KL Corrected Variational Inference for Gaussian Processes

Nathaniel J. King and Neil D. Lawrence

10th June 2005

Overview

- Variational Inference in Gaussian Processes
- Modified Variational Inference
  - Probabilistic Point Assimilation (PPA) 'more tractable'
- KL Correction of the Variational Bound
- Results
- Speculation

Notation

- Labels $\mathbf{y} = [y_1 \dots y_N]^{\mathrm{T}}$.
- Input vector $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]^{\mathrm{T}}$
- Gaussian distribution over $\mathbf{y}$ is $N(\mathbf{y}|\boldsymbol{\mu}, \Sigma)$ with mean $\boldsymbol{\mu}$ and covariance Σ.
- Process variable (the function) $\mathbf{f} = [f_1 \dots f_N]^{\mathrm{T}}$ and $\bar{\mathbf{f}} = \left[\bar{f}_1 \dots \bar{f}_N\right]^{\mathrm{T}}$.
- The notation $\mathbf{f}_{\setminus n}$ represents the vector without the nth element.

Gaussian Process Graph

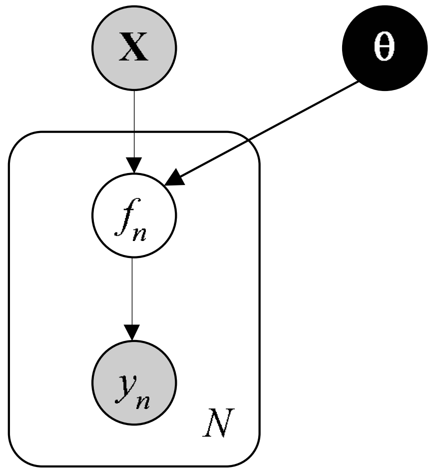


Figure 1: Graphical model of Gaussian process.

$$\log p(\mathbf{y}) = \log \int \prod_{n=1}^{N} p(y_n|f_n)\, p(\mathbf{f}|\mathbf{X}, \theta)\, df$$

$p(y_n|f_n)$ is a noise model $p(\mathbf{f}|\mathbf{X}, \theta) = N(\mathbf{f}|\mathbf{0}, \mathbf{K})$

K is a covariance function parameterised by θ

Variational Inference (Vanilla)

$$\log p(\mathbf{y}) \geq \left\langle \sum_{n=1}^{N} \log p(y_n|f_n)\, p(\mathbf{f}|\mathbf{X},\theta) \right\rangle_{q(\mathbf{f})} - \langle \log q(\mathbf{f}) \rangle_{q(\mathbf{f})}$$

$$q(\mathbf{f}) \propto \prod_{n=1}^{N} p(y_n|f_n)\, p(\mathbf{f}|\mathbf{X},\theta)$$

- Constrain $q(\mathbf{f})$ to be Gaussian — Seeger [2000].
- Constrain covariance of $q(\mathbf{f})$ to have a FA style structure.
- Method is slow and not easily adjusted to new noise models.

Augmented Model — PPA

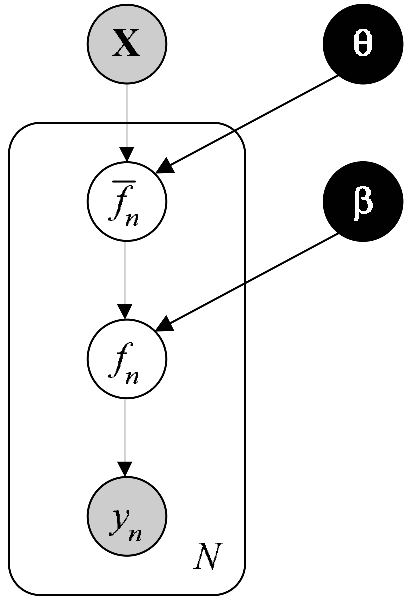


Figure 2: Graph of Augmented Model

$$\log p(\mathbf{y}) = \log \int \prod_{n=1}^{N} \int p(y_n|f_n)\, p\left(f_n|\bar{f}_n, \beta\right) p\left(\bar{\mathbf{f}}|\mathbf{X}, \theta\right) df$$

$$p\left(f_n|\bar{f}_n, \beta\right) = N\left(f_n|\bar{f}_n, \beta^{-1}\right)$$

Variational Inference (PPA)

$$\log p(\mathbf{y}) \geq \sum_{n=1}^{N} \left\langle \log p(y_n|f_n)\, p\left(f_n|\bar{f}_n, \beta\right) p\left(\bar{\mathbf{f}}|\mathbf{X}, \theta\right) \right\rangle_{\prod_{n=1}^{N} q(\bar{f}_n) q(\bar{\mathbf{f}})}$$
$$- \sum_{n=1}^{N} \langle \log q(f_n) \rangle_{q(f_n)} - \left\langle \log q\left(\bar{f}\right) \right\rangle_{q(\bar{\mathbf{f}})}$$

Maximised by

$$q\left(\bar{\mathbf{f}}\right) \propto \exp \left\langle \sum_{n=1}^{N} \log p\left(f_n|\bar{f}_n, \beta\right) \right\rangle_{\prod_{n=1}^{N} q(f_n)} p\left(\bar{\mathbf{f}}|\mathbf{X}, \theta\right)$$

and

$$q(f_n) \propto \exp \left\langle \log p\left(f_n|\bar{f}_n, \beta\right) \right\rangle_{q(\bar{\mathbf{f}})} p(y_n|f_n)$$

Expectations of $\log p\left(f_n|\bar{f}_n, \beta\right)$

- Since $\exp \left\langle \log p\left(f_n|\bar{f}_n, \beta\right)\right\rangle_{q(f_n)} \propto N\left(\langle f_n\rangle|\bar{f}_n, \beta^{-1}\right)$ we have

$$q\left(\bar{\mathbf{f}}\right) \propto \prod_{n=1}^{N} N\left(\langle f_n\rangle|\bar{f}_n, \beta^{-1}\right) p\left(\bar{\mathbf{f}}|\mathbf{X}, \theta\right)$$

- Since $\exp \left\langle \log p\left(f_n|\bar{f}_n, \beta\right)\right\rangle_{q(\bar{\mathbf{f}})} \propto N\left(f_n|\left\langle\bar{f}_n\right\rangle, \beta^{-1}\right)$ we have

$$q\left(f_n\right) \propto N\left(f_n|\left\langle\bar{f}_n\right\rangle, \beta^{-1}\right) p\left(y_n|f_n\right)$$

- So
  - $q\left(\bar{\mathbf{f}}\right)$ is a Gaussian process regardless of form of $p\left(y_n|f_n\right)$.
  - Moments of $q\left(f_n\right)$ are straightforward to compute for any $p\left(y_n|f_n\right)$ see *e.g. Csató [2002]*

Speed up Variational Method

- Variational Methods can be tediously slow. (yawn!)

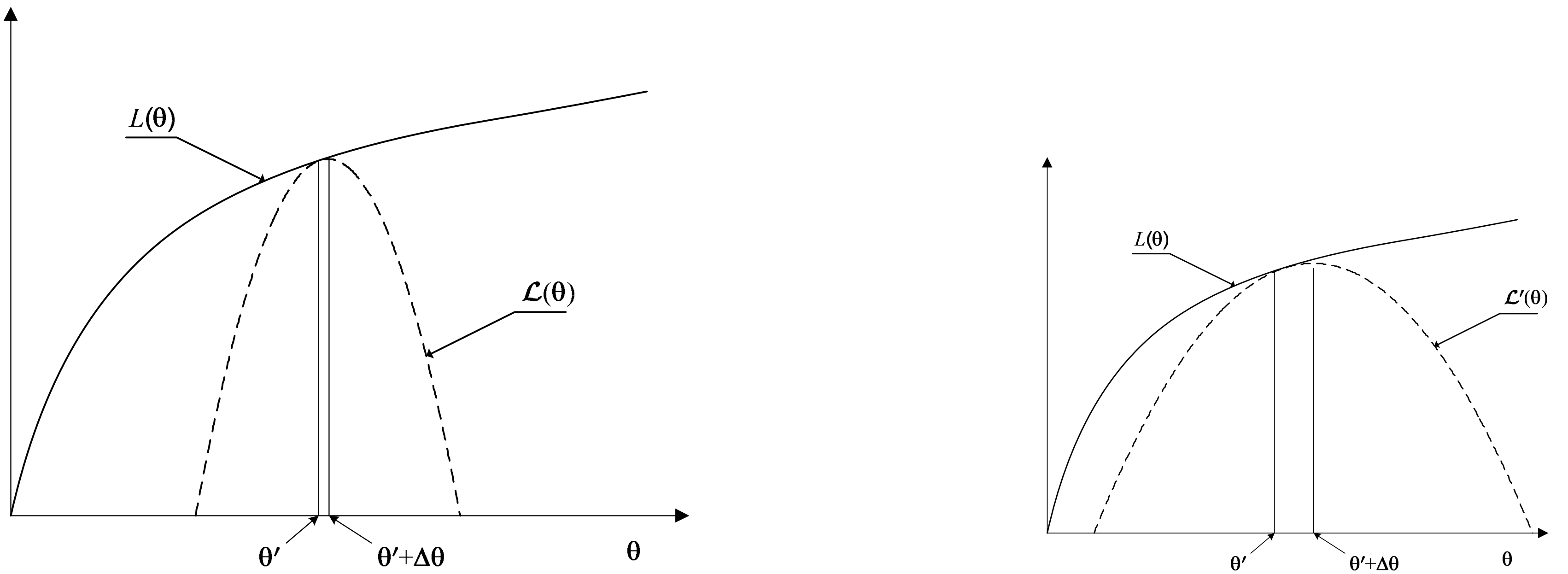


Figure 3: When variational methods are slow.

- Problem occurs when bound's quality degrades rapidly with parameter changes.

KL Corrected Variational Inference

- Updating Parameters: variational lower bound

$$L(\beta, \theta) = \sum_{n=1}^{N} \left\langle \log p\left(f_n|\bar{f}_n, \beta\right)\right\rangle_{q(\bar{f}_n)q(f_n)} + \left\langle \log p\left(\bar{\mathbf{f}}|\mathbf{X}, \theta\right)\right\rangle_{q(\bar{\mathbf{f}})}. \quad (1)$$

- Solution: make the quality of the bound responsive to changes in the parameters.

KL Corrected Variational Inference

- Ideally we would like to optimise the marginal likelihood,

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \beta) = L(\boldsymbol{\theta}) = \log \int \prod_{n=1}^{N} p\left(y_n|\bar{f}_n, \beta\right) p\left(\bar{\mathbf{f}}|\mathbf{X}, \boldsymbol{\theta}\right) d\bar{\mathbf{f}}, \tag{2}$$

- Substitute for noise model

$$\begin{aligned} \log p\left(y_n|\bar{f}_n, \beta\right) \geq & \left\langle \log p\left(y_n|f_n\right)\right\rangle_{q(f_n)} + \left\langle \log p\left(f_n|\bar{f}_n, \beta\right)\right\rangle_{q(f_n)} \\ & - \sum_{n=1}^{N} \left\langle \log q\left(f_n\right)\right\rangle_{q(f_n)}, \end{aligned}$$

KL Corrected Lower Bound

- A new lower bound is

$$\log p\left(\mathbf{y}|\mathbf{X},\theta,\beta\right) \geq \log \int \prod_{n=1}^{N} \exp \left\langle \log p\left(f_n|\bar{f}_n,\beta\right)\right\rangle_{q(f_n)} p\left(\bar{\mathbf{f}}|\mathbf{X},\theta\right) d\bar{\mathbf{f}} + \mathrm{const}$$

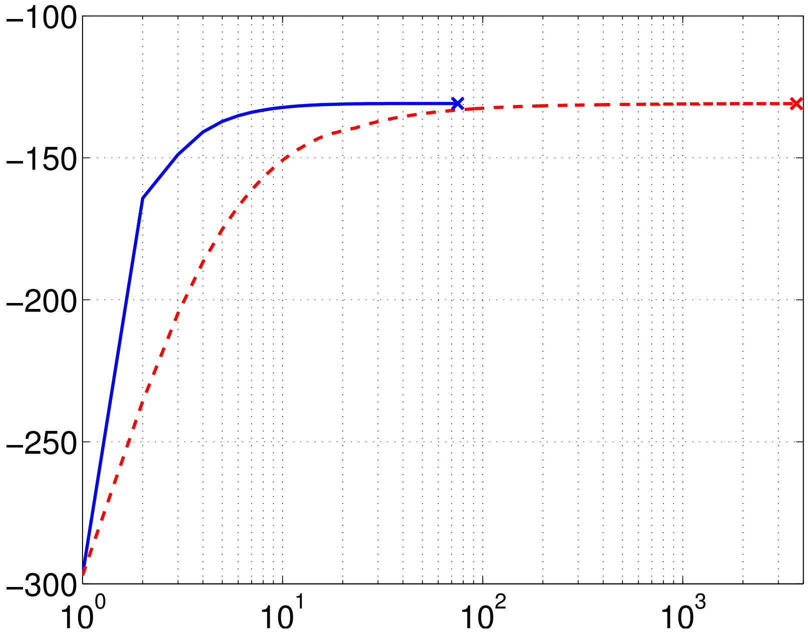
- Which leads to

$$\mathcal{L}'\left(\theta\right) = \log \int \prod_{n=1}^{N} N\left(\left\langle f_n\right\rangle|\bar{f}_n,\beta^{-1}\right) p\left(\bar{\mathbf{f}}|\mathbf{X}\right) d\bar{\mathbf{f}} + \mathrm{const},$$

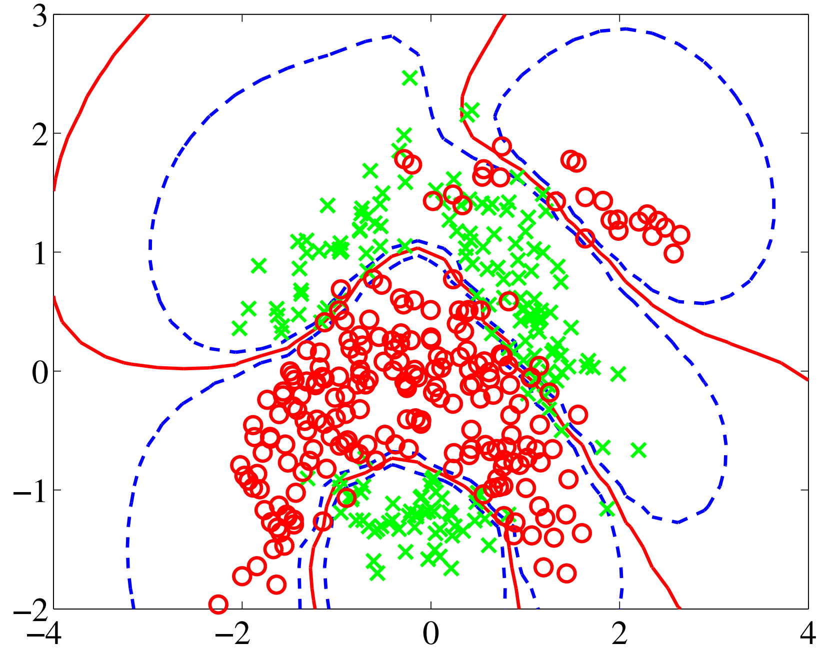
which does not depend on $q\left(\bar{\mathbf{f}}\right)$.

- This is the KL corrected bound.

Convergence Speed



(a)



(b)

Figure 4: (a) Plot of log-likelihood vs iteration number (log-scale) for the KL-corrected objective function (solid line) and the standard variational bound (dashed line). (b) The resulting classification of the **banana** data set.

- KL-corrected requires 74 iterations for convergence, standard variational inference (via PPA) requires 3697 iterations.

Alternative View Point

The marginal likelihood for the augmented model is

$$\log p\left(\mathbf{y}\right) = \int \int p\left(y_n|f_n\right) p\left(f_n|\bar{f}_n, \beta\right) df_n p\left(\bar{\mathbf{f}}|\mathbf{X}, \theta\right) d\bar{\mathbf{f}}$$

to make progress we insert a variational lower bound on the inner integral,

$$\begin{aligned}\log \int p\left(y_n|f_n\right) p\left(f_n|\bar{f}_n, \beta\right) df_n \quad \geq \quad & \left\langle \log p\left(y_n|f_n\right)\right\rangle_{q(f_n)} \\ & + \left\langle \log p\left(f_n|\bar{f}_n, \beta\right)\right\rangle_{q(f_n)} \\ & + \left\langle \log q\left(f_n\right)\right\rangle_{q(f_n)}\end{aligned}$$

New Bound

Substituting in this lower bound we have,

$$\begin{aligned}
\log p(\mathbf{y}) \geq\ & \log \int \prod_{n=1}^{N} \exp \left\langle \log p\left(f_n|\bar{f}_n, \beta\right)\right\rangle p\left(\bar{\mathbf{f}}|\mathbf{X}, \theta\right) d\bar{\mathbf{f}} \\
& + \sum_{n=1}^{N} \left\langle \log p\left(y_n|f_n\right)\right\rangle_{q(f_n)} \\
& + \sum_{n=1}^{N} \left\langle \log q\left(f_n\right)\right\rangle_{q(f_n)} \\
\doteq\ & \mathcal{L}'(\theta) \qquad (3)
\end{aligned}$$

Minimise directly wrt $q\left(f_n\right)$

Bound's dependence on $q\left(f_n\right)$ is summarised as

$$\begin{aligned}\mathcal{L}'_n\left(\theta\right) &= \frac{1}{2}\left(\beta-\frac{1}{\sigma_n^2}\right)\left(\left\langle f_n^2\right\rangle-\left\langle f_n\right\rangle^2\right)-\left\langle\log N\left(f_n|\mu_n,\sigma_n^2\right)\right\rangle\\ &+\left\langle\log p\left(y_n|f_n\right)\right\rangle+\left\langle\log q\left(f_n\right)\right\rangle_{q(f_n)}+\text{const}\end{aligned}$$

where

$$\mu_n=\mathbf{k}_n^{\mathrm{T}}\left(\mathbf{K}_{\backslash n}+\beta^{-1}\mathbf{I}\right)^{-1}\left\langle\mathbf{f}_{\backslash n}\right\rangle$$

and

$$\sigma_n^2=\beta^{-1}+k_{nn}-\mathbf{k}_n^{\mathrm{T}}\left(\mathbf{K}_{\backslash n}+\beta^{-1}\mathbf{I}\right)^{-1}\mathbf{k}_n$$

where If $\frac{1}{2}\left(\beta-\frac{1}{\sigma_n^2}\right)\left(\left\langle f_n^2\right\rangle-\left\langle f_n\right\rangle^2\right)$ is small then this implies

$$q\left(f_n\right)\propto p\left(y_n|f_n\right)N\left(f_n|\mu_n,\sigma_n^2\right)$$

Which is very similar to the approximating distribution that arises in ... EP

Conclusions

- Variational inference in GPs is practical.
  - Various noise models can be accommodated.
  - Slow convergence can be solved.
- Recent (Monday & Tuesday!) analysis suggests connections with EP.

References

Lehel Csató. *Gaussian Processes — Iterative Sparse Approximations*. PhD thesis, Aston University, 2002.

Matthias Seeger. Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 603–609, Cambridge, MA, 2000. MIT Press.