

Exercise 9.1

Reminder: A is pos. sem. def. iff $\forall z \in \mathbb{R}^m$ holds $z^T A z \geq 0$.

i. Is $k(x, y) = k_1(x, y) + k_2(x, y)$ a kernel if k_1 and k_2 are?

Take arbitrary n points from X . Now G_1 and G_2 are valid Gram-matrices w.r.t. k_1 and k_2 , and thus pos. sem. def. The candidate Gram-matrix G for k is clearly $G_1 + G_2$. Now

$$z^T(G)z = z^T(G_1 + G_2)z = z^T G_1 z + z^T G_2 z \geq 0. \quad (136)$$

Hence G is pos. sem. def., k is symmetric (as a sum of two symmetric functions), and its continuous as a sum of two continuous functions (by assumption). Now theorem 4.5 shows that k is a kernel.

ii. Is $k(x, y) = a k_1(x, y), \forall a > 0$ a kernel?

$$z^T(aG)z = a(z^T G z) \geq 0, \quad (137)$$

the result follows as before.

iii. Let $k(x, y) = \langle \phi(x), \phi(y) \rangle$ for the original ϕ . Lets expand the dot product related to the new mapping $\tilde{\phi}$,

$$\langle \tilde{\phi}(x), \tilde{\phi}(y) \rangle = \left\langle \frac{\phi(x)}{\|\phi(x)\|}, \frac{\phi(y)}{\|\phi(y)\|} \right\rangle \quad (138)$$

$$= \frac{1}{\|\phi(x)\|} \frac{1}{\|\phi(y)\|} \langle \phi(x), \phi(y) \rangle \quad (139)$$

$$= \frac{1}{\sqrt{k(x, x)}} \frac{1}{\sqrt{k(y, y)}} k(x, y) \quad (140)$$

$$= \frac{k(x, y)}{(k(x, x)k(y, y))^{\frac{1}{2}}} = \tilde{k}(x, y). \quad (141)$$

Hence \tilde{k} is a kernel as we could explicitly construct the related dot product of the new feature mapping, inheriting the suitable space from the original mapping ϕ .

□

Exercise 9.2

We know from (4.7.1)-(4.7.3) of the lecture notes that kernel sums, scalar multiplies and products of kernels are kernels.

i. It appears we have to restrict ourselves to polynomials with positive coefficients.

First, show that any single monomial of the polynomial is a kernel. This can be done by induction: We know k^1 is a kernel. Now suppose that $k' = k^n$ is a kernel. Due to (4.7.3), $k'k = k^{n+1}$ is a kernel. This completes the induction w.r.t. exponentiation of a kernel. Now (4.7.2) tells us that $ak' = ak^n$ is a kernel for all $a > 0$, as k' is a kernel. Finally, we know from (4.7.1) that $k_1 + k_2$ is a kernel for all kernels k_1, k_2 . Similar inductive argument shows that finite sums of kernels are kernels. This shows that we can construct any polynomial out of k while still retaining the kernel property.

ii. It can be taken as known that

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \forall x \in \mathfrak{R}. \tag{142}$$

Hence the required e^k is an infinite sum of kernels (the monomials are kernels due to previous exercise). In terms of Gram matrices, $G = \lim_{n \rightarrow \infty} \sum_1^n G_i$. We need to show that G is pos. sem. def. We have an infinite sum of non-negative terms

$$z^T G z = z^T G_1 z + z^T G_2 z + \dots \tag{143}$$

The limit is clearly non-negative as each of the terms are.

iii. $k(x, y) = \exp(-\|x - y\|^2 / (2\sigma^2))$, where $\|z\|^2 = \langle z, z \rangle = z^T z$. Is it a kernel?

First, lets show $\exp(x^T y / \sigma^2)$ is a kernel. This is true because $\langle \cdot, \cdot \rangle$ is a (linear) kernel, αk is a kernel (where $\alpha = 1/\sigma^2$), $\exp(k)$ is a kernel due to the previous part. We also know that a normalized kernel is a kernel (prev. ex.), so we normalize the exp kernel. A few manipulations show the result as follows,

$$\frac{\exp(\frac{x^T z}{\sigma^2})}{\exp(\frac{x^T x}{\sigma^2})^{\frac{1}{2}} \exp(\frac{z^T z}{\sigma^2})^{\frac{1}{2}}} = \dots \tag{144}$$

$$= \exp(\frac{x^T z}{\sigma^2}) \exp(\frac{x^T x}{\sigma^2} + \frac{z^T z}{\sigma^2})^{-\frac{1}{2}} \tag{145}$$

$$= \exp(-\frac{1}{2\sigma^2}(x^T x - x^T z + z^T z - x^T z)) \tag{146}$$

$$= \exp(-\frac{1}{2\sigma^2}(x^T(x - z) + z^T(z - x))) \tag{147}$$

$$= \exp(-\frac{1}{2\sigma^2}(x - z)^T(x - z)) \tag{148}$$

$$= \exp(-\frac{1}{2\sigma^2}\|x - z\|^2) \tag{149}$$

□

Exercise 9.3

i. The optimization problem can be written as

$$\begin{aligned} \min \quad & x^2 \\ \text{s.t.} \quad & (x-2)^2 - 1 \leq 0 \end{aligned}$$

The related Lagrangian is

$$L(x, \lambda) = x^2 + \lambda((x-2)^2 - 1) \quad (150)$$

The KKT conditions can be written as

$$\frac{\delta L}{\delta x} = 2x + \lambda(2x - 4) = 0 \quad (151)$$

$$(x-2)^2 - 1 \leq 0 \quad (152)$$

$$\lambda \geq 0 \quad (153)$$

$$\lambda((x-2)^2 - 1) = 0 \quad (154)$$

Suppose $\lambda \neq 0$. Then from condition 4 follows that $(x-2)^2 - 1 = 0 \Rightarrow x = 1$. Inserting this to the first condition reveals $2 + \lambda(2 - 4) = 0 \Rightarrow \lambda = 1$. The pair $(x = 1, \lambda = 1)$ can be seen to fulfill the KKT conditions, so we are done (supposing $\lambda = 0$ would result in condition 2 being violated).

The dual is specified as $d(\lambda) = \inf_x L(x, \lambda)$. Solving the derivative in the first KKT condition for x results in $x = \frac{2\lambda}{1+\lambda}$. Inserting this back to the Lagrangian and simplifying a little, results in dual

$$d(\lambda) = -\frac{\lambda(\lambda-3)}{1+\lambda} \quad (155)$$

(inserting $\lambda = 1$ to d returns 1).

ii. In the case $g(x) \leq 8$ the second and last conditions change to

$$(x-2)^2 - 8 \leq 0 \quad (156)$$

$$\lambda((x-2)^2 - 8) = 0 \quad (157)$$

The choice $\lambda \neq 0$ allows no solutions. Hence, $\lambda = 0$. Solving for this choice results in $x = 0$. With the new Lagrangian L' (omitted) the dual is $-\frac{4\lambda(2\lambda+1)}{1+\lambda}$, solved as previously.

Plotting the situation graphically instantly reveals the intuitive acceptability of these results: with $g(x) \leq 8$ the feasible region is large enough to include the lowest point of the figure $f(x) = x^2$, that is, it is no longer constraining the optimization. In the first part the $g(x) \leq 1$ constraint limited $x = 0$ outside the feasible region.

□

Exercise 9.4

In this exercise, W_t is the set of all linear models with the squared error for current example (x, y) less than R^2 . We wish to choose our new model w from this set so that it is as close as possible to our current model w_t . Drawing a picture can help here too.

This can be formulated as an optimization problem,

$$\begin{aligned} \min \quad & \|w_t - w\|^2 \\ \text{s.t.} \quad & (w^T x - y)^2 \leq R^2. \end{aligned}$$

This is a convex problem, the functions are differentiable and due to Slater's theorem, strong duality holds (we can satisfy its conditions), so the problem seems suitable for a KKT style solution. The Lagrangian is

$$L(w, \lambda) = \|w_t - w\|^2 + \lambda((w^T x - y)^2 - R^2). \quad (158)$$

$\nabla_w L = 0$ gives the first KKT condition, and writing out the rest,

$$-2w_t + w + 2\lambda((w^T x - y)x) = 0 \quad (159)$$

$$(w^T x - y)^2 - R^2 \leq 0 \quad (160)$$

$$\lambda \geq 0 \quad (161)$$

$$\lambda((w^T x - y)^2 - R^2) = 0. \quad (162)$$

Supposing $\lambda = 0$ results in the solution $w = w_t$. This means that w_t is already in the feasible region, and naturally it is the closest one to itself.

If $\lambda \neq 0$, we know from the last KKT condition that

$$(w^T x - y)^2 - R^2 = 0 \Leftrightarrow w^T x - y = kR, k \in \{-1, 1\}. \quad (163)$$

Inserting this to the first KKT condition gives

$$w = w_t - \lambda k R x. \quad (164)$$

Inserting this back to (163) returns after a few manipulations

$$\lambda = \frac{-kR - y + w_t^T x}{kR x^T x} \quad (165)$$

which we can insert to (164) to get the update rule

$$w = w_t + \frac{(kR + y - w_t^T x)x}{x^T x}. \quad (166)$$

Finally, we can solve k from the second KKT condition, leading to the selection of $k = 1$ if $w_t^T x - y \leq -R$ and -1 otherwise.

Note the close resemblance of the solution to the one found in exercise 3.4. The geometric intuition should be similar.

□

Exercise 9.5

This problem might be solvable more easily thru convexity, Jensen, demonstrating lower bound of 0 and showing that choice $p = q$ meets the bound. However, here we take another route to drill the KKT techniques.

Reformulating, we need to solve the problem

$$\begin{aligned} \min \quad & p^T \ln \frac{p}{q} \\ \text{s.t.} \quad & -p_i \leq 0, i \in \{1, \dots, n\} \\ & \sum p_i - 1 = 0 \end{aligned}$$

As a difference to the previous exercises, we now have n inequality constraints and one equality constraint. The Lagrangian can be written as

$$L(p, \alpha, \beta) = \sum_i p_i \ln \frac{p_i}{q_i} - \sum_i \alpha_i p_i + \beta (\sum_i p_i - 1) \quad (167)$$

$$= \sum_i p_i (\ln p_i - \ln q_i - \alpha_i + \beta) - \beta. \quad (168)$$

Now

$$\frac{\delta L}{\delta p_i} = \ln p_i - \alpha_i + \beta + 1 - \ln q_i. \quad (169)$$

Setting this to zero gives

$$p_i = q_i \exp(\alpha_i - \beta - 1). \quad (170)$$

Further, inserting this back to the Lagrangian reveals us the dual

$$g(\alpha, \beta) = - \sum q_i \exp(\alpha_i - \beta - 1) - \beta. \quad (171)$$

For this choice of p_i we know that

$$p_i = q_i \exp(\alpha_i - \beta - 1) > 0 \quad (172)$$

because $q_i > 0$ by assumption and $\exp()$ is always positive. This means that $\alpha_i = 0, \forall i$, because all the related constraints are inactive. From the condition requiring p to be a distribution, we get

$$1 = \sum p_i = \sum q_i \exp(\alpha_i - \beta - 1) \quad (173)$$

$$= \exp(-\beta - 1) \sum q_i \quad (174)$$

$$= \exp(-\beta - 1) \Rightarrow \beta = -1. \quad (175)$$

Inserting these to (170) results in $p_i = q_i, \forall i$. It can be checked that the KKT conditions are fulfilled by these choices.

□