UNIVERSITY OF CAMBRIDGE
Department of Engineering

Understanding Expectation Propagation

Author Name: Siddharth Swaroop
Supervisor: Dr. Richard Turner
Date: 30/05/2017

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

Signed _____ *date* _____

Understanding Expectation Propagation

Siddharth Swaroop, Churchill College

Technical Abstract

Understanding and characterising approximate inference schemes has recently become extremely important in Machine Learning, as datasets are becoming increasingly larger and more complex. This report continues work on characterising Expectation Propagation (EP), an approximate Bayesian inference scheme that approximates a complex distribution with a simpler distribution (by breaking the complex distribution into factors, each of which are approximated in turn). EP is widely used in the machine learning community without its properties extensively mathematically characterised or theoretically understood, due to the complex nature of the EP approximation. By looking at three toy cases of interest, this report focusses on a conjecture in the Machine Learning community: that EP's approximation for the model evidence is an underestimate of the true model evidence.

This report applies EP to the model of a Gaussian prior multiplied with Heaviside step functions, using a Gaussian approximating family. The first toy case considered is the symmetric box case, which is a counter-example to the aforementioned conjecture, as EP overestimates the model evidence. This overestimation is shown empirically and proven mathematically. Softening the Heaviside functions to probit functions and expanding to multiple dimensions still results in an overestimation of the model evidence.

EP is then applied to repeated Heavisides, the second toy case. This toy case provides a link to Power EP (minimising the more general α-Divergence instead of the $\mathcal{KL}$ divergence) and to classification, and a large underestimation of the true model evidence is observed. Adding jitter (noise) to this case reduces the underestimation slightly. Combining these two toy cases (symmetric box case and repeated Heavisides) in a simple binary classification example shows that the underestimation tends to dominate with larger and more realistic datasets, helping explain why the conjecture in the community is observed when EP is applied on real-world examples.

The third toy case considered is on a 'Fully Independent Training Condition' (FITC) example (a sparse approximation for Gaussian Process regression). Using the relationship between the FITC approximation and the EP algorithm, it is shown on a simple (1-dimensional, 2 input, 2 inducing input) example that FITC, and hence EP, can overestimate the model likelihood. This is another counter-example to the conjecture on EP's model evidence. This example also sheds light on properties of the FITC approximation, as it is found that FITC prefers to use only one inducing input, instead of the two inducing inputs allowed (and the two used in the true regression case). This result is analysed and explained mathematically; future work could involve expanding this simple toy case to bigger datasets, to better explain FITC's characteristics.

Contents

Acknowledgements

I would like to thank my supervisor, Dr Richard E Turner, for his help throughout the project, from providing ideas to helping me when I faced problems. I would also like to thank Yingzhen Li and Thang Bui for their contributions during discussions on my work.

1 Introduction

When applying a Machine Learning technique to draw conclusions and/or predictions from a dataset, a model, loss function and inference scheme all need to be chosen. Reasons for choosing a model and loss function are well-known and are typically a direct consequence of the specific problem; however, the choice of inference scheme is more subtle and often requires expert guidance. Choosing a good (approximate) inference scheme is particularly important when faced with large datasets or intractable integrals, both recent trends in Machine Learning applications. Figure 1 summarises the most used sets of approximate inference schemes.

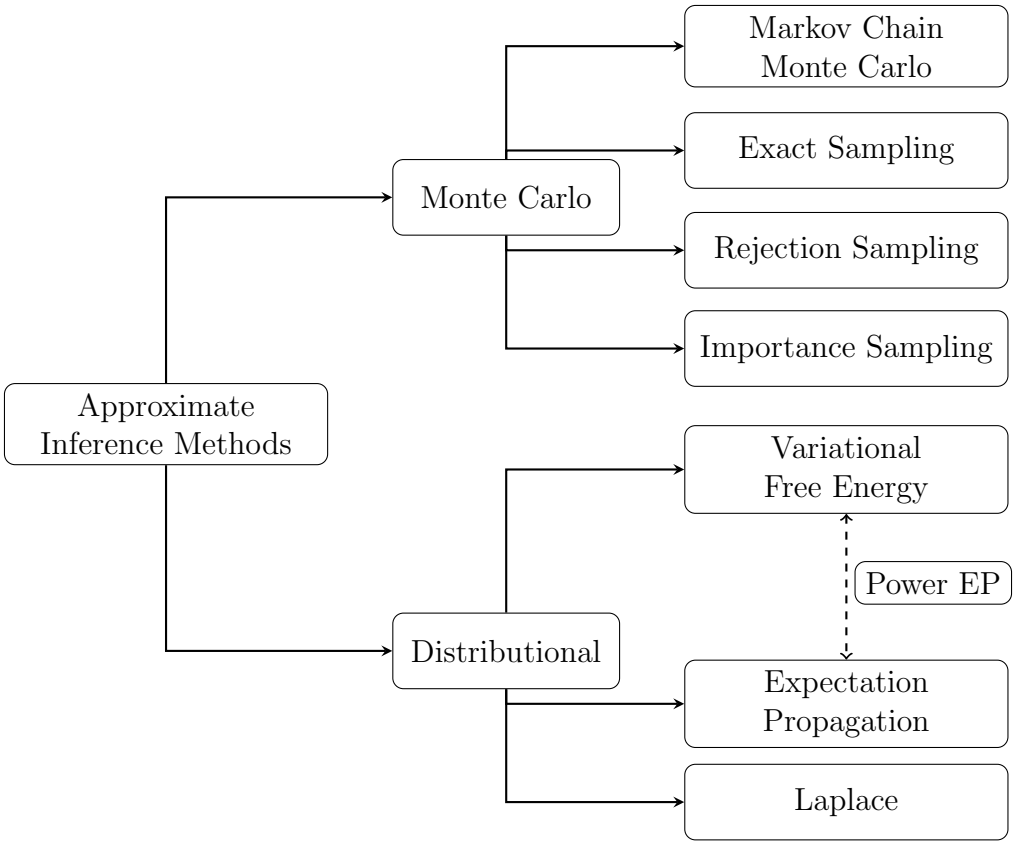


Figure 1: Approximate Inference Schemes

Approximate inference schemes can be categorised into two groups: Monte Carlo schemes and distributional schemes. Monte Carlo algorithms, of which examples are given in Figure 1, rely on random sampling from the true distribution. Distributional methods work more directly with properties of this distribution. For example, Variational Free Energy (VFE) methods and EP (or the more general Power Expectation Propagation (PEP)) attempt to match an

approximate distribution to the true distribution, with constraints placed on the approximate distribution to make manipulating it easier than manipulating the true distribution.

Each scheme is found to work well in certain cases, while performing poorly in others. It is hard to make generalising statements as to when certain schemes work better than others (it tends to be very problem-specific), but there are a few points we can make. Monte Carlo algorithms are usually computationally expensive, particularly in higher dimensions, but have the potential to provide very accurate approximations. The Laplace method can be expensive ($\mathcal{O}(N^3)$), and has empirically been found to be poor in certain scenarios. VFE methods also tend to fail in certain situations, such as when there are non-smooth or close to non-smooth potentials, or when applied on heavily factorised approximations on chains (such as mean-field). EP has empirically been found to work well in some of these situations, and is also relatively computationally cheap.

For a Machine Learning expert to choose the best scheme for an application, the schemes need to be understood better, and more of their properties characterised. Recently, the field in general has moved towards empirically evaluating sets of approximate inference methods. For example, sparse approximation methods for Gaussian Process (GP) regression models have been compared, such as the Fully Independent Training Condition (FITC) and VFE models [1]. New approximate inference schemes have also been developed, including a new framework of pseudo-point approximations for Gaussian Process models using PEP [2], relating FITC to EP. Other new approximate inference schemes recently introduced include Blackbox-α [6], which is more general than Variational Bayes (VB) or EP, and the variational Rényi bound [8]. Each scheme has been empirically shown to have uses in particular situations, and it is useful to continue to similarly characterise EP, to allow for comparison.

Although there is literature characterising many schemes, there is little theoretical understanding or mathematical explanation of EP's properties, despite many conjectures in the research community regarding its use. Researchers have attempted to characterise the behaviour of the EP loop, looking at its fixed points and comparing them to VB [5]. Generalising EP to PEP related the α-Divergence to EP and VB, allowing conclusions to be drawn about EP [11], but characterising more properties remains difficult. Attempts to empirically characterise the EP solution include studying the EP approximation for the model evidence, when applied to binary GP classification [7] and when applied to hyperrectangular and general polyhedral regions [4]. These studies led to conjectures regarding EP, such as saying EP always underestimates the true model evidence. However, it has been difficult to prove this conjecture [12], even though mathematical bounds exist in specific cases for belief

propagation [14] [15], with counter-examples to the conjecture introduced [3].

This project attempts to obtain a theoretical understanding of some of EP's properties by applying EP to toy cases of interest, before expanding the cases to more closely resemble real applications. See Figure 2 for a summary of the cases considered. We first introduce the EP algorithm in Section 2, and describe how it was applied with Heaviside step functions and probit functions in Section 3. We consider the symmetric box case counter-example to the conjecture that EP underestimates the model evidence, providing a possible intuitive explanation, and mathematically showing it is a counter-example, in Section 4. We also expand the symmetric box case by softening the Heaviside functions to probit functions, and by considering the multi-dimensional case. We then consider why the conjecture is likely to hold with real datasets, considering repeated functions and the α-Divergence in Section 5. Applying these two cases to a classification example in Section 6 brings together this work in a more realistic example. In Section 7, we consider another counter-example to the conjecture on EP's model evidence approximation, using the relationship between FITC and EP fixed points, and we also explore how this example affects the FITC algorithm. Section 8 provides conclusions on results obtained and possible future work.

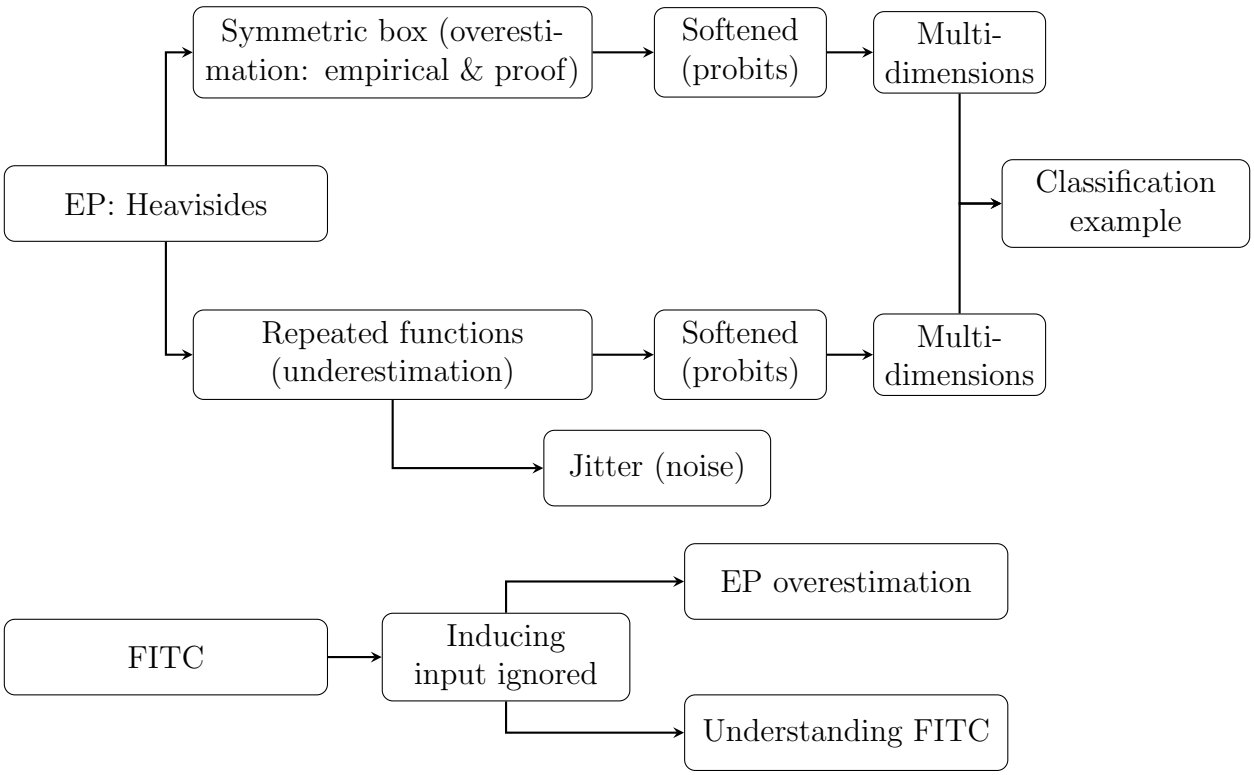


Figure 2: Summary of cases considered

2 Expectation Propagation

Expectation Propagation is an approximate Bayesian inference scheme [10]. Let the posterior requiring approximation be $p(\mathbf{x})$, and $q(\mathbf{x})$ be the approximating distribution. EP attempts to obtain an ideal approximation $q(\mathbf{x})$ by minimising $\mathcal{KL}$ divergences. One idea would be to minimise the global $\mathcal{KL}$ divergence between $p(\mathbf{x})$ and $q(\mathbf{x})$,

$$\mathcal{KL}\left(p(\mathbf{x}) \,\|\, q(\mathbf{x})\right).$$

Although minimising this global $\mathcal{KL}$ divergence is useful in some situations, it does not lead to a good approximation when $p(\mathbf{x})$ is strongly multi-modal: $q(\mathbf{x})$ tends to an average of the modes of $p(\mathbf{x})$, potentially causing most of $q(\mathbf{x})$'s mass to be in between $p(\mathbf{x})$'s modes [11], a poor approximation. A different idea breaks $p(\mathbf{x})$ into factors, $t_n(\mathbf{x})$. Each of these factors is approximated by $\tilde{t}_n(\mathbf{x})$ in $q(\mathbf{x})$. As an example (the case considered in this report), assume that $p(\mathbf{x})$ can be simply factorised into a product of of $t_n(\mathbf{x})$'s,

$$p(\mathbf{x}) = \frac{1}{Z_{\text{true}}} \left(\prod_{n=0}^{N} t_n(\mathbf{x}) \right) \approx \frac{1}{Z_{\text{EP}}} \left(\prod_{n=0}^{N} \tilde{t}_n(\mathbf{x}) \right) = q(\mathbf{x}), \tag{1}$$

$$\text{where } Z_{\text{true}} = \int \prod_{n=0}^{N} t_n(\mathbf{x}) \, d\mathbf{x}, \tag{2}$$

$$\text{and } Z_{\text{EP}} = \int \prod_{n=0}^{N} \tilde{t}_n(\mathbf{x}) \, d\mathbf{x}. \tag{3}$$

We can then attempt to approximate $p(\mathbf{x})$ by minimising a $\mathcal{KL}$ divergence for each $\tilde{t}_n(\mathbf{x})$ in turn,

$$\mathcal{KL}\left(p(\mathbf{x}) \,\middle\|\, \frac{p(\mathbf{x})}{t_n(\mathbf{x})} \tilde{t}_n(\mathbf{x}) \right) = \mathcal{KL}\left(\frac{p(\mathbf{x})}{t_n(\mathbf{x})} t_n(\mathbf{x}) \,\middle\|\, \frac{p(\mathbf{x})}{t_n(\mathbf{x})} \tilde{t}_n(\mathbf{x}) \right). \tag{4}$$

With $\tilde{t}_n(\mathbf{x})$ chosen from an exponential family, minimising the unnormalised $\mathcal{KL}$ divergence in Equation 4 is mathematically the same as matching relevant moments. However, it is often intractable to calculate these moments for the complete distribution $p(\mathbf{x})$. EP therefore replaces $\frac{p(\mathbf{x})}{t_n(\mathbf{x})}$ by its approximation $\frac{q(\mathbf{x})}{\tilde{t}_n(\mathbf{x})}$. In fact, this is arguably more sensible than the idealised approach: if $q(\mathbf{x})$ is the 'best' approximating distribution for $p(\mathbf{x})$, it would seem sensible to update $q(\mathbf{x})$ as a whole instead of breaking it into factors and updating them separately.

Overall, when EP is applied as an approximate inference technique, $p(\mathbf{x})$ is broken into

factors, $t_n(\mathbf{x})$. Each of these factors is approximated by $\tilde{t}_n(\mathbf{x})$ as in Equation 1, all of which are of the same exponential family, such as Gaussian, and are suitably initialised. EP then iteratively refines each $\tilde{t}_n(\mathbf{x})$ by minimising the local (unnormalised) $\mathcal{KL}$ divergence, seen in Equation 5, until convergence. Each approximating factor $\tilde{t}_n(\mathbf{x})$ is treated in turn, first removed (divided out) from the approximating posterior $q(\mathbf{x})$, giving the cavity distribution, before being replaced by $\tilde{t}_n^{\text{new}}(\mathbf{x})$, according to

$$\underset{\tilde{t}_n^{\text{new}}(\mathbf{x})}{\operatorname{argmin}} \mathcal{KL}\left(\frac{q(\mathbf{x})}{\tilde{t}_n(\mathbf{x})} t_n(\mathbf{x}) \,\middle\|\, \frac{q(\mathbf{x})}{\tilde{t}_n(\mathbf{x})} \tilde{t}_n^{\text{new}}(\mathbf{x})\right). \tag{5}$$

As mentioned, minimising this unnormalised $\mathcal{KL}$ Divergence is mathematically the same as matching relevant moments. For a Gaussian approximating family (the only case considered in this report), the properties of $\tilde{t}_n^{\text{new}}(x)$ are chosen such that the 0^{th}, 1^{st} and 2^{nd} order moments of the two terms in the $\mathcal{KL}$ Divergence are the same. To do so, we first need to calculate these moments for the left hand term in the $\mathcal{KL}$ Divergence in Equation 5, known as the tilted posterior, and then set $\tilde{t}_n^{\text{new}}(x)$ such that the right hand term matches these moments.

The normalising constants for $p(\mathbf{x})$ and $q(\mathbf{x})$, Z_{true} (defined in Equation 2) and Z_{EP} (defined in Equation 3) respectively, are often important. Z_{true} can be interpreted as the evidence for the model, an important quantity for many applications. Due to the typically complex nature of $p(\mathbf{x})$, calculating Z_{true} exactly is often intractable, and we work with the computationally easier Z_{EP}. Understanding and characterising the relationship between Z_{EP} and Z_{true} therefore allows the quality of the EP solution to be assessed, informing appropriate usage.

3 Applying Expectation Propagation

The toy cases considered in this project apply EP to a multi-dimensional Gaussian function prior multiplied by N factors. These factors are either Heaviside step functions or softened step functions (probit functions). A Gaussian approximating family is used. This section sets up these cases, and derives the equations required to implement the EP algorithm.

The general expression for the probability distribution $p(\mathbf{x})$, with Heaviside functions, is given in Equation 6. The Gaussian prior $p_0(\mathbf{x})$ (usually the standard normal function) has been explicitly shown in the equation to separate it from the Heaviside functions. The Heaviside functions are denoted by $h(\mathbf{w}_n^T\mathbf{x}+b_n)$, where $\frac{\mathbf{w}_n}{|\mathbf{w}_n|}$ indicates the direction of the function, and $\frac{b_n}{|\mathbf{w}_n|}$ denotes the cut-off point. We only use $|\mathbf{w}_n| = 1$ when considering Heaviside functions. An example in 1-dimension can be seen in Figure 3, where the thin purple line is the standard normal prior, and there are three Heaviside functions applied ($N = 3$): $h_1(x) = h(-x+2)$, $h_2(x) = h(x+1)$, $h_3(x) = h(x+3)$. When these are multiplied with the prior, we obtain the exact posterior $p(x)$, plotted in a thick purple line.

$$p(\mathbf{x}) = \frac{1}{Z_{\text{true}}}\left(p_0(\mathbf{x})\prod_{n=1}^{N} h(\mathbf{w}_n^T\mathbf{x}+b_n)\right) \approx \frac{1}{Z_{\text{EP}}}\left(p_0(\mathbf{x})\prod_{n=1}^{N}\tilde{t}_n(\mathbf{x})\right) = q(\mathbf{x}),\ \mathbf{x}\in\mathbb{R}^d \quad (6)$$

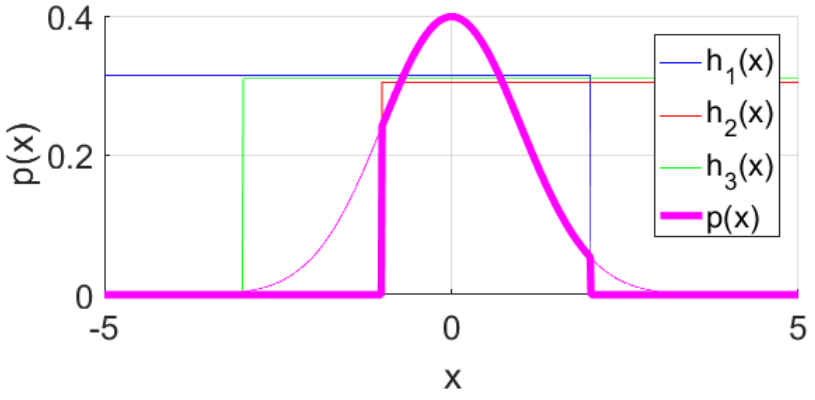


Figure 3: An example of the 1-dimensional toy case. Note that the step functions' magnitudes have been scaled.

In this EP approximation, each Heaviside function is approximated by a 1-dimensional Gaussian, $\tilde{t}_n(x)$, whose parameters need to be found. Note that $h_3(x)$ does not affect the form of $p(x)$ given $h_1(x)$ and $h_2(x)$; however, it is still approximated by $\tilde{t}_3(x)$ when EP is applied.

As we are dealing with a Gaussian approximating family, minimising the local unnormalised $\mathcal{KL}$ Divergence according to Equation 5 is equivalent to equating the 0^{th}, 1^{st} and 2^{nd} order moments. My first steps for this project were to mathematically derive the equations that equate these moments, and apply these equations as part of an EP algorithm in Matlab. As is well-known, it is non-trivial to implement EP, and the derivation that follows is long.

3.1 EP applied with Heaviside functions

This section derives the equations that iteratively implement EP in the multivariate case, where a Gaussian prior is multiplied with N Heaviside functions as in Equation 6. The approximating family is a multivariate Gaussian.

Consider the form of the multivariate Gaussian, written in terms of natural parameters,

$$\tilde{t}_n(\mathbf{x}) = e^{\alpha^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \mathbf{P} \mathbf{x} + \ln Z}, \; Z = \frac{|\mathbf{P}|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\alpha^T \mathbf{P}^{-1} \alpha}. \tag{7}$$

Let $\tilde{t}_n(\mathbf{x})$, the approximating multivariate Gaussian, have 0^{th} order moment Z_{n}, natural mean parameter α_{n}, and precision matrix $\mathbf{P}_{\text{n}}$. These properties need to be calculated every time EP updates $\tilde{t}_n^{\text{new}}(\mathbf{x})$. Let the cavity, defined in Equation 8, have natural parameters α_{cav} and $\mathbf{P}_{\text{cav}}$. Note that the cavity is unnormalised.

$$\text{Cavity} = p_0(\mathbf{x}) \prod_{n' \neq n} \tilde{t}_{n'}(\mathbf{x}). \tag{8}$$

$$\alpha_{\text{cav}} = \alpha_0 + \sum_{n' \neq n} \alpha_{n'}. \tag{9}$$

$$\mathbf{P}_{\text{cav}} = \mathbf{P}_0 + \sum_{n' \neq n} \mathbf{P}_{n'}. \tag{10}$$

First, we calculate the moments of the tilted posterior, defined in Equation 11: Z_0 (the normalising constant), $\mathbf{Z}_1$ and $\mathbf{Z}_2$. We relate $\mathbf{Z}_1$ and $\mathbf{Z}_2$ to Z_0, $\frac{dZ_0}{d\alpha_{\text{cav}}}$ and $\frac{d^2 Z_0}{d{\alpha_{\text{cav}}}^2}$, all of which are found analytically. These moments are then used to calculate the parameters of $\tilde{t}_n^{\text{new}}(\mathbf{x})$,

the new approximating distribution for the n'th Heaviside function.

$$\text{Tilted Posterior} = p_0(\mathbf{x}) \prod_{n' \neq n} \tilde{t}_{n'}(\mathbf{x})\, h(\mathbf{w}_n^T \mathbf{x} + b_n) \tag{11}$$

$$= e^{\alpha_{\text{cav}}^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \mathbf{P}_{\text{cav}} \mathbf{x} + \ln Z_{\text{cav}}}\, h(\mathbf{w}_n^T \mathbf{x} + b_n),\; Z_{\text{cav}} = \frac{|\mathbf{P}_{\text{cav}}|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\alpha_{\text{cav}}^T \mathbf{P}_{\text{cav}}^{-1} \alpha_{\text{cav}}}. \tag{12}$$

We first find the 0^{th} moment of the tilted posterior,

$$\begin{aligned} Z_0 &= \int e^{\alpha_{\text{cav}}^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \mathbf{P}_{\text{cav}} \mathbf{x} + \ln Z_{\text{cav}}}\, h(\mathbf{w}_n^T \mathbf{x} + b_n)\, d\mathbf{x} \\ &= \int \mathcal{N}(u; \mathbf{w}_n^T \mu_{\text{cav}}, \mathbf{w}_n^T \mathbf{\Sigma}_{\text{cav}} \mathbf{w}_n)\, h(u + b_n)\, du \\ &= 0.5 + 0.5 \operatorname{erf}\left(\frac{\sqrt{P_{\text{new}}}}{\sqrt{2}} \left(\frac{\alpha_{\text{new}}}{P_{\text{new}}} - b_n \right) \right) \\ \therefore Z_0 &= 0.5 + 0.5 \operatorname{erf}(z), \end{aligned} \tag{13}$$

$$\text{where } \alpha_{\text{new}} = \frac{\mathbf{w}_n^T \mu_{\text{cav}}}{\mathbf{w}_n^T \mathbf{\Sigma}_{\text{cav}} \mathbf{w}_n},\; P_{\text{new}} = \frac{1}{\mathbf{w}_n^T \mathbf{\Sigma}_{\text{cav}} \mathbf{w}_n},\; z = \frac{\sqrt{P_{\text{new}}}}{\sqrt{2}} \left(\frac{\alpha_{\text{new}}}{P_{\text{new}}} - b_n \right). \tag{14}$$

We then find the 1^{st} moment of the normalised tilted posterior,

$$\mathbf{Z}_1 = \frac{1}{Z_0} \int \mathbf{x}\, e^{\alpha_{\text{cav}}^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \mathbf{P}_{\text{cav}} \mathbf{x} + \ln Z_{\text{cav}}}\, h(\mathbf{w}_n^T \mathbf{x} + b_n)\, d\mathbf{x}.$$

$$\text{Consider } \frac{dZ_0}{d\alpha_{\text{cav}}} = \int \left(\mathbf{x}^T + \frac{d\left(\ln Z_{\text{cav}}\right)}{d\alpha_{\text{cav}}} \right) e^{\alpha_{\text{cav}}^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \mathbf{P}_{\text{cav}} \mathbf{x} + \ln Z_{\text{cav}}}\, h(\mathbf{w}_n^T \mathbf{x} + b_n)\, d\mathbf{x}.$$

$$\text{Now, } \frac{d\left(\ln Z_{\text{cav}}\right)}{d\alpha_{\text{cav}}} = -\left(\mathbf{P}_{\text{cav}}^{-1} \alpha_{\text{cav}}\right)^T,$$

$$\therefore \frac{dZ_0}{d\alpha_{\text{cav}}} = Z_0 \mathbf{Z}_1^T - Z_0 \left(\mathbf{P}_{\text{cav}}^{-1} \alpha_{\text{cav}}\right)^T,$$

$$\therefore \mathbf{Z}_1 = \frac{\left(\frac{dZ_0}{d\alpha_{\text{cav}}}\right)^T}{Z_0} + \mathbf{P}_{\text{cav}}^{-1} \alpha_{\text{cav}}. \tag{15}$$

We can find an analytical expression for $\frac{dZ_0}{d\alpha_{\text{cav}}}$ (Z_0 is defined in Equation 13, and z is defined

in Equation 14),

$$\begin{aligned}\frac{dZ_0}{d\alpha_{\text{cav}}} &= \frac{dZ_0}{d\mu_{\text{cav}}}\frac{d\mu_{\text{cav}}}{d\alpha_{\text{cav}}} \\ &= \frac{dZ_0}{dz}\frac{dz}{d\mu_{\text{cav}}}\mathbf{P}_{\text{cav}}^{-1}, \text{ with } \frac{d\left(\operatorname{erf} z\right)}{dz} = \frac{2}{\sqrt{\pi}}e^{-z^2}, \\ \therefore \frac{dZ_0}{d\alpha_{\text{cav}}} &= \mathbf{w}_n^T\mathbf{P}_{\text{cav}}^{-1} \times \frac{\sqrt{P_{\text{new}}}}{\sqrt{2\pi}}e^{-z^2}. \end{aligned} \tag{16}$$

Similarly, the 2$^{\text{nd}}$ moment, $\mathbf{Z}_2$, can be found,

$$\mathbf{Z}_2 = \frac{1}{Z_0}\int \mathbf{x}\mathbf{x}^T\, e^{\alpha_{\text{cav}}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{P}_{\text{cav}}\mathbf{x} + \ln Z_{\text{cav}}}\, h(\mathbf{w}_n^T\mathbf{x} + b_n)\, d\mathbf{x}.$$

$$\text{Consider } \frac{d^2Z_0}{d{\alpha_{\text{cav}}}^2} = \frac{d}{d\alpha_{\text{cav}}}\left[Z_0\mathbf{Z}_1^T - Z_0\left(\mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\right)^T\right].$$

$$\begin{aligned}\text{Now, } \frac{d\left(Z_0\mathbf{Z}_1\right)}{d\alpha_{\text{cav}}} &= \frac{d}{d\alpha_{\text{cav}}}\left[\int \mathbf{x}\, e^{\alpha_{\text{cav}}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{P}_{\text{cav}}\mathbf{x} + \ln Z_{\text{cav}}}\, h(\mathbf{w}_n^T\mathbf{x} + b_n)\, d\mathbf{x}\right] \\ &= Z_0\mathbf{Z}_2 - Z_0\mathbf{Z}_1\left(\mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\right)^T.\end{aligned}$$

$$\text{And, } \frac{d\left(Z_0\mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\right)}{d\alpha_{\text{cav}}} = Z_0\mathbf{P}_{\text{cav}}^{-1} + Z_0\mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\left(\mathbf{Z}_1 - \mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\right)^T.$$

$$\therefore \mathbf{Z}_2 = \frac{\left(\frac{d^2Z_0}{d{\alpha_{\text{cav}}}^2}\right)}{Z_0} + \mathbf{P}_{\text{cav}}^{-1} - \left(\mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\right)\left(\mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\right)^T + \mathbf{Z}_1\left(\mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\right)^T + \left(\mathbf{P}_{\text{cav}}^{-1}\alpha_{\text{cav}}\right)\mathbf{Z}_1^T, \tag{17}$$

as $\mathbf{Z}_2^T = \mathbf{Z}_2$. We can find an analytical expression for $\frac{d^2Z_0}{d{\alpha_{\text{cav}}}^2}$,

$$\begin{aligned}\frac{d^2Z_0}{d{\alpha_{\text{cav}}}^2} &= \frac{d}{d\mu_{\text{cav}}}\left(\frac{e^{-z^2}}{\sqrt{2\pi}}\sqrt{P_{\text{new}}}\left(\mathbf{w}_n^T\mathbf{P}_{\text{cav}}^{-1}\right)\right) \times \frac{d\mu_{\text{cav}}}{d\alpha_{\text{cav}}} \\ &= \left(\frac{-ze^{-z^2}}{\sqrt{\pi}}P_{\text{new}}\right) \times \mathbf{P}_{\text{cav}}^{-1}\mathbf{w}_n\mathbf{w}_n^T\mathbf{P}_{\text{cav}}^{-1},\end{aligned} \tag{18}$$

as $\left(\mathbf{P}_{\text{cav}}^{-1}\right)^T = \mathbf{P}_{\text{cav}}^{-1}$.

We now calculate the parameters of $\tilde{t}_n^{\text{new}}(x)$ using the computed $Z_0, \mathbf{Z}_1, \mathbf{Z}_2$. Consider $q(\mathbf{x})$, which has natural parameters $\alpha_q = \alpha_{\text{cav}} + \alpha_n$, $\mathbf{P}_q = \mathbf{P}_{\text{cav}} + \mathbf{P}_n$. We can write Z_0 in terms of the parameters of $\mathbf{P}_q$, and use this to calculate Z_n.

$$Z_0 = \int e^{\alpha_q^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{P}_q\mathbf{x} + \ln\left(Z_nZ_{\text{cav}}\right)}d\mathbf{x}, Z_{\text{cav}} \text{ from Equation 12.}$$

$$\text{Now, } \int q(\mathbf{x})\, d\mathbf{x} = \int e^{\alpha_q^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \mathbf{P}_q \mathbf{x} + \ln(Z_q)}\, d\mathbf{x} = 1,$$

$$\therefore \int e^{\alpha_q^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \mathbf{P}_q \mathbf{x}}\, d\mathbf{x} = Z_q^{-1}.$$

$$\therefore Z_n = \frac{Z_0 Z_q}{Z_{\text{cav}}},$$

$$\therefore Z_n = Z_0 \times \frac{|\mathbf{P}_q|^{\frac{1}{2}}}{|\mathbf{P}_{\text{cav}}|^{\frac{1}{2}}} \times e^{\frac{1}{2}\alpha_{\text{cav}}^T \mathbf{P}_{\text{cav}}^{-1} \alpha_{\text{cav}} - \frac{1}{2}\alpha_q^T \mathbf{P}_q^{-1} \alpha_q}. \tag{19}$$

$$\therefore \ln Z_n = \ln Z_0 + \ln Z_q - \ln Z_{\text{cav}}. \tag{20}$$

Natural logarithms of the normalising constants are used for numerical stability. We calculate $\ln(Z_q)$ and $\ln(Z_{\text{cav}})$ according to

$$\ln Z = -\alpha^T \mathbf{P}^{-1} \alpha + \frac{1}{2} \ln |\mathbf{P}| - \frac{d}{2} \ln(2\pi) \tag{21}$$

which is obtained by taking logs of Equation 7.

We can use the definitions of natural parameters to write down, and then solve, Equations 22 and 23. This gives expressions for α_q, α_n, $\mathbf{P}_q$ and $\mathbf{P}_n$. The update for $\tilde{t}_n^{\text{new}}(x)$ is ignored if the final updated precision, $\mathbf{P}_q$, is negative.

$$\text{Now, } \mathbf{Z}_1 = \mathbf{P}_q^{-1} \alpha_q = (\mathbf{P}_n + \mathbf{P}_{\text{cav}})^{-1} (\alpha_n + \alpha_{\text{cav}}), \tag{22}$$

$$\text{and, } \mathbf{Z}_2 = \mathbf{P}_q^{-1} + \mathbf{Z}_1 \mathbf{Z}_1^T = (\mathbf{P}_n + \mathbf{P}_{\text{cav}})^{-1} + \mathbf{Z}_1 \mathbf{Z}_1^T. \tag{23}$$

$$\alpha_q = \left(\mathbf{Z}_2 - \mathbf{Z}_1 \mathbf{Z}_1^T\right)^{-1} \mathbf{Z}_1. \tag{24}$$

$$\alpha_n = \alpha_q - \alpha_{\text{cav}}. \tag{25}$$

$$\mathbf{P}_q = \left(\mathbf{Z}_2 - \mathbf{Z}_1 \mathbf{Z}_1^T\right)^{-1}. \tag{26}$$

$$\mathbf{P}_n = \mathbf{P}_q - \mathbf{P}_{\text{cav}}. \tag{27}$$

Overall, applying EP in multi-dimensions with Heaviside functions involves iteratively updating $\ln(Z_n)$, α_n and $\mathbf{P}_n$ until convergence. We first calculate the natural parameters of the cavity distribution, given by Equations 9 and 10. These are used to calculate the moments of the tilted posterior: Z_0 is given by Equations 13 and 14, $\mathbf{Z}_1$ is given by Equations 15 and 16, and $\mathbf{Z}_2$ is given by Equations 17 and 18. We can now calculate the parameters of $\tilde{t}_n^{\text{new}}(x)$: $\ln(Z_n)$ is given by Equations 20, 21, 24 and 26, α_n is given by Equation 25, and $\mathbf{P}_n$ is given

by Equation 27.

Finally, after convergence, the parameters of the final EP approximation, $q(\mathbf{x})$, can be found. Finding the expression for Z_{EP} uses a similar trick as used to calculate Z_n in Equation 19, and $\ln(Z_{\text{final}})$ is given by Equation 21.

$$\alpha_{\text{EP}} = \sum_{n=0}^{N} \alpha_n, \tag{28}$$

$$\mathbf{P}_{\text{EP}} = \sum_{n=0}^{N} \mathbf{P}_n, \tag{29}$$

$$Z_{\text{EP}} = -\ln Z_{\text{final}} + \sum_{n=0}^{N} \ln Z_n. \tag{30}$$

3.2 EP applied with probit functions

After EP was implemented with Heaviside functions, these functions were softened to probit functions. The form of a probit function is plotted in Figure 4, and is

$$\text{probit}(\mathbf{w}_n^T\mathbf{x} + b_n) = \frac{1}{2} + \frac{1}{2}\,\text{erf}\,(\mathbf{w}_n^T\mathbf{x} + b_n), \tag{31}$$

where $|\mathbf{w}_n|$ affects the smoothness of the function.

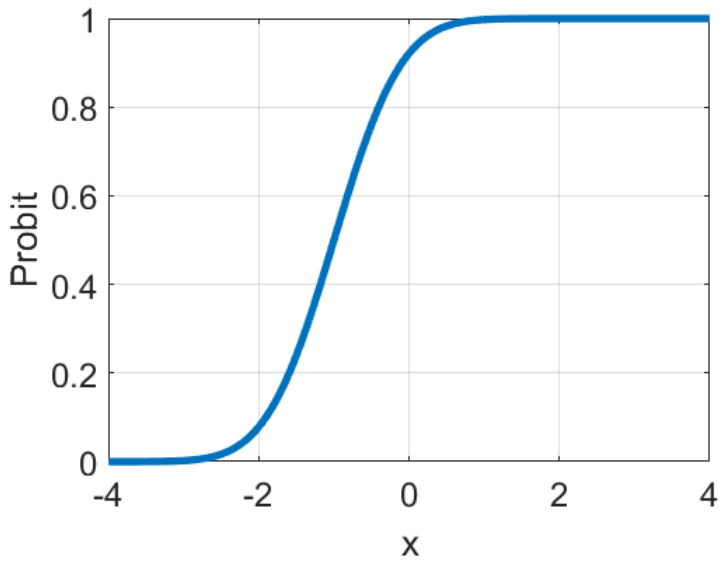


Figure 4: An example of the 1-dimensional probit function, with $w_n = b_n = 1$

Implementing softened step functions required replacing the expressions for Z_0, $\frac{dZ_0}{d\alpha_{\text{cav}}}$ and $\frac{d^2 Z_0}{d\alpha_{\text{cav}}{}^2}$. The rest of the implementation of EP is unchanged, and is as in Section 3.1.

$$\begin{aligned} Z_0 &= \int e^{\alpha_{\text{cav}}^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \mathbf{P}_{\text{cav}} \mathbf{x} + \ln Z_{\text{cav}}} \,\text{probit}(\mathbf{w}_n^T \mathbf{x} + b_n)\, d\mathbf{x} \\ &= 0.5 + 0.5 \int \mathcal{N}(u; \mathbf{w}_n^T \mu_{\text{cav}}, \mathbf{w}_n^T \mathbf{\Sigma}_{\text{cav}} \mathbf{w}_n)\, \text{erf}\,(u + b_n)\, du \end{aligned}$$

To calculate this integral, we first notice that the integral is zero when b_n is zero (the integrand is an odd function). We then calculate

$$\begin{aligned} Z_0 &= 0.5 + 0.5 \int_0^{b_n} \frac{\partial Z_0}{\partial b_n}\, db_n \\ &= 0.5 + 0.5\, \text{erf} \left(\frac{b_n + \frac{\alpha_{\text{new}}}{P_{\text{new}}}}{\sqrt{1 + \frac{2}{P_{\text{new}}}}} \right) \\ \therefore Z_0 &= 0.5 + 0.5\, \text{erf}\,(z) \end{aligned} \tag{32}$$

$$\text{Where } \alpha_{\text{new}} = \frac{\mathbf{w}_n^T \mu_{\text{cav}}}{\mathbf{w}_n^T \mathbf{\Sigma}_{\text{cav}} \mathbf{w}_n},\ P_{\text{new}} = \frac{1}{\mathbf{w}_n^T \mathbf{\Sigma}_{\text{cav}} \mathbf{w}_n},\ z = \frac{b_n + \frac{\alpha_{\text{new}}}{P_{\text{new}}}}{\sqrt{1 + \frac{2}{P_{\text{new}}}}}. \tag{33}$$

We can now find

$$\begin{aligned}\frac{dZ_0}{d\alpha_{\text{cav}}} &= \frac{dZ_0}{dz}\frac{dz}{d\mu_{\text{cav}}}\frac{d\mu_{\text{cav}}}{d\alpha_{\text{cav}}} \\ &= \mathbf{w}_n^T\mathbf{P}_{\text{cav}}^{-1} \times \frac{e^{-z^2}}{\sqrt{\pi\left(1+\frac{2}{P_{\text{new}}}\right)}}. \end{aligned} \tag{34}$$

$$\begin{aligned}\frac{d^2Z_0}{d{\alpha_{\text{cav}}}^2} &= \frac{d}{d\mu_{\text{cav}}}\left(\frac{e^{-z^2}}{\sqrt{\pi\left(1+\frac{2}{P_{\text{new}}}\right)}}\left(\mathbf{w}_n^T\mathbf{P}_{\text{cav}}^{-1}\right)\right) \times \frac{d\mu_{\text{cav}}}{d\alpha_{\text{cav}}} \\ &= \left(\frac{-2ze^{-z^2}}{\sqrt{\pi}\left(1+\frac{2}{P_{\text{new}}}\right)}\right) \times \mathbf{P}_{\text{cav}}^{-1}\mathbf{w}_n\mathbf{w}_n^T\mathbf{P}_{\text{cav}}^{-1}. \end{aligned} \tag{35}$$

where z and P_{new} are given in Equation 33. Therefore, to implement EP with probit functions, we replace Equations 13, 14, 16 and 18 with Equations 32, 33, 34 and 35 respectively.

4 Symmetric box case

One of the conjectures in the machine learning community regarding EP is that the normalising constant obtained via EP, Z_{EP}, is an underestimate of the true normalising constant, Z_{true}; equivalently, EP always underestimates the true model likelihood. Empirical evidence supporting this conjecture has been found [7]. However, it has been difficult to characterise the relationship between Z_{EP} and Z_{true}, even after making many approximations [12]. The symmetric box case is a toy case counter example to this conjecture, showing that it is not always true.

Equation 36 summarises the 1-dimensional setup for this example. The prior, $p_0(x)$, is the standard normal Gaussian. This is multiplied by two Heaviside functions that are symmetric about $x = 0$, with cut-offs at $+b$ and $-b$, as shown in Figure 5. This results in the true distribution, $p(x)$, shown in a thick purple line. In the EP approximation $q(x)$, each of the Heaviside functions are approximated by Gaussian functions, resulting in the final Gaussian approximation shown in blue.

$$p(x) = \frac{1}{Z_{\text{true}}}\left(p_0 x\right)h(x+b)h(-x+b)) \approx \frac{1}{Z_{\text{EP}}}\left(p_0(x)\tilde{t}_1(x)\tilde{t}_2(x)\right) = q(x). \tag{36}$$

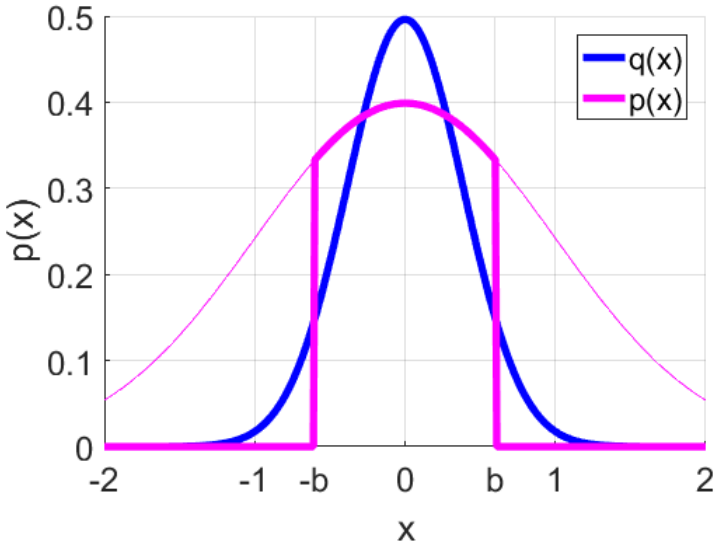


Figure 5: Symmetric box setup

Figure 6 plots the EP overestimate of the true model likelihood ($Z_{\text{EP}} > Z_{\text{true}}$), a finding

contrary to the community's conjecture, as a function of b. As $\log_e(b)$ decreases, the distributions get narrower around $x = 0$, and the overestimation ratio increases, with a maximum overestimation of about 10%.

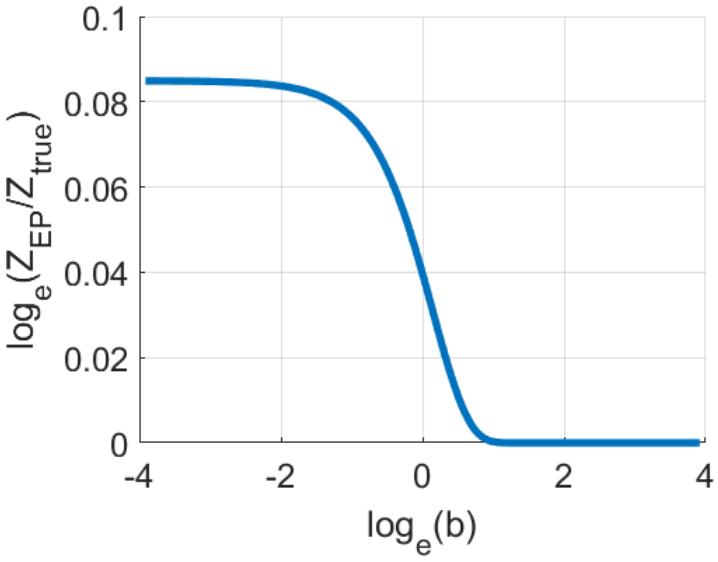


Figure 6: Overestimation of normalising constant, with Heaviside functions

Softening the Heaviside functions to probit functions still leads to an overestimation effect. This confirms that the overestimation seen is not solely due to the discontinuous nature of the Heaviside function. The setup for this case is

$$p(x) = \frac{1}{Z_{\text{true}}} \left(p_0(x)\text{probit}(w_n x + b)\text{probit}(-w_n x + b)\right) \approx \frac{1}{Z_{\text{EP}}} \left(p_0(x)\tilde{t}_1(x)\tilde{t}_2(x)\right) = q(x).$$

The prior is still the standard normal Gaussian, but this is now multiplied by two probit functions to give the true posterior. There is now an additional parameter, $|w_n|$, the weight of the probit functions. The overestimation as $|w_n|$ changes can be seen in Figure 7. As $|w_n|$ increases, the probit function tends towards the more non-smooth Heaviside function, and the graphs become similar at very large $|w_n|$, with an overestimation of about 10%. However, as can be seen, there is still an overestimation at smaller $|w_n|$.

Increasing the dimensionality of this example, but keeping pairs of Heaviside functions (or probit functions) uncorrelated in dimensions (so that the problem can be broken into uncorrelated 1D cases), increases this overestimation in a predictable fashion: if there are d dimensions, then the overestimation ratio becomes a factor of d larger in log-space. Increasing the dimensionality can therefore arbitrarily increase the overestimation observed.

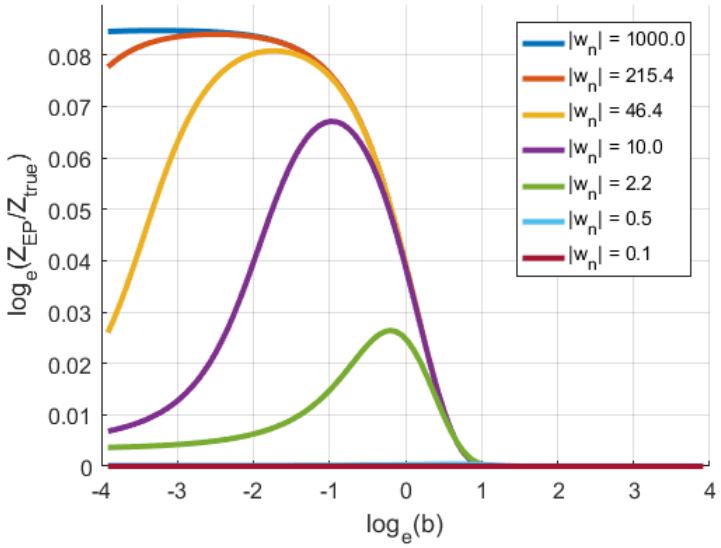


Figure 7: Overestimation of normalising constant, with probit functions

A possible intuitive reason as to why an overestimation is observed uses reasoning based on entropy. At each EP update, the entropy of the tilted distribution is larger than the entropy of the true distribution. This larger entropy can be seen schematically as the shaded area in Figure 8, where we are considering the update for $\tilde{t}_1(x)$. This additional entropy at each EP update could lead to an overall increase in approximate model evidence at convergence of the EP algorithm (Z_{EP}) as compared to the true model evidence.

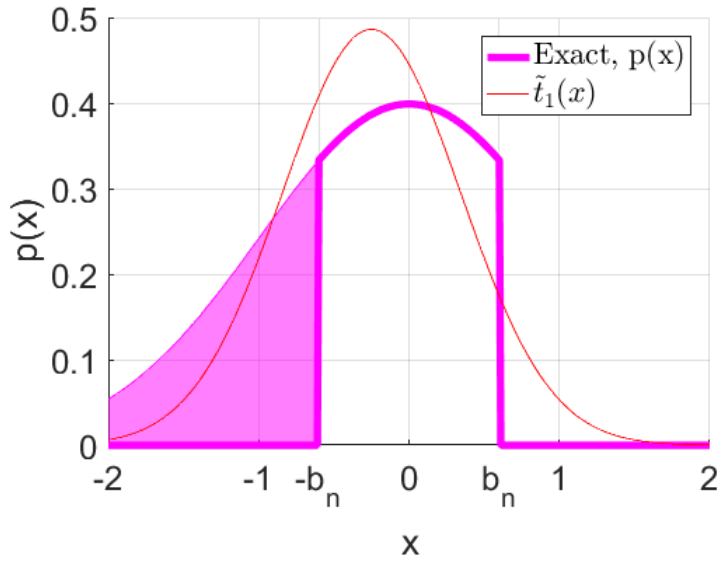


Figure 8: Additional entropy 'seen' by an EP update

In order to present this example of overestimation to the research community in a neater way, I have derived the expression for Z_{EP} from the equations satisfied at the fixed point of EP, and shown that at this fixed point, $Z_{\text{EP}} > Z_{\text{true}}$.

4.1 Proof of counter-example

This section mathematically shows that, in the 1-dimensional symmetric box case, $Z_{\text{EP}} > Z_{\text{true}}$. Showing this is true mathematically makes the result more presentable to the research community, as any replication of the result does not require running an algorithm to convergence (as in the previous section).

We first find expressions for Z_{EP} (and Z_{true}) at the fixed point of EP for the (1-dimensional) symmetric box case. This is represented by a system of non-linear equations. We then show that $Z_{\text{EP}} > Z_{\text{true}}$ at this fixed point.

The EP algorithm approximates $p(x)$ by $q(x)$, both given in Equation 36, by iteratively refining the two $\tilde{t}_n(x)$ according to Equation 5. Therefore, at the fixed point of EP (at convergence), we must have

$$\text{MOM}[q(x)] = \text{MOM}\left[\frac{\hat{q}(x)}{\tilde{t}_1(x)}h_1(x)\right] = \text{MOM}\left[\frac{\hat{q}(x)}{\tilde{t}_2(x)}h_2(x)\right] \tag{37}$$

where $\hat{q}(x) = \frac{q(x)}{\int q(x)\,dx}$ (normalised $q(x)$), and $\text{MOM}[q(x)]$ refers to the 0^{th}, 1^{st} and 2^{nd} moments of the distribution $q(x)$.

We let $\tilde{t}_1(x)$ have 0^{th} moment Z_{t1}, natural mean α_1 ($\alpha_1 > 0$) and precision P_1 at EP's fixed point. By symmetry, we can say that $\tilde{t}_2(x)$ has parameters $Z_{\text{t2}} = Z_{\text{t1}}$, $\alpha_2 = -\alpha_1$ and $P_2 = P_1$. We know that the unit normal prior $p_0(x)$ has natural mean $\alpha_0 = 0$ and precision $P_0 = 1$. We focus on first solving Equation 37 for α_1 and P_1, before using these parameters to find Z_{t1} and hence Z_{EP}.

We first consider solving

$$\text{MOM}[q(x)] = \text{MOM}\left[\frac{\hat{q}(x)}{\tilde{t}_1(x)}h_1(x)\right]. \tag{38}$$

$$\text{Now, } \alpha_q = \alpha_0 + \alpha_1 + \alpha_2 = 0,$$

$$\text{and, } P_q = P_0 + P_1 + P_2 = 1 + 2P_1.$$

We can use Equations 15 and 16 to obtain an expression for the tilted posterior's (the right hand side of Equation 38) 1st moment, $Z_{1,1}$, and Equations 17 and 18 to obtain an expression for its second moment, $Z_{2,1}$. The natural mean of the cavity $\frac{q(x)}{\tilde{t}_1(x)}$ is $\alpha_{\text{cav},1} = \alpha_0 + \alpha_1 = \alpha_1$, and the precision is $P_{\text{cav},1} = P_0 + P_1 = 1 + P_1$. Therefore

$$Z_{1,1} = \frac{-e^{-z^2}}{Z_0\sqrt{2\pi(1+P_1)}} + \frac{\alpha_1}{1+P_1},$$

$$\text{and } Z_{2,1} = \frac{1}{1+P_1} - \frac{\alpha_1^2}{(1+P_1)^2} + \frac{2\alpha_1 Z_{1,1}}{1+P_1} - \frac{ze^{-z^2}}{Z_0\sqrt{\pi}(1+P_1)},$$

$$\text{where } z = \frac{\sqrt{1+P_1}}{\sqrt{2}}\left(\frac{-\alpha_1}{1+P_1} + b\right), \tag{39}$$

$$\text{and } Z_0 = \frac{1}{2} + \frac{1}{2}\,\text{erf}\,(z) = \text{probit}(z). \tag{40}$$

We can now substitute this into Equation 38 and solve for α_1 and P_1, giving

$$\alpha_q = \frac{Z_{1,1}}{Z_{2,1} - Z_{1,1}^2} = 0.$$

$$\therefore Z_{1,1} = 0.$$

$$\therefore \frac{\alpha_1}{1+P_1} = \frac{e^{-z^2}}{Z_0\sqrt{2\pi(1+P_1)}}. \tag{41}$$

$$\text{And, } P_q = \frac{1}{Z_{2,1} - Z_{1,1}^2} = \frac{1}{1+2P_1},$$

$$\therefore Z_{2,1} = \frac{1}{1+2P_1}.$$

Substituting into Equation 41 and simplifying gives

$$\frac{1+P_1}{1+2P_1} = 1 - b\alpha_1,$$

$$\therefore \alpha_1 = \frac{P_1}{b(1+2P_1)}. \tag{42}$$

$$\frac{P_1}{b} = \frac{\sqrt{1+P_1}(1+2P_1)}{\sqrt{2\pi}} \frac{e^{-z^2}}{\frac{1}{2} + \frac{1}{2}\,\text{erf}\,(z)}. \tag{43}$$

Having obtained equations that can be solved for α_1 and P_1 (Equations 39, 42 and 43), we now use α_1 and P_1 to find Z_{EP}, by solving Equation 38 for the 0^{th} moment. First, consider

the left hand side of Equation 38,

$$\int q(x)\,dx = \int p_0(x)\,\tilde{t}_1(x)\,\tilde{t}_2(x)\,dx = \int e^{-\frac{1}{2}x^2(1+2P_1)+\ln Z_{\rm EP}}\,dx = Z_{\rm EP} \times \frac{\sqrt{2\pi}}{\sqrt{1+2P_1}} \tag{44}$$

as $\alpha_q = 0$.

Now consider the right hand side of Equation 38,

$$\begin{aligned}\int p_0(x)\,\tilde{t}_1(x)\,h_2(x)\,dx &= \int e^{\alpha_1 x - \frac{1}{2}x^2(1+P_1)+\ln(Z_{\rm prior}Z_{\rm t1})}\,h_2(x)\,dx \\ &= \frac{Z_{\rm prior}Z_0Z_{\rm t1}\sqrt{2\pi}}{\sqrt{1+P_1}e^{-\frac{\alpha_1^2}{2(1+P_1)}}}\end{aligned} \tag{45}$$

where Z_0 is given in Equation 40. We can also write

$$Z_{\rm EP} = Z_{\rm prior}Z_{\rm t1}Z_{\rm t2} = Z_{\rm prior}Z_{\rm t1}^2.$$

Substituting this in, and equating Equations 44 and 45, gives

$$Z_{\rm prior}Z_{\rm t1}^2\frac{\sqrt{2\pi}}{\sqrt{1+2P_1}} = \frac{Z_{\rm prior}Z_0Z_{\rm t1}\sqrt{2\pi}}{\sqrt{1+P_1}e^{\frac{-\alpha_1^2}{2(1+P_1)}}},$$

$$\therefore Z_{\rm EP} = \frac{\sqrt{1+2P_1}}{1+P_1} \times e^{-\frac{\alpha_1^2}{1+P_1}} \times \left(\frac{1}{2}+\frac{1}{2}\,{\rm erf}\,(z)\right)^2. \tag{46}$$

We can also write, from considering Figure 5,

$$Z_{\rm true} = {\rm erf}\left(\frac{b}{\sqrt{2}}\right). \tag{47}$$

We therefore have a system of equations that can be solved for $Z_{\rm EP}$ (Equations 39, 42, 43 and 46), and an equation for $Z_{\rm true}$ (Equation 47), which depend only on the cut-off value b. If desired, these equations can be numerically solved at suitable (small) values of b ($b = 0.1$ works well) to show $Z_{\rm EP} > Z_{\rm true}$.

We now show that this inequality holds around small positive P_1. We substitute an approximation for P_1 in terms of b (a zeroth order Taylor series expansion of Equation 43 around $P_1 = 0$) into a second order Taylor series expansion of Equation 46 around $P_1 = 0$, which yields an expression for $Z_{\rm EP}$ in terms of P_1 and b. We therefore obtain an expression for $Z_{\rm EP}$ in terms of b at small P_1, which is shown to be greater than $Z_{\rm true}$. Using the second order

Taylor series expansion ensures that we can confirm the proof numerically. We also derive the first order Taylor series expansion of P_1 (Equation 43), useful for numerical analysis.

We will need the following expressions, derived from Equation 39:

$$z|_{P_1=0} = \frac{b}{\sqrt{2}}, \tag{48}$$

$$\left.\frac{\partial z}{\partial P_1}\right|_{P_1=0} = \frac{b}{2\sqrt{2}} - \frac{1}{b\sqrt{2}}. \tag{49}$$

Let $f(P_1) = \frac{P_1}{b}$, as in Equation 43. We can therefore write

$$f(0) = \frac{e^{-\frac{b^2}{2}}}{\sqrt{2\pi} \times \text{probit}\left(\frac{b}{\sqrt{2}}\right)},$$

$$\therefore P = \frac{b}{\sqrt{2\pi}} \times \frac{e^{-\frac{b^2}{2}}}{\text{probit}\left(\frac{b}{\sqrt{2}}\right)} \tag{50}$$

where P is small and positive. At large values of b, we can see that $P = \mathcal{O}\left(e^{-\frac{b^2}{2}}\right)$. We therefore need to apply a Taylor series expansion on the equation for Z_{EP} (Equation 46) to the order required to incorporate all $\mathcal{O}\left(e^{-\frac{b^2}{2}}\right)$ terms, which is found to be the second order Taylor series expansion.

However, at values of b that can be numerically checked, we find that Equation 50 is not very precise, and we need to use the first order Taylor series expansion. The first order Taylor series expansion is

$$f(0+P) = f(0) + P \left.\frac{\partial f(P_1)}{\partial P_1}\right|_{P_1=0} = \frac{P}{b}$$

$$\therefore P = \left(\frac{1}{f(0)b} - \frac{\left.\frac{\partial f(P_1)}{\partial P_1}\right|_{P_1=0}}{f(0)}\right)^{-1}. \tag{51}$$

We can write

$$\begin{aligned}\left.\frac{\partial f(P_1)}{\partial P_1}\right|_{P_1=0} &= \frac{1}{\sqrt{2\pi}} \times \frac{1}{2\sqrt{1+P_1}} \times \left.\frac{e^{-z^2}}{\text{probit}\left(\frac{b}{\sqrt{2}}\right)}\right|_{P_1=0} \\ &\quad + \left.\frac{\sqrt{1+P_1}}{\sqrt{2\pi}} \times \frac{1}{\text{probit}\left(\frac{b}{\sqrt{2}}\right)} \times \left(-2z\frac{\partial z}{\partial P_1}\right) \times e^{-z^2}\right|_{P_1=0} \\ &\quad + \left.\frac{\sqrt{1+P_1}}{\sqrt{2\pi}} \times e^{-z^2} \times \frac{-1}{\left[\text{probit}\left(\frac{b}{\sqrt{2}}\right)\right]^2} \times \left(\frac{1}{2}\frac{\partial(\text{erf}\,(z))}{\partial z}\frac{\partial z}{\partial P_1}\right)\right|_{P_1=0} \\ &\quad + 2f(0) \\ &= 2f(0) + \frac{e^{-\frac{b^2}{2}}}{\text{probit}\left(\frac{b}{\sqrt{2}}\right)}\left(\frac{3-b^2}{\sqrt{8\pi}}\right) + \frac{e^{-b^2}}{\left[\text{probit}\left(\frac{b}{\sqrt{2}}\right)\right]^2}\left(\frac{\frac{2}{b}-b}{4\pi}\right).\end{aligned}$$

Substituting into Equation 51 gives

$$P = \left(\frac{1}{b} \times \sqrt{2\pi} \times e^{\frac{b^2}{2}} \times \text{probit}\left(\frac{b}{\sqrt{2}}\right) - \frac{7-b^2}{2} - \frac{e^{-\frac{b^2}{2}}}{\text{probit}\left(\frac{b}{\sqrt{2}}\right)}\left(\frac{\frac{2}{b}-b}{\sqrt{8\pi}}\right)\right)^{-1}. \tag{52}$$

At large values of b (at approximately $b > \sqrt{7}$), which is assumed as small precision P_1 is only achieved at large b,

$$P < \frac{b}{\sqrt{2\pi}} \times \frac{e^{-\frac{b^2}{2}}}{\text{probit}\left(\frac{b}{\sqrt{2}}\right)}$$

We can therefore see that the approximation in Equation 50 is an over-approximation. This will be useful later, when substituting this expression in. Equation 52 is also useful when numerically checking this derivation.

We now consider the second order Taylor series expansion of Z_{EP} around $P_1 = 0$. We will need the expressions in Equations 48 and 49, as well as the following expressions derived from Equation 42:

$$\begin{aligned}\alpha_1|_{P_1=0} &= 0, \\ \left.\frac{\partial \alpha_1}{\partial P_1}\right|_{P_1=0} &= \left.\frac{b(1+2P_1) - P_1(2b)}{b^2(1+2P_1)^2}\right|_{P_1=0} = \frac{1}{b},\end{aligned}$$

The second order Taylor series expansion of $Z_{\mathrm{EP}}(P_1)$ around $P_1 = 0$ is

$$Z_{\mathrm{EP}}(0+P) = Z_{\mathrm{EP}}(0) + P\left.\frac{\partial Z_{\mathrm{EP}}(P_1)}{\partial P_1}\right|_{P_1=0} + \frac{P^2}{2}\left.\frac{\partial^2 Z_{\mathrm{EP}}(P_1)}{\partial {P_1}^2}\right|_{P_1=0} \tag{53}$$

where P is small and positive.

We consider each of the three terms in Equation 53 in turn. Equation 46 gives

$$Z_{\mathrm{EP}}(0) = \left[\frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{b}{\sqrt{2}}\right)\right]^2 = \left[\mathrm{probit}\left(\frac{b}{\sqrt{2}}\right)\right]^2.$$

$$\begin{aligned}
\left.\frac{\partial Z_{\mathrm{EP}}(P_1)}{\partial P_1}\right|_{P_1=0} &= \left.\frac{1}{2}(1+2P_1)^{-\frac{1}{2}}(2)(1+P_1)^{-1}e^{\frac{\alpha_1^2}{1+P_1}}\left[\mathrm{probit}(z)\right]^2\right|_{P_1=0} \\
&\quad + \left.(1+2P_1)^{-\frac{1}{2}}(-1)(1+P_1)^{-2}e^{\frac{\alpha_1^2}{1+P_1}}\left[\mathrm{probit}(z)\right]^2\right|_{P_1=0} \\
&\quad + \left.(1+2P_1)^{-\frac{1}{2}}(1+P_1)^{-1}\left(\frac{2\alpha_1\frac{\partial\alpha_1}{\partial P_1}(1+P_1)-\alpha_1^2}{(1+P_1)^2}\right)\left[\mathrm{probit}(z)\right]^2\right|_{P_1=0} \\
&\quad + \left.(1+2P_1)^{-\frac{1}{2}}(1+P_1)^{-1}e^{\frac{\alpha_1^2}{1+P_1}}(2)\left[\mathrm{probit}(z)\right]\left(\frac{1}{2}\frac{\partial(\mathrm{erf}\,(z))}{\partial z}\frac{\partial z}{\partial P_1}\right)\right|_{P_1=0} \\
&= \frac{\mathrm{probit}\left(\frac{b}{\sqrt{2}}\right)}{\sqrt{2\pi}}e^{-\frac{b^2}{2}}\left(b-\frac{2}{b}\right).
\end{aligned}$$

Finding an expression for $\left.\frac{\partial^2 Z_{\mathrm{EP}}(P_1)}{\partial {P_1}^2}\right|_{P_1=0}$ is more complicated, but we can ignore terms of $\mathcal{O}(P) = \mathcal{O}(e^{-\frac{b^2}{2}})$ and higher, giving

$$\left.\frac{\partial^2 Z_{\mathrm{EP}}(P_1)}{\partial {P_1}^2}\right|_{P_1=0} = \left[\mathrm{probit}\left(\frac{b}{\sqrt{2}}\right)\right]^2\left(\frac{2}{b^2}-1\right) + \mathcal{O}(e^{-\frac{b^2}{2}}).$$

Substituting these expressions in Equation 53 gives

$$\begin{aligned}
Z_{\mathrm{EP}} = &\left[\mathrm{probit}\left(\frac{b}{\sqrt{2}}\right)\right]^2 + P\frac{\mathrm{probit}\left(\frac{b}{\sqrt{2}}\right)}{\sqrt{2\pi}}e^{-\frac{b^2}{2}}\left(b-\frac{2}{b}\right) \\
&+ \frac{P^2}{2}\left[\mathrm{probit}\left(\frac{b}{\sqrt{2}}\right)\right]^2\left(\frac{2}{b^2}-1\right) + \mathcal{O}(P^3).
\end{aligned}$$

We want to show that this expression for Z_{EP} satisfies $Z_{\mathrm{EP}} - Z_{\mathrm{true}} > 0$, where Z_{true} is given

in Equation 47. Now, for all $b > 0$,

$$\left[\text{probit}\left(\frac{b}{\sqrt{2}}\right)\right]^2 - \text{erf}\left(\frac{b}{\sqrt{2}}\right) > 0.$$

We therefore now need to show that

$$P \times \left[\text{probit}\left(\frac{b}{\sqrt{2}}\right)\right] \times \left[\frac{e^{-\frac{b^2}{2}}}{\sqrt{2\pi}}\left(b - \frac{2}{b}\right) + \frac{P \times \left[\text{probit}\left(\frac{b}{\sqrt{2}}\right)\right]}{2}\left(\frac{2}{b^2} - 1\right)\right] > 0.$$

To do so, we first note that $P > 0$, $\text{probit}\left(\frac{b}{\sqrt{2}}\right) > 0$. We then substitute in P, as given in Equation 50, noting that this is an over-approximation of P,

$$\left[\frac{e^{-\frac{b^2}{2}}}{\sqrt{2\pi}}\left(b - \frac{2}{b}\right) + \frac{P \times \left[\text{probit}\left(\frac{b}{\sqrt{2}}\right)\right]}{2}\left(\frac{2}{b^2} - 1\right)\right] > \frac{e^{-\frac{b^2}{2}}}{\sqrt{2\pi}}\left[\frac{b}{2} - \frac{1}{b}\right] > 0,$$

for $b > \sqrt{2}$, which is assumed as large b is assumed. We have therefore proved that, at small P_1 (and large b),

$$Z_{\text{EP}} > Z_{\text{true}}$$

in a way that can be verified computationally, for the 1-dimensional symmetric box case.

5 Repeated Heaviside functions

This section describes the toy case example of repeated Heaviside functions. As seen in Equation 54 and Figure 9, this example involves N Heaviside functions with cut-off values at $b = 0$, each of which are approximated by a Gaussian function $\tilde{t}_n(x)$.

$$p(x) = \frac{1}{Z_{\text{true}}} \left(p_0(x) \prod_{n=1}^{N} h(-x+0) \right) \approx \frac{1}{Z_{\text{EP}}} \left(p_0(x) \prod_{n=1}^{N} \tilde{t}_n(x) \right) = q(x). \tag{54}$$

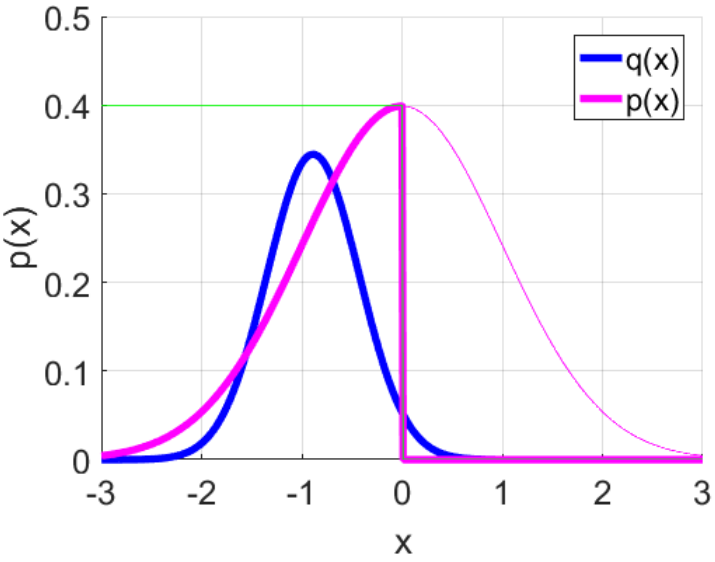


Figure 9: Three repeated Heaviside functions. Note that the magnitudes of the Heaviside functions have been scaled.

This toy case is useful for understanding how EP behaves when classifying well-separated data, where datapoints contribute similar Heaviside functions that multiply together, resulting in a distribution similar to the form of $p(x)$. Figure 10 summarises a simple 1-dimensional binary classification example, where there is well-separated data (the two classes are represented by crosses and circles), and therefore many Heaviside functions on top of each other (shown separated by a small amount).

Repeated Heaviside functions can also provide insight into Power EP (PEP). In PEP, a factor $t_n(x)$ from Equation 1 is replaced by N identical factors $t_n(x)^{\frac{1}{N}}$. Because the Heaviside function only has values 0 or 1, repeating identical Heaviside functions is an application of PEP. It can be shown that PEP attempts to minimise the more general α-Divergence instead

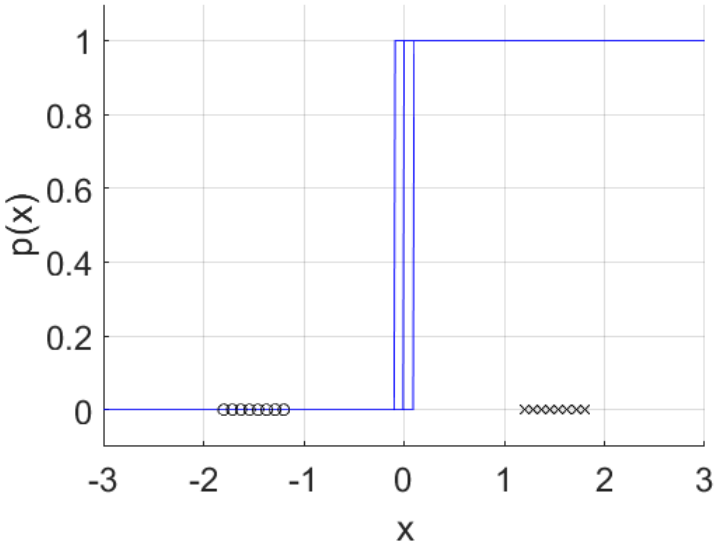


Figure 10: Binary classification example with well-separated data

of the $\mathcal{KL}$ Divergence in Equation 5 [9], with $\alpha = \frac{2}{N} - 1$. Increasing N reduces α, from $\alpha = 1$ (as in EP) towards $\alpha = -1$ (the Variational Inference (VI) case). Additionally, due to the non-smooth potentials due to the Heaviside functions used in this example, VI fails in this example, while EP does not, highlighting the importance of understanding EP's behaviour.

Figure 11 shows the normalising constant, Z_{EP}, increasingly underestimating the true model likelihood as the number of Heaviside functions, N, increases. This follows from a property of the α-Divergence. The underestimation is much greater than 10%, the maximum overestimation observed in the symmetric box case example. This provides a possible explanation for the conjecture in the community (that $Z_{\text{EP}} < Z_{\text{true}}$): in applied data sets, this underestimation could overpower any overestimation effects. This claim is tested in Section 6.

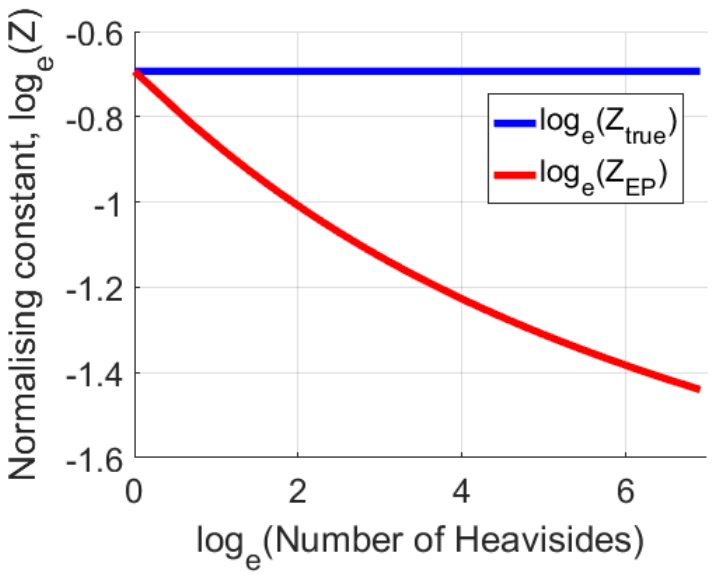


Figure 11: Underestimation of normalising constant

Repeating these tests with probit functions instead of Heaviside functions provides the same result: EP underestimates the model evidence (by more than the overestimation in Section 4). The underestimation increases with increasing number of repeated probit functions. The same underestimation is observed in multiple dimensions.

5.1 Repeated Heaviside functions with jitter

Adding jitter to the value of the Heaviside functions' cut-off value, b, results in a more realistic example (the jitter can be due to noise). This section considers two Heaviside functions in the 1-dimensional case, with a difference in cut-off value of Δb, arranged as in Figure 12 (where $\Delta b = 1$). One Heaviside function, with cut-off value $b = 0$, is held still, while the other is moved as Δb changes.

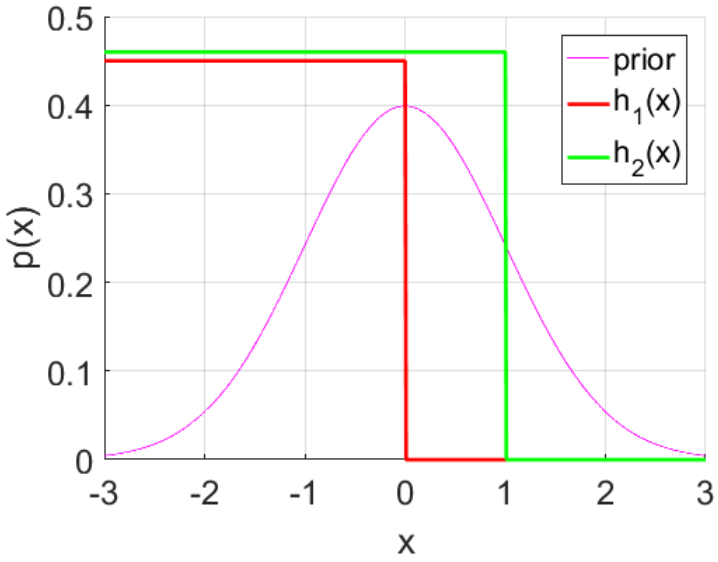


Figure 12: Repeated Heaviside functions with jitter. Note that the magnitudes of the Heaviside functions have been scaled.

Figure 13 shows the true and jitter normalising constants for this example. At $\Delta b = 0$, we have the repeated Heaviside functions case as in Figure 11 with two repeated Heaviside functions. As Δb gets large, Z_{jitter} tends to Z_{true} because the second Heaviside function hardly affects the EP approximation.

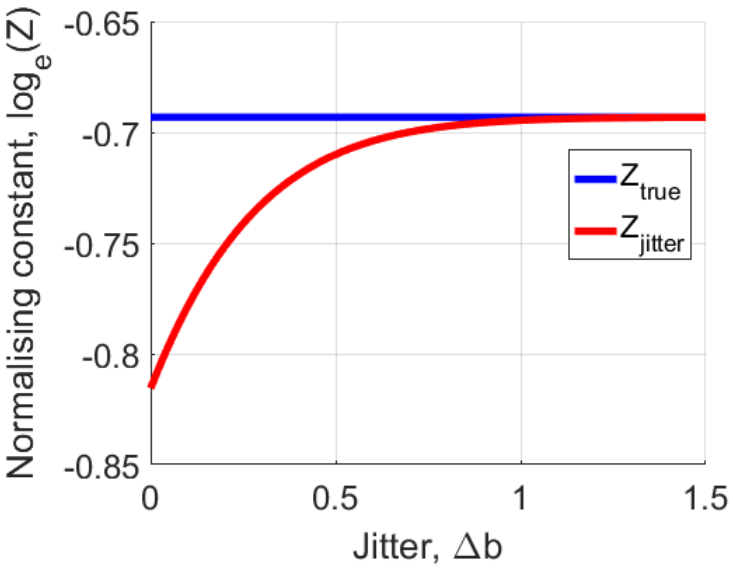


Figure 13: Underestimation of normalising constant with jitter

Therefore we can say that adding jitter of this form reduces the underestimation of the normalising constant, Z. This effect will also be seen when adding jitter in multiple dimensions (we can resolve along dimensions separately to recover this 1-dimensional case). Using probit functions instead of Heaviside functions gives a very similar reducing underestimation.

6 Simple classification example

This section expands the symmetric box case example in Section 4 and the repeated Heaviside example in Section 5 to a more realistic binary classification example, where we are interested in the EP approximation of the model evidence.

Consider a labelled data set, with N inputs $\mathbf{x}_i$, each associated with a label $y_i \in \{1, -1\}$. We would like to learn a probit classifier of the form $\text{probit}(\mathbf{w}^T\mathbf{x}_i)$, where the probit function is defined in Equation 31 (note that this example does not explicitly include the cut-off value b). We assume a unit normal Gaussian prior on $\mathbf{w}$.

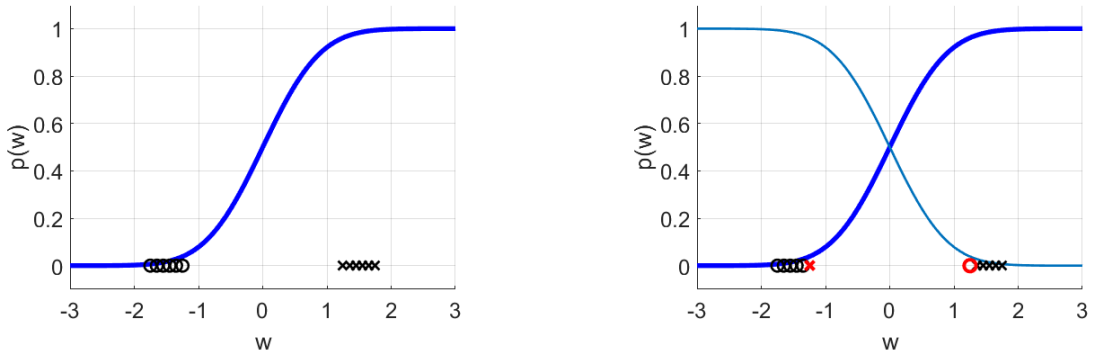
We start with the 1-dimensional case, and expand to multiple dimensions later. We write

$$P(y_i = 1|x_i, w) = \text{probit}(wx_i),\ P(y_i = -1|x_i, w) = 1 - \text{probit}(wx_i) = \text{probit}(-wx_i).$$

First, let us consider well-separated data, where we have 12 data points, 6 each at $(x, y) = (1, 1)$ and $(x, y) = (-1, -1)$. Figure 14a plots this case, with 12 identical probit functions (points labelled $y = 1$ are represented by crosses on the right hand side of the figure, while points labelled $y = -1$ are represented by circles on the left hand side). Running EP on this example is identical to EP on repeated probit functions (see Section 5): EP underestimates the model evidence. We then swap one pair of data points, so that there are 5 points at $(x, y) = (1, 1)$, 5 at $(x, y) = (-1, -1)$, 1 point at $(x, y) = (1, -1)$, and 1 point at $(x, y) = (-1, 1)$. See Figure 14b, where the swapped pair of points are highlighted in red, and there are now two probit functions in the opposite direction. The swapped pair of points now contribute probit functions acting in a similar way to the symmetric box case (see Section 4), and EP's underestimation of the model evidence reduces, as in Figure 15. We can swap two more pairs of data points. As seen in Figure 15, the underestimation is bigger than any overestimation, until the final swap, when the EP model evidence is almost exactly the true model evidence.

Increasing the data points' values of x_i has been considered in Section 4: it reduces the underestimation further. Adding some noise to the value of the data points' has also already been considered (see Section 5.1), also causing a decreasing underestimation. Expanding this example to multiple dimensions is the same as treating this example in 1-dimension (in the absence of noise).

The underestimation of the model evidence overpowers any overestimation due to the 'box



(a) Initial data set, well-separated data (b) Dataset after one pair of datapoints swapped

Figure 14: Classification example setup

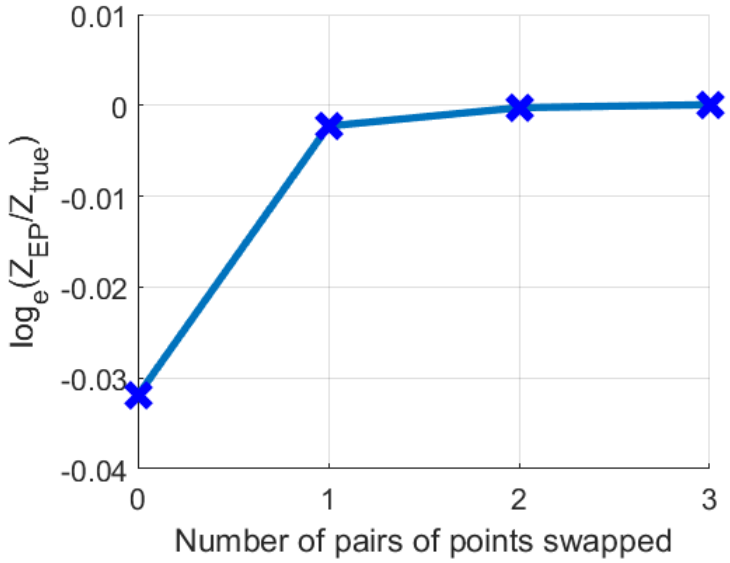


Figure 15: The difference in normalising constants as pairs of points are swapped

case' effect until the final swap. Relaxing the assumptions used for Figure 15, such as by adding some noise or increasing the values of $|\mathbf{x}_i|$, reduces this underestimation slightly, but will not cause an overall overestimation unless there is a strongly separated dataset, when an overall overestimation might be possible. However, such separated datasets are unlikely to be seen in practice. This illustrates why the conjecture that $Z_{\text{EP}} < Z_{\text{true}}$ holds on realistic datasets [7], where there are many data points that are mostly well-separated..

7 FITC

In this section, we use a sparse approximation method for regression models based on Gaussian Processes called the "Fully Independent Training Condition" (FITC) [13]. Applying FITC to a particular case produces another toy case counter-example of the conjecture that $Z_{\text{EP}} < Z_{\text{true}}$ (further to the symmetric box case presented in Section 4).

Gaussian Processes are simple probabilistic models that can be used for a variety of machine learning problems, such as regression and classification. We consider the regression case here. Let the data set consist of N pairs of inputs $\mathbf{x}_i$ and noisy real outputs y_i, related by

$$y_i = f(\mathbf{x}_i) + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma_{\text{n}}^2)$$

where $f(\mathbf{x})$ is an unknown function. The noise is additive, independent Gaussian noise.

When applying a Gaussian Process model, we assume a Gaussian Process prior over the functions $f(\mathbf{x})$. This can be summarised by

$$p(\mathbf{f}|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) \sim \mathcal{N}(\mathbf{0}, \mathbf{K_{ff}}) \tag{55}$$

with $\mathbf{f} = [f_1, f_2, ..., f_N]^T$, where $f_i = f(\mathbf{x}_i)$. $\mathbf{K_{ff}}$ is the covariance matrix, determined by the covariance function $\mathbf{K_{ff}}_{,ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

We can therefore write the objective function for the full Gaussian Process regression model,

$$\begin{aligned} p_{true}(\mathbf{y}|\theta) &\sim \mathcal{N}(\mathbf{0}, \mathbf{K_{ff}} + \sigma_{\text{n}}^2\mathbf{I}) \\ \therefore p_{true}(\mathbf{y}|\theta) &= -\left[\frac{N}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{K_{ff}} + \sigma_{\text{n}}^2\mathbf{I}| + \frac{1}{2}\mathbf{y}^T\left(\mathbf{K_{ff}} + \sigma_{\text{n}}^2\mathbf{I}\right)^{-1}\mathbf{y}\right]. \end{aligned} \tag{56}$$

The vector $\mathbf{y} = [y_1, y_2, ..., y_N]^T$ is the output vector, and θ incorporates the hyperparameters.

However, applying the full Gaussian Process regression model can be computationally expensive, especially with a large number of training cases (the complexity is $\mathcal{O}(n^3)$). Sparse approximations attempt to reduce this computation complexity. FITC achieves this by introducing a set of M inducing variables, $\mathbf{u} = [u_1, u_2, ..., u_M]^T$. These latent variables are the Gaussian Process values at location inputs $\mathbf{X_u}$ (the matrix incorporating all $\mathbf{x}_{u,i}$), the inducing inputs. This gives Equation 57. FITC assumes that the f_i are conditionally independent given $\mathbf{u}$, giving the objective function in Equation 58. The derivation of this

objective function can be followed in [1], and we have used the same notation, using $\mathbf{Q_{ff}}$ and $\mathbf{G}$.

$$p(\mathbf{u}|\mathbf{X_u}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K_{uu}}). \tag{57}$$

$$p_{\mathrm{FITC}}(\mathbf{y}|\theta) = -\left[\frac{N}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{Q_{ff}} + \mathbf{G}| + \frac{1}{2}\mathbf{y}^T(\mathbf{Q_{ff}} + \mathbf{G})^{-1}\mathbf{y}\right], \tag{58}$$

$$\text{with } \mathbf{Q_{ff}} = \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}},\ \mathbf{G} = \mathrm{diag}(\mathbf{K_{ff}} - \mathbf{Q_{ff}}) + \sigma_{\mathrm{n}}^2\mathbf{I}.$$

FITC attempts to maximise this objective function by changing the inducing inputs' locations, $\mathbf{X_u}$. It can be shown that $p_{FITC}(\mathbf{y}|\theta) = Z_{\mathrm{EP}}$ (and $p_{true}(\mathbf{y}|\theta) = Z_{\mathrm{true}}$), where the EP approximation [2] is

$$\frac{1}{Z_{\mathrm{true}}}\left(p_0(f)\prod_{n=1}^{N}p(y_n|\mathbf{f})\right) \approx \frac{1}{Z_{\mathrm{EP}}}\left(p_0(f)\prod_{n=1}^{N}\tilde{t}_n(\mathbf{u})\right).$$

7.1 FITC toy case

We consider a simple $N = 2$ case, with 1-dimensional input and outputs, approximated by $M = 2$ inducing inputs, with a squared exponential covariance function as in Equation 59. This means that $\mathbf{x} = [x_1, x_2]^T \in \mathbb{R}^2, \mathbf{y} \in \mathbb{R}^2, \mathbf{x}_{\mathrm{u}} = [x_{\mathrm{u},1}, x_{\mathrm{u},2}]^T \in \mathbb{R}^2$. The 2×2 matrices $\mathbf{K_{ff}}, \mathbf{K_{uu}}$ and $\mathbf{K_{fu}} = \mathbf{K_{uf}^T}$ follow, as do $\mathbf{Q_{ff}}$ and $\mathbf{G}$. These matrices depend on the hyperparameters $\theta = \{\sigma_{\mathrm{n}}, \sigma_{\mathrm{f}}, \ell\}$.

$$k(x_i, x_j) = \sigma_f^2\, e^{-\frac{(x_i - x_j)^2}{2\ell^2}}. \tag{59}$$

$$\mathbf{K_{ff}} = \begin{bmatrix} \sigma_f^2 & \sigma_f^2\, e^{-\frac{(x_1-x_2)^2}{2\ell^2}} \\ \sigma_f^2\, e^{-\frac{(x_2-x_1)^2}{2\ell^2}} & \sigma_f^2 \end{bmatrix}.$$

$$\mathbf{K_{uu}} = \begin{bmatrix} \sigma_f^2 & \sigma_f^2\, e^{-\frac{(x_{u,1}-x_{u,2})^2}{2\ell^2}} \\ \sigma_f^2\, e^{-\frac{(x_{u,2}-x_{u,1})^2}{2\ell^2}} & \sigma_f^2 \end{bmatrix}.$$

$$\mathbf{K_{fu}} = \begin{bmatrix} \sigma_f^2 & \sigma_f^2\, e^{-\frac{(x_1-x_{u,2})^2}{2\ell^2}} \\ \sigma_f^2\, e^{-\frac{(x_2-x_{u,1})^2}{2\ell^2}} & \sigma_f^2 \end{bmatrix} = \mathbf{K_{uf}^T}.$$

Because there are as many inducing inputs as true inputs, the approximation is exact at $\mathbf{x}_{\mathrm{u}} = \mathbf{x}$, while other values of $\mathbf{x}_{\mathrm{u}}$ lead to inexact approximations. In fact, we find that the FITC approximation (where $p_{FITC}(\mathbf{y}|\theta)$ is maximised with respect to $\mathbf{x}_{\mathrm{u}}$) may not be at the

exact solution. This was also found in [1], which states that the exact solution is a saddle point for FITC, not a global maximum. The consequences of this toy case's inexactness on the FITC algorithm are discussed in Section 7.2; here, we consider one case to show that $p_{FITC}(\mathbf{y}|\theta) = Z_{\text{EP}} > Z_{\text{true}} = p_{true}(\mathbf{y}|\theta)$, using it as a counter-example to the previously mentioned conjecture. This is achieved by choosing $\{\theta, \mathbf{x}, \mathbf{y}\}$ appropriately.

Consider the case where $\mathbf{y} = [0, 0]^T$, $\mathbf{x} = [1, 0]^T$, $\sigma_{\text{n}} = 0.1$, and $\sigma_{\text{f}} = \ell = 0.5$. Figure 16 shows $[\ln(p_{FITC}(\mathbf{y}|\theta)) - \ln(p_{true}(\mathbf{y}|\theta))]$ plotted as a function of $x_{\text{u},1}$ and $x_{\text{u},2}$. Note that $x_{\text{u},1}$ and $x_{\text{u},2}$ are interchangeable, and the plot is therefore symmetrical about $x_{\text{u},1} = x_{\text{u},2}$. As seen, at $\mathbf{x}_{\text{u}} = \mathbf{x} = [1, 0]^T$, marked by a red circle, the approximation is exact, and $p_{FITC}(\mathbf{y}|\theta) = p_{true}(\mathbf{y}|\theta)$. However, $p_{FITC}(\mathbf{y}|\theta) > p_{true}(\mathbf{y}|\theta)$ at certain values of $\mathbf{x}_{\text{u}}$. Plugging in $\mathbf{x}_{\text{u}} = [0.5, 0.5]^T$ or $\mathbf{x}_u = [0.5, \infty]^T$ into Equations 56 and 58 confirms this relationship.

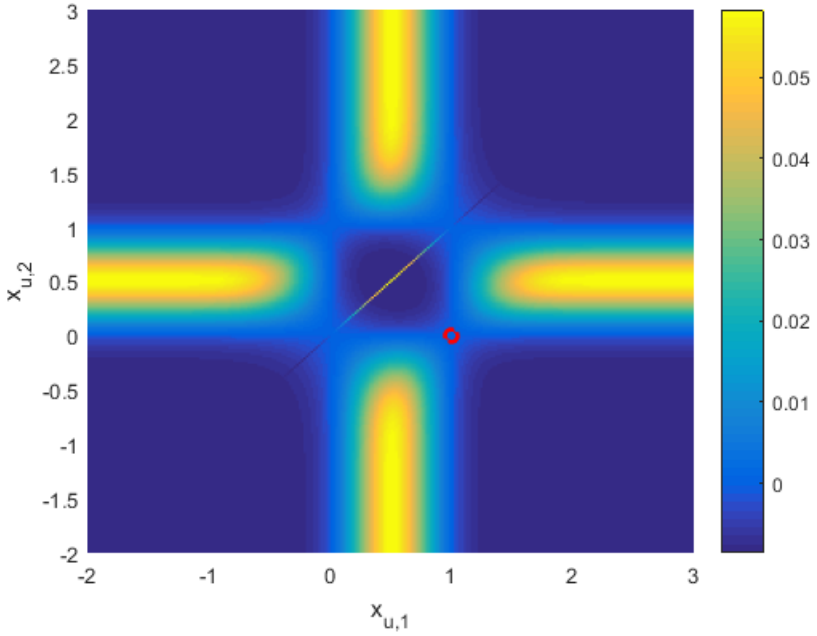


Figure 16: Difference in FITC and true log likelihoods

We have therefore found an example where $p_{FITC}(\mathbf{y}|\theta) > p_{true}(\mathbf{y}|\theta)$, and so $Z_{\text{EP}} > Z_{\text{true}}$. Section 7.2 discusses the consequences of this simple example on the FITC algorithm.

7.2 Understanding FITC

In this section, we consider the simple example introduced in Section 7.1, and attempt to explain some of FITC's properties using the example. Although the number of inducing inputs used for FITC is equal to the true number of inputs, as $M = N = 2$, it was shown that FITC prefers solutions away from the exact answer. In fact, as previously mentioned, the solution with $x_{\mathrm{u},1} = 0.5$ (the average of the 'true' inputs $\mathbf{x}$) and $x_{\mathrm{u},2}$ tending to ∞ leads to $p_{FITC}(\mathbf{y}|\theta) > p_{true}(\mathbf{y}|\theta)$, and is preferred as a solution to the FITC algorithm. In effect, this solution uses only one inducing input ($x_{\mathrm{u},1}$), and ignores the second inducing input.

Figure 17 plots the mean and 95% predictive error bars of both the true and approximating Gaussian Processes. The variance of the FITC model and the true Gaussian Process underlying the data do not appear to match well between $x = -1$ and $x = 2$, suggesting that the FITC approximation is locally poor.

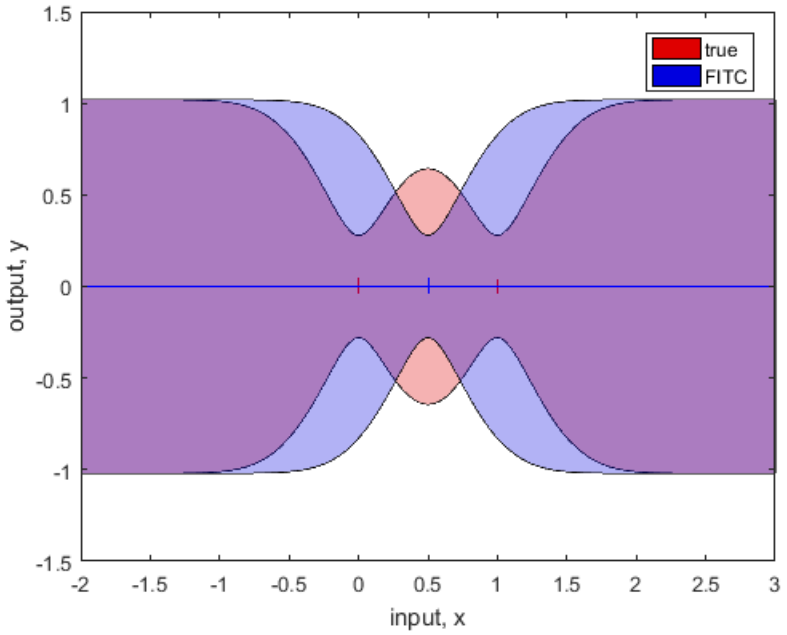


Figure 17: Mean and 95% predictive error bars of true and FITC Gaussian Processes

In order to understand why FITC prefers this approximation, we return to Equations 56 and 58. We can split the likelihood term into the 2π term (which we can ignore as it is cancelled out when considering $p_{\mathrm{FITC}}(\mathbf{y}|\theta) - p_{\mathrm{true}}(\mathbf{y}|\theta)$), a determinant term of the form $\frac{1}{2}\log|\mathbf{\Sigma}|$, and a quadratic term of the form $\frac{1}{2}\mathbf{y}^T(\mathbf{\Sigma})^{-1}\mathbf{y}$. In the example considered, $y = [0, 0]^T$, and so the quadratic term is zero, and only the determinant term affects $p_{\mathrm{FITC}}(\mathbf{y}|\theta)$. Maximising

$p_{\text{FITC}}(\mathbf{y}|\theta)$ therefore means minimising the determinant term. Because we are dealing with 2×2 matrices, and FITC enforces exact diagonal terms in the matrix ($\text{diag}\left(\mathbf{\Sigma}_{\text{FITC}}\right) = \text{diag}\left(\mathbf{\Sigma}_{\text{true}}\right)$), FITC therefore maximises the off-diagonal terms in the matrix $\mathbf{\Sigma}_{\text{FITC}} = \mathbf{Q_{ff}} + \mathbf{G}$. This directly leads to the behaviour observed. We can say that the determinant term will always bias $p_{\text{FITC}}(\mathbf{y}|\theta)$ to prefer this solution (in the 2 input, 2 inducing input case).

Let us now consider the quadratic term in the likelihood functions, with a non-zero output $\mathbf{y} = [y_1, y_2]^T$. Let

$$\mathbf{\Sigma}_{\text{FITC}} = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

where a is fixed (the diagonal terms are exact), and FITC changes b to minimise the quadratic term

$$\frac{1}{2}\mathbf{y}^T \left(\mathbf{\Sigma}_{\text{FITC}}\right)^{-1} \mathbf{y} = \frac{ay_1^2 + ay_2^2 - 2by_1y_2}{2(a^2 - b^2)}. \tag{60}$$

Let us first consider the case where $y_1 = y_2$. FITC therefore minimises

$$y_1^2 \left(\frac{a-b}{a^2-b^2}\right) = y_1^2 \left(\frac{1}{a+b}\right)$$

leading to maximising b. This is the same effect as observed by the determinant term, and increases the overestimation of Z_{EP}. The difference in quadratic terms in $p_{\text{FITC}}(\mathbf{y}|\theta)$ and $p_{\text{true}}(\mathbf{y}|\theta)$ has been plotted in Figure 18, for $y = [0.2, 0.2]^T$.

We now consider the case where $y_1 = -y_2$. FITC therefore minimises (referring to Equation 60)

$$y_1^2 \left(\frac{a+b}{a^2-b^2}\right) = y_1^2 \left(\frac{1}{a-b}\right)$$

leading to minimising b. This leads to the opposite effect to that observed so far. The difference in quadratic terms in $p_{\text{FITC}}(\mathbf{y}|\theta)$ and $p_{\text{true}}(\mathbf{y}|\theta)$ has been plotted in Figure 19, for $y = [0.2, -0.2]^T$. At these small absolute values of y_1 and y_2, we can see that the quadratic term has a smaller effect on the likelihood terms' difference than the determinant term (which can be seen in Figure 16). However, the quadratic term's effect increases proportional to the square of these absolute values, quickly leading to a larger effect.

All of the tests so far have been performed with $\sigma_{\text{f}} = 0.5, \sigma_{\text{n}} = 0.1, \ell = 0.5$. Changing these hyperparameters results predictable changes to plots of $p_{\text{FITC}}(\mathbf{y}|\theta) - p_{\text{true}}(\mathbf{y}|\theta)$. Increasing the characteristic length scale ℓ leads to a more smoothed out plot, with a larger overestimation, while reducing ℓ to small values (relative to $|x_2 - x_1|$) leads to $p_{\text{FITC}}(\mathbf{y}|\theta) = p_{\text{true}}(\mathbf{y}|\theta)$ at all $\mathbf{x}_{\text{u}}$, as the length scale is now small enough for the model to confidently incorporate any $\mathbf{x}$.

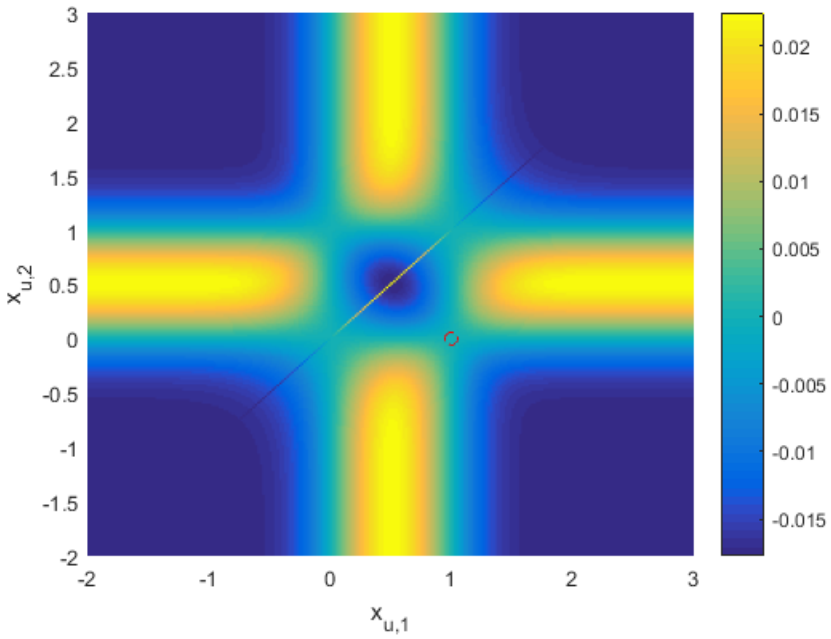


Figure 18: Difference in (log) likelihood quadratic terms for $\mathbf{y} = [0.2, 0.2]^T$

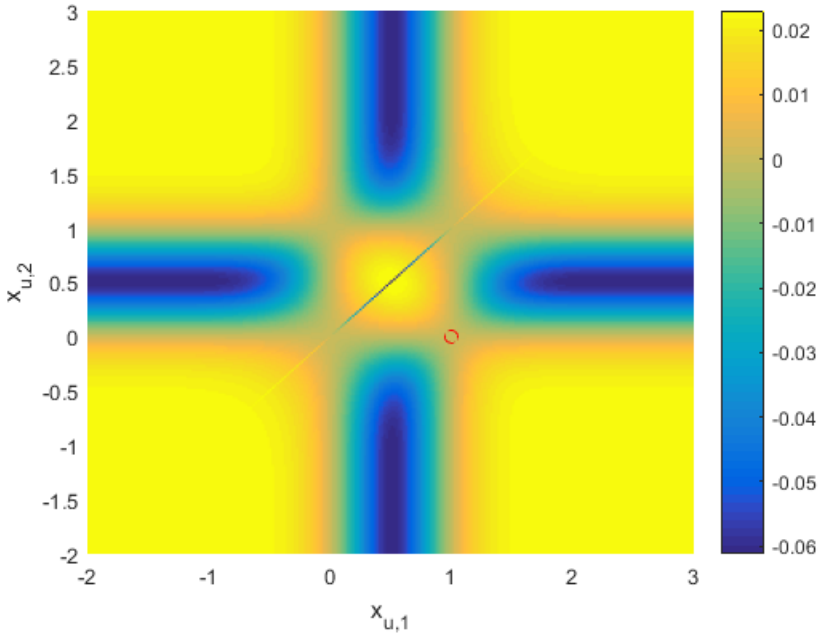


Figure 19: Difference in (log) likelihood quadratic terms for $\mathbf{y} = [0.2, -0.2]^T$

Changing the noise variance, σ_{n}, or the signal variance, σ_{f}, does not affect the shape of the plot. Increasing σ_{n} (or reducing σ_{f}) causes a smaller overestimation (allowing larger noise variance in the model means that the model is more confident about any inputs); conversely,

reducing σ_{n} (or increasing σ_{f}) causes a larger overestimation.

We have seen that, in a 2 input, 2 inducing input case, when the output $\mathbf{y} = \mathbf{0}$, FITC prefers keeping one inducing input and letting the other tend to ∞, a solution that is different from the exact solution. With non-zero $\mathbf{y}$, we get results that depend on the values of y_1 and y_2. It would be interesting to see how this toy case behaviour affects the FITC algorithm when run on bigger, more realistic datasets. We could begin by expanding this toy case example to 3 inputs, 2 inducing inputs, and seeing if FITC still prefers to ignore one inducing input. The effect of FITC ignoring inducing inputs has been noted in literature [1], and studying similar toy examples could provide further mathematical explanation for this effect.

8 Conclusions

We have looked in detail at the conjecture that $Z_{\text{EP}} < Z_{\text{true}}$. The symmetric box case is a counter-example to this conjecture, and we have presented empirical evidence and a mathematical proof to show $Z_{\text{EP}} > Z_{\text{true}}$. We also used the relationship between FITC and EP to find another counter-example to the conjecture. We can therefore conclude that the conjecture does not hold in the general case for EP. However, we also considered a binary classification example, illustrating why it is likely that the conjecture holds true on large, realistic datasets. Looking at the repeated Heaviside functions example showed that EP increasingly underestimates the model evidence when there are repeated functions, as expected from considering PEP and the α-Divergence, a conclusion that holds in the presence of jitter (noise). Softening the Heaviside functions to probit functions on all cases led to similar results. We also discussed the implications of the toy case FITC example on the FITC algorithm, showing mathematically why FITC can prefer to ignore an inducing input. This provides an interesting insight into the FITC algorithm, as one would hope for FITC's inducing inputs to perfectly match the true inputs.

Future work would involve considering other parts of the toy cases in more detail, as well as expanding them further. For example, one could consider EP's approximation of variance, studying whether EP underestimates or overestimates the true variance in the symmetric box case example and the repeated Heaviside example. We could also consider approximating families other than the Gaussian approximating family, and if the results observed in this report hold when we consider more moments. Future work could also expand the FITC example from the 2 input, 2 inducing input case to see how the observed results generalise to bigger datasets. It would be interesting to see if FITC locally prefers to approximate 2 input points as 1 inducing point on larger datasets, and if this behaviour can be explained mathematically as well as shown empirically.

References

[1] Matthias Stephan Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding Probabilistic Sparse Gaussian Process Approximations. *arXiv:1606.04820 [stat]*, June 2016. arXiv: 1606.04820.

[2] Thang D. Bui, Josiah Yan, and Richard E. Turner. A Unifying Framework for Sparse Gaussian Process Approximation using Power Expectation Propagation. *arXiv:1605.07066 [cs, stat]*, May 2016. arXiv: 1605.07066.

[3] John P. Cunningham. Expectation Propagation: Factorization and Entropy Approximation, May 2015.

[4] John P. Cunningham, Philipp Hennig, and Simon Lacoste-Julien. Gaussian Probabilities and Expectation Propagation. *arXiv:1111.6832 [stat]*, November 2011. arXiv: 1111.6832.

[5] Guillaume Dehaene. *Le statisticien neuronal*. PhD thesis, Paris Descartes University, September 2016.

[6] José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang D. Bui, and Richard E. Turner. Black-box α-divergence minimization. 2016.

[7] Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of machine learning research*, 6(Oct):1679–1704, 2005.

[8] Yingzhen Li and Richard E. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

[9] Thomas Minka. Power ep. Technical report, Technical report, Microsoft Research, Cambridge, 2004.

[10] Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[11] Tom Minka. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

[12] Ulrich Paquet, Adrian Weller, Ole Winther, and Nicholas Ruozzi. Towards (?) marginal likelihood lower bounds with EP.

[13] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

[14] Adrian Weller and Tony Jebara. Clamping variables and approximate inference. In *Advances in Neural Information Processing Systems*, pages 909–917, 2014.

[15] Alan S. Willsky, Erik B. Sudderth, and Martin J. Wainwright. Loop series and Bethe variational bounds in attractive graphical models. In *Advances in neural information processing systems*, pages 1425–1432, 2008.

Appendix A Risk Assessment retrospective

This mathematical and computer-based project did not cause or provide any hazards throughout the year. The risk assessment provided at the beginning of the project was accurate.