Understanding Expectation Propagation

Siddharth Swaroop
Department of Engineering
University of Cambridge
Cambridge, UK
ss2163@cam.ac.uk

Richard E. Turner
Department of Engineering
University of Cambridge
Cambridge, UK
ret26@cam.ac.uk

Abstract

Understanding and characterising the properties of approximate inference schemes is extremely important, but arguably under studied. This report continues work on characterising Expectation Propagation (EP), an approximate Bayesian inference scheme, looking at four toy cases of interest. We initially focus on the empirically motivated conjecture stating that EP's approximation for the model evidence is an underestimate of the true model evidence. The first two toy cases apply EP to a simple classification example. They indicate why EP tends to underestimate the model evidence on realistic datasets, even though there are counter-examples to the conjecture, which we show analytically for the first time. The third toy case uses the link between the Fully Independent Training Condition algorithm (FITC, a sparse approximation method for Gaussian Process regression) and EP to find another analytic counter-example. This toy case also raises interesting questions as to how and why FITC works, which we consider mathematically. The final toy example compares mean field EP to mean field and structured Variational Inference (VI) on a small time-series model. We find that EP's uncertainty estimates do not collapse pathologically as they do for mean field VI.

1 Introduction and Related Work

Expectation Propagation (EP) is a popular approximate Bayesian inference algorithm [11]. Its use cases span industry (such as for the TrueSkill model [8], or in infer.NET [14]), statistics [7], deep learning [10] and physics [15]. Considering the case where the true distribution $p(\mathbf{x})$ can be factorised into a product of factors $t_n(\mathbf{x})$, EP approximates the true distribution $p(\mathbf{x}) = \frac{1}{Z_{\text{true}}}\left(\prod_{n=0}^{N} t_n(\mathbf{x})\right)$ using a tractable distribution $q(\mathbf{x}) = \frac{1}{Z_{\text{EP}}}\left(\prod_{n=0}^{N} \tilde{t}_n(\mathbf{x})\right)$ by iteratively refining $\tilde{t}_n(\mathbf{x})$ according to an unnormalised $\mathcal{KL}$ divergence minimisation [13],

$$\underset{\tilde{t}_n^{\text{new}}(\mathbf{x})}{\operatorname{argmin}} \mathcal{KL}\left(\frac{q(\mathbf{x})}{\tilde{t}_n(\mathbf{x})} t_n(\mathbf{x}) \,\middle\|\, \frac{q(\mathbf{x})}{\tilde{t}_n(\mathbf{x})} \tilde{t}_n^{\text{new}}(\mathbf{x})\right). \tag{1}$$

The iterative nature of the EP update makes it difficult to theoretically analyse the final solution it returns. There has been work on characterising the asymptotic properties of EP [6], but relatively little on the finite sample properties. In this report, we first consider an empirically motivated conjecture in the machine learning community, stating that EP's approximation for the model evidence Z_{EP} is an underestimate of the true model evidence Z_{true} [9, 5]. It has been difficult to prove this conjecture [16], even though mathematical bounds exist in specific cases for belief propagation [19, 20], and empirical counter-examples have potentially been identified [4]. We then apply EP to a two dimensional two time-step time series model, using a fully factorised (mean field) approximation. Work in this field has previously applied mean field approximations with Variational Inference (VI) [18], and we compare VI's approximation to EP's.

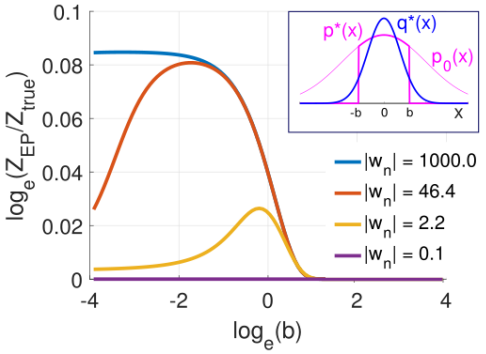
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 A Gaussian in a (soft) symmetric box

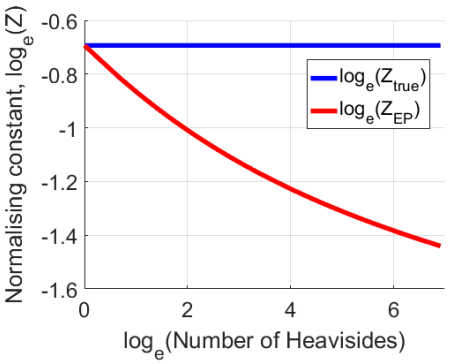
We apply EP on a toy case, in which a Gaussian prior $p_0(x)$ is multiplied by a pair of probit functions, $t_n(x) = \text{probit}(w_n^{\mathrm{T}} x + b_n) = 0.5 + 0.5\,\text{erf}\,(w_n^{\mathrm{T}} x + b_n)$, thereby restricting the Gaussian to a (soft) symmetric box (see Equation 2). This work is similar to that started in Cunningham et al. [5], but differs in being simpler, allowing an analytic counter-example to the conjecture.

$$p(x) = \frac{1}{Z_{\text{true}}} p_0(x)\text{probit}(w_n x + b)\text{probit}(-w_n x + b) \approx \frac{1}{Z_{\text{EP}}} p_0(x)\tilde{t}_1(x)\tilde{t}_2(x) = q(x). \tag{2}$$

The approximating factors $\tilde{t}_n(x)$ are Gaussian, and solving Equation 1 in this example provides analytic solutions for $\tilde{t}_n(x)$. We have plotted the difference in normalising constant in Figure 1a, as the cut-off point b and softness of the probit function w_n change. There is an overestimation of the normalising constant, of less than 10%. We can increase the dimensionality, d, of this example by having one pair of probit functions in each (uncorrelated) dimension. This causes the overestimation factor to be a factor of d larger in log-space. We have also mathematically shown that $Z_{\text{EP}} > Z_{\text{true}}$ in the 1-dimensional Heaviside function case [17], by first finding the EP fixed points and then finding expressions for the normalising constants.



(a) Overestimation of true normalising constant, with 1D probit functions (symmetric box case). $p^*(x)$ and $q^*(x)$ are unnormalised $p(x)$ and $q(x)$.



(b) Underestimation of true normalising constant as N increases (repeated Heaviside functions).

Figure 1: Comparing EP's overestimation and underestimation of the true model evidence.

3 Repeated Heaviside functions

The second toy case applies EP to a Gaussian prior multiplied by repeated Heaviside functions (Equation 3). This is equivalent to an application of Power EP (PEP) [12]: repeating the function N times does not change $p(x)$ ($N \geq 1$), and minimises the α-divergence (instead of the $\mathcal{KL}$ divergence) with $\alpha = \frac{2}{N} - 1$. Increasing N reduces α from $\alpha = 1$ (as in EP) towards $\alpha = -1$ (the Variational Inference case). As expected from properties of PEP [13], this approximation increasingly underestimates the normalising constant as N increases, seen in Figure 1b. We note that this underestimation is greater than the 10% overestimation observed in Figure 1a, although the underestimation magnitude is decreased slightly when probit functions are used (instead of Heaviside functions), and also when slight jitter (noise) is introduced in the cut-off value between functions (we have only considered $b = 0$ for all functions so far).

$$p(x) = \frac{1}{Z_{\text{true}}} \left(p_0(x) \prod_{n=1}^{N} h(-x + 0) \right) \approx \frac{1}{Z_{\text{EP}}} \left(p_0(x) \prod_{n=1}^{N} \tilde{t}_n(x) \right) = q(x). \tag{3}$$

Combining the two toy cases considered in a simple 1-dimensional binary classification example illustrates why $Z_{\text{EP}} \leq Z_{\text{true}}$ is often observed on realistic datasets. Real world data will tend to be fairly well separated and the likelihood functions in such cases will all have a similar form, meaning the repeated functions case is more prevalent than the symmetric box case. This would cause the

repeated functions case's underestimation to overpower any overestimation that may otherwise be observed. It is interesting to note that the two effects can cancel each other out slightly, leading to a value of Z_{EP} that is close to Z_{true}. In Swaroop [17] we show that this underestimation does indeed exist for small datasets, and the conjecture is that this expands to larger datasets.

4 FITC

The 'Fully Independent Training Condition' (FITC) algorithm is a sparse approximation for Gaussian Process (GP) regression [3], and it has been shown that the solution FITC reaches is the same as that of EP [2]. This is a nice application of EP as all factors are Gaussian in FITC, making it amenable to analysis. In fact, it is one of the only cases that the authors know where EP is approximate (it does not return exact inference) but still converges (in a single pass) to an analytic fixed point given by FITC.

The toy case considered here is another counter-example to $Z_{\mathrm{EP}} \leq Z_{\mathrm{true}}$. FITC approximates GP regression by using a set of M inducing points to approximate the true dataset, obtained by maximising the FITC likelihood (EP energy) with respect to the inducing points' locations. It can be shown that $p_{\mathrm{FITC}}(\mathbf{y}|\theta) = Z_{\mathrm{EP}}$, and $Z_{\mathrm{true}} = p_{\mathrm{true}}(\mathbf{y}|\theta)$ [2].

We consider the case with two training (1-dimensional) input and output pairs, $(x_1, y_1) = (1, 0)$ and $(x_2, y_2) = (0, 0)$, approximated by $M = 2$ inducing points, $(x_{u,1}, u_1)$ and (x_{u_2}, u_2), with a squared exponential covariance matrix. Figure 2 plots the difference in FITC and true likelihoods as we change the input values of the inducing points. The FITC solution is at the maximum of this plot, and we can see that this is not at the true training input locations (marked in red). Having $(x_{u,1}, x_{u,2}) = (0.5, 0.5)$ is a maximum, as is having one of the inducing input locations ignored (the inducing input has large magnitude) with the other at 0.5. These FITC solutions therefore satisfy $p_{\mathrm{FITC}}(\mathbf{y}|\theta) = Z_{\mathrm{EP}} > Z_{\mathrm{true}} = p_{\mathrm{true}}(\mathbf{y}|\theta)$.

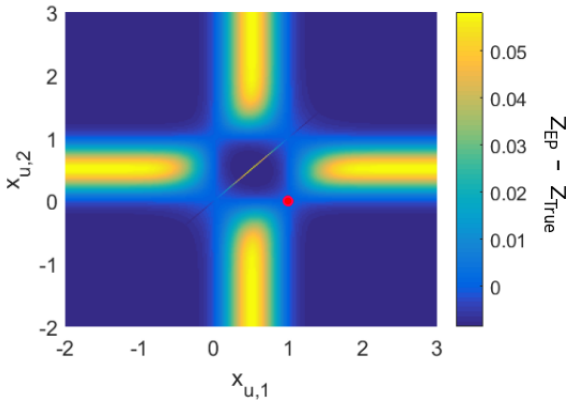


Figure 2: Difference in FITC and true likelihoods (for details, see main text).

This inexact approximation has been observed before on larger datasets [1, 2], and we can use this example to attempt to mathematically characterise it. We compare the behaviour of the terms in the likelihood function, noting their contributions as $\mathbf{x}$ and $\mathbf{y}$ change. We find that a determinant term, of the form $\frac{1}{2}\log|\mathbf{\Sigma}|$, prefers an overestimation of the normalising constant, while a quadratic term in $\mathbf{y}$, of the form $\frac{1}{2}\mathbf{y}^T(\mathbf{\Sigma})^{-1}\mathbf{y}$, prefers an overestimation if $y_1 = y_2$, and an underestimation if $y_1 = -y_2$ [17]. If the overestimation terms overpower other terms, such as in our example ($\mathbf{y} = 0$), then FITC prefers to ignore an inducing input. Future work could involve considering more training data in order to see how this effect scales mathematically onto larger datasets.

5 Time series

The final example applies EP and VI on a simple time series setup, comparing their behaviour. We consider two time-steps of a model comprising two-dimensional latent variables at each time-step (four latent variables in total). The priors are Gaussian $p(x_{d1}) = \mathrm{N}\left(x_{d1}; 0, \frac{\sigma_x^2}{1-\lambda^2}\right)$, the time-step

correlation is $p(x_{d2}|x_{d1}) = \mathrm{N}(x_{d2}; \lambda x_{d1}, \sigma_x^2)$, and at every time-step we observe y_t, $p(y_t|x_{d1}, x_{d2}) = \mathrm{N}(y_t; w_1 x_{1t} + w_2 x_{2t}, \sigma_y^2)$. The approximating factors are also Gaussian, ensuring the entire problem is tractable: we can directly compare the approximate inference algorithms with ground truth.

Figure 3 plots the correlation between x_{11} and x_{21} as we increase the observation noise σ_y^2, in the exact case, the EP mean field (MF) approximation (all four variables assumed uncorrelated), the VI MF approximation, and the structured VI approximation (factored across chains, unfactored across time). We have kept the observation weights w_1 and w_2 at unity for this plot, with $\lambda = 0.8$. We see that when σ_y^2 is very small, there is a high (anti-)correlation between the two variables. Here, the VI approximations are very certain of their predictions (their uncertainty is at the conditional variance of the variables), and EP is much more uncertain (but note that it does not return the marginal uncertainty of the true posterior as some authors suggest). This is a well-known undesired feature of the VI approximation [18]. As the observation noise increases, we see that all the approximations get more uncertain, with EP always being the most uncertain. At high σ_y^2, the exact distribution tends to the prior, as do the EP and structured VI approximations. The VI MF approximation, however, gets more certain than the prior after observing essentially meaningless data; this seems like a catastrophic failure on its part. For intermediate values of σ_y^2, both the EP and structured VI approximations seem to provide good approximations to the correlated exact distribution: their uncertainty is between being too certain (the VI MF approximation) and too uncertain (the prior). The nature of the downstream task will determine which of the two approximations is better. However, unlike the structured VI approximation, the EP mean field factorisation is coarser, which can yield computational advantages. See Appendix A for a comparison of EP and VI when optimising hyperparameters. We find that EP's estimate of model evidence is slightly better than VI's estimate, and EP overestimates the optimal λ^*.

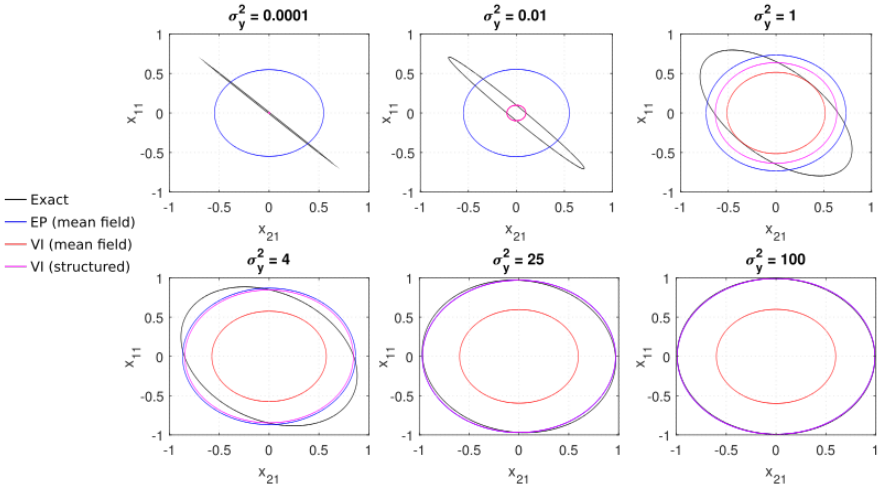


Figure 3: One standard deviation plots as σ_y^2 increases.

6 Conclusions

We initially looked in detail at the conjecture that $Z_{\mathrm{EP}} < Z_{\mathrm{true}}$, presenting the symmetric box case counter-example and arguing why this conjecture may be found on real classification datasets by also considering the repeated Heaviside functions case. The symmetric box case is the first analytic counter-example in the literature. We used the relationship between FITC and EP to find another counter-example to the conjecture, and also used this example to mathematically explain 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. In the final toy case, we compared EP and VI's approximations in a simple time series example, finding that the EP mean field approximation retains sensible uncertainty estimates in regimes where mean field VI fails catastrophically. We hope that this work will inspire further work applying approximate inference algorithms to small datasets in order to characterise and analyse them.

References

[1] M. S. Bauer, M. van der Wilk, and C. E. Rasmussen. Understanding Probabilistic Sparse Gaussian Process Approximations. *arXiv:1606.04820 [stat]*, June 2016. URL http://arxiv.org/abs/1606.04820. arXiv: 1606.04820.

[2] T. D. Bui, J. Yan, and R. E. Turner. A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation. *Journal of Machine Learning Research*, 18(104):1–72, 2017. URL http://jmlr.org/papers/v18/16-603.html.

[3] L. Csató and M. Opper. Sparse On-Line Gaussian Processes. *Neural Computation*, 14(3):641–668, Mar. 2002. ISSN 0899-7667. doi: 10.1162/089976602317250933. URL https://doi.org/10.1162/089976602317250933.

[4] J. P. Cunningham. Expectation Propagation: Factorization and Entropy Approximation, May 2015. URL http://gpss.cc/gpa15/assets/cunningham.pdf.

[5] J. P. Cunningham, P. Hennig, and S. Lacoste-Julien. Gaussian Probabilities and Expectation Propagation. *arXiv:1111.6832 [stat]*, Nov. 2011. URL http://arxiv.org/abs/1111.6832. arXiv: 1111.6832.

[6] G. P. Dehaene and S. Barthelmé. Bounding errors of Expectation-Propagation. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 244–252. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5912-bounding-errors-of-expectation-propagation.pdf.

[7] A. Gelman, A. Vehtari, P. Jylänki, T. Sivula, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. Robert. Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *arXiv:1412.4869 [stat]*, Dec. 2014. URL http://arxiv.org/abs/1412.4869. arXiv: 1412.4869.

[8] R. Herbrich, T. Minka, and T. Graepel. TrueSkill™ : A Bayesian Skill Rating System. In P. B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press, 2007. URL http://papers.nips.cc/paper/3079-trueskilltm-a-bayesian-skill-rating-system.pdf.

[9] M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of machine learning research*, 6(Oct):1679–1704, 2005. URL http://www.jmlr.org/papers/v6/kuss05a.html.

[10] Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems 28*, pages 2323–2331. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5760-stochastic-expectation-propagation.pdf.

[11] T. 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. URL http://dl.acm.org/citation.cfm?id=2074067.

[12] T. Minka. Power ep. Technical report, Technical report, Microsoft Research, Cambridge, 2004. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2004-149.pdf.

[13] T. Minka. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005. URL https://www.seas.harvard.edu/courses/cs281/papers/minka-divergence.pdf.

[14] T. Minka, J. Winn, J. Guiver, S. Webster, Y. Zaykov, B. Yangel, A. Spengler, and J. Bronskill. Infer.NET 2.6, 2014. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

[15] M. Opper, B. Çakmak, and O. Winther. A Theory of Solving TAP Equations for Ising Models with General Invariant Random Matrices. *Journal of Physics A: Mathematical and Theoretical*, 49(11): 114002, Mar. 2016. ISSN 1751-8113, 1751-8121. doi: 10.1088/1751-8113/49/11/114002. URL http://arxiv.org/abs/1509.01229. arXiv: 1509.01229.

[16] U. Paquet, A. Weller, O. Winther, and N. Ruozzi. Towards (?) marginal likelihood lower bounds with EP. URL http://gpss.cc/gpa15/assets/paquet.pdf.

[17] S. Swaroop. Understanding Expectation Propagation. Technical report, University of Cambridge, May 2017. URL https://github.com/siddharthswaroop/masters-project/blob/master/Swaroop%202017.pdf.

[18] R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. *Bayesian Time series models*, pages 115–138, 2011.

[19] A. Weller and T. Jebara. Clamping variables and approximate inference. In *Advances in Neural Information Processing Systems*, pages 909–917, 2014. URL http://papers.nips.cc/paper/5529-clamping-variables-and-approximate-inference.

[20] A. S. Willsky, E. B. Sudderth, and M. J. Wainwright. Loop series and Bethe variational bounds in attractive graphical models. In *Advances in neural information processing systems*, pages 1425–1432, 2008. URL http://papers.nips.cc/paper/3354-loop-series-and-bethe-variational-bounds-in-attractive-graphical-models.

Appendix A

This section compares EP mean field (MF) and VI MF approximations when used to optimise a hyperparameter of the time series model introduced in Section 5. Specifically, we consider optimising the value of λ when given the correct values of the other hyperparameters, $\{\sigma_x^2, \sigma_y^2, w_1, w_2\}$. To do this, we set a value for λ_{true}, and calculate the model evidence for EP MF and VI MF. We set $w_1 = w_2 = 1$, $\sigma_y^2 = 0.43$, $\sigma_x^2 = 1 - \lambda_{\text{true}}^2$, and use λ_{true} to calculate the sufficient statistics of y_1 and y_2.

Figure 4 shows EP MF, VI MF and exact model evidences as we sweep over different values of λ. The optimal λ^* is given by the maxima of the plots, and is marked with red crosses. As expected, the VI MF model evidence is always less than the true model evidence (VI provides a lower bound), and we also note that VI MF always underestimates λ_{true} [18]. The EP MF model evidence, however, is much closer to the exact model evidence at all values of λ_{true} and λ, while appearing to overestimate the exact model evidence, providing another (empirical) counter-example to the conjecture discussed in the paper. Additionally, the EP MF approximation for λ_{true} always seems to be an overestimate ($\lambda_{\text{EP}}^* > \lambda_{\text{true}}$), while also being closer to λ_{true} than the VI MF.

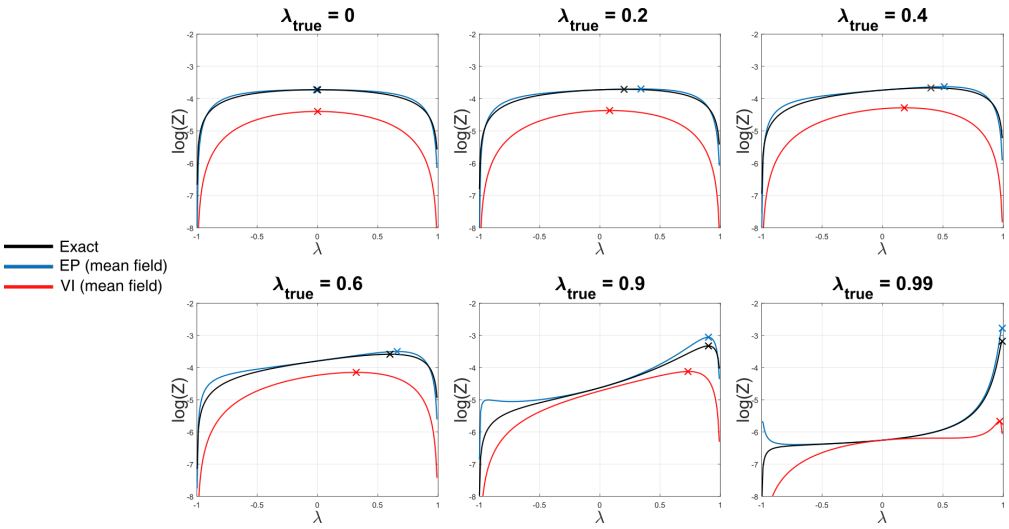


Figure 4: EP and VI model evidence plots for optimising λ.

These preliminary results indicate that the EP MF approximation would be better than VI MF when optimising hyperparameters. Future work would include comparing the EP MF approximations with the structured VI approximation, as in Section 5. We could also see if the EP MF approximation overestimates the model evidence and the optimal hyperparameter value for other hyperparameters in the model, and when we attempt to optimise more than one hyperparameter at a time. If an overestimation is observed, we could try applying mean field Power EP (with different values of α in the α-Divergence) instead of EP or VI to see if we obtain more accurate optimal hyperparameter values. We could also isolate the time correlation λ from the variance of the prior, a technique often used in practice, and repeat our tests. Specifically, the equation for the prior changes to $p(x_{d1}) = \mathrm{N}\left(x_{d1}; 0, \sigma_1^2\right)$, while the time-step correlation $p(x_{d2}|x_{d1})$ and observation $p(y_t|x_{d1}, x_{d2})$ are unchanged. It is not clear whether the results in this section will hold with the new time series model, or whether the results are specific to a particular parameterisation.