Ex 2 b) Explain how the solution of the resulting kernelized CCA are to be interpreted, and under which condition the solution can/cannot be expressed as directions in the input spaces $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$.

Classical CCA looks for linear mappings $w_x^T X$ and $w_y^T Y$ that achieve maximum correlation. Kernel CCA extends this approach by looking for functions $\Phi \in H_X$ and $\Psi \in H_Y$ such that the random variables $\Phi(X)$ and $\Psi(Y)$ have maximal correlation. The nonlinear mappings allow to clarify the dependency between X and Y, even though it cannot be captured by classical CCA (if they have no linear correlation). In Kernel CCA, we suppose that the original data are mapped into a feature space via nonlinear functions. Then linear CCA is applied in the feature space.

Therefore, analogously to classical CCA, where solutions $w_x, w_y$ are the weights of the linear combinations, which maximize the correlation between the variables X and Y, (i.e. in input space), then in Kernel CCA the solutions $\alpha_x$ and $\alpha_y$ are interpreted as the weights ~~which maximize the correlat~~ of the linear combinations that maximize the correlation between the (nonlinearly) transformed variables (i.e. in the feature space).

The solutions $\alpha_x$ and $\alpha_y$ can be expressed as directions in the input spaces $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ if they are transformed to

$$w_x = \Phi(X)\alpha_x \quad \text{and} \quad w_y = \Psi(Y)\alpha_y.$$

The sample KCCA crucially depends on the relation between the sample size and the dimensionalities of the space involved. Mappings into higher dimensional spaces are most likely to increase the Kernel Canonical Correlation coefficient relative to linear CCA between the input spaces. Therefore the KCCA has to be interpreted with caution (and should rather be considered as a geometrical algorithm to construct highly correlated features).