Homework for the Course
“Advanced Learninig Models”

Anja PANTOVIC
master MSIAM DS
anja.pantovic@grenoble-inp.org

Predrag PILIPOVIC
master MSIAM DS
predrag.pilipovic@grenoble-inp.org

1 Neural Networks

Let $\mathbf{X} = (x_{ij})_{ij}, i, j \in \{1, ..., 5\}$ denote the input of a convolutional layer with no bias. Let $\mathbf{W} = (w_{ij})_{ij}, i, j \in \{1, ..., 3\}$ denote the weights of the convolutional filter. Let $\mathbf{Y} = (y_{ij})_{ij}, i \in \{1, ..., I\}, j \in \{1, ..., J\}$ denote the output of the convolution operation.

1. **What is the output size (i.e. values of I and J) if:**

   (a) **the convolution has no padding and no stride?**

   The output size for convolution with no padding and no stride (stride = 1) is 3×3.

   (b) **the convolution has stride 2 and no padding?**

   The output size for stride 2 convolution when the padding is deactivated is 2×2.

   (c) **the convolution has no stride and padding 2?**

   The output size for convolution with no stride and padding 2 is 7×7.

2. **Let us suppose that we are in situation 1.(b) (i.e. stride 2 and no padding). Let us also assume that the output of the convolution goes through a ReLU activation, whose output is denoted by $\mathbf{Z} = (z_{ij})_{ij}, i \in \{1, ..., I\}, j \in \{1, ..., J\}$:**

   (a) **Derive the expression of the output pixels z_{ij} as a function of the input and the weights.**

   We have seen that $I = J = 2$ in our case. Let us see what will happen when we apply filter at the beginning of the layer, so we have

$$\begin{aligned} y_{11} &= w_{11}x_{11} + w_{12}x_{12} + w_{13}x_{13} \\ &\quad + w_{21}x_{21} + w_{22}x_{22} + w_{23}x_{23} \\ &\quad + w_{31}x_{31} + w_{32}x_{32} + w_{33}x_{33} \\ &= \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}x_{ij}. \end{aligned}$$

If we do the same thing for the rest y_{ij}, we will have

$$\begin{aligned}
y_{12} &= w_{11}x_{13} + w_{12}x_{14} + w_{13}x_{15} \\
&\quad + w_{21}x_{23} + w_{22}x_{24} + w_{23}x_{25} \\
&\quad + w_{31}x_{33} + w_{32}x_{34} + w_{33}x_{35} \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}x_{i,j+2}, \\
y_{21} &= w_{11}x_{31} + w_{12}x_{32} + w_{13}x_{33} \\
&\quad + w_{21}x_{41} + w_{22}x_{42} + w_{23}x_{43} \\
&\quad + w_{31}x_{51} + w_{32}x_{52} + w_{33}x_{53} \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}x_{i+2,j}, \\
y_{22} &= w_{11}x_{33} + w_{12}x_{34} + w_{13}x_{35} \\
&\quad + w_{21}x_{43} + w_{22}x_{44} + w_{23}x_{45} \\
&\quad + w_{31}x_{53} + w_{32}x_{54} + w_{33}x_{55} \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}x_{i+2,j+2}.
\end{aligned}$$

So, we can conclude that the general formula is

$$y_{lk} = \sum_{i=1}^{3}\sum_{j=1}^{3} w_{ij}x_{i+2(l-1),j+2(k-1)}.$$

Finally, we know that $z_{lk} = \sigma(y_{lk})$, where σ is ReLu activation function, more precisely $\sigma(x) = \max\{0, x\}$.

(b) **How many multiplications and additions are needed to compute the output (the forward pass)?**

As we saw in the first part of the question, for computing y_{lk} we need 9 multiplication and 8 additions. As there are 4 cells in the output of the convolution, we need $4\cdot(8+9)$ operations to compute the output of the convolution.

3. **Assume now that we are provided with the derivative of the loss w.r.t. the output of the convolution layer $\partial\mathcal{L}/\partial z_{ij}$, $\forall i \in \{1, ..., I\}, j \in \{1, ..., J\}$:**

(a) **Derive the expression of $\partial\mathcal{L}/\partial x_{ij}$, $\forall i, j \in \{1, ..., 5\}$.**

We will use the chain rule so we have

$$\frac{\partial\mathcal{L}}{\partial x_{ij}} = \sum_{l=1}^{3}\sum_{k=1}^{3} \frac{\partial\mathcal{L}}{\partial z_{lk}} \cdot \frac{\partial z_{lk}}{\partial y_{lk}} \cdot \frac{\partial y_{lk}}{\partial x_{ij}}.$$

We assumed to know $\partial\mathcal{L}/\partial z_{lk}$, so we need to compute the two last partial derivatives. We can easily see that

$$\frac{\partial z_{lk}}{\partial y_{lk}} = \begin{cases} 1, & y_{lk} > 0 \\ 0, & y_{lk} < 0 \end{cases}.$$

For the last partial derivative, we need to change the indexing in two sums in formula for y_{lk}, so we have

$$y_{lk} = \sum_{i=2l-1}^{3}\sum_{j=2k-1}^{3} w_{i-2(l-1),j-2(k-1)}x_{ij}.$$

Finally, we have that $\partial y_{lk}/\partial x_{ij} = w_{i-2(l-1),j-2(k-1)}$, which gives us

$$\frac{\partial\mathcal{L}}{\partial x_{ij}} = \sum_{l=1}^{3}\sum_{k=1}^{3} \frac{\partial\mathcal{L}}{\partial z_{lk}} \cdot \frac{\partial z_{lk}}{\partial y_{lk}} \cdot w_{i-2(l-1),j-2(k-1)}$$

(b) **Derive the expression of $\partial\mathcal{L}/\partial w_{ij}$, $\forall i, j \in \{1, ..., 3\}$.**
Similarly, we have

$$\frac{\partial\mathcal{L}}{\partial w_{ij}} = \sum_{l=1}^{3}\sum_{k=1}^{3} \frac{\partial\mathcal{L}}{\partial z_{lk}} \cdot \frac{\partial z_{lk}}{\partial y_{lk}} \cdot \frac{\partial y_{lk}}{\partial w_{ij}} = \sum_{l=1}^{3}\sum_{k=1}^{3} \frac{\partial\mathcal{L}}{\partial z_{lk}} \cdot \frac{\partial z_{lk}}{\partial y_{lk}} \cdot x_{i+2(l-1),j+2(k-1)}.$$

Let us now consider a fully connected layer, with two input and two output neurons, without bias and with a sigmoid activation. Let x_i, $i = 1, 2$ denote the inputs, and z_j , $j = 1, 2$ the output. Let w_{ij} denote the weight connecting input i to output j. Let us also assume that the gradient of the loss at the output $\partial\mathcal{L}/\partial z_j$, $j = 1, 2$ is provided.

4. **Derive the expressions for the following derivatives:**

(a) $\dfrac{\partial\mathcal{L}}{\partial x_i}$:

Firstly, we can introduce a sigmoid activation function

$$\sigma(x) = \frac{1}{1 + \exp(-x)},$$

from where we can derivate

$$\sigma'(x) = \frac{-\exp(-x)}{(1 + \exp(-x))^2}.$$

Also, we know that

$$z_j = \sigma(w_{1j}x_1 + w_{2j}x_2) = \frac{1}{1 + \exp(-(w_{1j}x_1 + w_{2j}x_2))}, \ \ j = 1, 2.$$

Again, using the chain rule, we get

$$\frac{\partial\mathcal{L}}{\partial x_i} = \sum_{j=1}^{2} \frac{\partial\mathcal{L}}{\partial z_j} \cdot \frac{\partial z_j}{\partial x_i} = \sum_{j=1}^{2} \frac{\partial\mathcal{L}}{\partial z_j} \cdot \frac{w_{ij}\exp(-(w_{1j}x_1 + w_{2j}x_2))}{(1 + \exp(-(w_{1j}x_1 + w_{2j}x_2)))^2}.$$

(b) $\dfrac{\partial\mathcal{L}}{\partial w_{ij}}$:
Similarly, we have

$$\frac{\partial\mathcal{L}}{\partial w_{ij}} = \frac{\partial\mathcal{L}}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}} = \frac{\partial\mathcal{L}}{\partial z_j} \cdot \frac{x_i\exp(-(w_{1j}x_1 + w_{2j}x_2))}{(1 + \exp(-(w_{1j}x_1 + w_{2j}x_2)))^2}.$$

But this time without the sum, because w_{ij} depends only on z_j.

(c) $\dfrac{\partial^2\mathcal{L}}{\partial w_{ij}^2}$:
Having in mind that $\partial\mathcal{L}/\partial z_j$ is a function of w_{ij}, we have

$$\frac{\partial^2\mathcal{L}}{\partial w_{ij}^2} = \frac{\partial}{\partial w_{ij}}\left(\frac{\partial\mathcal{L}}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}}\right) = \frac{\partial^2\mathcal{L}}{\partial z_j \partial w_{ij}} \cdot \frac{\partial z_j}{\partial w_{ij}} + \frac{\partial\mathcal{L}}{\partial z_j} \cdot \frac{\partial^2 z_j}{\partial w_{ij}^2}.$$

The only thing left to compute is $\partial^2 z_j/\partial w_{ij}^2$, because we assumed to know $\partial\mathcal{L}/\partial z_j$, hence we will know $\partial^2\mathcal{L}/\partial z_j\partial w_{ij}$. We will need the second derivative of σ, so

$$\sigma''(x) = \frac{2\exp(-2x)}{(1 + \exp(-x))^3} - \frac{\exp(-x)}{(1 + \exp(-x))^2}.$$

Finally, we have

$$\begin{aligned}\frac{\partial^2 z_j}{\partial w_{ij}^2} &= \frac{\partial}{\partial w_{ij}}\left(\frac{x_i\exp(-(w_{1j}x_1 + w_{2j}x_2))}{(1 + \exp(-(w_{1j}x_1 + w_{2j}x_2)))^2}\right)\\ &= x_i^2 \cdot \sigma''\left(w_{1j}x_1 + w_{2j}x_2\right).\end{aligned}$$

(d) $\dfrac{\partial^2 \mathcal{L}}{\partial w_{ij} w_{i'j'}}$, $i \neq i'$, $j \neq l'$: Again, from the chain rule we have

$$\frac{\partial^2 \mathcal{L}}{\partial w_{ij} \partial w_{i'j'}} = \frac{\partial}{\partial w_{ij}} \left(\frac{\partial \mathcal{L}}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{i'j'}} \right) = \frac{\partial^2 \mathcal{L}}{\partial z_j \partial w_{ij}} \cdot \frac{\partial z_j}{\partial w_{i'j'}} + \frac{\partial \mathcal{L}}{\partial z_j} \cdot \frac{\partial^2 z_j}{\partial w_{ij} \partial w_{i'j'}}.$$

But now, we can see from the previous exercises that the last term will always be zero for $j \neq j'$, because we do not have j' in $\partial z_j / \partial w_{ij}$, so we will have just the first term.

(e) **The elements in (c) and (d) are the entries of the Hessian matrix of $\mathcal{L}$ w.r.t the weight vector. Imagine now that storing the weights of a network requires 40 MB of disk space: how much would it require to store the gradient? And the Hessian?**
If storing the weights of a network requires 40 MB, and we have 4 weights, it means that one representation of number will require 10 MB. We know that for gradient we have 4 elements, so we can conclude it will require 40 MB, as well. Since Hessian is symmetric matrix we need to store just the upper triangle, which means we need $\frac{n(n+1)}{2}$ for the matrix of the size $n \times n$. In our case, $n = 4$, so we need 10 elements, or 100 MB.

2 Conditionally Positive Definite Kernels

Let $\mathcal{X}$ be a set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called conditionally positive definite (c.p.d.) if and only if it is symmetric and satisfies:

$$\sum_{i,j}^{n} a_i a_j k(x_i, x_j) \geq 0$$

for any $n \in \mathbb{N}$, $(x_1, x_2, ..., x_n) \in \mathcal{X}^n$ and $(a_1, a_2, ..., a_n) \in \mathbb{R}^n$ with $\sum_{i=1}^{n} a_i = 0$.

1. **Show that a positive definite (p.d.) function is c.p.d.**.
   Let k be a positive definite function. This means that it is symmetric and for any $n \in \mathbb{N}$, $(x_1, ..., x_n) \in \mathcal{X}^n$, $(a_1, ..., a_n) \in \mathbb{R}^n$

   $$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

   Since this holds for any $(a_1, ..., a_n) \in \mathbb{R}^n$, $n \in \mathbb{N}$, it holds for $(a_1, ..., a_n) \in \mathbb{R}^n$ with $\sum_{i=1}^{n} a_i = 0$ as well. Hence, any positive definite function is conditionally positive definite.

2. **Is a constant function p.d.? Is it c.p.d.?**
   Let k be a constant function

   $$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x_i, x_j) &\mapsto c \end{aligned}$$

   Since $k(x_i, x_j) = k(x_j, x_i) = c$, for all $(x_i, x_j) \in \mathcal{X} \times \mathcal{X}$, the symmetry holds.
   Let $n \in \mathbb{N}$, $(x_1, ..., x_n) \in \mathcal{X}^n$ and $(a_1, ..., a_n) \in \mathbb{R}^n$.

   $$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) = c \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j = c \left(\sum_{i=1}^{n} a_i \right)^2 \geq 0 \iff c \geq 0$$

   So k is positive definite if and only if $c \geq 0$. We already know that k is c.p.d. when $c > 0$ from the first question. Let us see if conditional positive definiteness of k works for any c.

Let $n \in \mathbb{N}, (x_1, ..., x_n) \in \mathcal{X}^n$ and $(a_1, ..., a_n) \in \mathbb{R}^n$ such that $\sum\limits_{i=1}^{n} a_i = 0$. We have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = c \cdot \left(\sum_{i=1}^{n} a_i\right)^2 = 0$$

Hence, k is a conditionally positive definite function for any $c \in \mathbb{R}$.

3. **If $\mathcal{X}$ is a Hilbert space, then is $k(x, y) = -||x - y||^2$ p.d.? Is it c.p.d.?**

Firstly, we can see that

$$k(x, y) = -||x - y||^2 = -||x||^2 - ||y||^2 + 2\langle x, y\rangle = -||y - x||^2 = k(y, x),$$

hence, k is symmetric. Let $n \in \mathbb{N}, (x_1, ..., x_n) \in \mathcal{X}^n$ and $(a_1, ..., a_n) \in \mathbb{R}^n$ such that $\sum\limits_{i=1}^{n} a_i = 0$.

$$\begin{aligned}\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) &= -\sum_{i,j=1}^{n} a_i a_j ||x_i||^2 - \sum_{i,j=1}^{n} a_i a_j ||x_j||^2 + 2\sum_{i,j=1}^{n} a_i a_j \langle x_i, x_j\rangle \\ &= -\underbrace{\sum_{j=1}^{n} a_j}_{0} \sum_{i=1}^{n} a_i ||x_i||^2 - \underbrace{\sum_{i=1}^{n} a_i}_{0} \sum_{j=1}^{n} a_j ||x_j||^2 + 2\sum_{i,j}^{n} a_i a_j \langle x_i, x_j\rangle \\ &= 2\sum_{i,j=1}^{n} a_i a_j \langle x_i, x_j\rangle = 2\left\|\sum_{i=1}^{n} a_i x_i\right\|^2 \geq 0.\end{aligned}$$

Thus, k is a conditionally positive definite function. We can see intuitively that k is not positive definite function. To prove that we need one counterexample. For that reason we can use $n = 2$, so we have

$$\begin{aligned}\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) &= a_1^2 \underbrace{k(x_1, x_1)}_{0} + 2a_1 a_2 k(x_1, x_2) + a_2^2 \underbrace{k(x_2, x_2)}_{0} \\ &= -2a_1 a_2 \|x_1 - x_2\|^2 \leq 0,\end{aligned}$$

where a_1 and a_2 are positive.

4. **Let $\mathcal{X}$ be a nonempty set, and $x_0 \in \mathcal{X}$ a point. For any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the function defined by**

$$\tilde{k}(x, y) = k(x, y) - k(x_0, x) - k(x_0, y) + k(x_0, x_0).$$

Show that k is c.p.d. if and only if $\tilde{k}$ is p.d.

($\Rightarrow$) Suppose $\tilde{k}$ is positive definite, i.e. for all $n \in \mathbb{N}$, $(a_1, ..., a_n) \in \mathbb{R}^n$, $(x_1, ..., x_n) \in \mathcal{X}^n$, we have

$$\sum_{i,j}^{n} a_i a_j \tilde{k}(x, y) \geq 0.$$

Let us fix $n \in \mathbb{N}$, and choose $(a_1, ..., a_n) \in \mathbb{R}^n$ such that $\sum\limits_{i=1}^{n} a_i = 0$. For $(x_1, ..., x_n) \in \mathcal{X}^n$, we have:

$$\begin{aligned}\sum_{i,j=1}^{n} a_i a_j \tilde{k}(x,y) &= \sum_{i,j=1}^{n} a_i a_j (k(x_i,x_j) - k(x_0,x_i) - k(x_0,x_j) + k(x_0,x_0)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i,x_j) - \underbrace{\sum_{j=1}^{n} a_j}_{0} \sum_{i=1}^{n} a_i k(x_0,x_i) \\ &- \underbrace{\sum_{i=1}^{n} a_i}_{0} \sum_{j=1}^{n} a_j k(x_0,x_j) + k(x_0,x_0) \underbrace{\left(\sum_{i=1}^{n} a_i\right)^2}_{0} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i,x_j).\end{aligned}$$

Hence, positive definiteness of $\tilde{k}$ implies conditional positive definiteness of k.

($\Leftarrow$) Let us now suppose that k is conditionally positive definite, i.e. for any $n \in \mathbb{N}$, $(x_0, ..., x_n) \in \mathcal{X}^{n+1}$, and for any $(a_0, ..., a_n) \in \mathbb{R}^{n+1}$ such that $\sum\limits_{i=0}^{n} a_i = 0$, we have

$$\sum_{i,j=0}^{n} a_i a_j k(x_i,x_j) \geq 0.$$

We want to show that for all $n \in \mathbb{N}$, $(a_1, ..., a_n) \in \mathbb{R}^n$, $(x_1, ..., x_n) \in \mathcal{X}^n$, it is

$$\sum_{i,j=1}^{n} a_i a_j \tilde{k}(x_i,x_j) \geq 0.$$

To use an assumption, we need to introduce $a_0 := -\sum\limits_{i=1}^{n} a_i$, so we get $\sum\limits_{i=0}^{n} a_i = 0$. Now, we have

$$\begin{aligned}\sum_{i,j=1}^{n} a_i a_j \tilde{k}(x_i,x_j) &= \sum_{i,j=1}^{n} a_i a_j k(x_i,x_j) - \sum_{i,j=1}^{n} a_i a_j k(x_0,x_j) \\ &- \sum_{i,j=1}^{n} a_i a_j k(x_0,x_i) + \sum_{i,j=1}^{n} a_i a_j k(x_0,x_0) \\ &= \sum_{i,j=1}^{n} a_i a_j k(x_i,x_j) - \underbrace{\sum_{i=1}^{n} a_i}_{-a_0} \sum_{j=1}^{n} a_j k(x_0,x_j) \\ &- \underbrace{\sum_{j=1}^{n} a_j}_{-a_0} \sum_{i=1}^{n} a_i k(x_i,x_0) + \underbrace{\left(\sum_{i=1}^{n} a_i\right)^2}_{a_0^2} k(x_0,x_0) \\ &= \sum_{i,j=0}^{n} a_i a_j \tilde{k}(x_i,x_j) \geq 0.\end{aligned}$$

Finally, we can conclude that k is c.p.d. if and only if $\tilde{k}$ is p.d.

5. **Let k be a c.p.d. kernel on $\mathcal{X}$ such that $k(x,x) = 0$ for any $x \in \mathcal{X}$. Show that there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that, for any $x, y \in \mathcal{X}$,**

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

Let k be a c.p.d. kernel on $\mathcal{X}$ such that $k(x,x) = 0$, for any $x \in \mathcal{X}$. From the previous question, we know how to construct the feature map from k which is positive definite. Let

$$\tilde{k}(x,y) := \frac{1}{2}\big(k(x,y) - k(x_0,x) - k(x_0,y) + k(x_0,x_0)\big),$$

where $x_0 \in \mathcal{X}$ is fixed. Then $\tilde{k}$ is p.d. and hence we can use Aronszajn theorem, which says that there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi : X \to \mathcal{H}$ such that, for any $x, y \in \mathcal{X}$

$$\tilde{k}(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}\,.$$

Now, only thing left to be proven is

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

For this part, we will use the assumption $k(x,x) = 0$, for any $x \in \mathcal{X}$. We have

$$\begin{aligned}
||\Phi(x) - \Phi(y)||^2 &= \Phi^2(x) - 2\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} + \Phi^2(y) \\
&= \tilde{k}(x,x) - 2\tilde{k}(x,y) + \tilde{k}(y,y) \\
&= \frac{1}{2}\Bigg(\underbrace{k(x,x)}_{0} - k(x_0,x) - k(x_0,x) + \underbrace{k(x_0,x_0)}_{0} \\
&\qquad - 2k(x,y) + 2k(x_0,x) + 2k(x_0,y) + \underbrace{2k(x_0,x_0)}_{0} \\
&\qquad + \underbrace{k(y,y)}_{0} - k(x_0,y) - k(x_0,y) + \underbrace{k(x_0,x_0)}_{0} \Bigg) \\
&- k(x_0,x) - k(x,y) + k(x_0,x) + k(x_0,y) - k(x_0,y) \\
&= -k(x,y).
\end{aligned}$$

6. **Show that if k is c.p.d., then the function $\exp(tk(x,y))$ is p.d. for all $t \geq 0$.**

Firstly, we need to show that the product of two p.d. functions is also a p.d. function. Let k_1, and k_2 be two p.d. functions, and $[k_1]$, $[k_2]$ the positive semidefinite similarity matrices of k_1, k_2, respectively. Since $[k_2]$ is symmetric positive semidefinite, it has a symmetric positive semidefinite square root S, i.e. $[k_2] = S^2$, more precisely

$$[k_2]_{ij} = \sum_{l=1}^{n} S_{il}S_{lj} = \sum_{l=1}^{n} S_{il}S_{jl},$$

for all $i, j = 1, 2, ..., n$. Therefore, for any $(a_1, a_2, ..., a_n) \in \mathbb{R}^n$, we have

$$\begin{aligned}
\sum_{i,j=1}^{n} a_i a_j [k_1]_{ij}[k_2]_{ij} &= \sum_{i,j=1}^{n} a_i a_j [k_1]_{ij} \left(\sum_{l=1}^{n} S_{il}S_{jl} \right) \\
&= \sum_{l=1}^{n} \underbrace{\left(\sum_{i,j=1}^{n} \underbrace{a_i S_{il}}_{\tilde{a}_i} \underbrace{a_j S_{jl}}_{\tilde{a}_j} [k_1]_{ij} \right)}_{\geq 0} \geq 0.
\end{aligned}$$

We used that the inner sum is nonnegative using the weights $(\tilde{a}_1, \tilde{a}_2, ...\tilde{a}_n) \in \mathbb{R}^n$ and the fact that k_1 is p.d. kernel. So, we proved that the similarity matrix $[k]$, defined by $[k]_{ij} = [k_1]_{ij}[k_2]_{ij}$ is positive semidefinite. The symmetry of kernel k is immediate consequence of the symmetry of k_1 and k_2. This means, we proved that the product kernel k is indeed a p.d. kernel.

We also need to prove that if a given sequence of p.d. kernels $\{k_n\}_{n\in\mathbb{N}}$ pointwise converges to k, i.e. for all $x, y \in \mathcal{X}$ is

$$\lim_{n\to\infty} k_n(x,y) = k(x,y),$$

then k is a p.d. kernel. First of all, by uniqueness of the limit, we can indeed define the pointwise limit k as a function. It is symmetric as a immediate consequence of the symmetry of all the kernels k_n. Let $m \in \mathbb{N}$, $(x_1, x_2, ..., x_m) \in \mathcal{X}$ and $(a_1, a_2, ..., a_m) \in \mathbb{R}^m$, then we have

$$\begin{aligned}
\sum_{i,j=1}^{m} a_i a_j k(x_i, x_j) &= \sum_{i,j}^{m} a_i a_j \lim_{n\to\infty} k_n(x_i, x_j \\
&= \sum_{i,j=1}^{m} \lim_{n\to\infty} a_i a_j k_n(x_i, x_j) \\
&= \lim_{n\to\infty} \underbrace{\left(\sum_{i,j}^{m} a_i a_j k_n(x_i, x_j) \right)}_{\geq 0} \geq 0.
\end{aligned}$$

This proves that k is also p.d. kernel.

Now, we can go back to our assignment. If k is c.p.d. we know that we can associate k with p.d. $\tilde{k}$, such that

$$\tilde{k}(x, y) = k(x, y) - k(x_0, x) - k(x_0, y) + k(x_0, x_0),$$

for any $x, y \in \mathcal{X}$, and some point $x_0 \in \mathcal{X}$. From the previous line it follows

$$k(x, y) = \tilde{k}(x, y) + k(x_0, x) + k(x_0, y) - k(x_0, x_0),$$

or

$$\exp(tk(x, y)) = \underbrace{\exp(t\tilde{k}(x, y))}_{k_1} \underbrace{\exp(tk(x_0, x)) \exp(tk(x_0, y)) \exp(-tk(x_0, x_0))}_{k_2}.$$

We know that $\tilde{k}$ is p.d., and using the Taylor expansion we can write

$$\exp(t\tilde{k}(x, y)) = \sum_{m=0}^{\infty} \frac{(t\tilde{k}(x, y))^m}{m!} = \lim_{m\to\infty} \sum_{i=0}^{m} \frac{(t\tilde{k}(x, y))^i}{i!}.$$

On the right hand side we have limit of the sum of the product of p.d., which means that $\exp(t\tilde{k}(x, y))$ is also p.d (obviously, sum of two p.d. is again p.d.). On the other hand, for all $(a_1, a_2, ..., a_n) \in \mathbb{R}^n$ we have

$$\begin{aligned}
&\sum_{i,j=1}^{n} a_i a_j \exp(tk(x_0, x_i)) \exp(t(x_0, x_j)) \exp(-tk(x_0, x_0)) \\
&= \exp(-tk(x_0, x_0)) \sum_{i,j=1}^{n} a_i \exp(tk(x_0, x_i)) a_j \exp(t(x_0, x_j)) \\
&= \exp(-tk(x_0, x_0)) \left\| \sum_{i=1}^{n} a_i \exp(tk(x_0, x_i)) \right\|^2 \geq 0.
\end{aligned}$$

Finally, we proved that k_1 and k_2 are p.d. hence $\exp(tk)$ is p.d. as a product of k_1 and k_2.

7. **Conversely, show that if the function $\exp(tk(x, y))$ is p.d. for any $t \geq 0$, then k is c.p.d.**
   We know that

$$k(x, y) = \lim_{t\to 0} \frac{\exp(tk(x, y)) - 1}{t},$$

for all $x, y \in \mathcal{X}$. We assummed that $\exp(tk)$ is p.d. so it must be c.p.d. also. Now, for any $n \in \mathbf{n}$, $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$, and any $(a_1, a_2, ..., a_n) \in \mathbb{R}^n$ such that $\sum\limits_{i=1}^{n} a_i = 0$, we have

$$\sum_{i,j=1}^{n} a_i a_j \frac{\exp(tk(x_i, x_j)) - 1}{t} = \frac{1}{t} \underbrace{\sum_{i,j=1}^{n} a_i a_j \exp(tk(x_i, x_j))}_{\geq 0} - \underbrace{\sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j}_{0} \geq 0.$$

Which means that $\frac{\exp(tk)-1}{t}$ is c.p.d. And finally,

$$\begin{aligned}\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i,j=1}^{n} a_i a_j \lim_{t\to 0} \frac{\exp(tk(x_i, x_j)) - 1}{t} \\ &= \lim_{t\to 0} \sum_{i,j=1}^{n} a_i a_j \frac{\exp(tk(x_i, x_j)) - 1}{t} \geq 0,\end{aligned}$$

so k is c.d.p.

8. **Show that the shortest-path distance on a tree is c.p.d over the set of vertices (a tree is an undirected graph without loops. The shortest-path distance between two vertices is the number of edges of the unique path that connects them). Is the shortest-path distance over graphs c.p.d. in general?**

Let $G = (V, E)$ a tree (V being a set of vertices and E set of edges) and $x_0 \in V$ its root. Let us represent each node $x \in V$ by $\Phi(x)$, as follows

$$\Phi : V \to \mathbb{R}^{|E|}$$

such that

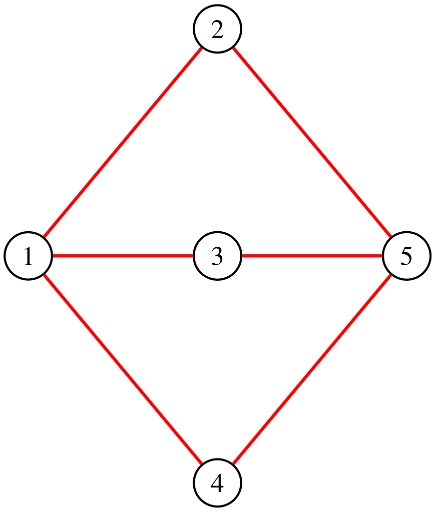
$$\Phi_i(x) = \begin{cases} 1, & \text{if is the } i\text{-th edge is in the path between } x \text{ and } x_0 \\ 0, & \text{otherwise} \end{cases}$$

We know that for each vertex $x \in V$, the path to x_0 is unique. Then, the graph distance $d_G(x, y)$ between any two vertices x and y (the length of the shortest path between x and y) is given by

$$d_G(x, y) = \|\Phi(x) - \Phi(y)\|^2.$$

In problem 2.3 we have seen that $-d_G$ is c.p.d. Now, using the previous we can conclude that $\exp(-td_G(x, y))$ is p.d. for all $t \geq 0$.

On the other hand, in general graphs do not have the property that $-d_G$ is c.p.d. We can see that on the counterexample. Let us look at the graph below



We can write down its shortest-path distance matrix (it is 5×5 matrix):

$$[d_G] = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 \\ 1 & 0 & 2 & 2 & 1 \\ 1 & 2 & 0 & 2 & 1 \\ 1 & 2 & 2 & 0 & 1 \\ 2 & 1 & 1 & 1 & 0 \end{bmatrix}$$

In order for shortest distance to correspond to a c.p.d function, $\exp(-td_G(x,y))$ must be p.d. for all $t \geq 0$. We can write down matrix $[\exp(-td_G(x,y))]$ and we shell show that it is not positive semi definite.
We have

$$[d_G] = \begin{bmatrix} 1 & e^{-t} & e^{-t} & e^{-t} & e^{-2t} \\ e^{-t} & 1 & e^{-2t} & e^{-2t} & e^{-t} \\ e^{-t} & e^{-2t} & 1 & e^{-2t} & e^{-t} \\ e^{-t} & e^{-2t} & e^{-2t} & 1 & e^{-t} \\ e^{-2t} & e^{-t} & e^{-t} & e^{-t} & 1 \end{bmatrix}$$

We can use Sylvester's criterion to show that this matrix is not positive semi definite. Let us calculate the corresponding determinant.

$$\begin{vmatrix} 1 & e^{-t} & e^{-t} & e^{-t} & e^{-2t} \\ e^{-t} & 1 & e^{-2t} & e^{-2t} & e^{-t} \\ e^{-t} & e^{-2t} & 1 & e^{-2t} & e^{-t} \\ e^{-t} & e^{-2t} & e^{-2t} & 1 & e^{-t} \\ e^{-2t} & e^{-t} & e^{-t} & e^{-t} & 1 \end{vmatrix} = e^{-10t}(e^{2t}-2)(e^{2t}-1)^4$$

Let $t = 0.2$. Then the determinant reads

$$e^{-10t}(e^2t-2)(e^2t-1)^4 = \underbrace{e^{-2}}_{>0}\underbrace{(e^{0.4}-2)}_{<0}\underbrace{(e^{0.4}-1)^4}_{>0} < 0$$

Hence, we can conclude that the shortest-path distance over graphs is not c.p.d. in general.