A Some derivations for Bayesian Kernel Embedding

A.1 Notation

Consider a dataset $x_1, \ldots, x_n \in \mathbb{R}^D$ and suppose that there exists some unknown probability distribution P for which the x_i are i.i.d.:

$$x_i \sim \mathsf{P} \ . \tag{18}$$

Denote by μ_θ the RKHS mean embedding element for a given kernel $k_\theta(\cdot,\cdot)$ with hyperparameter $\theta \in \mathbb{R}^Q$ and by $\widehat{\mu_\theta}(\cdot)$ the empirical mean embedding

$$\widehat{\mu_\theta}(\cdot) := \frac{1}{n}\sum_{i=1}^{n} k_\theta(x_i, \cdot) \ . \tag{19}$$

We posit as our model that μ_θ has a GP prior with covariance r_θ, where

$$r_\theta(x, y) = \int k_\theta(x, u) k_\theta(u, y) \nu(du) \ ,$$

where ν is a finite measure on $\mathbb{R}^D$ thus ensuring that $\mu_\theta \in \mathcal{H}_{k_\theta}$ when drawn from the prior

$$\mu_\theta | \theta \sim \mathcal{GP}(0, r_\theta(\cdot,\cdot)) \ . \tag{20}$$

In addition, we model the link between the population mean embedding and the empirical mean embedding functions at a given location x as follows

$$p(\widehat{\mu_\theta}(x) | \mu_\theta(x)) = \mathcal{N}(\widehat{\mu_\theta}(x); \mu_\theta(x), \tau^2/n) \tag{21}$$

where τ^2 is another hyperparameter.

A.2 Priors over RKHS

The results in this section have appeared in the literature before, but as they are not well known or collected in one place, we have included them for completeness. A similar discussion appears in Pillai et al. (2007), but without the construction of explicit GP priors over the RKHSs which we provide below.

It is well known that the sample paths of a GP with kernel k are almost surely outside RKHS $\mathcal{H}_k$, the result known as Kallianpur's 0-1 law (Kallianpur, 1970; Wahba, 1990). It is easiest to demonstrate this by considering a Mercer's expansion Rasmussen and Williams (2006, Section 4.3) of kernel k given by

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x'), \tag{22}$$

for the eigenvalue-eigenfunction pairs $\{(\lambda_i, e_i)\}_{i=1}^n$. Then, a representation of $f \sim \mathcal{GP}(0, k)$ is given by $f = \sum_{i=1}^{\infty} \sqrt{\lambda_i} Z_i e_i$, where $\{Z_i\}_{i=1}^{\infty}$ are independent and identically distributed standard normal random variables. However,

$$\|f\|_{\mathcal{H}_k}^2 = \sum_{i=1}^{\infty} \frac{\lambda_i Z_i^2}{\lambda_i} = \sum_{i=1}^{\infty} Z_i^2 = \infty, \quad a.s. \tag{23}$$

so $f \notin \mathcal{H}_k$ almost surely. This issue is often sidelined in the literature, cf. e.g. (Rasmussen and Williams, 2006, Section 6.1) – in GP regression, it is not necessary to ensure that the prior on the regression function is supported on $\mathcal{H}_k$ (the posterior mean will still lie in $\mathcal{H}_k$, however). However, since the object of our interest, kernel embedding, is by construction an element of $\mathcal{H}_k$ - we opt for an approach where the prior is indeed specified over the correct space. Fortunately, it is straightforward to construct a kernel r such that the realizations from a GP with kernel r are almost surely inside RKHS $\mathcal{H}_k$. For this, we will need notions of dominance and nuclear dominance for kernel functions.

Definition 1. *Kernel k is said to dominate kernel r (written $k \succ r$) if $\mathcal{H}_r \subseteq \mathcal{H}_k$.*

Lukić and Beder (2001, Theorem 1.1) characterise dominance $k \succ r$ via the existence of a certain positive, continuous and self-adjoint operator $L : \mathcal{H}_k \to \mathcal{H}_k$ for which

$$r(x, x') = \langle L[k(\cdot, x)], k(\cdot, x') \rangle_{\mathcal{H}_k}, \qquad \forall x, x' \in \mathcal{X}. \tag{24}$$

When L is also a trace class operator, dominance is termed *nuclear*, and denoted $k \succ\succ r$. The following theorem from Lukić and Beder (2001, Theorem 7.2) then fully characterises kernels that lead to valid GP priors over RKHS $\mathcal{H}_k$.

Theorem 1. *Let $\mathcal{H}_k$ be separable and let $m \in \mathcal{H}_k$. Then $\mathcal{GP}(0, r(\cdot,\cdot))$ has trajectories in $\mathcal{H}_k$ with probability 1 if and only if $k \succ\succ r$.*

Thus, we just need to specify a trace-class, positive, continuous and self-adjoint operator $L : \mathcal{H}_k \to \mathcal{H}_k$ and compute $\langle L[k(\cdot,x)], k(\cdot,x')\rangle_{\mathcal{H}_k}$. A convenient choice for a given bounded continuous kernel k can be defined as follows. Take the convolution operator $S_k : L^2(\mathcal{X};\nu) \to \mathcal{H}_k$ with respect to a finite measure ν, defined as

$$[S_k f](x) = \int f(u)k(x,u)\nu(du). \tag{25}$$

It is well known that the adjoint of S_k is the inclusion of $\mathcal{H}_k$ into L^2 (Steinwart and Christmann, 2008, Section 4.3). Thus, we let $L = S_k S_k^*$, which is the (uncentred) covariance operator $L = \int k(\cdot,u) \otimes k(\cdot,u)\nu(du)$ of ν. As a covariance operator, L is then positive, continuous and self-adjoint. It is also trace-class in most cases of interest – and in particular whenever $\int k(u,u)\nu(du) < \infty$ (Steinwart and Christmann, 2008, Theorem 4.27), and thus for every stationary kernel provided that ν is a finite measure. This leads to

$$\begin{aligned} r(x,x') &= \langle S_k S_k^*[k(\cdot,x)], k(\cdot,x')\rangle_{\mathcal{H}_k} \\ &= \langle S_k^*[k(\cdot,x)], S_k^* k(\cdot,x')\rangle_{L^2(\mathcal{X};\nu)} \\ &= \int k(x,u)k(u,x')\nu(du), \end{aligned}$$

so r can be simply computed as a convolution of k with itself, and we can use $\mathcal{GP}(0, r(\cdot,\cdot))$ as a prior over $\mathcal{H}_k$.

A.3 Covariance function r_θ

In this subsection, we derive the covariance function r_θ for squared exponential kernels. Consider a squared exponential kernel on $\mathcal{X} = \mathbb{R}^D$ with full covariance matrix Σ_θ defined by

$$k_\theta(x,y) = \exp\left(-\frac{1}{2}(x-y)^T\Sigma_\theta^{-1}(x-y)\right), \quad x,y \in \mathbb{R}^D. \tag{26}$$

While we have required in A.2 that ν is a finite measure for the covariance operator to be trace class when working with stationary kernels, let us for simplicity first consider the instructive case when ν is the Lebesgue measure. Then, we have

$$\begin{aligned} r_\theta(x,y) &= \int k_\theta(x,u)k_\theta(u,y)du \\ &= \int \exp\left(-\frac{1}{2}\left((x-u)^T\Sigma_\theta^{-1}(x-u) + (y-u)^T\Sigma_\theta^{-1}(y-u)\right)\right)du \end{aligned}$$

Note that

$$(x-u)^T\Sigma_\theta^{-1}(x-u) + (y-u)^T\Sigma_\theta^{-1}(y-u) = 2\left(u - \frac{x+y}{2}\right)^T\Sigma_\theta^{-1}\left(u - \frac{x+y}{2}\right) + \frac{1}{2}(x-y)^T\Sigma_\theta^{-1}(x-y)\ .$$

Then

$$\begin{aligned} r_\theta(x,y) &= \exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y)\right)\int \exp\left(-\frac{1}{2}\left(u - \frac{x+y}{2}\right)^T\left(\frac{1}{2}\Sigma_\theta\right)^{-1}\left(u - \frac{x+y}{2}\right)\right)du \\ &= \exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y)\right) \times (2\pi)^{D/2}|\Sigma_\theta/2|^{1/2} \\ &= \pi^{D/2}\ |\Sigma_\theta|^{1/2}\ \exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y)\right)\ . \end{aligned}$$

Thus r_θ is proportional to another squared exponential kernel with covariance $2\Sigma_\theta$. For the special case where the covariance matrix Σ_θ is diagonal – let $\Sigma_\theta = \theta I_D$ and $\theta = (\theta^{(1)}, \ldots, \theta^{(D)})^T$– we have

$$r_\theta(x,y) = \pi^{D/2}\ \left(\prod_{d=1}^D \theta^{(d)}\right)^{1/2} \exp\left(-\frac{1}{2}(x-y)^T(2\theta I_D)^{-1}(x-y)\right). \tag{27}$$

Now, take $\nu(du) = \exp\left(-\frac{\|u\|_2^2}{2\eta^2}\right) du$, i.e., ν is a finite measure and is proportional to a Gaussian measure on $\mathbb{R}^d$. In that case, we have

$$\begin{aligned} r_\theta(x,y) &= \int k_\theta(x,u)k_\theta(u,y)\nu(du) \\ &= \int \exp\left(-\frac{1}{2}\underbrace{\left((x-u)^T\Sigma_\theta^{-1}(x-u) + (y-u)^T\Sigma_\theta^{-1}(y-u) + \eta^{-2}u^\top u\right)}_{\mathsf{A}}\right) du. \end{aligned}$$

From standard Gaussian integration rules, it follows that

$$\mathsf{A} = \frac{1}{2}(x-y)^T\Sigma_\theta^{-1}(x-y) + (u-\mathsf{m})^\top \mathsf{S}^{-1}(u-\mathsf{m}) + \left(\frac{x+y}{2}\right)^\top \left(\frac{1}{2}\Sigma_\theta + \eta^2 I_D\right)^{-1}\left(\frac{x+y}{2}\right)$$

where $\mathsf{m} = \mathsf{S}^{-1}\Sigma_\theta^{-1}(x+y)$ and $\mathsf{S} = (2\Sigma_\theta^{-1} + \eta^{-2}I_D)^{-1}$. Therefore

$$\begin{aligned} r_\theta(x,y) &= (2\pi)^{D/2}|\mathsf{S}|^{1/2}\exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y) - \frac{1}{2}\left(\frac{x+y}{2}\right)^\top\left(\frac{1}{2}\Sigma_\theta + \eta^2 I_D\right)^{-1}\left(\frac{x+y}{2}\right)\right) \\ &= (2\pi)^{D/2}\left|2\Sigma_\theta^{-1} + \eta^{-2}I_D\right|^{-1/2}\exp\left(-\frac{1}{2}(x-y)^T(2\Sigma_\theta)^{-1}(x-y)\right) \\ &\qquad \times \exp\left(-\frac{1}{2}\left(\frac{x+y}{2}\right)^\top\left(\frac{1}{2}\Sigma_\theta + \eta^2 I_D\right)^{-1}\left(\frac{x+y}{2}\right)\right). \end{aligned}$$

Thus, we see that r_θ has a nonstationary component that penalises the norm of $\left(\frac{x+y}{2}\right)$. This is reminiscent of the well known locally stationary covariance functions (Silverman, 1957). However, for large values of η, the nonstationary component becomes negligible and r_θ reverts to being proportional to a standard squared exponential kernel with covariance $2\Sigma_\theta$, just like in the case of Lebesgue measure. We note that any choice of $\eta > 0$ gives a valid prior over $\mathcal{H}_k$. Treating η as another hyperparameter to be learned would be an interesting direction for future research.

A.4 Fast computation of the marginal pseudolikelihood

The marginal pseudolikelihood in Eq. (15) requires computation of the likelihood for an mn-dimensional normal distribution

$$\mathcal{N}\left(\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\}; \mathbf{0},\ \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}\right).$$

However, the Kronecker product structure in the covariance matrix $C = \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}$ allows efficient computation. We denote with $R_{\theta,\mathbf{zz}} = Q\Lambda Q^\top$ the eigendecomposition of the matrix $R_{\theta,\mathbf{zz}}$ with $\Lambda = \text{diag}\left[\lambda_1, \ldots, \lambda_m\right]$. Note that $\mathbf{1}_n\mathbf{1}_n^\top$ is a rank-one matrix with the eigenvalue equal to n. Therefore C has top m eigenvalues equal to $n\lambda_i + \tau^2$, $i = 1, \ldots, m$, and the remaining $n(m-1)$ all equal to τ^2. Thus, the log-determinant is simply

$$\log\det C = \sum_{i=1}^m \log(n\lambda_i + \tau^2) + m(n-1)\log\tau^2 = \log\det\left[R_{\theta,\mathbf{zz}} + (\tau^2/n)I_m\right] + m\log n + m(n-1)\log\tau^2. \quad (28)$$

Further, we need to compute $\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\}^\top C^{-1}\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\}$. By completing $b_1 = n^{-1/2}\mathbf{1}_n$ to an orthonormal basis $\{b_1, \ldots, b_n\}$ of $\mathbb{R}^n$ and forming the corresponding matrix $B = [b_1 \cdots b_n]$, and denoting by n an $n \times n$ matrix with $\mathsf{n}_{11} = n$ and $\mathsf{n}_{ij} = 0$ elsewhere, we have that

$$C^{-1} = (B \otimes Q)(\mathsf{n} \otimes \Lambda + \tau^2 I_{nm})^{-1}(B \otimes Q)^\top. \quad (29)$$

We now simply need to apply Kronecker identity $(B^\top \otimes Q^\top)\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\} = \text{vec}\left\{Q^\top K_{\theta,\mathbf{zx}} B\right\}$, to obtain

$$\begin{aligned} &\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\}^\top C^{-1}\text{vec}\left\{K_{\theta,\mathbf{zx}}\right\} = \text{vec}\left\{Q^\top K_{\theta,\mathbf{zx}}B\right\}^\top (\mathsf{n} \otimes \Lambda + \tau^2 I_{nm})^{-1}\text{vec}\left\{Q^\top K_{\theta,\mathbf{zx}}B\right\} \\ &= \sum_{j=1}^m \frac{n^{-1}\left[Q^\top K_{\theta,\mathbf{zx}}\mathbf{1}_n\right]_j^2}{n\lambda_j + \tau^2} + \frac{1}{\tau^2}\sum_{i=2}^n\sum_{j=1}^m \left[Q^\top K_{\theta,\mathbf{zx}} b_i\right]_j^2. \end{aligned} \quad (30)$$

For the first term, we have

$$\sum_{j=1}^{m} \frac{n^{-1}\left[Q^\top K_{\theta,\mathbf{zx}}\mathbf{1}_n\right]_j^2}{n\lambda_j+\tau^2} = \sum_{j=1}^{m} \frac{\left[Q^\top \hat{\mu}(\mathbf{z})\right]_j^2}{\lambda_j+\tau^2/n} = \sum_{j=1}^{m} \frac{\mathrm{Tr}\left[\hat{\mu}(\mathbf{z})\hat{\mu}(\mathbf{z})^\top q_j q_j^\top\right]}{\lambda_j+\tau^2/n}$$
$$= \hat{\mu}(\mathbf{z})^\top \left(R_{\theta,\mathbf{zz}} + (\tau^2/n)I_m\right)^{-1} \hat{\mu}(\mathbf{z}). \tag{31}$$

And for the second term:

$$\begin{aligned}
\frac{1}{\tau^2}\sum_{i=2}^{n}\sum_{j=1}^{m}\left[Q^\top K_{\theta,\mathbf{zx}} b_i\right]_j^2 &= \frac{1}{\tau^2}\sum_{j=1}^{m}\sum_{i=2}^{n}\left[q_j^\top K_{\theta,\mathbf{zx}} b_i\right]^2 \\
&= \frac{1}{\tau^2}\sum_{j=1}^{m}\left\{\left\|K_{\theta,\mathbf{xz}} q_j\right\|^2 - n\left(q_j^\top \hat{\mu}(\mathbf{z})\right)^2\right\} \\
&= \frac{1}{\tau^2}\left\|K_{\theta,\mathbf{xz}}\right\|_F^2 - \frac{n}{\tau^2}\left\|\hat{\mu}\left(\mathbf{z}\right)\right\|^2 .
\end{aligned} \tag{32}$$

Altogether, the log-likehood is given by

$$\begin{aligned}
\log\left\{\mathcal{N}\left(\mathrm{vec}\left\{K_{\theta,\mathbf{zx}}\right\}; \mathbf{0},\ \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}\right)\right\} &= -\frac{1}{2}\Bigg\{\log\det\left[R_{\theta,\mathbf{zz}} + (\tau^2/n)I_m\right] \\
&\quad + \hat{\mu}(\mathbf{z})^\top\left(R_{\theta,\mathbf{zz}} + (\tau^2/n)I_m\right)^{-1}\hat{\mu}(\mathbf{z}) \\
&\quad + \frac{1}{\tau^2}\left\|K_{\theta,\mathbf{xz}}\right\|_F^2 - \frac{n}{\tau^2}\left\|\hat{\mu}\left(\mathbf{z}\right)\right\|^2 \\
&\quad + m\log n + m(n-1)\log\tau^2 + mn\log(2\pi)\Bigg\}.
\end{aligned} \tag{33}$$

B Source for Stan model

```
functions {
  // phi should be m x n
  real kron_multi_normal(matrix K,matrix R,matrix Q1,vector e1,int m,int n,real sigma2) {
    vector[m*n] e;
    matrix[m,m] Q2;
    vector[m] e2;
    vector[m] ones;
    vector[m*n] mv2;
    real mvp;
    real logdet;
    Q2 <- eigenvectors_sym(R);
    e2 <- eigenvalues_sym(R);
    for(j in 1:m) {
      ones[j] <- 1;
      for(i in 1:n)
        e[(j-1)*n + i] <- 1/(e1[i] * e2[j] + sigma2);
    }
    mv2 <- to_vector((transpose(Q2) * transpose(K)) * Q1);
    mvp <- sum(mv2 .* e .* mv2);
    logdet <- sum(log(e2 .* (ones * n) + ones * sigma2)) + m * (n-1) * log(sigma2);

    return( - .5 * logdet - .5 * mvp);
  }
}

data {
  int<lower=1> n;
  int<lower=1> m;
  vector[n] x;
```

```
  vector[m] u;
}

transformed data {
  matrix[n,m] xu_dist2;
  matrix[m,m] u_dist2;
  matrix[n,n] ones;
  vector[n] zeros;
  matrix[n,n] Q1;
  vector[n] e1;

  for (i in 1:n) {
    zeros[i] <- 0;
    e1[i] <- 0;
    for (j in 1:n)
      ones[i,j] <- 1;
    for(j in 1:m)
      xu_dist2[i, j] <- square(x[i] - u[j]);
  }
  for(i in 1:m) {
    for(j in 1:m)
      u_dist2[i,j] <- square(u[i] - u[j]);
  }
  e1[1] <- n;
  Q1 <- eigenvectors_sym(ones);
}

parameters {
  real<lower=0> lengthscale;
  real<lower=0> sigma2;
}
transformed parameters {
  matrix[m,m] R;
  matrix[n,m] J;
  matrix[n,m] K;

// R <- lengthscale * sqrt(pi()) *
  R <- exp(- u_dist2/(4*lengthscale^2));
  K <- exp(- xu_dist2/(2*lengthscale^2));
  J <- K .* K .* xu_dist2 / lengthscale^4;
}

model {
  for(i in 1:n) // Jacobian
    increment_log_prob(log(.5 * sum(J[i])));

  increment_log_prob(kron_multi_normal(K, R, Q1, e1, m, n, sigma2));
  lengthscale ~ gamma(1,1);
  sigma2 ~ gamma(1,1);
}
```