To be completed individually or in groups of two people. **Please be sure matriculation numbers are clearly included at the top of your submission.** Submissions can be handwritten or in LaTeX formatting, but hard-to-read handwritten submissions will not be graded.

Please submit via Ilias. Submissions should be a single PDF document (note that Jupyter notebooks can and should also be downloaded as PDFs, and not submitted as .ipynb files).

Each question will be graded "pass" (full points) or "fail" (no points). We award 0.5 bonus points for the exam for each theory and practical question solved. You must complete 50% of all exercises to enter the final exam.

1. **EXAMple Question**

   Remember that a symmetric $N \times N$ matrix K is positive (semi-)definite if and only if $\boldsymbol{v}^T K \boldsymbol{v} > (\geq) 0$ for all vectors $\boldsymbol{v} \in \mathbb{R}^N / \mathbf{0}$, and that a kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a Mercer kernel, if for any finite collection $X = [x_1, \ldots, x_N] \subset \mathbb{X}$, the matrix $k_{XX} \in \mathbb{R}^{N \times N}$ with $[k_{XX}]_{ij} = k(x_i, x_j)$ is positive semidefinite.

   Prove or disprove the following statements:

   (a) for $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^D$, the function $k(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^T \boldsymbol{b}$ is a Mercer Kernel.

   (b) If k is a Mercer kernel, then $\alpha^2 k$ is also a Mercer kernel, for $\alpha \in \mathbb{R}$.

   (c) If k and h are Mercer kernels, then $k + h$ is also a Mercer kernel.

2. **Theory Question** Recall that the predictive variance of a Gaussian Process regression model at a single test data point x given a training dataset X of size n is given by:

$$\mathrm{var}_n(f(\boldsymbol{x})) = k(x, x) - k_{xX}(k_{XX} + \sigma^2 I)^{-1} k_{Xx}$$

   where k_{xX} denotes the vector of covariances between the test point and the n training points, $k_{XX} = K(X, X)$ is the covariance matrix of the training data, σ^2 is the observation noise and I is the $n \times n$ identity matrix. Let $\mathrm{var}_{n-1}(f(x))$ be the corresponding predictive variance using only the first $n - 1$ training points of X. Show that $\mathrm{var}_n(f(x)) \leq \mathrm{var}_{n-1}(f(x))$, i.e., that the predictive variance at x cannot increase as more training data is obtained.

   Hints:

   (a) You may use the following matrix identity for partitioned matrices. Let the invertible $n \times n$ matrix A and its inverse A^{-1} be partitioned into:

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix}$$

   where P and $\tilde{P}$ are $n_1 \times n_1$ matrices and S and $\tilde{S}$ are $n_2 \times n_2$ matrices with $n = n_1 + n_2$. The submatrices of A^{-1} are given as

$$\begin{aligned} \tilde{P} &= P^{-1} + P^{-1} Q M R P^{-1} \\ \tilde{Q} &= -P^{-1} Q M \\ \tilde{R} &= -M R P^{-1} \\ \tilde{S} &= M \end{aligned}$$

   where $M = (S - R P^{-1} Q)^{-1}$.

(b) The inverse of a symmetric positive definite matrix A is also symmetric and positive definite.

(c) Blockdiagonal submatrices of symmetric positive definite matrices are also symmetric positive definite.

3. **Practical Question** See `Exercise_04.ipynb`.