6 Proof for Optimal IDS Solution

Given data $X \in \mathbb{R}^{N \times d}$, and the data in RKHS of a Guassian kernel as $\Psi(X) \in \mathbb{R}^{N \times \infty}$. To maximize HSIC in IDS, we have the following formulation.

$$\max_{W} \quad \operatorname{Tr}\left[\Psi(X) W W^T \Psi(X)^T H K_Y H\right] \quad s.t: W^T W = I. \tag{42}$$

This objective has an alternate interpretation. Assuming that there exists a feature map Φ

$$\Phi(x) = \begin{bmatrix} \phi_1(x) & \phi_2(x) & .. & \phi_q(x) \end{bmatrix}, \quad \Phi(X) = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & .. & \phi_q(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & .. & \phi_q(x_2) \\ .. & & & \\ .. & & & \\ \phi_1(x_N) & \phi_2(x_N) & .. & \phi_q(x_N) \end{bmatrix} \tag{43}$$

where each ϕ_i is a bounded continuous function in the RKHS of an Gaussian kernel; we denote this RKHS as $\mathcal{H}$. Instead of using the Gaussian kernel Ψ, we wish to find the optimal kernel Φ that maximizes the HSIC. The new objective is reformulated as

$$\max_{\Phi \in \mathcal{H}} \quad \operatorname{Tr}\left[\Phi(X)\Phi(X)^T H K_Y H\right] \tag{44}$$

$$\max_{\Phi \in \mathcal{H}} \quad \operatorname{Tr}\left[\Phi(X)^T H K_Y H \Phi(X)\right] \tag{45}$$

$$\max_{\Phi \in \mathcal{H}} \quad \operatorname{Tr}\left[\Phi(X)^T H K_Y H \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & .. & \phi_q(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & .. & \phi_q(x_2) \\ .. & & & \\ .. & & & \\ \phi_1(x_N) & \phi_2(x_N) & .. & \phi_q(x_N) \end{bmatrix}\right]. \tag{46}$$

Since ϕ_i is a function within $\mathcal{H}$, we can apply the reproducing property where

$$\phi_i(x) = \langle \phi_i, \psi(x) \rangle. \tag{47}$$

It is important to note the difference between ϕ and ψ. While ψ is a feature map of a Gaussian kernel, $\phi_i \in \mathcal{H}$ is a function within the RKHS of a Gaussian kernel. Following the reproducing property, the formulation becomes

$$\max_{\Phi \in \mathcal{H}} \quad \operatorname{Tr}\left[\Phi(X)^T H K_Y H \begin{bmatrix} \langle \phi_1, \psi(x_1) \rangle & \langle \phi_2, \psi(x_1) \rangle & .. & \langle \phi_q, \psi(x_1) \rangle \\ \langle \phi_1, \psi(x_2) \rangle & \langle \phi_2, \psi(x_2) \rangle & .. & \langle \phi_q, \psi(x_2) \rangle \\ .. & & & \\ .. & & & \\ \langle \phi_1, \psi(x_N) \rangle & \langle \phi_2, \psi(x_N) \rangle & .. & \langle \phi_q, \psi(x_N) \rangle \end{bmatrix}\right]. \tag{48}$$

We next separate out ϕ.

$$\max_{\Phi \in \mathcal{H}} \quad \operatorname{Tr}\left[\Phi(X)^T H K_Y H \begin{bmatrix} \psi(x_1)^T \\ \psi(x_2)^T \\ .. \\ .. \\ \psi(x_N)^T \end{bmatrix} \begin{bmatrix} \phi_1 & \phi_2 & ... & \phi_q \end{bmatrix}\right]. \tag{49}$$

$$\max_{\Phi \in \mathcal{H}} \quad \operatorname{Tr}\left[\Phi(X)^T H K_Y H \Psi(X) \begin{bmatrix} \phi_1 & \phi_2 & ... & \phi_q \end{bmatrix}\right]. \tag{50}$$

$$\max_{\Phi \in \mathcal{H}} \quad \operatorname{Tr}\left[\begin{bmatrix} \phi_1^T \\ \phi_2^T \\ ... \\ \phi_q^T \end{bmatrix} \Psi(X)^T H K_Y H \Psi(X) \begin{bmatrix} \phi_1 & \phi_2 & ... & \phi_q \end{bmatrix}\right]. \tag{51}$$

Key Oberservation 1. Since ϕ_i is a function within the Gassuian RKHS, it has the property

$$\phi^T \phi = 1. \tag{52}$$

Therefore, the optimal feature map that is constrained on $\mathcal{H}$ is the most dominate eigenvector of the matrix

$$\mathcal{Q} = \Psi(X)^T H K_Y H \Psi(X) \tag{53}$$

Key Oberservation 2. We let $\overline{\Psi(X)}$ be the centered version of $\Psi(X)$ where

$$\overline{\Psi(X)} = H\Psi(X) = \begin{bmatrix} \bar{\psi}(x_1)^T \\ \bar{\psi}(x_2)^T \\ .. \\ \bar{\psi}(x_n)^T \end{bmatrix} \tag{54}$$

, then the $\mathcal{Q}$ matrix can be rewritten as

$$\mathcal{Q} = \begin{bmatrix} \bar{\psi}(x_1) & \bar{\psi}(x_2) & .. & \bar{\psi}(x_n) \end{bmatrix} YY^T \begin{bmatrix} \bar{\psi}(x_1)^T \\ \bar{\psi}(x_2)^T \\ .. \\ \bar{\psi}(x_n)^T \end{bmatrix} \tag{55}$$

. If we are facing a classification problem then Y is represented as a one-hot vector to indicate the class label. This implies that $\mathcal{Q}$ becomes

$$\mathcal{Q} = \begin{bmatrix} \sum_{i \in \mathcal{S}^1} \bar{\psi}(x_i) & .. & \sum_{i \in \mathcal{S}^c} \bar{\psi}(x_i) & \end{bmatrix} \begin{bmatrix} \sum_{i \in \mathcal{S}^1} \bar{\psi}(x_i) \\ .. \\ \sum_{i \in \mathcal{S}^c} \bar{\psi}(x_i) \end{bmatrix} \tag{56}$$

where $\mathcal{S}^i$ indicates all the sample indices within class i, and c is the number of classes. If we denote $\bar{\mathcal{U}}_i$ as the emperical kernel mean embedding of class i and α_i a constant associated with class i, then $\mathcal{Q}$ becomes

$$\mathcal{Q} = \begin{bmatrix} \alpha_1 \bar{\mathcal{U}}_1 & .. & \alpha_c \bar{\mathcal{U}}_c \end{bmatrix} \begin{bmatrix} \alpha_1 \bar{\mathcal{U}}_1^T \\ .. \\ \alpha_c \bar{\mathcal{U}}_c{}^T \end{bmatrix}. \tag{57}$$

Remember from KChain that a closed form solution of a network is W_s where

$$W_s = \begin{bmatrix} \alpha_1 \mathcal{U}_1 & .. & \alpha_c \mathcal{U}_c \end{bmatrix}. \tag{58}$$

There difference between $\bar{\mathcal{U}}_i$ and $\mathcal{U}_i$ is $H\Psi(X)$ and $\Psi(X)$. So the solution W_s is actually very close to the optimal solution. Indeed, if we had centered $\Psi(X)$ and found its most dominant eigenvectors, we would have achieved the optimal solution.

Key Oberservation 3. We first go back to a previous form of the objective where we maximize

$$\max_{\Phi \in \mathcal{H}} \quad \mathrm{Tr} \left[\begin{bmatrix} \phi_1^T \\ \phi_2^T \\ ... \\ \phi_q^T \end{bmatrix} \Psi(X)^T H K_Y H \Psi(X) \begin{bmatrix} \phi_1 & \phi_2 & ... & \phi_q \end{bmatrix} \right] \tag{59}$$

Without a loss of generality, let's only look at ϕ_1, the equation becomes

$$\max_{\Phi \in \mathcal{H}} \quad \mathrm{Tr} \left[\phi_1^T \Psi(X)^T H K_Y H \Psi(X) \phi_1 \right]. \tag{60}$$

Since ϕ_1 is a function in $\mathcal{H}$, we can approximate this function as

$$\phi \approx \sum_i^N \beta_i \psi(x_i) = \Psi(X)^T \beta. \tag{61}$$

We can now apply Eq. (61) to Eq. (60) to obtain

$$\max_{\beta} \quad \mathrm{Tr}\left[\beta^T \Psi(X)\Psi(X)^T H K_Y H \Psi(X)\Psi(X)^T \beta\right] \tag{62}$$

$$\max_{\beta} \quad \mathrm{Tr}\left[\beta^T K_X H K_Y H K_X \beta\right]. \tag{63}$$

We are able to approximate the global optimal solution via β by constraining it to $\beta^T\beta = 1$. Given β, the optimal kernel feature map ϕ^* becomes

$$\phi^* = \sum_{i=1}^{N} \beta_i \psi(x_i). \tag{64}$$

RFF is no longer necessary.