Appendix A. Compositional Kernel Search Algorithm

Algorithm 1: Compositional Kernel Search Algorithm

Input: data $x_1, \ldots, x_n \in \mathbb{R}^D, y_1, \ldots, y_n \in \mathbb{R}$,base kernel set $\mathcal{B}$
Output: k, the resulting kernel
For each base kernel on each dimension, fit GP to data (i.e. optimise hyperparams by ML-II) and set k to be kernel with smallest BIC.
for *depth=1:T (either fix T or repeat until BIC no longer decreases)* **do**
  Fit GP to following kernels and set k to be the one with lowest BIC:
  (1) All kernels of form $k + B$ where B is any base kernel on any dimension
  (2) All kernels of form $k \times B$ where B is any base kernel on any dimension
  (3) All kernels where a base kernel in k is replaced by another base kernel
end

Appendix B. Base Kernels

$$\begin{aligned}
\mathrm{LIN}(x, x') &= \sigma^2 (x - l)(x' - l) \\
\mathrm{SE}(x, x') &= \sigma^2 \exp\Big(-\frac{(x - x')^2}{2l^2} \Big) \\
\mathrm{PER}(x, x') &= \sigma^2 \exp\Big(-\frac{2\sin^2(\pi(x - x')/p)}{l^2} \Big) \\
\mathrm{RQ}(x, x') &= \sigma^2 \Big(1 + \frac{(x - x')^2}{2\alpha l^2} \Big)^{-\alpha}
\end{aligned}$$

Appendix C. Random Fourier Features for Sums and Products of Kernels

Theorem 1 (Bochner's Theorem (Rudin, 1964)) *A stationary kernel k(d) is positive definite if and only if k(d) is the Fourier transform of a non-negative measure.*

For RFF the kernel can be approximated by the inner product of random features given by samples from its spectral density, in a Monte Carlo approximation, as follows:

$$\begin{aligned}
k(x - y) = \int_{\mathbb{R}^D} e^{iv^T(x-y)} d\mathbb{P}(v) \propto \int_{\mathbb{R}^D} p(v) e^{iv^T(x-y)} dv &= \mathbb{E}_{p(v)}[e^{iv^T x}(e^{iv^T y})^*] \\
&= \mathbb{E}_{p(v)}[Re(e^{iv^T x}(e^{iv^T y})^*)] \\
&\approx \frac{1}{m} \sum_{k=1}^{m} Re(e^{iv_k{}^T x}(e^{iv_k{}^T y})^*) \\
&= \phi(x)^T \phi(y)
\end{aligned}$$

where $\phi(x) = \sqrt{\frac{1}{m}}(cos(v_1{}^T x + b_1), \ldots, cos(v_m{}^T x + b_m))$ with spectral frequencies v_k iid samples from $p(v)$ and b_k iid samples from $U[0, 2\pi]$.
Let k_1, k_2 be two stationary kernels, with respective spectral densities p_1, p_2 so that

$k_1(d) = a_1\hat{p_1}(d), k_2(d) = a_2\hat{p_2}(d)$, where $\hat{p}(d) := \int_{\mathbb{R}^D} p(v)e^{iv^T d}dv$. We use this convention as the Fourier transform. Note $a_i = k_i(0)$.

$$(k_1 + k_2)(d) = a_1 \int p_1(v)e^{iv^T d}dv + a_2 \int p_2(v)e^{iv^T d}dv = (a_1 + a_2)\hat{p_+}(d)$$

where $p_+(v) = \frac{a_1}{a_1+a_2}p_1(v) + \frac{a_2}{a_1+a_2}p_2(v)$, a mixture of p_1 and p_2. So to generate RFF for $k_1 + k_2$, generate $v \sim p_+$ by generating $v \sim p_1$ w.p. $\frac{a_1}{a_1+a_2}$ and $v \sim p_2$ w.p. $\frac{a_2}{a_1+a_2}$
Now for the product, suppose

$$(k_1 \cdot k_2)(d) = a_1 a_2 \hat{p_1}(d)\hat{p_2}(d) = a_1 a_2 \hat{p_*}(d)$$

Then $p_*(d)$ is the inverse fourier transform of $\hat{p_1}\hat{p_2}$, which is the convolution $p_1 * p_2(d) := \int_{\mathbb{R}^D} p_1(z)p_2(d-z)dz$. So to generate RFF for $k_1 \cdot k_2$, generate $v \sim p_*$ by generating $v_1 \sim p_1, v_2 \sim p_2$ and setting $v = v_1 + v_2$.
This is not applicable for non-stationary kernels, such as the linear kernel. We show how this is dealt with in Appendix D.

Appendix D. Random Features for Sums and Products of Kernels

Suppose ϕ_1, ϕ_2 are random features such that
$k_1(x, x') = \phi_1(x)^T\phi_1(x'), \phi_2(x)^T\phi_2(x'), \phi_i : \mathbb{R}^D \to \mathbb{R}^m$.
It is straightforward to verify that

$$(k_1 + k_2)(x, x') = \phi_+(x)^T\phi_+(x') \text{ where } \phi_+(\cdot) = (\phi_1(\cdot)^T, \phi_2(\cdot)^T)^T$$
$$(k_1 \cdot k_2)(x, x') = \phi_*(x)^T\phi_*(x') \text{ where } \phi_*(\cdot) = \phi_1(\cdot) \otimes \phi_2(\cdot)$$

However we do not want the number of features to grow as we add or multiply kernels, since it will grow exponentially. We want to keep it to be m features. So we subsample m entries from ϕ_+ (or ϕ_*) and scale by factor $\sqrt{2}$ ($\sqrt{m}$ for ϕ_*), which will still give us unbiased estimates of the kernel provided that each term of the inner product $\phi_+(x)^T\phi_+(x')$ (or $\phi_*(x)^T\phi_*(x')$) is an unbiased estimate of $(k_1 + k_2)(x, x')$(or $(k_1 \cdot k_2)(x, x')$).
This is how we generate random features for linear kernels combined with other stationary kernels, using the features $\phi(x) = \frac{\sigma}{\sqrt{m}}(x - l, \ldots, x - l)^T$.

Appendix E. Spectral Densities for RQ and PER

From Solin and Sarkka (2014), we have that the spectral density of the RQ kernel is:

$$p(v) = l\sqrt{\frac{2\alpha}{\pi}}\left(l|v|\sqrt{\frac{\alpha}{2}}\right)^{\alpha-\frac{1}{2}} K_{\alpha-\frac{1}{2}}(l|v|\sqrt{2\alpha})/\Gamma(\alpha)$$

where K is the modified Bessel function of the second kind.
From Solin and Särkkä (2014), we have that the spectral density of the PER kernel is:

$$\sum_{n=-\infty}^{\infty} \frac{I_n(l^{-2})}{\exp(l^{-2})}\delta\left(v - \frac{2\pi n}{p}\right)$$

where I is the modified Bessel function of the first kind.

Appendix F. Matrix Identities

Lemma 2 (Woodbury's Matrix Inversion Lemma)
$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}$

Lemma 3 (Sylverster's Determinant Theorem)
$\det(I + AB) = \det(I + BA)\ \forall A \in \mathbb{R}^{m\times n} \forall B \in \mathbb{R}^{n\times m}$