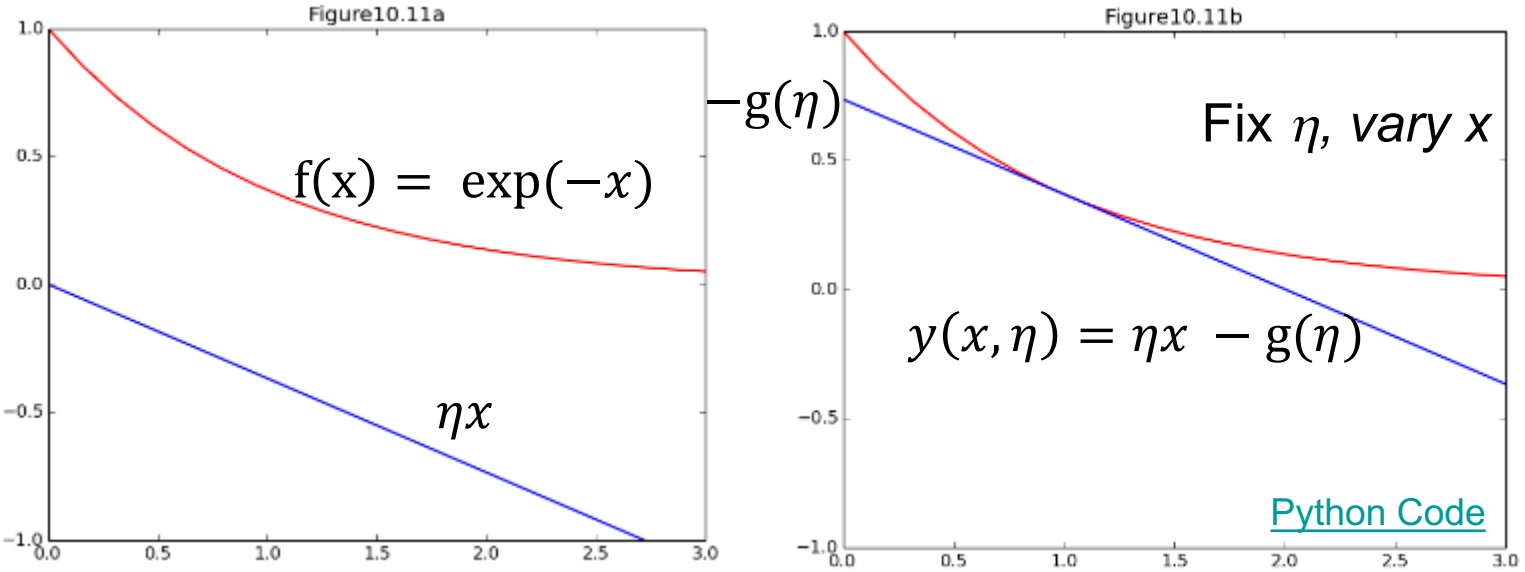


Convex Duality

This approach can be generalized using the concept of *convex duality*. Consider the illustration of a convex function *f*(*x*) shown in the left-hand plot.

The function ηx is a lower bound on *f*(*x*) but it is not the best lower bound that can be achieved by a linear function having slope η, because the tightest bound is given by the tangent line. Let us write the equation of the tangent line, having slope η as $\eta x - g(\eta)$. To determine the intercept, -*g*(η), note that the line must be moved vertically by an amount equal to the smallest vertical distance between the line and the function. Thus

$$g(\eta) = -\min_x \left(f(x) - \eta x \right) = \max_x \left(\eta x - f(x) \right)$$



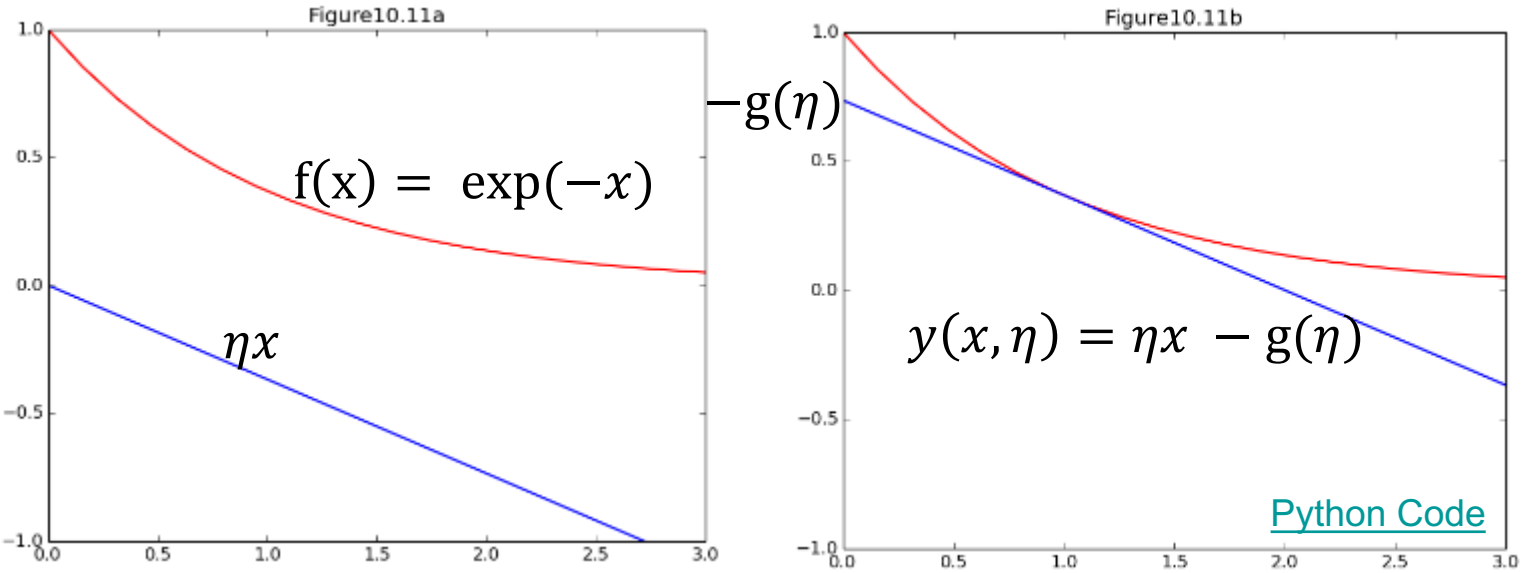
Convex Duality

Now, instead of fixing η and varying *x*, we can consider a particular *x* and then adjust η until the tangent plane is tangent at that particular *x*. Because the *y* value of the tangent line at a particular *x* is maximized when that value coincides with its contact point, we have

$$f(x) = \max_{\eta}\{\eta x - g(\eta)\}$$ Fix x, vary η

The functions *f*(*x*) and *g*(η) play a dual role

$$g(\eta) = -\min_{x}\left(f(x) - \eta x\right) = \max_{x}\left(\eta x - f(x)\right)$$ Fix η, *vary x*

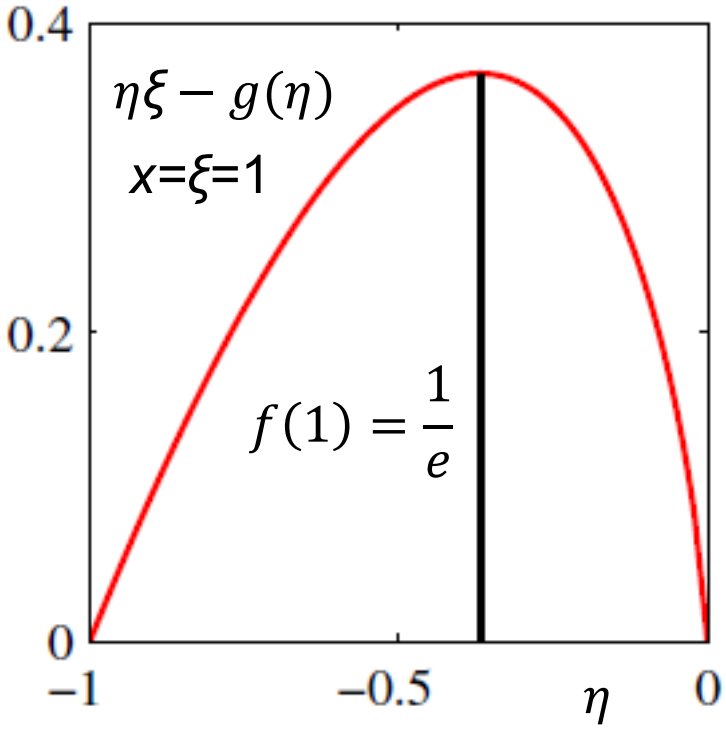


Local Variational Methods

For the example $f(x) = \exp(-x)$, $g(\eta) = \max_x\{\eta x - f(x)\}$ gives the maximizing value of x as $\xi = -\ln(-\eta)$, and back-substituting we obtain the conjugate function $g(\eta)$ in the form

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

Substituting into $f(x) = \max_\eta\{\eta x - g(\eta)\}$ gives the maximizing value of $\eta = -\exp(-x)$, and back-substituting then recovers $f(x) = \exp(-x)$.



$$f(x) = \max_\eta\{\eta x - g(\eta)\}$$

For $x = 1$, the max occurs when: $\eta = -e^{-1}$

$$f(1) = -e^{-1} - \left(-e^{-1} + e^{-1}ln(e^{-1})\right) = \frac{1}{e}$$

Convex Duality for the function f(x)=lnx

ln(x) is a concave function: $\frac{d^2 lnx}{dx^2} = -\frac{1}{x^2} < 0.$

Using $g(\eta) = \min_x \left(\eta x - f(x)\right) = \min_x \left(\eta x - \ln(x)\right)$, we compute x=1/ η, and thus: $g(\eta)=1-ln(1/\eta)$.

From $f(x) = \min_\eta \{\eta x - g(\eta)\} = \min_\eta \{\eta x - 1 + ln\frac{1}{\eta}\}$. This is maximized for $\eta = \frac{1}{x}$ and back substitution gives the obvious result: $f(x) = \frac{1}{x}x - 1 + lnx = lnx.$

We can write the upper bound of ln(x) as follows:

$$f(x) \leq \eta x - g(\eta) = \eta x - 1 + ln\frac{1}{\eta}$$

Upper Bound to the Logistic Sigmoid

If the function of interest is not convex (or concave), we can first seek invertible transformations which change it into a convex form. We then calculate the conjugate function and then transform back to the original variables.

We can apply this idea to the sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$

This function is neither convex nor concave, but the log is concave ($\frac{d^2 \ln(\sigma)}{dx^2} = -\sigma(x)^2 e^{-x} < 0$).
We therefore need to consider the 'reverse' of the method described previously:

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

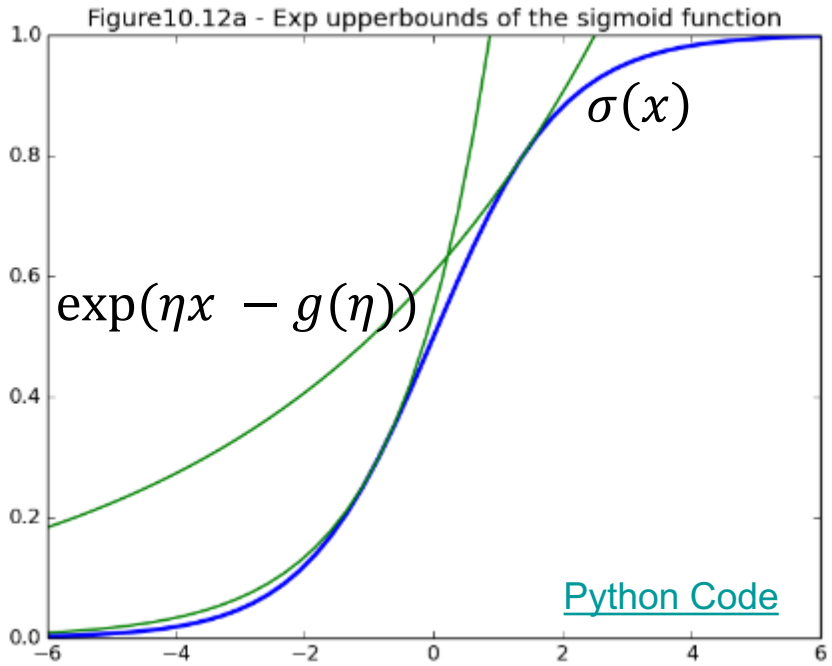
From the second Eq. above, derive the maximizing x=-lnη+ln(1-η), and plugging in: $g(\eta) = -\eta ln\eta - (1-\eta)\ln(1-\eta)$. Using this, we can obtain an upper bound to the logistic sigmoid:

$$f(x) = \min_{\eta}\{\eta x - g(\eta)\} \Rightarrow \ln \sigma(x) \leq \eta x - g(\eta) \text{ i.e.}$$
$$\sigma(x) \leq \exp(\eta x - g(\eta))$$

The same result can be obtained using a Taylor series expansion and that lnσ(x) is concave: ln$\sigma(x) \leq$ ln$\sigma(\xi)$+(x- ξ) $\sigma(\xi)e^{-\xi}$. By setting $\eta = \sigma(\xi)e^{-\xi}$, the upper bound is derived.

Note: $\eta = \frac{e^{-\xi}}{1+e^{-\xi}}, ln\sigma(\xi) = -\ln\left(1+e^{-\xi}\right) = \ln(1-\eta), \xi = -ln\eta + \ln(1-\eta)$, ln$\sigma(\xi)$+(x- ξ) $\eta = \ln(1-\eta) + \eta x - (-ln\eta + \ln(1-\eta))\,\eta = \eta x - g(\eta)$.

Upper Upper Bound to the Logistic Sigmoid



- Gibbs, M. N. (1997). *Bayesian Gaussian processes for regression and classification*. Phd thesis, University of Cambridge.

Lower Bound to the Logistic Sigmoid

To obtain a lower bound

First observe that

$\ln \sigma(x) = -\ln(1+e^{-x}) = -\ln\left\{e^{-\frac{x}{2}}\left(e^{\frac{x}{2}}+e^{-\frac{x}{2}}\right)\right\} = \frac{x}{2} - \ln\left(e^{\frac{x}{2}}+e^{-\frac{x}{2}}\right)$ where $f(x) = -\ln\left(e^{\frac{x}{2}}+e^{-\frac{x}{2}}\right)$ is a convex function of x^2 (For $y=x^2$, note that $\frac{d^2f}{dy^2} = \frac{1}{8y}\left(tanh\frac{\sqrt{y}}{2}\left(\frac{1}{\sqrt{y}}+\frac{1}{2}tanh\frac{\sqrt{y}}{2}\right)-\frac{1}{2}\right) > 0$).

We can obtain a lower bound on $f(x)$ that is a linear function of x^2 whose conjugate function is given by

$$g(\eta) = \max_{x^2}\{\eta x^2 - f(\sqrt{x^2})\}$$

The stationary condition leads to

$$0 = \eta - \frac{dx}{dx^2}\frac{d}{dx}f(x) = \eta + \frac{1}{4x}\tanh\frac{x}{2}$$

$$\tanh\alpha = \frac{e^{\alpha}-e^{-\alpha}}{e^{\alpha}+e^{-\alpha}}$$

We call this particular value of x as ξ (corresponding to the contact point of the tangent line for this particular value of η):

$$\eta = -\frac{1}{4\xi}\tanh\frac{\xi}{2} = -\frac{1}{2\xi}\left[\sigma(\xi)-\frac{1}{2}\right] \equiv -\lambda(\xi)$$

Defining $\lambda = -\eta$ we get

$$g(\lambda(\xi)) = -\lambda(\xi)\xi^2 - f(\xi) = -\lambda(\xi)\xi^2 + \ln(e^{\frac{\xi}{2}}+e^{\frac{-\xi}{2}})$$

Lower Bound to the Logistic Sigmoid

Thus from $f(x) = \max_{\eta}\{\eta x - g(\eta)\}$, the bound can be written as

$$f(x) \geq -\lambda(\xi)x^2 - g\big(\lambda(\xi)\big) = -\lambda(\xi)x^2 + \lambda(\xi)\xi^2 - \ln\left(e^{\frac{\xi}{2}} + e^{\frac{-\xi}{2}}\right)$$

Then using $\ln \sigma(x) = -\ln(1 + e^{-x}) = \frac{x}{2} + f(x)$, the lower bound on the sigmoid becomes:

$$\sigma(x) \geq \sigma(\xi)\exp\{\frac{(x-\xi)}{2} - \lambda(\xi)(x^2 - \xi^2)\}$$

where

$$\lambda(\xi) \equiv \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]$$

This lower bound has the form of the exponential of a quadratic function of *x*, which is useful when we seek Gaussian representations of posterior distributions defined through logistic sigmoid functions.

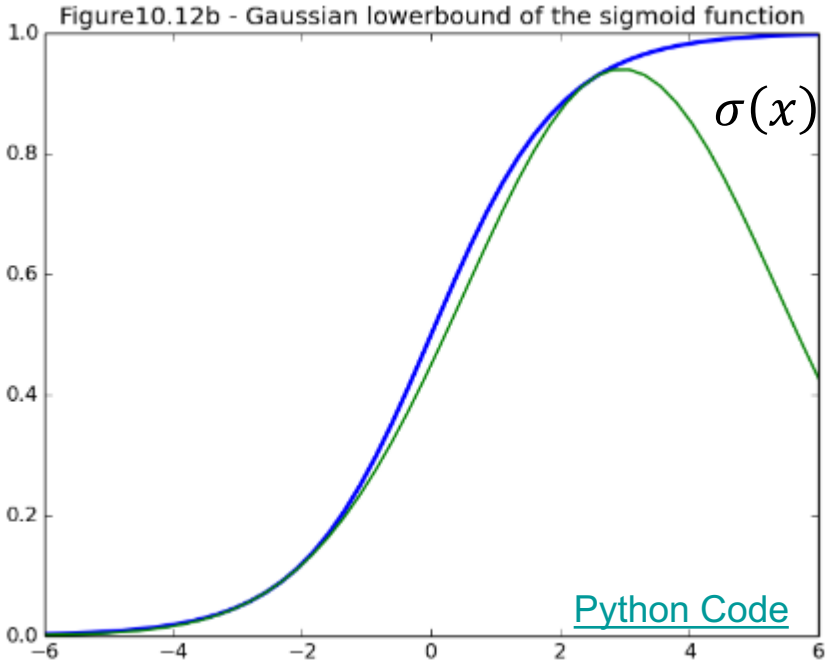
Note: the quadratic bound given here can be useful in evaluating approximations to integrals of the form $I=\int \sigma(a)p(a)da \geq \int f(a,\xi)p(a)da = F(\xi)$ (p(a) can be the posterior (e.g. Gaussian) of some parameters and I the predictive distribution). The variational parameter ξ can be chosen to maximize $F(\xi)$. However note that the optimal ξ*(a) in $\sigma(a) \geq f(a,\xi)$. Thus a compromise value ξ* is needed to perform the integration.

Lower Bound to the Logistic Sigmoid

Lower $Bound$:

$$\sigma(\xi)\exp\{\frac{(x-\xi)}{2}-\lambda(\xi)(x^2-\xi^2)\}$$

$Exact\ at\ \xi = \pm 2.5$

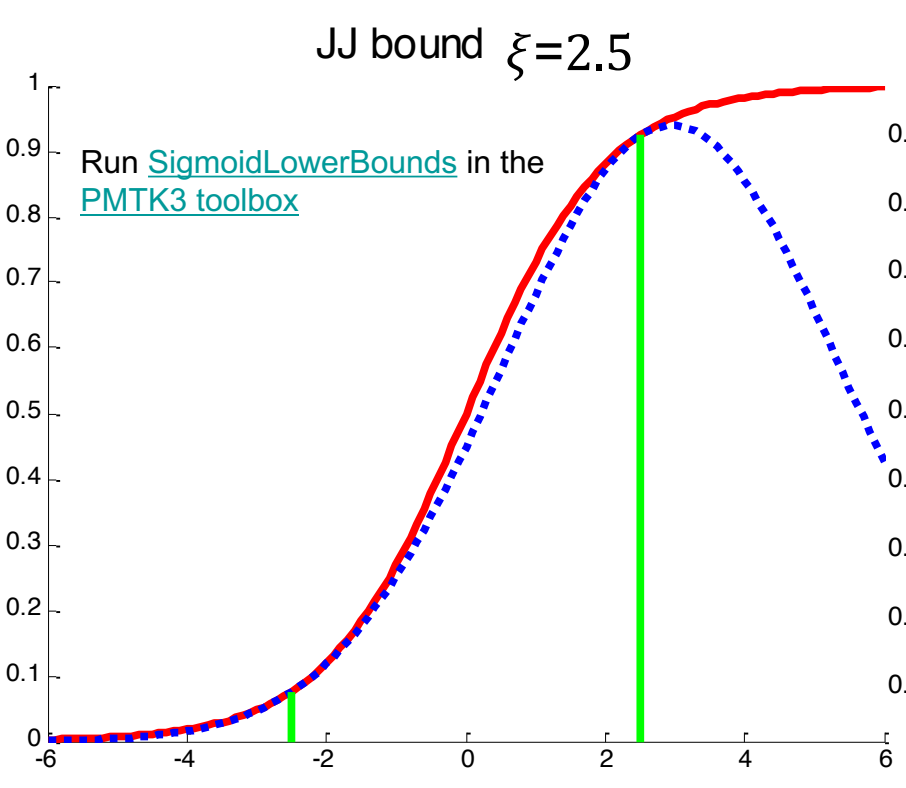


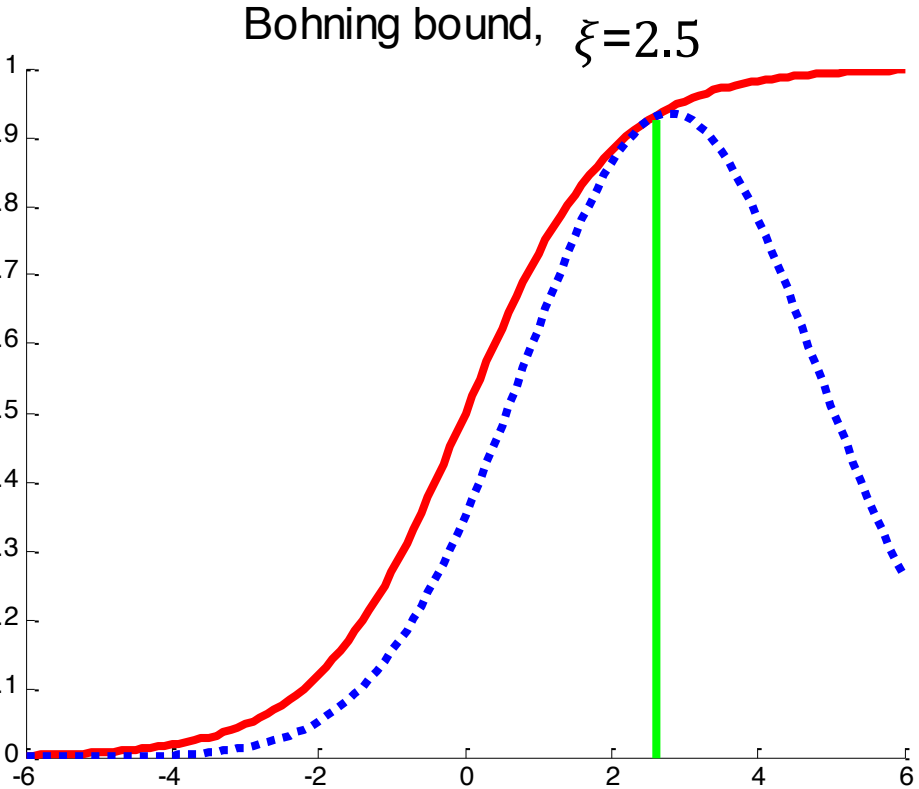
For the construction of bounds to the softmax (multiclass distribution) see reference by Gibbs.

- Gibbs, M. N. (1997). *Bayesian Gaussian processes for regression and classification*. Phd thesis, University of Cambridge.

Lower Bound to the Logistic Sigmoid

- The bound discussed is referred to as the JJ (Jaakkola-Jordan bound). The Bohning Bound shown on the right will be discussed shortly.





- Jaakkola, T. and M. Jordan (1996b). A variational approach to Bayesian logistic regression problems and their extensions. In *AI + Statistics*.
- Jaakkola, T. S. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing 10*, 25–37.

Variational Logistic Regression

In an earlier lecture, we used a Laplace approximation to approximate the posterior distribution of the Logistic Regression problem.

Local Variational methods can also be applied to evaluate the posterior of the Logistic Regression problem. Both methods lead to a Gaussian approximation of the posterior.

Generally speaking VI leads to improved accuracy when compared with the Laplace approximation. It optimizes a well defined bound to the model evidence.

The logistic regression problem concerns the evaluation of a two-class classification posterior under supervised learning.

We denote the training set observations by (ϕ_n, t_n). We use a prior $p(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$ where we take the hyperparameters $\mathbf{m}_0$, $\mathbf{S}_0$ as constant.

- Jaakkola, T. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.
- Dybowski, R. and S. Roberts (2005). An anthology of probabilistic models for medical informatics. In D. Husmeier, R. Dybowski, and S. Roberts Eds.), *Probabilistic Modeling in Bioinformatics and Medical Informatics*, pp. 297–349. Springer.

Variational Logistic Regression

Under the VI framework, we seek to maximize a lower bound on the marginal likelihood. For the Bayesian logistic regression model this takes the form

$$p(\boldsymbol{t}) = \int p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \int \left[\prod_{n=1}^{N} p(t_n|\boldsymbol{w})\right] p(\boldsymbol{w})d\boldsymbol{w}$$

Defining $a = \boldsymbol{w}^T\boldsymbol{\phi}$, the conditional distribution can be written as:

$$p(t|\boldsymbol{w}) = \sigma(a)^t\big(1-\sigma(a)\big)^{1-t} = \left(\frac{1}{1+e^{-a}}\right)^t\left(1-\frac{1}{1+e^{-a}}\right)^{1-t} = \frac{e^{at}e^{-a}}{1+e^{-a}} = e^{at}\sigma(-a)$$

We now use the previously derived result:

$$\sigma(x) \geq \sigma(\xi)\exp\{\frac{(x-\xi)}{2} - \lambda(\xi)(x^2-\xi^2)\}$$

where

$$\frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right] = \lambda(\xi)$$

This bound does not generalize to the multiclass case K>2 (see Gibbs, 1997).

- Gibbs, M. N. (1997). *Bayesian Gaussian processes for regression and classification*. Phd thesis, University of Cambridge.

Variational Logistic Regression

We can therefore write

$$p(t|\boldsymbol{w}) = e^{at}\sigma(-a) \geq e^{at}\sigma(\xi)\exp\{\frac{-(a+\xi)}{2} - \lambda(\xi)(a^2 - \xi^2)\}$$

We can then write the joint distribution as:

$$p(\boldsymbol{t}, \boldsymbol{w}) = p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}) \geq h(\boldsymbol{w}, \boldsymbol{\xi})p(\boldsymbol{w})$$

where $\boldsymbol{\xi}$ denotes the set of variational parameters ξ_n,n=1,...,N and

$$h(\boldsymbol{w}, \boldsymbol{\xi}) = \prod_{n=1}^{N} \sigma(\xi_n)\exp\left(\boldsymbol{w}^T\boldsymbol{\phi}_n t_n - \frac{\boldsymbol{w}^T\boldsymbol{\phi}_n + \xi_n}{2} - \lambda(\xi_n)([\boldsymbol{w}^T\boldsymbol{\phi}_n]^2 - \xi_n^2)\right)$$

Evaluation of the exact posterior distribution would require normalization of the left hand side of $p(\boldsymbol{t}, \boldsymbol{w}) \geq h(\boldsymbol{w}, \boldsymbol{\xi})p(\boldsymbol{w})$. Because this is intractable, we work instead with the right-hand side.

The function $h(\boldsymbol{w}, \boldsymbol{\xi})p(\boldsymbol{w})$ on the right-hand side cannot be interpreted as a probability density because it is not normalized. Once it is normalized to give a variational posterior distribution $q(\mathbf{w})$, however, it no longer represents a bound.

Variational Logistic Regression

Due to monotonicity of the log function, $A \geq B$ implies $\ln A \geq \ln B$, we can thus take logs of both sides and preserve the earlier inequality. This gives a lower bound on the log of the joint distribution of **t**,**w**:

$$\ln\{p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})\} \geq \ln\{p(\boldsymbol{w})\} + \sum_{n=1}^{N}\left\{\ln\sigma(\xi_n) + \boldsymbol{w}^T\boldsymbol{\phi}_n t_n - \frac{(\boldsymbol{w}^T\boldsymbol{\phi}_n + \xi_n)}{2} - \lambda(\xi_n)\left((\boldsymbol{w}^T\boldsymbol{\phi}_n)^2 - \xi_n^2\right)\right\}$$

Substituting the prior $p(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$, the right hand side can be expressed in the form:

$$\ln\{p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})\} \geq -\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0) + \sum_{n=1}^{N}\left\{\boldsymbol{w}^T\boldsymbol{\phi}_n\left(t_n - \frac{1}{2}\right) - \lambda(\xi_n)\boldsymbol{w}^T(\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T)\boldsymbol{w}\right\} + const$$

This gives a Gaussian of the form

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$$

where

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\left(t_n - \frac{1}{2}\right)\boldsymbol{\phi}_n\right)$$

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T$$

Note that this bound is only applicable for the case where K = 2.

Variational Logistic Regression

As with the Laplace approximation, we obtain a Gaussian approximation to the posterior.

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\left(t_n - \frac{1}{2}\right)\boldsymbol{\phi}_n\right), \boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T$$

This VI framework provides extra flexibility and accuracy by the presence of the variational parameters ξ_n.

The formulation of this variational approach for the sequential case (the data are processed one at a time and then discarded) is given below:

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_{N-1}^{-1}\boldsymbol{m}_{N-1} + \left(t_N - \frac{1}{2}\right)\boldsymbol{\phi}_N\right), \qquad \boldsymbol{S}_N^{-1} = \boldsymbol{S}_{N-1}^{-1} + 2\lambda(\xi_N)\boldsymbol{\phi}_N\boldsymbol{\phi}_N^T$$

- Jaakkola, T. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.
- Gibbs, M. N. (1997). *Bayesian Gaussian processes for regression and classification*. Phd thesis, University of Cambridge.
- Bishop, C. M. and M. Svensén (2003). Bayesian hierarchical mixtures of experts. In U. Kjaerulff and C. Meek (Eds.), *Proceedings Nineteenth Conference on Uncertainty in Artificial Intelligence*, pp. 57–64. Morgan Kaufmann.

Optimizing the Variational Parameters

Now we have a Gaussian approximation $q(\boldsymbol{w})$ to the posterior, we need to obtain the variational parameters ξ_n by maximizing the lower bound on the marginal likelihood.

$$\ln p(\boldsymbol{t}) = \ln \int p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} \geq \ln \int h(\boldsymbol{w},\boldsymbol{\xi})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{L}(\boldsymbol{\xi})$$

As before, we can adopt an EM-based approach (view $\boldsymbol{w}$ as a latent variable).

Start with some guess of the parameters $\boldsymbol{\xi}^{old}$. In the E-step we use the parameter values to obtain $q(\boldsymbol{w})$ as $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$.

In the M-step we maximize the expected complete data log likelihood, where the expectation is taken wrt q($\mathbf{w}$) evaluated using $\boldsymbol{\xi}^{old}$:

$$Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{old}) = \mathbb{E}[\ln\{h(\boldsymbol{w}|\boldsymbol{\xi})p(\boldsymbol{w})\}] = \sum_{n=1}^{N}\{\ln\sigma(\xi_n) - \frac{\xi_n}{2} - \lambda(\xi_n)\,(\phi_n^T\,\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\phi_n - \xi_n^2)\} + const$$

Setting the derivative with respect to ξ_n and noting that $\lambda'(\xi)$ is non-zero, we obtain

$$\frac{\partial Q}{\partial \xi_n} = \frac{1}{\sigma(\xi_n)}\sigma'(\xi_n) - \frac{1}{2} - \lambda'(\xi_n)(\phi_n^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\phi_n - \xi_n^2) + \lambda(\xi_n)2\xi_n = \frac{1}{\sigma(\xi_n)}\sigma(\xi_n)\big(1-\sigma(\xi_n)\big) - \frac{1}{2} -$$
$$\frac{1}{2}\lambda'(\xi_n)(\phi_n^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\phi_n - \xi_n^2) + \frac{1}{2\xi_n}\Big(\sigma(\xi_n)) - \frac{1}{2}\Big)2\xi_n = -\lambda'(\xi_n)(\phi_n^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\phi_n - \xi_n^2) = 0$$

from which using $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$:

$$(\xi_n^{new})^2 = \boldsymbol{\phi}_n^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n = \boldsymbol{\phi}_n^T(\boldsymbol{S}_N + \boldsymbol{m}_N\boldsymbol{m}_N^T)\boldsymbol{\phi}_n$$

Variational Logistic Regression

- E-Step

Compute the following:

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\left(t_n - \frac{1}{2}\right)\boldsymbol{\phi}_n\right)$$

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T$$

- M-Step

Compute the following:

$$(\xi_n^{new})^2 = \boldsymbol{\phi}_n^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n$$
$$= \boldsymbol{\phi}_n^T(\boldsymbol{S}_N + \boldsymbol{m}_N\boldsymbol{m}_N^T)\boldsymbol{\phi}_n$$

Note: Instead of the batch implementation where all variational parameters are updated using statistics based on all data points, in a sequential implementation, we alternate between updating $\boldsymbol{m}_N$ and $\boldsymbol{S}_N$ with fixed ξ_n and updating ξ_n with $\boldsymbol{m}_N$ and $\boldsymbol{S}_N$ kept fixed.

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_{N-1}^{-1}\boldsymbol{m}_{N-1} + \left(t_N - \frac{1}{2}\right)\boldsymbol{\phi}_N\right), \qquad \boldsymbol{S}_N^{-1} = \boldsymbol{S}_{N-1}^{-1} + 2\lambda(\xi_N)\boldsymbol{\phi}_N\boldsymbol{\phi}_N^T$$

The update formula for the parameters $(\xi_n^{new})^2 = \boldsymbol{\phi}_n^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n = \boldsymbol{\phi}_n^T(\boldsymbol{S}_N + \boldsymbol{m}_N\boldsymbol{m}_N^T)\boldsymbol{\phi}_n$ remains the same but each parameter is updated only once.

Variational Logistic Regression

- E-Step

Compute the following:

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\left(t_n - \frac{1}{2}\right)\boldsymbol{\phi}_n\right)$$

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T$$

- M-Step

Compute the following:

$$(\xi_n^{new})^2 = \boldsymbol{\phi}_n^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n$$
$$= \boldsymbol{\phi}_n^T(\boldsymbol{S}_N + \boldsymbol{m}_N\boldsymbol{m}_N^T)\boldsymbol{\phi}_n$$

Note: An alternative approach to obtaining re-estimation equations for $\boldsymbol{\xi}$ is to note that in the integral over **w** in the definition $\ln \int h(\boldsymbol{w}, \boldsymbol{\xi})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{L}(\boldsymbol{\xi})$ of the lower bound $\mathcal{L}(\boldsymbol{\xi})$, the integrand has a Gaussian-like form and so the integral can be evaluated analytically. Having evaluated the integral, we can then differentiate with respect to ξ_n. It turns out that this gives rise to exactly the same re-estimation equations as does the EM approach given by $(\xi_n^{new})^2 = \boldsymbol{\phi}_n^T\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n = \boldsymbol{\phi}_n^T(\boldsymbol{S}_N + \boldsymbol{m}_N\boldsymbol{m}_N^T)\boldsymbol{\phi}_n$.

Optimizing the Variational Parameters

The integration over **w** in $\mathcal{L}(\boldsymbol{\xi})$ can be performed analytically by noting that $p(\mathbf{w})$ is Gaussian and $h(\boldsymbol{w},\boldsymbol{\xi}) = \prod_{n=1}^{N}\sigma(\xi_n)\exp\left(\boldsymbol{w}^T\boldsymbol{\phi}_n t_n - \frac{\boldsymbol{w}^T\boldsymbol{\phi}_n+\xi_n}{2} - \lambda(\xi_n)([\boldsymbol{w}^T\boldsymbol{\phi}_n]^2 - \xi_n^2)\right)$ is the exponential of a quadratic function of **w**.

$$p(\boldsymbol{w})h(\boldsymbol{w},\boldsymbol{\xi}) = (2\pi)^{-W/2}|\boldsymbol{S}_0|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{w}^T(\boldsymbol{S}_0^{-1} + 2\sum_{n1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T)\boldsymbol{w} + \boldsymbol{w}^T\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\boldsymbol{\phi}_n\left(t_n - \frac{1}{2}\right)\right)\right\}\exp\left\{-\frac{1}{2}\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\left(-\frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2\right)\right\}\prod_{n=1}^{N}\sigma(\xi_n)$$

We can complete the square over **w** using $\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\left(t_n - \frac{1}{2}\right)\boldsymbol{\phi}_n\right)$ and $\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T$ to yield:

$$p(\boldsymbol{w})h(\boldsymbol{w},\boldsymbol{\xi}) = (2\pi)^{-W/2}|\boldsymbol{S}_0|^{-1/2}\prod_{n=1}^{N}\sigma(\xi_n)\exp\left\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T\boldsymbol{S}_N^{-1}\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)\right\}\exp\left\{\frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{S}_N^{-1}\boldsymbol{m}_N - \frac{1}{2}\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\left\{-\frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2\right\}\right\}$$

We can now integrate in $\boldsymbol{w}$ in $\ln p(\boldsymbol{t}) \geq \ln\int h(\boldsymbol{w},\boldsymbol{\xi})p(\boldsymbol{w})d\boldsymbol{w} = \mathcal{L}(\boldsymbol{\xi})$ and replace the first exponential with the normalization factor $(2\pi)^{W/2}|\boldsymbol{S}_N|^{1/2}$.

Thus, by completing the square and making use of the standard result for the normalization coefficient of a Gaussian, we can obtain a closed form solution which takes the form

$$\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{2}\ln\frac{|\boldsymbol{S}_N|}{|\boldsymbol{S}_0|} + \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{S}_N^{-1}\boldsymbol{m}_N - \frac{1}{2}\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\left\{\ln\sigma(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2\right\}$$

Reestimation Eqs by maximization of $\mathcal{L}(\boldsymbol{\xi})$

$$\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{2}\ln\frac{|\boldsymbol{S}_N|}{|\boldsymbol{S}_0|} + \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{S}_N^{-1}\boldsymbol{m}_N - \frac{1}{2}\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\{\ln\sigma(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2\}$$

Using the identities $\frac{d}{d\alpha}\ln|\mathbf{A}| = \mathrm{Tr}\left(\boldsymbol{A}^{-1}\frac{d\boldsymbol{A}}{d\alpha}\right)$, $\sigma'(\xi_n) = \sigma(\xi_n)\big(1-\sigma(\xi_n)\big)$, $\lambda(\xi) = \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]$, we can differentiate $\mathcal{L}(\boldsymbol{\xi})$:

$$\frac{\partial\mathcal{L}(\xi)}{\partial\xi_n} = \frac{1}{2}\,Tr\left(\boldsymbol{S}_N^{-1}\frac{\partial\boldsymbol{S}_N}{\partial\xi_n}\right) + \frac{1}{2}\,Tr\left(\boldsymbol{a}_N\boldsymbol{a}_N^T\frac{\partial\boldsymbol{S}_N}{\partial\xi_n}\right) + \lambda'(\xi_n)\xi_n^2 = 0, \text{ where: } \boldsymbol{a}_N = \boldsymbol{S}_N^{-1}\boldsymbol{m}_N$$

Using $\frac{\partial\boldsymbol{S}_N^{-1}}{\partial\xi_n} = -\boldsymbol{S}_N^{-1}\frac{\partial\boldsymbol{S}_N}{\partial\xi_n}\boldsymbol{S}_N^{-1}$ and $\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T$, we can derive:

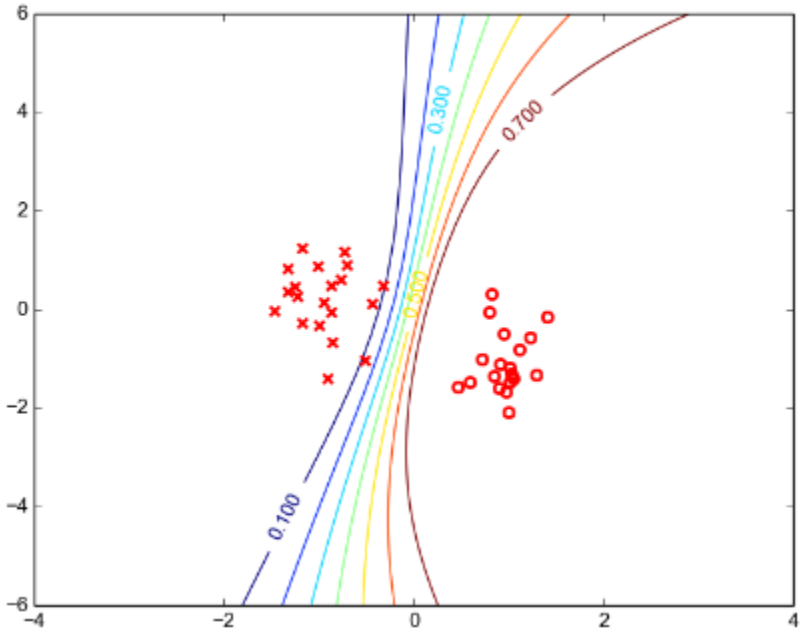
$$\frac{\partial\boldsymbol{S}_N}{\partial\xi_n} = \frac{\partial(\boldsymbol{S}_N^{-1})^{-1}}{\partial\xi_n} = -\boldsymbol{S}_N\frac{\partial\boldsymbol{S}_N^{-1}}{\partial\xi_n}\boldsymbol{S}_N = -2\boldsymbol{S}_N\lambda'(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\boldsymbol{S}_N$$

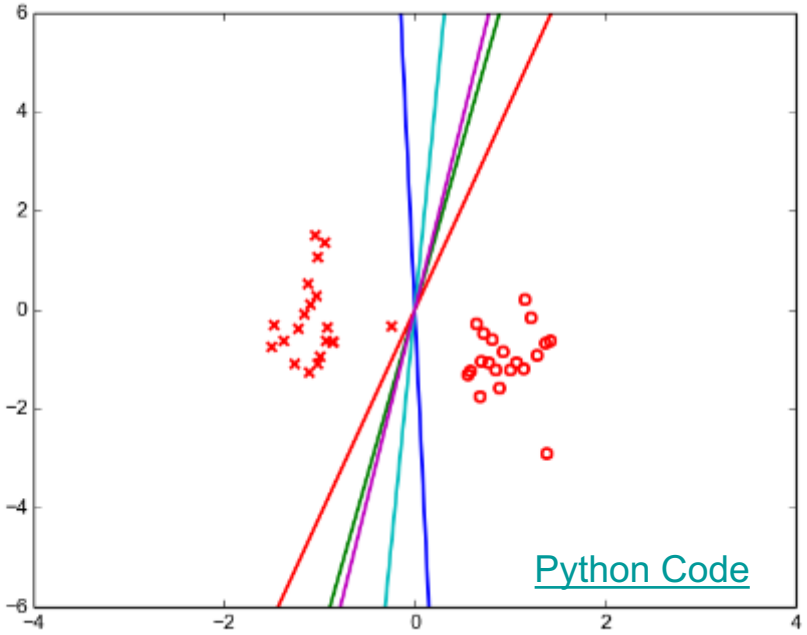
Thus:

$$\frac{\partial\mathcal{L}(\xi)}{\partial\xi_n} = \frac{1}{2}\,Tr\left((\boldsymbol{S}_N^{-1} + \boldsymbol{a}_N\boldsymbol{a}_N^T)\frac{\partial\boldsymbol{S}_N}{\partial\xi_n}\right) + \lambda'(\xi_n)\xi_n^2 = -\frac{1}{2}\,Tr\big((\boldsymbol{S}_N^{-1} + \boldsymbol{a}_N\boldsymbol{a}_N^T)2\boldsymbol{S}_N\lambda'(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\boldsymbol{S}_N\big) + \lambda'(\xi_n)\xi_n^2 = 0$$

$$(\xi_n)^2 = \boldsymbol{\phi}_n^T(\boldsymbol{S}_N + \boldsymbol{m}_N\boldsymbol{m}_N^T)\boldsymbol{\phi}_n$$

Variational Logistic Regression





- Bayesian approach to logistic regression for a simple linearly separable data set.
- The plot on the left shows the predictive distribution obtained using VI. We see that the decision boundary lies roughly mid way between the clusters of data points, and that the contours of the predictive distribution splay out away from the data reflecting the greater uncertainty in the classification of such regions.
- The plot on the right shows the decision boundaries corresponding to five samples of the parameter vector $\mathbf{w}$ drawn from the posterior distribution

Optimizing the Variational Parameters

$$\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{2}\ln\frac{|\boldsymbol{S}_N|}{|\boldsymbol{S}_0|} + \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{S}_N^{-1}\boldsymbol{m}_N - \frac{1}{2}\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N}\{\ln\sigma(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2\}$$

This variational framework can also be applied to situations in which the data is arriving sequentially (Jaakkola and Jordan, 2000).

In this case we maintain a Gaussian posterior distribution over **w**, which is initialized using the prior $p(\mathbf{w})$. As each data point arrives, the posterior is updated by making use of the bound

$$p(t|\boldsymbol{w}) = e^{at}\sigma(-a) \geq e^{at}\sigma(\xi)\exp\{\frac{-(a+\xi)}{2} - \lambda(\xi)(a^2-\xi^2)\}$$

and then normalized to give an updated posterior distribution.

- Jaakkola, T. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.

Inference of Hyperparameters

We now extend the Bayesian logistic regression model to allow the value of this parameter α to be inferred from the data set. This can be achieved by combining the global and local variational approximations into a single framework, so as to maintain a lower bound on the marginal likelihood at each stage.

Consider $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$. Our analysis is readily extended to more general Gaussian priors, for instance if we wish to associate a different hyperparameter with different subsets of the parameters w_j . As usual, we consider a conjugate hyperprior over α given by a gamma distribution $p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0)$

The marginal likelihood for this model now takes the form

$$p(\mathbf{t}) = \iint p(\mathbf{w}, \alpha, \mathbf{t})\, d\mathbf{w} d\alpha$$

where the joint distribution is given by

$$p(\mathbf{w}, \alpha, \mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha)$$

We are now faced with an analytically intractable integration over $\mathbf{w}$ and α, which we shall tackle by using both the local and global variational approaches in the same model.

- Bishop, C. M. and M. Svensén (2003). Bayesian hierarchical mixtures of experts. In U. Kjaerulff and C. Meek (Eds.), *Proceedings Nineteenth Conference on Uncertainty in Artificial Intelligence*, pp. 57–64. Morgan Kaufmann.

Inference of Hyperparameters

To begin with, we introduce a variational distribution $q(\mathbf{w}, \alpha)$, and then apply the decomposition $\ln p(\boldsymbol{t}) = \mathcal{L}(q) + KL(q||p)$ where the lower bound $\mathcal{L}(q)$ and the Kullback-Leibler divergence KL(q||p) are defined by

$$\mathcal{L}(q) = \int \int q(\mathbf{w}, \alpha) \ln \left\{ \frac{p(\mathbf{w}, \alpha, \boldsymbol{t})}{q(\mathbf{w}, \alpha)} \right\} d\boldsymbol{w}\, d\alpha$$

$$KL(q||p) = -\int \int q(\mathbf{w}, \alpha) \ln \left\{ \frac{p(\mathbf{w}, \alpha|\boldsymbol{t})}{q(\mathbf{w}, \alpha)} \right\} d\boldsymbol{w}\boldsymbol{d}\alpha$$

At this point, the lower bound $\mathcal{L}(q)$ is still intractable due to the form of the likelihood factor $p(\mathbf{t}|\mathbf{w})$. We therefore apply the local variational bound to each of the logistic sigmoid factors as before. This allows us to use $p(\boldsymbol{t}, \boldsymbol{w}) = p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}) \geq h(\boldsymbol{w}, \boldsymbol{\xi})p(\boldsymbol{w})$ and place a lower bound on $\mathcal{L}(q)$which will therefore also be a lower bound on the log marginal likelihood

$$\ln p(\boldsymbol{t}) \geq \mathcal{L}(q) \geq \tilde{\mathcal{L}}(q, \boldsymbol{\xi}) = \int \int q(\mathbf{w}, \alpha) \ln \left\{ \frac{h(\boldsymbol{w}, \boldsymbol{\xi})p(\boldsymbol{w}|\alpha)p(\alpha)}{q(\mathbf{w}, \alpha)} \right\} d\boldsymbol{w}\, d\alpha$$

Assume a factorization: $q(\mathbf{w}, \alpha) = q(\mathbf{w})\, q(\alpha)$

Inference of Hyperparameters

Appealing to the general VI results:

$$\ln q(\boldsymbol{w}) = \mathbb{E}_{\alpha}[ln\{h(\boldsymbol{w},\boldsymbol{\xi})p(\boldsymbol{w}|\alpha)p(\alpha)\}] + const = \ln h(\boldsymbol{w},\boldsymbol{\xi}) + \mathbb{E}_{\alpha}[ln\{p(\boldsymbol{w}|\alpha)\}] + \text{const}$$

Substitution of $h(\boldsymbol{w},\boldsymbol{\xi}) = \prod_{n=1}^{N}\sigma(\xi_n)\exp\left(\boldsymbol{w}^T\boldsymbol{\phi}_n t_n - \frac{\boldsymbol{w}^T\boldsymbol{\phi}_n+\xi_n}{2} - \lambda(\xi_n)([\boldsymbol{w}^T\boldsymbol{\phi}_n]^2 - \xi_n^2)\right)$ and $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$ gives:

$$\ln q(\boldsymbol{w}) = -\frac{\mathbb{E}[\alpha\}}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N}\left\{\boldsymbol{w}^T\boldsymbol{\phi}_n\left(t_n - \frac{1}{2}\right) - \lambda(\xi_n)\boldsymbol{w}^T(\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T)\boldsymbol{w}\right\} + const$$

This is a quadratic function of $\boldsymbol{w}$ and so the solution for q($\boldsymbol{w}$) will be Gaussian. Completing the square:

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$$

$$\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N = \sum_{n=1}^{N}\left\{\boldsymbol{\phi}_n\left(t_n - \frac{1}{2}\right)\right\}$$

$$\boldsymbol{\Sigma}_N^{-1} = \mathbb{E}[\alpha]\boldsymbol{I} + 2\sum_{n=1}^{N}\{\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\}$$

Re-estimation Eqs for $q(\mathbf{w})$, $q(\alpha)$ and ξ

Similarly

$$\ln q(\alpha) = \mathbb{E}_{\boldsymbol{w}}[ln\{p(\boldsymbol{w}|\alpha)\}] + lnp(\alpha) + const$$

Substitution of $p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0)$ and $p(\mathbf{w}|\alpha) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$ gives:

$$\ln q(\alpha) = \frac{M}{2} ln\alpha - \frac{\alpha}{2}\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}^T\boldsymbol{w}] + (a_0 - 1)ln\alpha - b_0\alpha + const$$

This is the log of a Gamma distribution and thus:

$$q(\alpha) = Gamma(\alpha|a_N, b_N) = \frac{1}{\Gamma(a_N)} a_N^{b_N} \alpha^{a_N - 1} e^{-b_N\alpha}$$

$$a_N = a_0 + \frac{M}{2}, b_N = b_0 + \frac{1}{2}\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}^T\boldsymbol{w}]$$

We also need to optimize ξ_n by maximizing $\tilde{\mathcal{L}}(q, \boldsymbol{\xi}) = \int q(\mathbf{w}) \ln h(\boldsymbol{w}, \boldsymbol{\xi}) d\boldsymbol{w}$ +const

This is the same form as $Q(\boldsymbol{\xi}, \boldsymbol{\xi}^{old}) = \mathbb{E}[\ln\{h(\boldsymbol{w}|\boldsymbol{\xi})p(\boldsymbol{w})\}]$ and thus from earlier results:

$$(\xi_n^{new})^2 = \boldsymbol{\phi}_n^T \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\phi}_n = \boldsymbol{\phi}_n^T(\boldsymbol{S}_N + \boldsymbol{m}_N\boldsymbol{m}_N^T)\boldsymbol{\phi}_n$$

In the iterative process, the required moments are $\mathbb{E}[\alpha] = \frac{a_N}{b_N}$, $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N\boldsymbol{\mu}_N^T$

Bohning's Bound to the log-sum-exp

The earlier examples considered require dealing with multiplying a Gaussian prior by a multinomial likelihood; this is difficult because of the log-sum-exp (lse) term. Here, we derive a "Gaussian-like" lower bound on this likelihood.

Consider a Taylor series expansion of the lse function $lse(\boldsymbol{\eta}_i) = ln(1 + \sum_{m=1}^{M} e^{\eta_{im}})$ around $\boldsymbol{\psi}_i \in \mathbb{R}^M$:

$$lse(\boldsymbol{\eta}_i) = lse(\boldsymbol{\psi}_i) + (\boldsymbol{\eta}_i - \boldsymbol{\psi}_i)^T \mathrm{g}(\boldsymbol{\psi}_i) + \frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\psi}_i)^T \mathbf{H}(\boldsymbol{\psi}_i)(\boldsymbol{\eta}_i - \boldsymbol{\psi}_i)$$

$$\mathrm{g}(\boldsymbol{\psi}_i) = \exp[\boldsymbol{\psi}_i - lse(\boldsymbol{\psi}_i)] = \mathcal{S}(\boldsymbol{\psi}_i)$$

$$\mathbf{H}(\boldsymbol{\psi}_i) = diag\big(\mathrm{g}(\boldsymbol{\psi}_i)\big) - \mathrm{g}(\boldsymbol{\psi}_i)\mathrm{g}(\boldsymbol{\psi}_i)^T$$

where **g** and **H** are the gradient and Hessian of lse, and the vector of variational parameters $\boldsymbol{\psi}_i \in \mathbb{R}^M$ is chosen such that equality holds. An upper bound to lse can be found by replacing the Hessian matrix $\mathbf{H}(\boldsymbol{\psi}_i)$ with a matrix $\mathbf{A}_i$ such that $\mathbf{A}_i \prec \mathbf{H}(\boldsymbol{\psi}_i)$.This can be achieved if we use the matrix $\boldsymbol{A}_i = \frac{1}{2}\left[\boldsymbol{I}_M - \frac{1}{M+1}\mathbf{1}_M\mathbf{1}_M^T\right]$. Recall M+1=C is the number of classes.

Note that $\boldsymbol{A}_i$ is independent of $\boldsymbol{\psi}_i$, however, we still write it as $\boldsymbol{A}_i$ (rather than dropping the *i* subscript), since other bounds that we consider below will have a data-dependent curvature term. The upper bound on lse therefore becomes

$$lse(\boldsymbol{\eta}_i) \leq \frac{1}{2}\boldsymbol{\eta}_i^T \boldsymbol{A}_i \boldsymbol{\eta}_i - \boldsymbol{b}_i^T \boldsymbol{\eta}_i + c_i, \boldsymbol{b}_i = \boldsymbol{A}_i \boldsymbol{\psi}_i - \mathrm{g}(\boldsymbol{\psi}_i), c_i = \frac{1}{2}\boldsymbol{\psi}_i^T \boldsymbol{A}_i \boldsymbol{\psi}_i - \mathrm{g}(\boldsymbol{\psi}_i)^T \boldsymbol{\psi}_i + lse(\boldsymbol{\psi}_i)$$

- Bohning, D. (1992). Multinomial logistic regression algorithm. *Annals of the Inst. of Statistical Math. 44*, 197– 200

Bohning's Bound to the log-sum-exp

$$lse(\boldsymbol{\eta}_i) \leq \frac{1}{2}\boldsymbol{\eta}_i^T \boldsymbol{A}_i \boldsymbol{\eta}_i \; - \boldsymbol{b}_i^T \boldsymbol{\eta}_i \; + c_i$$

We can use the above result to get the following lower bound on the softmax likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \; = \prod_{i=1}^{N} exp(\boldsymbol{y}^T \boldsymbol{\eta}_i - \; lse(\boldsymbol{\eta}_i))$, $\eta_j = \boldsymbol{x}^T \boldsymbol{w}_j$, $\mathbf{X}_i$ = blockdiag($\boldsymbol{x}_i^T$):

$$lnp(y_i = c|\boldsymbol{x}_i, \boldsymbol{w}) \geq \left[\boldsymbol{y}_i^T \boldsymbol{X}_i \boldsymbol{w} - \frac{1}{2}\boldsymbol{w}^T \boldsymbol{X}_i \boldsymbol{A}_i \boldsymbol{X}_i \boldsymbol{w} + \boldsymbol{b}_i^T \boldsymbol{X}_i \boldsymbol{w} - c_i\right]_c$$

To simplify notation, define the pseudo-measurement

$$\tilde{\boldsymbol{y}}_i \equiv \boldsymbol{A}_i^{-1}(\boldsymbol{b}_i + \boldsymbol{y}_i)$$

Then we can get a "Gaussianized" version of the observation model:

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}) \geq f(\boldsymbol{x}_i, , \boldsymbol{\psi}_i)\mathcal{N}(\tilde{\boldsymbol{y}}_i|\boldsymbol{X}_i \boldsymbol{w}, \boldsymbol{A}_i^{-1})$$

where $f(\boldsymbol{x}_i, , \boldsymbol{\psi}_i)$ is some function that does not depend on **w**. Given this, it is easy to compute the posterior $q(\mathbf{w}) = \mathcal{N}(\mathbf{m}_N, \mathbf{V}_N)$, using Bayes rule for Gaussians. Below we will explain how to update the variational parameters $\boldsymbol{\psi}_i$.

Multinomial Logistic Regression

Let us see how to apply this bound to multinomial logistic regression. From J(q)=E_q [−log $p(D|\mathbf{x})$] + KL($q(\mathbf{x})||p(\mathbf{x})$), we can define the goal of variational inference as maximizing

$$L(q) = -\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|D)) + \mathbb{E}_q\left[\sum_{i=1}^N \log p(y_i|\boldsymbol{x}_i,\mathbf{w})\right] = -\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|D)) + \mathbb{E}_q\left[\sum_{i=1}^N \boldsymbol{y}_i^T\boldsymbol{\eta}_i - lse(\boldsymbol{\eta}_i)\right] = -\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|D)) + \sum_{i=1}^N \boldsymbol{y}_i^T\mathbb{E}_q[\boldsymbol{\eta}_i] - \sum_{i=1}^N \mathbb{E}_q[lse(\boldsymbol{\eta}_i)]$$

where $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{V}_N)$ is the approximate posterior. The first term is just the KL divergence between two Gaussians, which is given by

$$-KL\left(\mathcal{N}(\boldsymbol{m}_N,\boldsymbol{V}_N)\,||\mathcal{N}(\boldsymbol{m}_0,\boldsymbol{V}_0)\right) =$$

$$-\frac{1}{2}\left[tr\left(\boldsymbol{V}_N\boldsymbol{V}_0^{-1}\right) - \ln\left|\boldsymbol{V}_N\boldsymbol{V}_0^{-1}\right| + (\boldsymbol{m}_N-\boldsymbol{m}_0)^T V_0^{-1}(\boldsymbol{m}_N-\boldsymbol{m}_0) \quad - DM\right]$$

DM is the dimensionality of the Gaussian, and we assume a prior $p(\mathbf{w}) = \mathcal{N}(\boldsymbol{m}_0,\boldsymbol{V}_0)$ where typically $\boldsymbol{\mu}_0 = \mathbf{0}_{DM}$, and $\mathbf{V}_0$ is block diagonal. The second term using $\boldsymbol{\eta}_i = \boldsymbol{x}_i^T\boldsymbol{w}$ is simply

$$\sum_{i=1}^N \boldsymbol{y}_i^T\mathbb{E}_q[\boldsymbol{\eta}_i] = \sum_{i=1}^N \boldsymbol{y}_i^T\widetilde{\boldsymbol{m}}_i \;,\; \widetilde{\boldsymbol{m}}_i = \boldsymbol{X}_i\boldsymbol{m}_N$$

The final term can be lower bounded by taking expectations of our quadratic upper bound on lse $lse(\boldsymbol{\eta}_i) \le \frac{1}{2}\boldsymbol{\eta}_i^T\boldsymbol{A}_i\boldsymbol{\eta}_i - \boldsymbol{b}_i^T\boldsymbol{\eta}_i + c_i$ as follows:

$$-\sum_{i=1}^N \mathbb{E}_q[lse(\boldsymbol{\eta}_i)] \ge -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}_i\widetilde{\boldsymbol{V}}_i\right) - \frac{1}{2}\widetilde{\boldsymbol{m}}_i A_i\widetilde{\boldsymbol{m}}_i + \boldsymbol{b}_i^T\widetilde{\boldsymbol{m}}_i - c_i,\; \widetilde{\boldsymbol{V}}_i = \boldsymbol{X}_i\boldsymbol{V}_i\boldsymbol{X}_i^T$$

Multinomial Logistic Regression

Putting it all together we have:

$$L_{QJ}(q)$$

$$\geq -\frac{1}{2}\left[tr\left(\boldsymbol{V}_N \boldsymbol{V}_0^{-1}\right) - \ln\left|\boldsymbol{V}_N \boldsymbol{V}_0^{-1}\right| + (\boldsymbol{m}_N - \boldsymbol{m}_0)^T V_0^{-1}(\boldsymbol{m}_N - \boldsymbol{m}_0)\right] - \frac{DM}{2} + \sum_{i=1}^{N} \boldsymbol{y}_i^T \widetilde{\boldsymbol{m}}_i$$

$$-\frac{1}{2} tr\left(\boldsymbol{A}_i \widetilde{\boldsymbol{V}}_i\right) + \boldsymbol{b}_i^T \widetilde{\boldsymbol{m}}_i - c_i$$

This lower bound combines Jensen's inequality (as in mean field inference), plus the quadratic lower bound due to the lse term, so we write it as L_{QJ}.

We will use coordinate ascent to optimize this lower bound. That is, we update the variational posterior parameters $\mathbf{V}_N$ and $\mathbf{m}_N$, and then the variational likelihood parameters $\boldsymbol{\psi}_i$.
We just state the results.

$$\boldsymbol{V}_N = \left(\boldsymbol{V}_0^{-1} + \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{A}_i \boldsymbol{X}_i\right)^{-1}, \boldsymbol{m}_N = \boldsymbol{V}_N \left(\boldsymbol{V}_0^{-1} m_0 + \sum_{i=1}^{N} \boldsymbol{X}_i^T (\boldsymbol{y}_i + \boldsymbol{b}_i)\right), \boldsymbol{\psi}_i = \widetilde{\boldsymbol{m}}_i = \boldsymbol{X}_i \boldsymbol{m}_N$$

We can exploit the fact that $\mathbf{A}_i$ is a constant matrix plus the fact that $\mathbf{X}_i$ has block structure :

$$\boldsymbol{V}_N = \left(\boldsymbol{V}_0^{-1} + \boldsymbol{A} \otimes \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1}, \quad \boldsymbol{m}_N = \boldsymbol{V}_N \left(\boldsymbol{V}_0^{-1} m_0 + \sum_{i=1}^{N} (\boldsymbol{y}_i + \boldsymbol{b}_i) \otimes \boldsymbol{x}_i\right)$$

Multinomial Logistic Regression

1 Input: $y_i \in \{1, \ldots, C\}$, $\boldsymbol{x}_i \in \mathbb{R}^D$, $i = 1 : N$, prior $\mathbf{m}_0$, $\mathbf{V}_0$;
2 Define $M := C - 1$; dummy encode $\mathbf{y}^i \in \{0, 1\}^M$; define $\mathbf{X}_i = \text{blockdiag}(\boldsymbol{x}_i^T)$;
3 Define $\mathbf{y} := [\mathbf{y}_1; \ldots; \mathbf{y}_N]$, $\mathbf{X} := [\mathbf{X}_1; \ldots; \mathbf{X}_N]$ and $\boldsymbol{A} = \frac{1}{2}\left[\boldsymbol{I}_M - \frac{1}{M+1}\mathbf{1}_M\mathbf{1}_M^T\right]$
4 $\boldsymbol{V}_N = \left(\boldsymbol{V}_0^{-1} + \boldsymbol{A} \otimes \sum_{i=1}^N \boldsymbol{x}_i\, \boldsymbol{x}_i^T\right)^{-1}$
5 Initialize $\mathbf{m}_N := \mathbf{m}_0$;
6 repeat
7 $\boldsymbol{\psi} := \mathbf{X}\mathbf{m}_N$;
8 $\boldsymbol{\Psi} := \text{reshape}(\mathbf{m}, M, N)$;
9 $\mathbf{G} := \exp(\boldsymbol{\Psi} - \text{lse}(\boldsymbol{\Psi}))$;
10 $\mathbf{B} := \mathbf{A}\boldsymbol{\Psi} - \mathbf{G}$;
11 $\mathbf{b} := (\mathbf{B})$;
12 $\boldsymbol{m}_N = \boldsymbol{V}_N\left(\boldsymbol{V}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{X}^T(\boldsymbol{y} + \boldsymbol{b})\right)$
13 Compute the lower bound L_{QJ} using $L_{QJ}(q) \geq -\frac{1}{2}\left[tr\left(\boldsymbol{V}_N\boldsymbol{V}_0^{-1}\right) - \ln\left|\boldsymbol{V}_N\boldsymbol{V}_0^{-1}\right| + (\boldsymbol{m}_N - \boldsymbol{m}_0)^T V_0^{-1}(\boldsymbol{m}_N - \boldsymbol{m}_0)\right] - \frac{DM}{2} + \sum_{i=1}^N \boldsymbol{y}_i^T\tilde{\boldsymbol{m}}_i - \frac{1}{2}tr\left(A_i\tilde{V}_i\right) + b_i^T\tilde{\boldsymbol{m}}_i - c_i$
14 until *converged*;
15 Return $\mathbf{m}_N$ and $\mathbf{V}_N$;

http://www.cs.ubc.ca/~emtiyaz/software/catLGM.html

- Categorical Data Analysis with Latent Gaussian Models.
- Piecewise Bounds for Estimating Bernoulli-Logistic Latent Gaussian Models.
- Variational Bounds for Mixed-Data Factor Analysis.
- Blei and Lafferty 2003, Correlated topic models (PDF)

Download CatLGM.zip.

Alternative Bound for the Sigmoid Function

In many models, we just have binary data. In this case, we have $y_i \in \{0, 1\}$, $M = 1$ and $\eta_i = \mathbf{w}^T \mathbf{x}_i$ where $\mathbf{w} \in \mathbb{R}^D$ is a weight vector (not matrix). In this case, the Bohning bound $lse(\boldsymbol{\eta}_i) \leq \frac{1}{2}\boldsymbol{\eta}_i^T \boldsymbol{A}_i \boldsymbol{\eta}_i - \boldsymbol{b}_i^T \boldsymbol{\eta}_i + c_i, \boldsymbol{b}_i = \boldsymbol{A}_i \boldsymbol{\psi}_i - \mathrm{g}(\boldsymbol{\psi}_i), c_i = \frac{1}{2}\boldsymbol{\psi}_i^T \boldsymbol{A}_i \boldsymbol{\psi}_i - \mathrm{g}(\boldsymbol{\psi}_i)^T \boldsymbol{\psi}_i + lse(\boldsymbol{\psi}_i)$ becomes

$$\log(1 + e^{\eta}) \leq \tfrac{1}{2} a\eta^2 - b\eta + c$$
$$a = 1/4$$
$$b = A\psi - (1 + e^{-\psi})^{-1}$$
$$c = 1/2\, A\psi^2 - (1 + e^{-\psi})^{-1}\psi + \log(1 + e^{\psi})$$

We derived earlier an alternative quadratic bound for this case. This has the following form

$$\log(1 + e^{\eta}) \leq \lambda(\xi)(\eta^2 - \xi^2) + 1/2\,(\eta - \xi) + \log(1 + e^{\xi})$$
$$\lambda(\xi) = 1/4\xi \tanh(\xi/2) = 1/2\xi\,(\mathrm{sigm}(\xi) - \tfrac{1}{2})$$

We shall refer to this as the **JJ bound**, after its inventors. To facilitate comparison with Bohning's bound, let us rewrite the JJ bound as a quadratic form as follows

$$\log(1 + e^{\eta}) \leq \tfrac{1}{2}\; a(\xi)\eta^2 - b(\xi)\eta + c(\xi),\; a(\xi) = 2\lambda(\xi),\; b(\xi) = -1/2,\; c(\xi) = -\lambda(\xi)\xi^2 - \tfrac{1}{2}\,\xi + \log(1 + e^{\xi})$$

- Jaakkola, T. and M. Jordan (1996b). A variational approach to Bayesian logistic regression problems and their extensions. In *AI + Statistics*.
- Jaakkola, T. S. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing 10*, 25–37.

Alternative Bound for the Sigmoid Function

$\log(1 + e^{\eta}) \leq \frac{1}{2}\ a(\xi)\eta^2 - b(\xi)\eta + c(\xi)$, $a(\xi) = 2\lambda(\xi)$, $b(\xi) = -1/2$, $c(\xi) = -\lambda(\xi)\xi^2 - \frac{1}{2}\ \xi + \log(1 + e^{\xi})$

The JJ bound has an adaptive curvature term, since a depends on ξ. In addition, it is tight at two points. The Bohning bound is a constant curvature bound, and is only tight at one point.

We can use our general results and re-derive the results when using the JJ bound for binary logistic regression. First, we use the new definitions for a_i, b_i and c_i. The fact that a_i is not constant when using the JJ bound, unlike when using the Bohning bound, means we cannot compute $\mathbf{V}_N$ outside of the main loop, making the method a constant factor slower. Next we note that $\mathbf{X}_i = \boldsymbol{x}_i^T$, so using, $\boldsymbol{V}_N = \left(\boldsymbol{V}_0^{-1} + \sum_{i=1}^N \boldsymbol{X}_i^T \boldsymbol{A}_i \boldsymbol{X}_i\right)^{-1}$, $\boldsymbol{m}_N = \boldsymbol{V}_N\left(\boldsymbol{V}_0^{-1} m_0 + \sum_{i=1}^N \boldsymbol{X}_i^T\ (\boldsymbol{y}_i + \boldsymbol{b}_i)\right)$, the updates for the posterior become

$$\boldsymbol{V}_N^{-1} = \boldsymbol{V}_0^{-1} + 2\sum_{i=1}^N \lambda(\xi_i)\boldsymbol{x}_i\ \boldsymbol{x}_i^T, \boldsymbol{m}_N = \boldsymbol{V}_N\left(\boldsymbol{V}_0^{-1} m_0 + \sum_{i=1}^N \boldsymbol{x}_i\ \left(\boldsymbol{y}_i - \frac{1}{2}\right)\right)$$

Finally, to compute the update for ξ_i, we isolate the terms in L_{QJ} that depend on ξ_i to get

$$L(\boldsymbol{\xi}) = \sum_{i=1}^N \left(lnsigm(\xi_i) - \frac{\xi_i}{2} - \lambda(\xi_i)\left(\boldsymbol{x}_i^T \mathbb{E}_q[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{x}_i\ - \xi_i^2\right)\right) + const$$

Optimizing this wrt ξ_i gives the equation $\lambda'(\xi_i)\left(\boldsymbol{x}_i^T \mathbb{E}_q[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{x}_i\ - \xi_i^2\right)=0$. Now $\lambda'(\xi_i)$ is monotonic for $\xi_i \geq 0$, and we do not need to consider negative values of ξ_i by symmetry of the bound around $\xi_i = 0$. Hence the only way to make the above expression 0 is if we have $\boldsymbol{x}_i^T \mathbb{E}_q[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{x}_i\ = \xi_i^2$.

Alternative Bound for the Sigmoid Function

Hence the update becomes

$$(\xi_i^{new})^2 = \boldsymbol{x}_i^T(\boldsymbol{V}_N + \boldsymbol{m}_N\boldsymbol{m}_N^T)\boldsymbol{x}_i$$

Although the JJ bound is tighter than the Bohning bound, sometimes it is not tight enough in order to estimate the posterior covariance accurately.

A more accurate approach, which uses a piecewise quadratic upper bound to lse, is described in Marlin et al.. By increasing the number of pieces, the bound can be made arbitrarily tight.

- Marlin, B., E. Khan, and K. Murphy (2011). Piecewise Bounds for Estimating Bernoulli-Logistic Latent Gaussian Models. In *Intl. Conf. on Machine Learning*.

Product of Sigmoids

There are several other bounds and approximations to the multiclass lse function which we can use.

They require numerical optimization methods to compute $\mathbf{m}_N$ and $\mathbf{V}_N$, making them more complicated to implement.

The approach in Bouchard 2007 exploits the fact that

$$ln \sum_{k=1}^{K} e^{\eta_k} \leq \alpha + \sum_{k=1}^{K} ln(1 + e^{\eta_k - \alpha})$$

It then applies the JJ bound to the term on the right.

$$ln \sum_{k=1}^{K} e^{\eta_k} \leq \alpha + \sum_{k=1}^{K} \left(\frac{x_k - a - \xi_k}{2} + \lambda(\xi_k)\left((x_k - a)^2 - \xi_k^2\right) + ln\left(1 + e^{\xi_k}\right) \right)$$

where $\lambda(\xi) = \frac{1}{2\xi}\left(\frac{1}{1+e^{-\xi}} - \frac{1}{2}\right)$

- Bouchard, G. (2007). Efficient bounds for the softmax and applications to approximate inference in hybrid models. In *NIPS 2007 Workshop on Approximate Inference in Hybrid Models* (presentation)

Jensen's Inequality

The approach in (Blei and Lafferty 2006a, 2007) uses Jensen's inequality as follows where the last term follows from $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{V}_N)$ and that the mean of a log-normal distribution is $e^{\mu+\sigma^2/2}$.

$$\mathbb{E}_q[(lse(\boldsymbol{\eta}_i))] = \mathbb{E}_q\left[ln\left(1 + \sum_{c=1}^{M} exp(\boldsymbol{x}_i^T \boldsymbol{w}_c)\right)\right] \leq ln\left(1 + \sum_{c=1}^{M} \mathbb{E}_q\left[exp(\boldsymbol{x}_i^T \boldsymbol{w}_c)\right]\right)$$

$$\leq ln\left(1 + \sum_{c=1}^{M} exp\left(\boldsymbol{x}_i^T \boldsymbol{m}_{N,c} + \frac{1}{2}\boldsymbol{x}_i^T \boldsymbol{V}_{N,cc} \boldsymbol{x}_i\right)\right)$$

- Blei, D. and J. Lafferty (2006b). Dynamic topic models. In *Intl. Conf. on Machine Learning*, pp. 113–120.
- Blei, D. and J. Lafferty (2007). A Correlated Topic Model of "Science". *Annals of Applied Stat. 1*(1), 17–35.

Multivariate Delta Method

The approach in (Ahmed and Xing 2007; Braun and McAuliffe 2010) uses the **multivariate delta method**, which is a way to approximate moments of a function using a Taylor series expansion.

In more detail, let $f(\mathbf{w})$ be the function of interest. Using a second-order approximation around $\mathbf{m}$ we have

$$f(\mathbf{w}) \approx f(\mathbf{m}) + (\mathbf{w}-\mathbf{m})^T \mathbf{g}(\mathbf{w}-\mathbf{m}) + 1/2\ (\mathbf{w}-\mathbf{m})^T \mathbf{H}(\mathbf{w}-\mathbf{m})$$

where $\mathbf{g}$ and $\mathbf{H}$ are the gradient and Hessian evaluated at $\mathbf{m}$. If $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{V})$, we have

$$\mathbb{E}_q[f(\mathbf{w})] \approx f(\mathbf{m}) + \tfrac{1}{2}\,\mathrm{tr}[\mathbf{HV}]$$

If we use $f(\mathbf{w}) = \mathrm{lse}(\mathbf{X}_i\mathbf{w})$, we get

$$\mathbb{E}_q[\mathrm{lse}(\mathbf{X}_i\mathbf{w})] \approx \mathrm{lse}(\mathbf{X}_i\mathbf{m}) + 1/2 tr\left[\boldsymbol{X}_i \boldsymbol{H} \boldsymbol{X}_i^T \boldsymbol{V}\right]$$

where $\mathbf{g}$ and $\mathbf{H}$ for the lse function are defined as

$$\mathrm{g}(\boldsymbol{\psi}_i) = \exp[\boldsymbol{\psi}_i - lse(\boldsymbol{\psi}_i)] = \mathcal{S}(\boldsymbol{\psi}_i)$$
$$\mathbf{H}(\boldsymbol{\psi}_i) = diag\big(\mathrm{g}(\boldsymbol{\psi}_i)\big) - \mathrm{g}(\boldsymbol{\psi}_i)\mathrm{g}(\boldsymbol{\psi}_i)^T$$

- Ahmed, A. and E. Xing (2007). On tight approximate inference of the logistic-normal topic admixture model. In *AI/Statistics*.
- Braun, M. and J. McAuliffe (2010). Variational Inference for Large-Scale Models of Discrete Choice. *J. of the Am. Stat. Assoc. 105*(489), 324–335.

Variational Inference Based on Upper Bounds

So far, we have been concentrating on lower bounds. However, sometimes we need to use an upper bound. For example, (Saul et al. 1996) derives a mean field algorithm for sigmoid belief nets, which are DGMs in which each CPD is a logistic regression function (Neal 1992). Unlike the case of Ising models, the resulting MRF is not pairwise, but contains higher order interactions.

This makes the standard mean field updates intractable. In particular, they turn out to involve computing an expression which requires evaluating

$$\mathbb{E}\left[ln\left(1+e^{-\sum_{j\in pa_i} w_{ij}x_j}\right)\right] = \mathbb{E}\left[-ln\sigma\left(\boldsymbol{w}_i^T \boldsymbol{x}_{pa(i)}\right)\right]$$

(Notice the minus sign in front.)

Saul et al. derive an upper bound on the sigmoid function so as to make this update tractable, resulting in a monotonically convergent inference procedure.

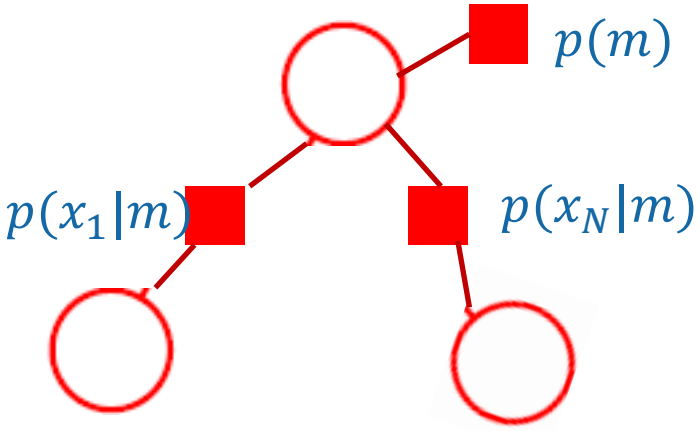
- Saul, L., T. Jaakkola, and M. Jordan (1996). Mean Field Theory for Sigmoid Belief Networks. *J. of AI Research 4*, 61–76.
- Neal, R. (1992). Connectionist learning of belief networks. *Artificial Intelligence 56*, 71–113.

Expectation Propagation

EP is another method to efficiently approximate posteriors. It is usually implemented in the context of factor graphs. It generalizes Assumed Density Filtering (from the Kalman Filtering community), and Loopy Belief Propagation.

For example, consider observations from a Gaussian with an unknown mean:

$$p(m|x_1, \ldots, x_N) \propto p(m) \prod_{i=1}^{N} p(x_i|m)$$



Given the data and factor graph above, what is the mean of x? One can address this with (1) Sampling, (2) Variable Elimination or (3) Message Passing (EP, Variational Bayes, etc.)

- Minka, T. (2001a). Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann.
- Minka, T. (2001b). *A family of approximate algorithms for Bayesian inference*. Ph. D. thesis, MIT.

The Clutter Problem

Suppose we have N i.i.d samples from the following distribution where all is known except from $\boldsymbol{\theta}$:

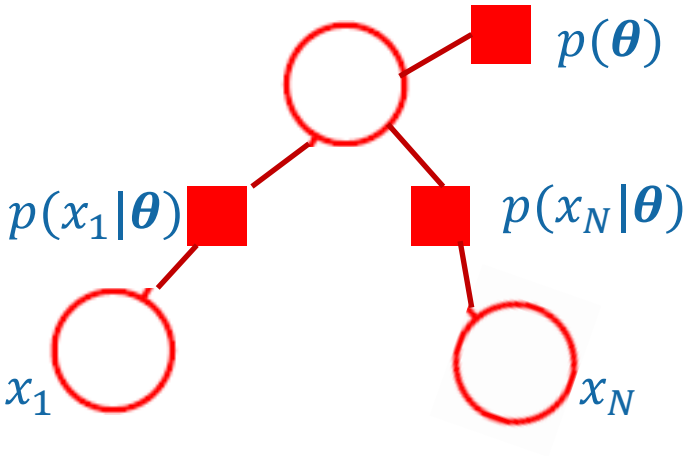
$$p(\boldsymbol{x}|\theta) = (1-w)\mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{I}) + w\mathcal{N}(\boldsymbol{x}|0, a\boldsymbol{I})$$

The prior taken over $\boldsymbol{\theta}$ is
$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{0}, b\boldsymbol{I})$

We take $a = 10, b = 100$ and $w = 0.5$

This gives the following joint distribution:

$$p(\mathcal{D}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(x_n|\boldsymbol{\theta})$$



- Minka, T. (2001b). *A family of approximate algorithms for Bayesian inference*. Ph. D. thesis, MIT.
- Minka, T. (2001a). Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann.
- Minka, T. (2008), EP: A Quick Reference

Recalling that the product of two Gaussians is a Gaussian, we note that one needs 2^N Gaussians to define $p(\mathcal{D}, \boldsymbol{\theta})$ and the posterior.

Still one expects with a lots of data, the exact posterior to look like a Gaussian.

Expectation Propagation

The expectation propagation algorithm is based on the idea of minimizing the reverse KL divergence wrt an approximating distribution $q(\boldsymbol{z})$ that is a member of the exponential family:

$$q(\boldsymbol{z};\boldsymbol{\eta}) = h(\boldsymbol{z})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T u(\boldsymbol{z}))$$

As a function of $\boldsymbol{\eta}$, the reverse KL divergence then becomes:

$$KL(p||q) = -\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T \mathbb{E}_{p(\boldsymbol{z})}[u(\boldsymbol{z})] + const$$

To minimize the reverse KL divergence we can set the gradient wrt $\boldsymbol{\eta}$ to zero.

$$Find\ \boldsymbol{\eta}^*\ s.t.: -\nabla \ln g(\boldsymbol{\eta}^*) = \mathbb{E}_{p(\boldsymbol{z})}$$

We know that for computing the moments of sufficient statistics for the exponential family, $-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}_{q(\boldsymbol{z})}[u(\boldsymbol{z})]$. Thus we conclude that the optimal solution is:

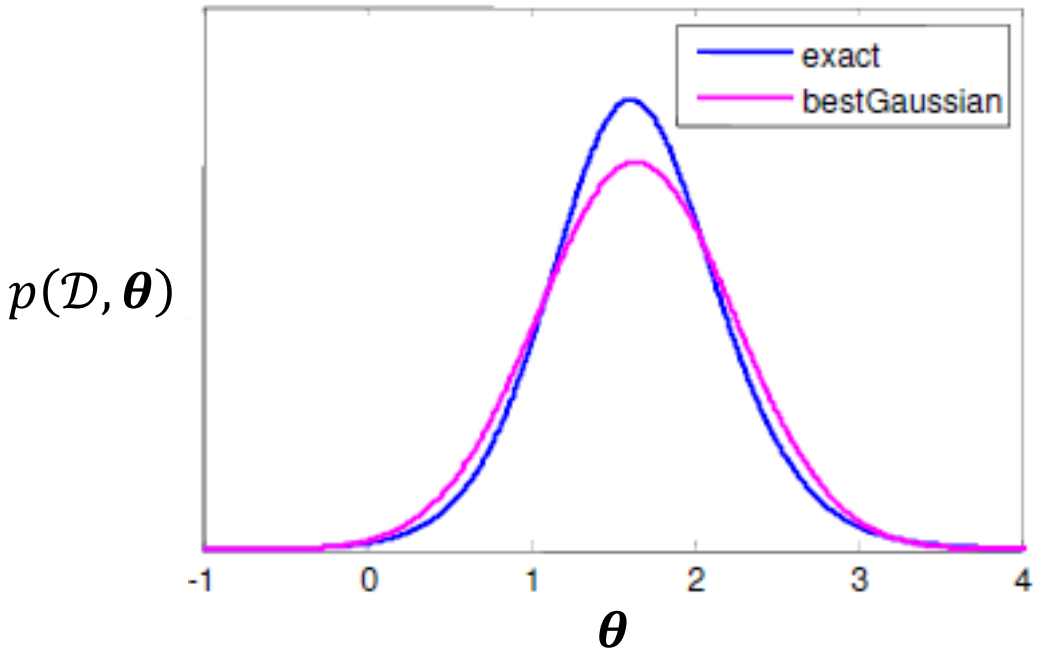
$$Choose\ \boldsymbol{\eta}^*\ s.t.: \int q(\boldsymbol{z};\boldsymbol{\eta}^*)u(\boldsymbol{z})d\boldsymbol{z} = \int p(\boldsymbol{z})u(\boldsymbol{z})d\boldsymbol{z}\ \ or\ \ \mathbb{E}_{q(\boldsymbol{z})}[u(\boldsymbol{z})] = \mathbb{E}_{p(\boldsymbol{z})}[u(\boldsymbol{z})]$$

Thus the optimum solution is obtained by 'moment matching'. E.g. if $q(\mathbf{z})$ is $N(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ then we set $\boldsymbol{\mu}$ equal to the mean of $p(\mathbf{z})$ and the covariance $\boldsymbol{\Sigma}$ equal to the covariance of $p(\mathbf{z})$.

- Minka, T. (2001a). Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann.
- Minka, T. (2001b). *A family of approximate algorithms for Bayesian inference*. Ph. D. thesis, MIT.

Best Gaussian by Moment Matching

Finding a Gaussian that fits the posterior by moment matching is easy for simple problems. However, in the clutter problem the posterior is a mixture of 2^N Gaussians.



- Approximate Inference, Tom Minka, Msft Research

Expectation Propagation

Now consider a model where the joint distribution is given by

$$p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta})$$

We denote here as $\boldsymbol{\theta}$ are the latent variables we hope to infer. E.g. i.i.d. data $f_i(\boldsymbol{\theta})$=p($\boldsymbol{x}_i$| $\boldsymbol{\theta}$) along with a prior $f_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ - this applies to directed graphical models with each factor being a conditional distribution for each node, and undirected graphs with each factor being a clique potential.

With posterior: $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \prod_i f_i(\boldsymbol{\theta})$

Model evidence: $p(\mathcal{D}) = \int \prod_i f_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}$

Then expectation propagation is based on the following approximate distribution:

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta})$$

Each $\tilde{f}_i(\boldsymbol{\theta})$ corresponds to a factor in the true posterior $\mathrm{f_i}(\boldsymbol{\theta})$ and is assumed to belong to the exponential family.

Thus the product of the factors also belongs to the exponential family and can be described by finite set of sufficient statistics.

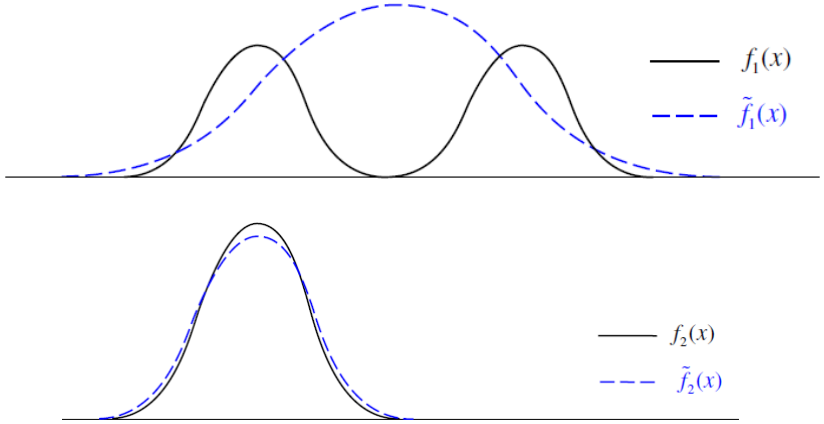
Expectation Propagation

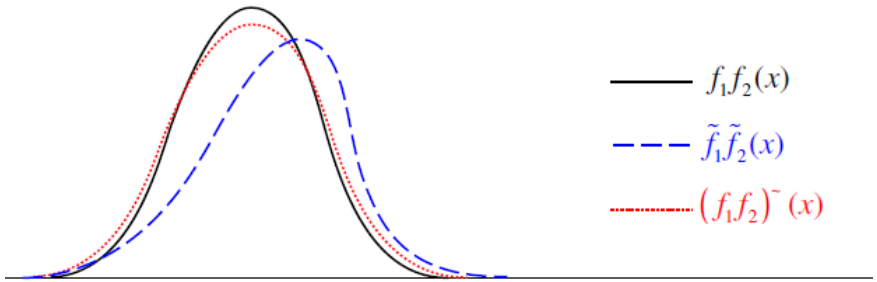
Minimization of the reverse KL divergence is analytically intractable since it involves averaging with respect to the true distribution.

$$KL(p||q) = KL\left(\frac{1}{p(\mathcal{D})}\prod_i f_i(\boldsymbol{\theta}) \;||\; \frac{1}{Z}\prod_i \tilde{f}_i(\boldsymbol{\theta})\right)$$

One could instead minimize the KL divergence between the pair $f_i(\boldsymbol{\theta})$ and $\tilde{f}_i(\boldsymbol{\theta})$ of factors.

This represents a simpler problem to solve, and is noniterative process. However, because each factor is individually approximated, the product of the factors could give a poor approximation.





Expectation Propagation overcomes this difficult by optimizing each factor in turn in the context of all the remaining factors. It cycles through the factors refining one at a time.

This is similar in nature to the Variational Bayes algorithm discussed previously.

Expectation Propagation

Suppose we want to refine a factor $\tilde{f}_i(\boldsymbol{\theta})$. We first remove the factor from the product and seek to determine a revised factor to ensure that the product

$$q(\boldsymbol{\theta}) \propto \tilde{f}_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$$

is as close as possible to

$$f_j(\boldsymbol{\theta}) \prod_i \tilde{f}_i(\boldsymbol{\theta})$$

Here we keep fixed the factors $\tilde{f}_i(\boldsymbol{\theta}), i \neq j$. **This ensures that the approximation is most accurate *in regions of high posterior probability as defined by the remaining factors.***

To achieve this, we define:

$$q^{\setminus j}(\boldsymbol{\theta}) = q(\boldsymbol{\theta}) / \tilde{f}_j(\boldsymbol{\theta})$$

This is combined with $f_j(\boldsymbol{\theta})$ to give the distribution.

$$\frac{1}{Z_j} f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta})$$

with normalizing constant

$$Z_j = \int f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

Expectation Propagation

We are now left with the easier task of minimizing the KL divergence between $\frac{1}{Z_j} f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$:

$$q^{new}(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})}{\operatorname{argmin}}\, KL\left(\frac{1}{Z_j} f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta}) \,||\, q(\boldsymbol{\theta})\right), \text{ where } q(\boldsymbol{\theta}) \propto \tilde{f}_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$$

As discussed previously, for $\tilde{f}_i(\boldsymbol{\theta})$ in the exponential family, this is just a case of moment matching.

The parameters of $q(\boldsymbol{\theta})$ can be found by matching its expected sufficient statistics to the corresponding moments of $\frac{1}{Z_j} f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta})$. From $q(\boldsymbol{\theta}) \propto \tilde{f}_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$, $\tilde{f}_j(\boldsymbol{\theta})$ can then be found as:

$$\tilde{f}_j(\theta) = \frac{K q^{new}(\boldsymbol{\theta})}{q^{\setminus j}(\boldsymbol{\theta})}$$

From the equation above and assuming that $q^{new}(\boldsymbol{\theta})$ is normalized:

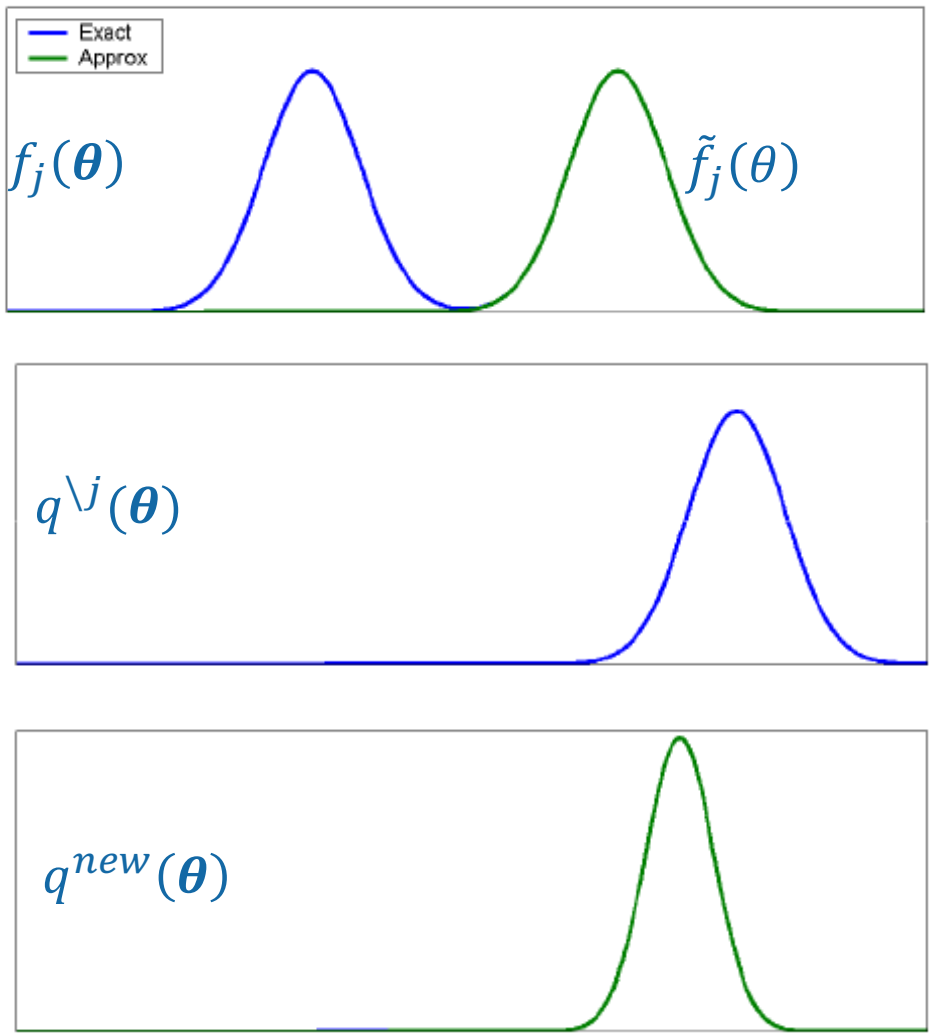
$$\int \tilde{f}_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int K q^{new}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = K$$

Using zero-order moment matching: $\mathrm{K} = \int \tilde{f}_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta}) d\boldsymbol{\theta} = Z_j$. Thus:

$$\tilde{f}_j(\theta) = \frac{Z_j q^{new}(\boldsymbol{\theta})}{q^{\setminus j}(\boldsymbol{\theta})}, \text{ where: } Z_j = \int f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Back to the Clutter Problem

- We approximate $f_j(\boldsymbol{\theta})$ in the context of $q^{\setminus j}(\boldsymbol{\theta})$
- We want in principle
$$f_j(\boldsymbol{\theta})q^{\setminus j}(\boldsymbol{\theta}) \approx \tilde{f}_j(\theta)q^{\setminus j}(\boldsymbol{\theta})$$
- Here $f_j(\boldsymbol{\theta})$ can be any distribution (mixture of two Gaussians) and $\tilde{f}_j(\theta)$ is taken as a Gaussian.
- Define $proj[p(\theta)] = \tilde{p}(\theta)$ as the operator that gives you a Gaussian with the same moments (mean and variance). We can then write:
$$\tilde{f}_j(\theta) = \frac{proj\left[f_j(\boldsymbol{\theta})q^{\setminus j}(\boldsymbol{\theta})\right]}{q^{\setminus j}(\boldsymbol{\theta})}$$
- The ratio of two Gaussians is a Gaussian – but note that the variance of $\tilde{f}_j(\theta)$ can be negative.



Gaussian Multiplication/Division Formulas

For multiplication:

$$\mathcal{N}(x; m_1, v_1)\mathcal{N}(x; m_2, v_2) = \mathcal{N}(m_1; m_2, v_1 + v_2)\mathcal{N}(x; m, v),$$

$$v = \frac{1}{\frac{1}{v_1} + \frac{1}{v_2}}, \qquad m = v\left(\frac{m_1}{v_1} + \frac{m_2}{v_2}\right)$$

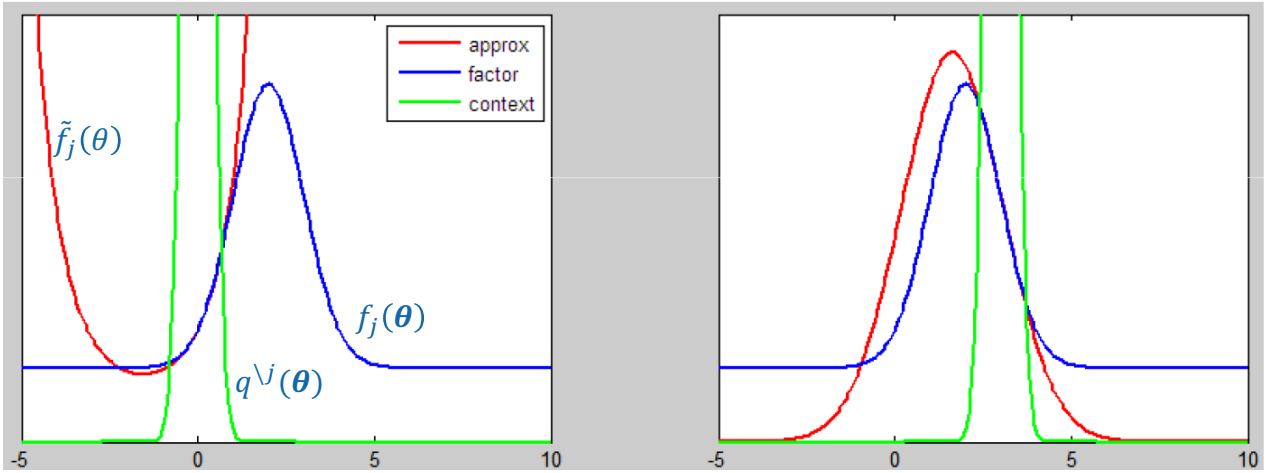
For division (note v can be negative):

$$\frac{\mathcal{N}(x; m_1, v_1)}{\mathcal{N}(x; m_2, v_2)} = \frac{v_2\mathcal{N}(x; m, v)}{(v_2 - v_1)\mathcal{N}(m_1; m_2, v_2 - v_1)},$$

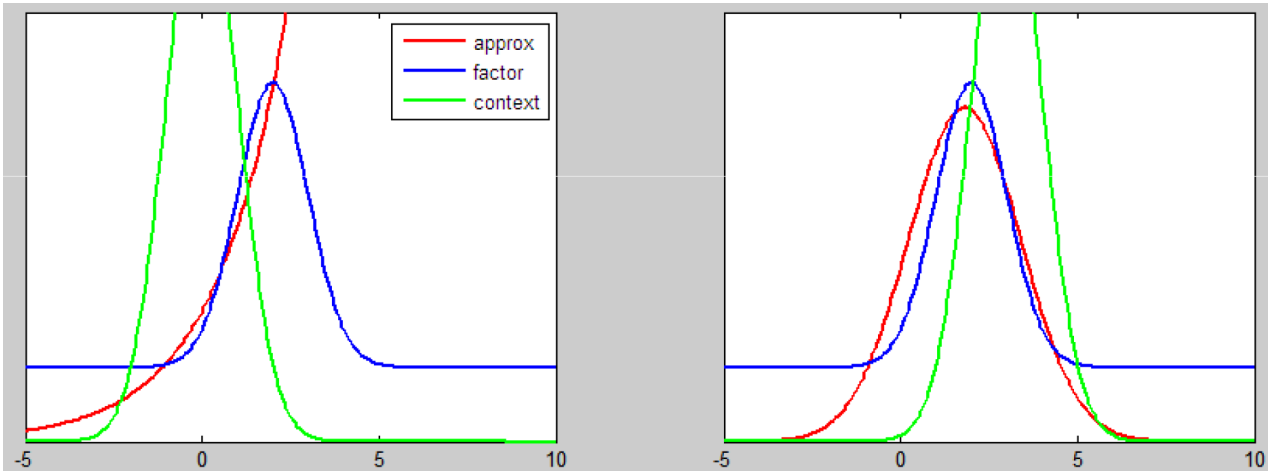
$$v = \frac{1}{\frac{1}{v_1} - \frac{1}{v_2}}, \qquad m = v\left(\frac{m_1}{v_1} - \frac{m_2}{v_2}\right)$$

Clutter Problem

- EP approximations with narrow context



- EP with medium context



Expectation Propagation: Algorithm

We are given $p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta})$ and wish to approximate the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ by a distribution $q(\boldsymbol{\theta}) = \frac{1}{Z}\prod_i \tilde{f}_i(\boldsymbol{\theta})$. The EP algorithm can be summarised as follows:

1) Initialize all $\tilde{f}_i(\boldsymbol{\theta})$ approximating factors

2) Initialize the posterior approximation $q(\boldsymbol{\theta}) \propto \prod_i \tilde{f}_i(\boldsymbol{\theta})$

3) Until convergence:

   a) Choose a factor $\tilde{f}_j(\boldsymbol{\theta})$ to refine

   b) Remove this factor from the posterior: $q^{\setminus j}(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/\tilde{f}_j(\boldsymbol{\theta})$

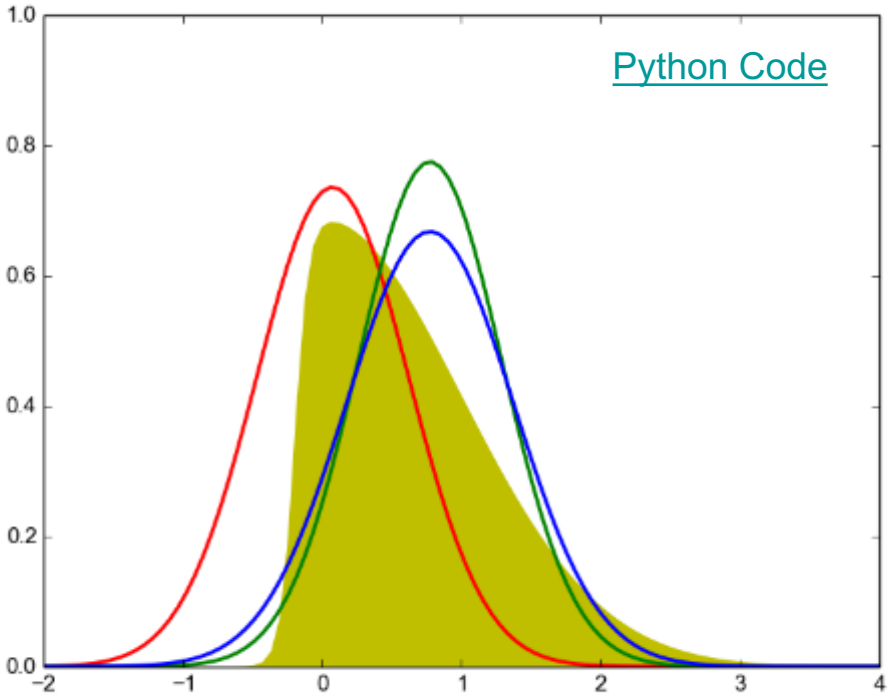
   c) Evaluate the new posterior by equating moments of $f_j(\boldsymbol{\theta})q^{\setminus j}(\boldsymbol{\theta})$ and $q^{new}(\boldsymbol{\theta})$ and evaluating $Z_j = \int f_j(\boldsymbol{\theta})q^{\setminus j}(\boldsymbol{\theta})d\boldsymbol{\theta}$

   d) Evaluate $\tilde{f}_j(\boldsymbol{\theta}) = \frac{Z_j q^{new}(\boldsymbol{\theta})}{q^{\setminus j}(\boldsymbol{\theta})}$

4) Approximate model evidence

$$p(\mathcal{D}) \sim \int \prod \tilde{f}_j(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Illustration of the EP Approximation



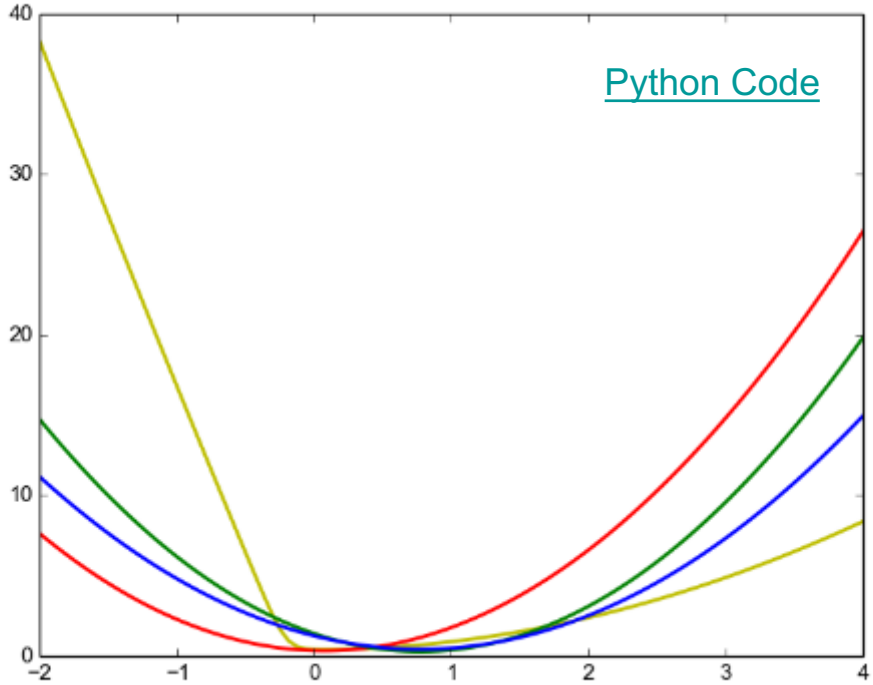


Illustration of the expectation propagation approximation using a Gaussian approximation for $p(z) \propto \exp(-\frac{z^2}{2})\sigma(20z + 4)$. The left-hand plot shows the original distribution (yellow) along with the Laplace (red), global variational (green), and EP (blue) approximations, and the right-hand plot shows the corresponding negative logarithms of the distributions.

Note that *the EP distribution is broader than that of the variational inference*, as a consequence of the different form of KL divergence.

Expectation Propagation

Consider the EP algorithm and suppose that one of the factors $f_0(\boldsymbol{\theta})$ in $p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta})$ has the same exponential family functional form as the approximating distribution $q(\boldsymbol{\theta})$.

We can show that if the factor $\tilde{f}_0(\boldsymbol{\theta})$ is initialized to be $f_0(\boldsymbol{\theta})$, then an EP update to refine $\tilde{f}_0(\boldsymbol{\theta})$ leaves $\tilde{f}_0(\boldsymbol{\theta})$ unchanged.

This situation typically arises when one of the factors is the prior $p(\boldsymbol{\theta})$, and so we see that the prior factor can be incorporated once exactly and does not need to be refined.

1) The initial $q(\boldsymbol{\theta})$ takes the form: $q_{init}(\boldsymbol{\theta}) \propto \tilde{f}_0(\boldsymbol{\theta}) \prod_{i \neq 0} \tilde{f}_i(\boldsymbol{\theta})$ with $\tilde{f}_0(\boldsymbol{\theta}) = f_0(\boldsymbol{\theta})$

2) Remove this factor from the posterior: $q^{\backslash 0}(\boldsymbol{\theta}) = q(\boldsymbol{\theta}) / \tilde{f}_0(\boldsymbol{\theta}) = \prod_{i \neq 0} \tilde{f}_i(\boldsymbol{\theta})$

3) Evaluate the new posterior $q^{new}(\boldsymbol{\theta})$ by equating sufficient statistics against $q^{\backslash 0}(\boldsymbol{\theta}) f_0(\boldsymbol{\theta}) = q_{init}(\boldsymbol{\theta})$. Since by definition this belongs to the same exponential family form as $q^{new}(\boldsymbol{\theta})$ it follows that:

$$q^{new}(\boldsymbol{\theta}) = q^{\backslash 0}(\boldsymbol{\theta}) f_0(\boldsymbol{\theta}) = q_{init}(\boldsymbol{\theta}).$$

4) Thus $\tilde{f}_0(\boldsymbol{\theta}) = \frac{Z_0 q^{new}(\boldsymbol{\theta})}{q^{\backslash 0}(\boldsymbol{\theta})} = Z_0 f_0(\boldsymbol{\theta})$, where $Z_0 = \int f_0(\boldsymbol{\theta}) q^{\backslash 0}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int q^{new}(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$ and $\tilde{f}_0(\boldsymbol{\theta}) = f_0(\boldsymbol{\theta})$.

Assumed Density Filtering

A special case of EP, known as *assumed density filtering* (ADF) or *moment matching*, is obtained by initializing all of the approximating factors except the first to unity and then making one pass through the factors updating each of them once.

Assumed density filtering can be *appropriate for on-line learning* in which data points are arriving in a sequence and we need to learn from each data point and then discard it before considering the next point.

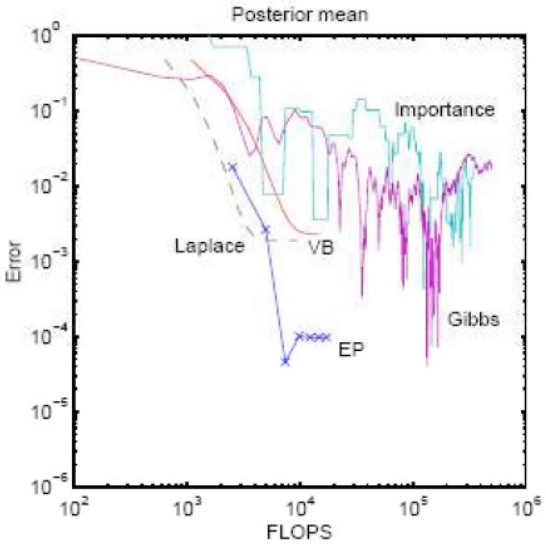
However, in a batch setting we have the opportunity to re-use the data points many times in order to achieve improved accuracy, and it is this idea that is exploited in expectation propagation.

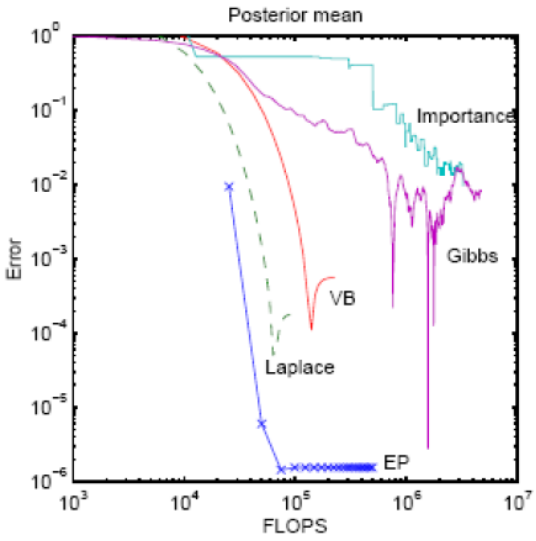
When applying ADF to batch data, the results will have an undesirable dependence on the (arbitrary) order in which the data points are considered, which again EP can overcome.

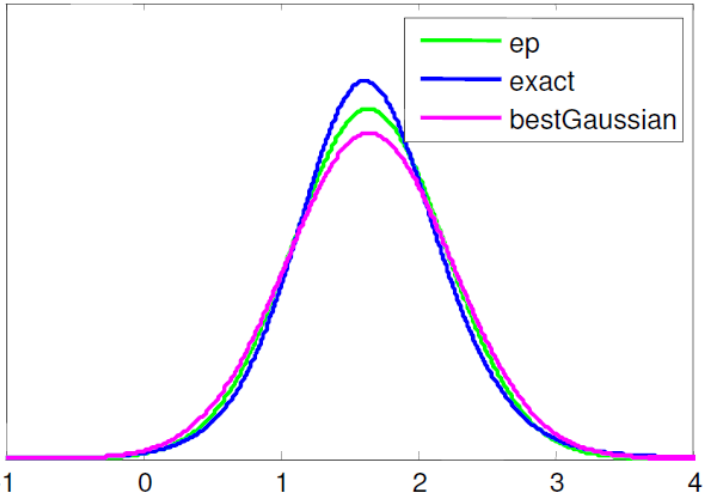
- Maybeck, P. S. (1982). *Stochastic models, estimation and control*. Academic Press.
- Lauritzen, S. L. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association* **87**, 1098–1108.
- Boyen, X. and D. Koller (1998). Tractable inference for complex stochastic processes. In G. F. Cooper and S. Moral (Eds.), *Proceedings 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 33–42. Morgan Kaufmann.
- Opper, M. and O. Winther (1999). A Bayesian approach to on-line learning. In D. Saad (Ed.), *On-Line Learning in Neural Networks*, pp. 363–378. Cambridge University Press. rch Cambridge.

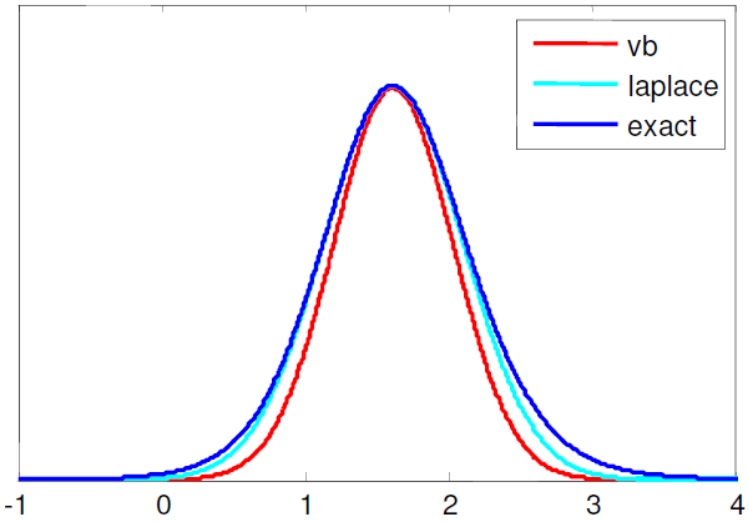
Cost Vs Accuracy

- Posterior mean:
  exact = 1.64864
  ep = 1.64514
  laplace = 1.61946
  vb = 1.61834
- Posterior variance:
  exact = 0.359673
  ep = 0.311474
  laplace = 0.234616
  vb = 0.171155









- Note errors remain the same for sampling methods as the number of data increases, while the error for EP becomes extremely small.

- Approximate Inference, Tom Minka, Msft Research

Assumed Density Filtering

Let us denote the joint distribution corresponding to the first j factors as $p_j(\boldsymbol{\theta}, \mathcal{D})$ with corresponding evidence $p_j(\mathcal{D})$. Then we can derive:

$$p_j(\mathcal{D}) = \int p_j(\boldsymbol{\theta}, \mathcal{D}) d\boldsymbol{\theta} = \int p_{j-1}(\boldsymbol{\theta}, \mathcal{D}) f_j(\boldsymbol{\theta}) d\boldsymbol{\theta} = p_{j-1}(\mathcal{D}) \int p_{j-1}(\boldsymbol{\theta}|\mathcal{D}) f_j(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$\cong p_{j-1}(\mathcal{D}) \int q_{j-1}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta}) d\boldsymbol{\theta} = p_{j-1}(\mathcal{D}) Z_j$$

This result applied recursively (initializing $p_0(\mathcal{D}) = 1$) leads to :

$$p(\mathcal{D}) = \prod_j Z_j$$

Thus in ADF, the evidence is given by the product of the normalization constants.

Convergence of Expectation Propagation

One disadvantage of EP is that there is no guarantee that the iterations will converge.

However, for approximations $q(\boldsymbol{\theta})$ in the exponential family, if the iterations do converge, the resulting solution will be a stationary point of a particular energy function (Minka, 2001a), although each iteration of EP does not necessarily decrease the value of this energy function.

Recall that in the variational Bayes, we iteratively maximize a lower bound on the log marginal likelihood, and in each iteration we guarantee not to decrease the bound.

It is possible to optimize the EP cost function directly, in which case it is guaranteed to converge, although the resulting algorithms can be slower and more complex to implement.

- Minka, T. (2001a). Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann.

Performance of Expectation Propagation

Another difference between variational Bayes and EP arises from the form of KL divergence that is minimized by the two algorithms, because the former minimizes KL($q||p$) whereas the latter minimizes KL($p||q$).

As we saw earlier for distributions $p(\boldsymbol{\theta})$ which are multimodal, minimizing KL($p||q$) can lead to poor approximations.

In particular, *if EP is applied to mixtures the results are not sensible because the approximation tries to capture all of the modes of the posterior distribution.*

Conversely, in logistic-type models, EP often out-performs both local variational methods and the Laplace approximation.

- Kuss, M. and C. Rasmussen (2006). Assessing approximations for Gaussian process classification. In *Advances in Neural Information Processing Systems*, Number 18. MIT Press. in press.

The Clutter Problem

In this example the goal is to infer the mean $\boldsymbol{\theta}$ of a multivariate Gaussian distribution over a variable **x** given a set of observations drawn from that distribution.

To make the problem more interesting, the observations are embedded in background clutter, which itself is also Gaussian distributed. The distribution of observed values **x** is therefore a mixture of Gaussians, which we take to be of the form

$$p(\boldsymbol{x}|\theta) = (1-w)\mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{I}) + w\mathcal{N}(\boldsymbol{x}|0, a\boldsymbol{I})$$

where w is the proportion of background clutter and is assumed to be known.

The prior over $\boldsymbol{\theta}$ is taken to be Gaussian

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, b\boldsymbol{I})$$

Following Minka, in the example to be shown later on are taken as a = 10, b = 100 and w = 0.5.

The joint distribution of N observations $\mathcal{D}$ = {$\mathbf{x}_1$, . . . , $\mathbf{x}_N$} and $\boldsymbol{\theta}$ is given by

$$p(\mathcal{D}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(x_n|\boldsymbol{\theta})$$

The posterior distribution comprises a mixture of 2^N Gaussians. Thus the computational cost of solving this problem exactly would grow exponentially with the size of the data set, and so an exact solution is intractable for moderately large N.

The Clutter Problem

Suppose we have N i.i.d samples from the following distribution where all is known except from $\boldsymbol{\theta}$:

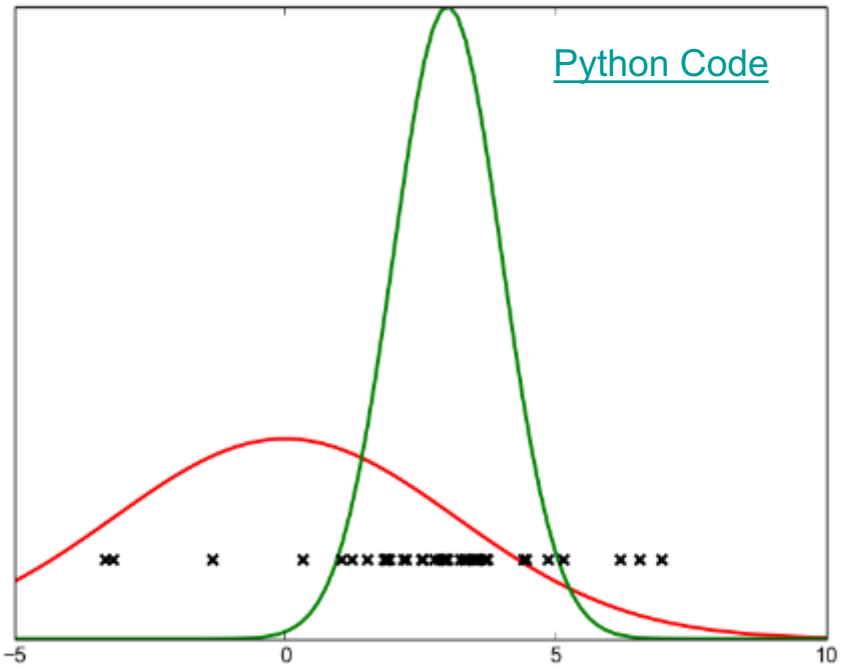
$$p(\boldsymbol{x}|\theta) = (1-w)\mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{I}) + w\mathcal{N}(\boldsymbol{x}|0, a\boldsymbol{I})$$

The prior taken over $\boldsymbol{\theta}$ is
$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, b\boldsymbol{I})$

We take $a = 10, b = 100$ and $w = 0.5$

This gives the following joint distribution:

$$p(\mathcal{D}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{n=1}^{N} p(x_n|\boldsymbol{\theta})$$



- Minka, T. (2001b). *A family of approximate algorithms for Bayesian inference*. Ph. D. thesis, MIT.
- Minka, T. (2001a). Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann.
- Minka, T. (2008), EP: A Quick Reference

The Clutter Problem

To apply EP to the clutter problem, we first identify the factors $f_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ and $f_n(\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})$.

Next we select an approximating distribution from the exponential family, and here we choose a spherical Gaussian

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, v\mathbf{I})$$

The factor approximations will therefore take the form of exponential-quadratic functions of the form

$$\tilde{f}_i(\boldsymbol{\theta}) = s_n exp\left(-\frac{1}{2v_n}(\boldsymbol{\theta} - \boldsymbol{m}^n)^T(\boldsymbol{\theta} - \boldsymbol{m}^n)\right)$$

where $n = 1, \ldots, N$, and we set $\tilde{f}_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$.

Note: You can write this alternatively as $\tilde{f}_i(\boldsymbol{\theta}) = s_n\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}_n, v_n\boldsymbol{I})$ but note $\mathcal{N}(\boldsymbol{\theta}|\cdot, \cdot)$ does not imply that the rhs is a well-defined Gaussian density (in fact the variance parameter v_n can be negative) but is simply a convenient shorthand notation.

For $\tilde{f}_i(\boldsymbol{\theta}) = s_n\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}_n, v_n\boldsymbol{I})$, the approximations $\tilde{f}_i(\boldsymbol{\theta})$ for $n = 1, \ldots, N$, can be initialized to unity, corresponding to $s_n = (2\pi v_n)^{D/2}$, $v_n \to \infty$ and $\boldsymbol{m}_n = 0$, where D is the dimensionality of $\mathbf{x}$ and hence of $\boldsymbol{\theta}$. The initial $q(\boldsymbol{\theta}) = \frac{1}{Z}\prod_i \tilde{f}_i(\boldsymbol{\theta})$ is therefore equal to the prior.

The Clutter Problem

We then iteratively refine the factors by taking one factor $\tilde{f}_i(\boldsymbol{\theta})$ at a time and applying

$$q^{\backslash j}(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/\tilde{f}_j(\boldsymbol{\theta})$$
$$Z_j = \int f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
$$\tilde{f}_j(\boldsymbol{\theta}) = \frac{Z_j q^{new}(\boldsymbol{\theta})}{q^{\backslash j}(\boldsymbol{\theta})}$$

Note that we do not need to revise the term $f_0(\boldsymbol{\theta})$ because an EP update will leave this term unchanged.

First we remove the current estimate $\tilde{f}_n(\boldsymbol{\theta})$ from $q(\boldsymbol{\theta})$ using $q^{\backslash n}(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/\tilde{f}_n(\boldsymbol{\theta})$ to give $q^{\backslash n}(\boldsymbol{\theta}) \propto exp\left\{-\frac{1}{2v}||\boldsymbol{\theta}-\boldsymbol{m}||^2 + \frac{1}{2v_n}||\boldsymbol{\theta}-\boldsymbol{m}_n||^2\right\} \propto exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}\left(\frac{1}{v}-\frac{1}{v_n}\right) + \boldsymbol{\theta}^T\left(\frac{1}{v}\boldsymbol{m} - \frac{1}{v_n}\boldsymbol{m}_n\right)\right\}$

which has inverse variance and mean given by

$$\left(v^{\backslash n}\right)^{-1} = v^{-1} - v_n^{-1}$$
$$\boldsymbol{m}^{\backslash n} = v^{\backslash n}(v^{-1}\boldsymbol{m} - v_n^{-1}\boldsymbol{m}_n) = v^{\backslash n}(v^{-1}\boldsymbol{m} - v_n^{-1}\boldsymbol{m}_n) + v^{\backslash n}v_n^{-1}\boldsymbol{m} - v^{\backslash n}v_n^{-1}\boldsymbol{m}$$
$$= \boldsymbol{m} + v^{\backslash n}v_n^{-1}(\boldsymbol{m} - \boldsymbol{m}_n)$$

We evaluate the normalization constant Z_n using $Z_n = \int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{m}^{\backslash n}, v^{\backslash n}\boldsymbol{I}\right)\left((1-w)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\theta}, \boldsymbol{I}) + w\mathcal{N}(\boldsymbol{x}_n|0, a\boldsymbol{I})\right)d\boldsymbol{\theta} = (1-w)\mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{m}^{\backslash n}, \left(v^{\backslash n}+1\right)\boldsymbol{I}\right) + w\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{0}, \alpha\boldsymbol{I})$ where for the 1st term we use familiar eqs for linear Gaussian models.

The Clutter Problem

We compute the mean of $q^{new}(\boldsymbol{\theta})$ by finding the mean and variance of $q^{\backslash n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})$ to give

$$\boldsymbol{m}^{new} = \boldsymbol{m}^{\backslash n} + \rho_n \frac{v^{\backslash n}}{v^{\backslash n} + 1}\left(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}\right)$$

where: $\rho_n = 1 - \frac{w}{Z_n}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{0}, \alpha\boldsymbol{I})$ is interpreted as the probability of the point $\mathbf{x}_n$ not being clutter.

To prove the Eq. for $\boldsymbol{m}^{new}$, using $q^{new}(\boldsymbol{\theta}) = \frac{q^{\backslash n}(\boldsymbol{\theta})\boldsymbol{f_n}(\boldsymbol{\theta})}{Z_n}$, note the following:

$$\nabla_{\boldsymbol{m}^{\backslash n}} lnZ_n = \frac{1}{Z_n}\nabla_{\boldsymbol{m}^{\backslash n}} \int q^{\backslash n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})\mathrm{d}\,\boldsymbol{\theta} = \frac{1}{Z_n}\int q^{\backslash n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})\left(-\frac{1}{v^{\backslash n}}\left(\boldsymbol{m}^{\backslash n} - \boldsymbol{\theta}\right)\right)\mathrm{d}\,\boldsymbol{\theta} = -\frac{\boldsymbol{m}^{\backslash n}}{v^{\backslash n}} + \frac{\mathbb{E}[\boldsymbol{\theta}]}{v^{\backslash n}}$$

where $\mathbb{E}[\boldsymbol{\theta}] = \boldsymbol{m}^{new}$ with $q^{new}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}^{new}, v^{new}\boldsymbol{I})$.

By substituting $Z_n = (1-w)\mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{m}^{\backslash n}, \left(v^{\backslash n}+1\right)\boldsymbol{I}\right) + w\mathcal{N}(x_n|0, \alpha\boldsymbol{I})$, we simplify as:

$$\nabla_{\boldsymbol{m}^{\backslash n}} lnZ_n = \frac{1}{Z_n}(1-w)\mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{m}^{\backslash n}, \left(v^{\backslash n}+1\right)\boldsymbol{I}\right)\frac{1}{v^{\backslash n}+1}\left(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}\right) = \rho_n \frac{1}{v^{\backslash n}+1}\left(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}\right)$$

where we defined:

$$\rho_n = \frac{1}{Z_n}(1-w)\mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{m}^{\backslash n}, \left(v^{\backslash n}+1\right)\boldsymbol{I}\right) = 1 - \frac{1}{Z_n}w\mathcal{N}(x_n|0, a\boldsymbol{I})$$

Thus: $-\frac{\boldsymbol{m}^{\backslash n}}{v^{\backslash n}} + \frac{\mathbb{E}[\boldsymbol{\theta}]}{v^{\backslash n}} = \rho_n \frac{\mathbf{1}}{v^{\backslash n}+1}\left(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}\right)$ from which:

$$\boldsymbol{m}^{new} \equiv \mathbb{E}[\boldsymbol{\theta}] = \boldsymbol{m}^{\backslash n} + \rho_n \frac{v^{\backslash n}}{v^{\backslash n}+1}\left(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}\right)$$

The Clutter Problem

Similarly, we compute the variance of $q^{\text{new}}(\boldsymbol{\theta})$ by finding the variance of $q^{\backslash n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})$ to give

$$v^{new} = v^{\backslash n} - \rho_n \frac{(v^{\backslash n})^2}{v^{\backslash n}+1} + \rho_n(1-\rho_n)\frac{(v^{\backslash n})^2}{D(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2$$

Note:

$$\nabla_{v^{\backslash n}} lnZ_n = \frac{1}{Z_n}\nabla_{v^{\backslash n}} \int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \frac{1}{Z_n}\int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \left(\frac{1}{2(v^{\backslash n})^2}(\boldsymbol{m}^{\backslash n} - \boldsymbol{\theta})^T(\boldsymbol{m}^{\backslash n} - \boldsymbol{\theta}) - \frac{D}{2v^{\backslash n}}\right) \mathrm{d}\boldsymbol{\theta} = \frac{1}{2(v^{\backslash n})^2}\{\mathbb{E}[\boldsymbol{\theta}^T\boldsymbol{\theta}] - 2\mathbb{E}[\boldsymbol{\theta}^T]\boldsymbol{m}^{\backslash n} + ||\boldsymbol{m}^{\backslash n}||^2\} - \frac{D}{2v^{\backslash n}}$$

We can also derive: $\nabla_{v^{\backslash n}} lnZ_n = \frac{1}{Z_n}(1-w)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{m}^{\backslash n}, (v^{\backslash n}+1)\boldsymbol{I})\left(\frac{1}{2(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2 - \frac{D}{2(v^{\backslash n}+1)}\right) = \rho_n\left(\frac{1}{2(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2 - \frac{D}{2(v^{\backslash n}+1)}\right)$

The needed variance is: $v^{new}\boldsymbol{I} = \mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^T] - \mathbb{E}[\boldsymbol{\theta}]\mathbb{E}[\boldsymbol{\theta}^T]$ and taking the trace:

$$v^{new}\,\mathrm{D} = \mathbb{E}[\boldsymbol{\theta}^T\boldsymbol{\theta}] - \mathbb{E}[\boldsymbol{\theta}^T]\mathbb{E}[\boldsymbol{\theta}]$$

The Clutter Problem

It remains to combine the following Eqs. $v^{new} = \mathbb{E}[\boldsymbol{\theta}^T\boldsymbol{\theta}] - \mathbb{E}[\boldsymbol{\theta}^T]\mathbb{E}[\boldsymbol{\theta}]$ and

$$\frac{1}{2(v^{\backslash n})^2}\left\{\mathbb{E}[\boldsymbol{\theta}^T\boldsymbol{\theta}] - 2\mathbb{E}[\boldsymbol{\theta}^T]\boldsymbol{m}^{\backslash n} + ||\boldsymbol{m}^{\backslash n}||^2\right\} - \frac{D}{2v^{\backslash n}}$$

$$= \rho_n\left(\frac{1}{2(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2 - \frac{D}{2(v^{\backslash n}+1)}\right)$$

$$v^{new} = \rho_n\left(\frac{2(v^{\backslash n})^2}{2(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2 - \frac{2(v^{\backslash n})^2 D}{2(v^{\backslash n}+1)}\right) + 2\mathbb{E}[\boldsymbol{\theta}^T]\boldsymbol{m}^{\backslash n} - \left|\left|\boldsymbol{m}^{\backslash n}\right|\right|^2 + D\,v^{\backslash n} - \mathbb{E}[\boldsymbol{\theta}^T]\mathbb{E}[\boldsymbol{\theta}]$$

where $\mathbb{E}[\boldsymbol{\theta}] = \boldsymbol{m}^{\backslash n} + \rho_n\frac{v^{\backslash n}}{v^{\backslash n}+1}(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n})$ from which:

$$\mathbb{E}[\boldsymbol{\theta}^T]\mathbb{E}[\boldsymbol{\theta}] = (\rho_n)^2\frac{(v^{\backslash n})^2}{(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2 + (\boldsymbol{m}^{\backslash n})^T\boldsymbol{m}^{\backslash n} + 2\rho_n\frac{v^{\backslash n}}{v^{\backslash n}+1}(\boldsymbol{m}^{\backslash n})^T(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n})$$

From the last three equations, we can derive:

$$v^{new} = \rho_n\left(\frac{2(v^{\backslash n})^2}{2(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2 - \frac{2(v^{\backslash n})^2 D}{2(v^{\backslash n}+1)}\right) + 2(\boldsymbol{m}^{\backslash n})^T\left(\boldsymbol{m}^{\backslash n} + \rho_n\frac{v^{\backslash n}}{v^{\backslash n}+1}(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n})\right) -$$

$$\left|\left|\boldsymbol{m}^{\backslash n}\right|\right|^2 - D\,v^{\backslash n} - (\rho_n)^2\frac{(v^{\backslash n})^2}{(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2 - (\boldsymbol{m}^{\backslash n})^T\boldsymbol{m}^{\backslash n} - 2\rho_n\frac{v^{\backslash n}}{v^{\backslash n}+1}(\boldsymbol{m}^{\backslash n})^T(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n})$$

which is simplified as: $v^{new} = v^{\backslash n} - \rho_n\frac{(v^{\backslash n})^2}{v^{\backslash n}+1} + \rho_n(1-\rho_n)\frac{(v^{\backslash n})^2}{D(v^{\backslash n}+1)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2$

The Clutter Problem

Thus the mean and variance of $q^{new}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}^{new}, v^{new}\boldsymbol{I})$ by finding the mean and variance of $q^{\backslash n}(\boldsymbol{\theta})f_n(\boldsymbol{\theta})$ is;

$$\boldsymbol{m}^{new} = \boldsymbol{m}^{\backslash n} + \rho_n \frac{v^{\backslash n}}{v^{\backslash n} + 1}\left(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}\right)$$

$$v^{new} = v^{\backslash n} - \rho_n \frac{\left(v^{\backslash n}\right)^2}{v^{\backslash n}+1} + \rho_n(1-\rho_n)\frac{\left(v^{\backslash n}\right)^2}{D\left(v^{\backslash n}+1\right)^2}||\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}||^2$$

where: $\rho_n = 1 - \frac{w}{Z_n}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{0}, \alpha\boldsymbol{I})$ has a simple interpretation as the probability of the point $\mathbf{x}_n$ not being clutter. Then we use $\tilde{f}_n(\boldsymbol{\theta}) = \frac{Z_n q^{new}(\boldsymbol{\theta})}{q^{\backslash n}(\boldsymbol{\theta})}$ and $q^{\backslash n}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}^{\backslash n}, v^{\backslash n}\boldsymbol{I})$, $\left(v^{\backslash n}\right)^{-1} = v^{-1} - v_n^{-1}$, $\boldsymbol{m}^{\backslash n} = \boldsymbol{m} + v^{\backslash n}v_n^{-1}(\boldsymbol{m} - \boldsymbol{m}_n)$ to compute the refined factor $\tilde{f}_n(\boldsymbol{\theta}) = s_n exp\left(-\frac{1}{2v_n}(\boldsymbol{\theta} - \boldsymbol{m}^n)^T(\boldsymbol{\theta} - \boldsymbol{m}^n)\right)$ whose parameters are given by

$$v_n^{-1} = (v^{new})^{-1} - \left(v^{\backslash n}\right)^{-1}$$

$$\boldsymbol{m}_n = \boldsymbol{m}^{\backslash n} + \left(v_n + v^{\backslash n}\right)\left(v^{\backslash n}\right)^{-1}\left(\boldsymbol{m}^{new} - \boldsymbol{m}^{\backslash n}\right)$$

$$s_n = \frac{Z_n}{\left(2\pi v_n\right)^{D/2}\mathcal{N}\left(\boldsymbol{m}_n|\boldsymbol{m}^{\backslash n}, \left(v_n + v^{\backslash n}\right)\boldsymbol{I}\right)}$$

The first two equations are obtained in the order they are presented above directly by substituting $q^{new}(\boldsymbol{\theta})$ and $q^{\backslash n}(\boldsymbol{\theta})$ at $\tilde{f}_n(\boldsymbol{\theta}) = \frac{Z_n q^{new}(\boldsymbol{\theta})}{q^{\backslash n}(\boldsymbol{\theta})}$ and completing the square. We prove next the Eq. for s_n.

The Clutter Problem

Using $\tilde{f}_n(\boldsymbol{\theta}) = \frac{Z_n q^{new}(\boldsymbol{\theta})}{q^{\backslash n}(\boldsymbol{\theta})}$ and $\tilde{f}_n(\boldsymbol{\theta}) = s_n exp\left(-\frac{1}{2v_n}(\boldsymbol{\theta} - \boldsymbol{m}^n)^T(\boldsymbol{\theta} - \boldsymbol{m}^n)\right)$ and the expressions for $q^{new}(\boldsymbol{\theta})$ and $q^{\backslash n}(\boldsymbol{\theta})$ we can write:

$$s_n = Z_n \left(\frac{2\pi v^{\backslash n}}{2\pi v^{new}}\right)^{D/2} exp\left(-\frac{(\boldsymbol{m}^{new})^T \boldsymbol{m}^{new}}{2v^{new}} + \frac{(\boldsymbol{m}^{\backslash n})^T \boldsymbol{m}^{\backslash n}}{2v^{\backslash n}} + \frac{(\boldsymbol{m}^n)^T \boldsymbol{m}^n}{2v_n}\right)$$

In the exponential, the first two terms come from $q^{new}(\boldsymbol{\theta})$ and $q^{\backslash n}(\boldsymbol{\theta})$ (not used to complete the square) and the last term from the $\pm\frac{(\boldsymbol{m}^n)^T \boldsymbol{m}^n}{2v_n}$ term needed to complete the square in $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}_n, v_n \boldsymbol{I})$. We will simplify by using: $\frac{1}{v^{new}} = \frac{1}{v_n} + \frac{1}{v^{\backslash n}}$ or $\frac{v^{\backslash n}}{v^{new}} = \frac{v_n + v^{\backslash n}}{v_n}$ and using $\boldsymbol{m_n} = \boldsymbol{m}^{\backslash n} + (v_n + v^{\backslash n})(v^{\backslash n})^{-1}(\boldsymbol{m}^{new} - \boldsymbol{m}^{\backslash n})$, $\boldsymbol{m}^{new} = \frac{\boldsymbol{m_n} - \boldsymbol{m}^{\backslash n}}{v_n + v^{\backslash n}} v^{\backslash n} + \boldsymbol{m}^{\backslash n} = \frac{\boldsymbol{m_n} v^{\backslash n} + \boldsymbol{m}^{\backslash n} v_n}{v_n + v^{\backslash n}}$

$$s_n = Z_n \left(\frac{v_n + v^{\backslash n}}{v_n}\right)^{D/2} exp\left(-\frac{v_n + v^{\backslash n}}{2v_n v^{\backslash n}}\left(\frac{\boldsymbol{m_n} v^{\backslash n} + \boldsymbol{m}^{\backslash n} v_n}{v_n + v^{\backslash n}}\right)^T \left(\frac{\boldsymbol{m_n} v^{\backslash n} + \boldsymbol{m}^{\backslash n} v_n}{v_n + v^{\backslash n}}\right) + \frac{(\boldsymbol{m}^{\backslash n})^T \boldsymbol{m}^{\backslash n}}{2v^{\backslash n}} + \frac{(\boldsymbol{m}^n)^T \boldsymbol{m}^n}{2v_n}\right)$$

or finally: $s_n = Z_n \left(\frac{2\pi(v_n + v^{\backslash n})}{2\pi v_n}\right)^{D/2} exp\left(\frac{1}{2(v_n + v^{\backslash n})}(\boldsymbol{m_n} - \boldsymbol{m}^{\backslash n})^T(\boldsymbol{m_n} - \boldsymbol{m}^{\backslash n})\right) = \frac{Z_n}{(2\pi v_n)^{D/2}} \mathcal{N}(\boldsymbol{m_n}|\boldsymbol{m}^{\backslash n}, (v_n + v^{\backslash n})\boldsymbol{I})$

The Clutter Problem

Thus the refined factor $\tilde{f}_n(\boldsymbol{\theta}) = s_n exp\left(-\frac{1}{2v_n}(\boldsymbol{\theta}-\boldsymbol{m}^n)^T(\boldsymbol{\theta}-\boldsymbol{m}^n)\right)$ has parameters given by

$$v_n^{-1} = (v^{new})^{-1} - \left(v^{\backslash n}\right)^{-1}$$

$$\boldsymbol{m_n} = \boldsymbol{m}^{\backslash n} + \left(v_n + v^{\backslash n}\right)\left(v^{\backslash n}\right)^{-1}\left(\boldsymbol{m}^{new} - \boldsymbol{m}^{\backslash n}\right)$$

$$s_n = \frac{Z_n}{\left(2\pi v_n\right)^{D/2}\mathcal{N}\left(\boldsymbol{m_n}\middle|\boldsymbol{m}^{\backslash n}, \left(v_n + v^{\backslash n}\right)\boldsymbol{I}\right)}$$

This refinement process is repeated until the maximum change in parameter values resulting from a complete pass through all factors is less than some threshold.

The Clutter Problem

Finally using $p(\mathcal{D}) \approx \int \prod_i \tilde{f}_j(\boldsymbol{\theta}) d\boldsymbol{\theta}$ with $\tilde{f}_n(\boldsymbol{\theta}) = s_n exp\left(-\frac{1}{2v_n}(\boldsymbol{\theta} - \boldsymbol{m}^n)^T(\boldsymbol{\theta} - \boldsymbol{m}^n)\right)$ to approximate the model evidence:

$$p(\mathcal{D}) \approx (2\pi v^{new})^{\frac{D}{2}} \exp(\frac{B}{2}) \prod_{n=1}^{N} \left\{ s_n \left(2\pi v_n\right)^{-D/2} \right\}$$

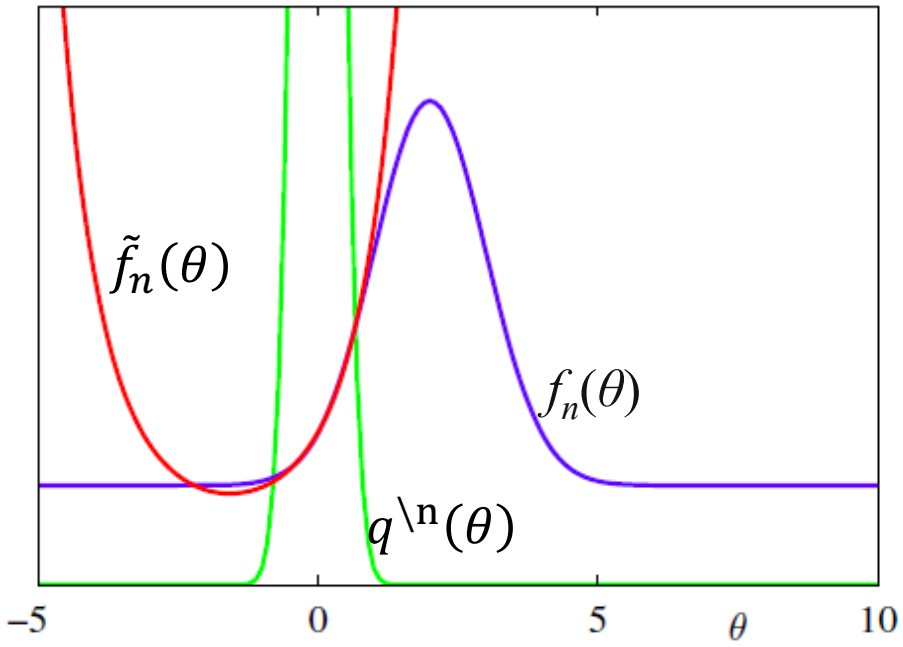
where

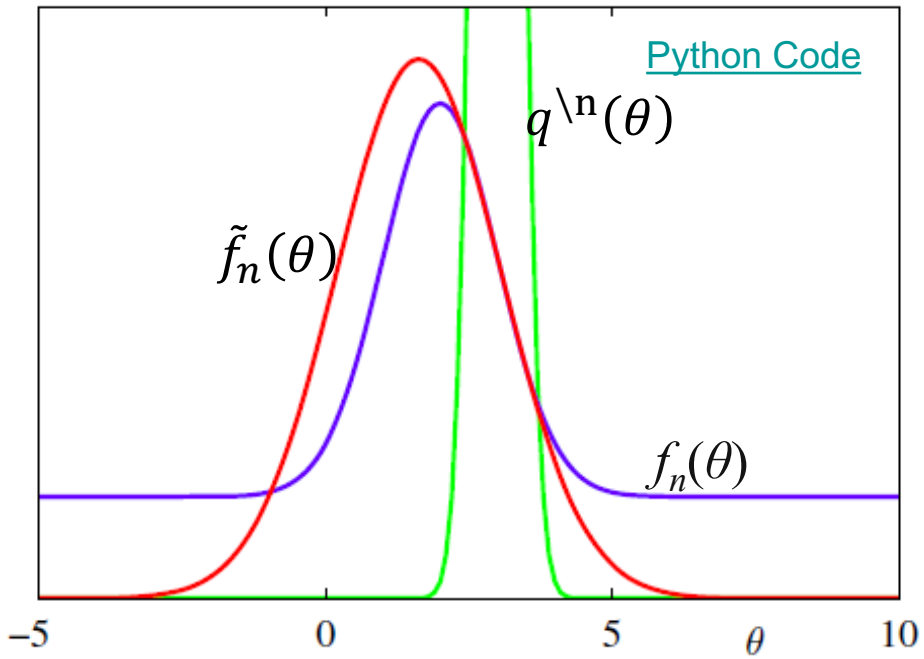
$$B = \frac{(\boldsymbol{m}^{new})^T \boldsymbol{m}^{new}}{v^{new}} - \sum_{n=1}^{N} \frac{(\boldsymbol{m}^{\mathrm{n}})^T \boldsymbol{m}^{\mathrm{n}}}{v_n}$$

Examples factor approximations for the clutter problem with 1D parameter space θ are shown next.

Note that the factor approximations can have infinite or even negative values for the 'variance' parameter v_n . This simply corresponds to approximations that curve upwards instead of downwards and are not necessarily problematic provided the overall approximate posterior $q(\boldsymbol{\theta})$ has positive variance.

The Clutter Problem

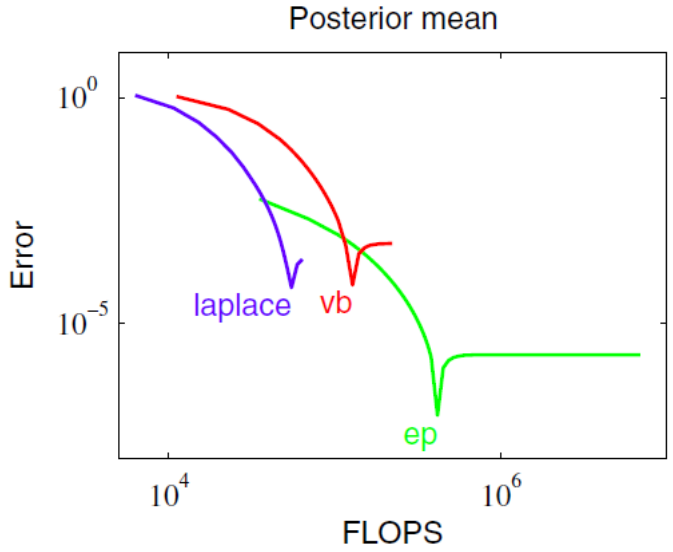


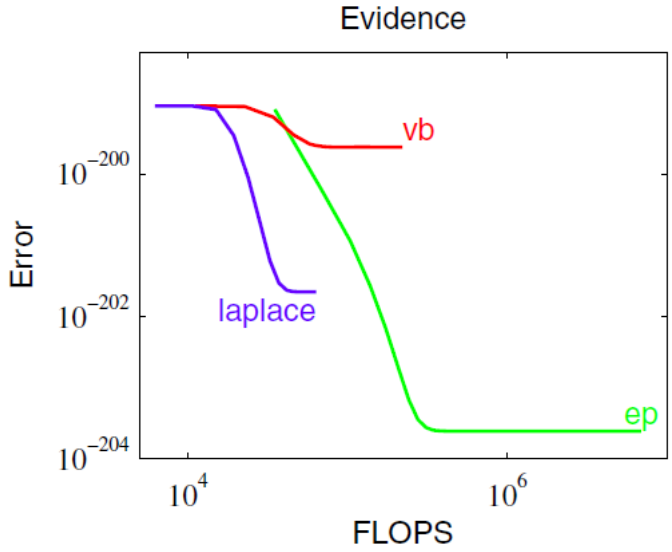


Examples of the approximation of specific factors for a one-dimensional version of the clutter problem, showing $f_n(\theta)$ in blue, $\tilde{f}_n(\theta)$ in red, and $q^{\backslash n}(\theta)$ in green. Notice that the current form for $q^{\backslash n}(\theta)$ controls the range of θ over which , $\tilde{f}_n(\theta)$ will be a good approximation to $f_n(\theta)$.

The factor approximations $\tilde{f}_n(\theta)$ can have infinite or even negative values for the 'variance' parameter v_n. This corresponds to $\tilde{f}_n(\theta)$ that curve upwards. This is not problematic provided the overall approximate posterior $q(\boldsymbol{\theta})$ has positive variance.

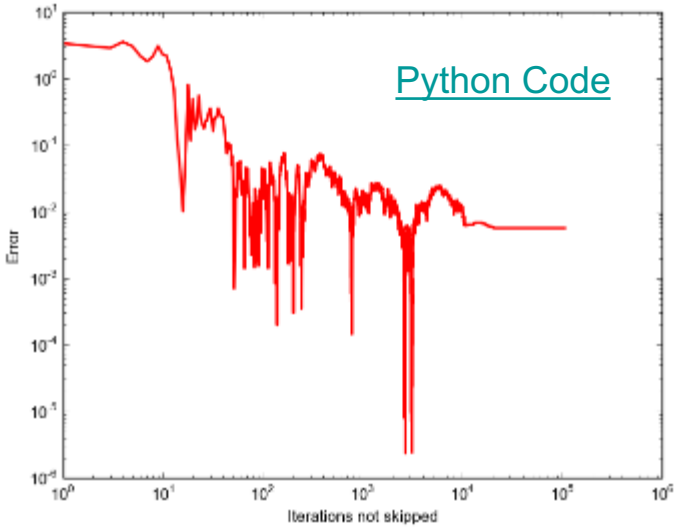
The Clutter Problem





Comparison of EP, variational inference (mean field approximation), and the Laplace approximation on the clutter problem.

The left-hand plot on the top shows the error in the predicted posterior mean versus the number of floating point operations, and the right-hand plot shows the corresponding results for the model evidence.



Expectation Propagation in Graphs

So far, we consider factors $f_i(\boldsymbol{\theta})$ in $p(\boldsymbol{\theta})$ to be functions of all of the components of $\boldsymbol{\theta}$, and similarly for the approximating factors $\tilde{f}_j(\boldsymbol{\theta})$ in the approximating distribution $q(\boldsymbol{\theta})$.

We now consider situations in which the factors depend only on subsets of the variables.

Such restrictions can be conveniently expressed using probabilistic graphical models.

Here we use a factor graph representation since it encompasses both directed and undirected graphs.

In the case the approximating distribution is fully factorized, expectation propagation reduces to loopy belief propagation.

Recall that if we minimize the Kullback-Leibler divergence KL($p||q$) with respect to a factorized distribution q, then the optimal solution for each factor is simply the corresponding marginal of p.

Factorized q(Z) and min of KL(p||q)

Consider the reverse KL divergence:

$$KL(p||q) = -\int p(\boldsymbol{Z}) \ln\left\{\frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})}\right\} d\boldsymbol{Z} = \int p(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z}$$

Note that in general $KL(p||q) \neq KL(q||p)$ and therefore using the reverse KL divergence would yield different results. Keeping only terms in $q_j(\boldsymbol{Z}_j)$, we can write:

$KL(p||q) = -\int p(\boldsymbol{Z}) \sum_i \ln q_i(\boldsymbol{Z}_i)\, d\boldsymbol{Z} + const = -\int p(\boldsymbol{Z}) \ln q_j(\boldsymbol{Z}_j)\, d\boldsymbol{Z} + const =$
$-\int \ln q_j(\boldsymbol{Z}_j) [\int p(\boldsymbol{Z}) \prod_{i\neq j} d\boldsymbol{Z}_i]\, d\boldsymbol{Z}_j + const = -\int \ln q_j(\boldsymbol{Z}_j) F_j(\boldsymbol{Z}_j)\, d\boldsymbol{Z}_j + const$, with $F_j(\boldsymbol{Z}_j) = \int p(\boldsymbol{Z}) \prod_{i\neq j} d\boldsymbol{Z}_i$.

Optimize wrt $q_j(\boldsymbol{Z}_j)$: $-\int \ln q_j(\boldsymbol{Z}_j) F_j(\boldsymbol{Z}_j)\, d\boldsymbol{Z}_j + \lambda\left(\int q_j(\boldsymbol{Z}_j)\, d\boldsymbol{Z}_j - 1\right)$.

This gives: $\lambda = \frac{F_j(\boldsymbol{Z}_j)}{q_j(\boldsymbol{Z}_j)}$. Integrating over $\boldsymbol{Z}_j$ gives $\lambda = 1$. Thus we conclude:

$$q_j^*(\boldsymbol{Z}_j) = F_j(\boldsymbol{Z}_j) = \int p(\boldsymbol{Z}) \prod_{i\neq j} d\boldsymbol{Z}_i = p(\boldsymbol{Z}_j) \text{ (marginal, no iteration needed)}$$

Expectation Propagation in Graphs

Now consider the factor graph shown.

The joint distribution is given by

$$p(\mathbf{x}) = f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4)$$

We seek an approximation $q(\mathbf{x})$ that has the same factorization, so that

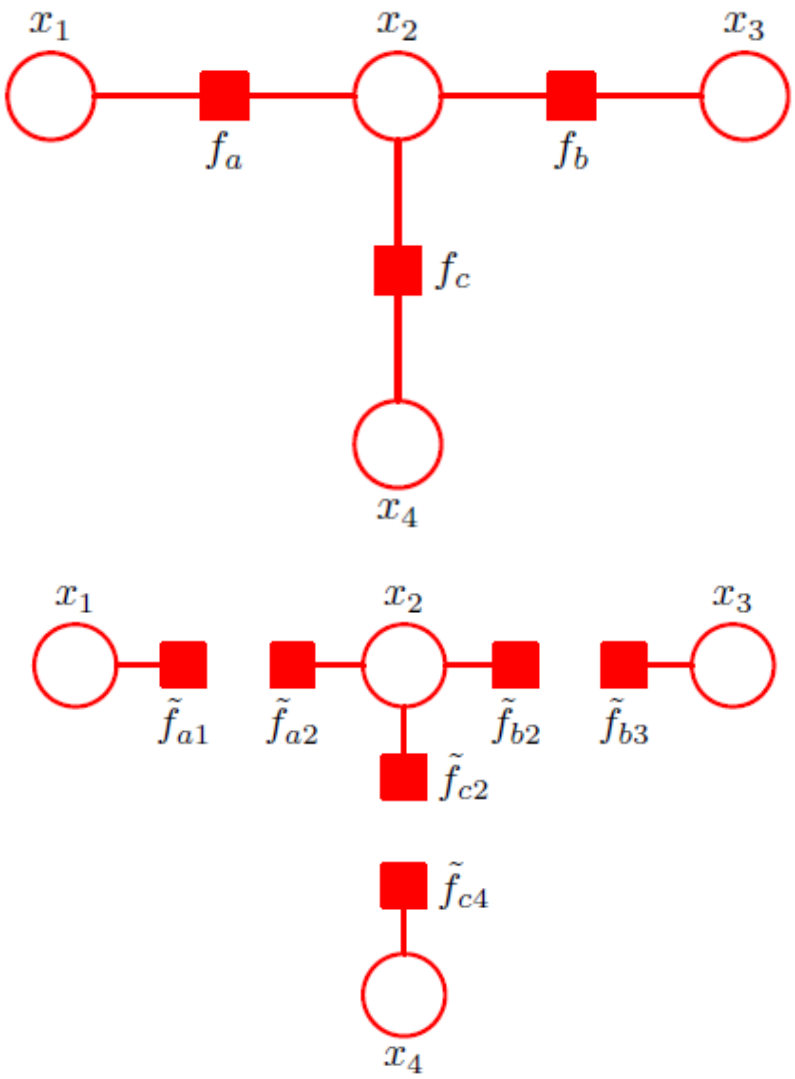
$$q(\mathbf{x}) \propto \tilde{f}_a(x_1, x_2)\, \tilde{f}_b(x_2, x_3)\, \tilde{f}_c(x_2, x_4)$$

Note that normalization constants have been omitted, and these can be re-instated at the end by local normalization, as is generally done in belief propagation.

Now suppose we restrict attention to approximations in which the factors themselves factorize with respect to the individual variables so that

$$q(\mathbf{x}) \propto \tilde{f}_{a1}(x_1)\, \tilde{f}_{a2}(x_2)\, \tilde{f}_{b2}(x_2)\, \tilde{f}_{b3}(x_3)\, \tilde{f}_{c2}(x_2)\, \tilde{f}_{c4}(x_4)$$

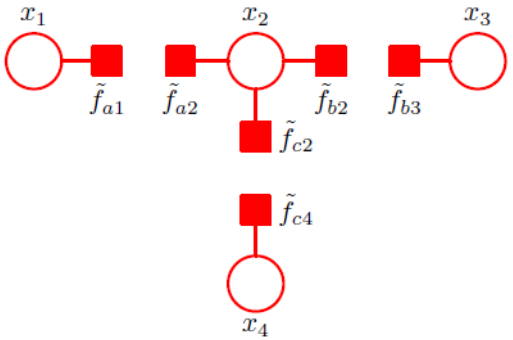
which corresponds to the factor graph shown on the right.



Expectation Propagation in Graphs

Now we apply the EP algorithm using the fully factorized approximation. Suppose that we have initialized all of the factors and that we choose to refine factor $\tilde{f}_b(x_2, x_3) = \tilde{f}_{b2}(x_2)\,\tilde{f}_{b3}(x_3)$. We first remove this factor from the approximating distribution to give

$$q^{\backslash b}(\mathbf{x}) \propto \tilde{f}_{a1}(x_1)\,\tilde{f}_{a2}(x_2)\,\tilde{f}_{c2}(x_2)\,\tilde{f}_{c4}(x_4)$$



We then multiply this by the exact factor $f_b(x_2, x_3)$ to give

$$\hat{p}(\boldsymbol{x}) = q^{\backslash b}(\mathbf{x}) f_b(x_2, x_3) = \tilde{f}_{a1}(x_1)\,\tilde{f}_{a2}(x_2)\,\tilde{f}_{c2}(x_2)\,\tilde{f}_{c4}(x_4) f_b(x_2, x_3)$$

We now find $q^{\text{new}}(\mathbf{x})$ by minimizing the Kullback-Leibler divergence $\text{KL}(\hat{p}||q^{\text{new}})$. The result is that $q^{\text{new}}(\mathbf{z})$ comprises the product of factors, one for each variable x_i, in which each factor is given by the corresponding marginal of $\hat{p}(\mathbf{x})$. These four marginals are given by

$$\hat{p}(x_1) \propto \tilde{f}_{a1}(x_1)$$

$$\hat{p}(x_2) \propto \tilde{f}_{a2}(x_2)\,\tilde{f}_{c2}(x_2) \sum_{x_3} f_b(x_2, x_3)$$

$$\hat{p}(x_3) \propto \sum_{x_2} f_b(x_2, x_3)\,\tilde{f}_{a2}(x_2)\,\tilde{f}_{c2}(x_2)$$

$$\hat{p}(x_4) \propto \tilde{f}_{c4}(x_4)$$

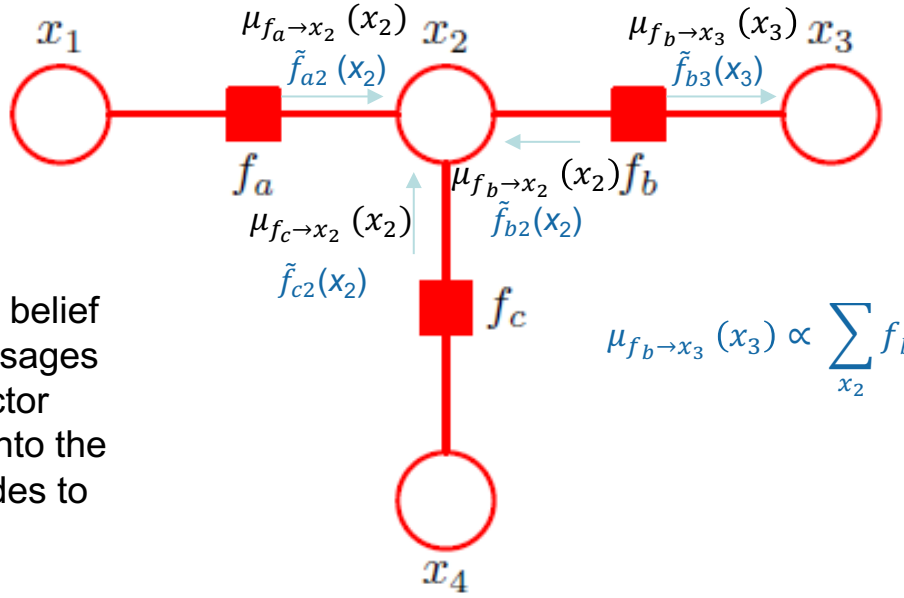
$$q^{\text{new}}(\mathbf{x}) = \hat{p}(x_1)\,\hat{p}(x_2)\,\hat{p}(x_3)\,\hat{p}(x_4)$$

Expectation Propagation in Graphs

The only factors in q(x) that change when we update $\tilde{f}_b(x_2, x_3)$ are those that involve the variables in f_b namely x_2 and x_3. To obtain the refined factor $\tilde{f}_b(x_2, x_3) = \tilde{f}_{b2}(x_2)\,\tilde{f}_{b3}(x_3)$ we simply divide $q^{new}(\mathbf{x})$ by $q^{\backslash b}(\boldsymbol{x})$ which gives:

$$\tilde{f}_{b2}(x_2) \propto \sum_{x_3} f_b(x_2, x_3)$$

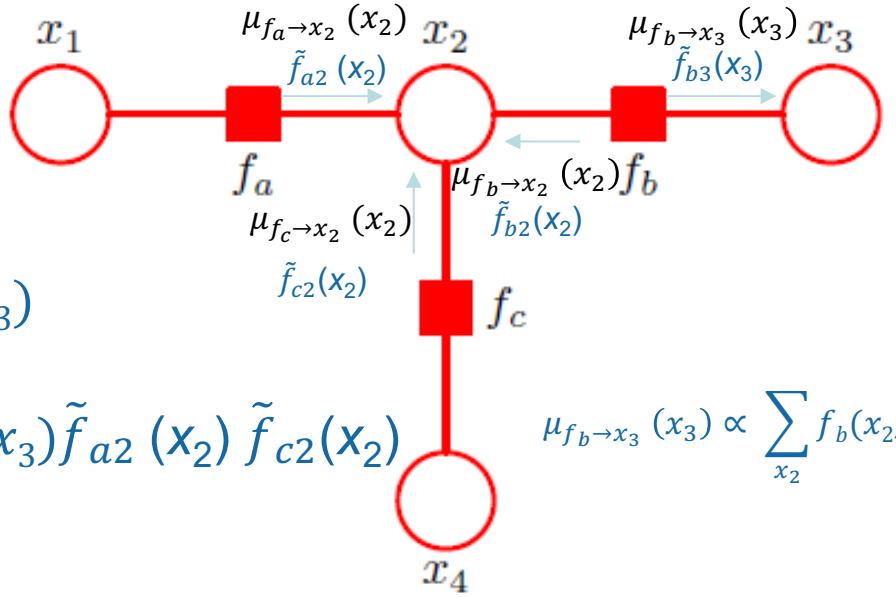
$$\tilde{f}_{b3}(x_3) \propto \sum_{x_2} f_b(x_2, x_3)\tilde{f}_{a2}(x_2)\,\tilde{f}_{c2}(x_2)$$



$$\mu_{f_b \to x_3}(x_3) \propto \sum_{x_2} f_b(x_2, x_3)\mu_{f_a \to x_2}(x_2)\mu_{f_c \to x_2}(x_2)$$

These are precisely the messages obtained using belief propagation in which messages from variable nodes to factor nodes have been folded into the messages from factor nodes to variable nodes.

Expectation Propagation in Graphs



$$\tilde{f}_{b2}(x_2) \propto \sum_{x_3} f_b(x_2, x_3)$$

$$\tilde{f}_{b3}(x_3) \propto \sum_{x_2} f_b(x_2, x_3) \tilde{f}_{a2}(x_2)\, \tilde{f}_{c2}(x_2)$$

$$\mu_{f_b \to x_3}(x_3) \propto \sum_{x_2} f_b(x_2, x_3) \mu_{f_a \to x_2}(x_2) \mu_{f_c \to x_2}(x_2)$$

$\tilde{f}_{b2}(x_2)$ corresponds to

$$\mu_{f_b \to x_2}(x_2) = \sum_{x_3} f_b(x_2, x_3) \qquad (\tilde{f}_{b2}(x_2) \propto \sum_{x_3} f_b(x_2, x_3))$$

Similarly, if we substitute $\mu_{x_2 \to f_b}(x_2) = \mu_{f_a \to x_2}(x_2) \mu_{f_c \to x_2}(x_2)$ in $\mu_{f_b \to x_3}(x_3) = \sum_{x_2} f_b(x_2, x_3)\, \mu_{x_2 \to f_b}(x_2)$ we obtain $\tilde{f}_{b3}(x_3) \propto \sum_{x_2} f_b(x_2, x_3) \tilde{f}_{a2}(x_2)\, \tilde{f}_{c2}(x_2)$ in which $\tilde{f}_{a2}(x_2)$ corresponds to $\mu_{f_a \to x_2}(x_2)$ and $\tilde{f}_{c2}(x_2)$ corresponds to $\mu_{f_c \to x_2}(x_2)$ giving the message $\tilde{f}_{b3}(x_3)$ which corresponds to $\mu_{f_b \to x_3}(x_3)$.

Expectation Propagation in Graphs

This result differs slightly from standard belief propagation in that messages in EP are passed in both directions at the same time.

We can modify EP to give the standard form of the sum-product algorithm by updating just one of the factors at a time. E.g. if we refine only $\tilde{f}_{b3}(x_3)$ then $\tilde{f}_{b2}(x_2)$ is unchanged by definition, while the refined version of $\tilde{f}_{b3}(x_3)$ is again given by

$$\tilde{f}_{b3}(x_3) \propto \sum_{x_2} f_b(x_2, x_3)\tilde{f}_{a2}(x_2)\,\tilde{f}_{c2}(x_2)$$

If we are refining only one term at a time, then we can choose the order in which the refinements are done as we wish.

In particular, for a tree-structured graph we can follow a two-pass update scheme, corresponding to the standard belief propagation schedule, which will result in exact inference of the variable and factor marginals.

Expectation Propagation in Graphs

Now let us consider a general factor graph corresponding to the distribution

$$p(\boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ represents the subset of variables associated with factor f_i. We approximate this using a fully factorized distribution of the form

$$q(\boldsymbol{\theta}) \propto \prod_i \prod_k \tilde{f}_{ik}(\theta_k)$$

where θ_k corresponds to an individual variable node. Suppose that we wish to refine the particular term $\tilde{f}_{jl}(\theta_\ell)$ keeping all other terms fixed. We first remove the term $\tilde{f}_j(\boldsymbol{\theta}_j)$ from $q(\boldsymbol{\theta})$ to give

$$q^{\setminus j}(\boldsymbol{\theta}) \propto \prod_{i \neq j} \prod_k \tilde{f}_{ik}(\theta_k)$$

and then multiply by the exact factor $f_j(\boldsymbol{\theta}_j)$. To determine the refined tern $\tilde{f}_{jl}(\theta_\ell)$ we need only consider the functional dependence on θ_ℓ and find the corresponding marginal of $q^{\setminus j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta}_j)$

Expectation Propagation in Graphs

Up to a multiplicative constant, this involves taking the marginal of $f_j(\boldsymbol{\theta}_j)$ multiplied by any terms from $q^{\setminus j}(\boldsymbol{\theta})$ that are functions of any of the variables in $\boldsymbol{\theta}_j$. Terms that correspond to other factors $\tilde{f}_i(\boldsymbol{\theta}_i)$ for $i \neq j$ cancel between numerator and denominator when we subsequently divide by $q^{\setminus j}(\boldsymbol{\theta})$. We obtain

$$\tilde{f}_{j\ell}(\theta_\ell) \propto \sum_{\theta_{m\neq\ell} \in \boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \prod_k \prod_{m\neq\ell} \tilde{f}_{km}(\theta_m)$$

We recognize this as the sum-product rule in the form in which messages from variable nodes to factor nodes have been eliminated.

The quantity $\tilde{f}_{km}(\theta_m)$ corresponds to the message $\mu_{f_k \to \theta_m}(\theta_m)$, which factor node k sends to variable node m, and the product over j in the equation above is over all factors that depend on the variables θ_m that have variables (other than variable θ_ℓ) in common with factor $f_j(\boldsymbol{\theta}_j)$.

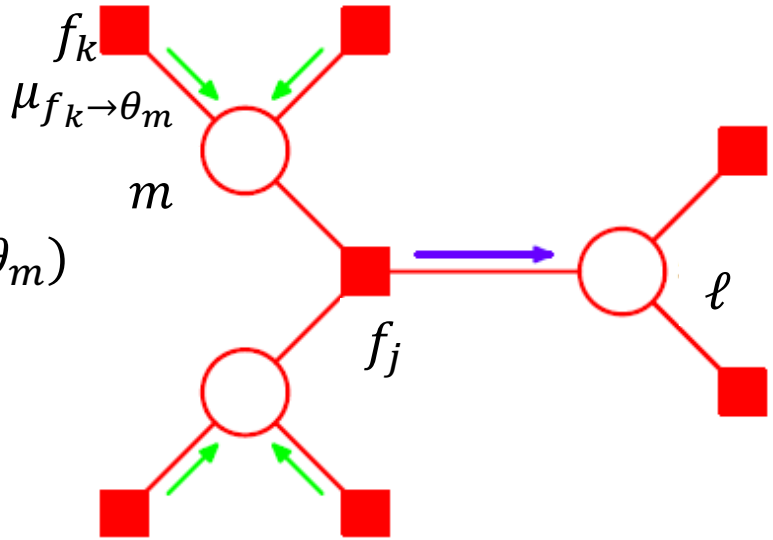
Expectation Propagation in Graphs

The quantity $\tilde{f}_{km}(\theta_m)$ corresponds to the message $\mu_{f_k \to \theta_m}(\theta_m)$, which factor node *k* sends to variable node *m*, and the product over is over all factors that depend on the variables θ_m that have variables (other than variable θ_ℓ) in common with factor $f_j(\boldsymbol{\theta}_j)$.

$$\tilde{f}_{j\ell}(\theta_\ell) \propto \sum_{\theta_{m \neq \ell} \in \boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \prod_k \prod_{m \neq \ell} \mu_{f_k \to \theta_m}(\theta_m)$$

E.g. to compute the outgoing message from a factor node, we take the product of all the incoming messages from other factor nodes, multiply by the local factor, and then marginalize.

$$\tilde{f}_{j\ell}(\theta_\ell) \propto \sum_{\theta_{m \neq \ell} \in \boldsymbol{\theta}_j} f_j(\boldsymbol{\theta}_j) \prod_k \prod_{m \neq \ell} \tilde{f}_{km}(\theta_m)$$



Expectation Propagation in Graphs

Thus, the sum-product algorithm arises as a special case of expectation propagation if we use an approximating distribution that is fully factorized.

This suggests that more flexible approximating distributions, corresponding to partially disconnected graphs, could be used to achieve higher accuracy.

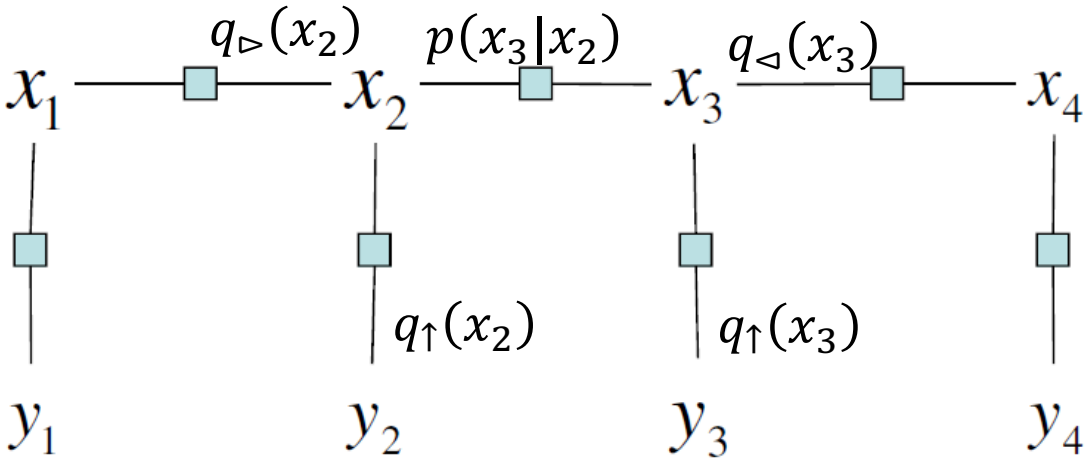
Another generalization is to group factors $f_i(\boldsymbol{\theta}_i)$ together into sets and to refine all the factors in a set together at each iteration. Both of these approaches can lead to improvements in accuracy

The problem of choosing the best combination of grouping and disconnection is an open research issue.

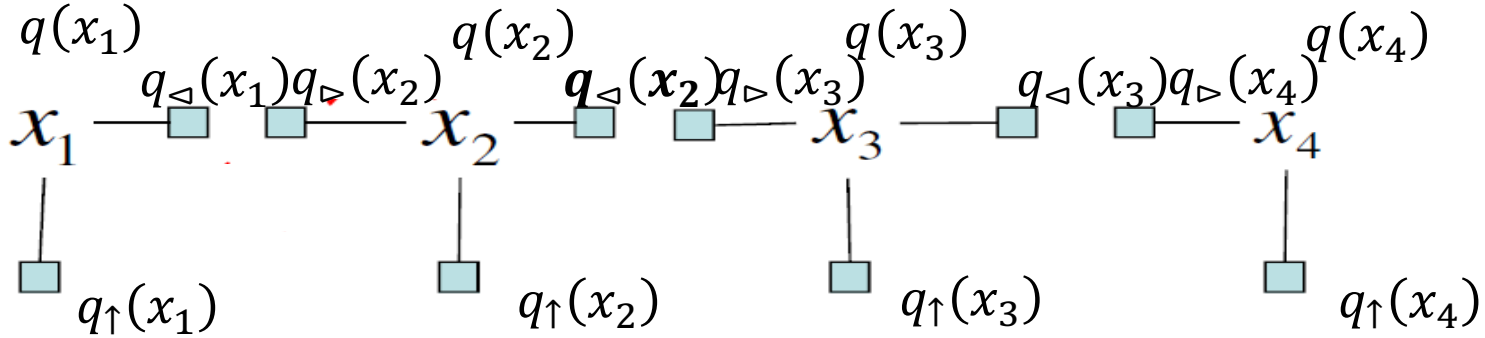
- Minka, T. (2001). *A family of approximate algorithms for Bayesian inference*. Ph. D. thesis, MIT.

Tracking Problem

Consider $x_t = x_{t-1} + v_t$ (random walk) and noisy observations $y_y = x_t + noise$. We are interested in the distribution of the x's given the y's.

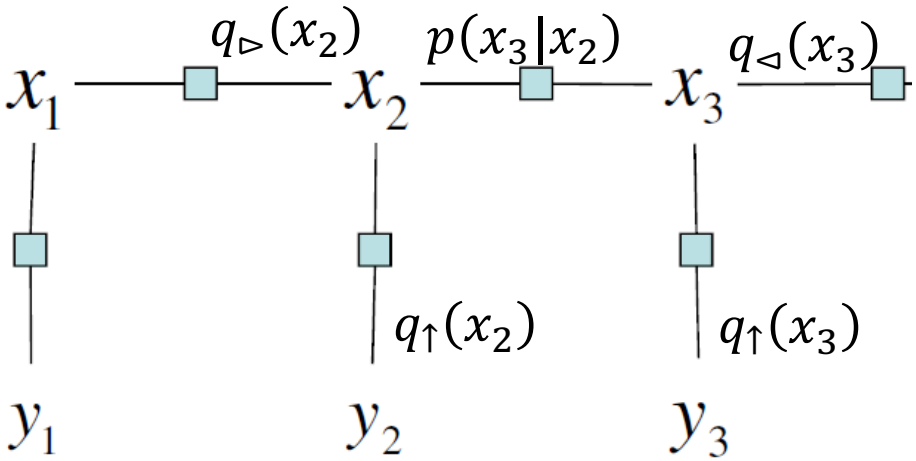


We use the following approximate factor graph: $p(x_2|x_1) \approx q_{\triangleleft}(x_1)q_{\triangleright}(x_2)$

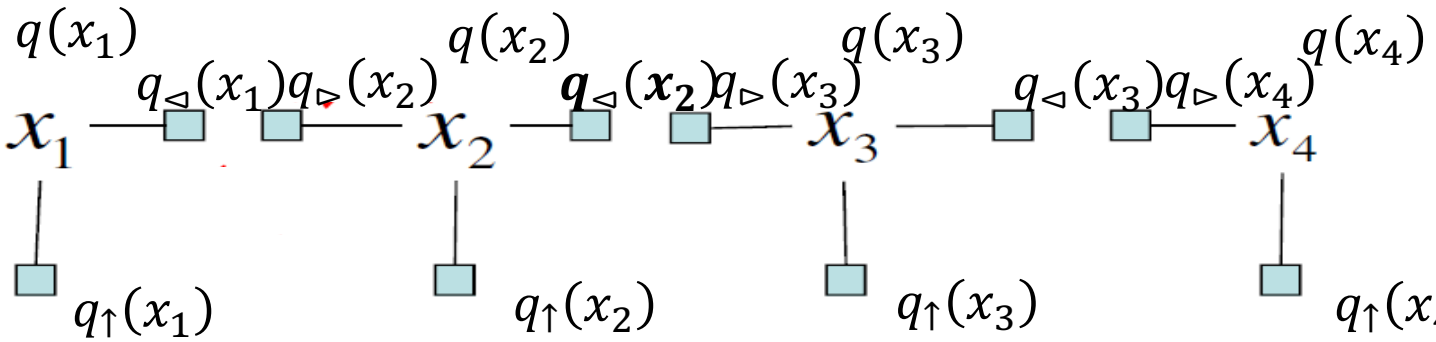


Use backward & forward Gaussian messages for each state in approximating $p(x_{t+1}|x_t)$.

Tracking Problem



$$q_{\triangleright}(x_2)q_{\uparrow}(x_2)\int p(x_3|x_2)q_{\triangleleft}(x_3)q_{\uparrow}(x_3)dx_3$$
$$\approx q_{\triangleright}(x_2)q_{\uparrow}(x_2)\boldsymbol{q}_{\triangleleft}(\boldsymbol{x_2})$$
$$\downarrow$$
$$\boldsymbol{q}_{\triangleleft}(\boldsymbol{x_2}) = \frac{proj\left[q_{\triangleright}(x_2)q_{\uparrow}(x_2)\int p(x_3|x_2)q_{\triangleleft}(x_3)q_{\uparrow}(x_3)dx_3\right]}{q_{\triangleright}(x_2)q_{\uparrow}(x_2)}$$
$$\downarrow$$
$$q_{\triangleleft}(x_2) \approx \int p(x_3|x_2)q_{\triangleleft}(x_3)q_{\uparrow}(x_3)dx_3 \quad \text{and similarly}$$
$$q_{\triangleright}(x_2) \approx \int p(x_2|x_1)q_{\triangleleft}(x_1)q_{\uparrow}(x_1)dx_1$$



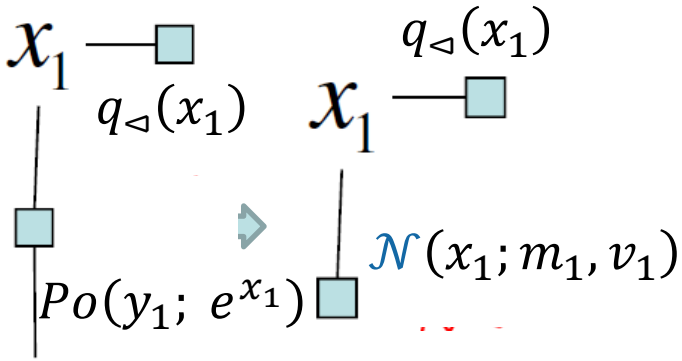
The messages for linear Gaussian model are the same as for the Kalman Filter. In sweeping once through the graph and computing for each factor the forward and backward messages, we can capture the exact posterior (no need to iterate)

Poisson Tracking

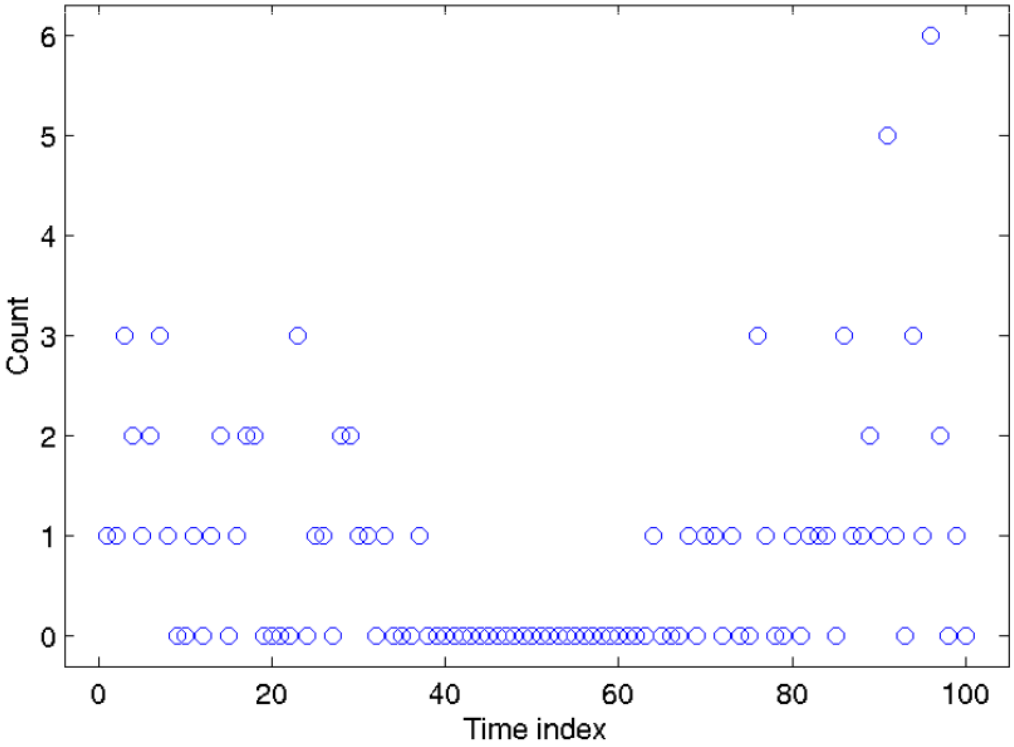
We consider a non-linear likelihood – y_t is a Poisson distributed integer with mean $\lambda = \exp(x_t)$

$$p(x_1) \sim \mathcal{N}(0, 100),$$
$$p(x_t|x_{t-1}) \sim \mathcal{N}(x_{t-1}, 0.01)$$
$$p(y_t|x_t) = \exp(y_t x_t - e^{x_t}) / y_t!$$

In addition to approximating the conditionals $p(x_t|x_{t-1})$, we need to approximate the likelihood:

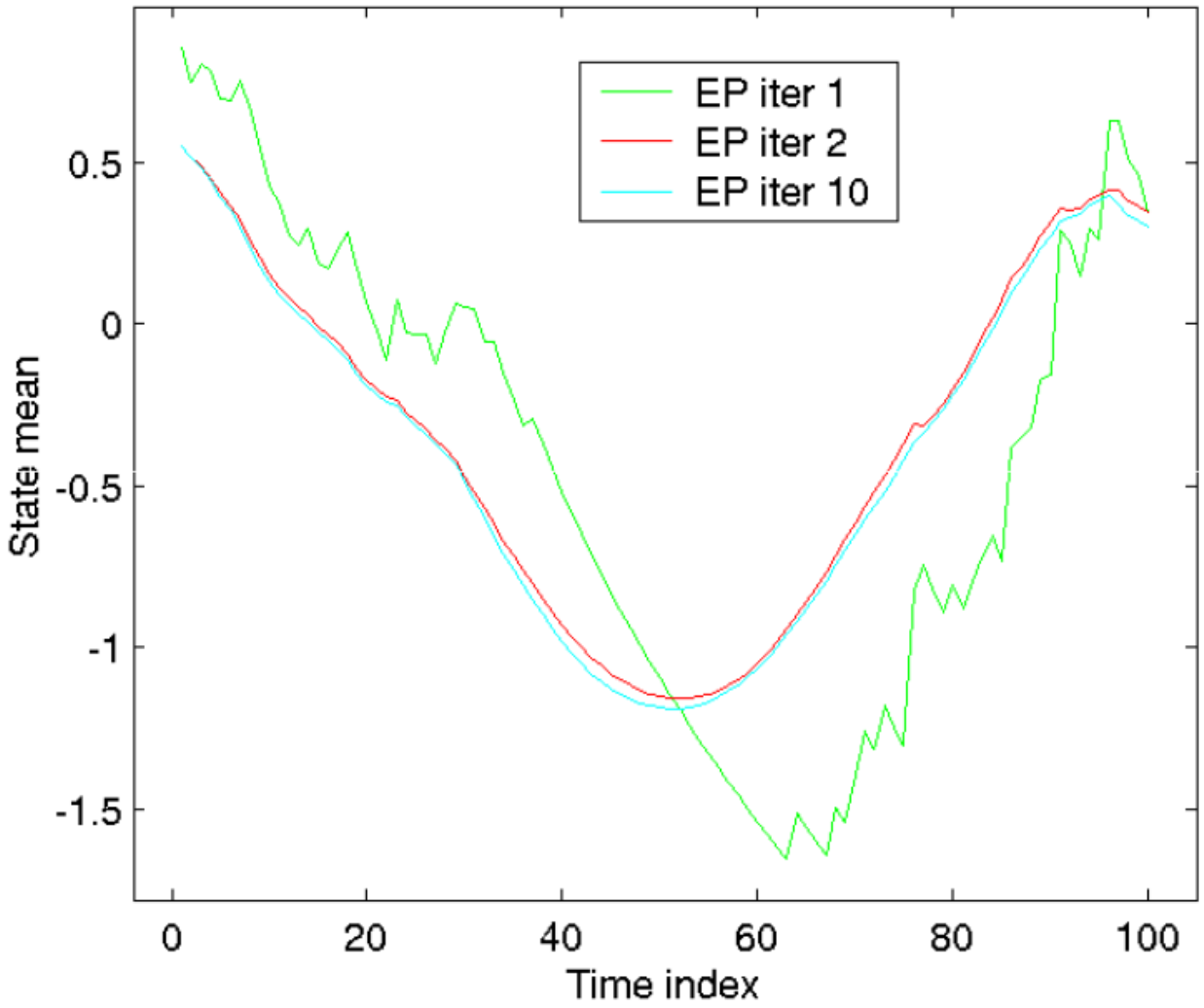


$$\mathcal{N}(x_1; m_1, v_1) = \frac{proj[Po(y_1;\ e^{x_1})\, q_\triangleleft(x_1)]}{q_\triangleleft(x_1)}$$



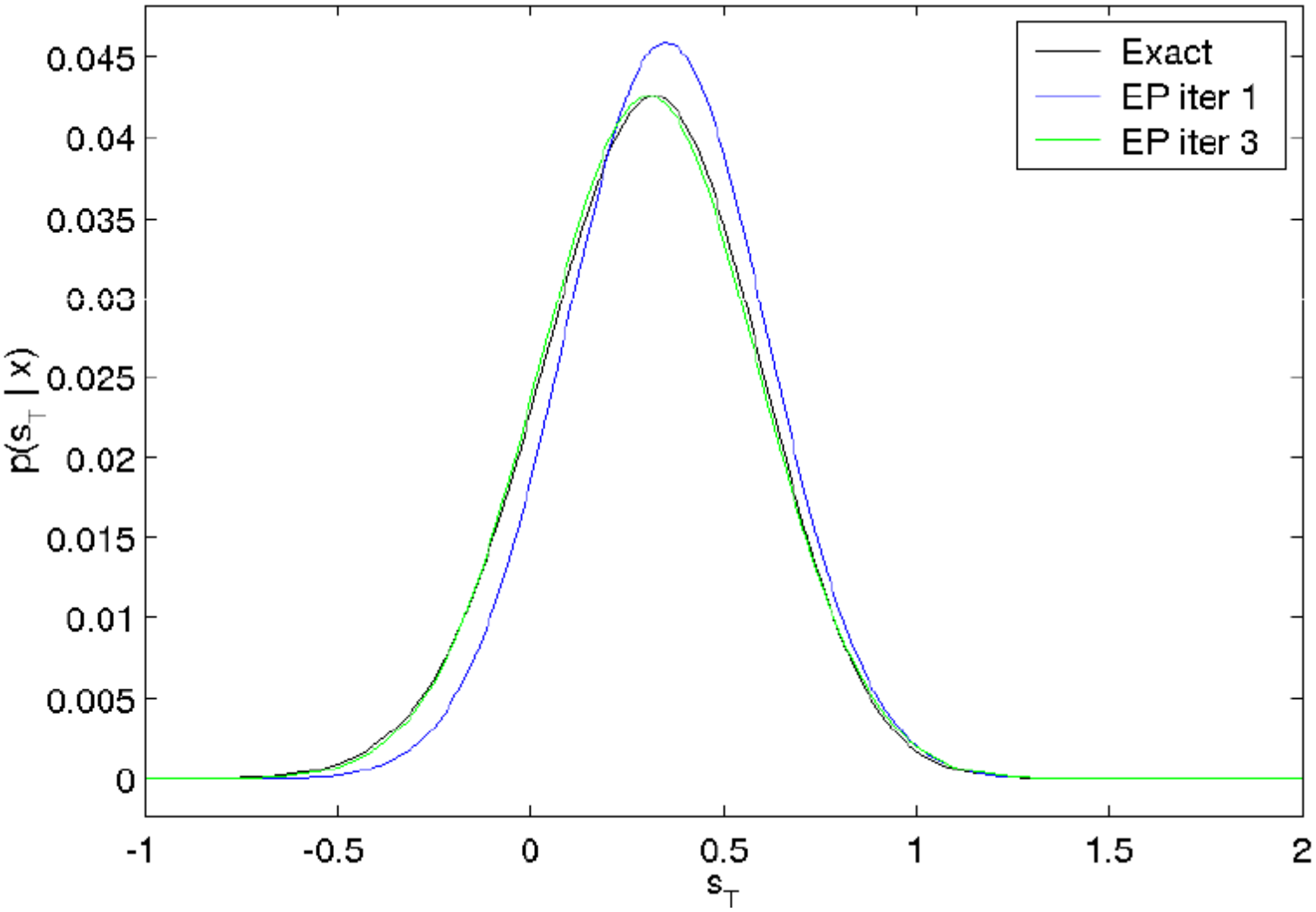
- Approximate Inference, Tom Minka, Msft Research

Poisson Tracking



- Approximate Inference, Tom Minka, Msft Research

Poisson Tracking



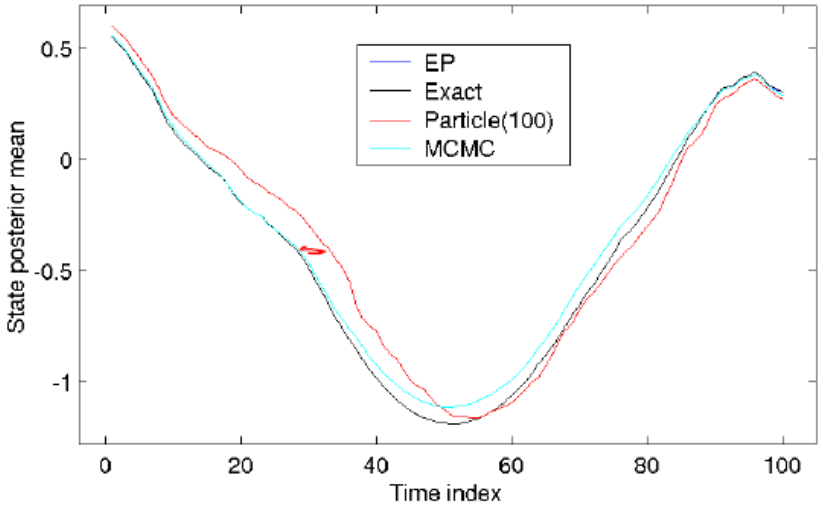
- Approximate Inference, Tom Minka, Msft Research

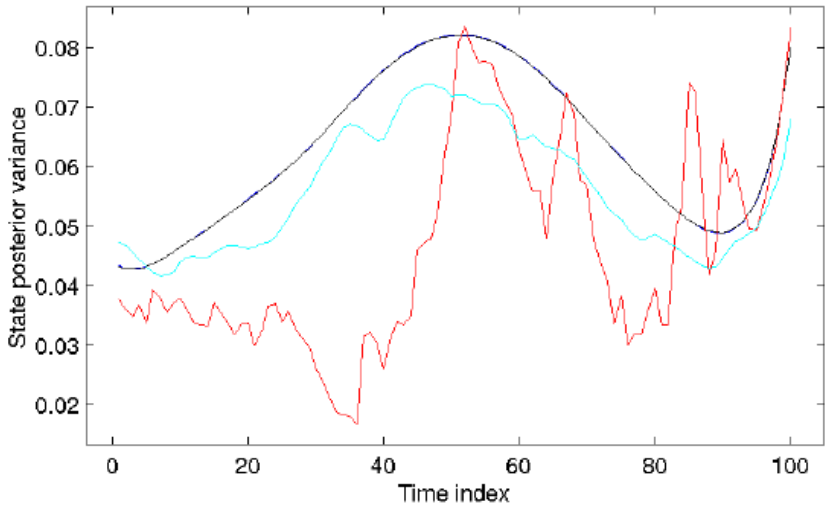
Poisson Tracking

Note that EP gives almost the exact solution

Particle filtering is good at filtering (predicting final state) but not at smoothing (going backwards)

The CPU time for the sampling methods is order of magnitude higher than that of EP.



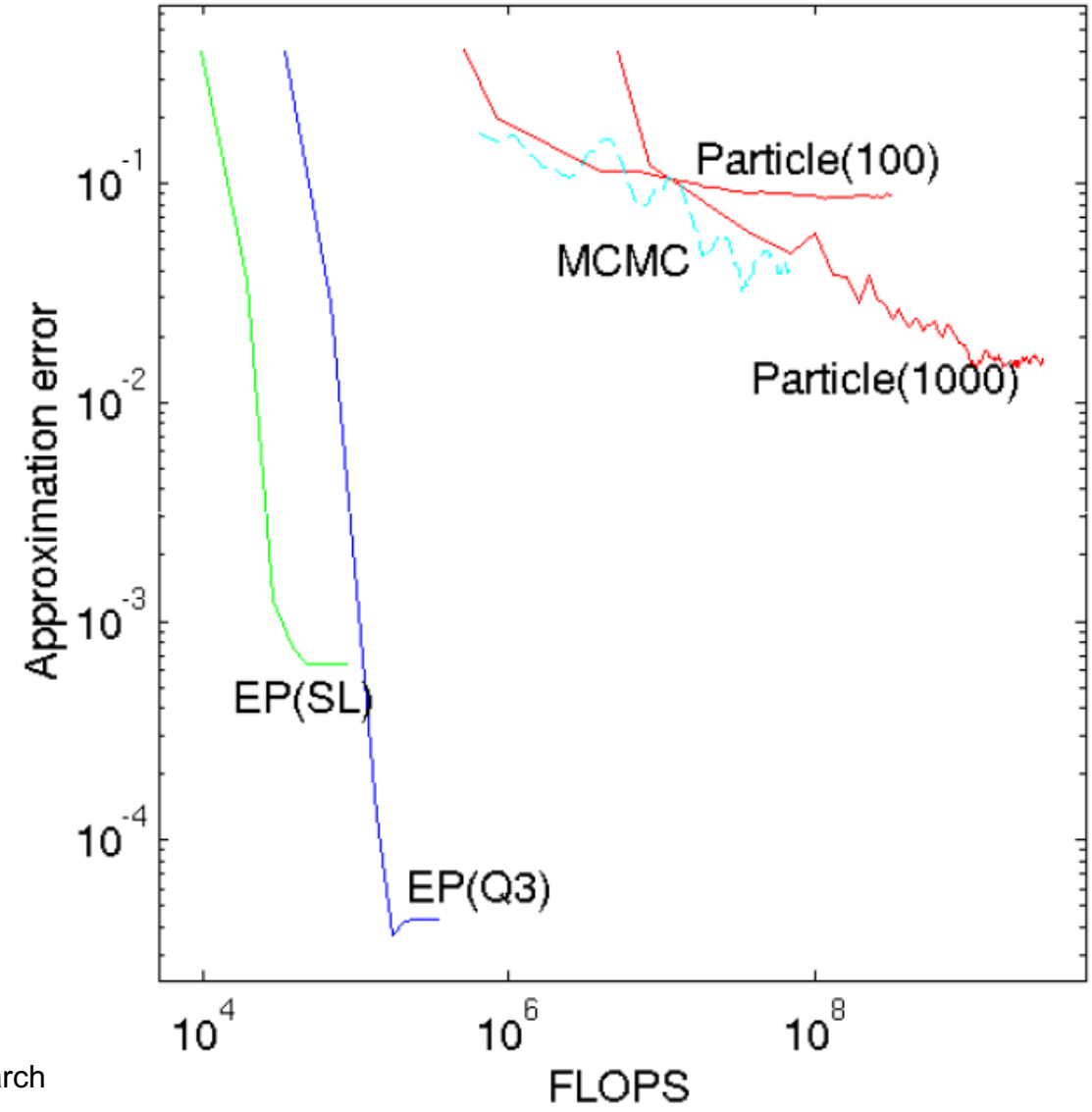


- Approximate Inference, Tom Minka, Msft Research

Poisson Tracking

The two EP approximations shown differ in the approximation of the Poisson convolution with Gaussians.

EP(SL) is the fast but less accurate of the two approximations.



- Approximate Inference, Tom Minka, Msft Research

Expectation Propagation in Graphs

Variational message passing and expectation propagation optimize two different forms of the Kullback-Leibler divergence.

Minka (2005) has shown that a broad range of message passing algorithms can be derived from a common framework involving minimization of members of the alpha family of divergences.

These include variational message passing, loopy belief propagation, expectation propagation, *tree-reweighted message passing*, *fractional belief propagation*, and *power EP*.

- Wainwright, M. J., T. S. Jaakkola, and A. S. Willsky (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory* **51**, 2313–2335.
- Wiegerinck, W. and T. Heskes (2003). Fractional belief propagation. In S. Becker, S. Thrun, and K. Obermayer (Eds.), *Advances in Neural Information Processing Systems*, Volume 15, pp. 455– 462. MIT Press.
- Minka, T. (2004). Power EP. Technical Report MSR-TR-2004-149, Microsoft Research Cambridge.
- Minka, T. (2005). Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Cambridge.
- M. Seeger, EP for Approximate Bayesian Inference, On line lecture (2010).
- A roadmap to research EP, Msft Research (video Presentation)