3. **Kernels (22 pts).**

(a) (4 pts) To what feature map Φ does the kernel

$$k(\boldsymbol{u}, \boldsymbol{v}) = (\langle u, v\rangle + 1)^3$$

correspond? Assume the inputs have an arbitrary dimension d and the inner product is the dot product.

- Here is my approach: solve this for $d = 2$ such that $u, v \in \mathbb{R}^2$, where $\langle u, v\rangle = \left(u^{(1)}v^{(1)} + u^{(2)}v^{(2)}\right)$, then extrapolate from inputs $\in \mathbb{R}^2$ to inputs $\in \mathbb{R}^d$.

- $\left(\langle u, v\rangle + 1\right)^3 = \left(u^{(1)}v^{(1)} + u^{(2)}v^{(2)} + 1\right)^3$. This will be messy so let $a = u^{(1)}v^{(1)}$, $b = u^{(2)}v^{(2)}$, and $c = 1$.

- $(a+b+c)^3 = (a+b+c)^2(a+b+c) = a^3+b^3+c^3+3a^2b+3a^2c+3ab^2+3b^2c+3ac^2+3bc^2+6abc$

- Here, $\Phi(u) = \Big[(u^{(1)})^3, (u^{(2)})^3, 1, \sqrt{3}(u^{(1)})^2(u^{(1)}), \sqrt{3}(u^{(1)})^2,$
$\sqrt{3}(u^{(1)})(u^{(1)})^2, \sqrt{3}(u^{(2)})^2, \sqrt{3}(u^{(1)}), \sqrt{3}(u^{(2)}), \sqrt{6}(u^{(1)})(u^{(2)})\Big]$

  and $\Phi(v)$ is the same, except with "v" in place of "u" $\forall u$.

- NOTE: When attempting to extrapolate this to higher dimensions, I found additional terms not present for $\mathbb{R}^2$ when experimenting with $d = 3$ and $d = 4$. I successfully detected and characterized this pattern for $d = d$ through a lot of scratch work. For my sake (I can't spend 2+ hours LaTex-ing scratch work) and your sake (a lot of extra reading), I have omitted this scratch work. The feature map can be found on the next page (it takes up enough space I put it on its own page).

- NOTE: I have intentionally formatted this a bit oddly, but it's done with the intent of grouping like elements of the feature map. See explanation below feature map definition for details. Extrapolating our $d = 2$ feature map to higher, arbitrary dimension d we'll have $\Phi(u) =$

  $\Big[(u^{(1)})^3, (u^{(2)})^3, ..., (u^{(d)})^3,$

  $1,$

  $\sqrt{3}(u^{(1)})^2(u^{(2)}), \sqrt{3}(u^{(1)})^2(u^{(3)}), ..., \sqrt{3}(u^{(1)})^2(u^{(d)}), ..., \sqrt{3}(u^{(d)})^2(u^{(1)}), ..., \sqrt{3}(u^{(d)})^2(u^{(d-1)}),$

  $\sqrt{3}(u^{(1)})^2, \sqrt{3}(u^{(2)})^2, ..., \sqrt{3}(u^{(d)})^2,$

  $\sqrt{3}(u^{(1)}), \sqrt{3}(u^{(2)}), ..., \sqrt{3}(u^{(d)}),$

  $\sqrt{6}\big(u^{(1)}u^{(2)}\big), \sqrt{6}\big(u^{(1)}u^{(3)}\big), ..., \sqrt{6}\big(u^{(1)}u^{(d)}\big), \sqrt{6}\big(u^{(2)}u^{(3)}\big), \sqrt{6}\big(u^{(2)}u^{(4)}\big)..., \sqrt{6}\big(u^{(d-1)}u^{(d)}\big),$

  $\sqrt{6}\big(u^{(1)}u^{(2)}\cdots u^{(d-1)}\big), \sqrt{6}\big(u^{(1)}u^{(2)}\cdots u^{(d-2)}u^{(d)}\big), ..., \sqrt{6}\big(u^{(1)}u^{(3)}\cdots u^{(d-1)}u^{(d)}\big)\Big]$

  i. The first line of $\Phi(u)$ has d elements
  ii. The second line of $\Phi(u)$ has 1 element
  iii. The third line of $\Phi(u)$ has $d(d-1)$ elements
  iv. The fourth line of $\Phi(u)$ has d elements
  v. The fifth line of $\Phi(u)$ has d elements
  vi. The sixth line of $\Phi(u)$ has $\frac{d(d-1)}{2}$ elements
  vii. The seventh line of $\Phi(u)$ has $\frac{(d-2)(d-1)(d)}{6}$ elements

  There end up being a total of $\left[\frac{(d+1)(d+2)(d+3)}{6}\right]$ elements in the feature map $\Phi(u)$.
  For example, for $d = 2$ there are 10 elements, which matches what we got for the work at the beginning of the problem. I checked this for higher-dimensional d and it holds. (Again, this comes from the ugly scratch work that isn't worth LaTex-ing).

  The feature map $\Phi(v)$ looks the same as $\Phi(u)$ except that all "u's" are replaced with "v's".

  Thus the desired feature map has been defined for $k(u, v) = \langle\Phi(u), \Phi(v)\rangle$ and we're done.

(b) (18 pts) Let k_1, k_2 be symmetric, positive-definite kernels over $\mathbb{R}^D \times \mathbb{R}^D$, let $a \in \mathbb{R}^+$ be a positive real number, let $f : \mathbb{R}^D \to \mathbb{R}$ be a real-valued function, and let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial with positive coefficients. For each of the functions k below, state whether it is necessarily a positive-definite kernel. If you think it is, prove it. If you think it is not, give a counterexample.

NOTE: for the remainder of this problem, I'll be using abbreviations to cut down on writing:

- PD = positive definite
- PSD = positive semi-definite
- SPD = symmetric positive definite
- IP = inner product

i. $k(x, z) = k_1(x, z) + k_2(x, z)$

- In this case the function k **IS** necessarily a PD kernel.
- Let's prove that k's corresponding kernel matrix is PSD $\forall \boldsymbol{y_i} \in \mathbb{R}^d$.
- From our notes, we say that a kernel is PD if its kernel matrix is PSD $\forall \boldsymbol{y}$.
- Since we already have $\boldsymbol{x}$'s in the problem, I'm going to use y for the next part.

Let $K_1 = \begin{bmatrix} k_1(y_1, y_1) & k_1(y_1, y_2) & \cdots & k_1(y_1, y_n) \\ k_1(y_2, y_1) & k_1(y_2, y_2) & \cdots & k_1(y_2, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ k_1(y_n, y_1) & k_1(y_n, y_2) & \cdots & k_1(y_n, y_n) \end{bmatrix}$

Also, define the matrix K_2 similarly, replacing all k_1's with k_2's. These matrices – K_1 and K_2 – are the respective kernel matrices for $k_1(x, z)$ and $k_2(x, z)$ respectively.

- We know that $\left(\boldsymbol{y}^T \left[K_1 + K_2\right] \boldsymbol{y}\right) = \left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right) + \left(\boldsymbol{y}^T \left[K_2\right] \boldsymbol{y}\right)$. Since it is given that k_1 and k_2 are SPD kernels, we know that both kernels are PD kernels, and have kernel matrices (K_1 and K_2) that are PSD by definition.
- Therefore, we know that $\left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y} \geq 0\right)$ and $\left(\boldsymbol{y}^T \left[K_2\right] \boldsymbol{y} \geq 0\right)$

$\implies \boldsymbol{y}^T \left[K_1 + K_2\right] \boldsymbol{y} \geq 0 \implies [K_1 + K_2]$ is a PSD kernel matrix $\implies k_1(x, z) + k_2(x, z)$ is a PD kernel $\implies$ the function $k(x, z)$ is necessarily a PD kernel.

ii. $k(x,z) = k_1(x,z) - k_2(x,z)$

- In this case the function k **IS NOT** necessarily a PD kernel.
- Counterexample
  Define K_1 and K_2 (kernel matrices for $k_1(x,z)$ and $k_2(x,z)$) as we did in part (i.) of this problem. However, let's specify that $K_2 = 2 \cdot K_1$. Also, let's specify that K_1 and K_2 are both PD matrices.
  By definition, K_1 and K_2 must be PSD matrices, but we are looking at a specific case where K_1 and K_2 are both PD. Since {PD matrices} $\subset$ {PSD matrices} this means K_1 and K_2 are still (also) PSD matrices.
- Now, using the kernel matrix for $k(x,z)$, $[K_1 - K_2] = [K_1 - 2 \cdot K_1]$ examining
  $\left(\boldsymbol{y}^T \left[K_1 - 2 \cdot K_1\right] \boldsymbol{y}\right) = \left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right) - 2\left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right)$.
  Since we chose K_1 to be PD, we know that $\left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right) > 0 \ \forall \boldsymbol{y} \in \mathbb{R}^d$ and $\boldsymbol{y} \neq \boldsymbol{0}$.
  Let $c = \left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right) \implies (c > 0)$. Knowing that $c > 0$ we can reduce the value of the kernel matrix for $k(x,z)$ from $\left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right) - 2\left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right)$ down to $\left(c - 2c\right) = -c$.
  Since $(c > 0) \implies (-c < 0) \implies$ the kernel matrix $[K_1 - K_2] = [K_1 - 2 \cdot K_1]$ is not PSD $\implies$ k is not a PD kernel in this case.

iii. $k(x,z) = ak_1(x,z)$

- In this case the function k **IS** necessarily a PD kernel.
- For this function let's show that its corresponding kernel matrix is PSD.
- Let $K_1 = \begin{bmatrix} k_1(y_1,y_1) & k_1(y_1,y_2) & \cdots & k_1(y_1,y_n) \\ k_1(y_2,y_1) & k_1(y_2,y_2) & \cdots & k_1(y_2,y_n) \\ \vdots & \vdots & \ddots & \vdots \\ k_1(y_n,y_1) & k_1(y_n,y_2) & \cdots & k_1(y_n,y_n) \end{bmatrix}$
  By definition, since we know that $k_1(x,z)$ is an SPD kernel, we know that its kernel matrix K_1 is PSD for all $\boldsymbol{y}_i \in \mathbb{R}^d$, mathematically notated as $\left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right) \geq 0$.
- We know that multiplying $a \cdot K_1$ will give a PSD matrix since
  $\left(\boldsymbol{y}^T \left[a \cdot K_1\right] \boldsymbol{y}\right) = a \cdot \left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right)$ and we know that $\left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right) \geq 0$. Since a is defined to be a positive real number (in prompt), we know that $(a) \cdot (\text{“} \geq 0\text{”}) \geq 0$, and therefore $a \cdot \left(\boldsymbol{y}^T \left[K_1\right] \boldsymbol{y}\right) \geq 0 \implies \left(\boldsymbol{y}^T \left[a \cdot K_1\right] \boldsymbol{y}\right) \geq 0 \implies [a \cdot K_1]$ is a PSD matrix.
- Since $[a \cdot K_1]$ is the kernel matrix for $ak_1(x,z)$ and it is a PSD matrix, we know that $k(x,z) = ak_1(x,z)$ is necessarily a PD kernel.

iv. $k(x,z) = k_1(x,z)k_2(x,z)$

- In this case the function k **IS** necessarily a PD kernel.
- We will show this by proving $k(x,z)$ is an IP kernel.
- From the prompt, it's given that k_1 and k_2 are SPD kernels. Using the theorem (given in notes) that [k is an SPD kernel] $\iff$ [k is an IP kernel] we know that k_1 and k_2 are IP kernels. As such, k_1 and k_2 can each be expressed as an IP of feature maps.
- Using these facts, define $k_1(x,z) = [\phi^{(1)}(x)]^T[\phi^{(1)}(z)] = \sum_{j=1}^{d}\left(\phi_j^{(1)}(x)\right)\left(\phi_j^{(1)}(z)\right)$
 $= \left[\left(\phi_1^{(1)}(x)\right)\left(\phi_1^{(1)}(z)\right) + \left(\phi_2^{(1)}(x)\right)\left(\phi_2^{(1)}(z)\right) + \ldots + \left(\phi_d^{(1)}(x)\right)\left(\phi_d^{(1)}(z)\right)\right]$
- Similarly, $k_2(x,z) = \left[\left(\phi_1^{(2)}(x)\right)\left(\phi_1^{(2)}(z)\right)+\left(\phi_2^{(2)}(x)\right)\left(\phi_2^{(2)}(z)\right)+\ldots+\left(\phi_d^{(2)}(x)\right)\left(\phi_d^{(2)}(z)\right)\right]$
- Now, using these definitions, let's see what $k_1(x,z)k_2(x,z)$ is:
 $k_1(x,z)k_2(x,z) =$
 $\left[\left((\phi_1^{(1)}(x))(\phi_1^{(2)}(x))\right)\left((\phi_1^{(1)}(z))(\phi_1^{(2)}(z))\right)\right]+\left[\left((\phi_1^{(1)}(x))(\phi_2^{(2)}(x))\right)\left((\phi_1^{(1)}(z))(\phi_2^{(2)}(z))\right)\right]$
 $+\ldots+\left[\left((\phi_1^{(1)}(x))(\phi_d^{(2)}(x))\right)\left((\phi_d^{(1)}(z))(\phi_d^{(2)}(z))\right)\right]+$
 $\left[\left((\phi_2^{(1)}(x))(\phi_1^{(2)}(x))\right)\left((\phi_2^{(1)}(z))(\phi_1^{(2)}(z))\right)\right]+\left[\left((\phi_2^{(1)}(x))(\phi_2^{(2)}(x))\right)\left((\phi_2^{(1)}(z))(\phi_2^{(2)}(z))\right)\right]$
 $+\ldots+\left[\left((\phi_2^{(1)}(x))(\phi_d^{(2)}(x))\right)\left((\phi_2^{(1)}(z))(\phi_d^{(2)}(z))\right)\right]$
 $+\ldots+\left[\left((\phi_d^{(1)}(x))(\phi_1^{(2)}(x))\right)\left((\phi_d^{(1)}(z))(\phi_1^{(2)}(z))\right)\right]+\left[\left((\phi_d^{(1)}(x))(\phi_2^{(2)}(x))\right)\left((\phi_d^{(1)}(z))(\phi_2^{(2)}(z))\right)\right]$
 $+\ldots+\left[\left((\phi_d^{(1)}(x))(\phi_d^{(2)}(x))\right)\left((\phi_d^{(1)}(z))(\phi_d^{(2)}(z))\right)\right]$
- Thus, to show that $k_1(x,z)k_2(x,z)$ can be expressed as the inner product of two feature maps (and is hence an IP kernel), we can use the above expression of k_1k_2 to show $k_1(x,z)k_2(x,z) = [\psi(x)]^T[\psi(z)]$.
- From the above expression, $\psi(x) =$
 $\left[\left((\phi_1^{(1)}(x))(\phi_1^{(2)}(x))\right), \left((\phi_1^{(1)}(x))(\phi_2^{(2)}(x))\right), \ldots, \left((\phi_1^{(1)}(x))(\phi_d^{(2)}(x))\right),\right.$
 $\left((\phi_2^{(1)}(x))(\phi_1^{(2)}(x))\right), \left((\phi_2^{(1)}(x))(\phi_2^{(2)}(x))\right), \ldots, \left((\phi_2^{(1)}(x))(\phi_d^{(2)}(x))\right),$
 $\left.\ldots\ldots, \left((\phi_d^{(1)}(x))(\phi_1^{(2)}(x))\right), \left((\phi_d^{(1)}(x))(\phi_2^{(2)}(x))\right), \ldots, \left((\phi_d^{(1)}(x))(\phi_d^{(2)}(x))\right)\right]$
 NOTE: the feature map is purposely broken up onto different lines to more clearly illustrate the patterns associated with the indices of the ϕ functions.
- Similarly, $\psi(z)$ will look identical to $\psi(x)$, except all x's will be replaced with z's.
- Since we are able to express $k(x,z) = k_1(x,z)k_2(x,z) = \langle\psi(x),\psi(z)\rangle\ \forall x,z \in \mathbb{R}^d$ we know that $k(x,z)$ is an IP kernel. Using the theorem
 (k is an IP kernel) $\iff$ (k is an SPD kernel), k must be an SPD kernel. Since all SPD kernels are also PD kernels, k must necessarily be a PD kernel.

v. $k(x,z) = f(x)f(z)$

- In this case the function k **IS** necessarily a PD kernel.
- To begin, here is a definition from course notes:
  We say $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is an IP kernel is $\exists$ an IP space V and a feature map $\Phi : \mathbb{R}^d \to V$ such that $k(x,z) = \langle \Phi(x), \Phi(z) \rangle \ \forall x, z \in \mathbb{R}^d$.
- Let's define feature map $\Phi(x) = f(x)$. Since $f(x) : \mathbb{R}^d \to \mathbb{R}$, from the definition of the feature map above ($\Phi : \mathbb{R}^d \to V$) we can deduce that $\mathbb{R}$ is our IP space V. Similarly, define $\Phi(z) = f(z)$. This $f(z)$ also maps $\mathbb{R}^d \to V$, where V is $\mathbb{R}$.
- Now that we've defined $\Phi(x)$ and $\Phi(z)$, we can use the definition to go from $k(x,z) = \langle \Phi(x), \Phi(z) \rangle$ to $k(x,z) = f(x)f(z)$.
- Since $f(x) : \mathbb{R}^d \to \mathbb{R}$ and $f(z) : \mathbb{R}^d \to \mathbb{R}$, we know that the result of $f(x)f(z)$ is $a \cdot b = c$, where $[f(x) = a] \in \mathbb{R}$, $[f(z) = b] \in \mathbb{R}$, and $c \in \mathbb{R}$ since the space $\mathbb{R}$ is closed under multiplication. Therefore, we've found a feature map $\Phi : \mathbb{R}^d \to V$ (where V is $\mathbb{R}$) such that $k(x,z) = f(x)f(z) \ \forall x, z \in \mathbb{R}^d$.
- Hence, we've shown that $k(x,z)$ is an IP kernel. Invoking the theorem (k is an IP kernel) $\iff$ (k is an SPD kernel), we conclude that k is also an SPD kernel. Since all SPD kernels are also PD kernels, we conclude that k must necessarily be a PD kernel.

vi. $k(x,z) = p\left(k_1(x,z)\right)$

- In this case the function k **IS** necessarily a PD kernel.
- Definition of a polynomial with degree p is $p(t) = \sum_{i=0}^{p} a_i t^i$. For this problem we're given that $a_i \in \mathbb{R}^+ \ \forall i$.
- We can express $k(x,z)$ as $p(k_1(x,z))$, where $p(k_1(x,z)) = \sum_{i=0}^{p} a_i \big(k_1(x,z)\big)^i$ where $\sum_{i=0}^{p} a_i \big(k_1(x,z)\big)^i = a_0 \big(k_1(x,z)\big)^0 + a_1 \big(k_1(x,z)\big)^1 + a_2 \big(k_1(x,z)\big)^2 + \ldots + a_p \big(k_1(x,z)\big)^p$.
- We know, by definition in the prompt, that $k_1(x,z)$ is an SPD kernel. From **part (iv)** of this problem, we've shown the property that the product of two SPD kernels is a PD kernel. We can extend this property to each of the terms $\big(k_1(x,z)\big)^i$ in our sum. So we know that $\big(k_1(x,z)\big)^i$ is a PD kernel for $i \in \{1, 2, \ldots, p\}$.
- Now that we know this, we can apply the property shown in **part (iii)** of this problem to know that $a_i \big(k_1(x,z)\big)^i$ is a PD kernel, since $a_i \in \mathbb{R}^+$ and we've just shown that $\big(k_1(x,z)\big)^i$ is a PD kernel for $i \in \{1, 2, \ldots, p\}$.
- Now that we know each term $a_i \big(k_1(x,z)\big)^i$ is a PD kernel $\forall i$, we know that each term in the sum $p(k_1(x,z)) = \sum_{i=0}^{p} a_i \big(k_1(x,z)\big)^i = a_0 \big(k_1(x,z)\big)^0 + a_1 \big(k_1(x,z)\big)^1 + a_2 \big(k_1(x,z)\big)^2 + \ldots + a_p \big(k_1(x,z)\big)^p$ is a PD kernel. From **part (i)** of this problem, we've shown that the sum of PD kernels is a PD kernel. Applying this to our sum $\sum_{i=0}^{p} a_i \big(k_1(x,z)\big)^i$, we know that this is necessarily a PD kernel.