Variational Inference for Sparse Spectrum Approximation in Gaussian Process Regression – Appendix

Yarin Gal YG279@CAM.AC.UK
Richard Turner RET26@CAM.AC.UK
University of Cambridge

A. Appendix

Identity 1.

$$\cos(x-y) = \int_0^{2\pi} \frac{1}{2\pi} \sqrt{2}\cos(x+b)\sqrt{2}\cos(y+b)db$$

Proof. We first evaluate the term inside the integral. We have

$$\begin{aligned}
&\cos(x+b)\cos(y+b) \\
&\quad = (\cos(x)\cos(b) - \sin(x)\sin(b)) \\
&\qquad \cdot (\cos(y)\cos(b) - \sin(y)\sin(b)) \\
&\quad = (\cos(x)\cos(y))\cos^2(b) + (\sin(x)\sin(y))\sin^2(b) \\
&\qquad - (\sin(x)\cos(y) + \cos(x)\sin(y))\sin(b)\cos(b).
\end{aligned}$$

Now, since $\int \cos^2(b)\mathrm{d}b = \frac{b}{2} + \frac{1}{4}\sin(2b)$, as well as $\int \sin^2(b)\mathrm{d}b = \frac{b}{2} - \frac{1}{4}\sin(2b)$, and $\int \sin(b)\cos(b)\mathrm{d}b = -\frac{1}{4}\cos(2b)$, we have

$$\begin{aligned}
&\int_0^{2\pi} \frac{1}{2\pi}\sqrt{2}\cos(x+b)\sqrt{2}\cos(y+b)\mathrm{d}b \\
&\quad = \frac{1}{\pi}(\cos(x)\cos(y)(\pi - 0) \\
&\qquad + \sin(x)\sin(y)(\pi - 0) \\
&\qquad - (\sin(x)\cos(y) + \cos(x)\sin(y)) \cdot 0 \\
&\quad = \cos(x-y)
\end{aligned}$$

□

Identity 2.

$$E_{\mathcal{N}(\mathbf{w};\mu,\Sigma)}\big(\cos(\mathbf{w}^T\mathbf{x}+b)\big) = e^{-\frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}}\cos(\mu^T\mathbf{x}+b)$$

Proof. We rely on the characteristic function of the Gaussian distribution to prove this identity.

$$\begin{aligned}
&E_{\mathcal{N}(\mathbf{w};\mu,\Sigma)}\big(\cos(\mathbf{w}^T\mathbf{x}+b)\big) \\
&\quad = \Re\Big(e^{ib}E_{\mathcal{N}(\mathbf{w};\mu,\Sigma)}\big(e^{i\mathbf{w}^T\mathbf{x}}\big)\Big) \\
&\quad = \Re(e^{ib}e^{i\mathbf{w}^T\mu - \frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}}) \\
&\quad = e^{-\frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}}\cos(\mu^T\mathbf{x}+b)
\end{aligned}$$

where $\Re(\cdot)$ is the real part function, and the transition from the second to the third lines uses the characteristic function of a multivariate Gaussian distribution. □

Identity 3.

$$\begin{aligned}
&E_{\mathcal{N}(\mathbf{w};\mu,\Sigma)}\big(\cos(\mathbf{w}^T\mathbf{x}+b)^2\big) \\
&\qquad = \frac{1}{2}e^{-2\mathbf{x}^T\Sigma\mathbf{x}}\cos(2\mu^T\mathbf{x}+2b) + \frac{1}{2}
\end{aligned}$$

Proof. Following the identity $\cos(\theta)^2 = \frac{\cos(2\theta)+1}{2}$,

$$\begin{aligned}
&E_{\mathcal{N}(\mathbf{w};\mu,\Sigma)}\big(\cos(\mathbf{w}^T\mathbf{x}+b)^2\big) \\
&= \frac{1}{2}E_{\mathcal{N}(\mathbf{w};\mu,\Sigma)}\big(\cos(2\mathbf{w}^T\mathbf{x}+2b)\big) + \frac{1}{2} \\
&= \frac{1}{2}e^{-2\mathbf{x}^T\Sigma\mathbf{x}}\cos(2\mu^T\mathbf{x}+2b) + \frac{1}{2}
\end{aligned}$$

□

Proposition 1. *Given a sum of covariance functions with L components (with each corresponding to Φ_i an $N \times K$ matrix) we have $\Phi = [\Phi_i]_{i=1}^L$ an $N \times LK$ matrix.*

Proof. We extend the derivation of equation 2 to sums of covariance functions. Given a sum of covariance functions with L components

$$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^L \sigma_i^2 K_i(\mathbf{x},\mathbf{y}),$$

following equation 1 we have

$$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^L \int_{\mathbb{R}^Q} \sigma_i^2 p_i(\mathbf{w})\cos(2\pi\mathbf{w}^T(\mathbf{x}-\mathbf{y}))\mathrm{d}\mathbf{w},$$

where we write σ_i^2 instead of $\sigma\sigma_i^2$ for brevity (with σ_i^2 not having to sum to one).

Following the derivations of equation 2, for each component i in the sum we get Φ_i an $N \times K$ matrix. Writing $\Phi = [\Phi_i]_{i=1}^L$ an $N \times LK$ matrix, we have that the sum of covariance matrices can be expressed with a single term after marginalizing $\mathbf{F}$ out,

$$\sum_{i=1}^{L} \Phi_i \Phi_i^T + \tau^{-1}\mathbf{I} = \Phi\Phi^T + \tau^{-1}\mathbf{I},$$

thus identity 2 still holds. □

Proposition 2. *Performing a change of variables to a sum of SE covariance functions results in $p(\mathbf{w})$ a standard normal distribution with covariance function hyperparameters expressed in Φ.*

Proof. The sum of SE covariance functions' corresponding probability measure $p(\mathbf{w})$ is expressed as a mixture of Gaussians,

$$\begin{aligned} p(\mathbf{w}) &= \sum_{i=1}^{L} \sigma_i^2 \prod_{q=1}^{Q} \sqrt{2\pi} l_{iq} e^{-\frac{(2\pi l_{iq})^2}{2} w_q^2} \\ &= \sum_{i=1}^{L} \sigma_i^2 \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{L}_i^{-2}), \end{aligned}$$

with σ_i^2 summing to one.

Following equation 1 with the above $p(\mathbf{w})$ we perform a change of variables to get,

$$\begin{aligned} &K(\mathbf{x}, \mathbf{y}) \\ &= \sum_{i=1}^{L} \int_{\mathbb{R}^Q} \sigma_i^2 \mathcal{N}(\mathbf{w}'; \mathbf{0}, \mathbf{L}_i^{-2}) \cos(2\pi \mathbf{w}'^T(\mathbf{x} - \mathbf{y})) \mathrm{d}\mathbf{w}' \\ &= \sum_{i=1}^{L} \int_{\mathbb{R}^Q} \sigma_i^2 \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I}) \cos(2\pi (\mathbf{L}_i^{-1}\mathbf{w})^T(\mathbf{x} - \mathbf{y})) \mathrm{d}\mathbf{w} \end{aligned}$$

for $\mathbf{w}' = \mathbf{L}_i^{-1}\mathbf{w}$.

For each component i we get Φ_i an $N \times K$ matrix with elements

$$\sqrt{\frac{2\sigma_i^2}{K}} \cos\big(2\pi (\mathbf{L}_i^{-1}\mathbf{w}_k)^T(\mathbf{x} - \mathbf{z}_k) + b_k\big),$$

where for simplicity, we index $\mathbf{w}_k$ and b_k with $k = 1, ..., LK$ as a function of i. □

Proposition 3. *Let $p(\mathbf{a}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The optimal distribution $q(\mathbf{a})$ solving*

$$\int q(\mathbf{a}) \int q(\boldsymbol{\omega}) \log p(\mathbf{y}|\mathbf{a}, \mathbf{X}, \boldsymbol{\omega}) d\boldsymbol{\omega} d\mathbf{a} - KL(q(\mathbf{a})||p(\mathbf{a})) - KL(q(\boldsymbol{\omega})||p(\boldsymbol{\omega}))$$

is given by

$$q(\mathbf{a}_d) = \mathcal{N}(\boldsymbol{\Sigma} E_{q(\boldsymbol{\omega})}(\Phi^T)\mathbf{y}_d,\ \tau^{-1}\boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma} = (E_{q(\boldsymbol{\omega})}(\Phi^T\Phi) + \tau^{-1}I)^{-1}$.

The lower bound to optimise then reduces to

$$\begin{aligned} \mathcal{L} = \sum_{d=1}^{D} \Big(&- \frac{N}{2}\log(2\pi\tau^{-1}) - \frac{\tau}{2}\mathbf{y}_d^T\mathbf{y}_d \\ &+ \frac{1}{2}\log(|\tau^{-1}\boldsymbol{\Sigma}|) \\ &+ \frac{1}{2}\tau \mathbf{y}_d^T E_{q(\boldsymbol{\omega})}(\Phi)\boldsymbol{\Sigma} E_{q(\boldsymbol{\omega})}(\Phi^T)\mathbf{y}_d \Big) \\ - KL(q(\boldsymbol{\omega})||p(\boldsymbol{\omega})). \end{aligned}$$

Proof. Let

$$\begin{aligned} \mathcal{L} = &\int q(\mathbf{a}) \int q(\boldsymbol{\omega}) \log p(\mathbf{y}|\mathbf{a}, \mathbf{X}, \boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \mathrm{d}\mathbf{a} \\ &- \int q(\mathbf{a}) \log \frac{q(\mathbf{a})}{p(\mathbf{a})} \mathrm{d}\mathbf{a} - \int q(\boldsymbol{\omega}) \log \frac{q(\boldsymbol{\omega})}{p(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega}. \end{aligned}$$

We want to solve

$$\frac{\mathrm{d}(\mathcal{L} + \lambda \int (\int q(\mathbf{a}) \mathrm{d}\mathbf{a} - 1))}{\mathrm{d}q(\mathbf{a})} = 0$$

for some λ. I.e.

$$\int q(\boldsymbol{\omega}) \log p(\mathbf{y}|\mathbf{a}, \mathbf{X}, \boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} - \log \frac{q(\mathbf{a})}{p(\mathbf{a})} - 1 + \lambda = 0.$$

This means that

$$\begin{aligned} q(\mathbf{a}) &= e^{\lambda - 1} e^{\int q(\boldsymbol{\omega}) \log p(\mathbf{y}|\mathbf{a}, \mathbf{X}, \boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}} p(\mathbf{a}) \\ &= \exp\Big(- \frac{1}{2}\mathbf{a}^T \tau (E(\Phi^T\Phi) + \tau^{-1}I)\mathbf{a} \\ &\qquad + \big(\tau \mathbf{y}^T E(\Phi)\big)\mathbf{a} + ... \Big) \end{aligned}$$

and since $q(\mathbf{a})$ is Gaussian, it must be equal to

$$q(\mathbf{a}) = \mathcal{N}(\boldsymbol{\Sigma} E_{q(\boldsymbol{\omega})}(\Phi^T)\mathbf{y},\ \tau^{-1}\boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma} = (E_{q(\boldsymbol{\omega})}(\Phi^T\Phi) + \tau^{-1}I)^{-1}$.

Writing $p(\mathbf{a})$ and $q(\mathbf{a})$ explicitly and simplifying results in the required lower bound.

□

Proposition 4. *Denoting $\mathbf{M} = [\mathbf{m}_d]_{d=1}^D$, we have*

$$E_{q(\mathbf{y}^*|\mathbf{x}^*)}(\mathbf{y}^*) = E_{q(\boldsymbol{\omega})}\big(\phi_*\big)\mathbf{M}.$$

Proof. The d'th output y_d^* of the mean of the distribution is given by (writing $\phi_* = \phi(\mathbf{x}^*, \boldsymbol{\omega})$)

$$\begin{aligned} E_{q(y_d^*|\mathbf{x}^*)}(y_d^*) &= \int y_d^* p(y_d^*|\mathbf{x}^*, \mathbf{A}, \boldsymbol{\omega}) q(\mathbf{A}, \boldsymbol{\omega}) \mathrm{d}\mathbf{A}\mathrm{d}\boldsymbol{\omega}\mathrm{d}y_d^* \\ &= \int \big(\phi_* \mathbf{a}_d\big) q(\mathbf{A}, \boldsymbol{\omega}) \mathrm{d}\mathbf{A}\mathrm{d}\boldsymbol{\omega} \\ &= \int \phi_* q(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \int \mathbf{a}_d q(\mathbf{A}) \mathrm{d}\mathbf{A} \\ &= E_{q(\boldsymbol{\omega})}\big(\phi_*\big)\mathbf{m}_d, \end{aligned}$$

which can be evaluated analytically. □

Proposition 5. *The variance of the predictive distribution is given by*

$$\begin{aligned} &Var_{q(\mathbf{y}^*|\mathbf{x}^*)}(\mathbf{y}^*) = \tau^{-1}\mathbf{I}_D + \Psi \\ &\quad + \mathbf{M}^T\big(E_{q(\boldsymbol{\omega})}\big(\phi_*^T\phi_*\big) - E_{q(\boldsymbol{\omega})}\big(\phi_*\big)^T E_{q(\boldsymbol{\omega})}\big(\phi_*\big)\big)\mathbf{M} \end{aligned}$$

with $\Psi_{i,j} = tr\big(E_{q(\boldsymbol{\omega})}\big(\phi_*^T\phi_*\big) \cdot \mathbf{s}_i\big) \cdot \mathbb{1}[i=j]$.

Proof. The raw second moment of the distribution is given by (remember that $\mathbf{y}^*$ is a $1 \times D$ row vector)

$$\begin{aligned} &E_{q(\mathbf{y}^*|\mathbf{x}^*)}((\mathbf{y}^*)^T(\mathbf{y}^*)) \\ &= \int \left((\mathbf{y}^*)^T(\mathbf{y}^*) p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{A}, \boldsymbol{\omega}) \mathrm{d}\mathbf{y}^* \right) q(\mathbf{A}, \boldsymbol{\omega}) \mathrm{d}\mathbf{A}\mathrm{d}\boldsymbol{\omega} \\ &= \int \big(\mathrm{Cov}_{p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{A}, \boldsymbol{\omega})}(\mathbf{y}^*) \\ &\quad + E_{p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{A}, \boldsymbol{\omega})}(\mathbf{y}^*)^T E_{p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{A}, \boldsymbol{\omega})}(\mathbf{y}^*)\big) q(\mathbf{A}, \boldsymbol{\omega}) \mathrm{d}\mathbf{A}\mathrm{d}\boldsymbol{\omega} \\ &= \tau^{-1}\mathbf{I}_D + E_{q(\mathbf{A})q(\boldsymbol{\omega})}\big(\mathbf{A}^T\phi_*^T\phi_*\mathbf{A}\big). \end{aligned}$$

Now, for $i \neq j$ between 1 and D,

$$\begin{aligned} \left(E_{q(\mathbf{A})q(\boldsymbol{\omega})}\big(\mathbf{A}^T\phi_*^T\phi_*\mathbf{A}\big)\right)_{i,j} &= E_{q(\mathbf{A})q(\boldsymbol{\omega})}\big(\mathbf{a}_i^T\phi_*^T\phi_*\mathbf{a}_j\big) \\ &= \mathbf{m}_i^T E_{q(\boldsymbol{\omega})}\big(\phi_*^T\phi_*\big)\mathbf{m}_j, \end{aligned}$$

and for $i = j$ between 1 and D,

$$\begin{aligned} \left(E_{q(\mathbf{A})q(\boldsymbol{\omega})}\big(\mathbf{A}^T\phi_*^T\phi_*\mathbf{A}\big)\right)_{i,i} &= E_{q(\mathbf{A})q(\boldsymbol{\omega})}\big(\mathbf{a}_i^T\phi_*^T\phi_*\mathbf{a}_i\big) \\ &= \mathbf{m}_i^T E_{q(\boldsymbol{\omega})}\big(\phi_*^T\phi_*\big)\mathbf{m}_i \\ &\quad + \mathrm{tr}\Big(E_{q(\boldsymbol{\omega})}\big(\phi_*^T\phi_*\big) \cdot \mathbf{s}_i\Big). \end{aligned}$$

Taking the difference between the raw second moment and the outer product of the mean we get that the variance of the predictive distribution is given by

$$\begin{aligned} &\mathrm{Var}_{q(\mathbf{y}^*|\mathbf{x}^*)}(\mathbf{y}^*) = \tau^{-1}\mathbf{I}_D + \Psi \\ &\quad + \mathbf{M}^T\big(E_{q(\boldsymbol{\omega})}\big(\phi_*^T\phi_*\big) - E_{q(\boldsymbol{\omega})}\big(\phi_*\big)^T E_{q(\boldsymbol{\omega})}\big(\phi_*\big)\big)\mathbf{M} \end{aligned}$$

with $\Psi_{i,j} = \mathrm{tr}\big(E_{q(\boldsymbol{\omega})}\big(\phi_*^T\phi_*\big) \cdot \mathbf{s}_i\big) \cdot \mathbb{1}[i=j]$. □

Discussion 1. We discuss some of the key properties of the VSSGP, fVSSGP, and sfVSSGP. Due to space constraints, this discussion was moved to the appendix.

Unlike the sparse pseudo-input approximation, where the variational uncertainty is over the locations of a sparse set of inducing points in the output space, the uncertainty in our approximation is over a sparse set of function frequencies. As the uncertainty over a frequency (Σ_k) grows, the exponential decay term in the expectation of Φ decreases, and the expected magnitude of the feature ($[(E_{q(\boldsymbol{\omega})}(\Phi))_{n,k}]_{n=1}^N$) tends to zero for points $\mathbf{x}_n$ far from $\mathbf{z}_k$. Conversely, as the uncertainty over a frequency decreases, the exponential decay term increases towards one, and the expected magnitude of the feature does not diminish for points $\mathbf{x}_n$ far from $\mathbf{z}_k$.

With the predictive uncertainty in equation 15 we preserve many of the GP characteristics. As an example, consider the SE covariance function[1]. In full GPs the variance increases towards $\sigma^2 + \tau^{-1}$ far away from the data. This property is key to Bayesian optimisation for example where this uncertainty is used to decide what action to take given a GP posterior.

With the SE covariance function, our expression for ϕ_* contains an exponential decay term $\exp(-\frac{1}{2}(\mathbf{x}_n - \mathbf{z}_k)^T\Sigma_k(\mathbf{x}_n - \mathbf{z}_k))$. This term tends to zero as $\mathbf{x}_n$ diverges from $\mathbf{z}_k$. For $\mathbf{x}_n$ far away from $\mathbf{z}_k$ for all k we get that the entire matrix Φ tends to zero, and that $E_{q(\boldsymbol{\omega})}\big(\phi_*^T\phi_*\big)$ tends to $\frac{\sigma^2}{K}\mathbf{I}_k$.

For fVSSGP, equation 15 then collapses to

$$\mathrm{Var}_{q(\mathbf{y}^*|\mathbf{x}^*)}(\mathbf{y}^*) = \tau^{-1}\mathbf{I}_D + \Psi'$$

with $\Psi'_{i,j} = \sigma^2\frac{1}{K}\sum_{k=1}^K(\mu_{ik}\mu_{jk} + s_{ik}^2\mathbb{1}[i=j])$.

This term leads to identical predictive variance to that of the full GP when $\mathbf{A}$ is fixed and follows the prior. It is larger than the predictive variance of a full GP when $s_{di}^2 > 1 - \mu_{di}^2$ on average, and smaller otherwise.

Unlike the SE GP, the predictive mean in the VSSGP with a SE covariance function does not tend to zero quickly far from the data. This is because the model can have high confidence in some frequencies, driving the inducing frequency variances (Σ_k) to zero. This in turn requires $\mathbf{x}_n - \mathbf{z}_k$ to be much larger for the exponential decay term to tend to zero. The frequencies the model is confident about will be used far from the data as well.

Unlike the SSGP, the approximation presented here is not periodic. This is one of the theoretical limitations of the sparse spectrum approximation (although in practice the period was observed to often be larger than the range of the

[1] Given by $\sigma^2 \exp\big(-\frac{1}{2}\sum_{q=1}^Q \frac{(x_q - y_q)^2}{l_q^2}\big)$

data). The limitation arises from the fact that the covariance is represented as a weighted sum of cosines in SSGP. In the approximation we present here this is avoided by decaying the cosines to zero.

It is interesting to note that although our approximated covariance function $K(\mathbf{x}, \mathbf{y})$ has to be stationary (i.e. it can be represented as $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{x} - \mathbf{y})$), the approximate posterior is not. This is because stationarity entails that for all $\mathbf{x}$ it must hold that $K(\mathbf{x}, \mathbf{x}) = K(\mathbf{x} - \mathbf{x}) = K(\mathbf{0})$. But for $E_{q(\boldsymbol{\omega})}(\widehat{\mathbf{K}}(\mathbf{X}, \mathbf{X})) = E_{q(\boldsymbol{\omega})}(\Phi\Phi^T)$ we have that the diagonal terms depend on $\mathbf{x}$:

$$\left(E_{q(\boldsymbol{\omega})}(\Phi\Phi^T)\right)_{n,n} = \sum_{k=1}^{K} \frac{2\sigma_i^2}{K} e^{-\overline{\mathbf{x}}_{nk}^T \Sigma_k \overline{\mathbf{x}}_{nk}} \cdot E_{q(b_k)}\big(\cos(\mu_k^T \overline{\mathbf{x}}_{nk}) + \overline{b}_{nk})\big)^2.$$

This is in comparison to the SSGP approximation, where the approximate model is stationary.

It is also interesting to note that the lower bound in equation 10 is equivalent to that of equation 11 for $\mathbf{s}_d$ non-diagonal. For $\mathbf{s}_d$ diagonal the lower bound is looser, but offers improved time complexity.

The use of the factorised lower bound allows us to save on the expensive computation of $\mathbf{A}$ for small updates of $\boldsymbol{\omega}$. Intuitively, this is because small updates in $\boldsymbol{\omega}$ would result in small updates to $\mathbf{A}$. Thus solving for $\mathbf{A}$ analytically at every time point without re-using previous computations is very wasteful. Optimising over $\mathbf{A}$ to solve the linear system of equations (given $\boldsymbol{\omega}$) allows us to use optimal $\mathbf{A}$ from previous steps, adapting it accordingly.

Even though it is possible to analytically integrate over $\mathbf{A}$, we can't analytically integrate $\boldsymbol{\omega}$. This is because $\boldsymbol{\omega}$ appears inside a cosine inside an exponent in equation 2.

Finally, we can approximate our approach to achieve a much more scalable implementation by only using the K' nearest inducing inputs for each data point. This is following the observation that for short length-scales and large Σ, the features will decay to zero exponentially fast with the distance of the data points from the inducing inputs.