8 Nov 2021 / 14 Nov 2021 / 15 Nov 2021

Chap 6: Kernel Methods

1. Motivation

(1) Kernels are real-valued binary functions on the input space, i.e., functions of type $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In the context of machine learning, they are used to express the similarity of two inputs. Their value on (x, x') becomes larger as x and x' become more similar from the perspective of a given classification or regression task. We will consider kernels that satisfy two properties, called symmetry and positive definiteness, which ensure that kernels indeed behave like similarity measures.

(2) Why do we study kernels? There are two reasons for this. First, multiple ML algorithms, such as SVM, can be generalised or modified to work with kernels. This means that all the information about the inputs used by such a modified algorithm comes from the application of a kernel on those inputs, both during the optimisation for training and during the prediction on a new input. This kernelisation dramatically increases the expressiveness of a hypothesis set or sets used by the algorithm, in particular, making the set include non-linear functions even when the algorithm is originally designed only for linear functions.

Second, kernels have a beautiful mathematical theory, and can be used to analyse ML algorithms in general.

For instance, the recent theoretical analysis of infinite-width neural networks uses kernels induced by those networks and proves non-trivial facts about neural networks, such as the identification of a case when the network training achieves zero loss eventually.

(3) We will focus on the first reason. If you are interested in the second, I encourage you to search for articles containing phrases like: neural network Gaussian process (NNGP) and neural tangent kernel (NTK).

2 Positive definite symmetric kernel

(1) $\mathcal{X}$ ... space for inputs.

Def. A kernel is a function k of type $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

A kernel k is positive definite symmetric (PDS) if:

(i) $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$; and

(ii) for all $m \in \mathbb{N}$ and all $x_1, \dots, x_m \in \mathcal{X}$, $c_1, \dots, c_m \in \mathbb{R}$,

$$\sum_{i,j=1}^{m} c_i c_j \, k(x_i, x_j) \geq 0.$$

(when k is PDS.)

(1) A good intuition is that $k(x, x')$ computes the inner product after embedding x and x' into an inner product space. That is, $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. The function Φ here is a feature map converting each input to a feature vector, and $\langle \cdot, \cdot \rangle$ measure similarity of two feature vectors. Such a k satisfies (i) and (ii).

② The two conditions (i) and (ii) can be expressed equivalently via kernel matrix or Gram matrix K associated to k and $S = (x_1, \dots, x_m)$:

$$K \in \mathbb{R}^{m \times m}, \quad (K)_{ij} = k(x_i, x_j).$$

That is, (i) and (ii) $\Leftrightarrow$ K is symmetric and positive semidefinite (SPSD)

("pos. semidefinite" means that for all $\vec{c} \in \mathbb{R}^m$, $(\vec{c})^T K \vec{c} \geq 0$.)

$\Leftrightarrow$ K is symmetric and has non-negative real eigenvalues.

We will use these equivalences. In particular, we will use the decomposition of K into $U \Sigma U^T$ where the columns of $U \in \mathbb{R}^{m \times m}$ are orthonormal eigenvectors and Σ is a diagonal matrix of non-negative eigenvalues

③ Don't be confused about somewhat inconsistent use of: "positive definite" (for kernels) and "positive semidefinite" (for matrices)

The authors of the textbook say that we are following the official terminology.

(2). Examples.

① Polynomial kernels. $c > 0$, $d \in \mathbb{N}$.

A polynomial kernel of degree d is a kernel k over $\mathbb{R}^N$ that has the following form:

$$k(x, x') = (\langle x, x' \rangle + c)^d \quad \text{for some } c \geq 0 \text{ (non-negative)}.$$

Can we express k as $\langle \Phi(\cdot), \Phi(\cdot) \rangle$? Yes.

Consider $d=2$, $N=2$ case. Define $\Phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2c}x_1, \sqrt{2c}x_2, c)$. $(x_1, x_2) \in \mathbb{R}^2$

Then, $\langle \Phi(x), \Phi(x') \rangle = x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2x_1x_1'x_2x_2' + 2c x_1x_1' + 2c x_2x_2' + c^2$.

$= (x_1x_1' + x_2x_2' + c)^2 = (\langle x, x' \rangle + c)^2$.

So, k computes the inner product in the feature space of degree ≤ 2 monomials. As we will see later, the SVM with k computes a linear classifier (or hypo.) on this feature space, so that from the view on the original input space $\mathbb{R}^2$, the result of the SVM is a non-linear classifier.

② Gaussian kernel or radial basis kernel (RBF). k

$k : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$

$k(x, x') = \exp\left(- \frac{\|x - x'\|^2}{2\sigma^2} \right)$. for some fixed $\sigma > 0$.

This is a very popular kernel.

What similarity does this capture? One informal answer is that k measures similarity using monomials of any degree. This is because k can be expressed as the normalized infinite sum of polynomial kernels. We will come back to this later.

③ Sigmoid kernel. k ..... this has a relationship with neural networks.

$k : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, $k(x, x') = \tanh\left(a\langle x, x' \rangle + b \right)$

$= \frac{e^{a\langle x, x' \rangle + b} - e^{-a\langle x, x' \rangle - b}}{e^{a\langle x, x' \rangle + b} + e^{-a\langle x, x' \rangle - b}}$

for some $a, b \geq 0$.

3. Reproducing Kernel Hilbert Space (RKHS)

(1) When explaining the def'n of a PDS kernel, we mentioned that intuitively such a kernel computes the inner product after embedding its arguments into some inner product space. Actually, this is not an informal statement. Indeed, there is such an inner product space. RKHS is a canonical such space that satisfies an additional requirement for being a Hilbert space (i.e., completeness). We will study the construction of RKHS.

(2) We will start with a lemma that indicates that a PDS k behaves like an inner product.

Lemma 6.7. [Cauchy-Schwartz inequality for PDS kernels]

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ... PDS kernel.

$\Rightarrow$ for all $x, x' \in \mathcal{X}$,

$$k(x,x')^2 \leq k(x,x)\,k(x',x').$$

Proof: Pick $x, x' \in \mathcal{X}$. Construct the Gram matrix K:

$$K = \begin{bmatrix} k(x,x) & k(x,x') \\ k(x,x') & k(x',x') \end{bmatrix}.$$

Then, K has non-negative real eigenvalues. Since the determinant of K is the product of eigenvalues,

$$\det(K) \geq 0.$$

But $\det(K) = k(x,x)k(x',x') - k(x,x')^2$.

$\therefore$ The claimed inequality holds. Q.

(3) we prove the following theorem about the existence of a reproducing kernel Hilbert space $\mathbb{H}$ associated to a PDS k.

Thm 8. $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ... a PDS kernel. → called reproducing kernel Hilbert space (RKHS)

$\Rightarrow$ $\exists$ a Hilbert space $\mathbb{H}$ and a map $\Phi: \mathcal{X} \to \mathbb{H}$ s.t.

(i) $\mathbb{H} \subseteq [\mathcal{X} \to \mathbb{R}]$;

(ii) $\Phi(x) = k(x, -)$ (i.e., $x' \longmapsto k(x, x')$) ;

(iii) $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{H}}$;

(iv) $\forall h \in \mathbb{H}\ \forall x \in \mathcal{X}$ $\quad h(x) = \langle h, \Phi(x) \rangle$.

↓ called reproducing property.

Proof. We first build an inner product space $\mathbb{H}_0$ that satisfies (i)–(iv), and then construct $\mathbb{H}$ out of $\mathbb{H}_0$ via completion (which adds limits of Cauchy sequences in $\mathbb{H}_0$) (→ a Hilbert space) We show only the construction of $\mathbb{H}_0$.

Define $\mathbb{H}_0 \overset{\text{def.}}{=} \operatorname{span}\{ k(x, -) \mid x \in \mathcal{X} \}$.

$$= \left\{ \sum_{i=1}^{m} a_i k(x_i, -) \;\middle|\; m \in \mathbb{N},\ a_i \in \mathbb{R},\ x_i \in \mathcal{X} \text{ for all } i \in [m] \right\}.$$

and $\Phi(x) \overset{\text{def}}{=} k(x, -)$.

Then, $\mathbb{H}_0$ is a vector space over $\mathbb{R}$,

$\mathbb{H}_0 \subseteq [\mathcal{X} \to \mathbb{R}]$, and $\Phi: \mathcal{X} \to \mathbb{H}_0$.

Now define $\langle \cdot, \cdot \rangle: \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ by

$$\left\langle \sum_{i=1}^{m} a_i k(x_i, -),\ \sum_{j=1}^{n} b_j k(x'_j, -) \right\rangle \overset{\text{def}}{=} \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j k(x_i, x'_j)$$

This is well-defined, because if $\sum_{i=1}^{m} a_i k(x_i, -) = \sum_{i=1}^{m'} a'_i k(y_i, -)$, then

$b_j \sum_{i=1}^{m} a_i k(x_i, x'_j) = b_j \sum_{i=1}^{m'} a'_i k(y_i, x'_j)$ for all $j \in [n]$, and so

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j k(x_i, x'_j) = \sum_{j=1}^{n} b_j \sum_{i=1}^{m} a_i k(x_i, x'_j) = \sum_{j=1}^{n} b_j \sum_{i=1}^{m'} a'_i k(y_i, x'_j)$$

$$= \sum_{i=1}^{m'}\sum_{j=1}^{n} a'_i b_j k(y_i, x'_j).$$

(Strictly speaking, we need to show a similar fact for the second argument, which we skip). Also, $\langle \cdot, \cdot \rangle$ is symmetric by its def'n and the symmetry of k, and it is bilinear again by definition.

To prove that $\langle \cdot, \cdot \rangle$ is an inner product, it remains to show that for all $h \in \mathcal{H}_0$,

$$\langle h, h \rangle \geq 0 \quad \text{and if } \langle h, h \rangle = 0, \text{ then } h = 0$$

(constant-zero fn.)

Pick $h \in \mathcal{H}_0$. Then,

$$h = \sum_{i=1}^{m} a_i k(x_i, -) \quad \text{for some } a_i, x_i, m.$$

$$\langle h, h \rangle = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j k(x_i, x_j) \geq 0 \quad \text{because } k \text{ is pos. definite.}$$

Now assume that $\langle h, h \rangle = 0$. Then, for all $x \in \mathcal{X}$,

$$h(x) = \sum_{i=1}^{m} a_i k(x_i, x) = \left\langle \sum_{i=1}^{m} a_i k(x_i, -), \underbrace{\Phi(x)}_{k(x,-)} \right\rangle = \langle h, \Phi(x) \rangle$$

But since $\langle \cdot, \cdot \rangle$ is bilinear and symmetric, and holds, the Cauchy–Schwartz inequality holds for $\langle \cdot, \cdot \rangle$. So,

$$h(x)^2 = \langle h, \Phi(x) \rangle^2 \leq \langle h, h \rangle \langle \Phi(x), \Phi(x) \rangle = 0.$$

$\therefore$ h should be the constant-zero function.

So, $\mathcal{H}_0$ is an inner-product space.

It remains to show (iii) and (iv). For (iii),

$$\langle \Phi(x), \Phi(x') \rangle = \langle k(x, -), k(x', -) \rangle = k(x, x') \quad \text{for any } x, x' \in \mathcal{X}.$$

For (iv), for any $h = \sum_{i=1}^{m} a_i k(x_i, -) \in \mathcal{H}_0$ and $x \in \mathcal{X}$,

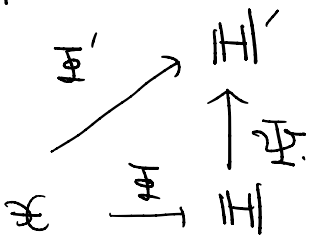
$$h(x) = \sum_{i=1}^{m} a_i k(x_i, x) = \left\langle \sum_{i=1}^{m} a_i k(x_i, -), k(x, -) \right\rangle = \langle h, \Phi(x) \rangle.$$

(4) The RKHS $\mathcal{H}$ constructed in Thm 6.8 is universal in the following sense:

For any Hilbert space $\mathcal{H}'$ and map $\Phi' : \mathcal{X} \to \mathcal{H}'$

s.t. $\langle \Phi'(x), \Phi'(x') \rangle_{\mathcal{H}'} = k(x, x')$ for all $x, x' \in \mathcal{X}$,

there exists a unique inner-product-preserving linear map $\Psi : \mathcal{H} \to \mathcal{H}'$ s.t. the following diagram commutes:



① The unique Ψ is constructed by first building it on $\mathcal{H}_0$ and then extending the built Ψ on $\mathcal{H}$ via limit preservation. We only show the first step:

$$\Psi\left(\sum_{i=1}^{m} a_i\, k(x_i, -)\right) \overset{\text{def}}{=} \sum_{i=1}^{m} a_i \Phi'(x_i).$$

Then, this is a well-defined inner-product-preserving linear map from $\mathcal{H}_0$ to $\mathcal{H}'$ that makes the above diagram commute. In fact, it is the only such map.

② This universality means that the RKHS $\mathcal{H}$ is built without using any information other than the one in k.

(5) The RKHS $\mathcal{H}$ often has an infinite dimension. But even in that case, it can be used to derive or analyse an algorithm that uses the kernel k. Also, it is useful to prove properties about k.

4. Recipes for constructing PDS kernels.

(1) Normalization

① Given a PDS kernel $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

the normalized kernel k' of k is:

$$k'(x,x') \overset{\text{def.}}{=} \begin{cases} 0 & \text{if } k(x,x)=0 \text{ or } k(x',x')=0 \\ \dfrac{k(x,x')}{\sqrt{k(x,x)k(x',x')}} & \text{otherwise} \end{cases}.$$

② Let $\mathbb{H}$ be the RKHS of k. and $\Phi: \mathcal{X} \to \mathbb{H}$ be $x \mapsto k(x,\cdot)$.

Then, k' can be understood as using the same $\mathbb{H}$ but changing Φ to the following Φ':

$$\Phi'(x) \overset{\text{def.}}{=} \begin{cases} 0 & \text{if } \Phi(x)=0 \quad \{ \text{constant-zero fn.} \\ \dfrac{\Phi(x)}{\|\Phi(x)\|_{\mathbb{H}}} & \text{otherwise.} \end{cases}$$

$$\left(\|\Phi(x)\|_{\mathbb{H}} \overset{\text{def.}}{=} \sqrt{\langle \Phi(x), \Phi(x) \rangle_{\mathbb{H}}} \right)$$

So, using k' means that we use 0 or points in the unit sphere in $\mathbb{H}$ to encode the elements in $\mathcal{X}$.

③ Example

$$k(x,x') = \exp\left(\frac{\langle x, x' \rangle}{\sigma^2} \right) \quad \text{for } x, x' \in \mathbb{R}^N.$$

Then, the normalization of k is the RBF kernel k':

$$k'(x,x') = \exp\left(\frac{\langle x, x' \rangle}{\sigma^2} \right) \Big/ \sqrt{\exp\left(\frac{\|x\|^2}{\sigma^2} \right) \exp\left(\frac{\|x'\|^2}{\sigma^2} \right)}$$

$$= \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2} \right).$$

Lemma 6.9. The normalized k' of a PDS kernel k is PDS.

Proof: By ② above. More precisely, $k'(x,x') = \langle \Phi'(x), \Phi'(x') \rangle$ ∎

Now the symmetry and positive definiteness of k' follows from the corresponding properties of $\langle \cdot, \cdot \rangle_{\mathbb{H}}$. □

(2) Empirical kernel map and empirically-induced kernels.

① Given a PDS k and a sample $S = (x_1, \dots, x_m)$, is there any way to construct $\Phi' : \mathcal{X} \to \mathbb{R}^m$ s.t.

$$k'(x,x') = \langle \Phi'(x), \Phi'(x') \rangle_{\mathbb{R}^m}$$

is closely related to k? The empirical kernel map and its cousin is such Φ'.

② The empirical kernel map Φ' associated to k and S is:

$$\Phi' : \mathcal{X} \to \mathbb{R}^m$$

$$\Phi'(x) \overset{\text{def.}}{=} (k(x,x_1), \dots, k(x,x_m))$$

Define $K \in \mathbb{R}^{m \times m}$ by $K_{ij} \overset{\text{def.}}{=} k(x_i, x_j)$.

Then, $\underbrace{\langle \Phi'(x_i), \Phi'(x_j) \rangle}_{k'(x_i,x_j) \overset{\text{def.}}{=} K'} = \sum_{l=1}^{m} k(x_i,x_l) k(x_j,x_l) = (K^2)_{ij}$.

So, the kernel induced by Φ' is related to k, but not as much as we want. We want $k'(x_i,x_j)$ to be K_{ij}, not K_{ij}^2. How should we modify Φ' to achieve this?

③ Eigen-decomposition of $K = U \Sigma U^T$ with $\Sigma_{ii} \geq 0$.

Let $(K^+)^{\frac{1}{2}} = U \Sigma^{-\frac{1}{2}} U^T$ where $\Sigma^{-\frac{1}{2}}$ is the diagonal matrix with

$$\left(\Sigma^{-\frac{1}{2}}\right)_{ii} = \begin{cases} 0 & \text{if } \Sigma_{ii} = 0. \\ \frac{1}{\sqrt{\Sigma_{ii}}} & \text{otherwise.} \end{cases}$$

Define $\Phi''(x) \overset{\text{def}}{=} (K^+)^{\frac{1}{2}} \Phi'(x)$ and $k''(x,x') = \langle \Phi''(x), \Phi''(x') \rangle$.

Lemma: $\forall i,j \in [m]$, $k''(x_i, x_j) = K_{ij}$

Proof Let $K^+ = U \Sigma^+ U^T$ where $\Sigma^+ = (\Sigma^+)^{\frac{1}{2}} (\Sigma^+)^{\frac{1}{2}}$.

Also, let $e_i = (0, 0, \ldots, 0, \underset{\uparrow}{1}, 0, \ldots 0) \in \mathbb{R}^m$
i-th component

Then, $KK^+K = K$. and $\Phi'(x_i) = K e_i$

So, $k''(x_i, x_j) = \langle \Phi''(x_i), \Phi''(x_j) \rangle = \left((K^+)^{\frac{1}{2}} \Phi'(x_i) \right)^T \left((K^+)^{\frac{1}{2}} \Phi'(x_j) \right)$

$= \Phi'(x_i)^T \left((K^+)^{\frac{1}{2}} \right)^T \left((K^+)^{\frac{1}{2}} \right) \Phi'(x_j)$

$(K^+)^{\frac{1}{2}}$ is symmetric

$= \Phi'(x_i)^T (K^+)^{\frac{1}{2}} (K^+)^{\frac{1}{2}} \Phi'(x_j)$

$K^+ = (K^+)^{\frac{1}{2}} (K^+)^{\frac{1}{2}}$

$= \Phi'(x_i)^T K^+ \Phi'(x_j)$ ← K is symmetric.

$= e_i^T K^T K^+ K e_j = e_i K K^+ K e_j$

$KK^+K = K$

$= e_i K e_j = K_{ij}$ □

(3) Sum, product, tensor product, pointwise limit, and composition with a power series

① k, k' ... PDS kernels on $\mathcal{X}$.

$(k + k')(x, x') \overset{\text{def}}{=} k(x, x') + k'(x, x')$ .... sum

$(k k')(x, x') \overset{\text{def}}{=} k(x, x') k'(x, x')$ ... product.

As we will show shortly, $k + k'$ and kk' are PDS kernels.

A good way to see what's going on in these constructions is to think in terms of feature maps Φ, Φ' associated to k, k'.

Supp. $\Phi : \mathcal{X} \to \mathbb{R}^n$, $\Phi' : \mathcal{X} \to \mathbb{R}^{n'}$, $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$,

$k'(x, x') = \langle \Phi'(x), \Phi'(x') \rangle$

Let $\Phi_+ : \mathcal{X} \to \mathbb{R}^{n+n'}$, $\Phi_+(x) = \langle \Phi(x), \Phi'(x) \rangle$

and $\Phi_\times : \mathcal{X} \to \mathbb{R}^{nn'}$, $\Phi_\times(x)_{(i-1)n'+j} = \Phi(x)_i \, \Phi'(x)_j$

Then, $(k+k')(x,x') = \langle \Phi_+(x), \Phi_+(x') \rangle_{\mathbb{R}^{n+n'}}$

$(kk')(x,x') = \langle \Phi_\times(x), \Phi_\times(x') \rangle_{\mathbb{R}^{nn'}}$

Thus, $k+k'$ and kk' correspond to two ways of combining feature maps of k and k'.

② k_1 ... PDS kernel on $\mathcal{X}_1$

k_2 ... PDS kernel on $\mathcal{X}_2$

$k_1 \otimes k_2 : (\mathcal{X}_1 \times \mathcal{X}_2) \times (\mathcal{X}_1 \times \mathcal{X}_2) \to \mathbb{R}$

$(k_1 \otimes k_2)\big((x_1,x_2),(x_1',x_2')\big) \overset{\text{def.}}{=} k_1(x_1,x_1') \, k_2(x_2,x_2')$.

... tensor product.

From the perspective of feature maps Φ_1 of k_1 and Φ_2 of k_2, the tensor product works like product in ①. If $\Phi_1 : \mathcal{X}_1 \to \mathbb{R}^{n_1}$, $\Phi_2 : \mathcal{X}_2 \to \mathbb{R}^{n_2}$, $k_1(x_1,x_1') = \langle \Phi_1(x_1), \Phi_1(x_1') \rangle$ and $k_2(x_2,x_2') = \langle \Phi_2(x_2), \Phi_2(x_2') \rangle$, then

$(k_1 \otimes k_2)\big((x_1,x_2),(x_1',x_2')\big) = \langle \Phi_\otimes(x_1,x_2), \Phi_\otimes(x_1',x_2') \rangle$

where $\Phi_\otimes : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}^{n_1 n_2}$.

$\Phi_\otimes(x_1,x_2)_{(i-1)n_2+j} = \Phi_1(x_1)_i \, \Phi_2(x_2)_j$.

③ k_i ... PDS kernels on $\mathcal{X}$ s.t. $k_i(x,x') \to k(x,x')$ for all $i \in \mathbb{N}$ as $i \to \infty$, for all $x, x' \in \mathcal{X}$.

Then, k is also PDS. Intuitively, k is defined in terms of the limit feature map of those of the k_i's, and using it often amounts to using an infinite feature space.

⑥ $\sum_{i=0}^{\infty} a_i x^i$ with $a_i \geq 0$ and radius of convergence $\rho > 0$. (i.e., if $x \in (-\rho, \rho)$, $\sum_{i=0}^{n} a_i x^i$ converges as $n \to \infty$).

k ... PDS kernel on $\mathcal{X}$ s.t. $k(x,x') \in (-\rho, \rho)$ for all $x, x' \in \mathcal{X}$.

$\Rightarrow$

$k'(x,x') \overset{\text{def.}}{=} \sum_{i=0}^{\infty} a_i k(x,x')^i$ is a PDS kernel.

.... composition with a power series.

⑦ Using what we learnt, we can prove that the RBF kernel $k_G(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ is PDS.

Note $k_0(x,x') = \langle x, x' \rangle$ is a PDS kernel.

Thus, by ⑥,

$k_1(x,x') = \exp\left(\frac{\langle x,x' \rangle}{\sigma^2}\right) = \sum_{i=0}^{\infty} \frac{1}{i!\,\sigma^{2i}} \left(\langle x,x' \rangle\right)^i$

is PDS. Finally, by normalization, the RBF kernel

$k_G(x,x') = k_1(x,x') \Big/ \sqrt{k_1(x,x)}\sqrt{k_1(x',x')}$

is PDS.

Thm 6.10 PDS kernels are closed under sum, product, tensor product, pointwise limit, and composition with power series.

Proof. The symmetry of the constructed kernel in each of the five cases is easy to check. So, we show only the positive definiteness of the constructed kernel.

(i) Let k, k' be PDS kernels on $\mathcal{X}$. Consider $c_1, \dots, c_m \in \mathbb{R}$ and $x_1, \dots, x_m \in \mathcal{X}$.

$$\sum_{i,j=1}^{m} c_i c_j (k+k')(x_i, x_j) = \sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) + \sum_{i,j=1}^{m} c_i c_j k'(x_i, x_j) \geq 0.$$

Let K be the Gram matrix of k and $(x_1, \dots, x_m)$.

Since the matrix K is symmetric and positive semi-definite, K has non-negative real eigenvalues. That is,

$$K = U \Sigma U^T$$

for an orthonormal matrix U and a diagonal matrix Σ with non-negative entries. Let $\Sigma^{\frac{1}{2}}$ be the diagonal matrix with $(\Sigma^{\frac{1}{2}})_{ii} = \sqrt{\Sigma_{ii}}$.

and $K^{\frac{1}{2}} = U \Sigma^{\frac{1}{2}} U^T$. Then $(K^{\frac{1}{2}})^T (K^{\frac{1}{2}}) = K$.

Thus,

$$\sum_{i,j=1}^{m} c_i c_j (kk')(x_i, x_j) = \sum_{i,j=1}^{m} c_i c_j K_{ij} k'(x_i, x_j).$$

$$= \sum_{i,j=1}^{m} c_i c_j \left(\sum_{k=1}^{m} (K^{\frac{1}{2}})_{ki} (K^{\frac{1}{2}})_{kj} \right) k'(x_i, x_j)$$

$$= \sum_{k=1}^{m} \sum_{i,j=1}^{m} \left(c_i (K^{\frac{1}{2}})_{ki} \right) \left(c_j (K^{\frac{1}{2}})_{kj} \right) k'(x_i, x_j) \geq 0.$$

(ii) Let k_1 and k_2 be PDS kernels on $\mathcal{X}_1$ and $\mathcal{X}_2$.

For $i \in \{1,2\}$, $k_i' : (\mathcal{X}_1 \times \mathcal{X}_2) \times (\mathcal{X}_1 \times \mathcal{X}_2) \to \mathbb{R}$

$$k_i'((x_1, x_2), (x_1', x_2')) = k_i(x_i, x_i').$$

is also PDS. The symmetry comes from the symmetry of k_i. For the positive definiteness,

$$\sum_{k,l=1}^{m} c_k c_l k_i'((x_{k,1}, x_{k,2}), (x_{l,1}, x_{l,2}))$$

$$= \sum_{k,l=1}^{m} c_k c_l k_i(x_{k,i}, x_{l,i}) \geq 0. \quad (\because k_i \text{ is positive definite}).$$

Then, $k_1 \otimes k_2 = k_1' k_2'$, and so by what we proved in (i), $k_1 \otimes k_2$ is positive definite.

(iii) Let $(k_i)_{i \in \mathbb{N}}$ be a sequence of PDS kernels on $\mathcal{X}$ s.t. for all $x, x' \in \mathcal{X}$, $k_i(x,x')$ converges as $i \to \infty$. Let $k(x,x') = \lim_{i \to \infty} k_i(x,x')$.

Consider $c_1, \dots, c_m \in \mathbb{R}$ and $x_1, \dots, x_m \in \mathcal{X}$.

Then $\sum_{k,l=1}^{m} c_k c_l k_i(x_k, x_l) \geq 0$ for all $i \geq 1$.

Also, as $i \to \infty$,

$$\sum_{k,l=1}^{m} c_k c_l k_i(x_k, x_l) \longrightarrow \sum_{k,l=1}^{m} c_k c_l k(x_k, x_l).$$

Thus " ≥ 0.

(iv) Consider a PDS kernel k on $\mathcal{X}$ and a power series $\sum_{i=0}^{\infty} a_i r^i$ s.t. $a_i \geq 0$ and the range of k falls into the radius of convergence of the series.

Note

$k_{a_i}(x,x') \overset{\text{def.}}{=} a_i$ and $k_i(x,x') \overset{\text{def.}}{=} (k(x,x'))^i$

are PDS kernels by (i). Again by (ii), $\sum_{i=0}^{n} k_{a_i} k_i(x,x') = \sum_{i=0}^{n} a_i (k(x,x'))^i$ is a PDS kernel. By (iii), the following limit is also PDS.

$$\lim_{n \to \infty} \sum_{i=0}^{n} a_i (k(x,x'))^i$$

R.

5 Kernel-based Algorithms.

(1) Several ML algorithms can be expressed in terms of inner products between inputs. Here we are considering both the training/learning part of such an algorithm and the learnt classifier of it. These ML algorithms can be generalised easily via kernels. We just need to replace inner products by the applications of a kernel.

(2) Let's see how this kernelisation (or kernel trick) works for SVM. Recall the following optimisation in the dual version of the SVM:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \quad (\langle x_i, x_j \rangle \to k(x_i, x_j))$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C \quad \text{for all } i \in [m] \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0.$$

and the learnt classifier as well:

$$h(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i^* y_i \langle x_i, x \rangle + y_{i_0} - \sum_{i=1}^{m} \alpha_i^* y_i \langle x_i, x_{i_0} \rangle \right)$$

where α^* is a solution of the optimisation pb. and $0 < \alpha_{i_0} < C$.

($\langle x_i, x \rangle \to k(x_i, x)$, $\langle x_i, x_{i_0} \rangle \to k(x_i, x_{i_0})$)

To kernelise this SVM algo., we just need to change $\langle \cdot, \cdot \rangle$ by $k(\cdot, \cdot)$. We show these changes in the above optimisation and the learnt classifier using comments in red color.

(3) What is going on here? When we change $\langle x_i, x_j \rangle$ in the above optimisation, what we do there is just to use the SVM

but not on the original input space, but on the new input space, namely, the RKHS of k.

(4) If we ignore the bias term "$y_{i_0} - \sum_{i=1}^{m} \alpha_i^* y_i \langle x_i, x_{i_0} \rangle$", the fn used to build h is:

$$f = \sum_{i=1}^{m} \alpha_i^* y_i k(x_i, -)$$

that is, a linear combination of the k-embedded examples in the RKHS $\mathcal{H}$, instead of some other fn $\mathcal{H}$. This is not a coincidence. Many optimizations over $\mathcal{H}$ have solutions of the above form. (By the way, the optimization of the kernelised SVM can be expressed as an opt. over $\mathcal{H}$). The next thm explains why.

Thm 6.11 [Representer Thm]. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ... PDS kernel, $\mathcal{H}$ ... the RKHS of k, $G : \mathbb{R} \to \mathbb{R}$ ... non-decreasing fn, $L : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ ... any fn. If

$$\operatorname{argmin}_{h \in \mathcal{H}} F(h) \quad \text{with} \quad F(h) = G(\|h\|_{\mathcal{H}}) + L(h(x_1), \dots, h(x_m))$$

has a solution, then it admits a solution of the form $h^* = \sum_{i=1}^{m} \beta_i k(x_i, -)$ for $\beta_i \in \mathbb{R}$. If G is further assumed to be increasing, every solution has this form.

Proof. Let $\mathcal{H}_1 = \operatorname{span}\{k(x_i, -) \mid i \in [m]\}$ and h be a solution of the optimization. Then, $\exists\, h_1 \in \mathcal{H}_1$ and $h_1^\perp$ s.t.

(i) $h = h_1 + h_1^\perp$ (ii) $\langle h_1^\perp, h' \rangle = 0$ for all $h' \in \mathcal{H}_1$.

Then, $\|h_1\| + \|h_1^\perp\| = \|h\|$. $\therefore\ G(\|h_1\|) \le G(\|h\|)$.

By the reproducing property, $h(x_i) = \langle h, k(x_i,-)\rangle$

$= \langle h_1 + h_1^\perp, k(x_i,-)\rangle$

$= \langle h_1, k(x_i,-)\rangle + \langle h_1^\perp, k(x_i,-)\rangle$

$= \langle h_1, k(x_i,-)\rangle = h_1(x_i)$.

$\therefore\ L(h(x_1),\cdots,h(x_m)) = L(h_1(x_1),\cdots,h_1(x_m))$

$\therefore\ F(h_1) \le F(h)$ $\therefore\ h_1 \in \mathcal{H}$ is also a sol.

If G is increasing, $G(\|h_1\|) < G(\|h\|)$ unless $h_1^\perp = 0$.

Since h is a solution, $h_1^\perp = 0$ $\therefore\ h \in \mathcal{H}_1$. Q.

One message of this thm is that we don't have to be afraid of using the RKHS $\mathcal{H}$ of k even when $\mathcal{H}$ is infinite dimensional. The thm says that if our learning algorithm uses a certain style of optimization, this optimization and its solution can be computed over $\mathcal{H}$.

(5) The margin theory for SVM continues to work for kernelized SVM.

Thm 6.12. $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ ... PDS kernel.

$\mathcal{H}$ ... RKHS of k.

$\Phi : \mathcal{X} \to \mathcal{H}$ ... feature map associated to k.

$S \in \{x \mid k(x,x) \le r^2\}^m$.

$\mathcal{F} = \{x \mapsto \langle w, \Phi(x)\rangle \mid \|w\|_{\mathcal{H}} \le \Lambda\}$.

$\Rightarrow$

$$\widehat{R}_S(\mathcal{F}) \le \frac{\Lambda\sqrt{\mathrm{Tr}[K]}}{m} \le \sqrt{\frac{r^2\Lambda^2}{m}}.$$

where $K \in \mathbb{R}^{m\times m}$, $K_{ij} = k(x_i, x_j)$.

Proof. The proof is essentially the one for Thm 5.10, except that we now use $\mathcal{H}$, instead of $\mathcal{X}$, for the feature space.

$$\widehat{R}_S(\mathcal{F}) = \frac{1}{m}\,\mathbb{E}_{\vec{\sigma}\in\mathrm{Unif}\{\pm1\}^m}\Big[\sup_{\|w\|_{\mathcal{H}}\le\Lambda}\sum_{i=1}^m \sigma_i\langle w, \Phi(x_i)\rangle_{\mathcal{H}}\Big]$$

$$= \frac{1}{m}\,\mathbb{E}_{\vec{\sigma}}\Big[\sup_{\|w\|_{\mathcal{H}}\le\Lambda}\Big\langle w, \sum_{i=1}^m \sigma_i\Phi(x_i)\Big\rangle_{\mathcal{H}}\Big]$$

Cauchy-Schwarz

$$\le \frac{\Lambda}{m}\,\mathbb{E}_{\vec{\sigma}}\Big[\Big\|\sum_{i=1}^m \sigma_i\Phi(x_i)\Big\|_{\mathcal{H}}\Big]$$

$$= \frac{\Lambda}{m}\,\mathbb{E}_{\vec{\sigma}}\Big[\Big\|\sum_{i=1}^m \sigma_i\Phi(x_i)\Big\|_{\mathcal{H}}^{2\times\frac{1}{2}}\Big]$$

$$\le \frac{\Lambda}{m}\,\mathbb{E}_{\vec{\sigma}}\Big[\sum_{i,j=1}^m \sigma_i\sigma_j\langle\Phi(x_i),\Phi(x_j)\rangle_{\mathcal{H}}\Big]^{\frac{1}{2}}.$$

$$= \frac{\Lambda}{m}\sqrt{\sum_{i=1}^m k(x_i,x_i)} = \frac{\Lambda}{m}\sqrt{\mathrm{Tr}[K]} \le \sqrt{\frac{\Lambda^2 r^2}{m}}. \quad \square$$

Now by Thm 6.12, Thm 5.8 and Thm 5.9, we get the following results:

Cor 6.13. $k: \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ … PDS kernel with $r^2 = \sup_{x\in\mathcal{X}} k(x,x)$.
$\mathcal{H}$, Φ … RKHS and feature mapping of k.
$\mathcal{F} = \{x\mapsto\langle w,\Phi(x)\rangle \mid \|w\|_{\mathcal{H}}\le\Lambda\}$.
$\rho>0$, $\delta>0$, $D\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})$, $m\in\mathbb{N}$.

$\Rightarrow$

$$\mathbb{P}_{S\sim D^m}\Big[\forall f\in\mathcal{F}.\ R(\mathrm{sgn}\circ f) \le \widehat{R}_{S,\rho}(f) + \frac{2}{\rho}\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\Big] \ge 1-\delta.$$

$$\mathbb{P}_{S\sim D^m}\Big[\forall f\in\mathcal{F}.\ R(\mathrm{sgn}\circ f) \le \widehat{R}_{S,\rho}(f) + \frac{2}{\rho}\sqrt{\frac{\Lambda^2\,\mathrm{Tr}[K]}{m}} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\Big] \ge 1-\delta.$$

Cor $k, \mathcal{H}, \Phi, \mathcal{F}, \delta > 0, D, m \in \mathbb{N}$ ... as before. $r' > 0$.

$\Rightarrow$

$$\mathbb{P}\left[\forall f \in \mathcal{F}\ \forall \rho \in (0, r'] \text{ s.t. } R(\operatorname{sgn} \circ f) \le \hat{R}_{S,\rho}(f) + \frac{4}{\rho}\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\log\log_2 \frac{2r'}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \ge 1-\delta$$

and

$$\mathbb{P}\left[\forall f \in \mathcal{F}\ \forall \rho \in (0, r'] \text{ s.t. } R(\operatorname{sgn} \circ f) \le \hat{R}_{S,\rho}(f) + \frac{4}{\rho}\sqrt{\frac{\operatorname{Tr}[K]\Lambda^2}{m}} + \sqrt{\frac{\log\log_2 \frac{2r'}{\rho}}{m}} + 3\sqrt{\frac{\log\frac{4}{\delta}}{2m}}\right] \ge 1-\delta$$

6. Negative Definite Symmetric Kernel

(1) So far we have used inner product as an exemplary similarity measure, and explained kernels as inner product operations over some rich-feature spaces. Another concept often used to express similarity is distance. In fact, the RBF kernel is defined in terms of distance:

$$k_{RBF}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right).$$

In this part of the chapter, we will define negative definite squared symmetric kernel k, which computes the distance in some induced Hilbert space. Then, we will show that:

$$k'(x, x') = \exp(-tk(x, x')) \quad \text{for } t > 0$$

is a PDS kernel, generalizing the construction used in the RBF kernel.

(2).

Def A kernel $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is negative-definite symmetric (CNDS) if k is symmetric and for all $c \in \mathbb{R}^m$ s.t. $\sum_{i=1}^m c_i = 0$ and for all $x_1, \dots, x_m \in \mathcal{X}$, $\sum_{i,j=1}^m c_i c_j k(x_i, x_j) \leq 0$.

① example.

$\mathcal{X} = \mathbb{R}^N$ (or more generally some Hilbert space $\mathcal{H}$)

$k(x, x') = \|x - x'\|^2$.

Then, k is symmetric. It is also negative definite.

as shown below: for $x_1, \dots, x_m \in \mathcal{X}$ and $c \in \mathbb{R}^m$ s.t. $\sum_{i=1}^m c_i = 0$,

$$\sum_{i,j=1}^m c_i c_j k(x_i, x_j) = \sum_{i,j=1}^m c_i c_j \|x_i - x_j\|^2$$

$$= \sum_{i,j=1}^m c_i c_j \left(\|x_i\|^2 - 2\langle x_i, x_j \rangle + \|x_j\|^2 \right)$$

$$= \sum_{i=1}^m c_i \|x_i\|^2 \left(\sum_{j=1}^m c_j \right) - 2 \left\| \sum_{i=1}^m c_i x_i \right\|^2 + \sum_{j=1}^m c_j \|x_j\|^2 \left(\sum_{i=1}^m c_i \right)$$

$$= -2 \left\| \sum_{i=1}^m c_i x_i \right\|^2 \leq 0.$$

② Relationship with PDS kernels.

Thm 6.16 $x_0 \in \mathcal{X}$, $k, k': \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t. k is symmetric.

For all x, x', $k'(x, x') = k(x, x_0) + k(x', x_0) - k(x, x') - k(x_0, x_0)$

$\Rightarrow$

k is NDS iff k' is PDS.

Proof. Assume k' is PDS. Pick $x_1, \dots, x_m \in \mathcal{X}$ and $c \in \mathbb{R}^m$ s.t. $\sum_{i=1}^m c_i = 0$.

$$\sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) = \sum_{i,j=1}^{m} c_i c_j \left(k(x_i, x_0) + k(x_j, x_0) - k(x_0, x_0) - k'(x_i, x_j) \right)$$

$$= \sum_i c_i k(x_i, x_0) \left(\sum_j c_j \right) + \sum_j c_j k(x_j, x_0) \left(\sum_i c_i \right) - k(x_0, x_0) \left(\sum_i c_i \right) \left(\sum_j c_j \right)$$
$$- \sum_{i,j} c_i c_j k'(x_i, x_j)$$

$$= - \sum_{i,j} c_i c_j k'(x_i, x_j) \leq 0.$$

Now assume that k is NDS. Then, k' is symmetric by the equation in the thm. To show the pos. definiteness of k', consider $x_1, \dots, x_m \in \mathcal{X}$ and $c_1, \dots, c_m \in \mathbb{R}$. Let $c_0 = -\sum_{i=1}^{m} c_i$.

Then,

$$\sum_{i,j=0}^{m} c_i c_j k(x_i, x_j) \leq 0.$$

But $\sum_{i,j=0}^{m} c_i c_j \left(k(x_i, x_0) + k(x_j, x_0) - k'(x_i, x_j) - k(x_0, x_0) \right)$

$$= \sum_{i=0}^{m} c_i k(x_i, x_0) \left(\sum_{j=0}^{m} c_j \right) + \sum_{j=0}^{m} c_j k(x_j, x_0) \left(\sum_{i=0}^{m} c_i \right)$$
$$- \sum_{i,j=0}^{m} c_i c_j k'(x_i, x_j) - k(x_0, x_0) \left(\sum_{i=0}^{m} c_i \right) \left(\sum_{j=0}^{m} c_j \right)$$

$$= - \sum_{i,j=0}^{m} c_i c_j k'(x_i, x_j) = - \sum_{i,j=1}^{m} c_i c_j k'(x_i, x_j) - 2c_0 \sum_{j=0}^{m} c_j k'(x_0, x_j) + c_0^2 k'(x_0, x_0)$$

$$\therefore \quad c_0^2 k'(x_0, x_0) - 2c_0 \sum_{j=0}^{m} c_j k'(x_0, x_j) \leq \sum_{i,j=0}^{m} c_i c_j k'(x_i, x_j).$$

($c_0^2 k'(x_0, x_0) = 0$)

Because $k'(x, x_0) = 0$ by the assumed equality. □

Thm 6.17. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ··· symmetric kernel.

$\Rightarrow$ k is NDS iff. $\exp(-tk)$ is PDS for all $t > 0$.

※ The textbook says that exercises 6.17 and 6.18. tell us how to prove this theorem. and the next thm.

Thm 6.18. $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ... NDS kernel.
$k(x, x') = 0$ iff $x = x'$.

$\Rightarrow$

$\exists$ a Hilbert space $\mathbb{H}$ and a map $\Phi: \mathcal{X} \to \mathbb{H}$ s.t.

$\forall x, x' \in \mathcal{X}$ $\quad k(x, x') = \| \Phi(x) - \Phi(x') \|^2$.

Thus, $\sqrt{k}$ is a metric

Proof: Actually, we can prove Thm 6.18 using Thm 6.16. Pick $x_0 \in \mathcal{X}$

Defn $k'(x, x') = k(x, x_0) + k(x', x_0) - k(x, x') - k(x_0, x_0)$.

By Thm 6.16, k' is PDS. Let $\mathbb{H}'$, Φ' be the RKHS of k' and the associated feature map. Note that:

$$\| \Phi'(x) - \Phi'(x') \|^2 = k'(x, x) - 2k'(x, x') + k'(x', x')$$

$$= 2k(x, x_0) - k(x, x) - k(x_0, x_0)$$

$$- 2k(x, x_0) - 2k(x', x_0) + 2k(x, x') + 2k(x_0, x_0)$$

$$+ 2k(x', x_0) - k(x', x') - k(x_0, x_0)$$

$$= - k(x, x) + 2k(x, x') - k(x', x')$$

$$= 2k(x, x').$$

$\therefore$ The desired $\mathbb{H}$ and Φ are $\mathbb{H}'$ and $\frac{1}{\sqrt{2}} \Phi'$. D.

1). Approximate Kernel Feature Maps.

(1) Kernel-based algorithms are too slow for some problems because they often have to compute $k(x_i, x_j)$ for all inputs x_i, x_j in the sample or $k(x, x_i)$. which becomes very expensive when the number m of inputs in the sample (or training set) is too large.

(2) In such cases, using an approximate k' defined by

$$k'(x, x') = \langle \Phi(x), \Phi(x') \rangle \quad \approx \quad k(x, x')$$

for $\Phi : \mathcal{X} \to \mathbb{R}^D$ for reasonably small D (in particular, $D \ll m$) is a promising approach. Note that the empirical kernel that we looked at before is not good here, because $D = m$ in that case. Is there a good principled way to define k'?

(3) We will study one such approach. It is based on the following thm:

Thm 6.24 [Bochner's Thm]

$\mathcal{X} = \mathbb{R}^N$, $G : \mathcal{X} \to \mathbb{R}$ ... continuous fn. with $G(0) = 1$

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k(x, x') \overset{\text{def}}{=} G(x - x')$

k is symmetric

$\Rightarrow$ k is positive definite iff G is the Fourier transform of a probability distribution.

that is, G has the following form:

$$G(x) = \int_{\mathcal{X}} e^{i\langle \omega, x \rangle} p(d\omega)$$

where p is a probability distribution.

(1) Instances:

	$G(x - x')$	$p(\omega)$ ... density.
Gaussian	$\exp\left(-\frac{\|x - x'\|^2}{2}\right)$	$\frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left(-\frac{\|\omega\|^2}{2}\right)$
Laplacian	$\exp\left(-\|x - x'\|_1\right)$	$\prod_{i=1}^{N} \frac{1}{\pi(1 + \omega_i^2)}$.
Cauchy	$\prod_{i=1}^{N} \frac{2}{1 + (x_i - x_i')^2}$.	$\exp\left(-\|\omega\|_1\right)$.

② The next proposition tells us how to approximate k using random samples.

Prop 6.25: Same setup for $\mathcal{X}, k, G$ as the one of Thm 6.24. Let p be the prob. distribution of the theorem. Then,

$$\mathbb{E}_{w\sim p}\left[\langle (\cos\langle w,x\rangle, \sin\langle w,x\rangle), (\cos\langle w,x'\rangle, \sin\langle w,x'\rangle)\rangle\right] = k(x,x') \quad \text{for all } x,x'\in\mathcal{X}.$$

This proposition leads to a randomized algorithm for computing an approximation of k. First, we sample

$$w_1, \cdots, w_L$$

independently from p. Then, we define a new kernel k' by

$$k'(x,x') = \frac{1}{L}\sum_{i=1}^{L}\langle (\cos\langle w_i,x\rangle, \sin\langle w_i,x\rangle), (\cos\langle w_i,x'\rangle, \sin\langle w_i,x'\rangle)\rangle$$

real part of a complex number.

The proposition ensures that on average, $k' = k$.

Proof of Prop 6.25: By Thm 6.24,

$$\begin{aligned} k(x,x') &= \mathrm{Re}\left[k(x,x')\right] \\ &= \int_{\mathcal{X}} \mathrm{Re}\left[e^{i\langle w, x-x'\rangle}\right] p(dw) \\ &= \int_{\mathcal{X}} \cos\langle w, x-x'\rangle \, p(dw) \\ &= \int_{\mathcal{X}} \cos\left(\langle w,x\rangle - \langle w,x'\rangle\right) p(dw) \\ &= \int_{\mathcal{X}} \cos(\langle w,x\rangle)\cos(\langle w,x'\rangle) + \sin(\langle w,x\rangle)\sin(\langle w,x'\rangle) \\ &\qquad p(dw). \end{aligned}$$

$$= \mathop{\mathbb{E}}_{\omega \sim p} \left[\left\langle (\cos\langle \omega, x \rangle, \sin\langle \omega, x \rangle), (\cos\langle \omega, x' \rangle, \sin\langle \omega, x' \rangle) \right\rangle \right]. \qquad \text{R.}$$

(3) The textbook has further more involved analysis on the error of k' from above. If you are interested, look at Lemma 6.26, Lemma 6.27, Thm 6.28.