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 $\epsilon \sim_{iid} \mathcal{N}(0, \sigma^2)$. Augment the model with inducing variables $u$ where $\text{dim}(u) = m \ll \text{dim}(f) = n$, we have the posterior of $f$ and $u$: $p(f, u | y) = p(f | u) p(u) p(y | f)$.

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, $u$ can be analytically integrated out to obtain the joint distribution of $f$ and $y$: $p(f, y) = \int du p(f, u, y)$.

The variational approach

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

$$\text{KL}(q(f, u) || p(f, u | y)) = \int du df q(f, u) \log \frac{q(f, u)}{p(f, u | y)} = \log p(y) + \int du df q(f, u) \log \frac{q(f, u)}{p(f, u, y)}$$

hence, $\log p(y) = \int du df q(f, u) \log \frac{p(f, u, y)}{q(f, u)} + \text{KL}(q(f, u) || p(f, u | y))$, or, $\mathcal{F}(q(u)) = \int du df q(f, u) \log \frac{p(f, u, y)}{q(f, u)}$ is the evidence lower bound (ELBO).

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

$$\log p(y) = \log \int du df p(y, f, u)$$

$$= \log \int du df q(f, u) \frac{p(y, f, u)}{q(f, u)}$$

$$\geq \int du df q(f, u) \log \frac{p(y, f, u)}{q(f, u)}$$
Titsias, 2009 chose a variational distribution \( q(f, u) \) such that,

\[
q(f, u) = p(f|u)q(u) \tag{8}
\]

In general \( q(f, u) = q(f|u)q(u) \), here the term \( q(f|u) \) is replaced by \( p(f|u) \) which is the prior conditional distribution. This particular choice means that now the only way \( q \) to affect \( f \) is through \( u \), as opposed to \( f \) separates \( u \) and \( y \) 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 \( q(f, u) \) into the ELBO,

\[
\mathcal{F}(q(u)) = \int dufdp(f|u)q(u) \log \frac{p(f|u)p(y|f)}{p(f|u)q(u)} \tag{9}
\]

\[
= \int dufdp(f|u)q(u) \log \frac{p(u)p(y|f)}{q(u)} \tag{10}
\]

\[
= \int dufdp(f|u)q(u) \log \frac{p(u)}{q(u)} + \int dufdp(f|u)q(u) \log p(y|f). \tag{11}
\]

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

\[
\frac{d}{dq(u)} \mathcal{F} + \lambda = \int df dp(f|u) [\log p(u) - \log q(u) - 1] + \int df dp(f|u) \log p(y|f) + \lambda \tag{12}
\]

Letting \( \frac{d}{dq(u)} \mathcal{F} + \lambda \) equal 0 gives

\[
q(u) = \frac{p(u)}{Z} \exp \left( \int df dp(f|u) \log p(y|f) \right), \tag{13}
\]

or \( q(u) = p(u)H(y, u)/Z \) where \( H(y, u) = \exp \left( \int df dp(f|u) \log p(y|f) \right) \)). Substitute this into the ELBO in \( \mathcal{F}(q(u)) \)

\[
\mathcal{F}(q(u)) = \int du dp(f|u)q(u) \log \frac{p(u)p(y|f)}{p(u)H(y, u)/Z} \tag{14}
\]

\[
= \int du dp(f|u)q(u) [\log p(y|f) - \log H(y, u) + \log Z] \tag{15}
\]

\[
= \int duq(u) \left[ \log Z - \log H(y, u) + \int df dp(f|u) \log p(y|f) \right] \tag{16}
\]

\[
= \log Z \tag{17}
\]

Consider the integral inside the exponential in the optimal \( q(u) \):

\[
M = \int df dp(f|u) \log p(y|f) \tag{18}
\]

\[
= \int dfN(f; K_ufK_{uu}^{-1}u, K_ff - K_ufK_{uu}^{-1}K_{uf}) \log[N(y; f, \sigma^2 I)] \tag{19}
\]

\[
= \int dfN(f; A, B) \left[ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - f)^\top I (y - f) \right] \tag{20}
\]

\[
= \int dfN(f; A, B) \left[ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} Tr(yy^\top - 2yf^\top + ff^\top) \right] \tag{21}
\]

\[
= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} Tr(B) + \log[N(y; A, \sigma^2 I)], \tag{22}
\]

\[
= -\frac{1}{2\sigma^2} Tr(B) + \log[N(y; A, \sigma^2 I)], \tag{23}
\]

2
where \( A = K_fuK_u^{-1}u \) and \( B = K_fuK_u^{-1}K_{uf} \). Hence the optimal form of \( q(u) \) can be found analytically as follows,

\[
q(u) = \mathcal{N}(K_{mn}(K_{mm}K_{mm}^{-1}K_{mm} + \sigma^2 I)y, K_{mm} - K_{mn}(K_{mm}K_{mm}^{-1}K_{mm} + \sigma^2 I)K_{mn})
\]

\( (24) \)

\[
= \mathcal{N}(\sigma^{-2}K_{mm}\Sigma K_{mm}y, K_{mm}\Sigma K_{mm}),
\]

\( (25) \)

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

\[
F(q(u)) = \log \mathcal{N}(y|0, \sigma^2 I + K_{mm}K_{mm}^{-1}K_{mm}) - \frac{1}{2\sigma^2} \text{Tr}(K_{mm}^{-1}K_{mm}^{-1}K_{mm}^{-1}).
\]

\( (26) \)

**Relationship with the DTC approximation**

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

\[
q(y|u) \leftarrow \arg\min_{q(y|u)} \text{KL}(q(f, u|y)||p(f, u|y)),
\]

\( (27) \)

where,

\[
q(f, u|y) = \frac{q(y|u)p(f|u)p(u)}{q(y)},
\]

\( (28) \)

\[
p(f, u|y) = \frac{p(y|f)p(f|u)p(u)}{p(y)}.
\]

\( (29) \)

and,

\[
q(y) = \int du p(u)p(y|u).
\]

\( (30) \)

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

\[
\mathcal{L} = \text{KL}(q(f, u|y)||p(f, u|y)) + \lambda \left( \int dq(y|u) - 1 \right)
\]

\( (31) \)

\[
= \log p(y) - \log q(y) + \int df du \frac{q(y|u)p(f|u)p(u)}{q(y)} \log \frac{q(y|u)}{p(y|f)} + \lambda \left( \int dq(y|u) - 1 \right).
\]

\( (32) \)

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

\[
\frac{\partial}{\partial q(y|u)} = \frac{1}{q(y)} p(u) + \int df \frac{p(f|u)p(u)}{q(y)} \log \frac{q(y|u)}{p(y|f)} + \int df \frac{q(y|u)p(f|u)p(u)}{q(y)} \frac{1}{q(y|u)} + \lambda
\]

\( (33) \)

\[
= \frac{p(u)}{q(y)} \left[ \log q(y|u) - \int df p(f|u) \log p(y|f) \right] + \lambda.
\]

\( (34) \)

Setting \( 34 \) to zero gives,

\[
q(y|u) = \exp \left( -\lambda q(y) \right) \exp \left( \int df p(f|u) \log p(y|f) \right).
\]

\( (35) \)

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

\[
q(y|u) = \mathcal{N}(y; K_fuK_u^{-1}u, \sigma^2 I)
\]

\( (36) \)
As noted by Snelson, 2007, the same result can be obtained by optimising the KL divergence between
the joint models of $y$, $f$ and $u$: $KL(q(y|u)p(f|u)p(u)||p(y|f)p(f|u)p(u))$.

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:

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

(37)

Here it becomes clear that why the expression of the posterior of $u$ 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 in learning? 2. are $u$ 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.

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.