Understanding Expectation Propagation

Siddharth Swaroop and Richard E. Turner
University of Cambridge
{ss2163, ret26}@cam.ac.uk

Motivation and background

- Properties of approximate inference algorithms are important to understand

- Expectation Propagation (EP) approximation:

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

Iteratively refine $\tilde{t}_n(\mathbf{x})$:

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

- Empirically motivated conjecture [1, 2]:

$$Z_{\text{EP}} = \int \prod_{n=0}^{N} \tilde{t}_n(\mathbf{x})\, d\mathbf{x} < \int \prod_{n=0}^{N} t_n(\mathbf{x})\, d\mathbf{x} = Z_{\text{true}}$$

- We consider toy cases:
  - Show a counterexample to conjecture
  - Why conjecture may hold on real-world datasets
  - Compare EP and VI on time series example

(Soft) symmetric box

- Setup ($\text{probit}(w_n x + b_n) = 0.5 + 0.5\,\text{erf}(w_n x + b_n)$):

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

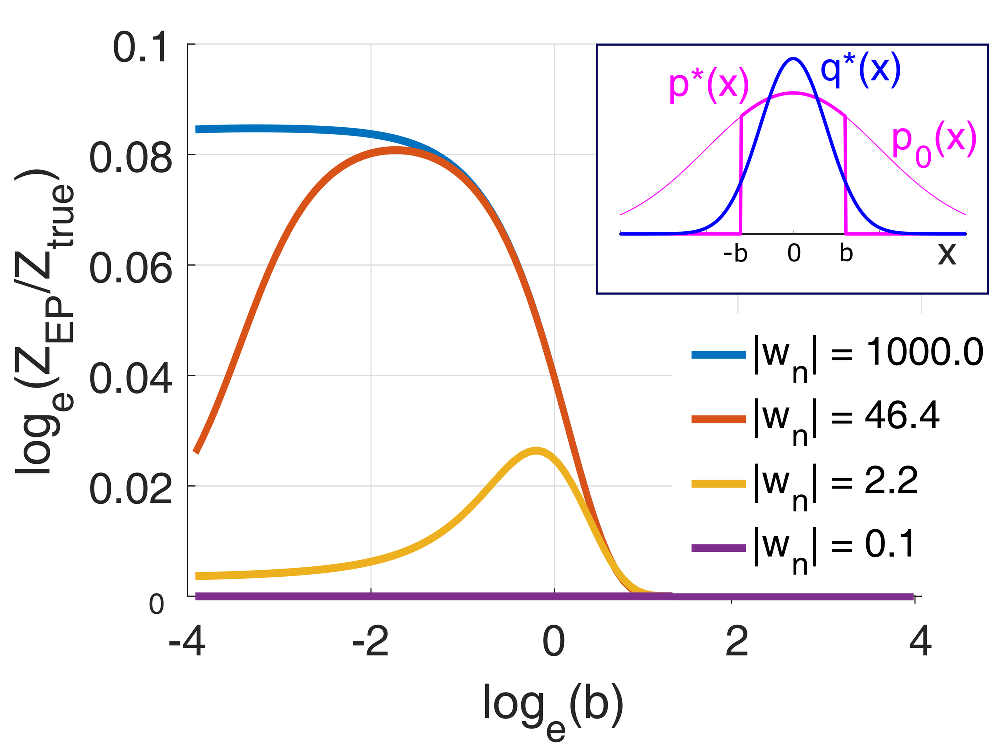

Figure 1: Overestimation of true normalising constant

- Have mathematically shown overestimation in 1D Heaviside function case

Repeated Heaviside functions

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

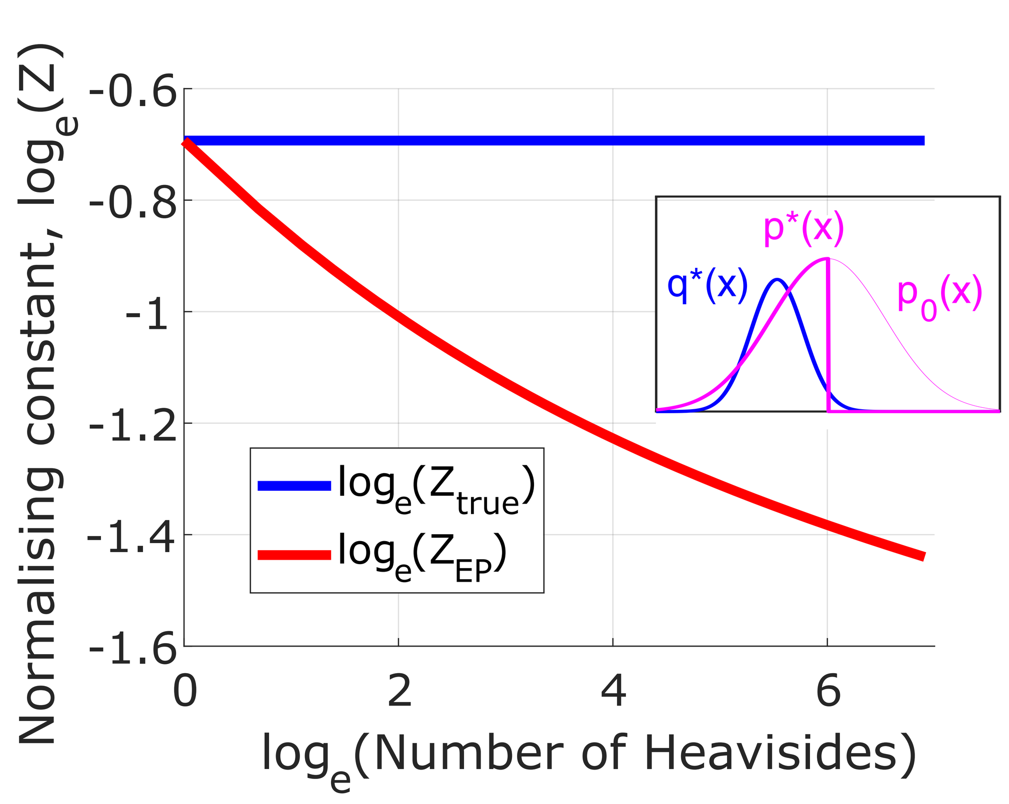

Figure 2: Underestimation of true normalising constant

[1] M. Kuss and C.E. Rasmussen. Assessing approximate inference for binary Gaussian process classification, JMLR Oct 2006
[2] John P. Cunningham, P. Hennig, and S. Lacoste-Julien. Gaussian Probabilities and Expectation Propagation, 2011
[3] R.E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models, Bayesian Time series models, 2011

Simple classification example

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

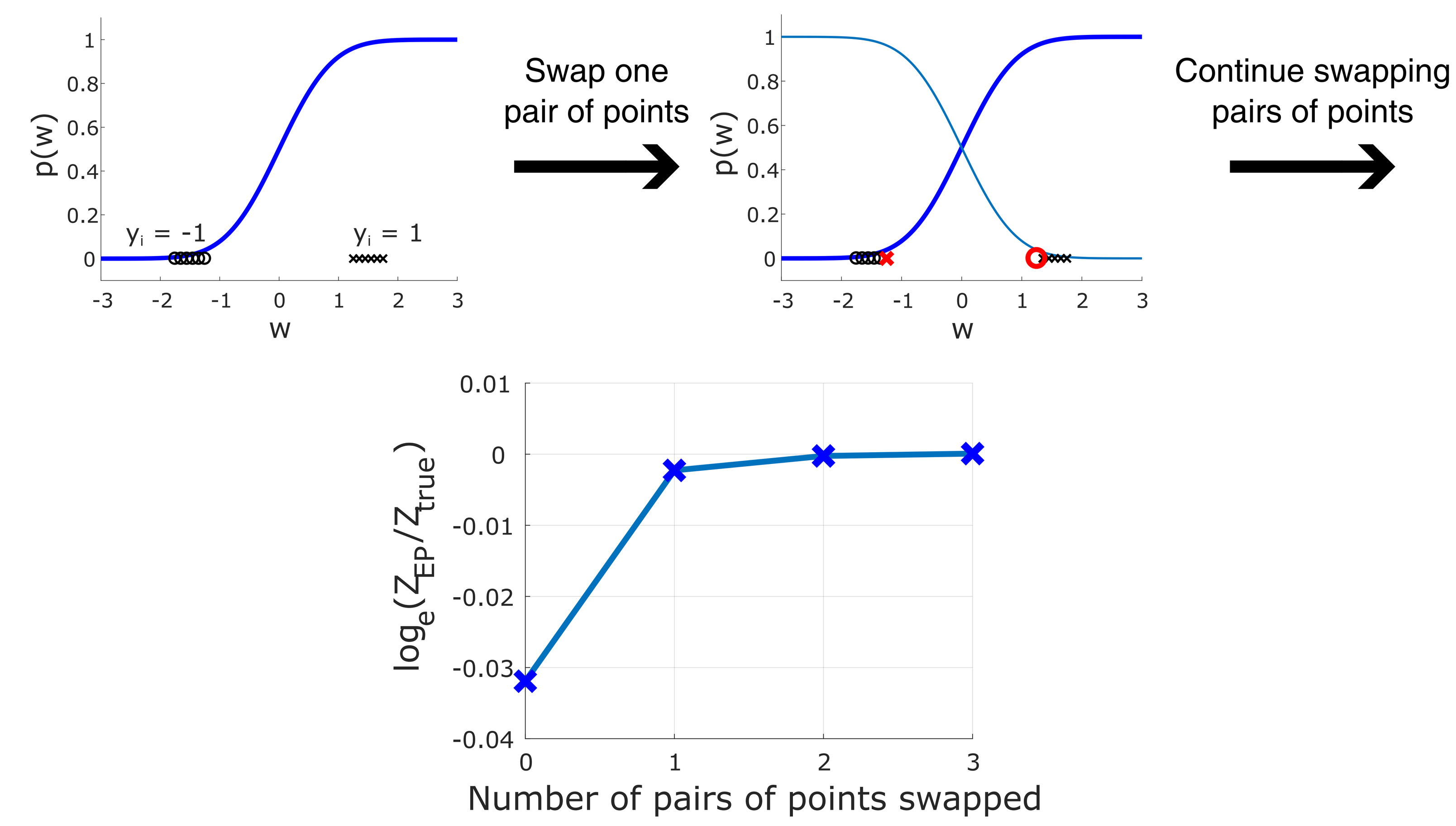

Figure 3: EP normalising constant approaches true value

- Realistic datasets tend to be fairly well-separated, and underestimation effect dominates

Time series

- 2 time-steps, 2-dimensional latent variables

$$p(x_{d1}) = \mathrm{N}\left(x_{d1}; 0, \frac{\sigma_x^2}{1-\lambda^2}\right)$$

$$p(x_{d2}|x_{d1}) = \mathrm{N}\left(x_{d2}; \lambda x_{d1}, \sigma_x^2\right)$$

$$p(y_t|x_{d1}, x_{d2}) = \mathrm{N}\left(y_t; w_1 x_{1t} + w_2 x_{2t}, \sigma_y^2\right)$$

Exact: $\boxed{\begin{matrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{matrix}}$

VI (structured): $\boxed{x_{11} \quad x_{12}}$ $\boxed{x_{21} \quad x_{22}}$

EP MF/VI MF: $\boxed{x_{11}}$ $\boxed{x_{12}}$ $\boxed{x_{21}}$ $\boxed{x_{22}}$

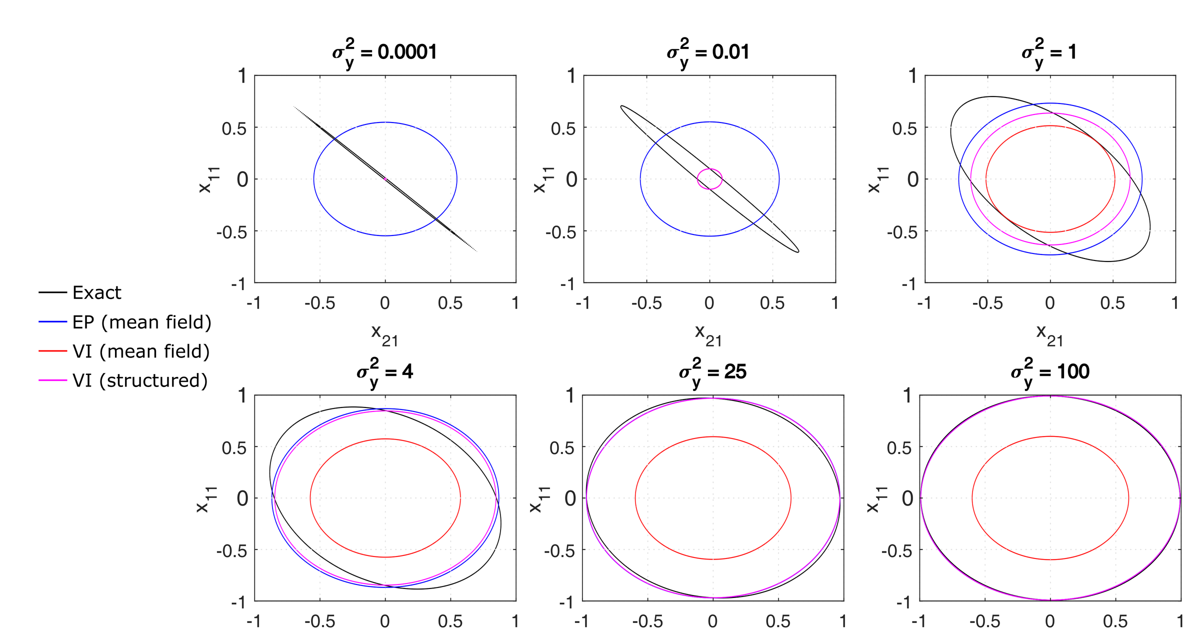

Figure 4: EP and VI [3] estimates of precision. VI (mean field) fails as observation noise increases.

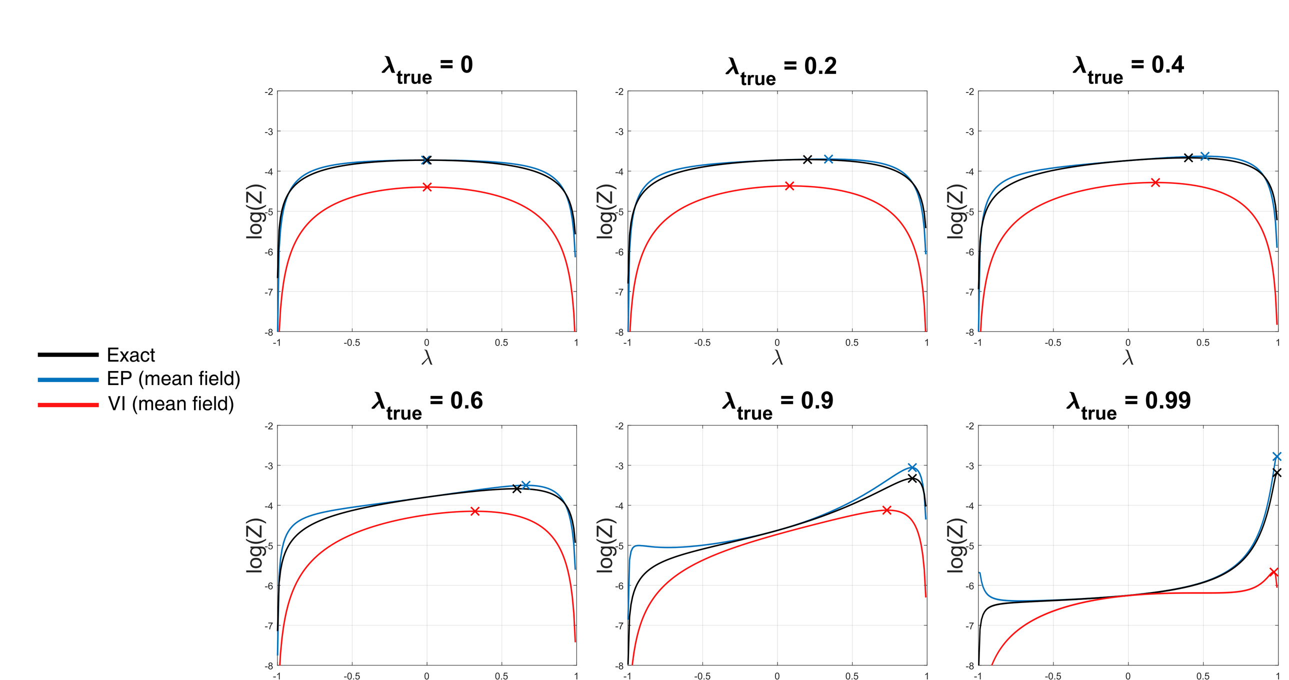

Figure 5: EP and VI plots of normalising constant, to estimate λ_{true}. EP tends to overestimate, while VI underestimates.

Further work

- See paper for details on Gaussian Process approximate inference results (FITC algorithm)

- Do model evidence results hold with other approximating families?

- Would Power EP provide better results in time series?