On the paper: Variational Learning of Inducing Variables in Sparse Gaussian Processes (Titsias, 2009)

Thang Bui and Richard Turner

May 25, 2014

Abstract

This summary was prepared for our internal reading club and serves as notes on the sparse GP regression using the variational method (Titsias, 2009). We also discuss why this approximation can be viewed as the corrected version of the Projected Process or Deterministic Training Conditional (DTC) approximation (Seeger, 2003).

Consider the following regression problem: $y_i = f(x_i) + \epsilon_i$, where $f \sim \mathcal{GP}(0, k(x_i, x_j))$ and $\epsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$. Augment the model with inducing variables $\mathbf{u}$ where $\dim(u) = m \ll \dim(f) = n$, we have the posterior of $\mathbf{f}$ and $\mathbf{u}$:

$$p(\mathbf{f}, \mathbf{u}|\mathbf{y}) = \frac{p(\mathbf{f}|\mathbf{u})p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{p(\mathbf{y})}. \tag{1}$$

The reason for this augmentation is that it allows us to produce tractable approximations (in the same spirit with the FITC family). As seen in the above expression, $\mathbf{u}$ can be analytically integrated out to obtain the joint distribution of $\mathbf{f}$ and $\mathbf{y}$: $p(\mathbf{f}, \mathbf{y}) = \int \mathrm{d}\mathbf{u} p(\mathbf{f}, \mathbf{u}, \mathbf{y})$.

The variational approach

- Minimising the KL divergence between $q(\mathbf{f}, \mathbf{u})$ and $p(\mathbf{f}, \mathbf{u}|\mathbf{y})$:

$$\mathrm{KL}(q(\mathbf{f}, \mathbf{u})||p(\mathbf{f}, \mathbf{u}|\mathbf{y})) = \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} q(\mathbf{f}, \mathbf{u}) \log \frac{q(\mathbf{f}, \mathbf{u})}{p(\mathbf{f}, \mathbf{u}|\mathbf{y})} \tag{2}$$

$$= \log p(\mathbf{y}) + \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} q(\mathbf{f}, \mathbf{u}) \log \frac{q(\mathbf{f}, \mathbf{u})}{p(\mathbf{f}, \mathbf{u}, \mathbf{y})} \tag{3}$$

$$\text{hence,} \quad \log p(\mathbf{y}) = \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} q(\mathbf{f}, \mathbf{u}) \log \frac{p(\mathbf{f}, \mathbf{u}, \mathbf{y})}{q(\mathbf{f}, \mathbf{u})} + \mathrm{KL}(q(\mathbf{f}, \mathbf{u})||p(\mathbf{f}, \mathbf{u}|\mathbf{y})), \tag{4}$$

  or, $\mathcal{F}(q(\mathbf{u})) = \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} q(\mathbf{f}, \mathbf{u}) \log \frac{p(\mathbf{f}, \mathbf{u}, \mathbf{y})}{q(\mathbf{f}, \mathbf{u})}$ is the evidence lower bound (ELBO).

- An alternative, equivalent way to obtain the ELBO is by using the Jensen's inequality:

$$\log p(\mathbf{y}) = \log \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} p(\mathbf{y}, \mathbf{f}, \mathbf{u}) \tag{5}$$

$$= \log \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} q(\mathbf{f}, \mathbf{u}) \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})} \tag{6}$$

$$\geq \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} q(\mathbf{f}, \mathbf{u}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})} \tag{7}$$

Titsias, 2009 chose a variational distribution $q(\mathbf{f}, \mathbf{u})$ such that,

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}). \tag{8}$$

In general $q(\mathbf{f}, \mathbf{u}) = q(\mathbf{f}|\mathbf{u})q(\mathbf{u})$, here the term $q(\mathbf{f}|\mathbf{u})$ is replaced by $p(\mathbf{f}|\mathbf{u})$ which is the prior conditional distribution. This particular choice means that now the only way $\mathbf{y}$ to affect $\mathbf{f}$ is through $\mathbf{u}$, as opposed to $\mathbf{f}$ separates $\mathbf{u}$ and $\mathbf{y}$ as in the original model. This also helps us mathematically as the optimal distribution $q(u)$ can be obtained analytically. This however limits the applicability of this tricks to derive an extension or a slightly different approximation. Substitute $q(\mathbf{f}, \mathbf{u})$ into the ELBO,

$$\begin{aligned}
\mathcal{F}(q(\mathbf{u})) &= \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \frac{p(\mathbf{f}|\mathbf{u})p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} && (9)\\
&= \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \frac{p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{q(\mathbf{u})} && (10)\\
&= \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \frac{p(\mathbf{u})}{q(\mathbf{u})} + \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}). && (11)
\end{aligned}$$

To find the optimal form of $q(\mathbf{u})$ that maximises the ELBO, consider the derivative of the ELBO w.r.t $q(\mathbf{u})$ with the addition of the Lagrange multiplier,

$$\frac{\mathrm{d}}{\mathrm{d}q(\mathbf{u})}\mathcal{F} + \lambda = \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \left[\log p(\mathbf{u}) - \log q(\mathbf{u}) - 1\right] + \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}) + \lambda \tag{12}$$

Letting 12 equal 0 gives

$$q(\mathbf{u}) = \frac{p(\mathbf{u})}{Z} \exp\left(\int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})\right), \tag{13}$$

or $q(\mathbf{u}) = p(\mathbf{u})H(\mathbf{y}, \mathbf{u})/Z$ where $H(\mathbf{y}, \mathbf{u}) = \exp\left(\int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})\right)$. Substitute this into the ELBO in 10,

$$\begin{aligned}
\mathcal{F}(q(\mathbf{u})) &= \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \frac{p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{p(\mathbf{u})H(\mathbf{y}, \mathbf{u})/Z} && (14)\\
&= \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \left[\log p(\mathbf{y}|\mathbf{f}) - \log H(\mathbf{y}, \mathbf{u}) + \log Z\right] && (15)\\
&= \int \mathrm{d}\mathbf{u} q(\mathbf{u}) \left[\log Z \underbrace{- \log H(\mathbf{y}, \mathbf{u}) + \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})}_{=0}\right] && (16)\\
&= \log Z && (17)
\end{aligned}$$

Consider the integral inside the exponential in the optimal $q(\mathbf{u})$:

$$\begin{aligned}
\mathrm{M} &= \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}) && (18)\\
&= \int \mathrm{d}\mathbf{f} \mathcal{N}(\mathbf{f}; K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mathbf{u}, K_{\mathbf{ff}} - K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}K_{\mathbf{uf}}) \log[\mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma^2 I)] && (19)\\
&= \int \mathrm{d}\mathbf{f} \mathcal{N}(\mathbf{f}; \mathbf{A}, \mathbf{B}) \left[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{f})^\intercal I(\mathbf{y} - \mathbf{f})\right] && (20)\\
&= \int \mathrm{d}\mathbf{f} \mathcal{N}(\mathbf{f}; \mathbf{A}, \mathbf{B}) \left[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\operatorname{Tr}(\mathbf{y}\mathbf{y}^\intercal - 2\mathbf{y}\mathbf{f}^\intercal + \mathbf{f}\mathbf{f}^\intercal)\right] && (21)\\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\operatorname{Tr}(\mathbf{y}\mathbf{y}^\intercal - 2\mathbf{y}\mathbf{A}^\intercal + \mathbf{A}\mathbf{A}^\intercal + \mathbf{B}) && (22)\\
&= -\frac{1}{2\sigma^2}\operatorname{Tr}(\mathbf{B}) + \log[\mathcal{N}(\mathbf{y}; \mathbf{A}, \sigma^2 I)], && (23)
\end{aligned}$$

where $\mathbf{A} = K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mathbf{u}$ and $\mathbf{B} = K_{\mathbf{ff}} - K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}K_{\mathbf{uf}}$. Hence the optimal form of $q(\mathbf{u})$ can be found analytically as follows,

$$q(\mathbf{u}) = \mathcal{N}(K_{mn}(K_{nm}K_{mm}^{-1}K_{mn} + \sigma^2 I)\mathbf{y}, K_{mm} - K_{mn}(K_{nm}K_{mm}^{-1}K_{mn} + \sigma^2 I)K_{nm}) \tag{24}$$

$$= \mathcal{N}(\sigma^{-2}K_{mm}\Sigma K_{mn}\mathbf{y}, K_{mm}\Sigma K_{mm}), \tag{25}$$

where $\Sigma = (K_{mm} + \sigma^{-2}K_{mn}K_{nm})^{-1}$. The lower bound on the marginal likelihood is:

$$\mathcal{F}(q(\mathbf{u})) = \underbrace{\log \mathcal{N}(\mathbf{y}|0, \sigma^2 I + K_{nm}K_{mm}^{-1}K_{mn})}_{\text{DTC log marginal likelihood}} \underbrace{- \frac{1}{2\sigma^2}\mathrm{Tr}(K_{nn} - K_{nm}K_{mm}^{-1}K_{mn})}_{\text{regulariser - avoid overfitting}}. \tag{26}$$

Relationship with the DTC approximation

The likelihood approximation presented in Csató and Opper, 2002; Seeger, 2003 can be justified by choosing a likelihood function $q(\mathbf{y}|\mathbf{u})$ to minimise the KL divergence,

$$q(\mathbf{y}|\mathbf{u}) \leftarrow \underset{q(\mathbf{y}|\mathbf{u})}{\arg\min}\ \mathrm{KL}(q(\mathbf{f}, \mathbf{u}|\mathbf{y})||p(\mathbf{f}, \mathbf{u}|\mathbf{y})), \tag{27}$$

where,

$$q(\mathbf{f}, \mathbf{u}|\mathbf{y}) = \frac{q(\mathbf{y}|\mathbf{u})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})}{q(\mathbf{y})}, \tag{28}$$

$$p(\mathbf{f}, \mathbf{u}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})}{p(\mathbf{y})}, \tag{29}$$

$$\text{and,} \quad q(\mathbf{y}) = \int \mathrm{d}\mathbf{u} p(\mathbf{u})p(\mathbf{y}|\mathbf{u}). \tag{30}$$

Consider the likelihood $q(\mathbf{y}|\mathbf{u})$ that is a normalised Gaussian or $\int \mathrm{d}y q(\mathbf{y}|\mathbf{u}) = 1$ Seeger, 2003, combining the reversed KL divergence and the normalisation assumption above gives us the Lagrangian:

$$\mathcal{L} = \mathrm{KL}(q(\mathbf{f}, \mathbf{u}|\mathbf{y})||p(\mathbf{f}, \mathbf{u}|\mathbf{y})) + \lambda\left(\int \mathrm{d}\mathbf{y} q(\mathbf{y}|\mathbf{u}) - 1\right) \tag{31}$$

$$= \log p(\mathbf{y}) - \log q(\mathbf{y}) + \int \mathrm{d}\mathbf{f}\, \mathrm{d}\mathbf{u} \frac{q(\mathbf{y}|\mathbf{u})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})}{q(\mathbf{y})} \log \frac{q(\mathbf{y}|\mathbf{u})}{p(\mathbf{y}|\mathbf{f})} + \lambda\left(\int \mathrm{d}y q(\mathbf{y}|\mathbf{u}) - 1\right). \tag{32}$$

The derivative $\mathcal{L}$ w.r.t $q(\mathbf{y}|\mathbf{u})$ is,

$$\frac{\partial}{\partial q(\mathbf{y}|\mathbf{u})} = \frac{1}{q(\mathbf{y})}p(\mathbf{u}) + \int \mathrm{d}\mathbf{f} \frac{p(\mathbf{f}|\mathbf{u})p(\mathbf{u})}{q(\mathbf{y})} \log \frac{q(\mathbf{y}|\mathbf{u})}{p(\mathbf{y}|\mathbf{f})} + \int \mathrm{d}\mathbf{f} \frac{q(\mathbf{y}|\mathbf{u})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})}{q(\mathbf{y})} \frac{1}{q(\mathbf{y}|\mathbf{u})} + \lambda \tag{33}$$

$$= \frac{p(\mathbf{u})}{q(\mathbf{y})}\left[\log q(\mathbf{y}|\mathbf{u}) - \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})\right] + \lambda. \tag{34}$$

Setting 34 to zero gives,

$$q(\mathbf{y}|\mathbf{u}) = \exp\left(-\frac{\lambda q(\mathbf{y})}{p(\mathbf{u})}\right) \exp\left(\int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})\right). \tag{35}$$

Similar as in the variational approach but now we have to use the normalisation constraint, the optimal form for the likelihood approximation is:

$$\boxed{q(\mathbf{y}|\mathbf{u}) = \mathcal{N}(\mathbf{y}; K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mathbf{u}, \sigma^2\,\mathrm{I})} \tag{36}$$

As noted by Snelson, 2007, the same result can be obtained by optimising the KL divergence between the joint models of y, $\mathbf{f}$ and $\mathbf{u}$: $\mathrm{KL}(q(\mathbf{y}|\mathbf{u})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})||p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}))$.

Let's remove the normalisation contraint of the likelihood term, this equivalently means that the Lagrange multiplier in the above expression is zero, or the optimal likelihood is:

$$\boxed{q(\mathbf{y}|\mathbf{u}) = \exp\left(-\frac{1}{2\sigma^2}\,\mathrm{Tr}(K_{\mathbf{ff}} - K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}K_{\mathbf{uf}})\right)\mathcal{N}(y; K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mathbf{u}, \sigma^2\,\mathrm{I})} \tag{37}$$

Here it becomes clear that why the expression of the posterior of $\mathbf{u}$ in DTC is exactly the same as in Titsias, 2009 and only the approximate marginal likelihood are different. Both are optimising the same KL divergence under the approximate likelihood regime, but Titsias, 2009 allows a free form for the likelihood (which turns out to be easily computed analytically) as opposed to a Gaussian likelihood in Seeger, 2003.

Open questions: 1. tighter bound, how biased is the variational approach is in learning? 2. are $\mathbf{u}$ truly variational parameters?

References

Csató, Lehel and Opper, Manfred (2002). "Sparse on-line Gaussian processes". In: *Neural computation* 14.3, pp. 641–668.

Seeger, Matthias (2003). "Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations". PhD thesis. School of Informatics, College of Science and Engineering, University of Edinburgh.

Snelson, Edward (2007). "Flexible and efficient Gaussian process models for machine learning". PhD thesis. Gatsby Computational Neuroscience Unit, University College London.

Titsias, Michalis K. (2009). "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In: *International Conference on Artificial Intelligence and Statistics*, pp. 567–574.