Sparse Gaussian Processes for Bayesian Optimization

Supplementary Material

March 1, 2016

Here we derive the result (14), the weighted KL divergence. Consider two sparse Gaussian processes, $GP \sim \mathrm{GP}(\boldsymbol{\alpha}, C)$, $\widehat{GP} \sim \mathrm{GP}(\hat{\boldsymbol{\alpha}}, \hat{C})$, which share a covariance function K. We can therefore write

$$GP \sim \mathcal{N}(\Phi\boldsymbol{\alpha}, I_{\mathcal{F}} + \Phi C \Phi^\top) \equiv \mathcal{N}(\boldsymbol{\mu}, \Sigma) \ , \tag{1}$$

and likewise for $\widehat{GP}$, where Φ is again the feature space representation of the inducing variables, which are shared. This assumption is made without loss of generality, since the inducing variables of each GP can simply be concatenated into a combined representation. In such a representation, if an inducing variable is only used by one of the GPs, the other will have zeros in the corresponding entries of $\boldsymbol{\alpha}$ and C. This is seen for example in online selection of inducing variables, in which one GP has $m+1$ inducing variables and the other uses a subset of m of these.

We restate Equation 10 for convenience:

$$D_{KL}^{f}(P\|Q) = \int_{\mathcal{F}} P(\boldsymbol{x}) \log\Big(\frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}\Big)^{f(\boldsymbol{x})} d\boldsymbol{x} = \int_{\mathcal{F}} f(\boldsymbol{x}) P(\boldsymbol{x}) \log\frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} d\boldsymbol{x} \ . \tag{2}$$

As before, we define $f^*(\boldsymbol{x})$ to be the prediction of GP at a point $\boldsymbol{x}$ in feature space:

$$f^*(\boldsymbol{x}) = \frac{\boldsymbol{x}^\top \Phi\boldsymbol{\alpha}}{|y^*|} = \frac{\boldsymbol{\mu}^\top \boldsymbol{x}}{|y^*|} \ .$$

Note that the weighting function f^* uses the full GP prediction rather than the reduced GP, and is therefore constant through the optimization of $D_{KL}^{f^*}(GP\|\widehat{GP})$ with respect to $\widehat{GP}$.

Abusing notation slightly, we write $GP(\boldsymbol{x})$ to denote the evaluation of the normal distribution corresponding to GP at a point $\boldsymbol{x}$ in the feature space, and likewise for $\widehat{GP}$. Our goal is then to find

$$D_{KL}^{f^*}(GP\|\widehat{GP}) = \int_{\mathcal{F}} f(\boldsymbol{x}) GP(\boldsymbol{x}) \log\frac{GP(\boldsymbol{x})}{\widehat{GP}(\boldsymbol{x})} d\boldsymbol{x} \ . \tag{3}$$

In the following, we will use $D_{KL}^{f^*}$ as a shorthand for $D_{KL}^{f^*}(GP\|\widehat{GP})$.

Evaluating the log term based on the normal distributions of GP and $\widehat{GP}$, this becomes

$$D_{KL}^{f^*} = \int_{\mathcal{F}} f(\boldsymbol{x}) GP(\boldsymbol{x}) \big[\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) - \frac{1}{2}\log(|\Sigma\hat{\Sigma}^{-1}|)\big] d\boldsymbol{x} \ . \quad (4)$$

Taking expectations over $GP \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, we have

$$\begin{aligned} 2D_{KL}^{f^*} &= E[f^*(\boldsymbol{x})(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})] - \\ &\quad E[f^*(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] - \log(|\Sigma\hat{\Sigma}^{-1}|) E[f^*(\boldsymbol{x})] \\ &\Rightarrow 2|y^*| D_{KL}^{f^*} = E[\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})] - \\ &\quad E[\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] - \log(|\Sigma\hat{\Sigma}^{-1}|)\|\boldsymbol{\mu}\|^2 \ . \quad (5) \end{aligned}$$

We use the following formula for Gaussian expectation:

$$\begin{aligned} E[\boldsymbol{x}(\boldsymbol{x}-\boldsymbol{m})^\top M(\boldsymbol{x}-\boldsymbol{m})] &= \Sigma M(\boldsymbol{\mu}-\boldsymbol{m}) + \Sigma M^\top(\boldsymbol{\mu}-\boldsymbol{m}) \\ &\quad + \mathrm{Tr}[\Sigma M^\top]\boldsymbol{\mu} + \boldsymbol{\mu}(\boldsymbol{\mu}-\boldsymbol{m})^\top M(\boldsymbol{\mu}-\boldsymbol{m}) \ . \quad (6) \end{aligned}$$

Since $\hat{\Sigma}$ is symmetric, this allows us to reduce (5) to

$$\begin{aligned} 2|y^*| D_{KL}^{f^*} &= 2\boldsymbol{\mu}^\top \Sigma\hat{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}) + \mathrm{Tr}[\Sigma\hat{\Sigma}^{-1} - I_{\mathcal{F}}]\|\boldsymbol{\mu}\|^2 \\ &\quad + (\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top \hat{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\|\boldsymbol{\mu}\|^2 - \log(|\Sigma\hat{\Sigma}^{-1}|)\|\boldsymbol{\mu}\|^2 \ . \quad (7) \end{aligned}$$

We now use the matrix inversion lemma to write

$$\hat{\Sigma}^{-1} = (I_{\mathcal{F}} + \Phi\hat{C}\Phi^\top)^{-1} = I_{\mathcal{F}} - \Phi(\hat{C}^{-1} + K)^{-1}\Phi^\top \ , \quad (8)$$

recalling that $K = \Phi^\top\Phi$. Then we can compute

$$\begin{aligned} \Phi^\top \Sigma\hat{\Sigma}^{-1}\Phi &= \\ &\quad \Phi^\top(I_{\mathcal{F}} + \Phi C\Phi^\top)(I_{\mathcal{F}} - \Phi(\hat{C}^{-1}+K)^{-1}\Phi^\top)\Phi \\ &= K - K(\hat{C}^{-1}+K)^{-1}K + KCK - KCK(\hat{C}^{-1}+K)^{-1}K \\ &\quad = (I+KC)(\hat{C}+K^{-1})^{-1} \ , \quad (9) \end{aligned}$$

where in the last line we have used the matrix inversion lemma again in the reverse direction. Likewise, we have

$$\Phi^\top\hat{\Sigma}^{-1}\Phi = K - K(\hat{C}^{-1}+K)^{-1}K = (\hat{C}+K^{-1})^{-1} \equiv \hat{V} \ , \quad (10)$$

In the same way, we can then compute

$$\mathrm{Tr}[\Sigma\hat{\Sigma}^{-1} - I_{\mathcal{F}}] = \mathrm{Tr}[(C-\hat{C})\hat{V}] \ , \log|\Sigma\hat{\Sigma}^{-1}| = \log|(C+K^{-1})\hat{V}| \ , \quad (11)$$

as shown in Csato (2002). Note that $\hat{C} = \hat{V}^{-1} - Q$; substituting then gives us w as defined previously:

$$w \equiv \mathrm{Tr}[\Sigma\hat{\Sigma}^{-1} - I_{\mathcal{F}}] - \log(|\Sigma\hat{\Sigma}^{-1}|) = \mathrm{Tr}[(C+Q)\hat{V} - I] - \log(|(C+Q)\hat{V}|) \ . \quad (12)$$

We can simplify (7) by substituting for $\boldsymbol{\mu}$, Σ, $\hat{\boldsymbol{\mu}}$, and $\hat{\Sigma}$:

$$\begin{aligned} 2|y^*|D_{KL}^{f^*} &= 2\boldsymbol{\alpha}^\top \Phi^\top \Sigma\hat{\Sigma}\Phi(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \\ &\quad (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^\top \Phi^\top (I_{\mathcal{F}} - \Phi(\hat{C}^{-1} + K)^{-1}\Phi^\top)\Phi(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\boldsymbol{\alpha}^\top \Phi^\top \Phi\boldsymbol{\alpha} \\ &\quad + w\boldsymbol{\alpha}^\top \Phi^\top \Phi\boldsymbol{\alpha} \\ &= 2\boldsymbol{\alpha}^\top (I + KC)\hat{V}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^\top \hat{V}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\boldsymbol{\alpha}^\top K\boldsymbol{\alpha} + w\boldsymbol{\alpha}^\top K\boldsymbol{\alpha} \ . \end{aligned} \quad (13)$$

We define

$$\Gamma = I + \frac{(I + KC)^\top}{\boldsymbol{\alpha}^\top K\boldsymbol{\alpha}} \ ,$$

and can then write

$$\begin{aligned} \frac{2|y^*|D_{KL}^{f^*}}{\boldsymbol{\alpha}^\top K\boldsymbol{\alpha}} &= 2\boldsymbol{\alpha}^\top (\Gamma^\top - I)\hat{V}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + \\ &\quad (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^\top \hat{V}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + w \\ &= \left[2\boldsymbol{\alpha}^\top (\Gamma^\top - I) + (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^\top\right]\hat{V}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + w \\ &= (2\Gamma\boldsymbol{\alpha} - (\boldsymbol{\alpha} + \hat{\boldsymbol{\alpha}}))^\top \hat{V}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}) + w \ . \end{aligned} \quad (14)$$

This is precisely Equation (14), and so we are done.

Similarly, we can compute the weighted KL divergence with arguments reversed:

$$\frac{2|y^*|D_{KL}^{f^*}(\widehat{GP}\|GP)}{\boldsymbol{\alpha}^\top K\hat{\boldsymbol{\alpha}}} = (2\hat{\Gamma}\boldsymbol{\alpha} + (\hat{\boldsymbol{\alpha}} - 3\boldsymbol{\alpha}))^\top V(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \hat{w} \ , \quad (15)$$

where

$$\begin{aligned} V = (C + K^{-1})^{-1}, \hat{\Gamma} = I + \frac{(I + K\hat{C})^\top}{\boldsymbol{\alpha}^\top K\hat{\boldsymbol{\alpha}}} \ , \\ \hat{w} = \mathrm{Tr}[(\hat{C} + K^{-1})V - I] - \log|(\hat{C} + K^{-1})V| \ . \end{aligned} \quad (16)$$