Appendix A Derivation for Equation 4

Given the objective function,

$$\begin{array}{ll} \max\limits_{U,W} & \mathrm{HSIC}(XW,U) - \lambda\,\mathrm{HSIC}(XW,Y) \\ s.t & W^TW = I, U^TU = I. \end{array}$$

Using the HSIC measure defined, the objective function can be rewritten as

$$\begin{aligned} \mathrm{HSIC}(XW,U) - \lambda\,\mathrm{HSIC}(XW,Y) &= \mathrm{Tr}(HUU^THD^{\frac{-1}{2}}K_{XW}D^{\frac{-1}{2}}) - \lambda\,\mathrm{Tr}(HYY^THD^{\frac{-1}{2}}K_{XW}D^{\frac{-1}{2}}) \\ &= \mathrm{Tr}(D^{\frac{-1}{2}}H(UU^T - \lambda YY^T)HD^{\frac{-1}{2}}K_{XW}) \\ &= \mathrm{Tr}(\gamma K_{XW}) \\ &= \textstyle\sum_{i,j}\gamma_{i,j}K_{X_{i,j}}. \end{aligned}$$

where γ is a symmetric matrix and $\gamma = H(UU^T - \lambda YY^T)H$. By substituting the Gaussian kernel for $K_{X_{i,j}}$, the objective function becomes

$$\min_{W} \quad -\sum_{i,j}\gamma_{i,j}e^{-\frac{\mathrm{Tr}[W^TA_{i,j}W]}{2\sigma^2}} \qquad s.t \quad W^TW = I.$$

Appendix B Proof for Lemma 2

Proof. Algorithm 2 sets the smallest q eigenvectors of $\Phi(W_k)$ as W_{k+1}. Since a fixed point W^* is reached when $W_k = W_{k+1}$, therefore W^* consists of the smallest eigenvectors of $\Phi(W^*)$ and Λ^* corresponds with a diagonal matrix of eigenvavlues. Since the eigenvectors of $\Phi(W^*)$ are orthonormal , ${W^*}^TW^* = I$ is also satisfied. □

Appendix C Proof for Lemma 3

Proof. Using Equation (4) as the objective function, the corresponding Lagrangian and its gradient is written as

$$\mathcal{L}(W,\Lambda) = -\sum_{i,j}\gamma_{i,j}e^{-\frac{\mathrm{Tr}(W^TA_{i,j}W)}{2\sigma^2}} - \frac{1}{2}\mathrm{Tr}(\Lambda(W^TW - I)), \tag{13}$$

and

$$\nabla_W\mathcal{L}(W,\Lambda) = \sum_{i,j}\frac{\gamma_{i,j}}{\sigma^2}e^{-\frac{\mathrm{Tr}(W^TA_{i,j}W)}{2\sigma^2}}A_{i,j}W - W\Lambda. \tag{14}$$

By setting the gradient of the Lagrangian to zero, and using the definition of $\Phi(W)$ from Equation (8), Equation (14) can be written as

$$\Phi(W)W = W\Lambda. \tag{15}$$

The gradient with respect to Λ is

$$\nabla_\Lambda\mathcal{L}(W,\Lambda) = W^TW - I. \tag{16}$$

Setting this gradient of the Lagrangian also to zero, condition (9b) is equivalent to

$$W^TW = I. \tag{17}$$

By Lemma 2, a fixed point W^* and its corresponding Λ^* satisfy (15) and (17), and the lemma follows. □

Appendix D Proof for Lemma 4

The proof for Lemma 4 relies on the following three sublemmas. The first two sublemmas demonstrate how the 2nd order conditions can be rewritten into a simpler form. With the simpler form, the third lemma demonstrates how the 2nd order conditions of a local minimum are satisfied given a large enough σ.

Lemma 4.1. *Let the directional derivative in the direction of Z be defined as*

$$\mathcal{D}f(W)[Z] := \lim_{t \to 0} \frac{f(W+tZ)-f(W)}{t}. \tag{18}$$

Then the 2nd order condition of Lemma 4 can be written as

$$\mathrm{Tr}(Z^T \mathcal{D}\nabla\mathcal{L}[Z]) = \left\{ \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} \left[\mathrm{Tr}(Z^T A_{i,j} Z) - \frac{1}{\sigma^2}\mathrm{Tr}(Z^T A_{i,j} W^*)^2 \right] \right\} - \mathrm{Tr}(Z^T Z \Lambda^*), \tag{19}$$

for all Z such that

$$Z^T W^* + W^{*T} Z = 0. \tag{20}$$

Proof. Observe first that

$$\nabla^2_{W^* W^*} \mathcal{L}(W^*, \Lambda^*) Z = \mathcal{D}\nabla\mathcal{L}[Z], \tag{21}$$

where the directional derivative of the gradient $\mathcal{D}\nabla\mathcal{L}[Z]$ is given by

$$\mathcal{D}\nabla\mathcal{L}[Z] = \lim_{t\to 0} \frac{\partial}{\partial t} \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}((W^*+tZ)^T A_{i,j}(W^*+tZ))}{2\sigma^2}} A_{i,j}(W^*+tZ) - (W^*+tZ)\Lambda.$$

This can be written as

$$\mathcal{D}\nabla\mathcal{L}[Z] = T_1 + T_2 - T_3,$$

where

$$T_1 = \lim_{t\to 0} \frac{\partial}{\partial t} \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}((W^*+tZ)^T A_{i,j}(W^*+tZ))}{2\sigma^2}} A_{i,j} W^* \tag{22}$$

$$= \lim_{t\to 0} \frac{\partial}{\partial t} \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^* + tZ^T A_{i,j} W^* + tW^{*T} A_{i,j} Z + t^2 Z^T A_{i,j} Z)}{2\sigma^2}} A_{i,j} W^* \tag{23}$$

$$= -\sum_{i,j} \frac{\gamma_{i,j}}{2\sigma^4} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} \mathrm{Tr}(Z^T A_{i,j} W^* + W^{*T} A_{i,j} Z) A_{i,j} W^* \tag{24}$$

$$= -\sum_{i,j} \frac{\gamma_{i,j}}{\sigma^4} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} \mathrm{Tr}(Z^T A_{i,j} W^*) A_{i,j} W^* \qquad \text{as } A_{i,j} = A_{i,j}^T, \tag{25}$$

$$T_2 = \lim_{t\to 0} \frac{\partial}{\partial t} \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} t e^{-\frac{\mathrm{Tr}((W^*+tZ)^T A_{i,j}(W^*+tZ))}{2\sigma^2}} A_{i,j} Z \tag{26}$$

$$= \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}(W^{*T} A_{i,j} W^*)}{2\sigma^2}} A_{i,j} Z, \tag{27}$$

$$T_3 = \lim_{t\to 0} \frac{\partial}{\partial t} (W^* + tZ)\Lambda \tag{28}$$

$$= Z\Lambda. \tag{29}$$

Hence, putting all three terms together yields

$$\mathcal{D}\nabla\mathcal{L}[Z] = \left\{ \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} \left[A_{i,j} Z - \frac{1}{\sigma^2} \mathrm{Tr}(Z^T A_{i,j} W^*) A_{i,j} W^* \right] \right\} - Z\Lambda. \tag{30}$$

Hence,

$$\mathrm{Tr}(Z^T \nabla^2_{W^* W^*} \mathcal{L}(W^*, \Lambda^*) Z) = \mathrm{Tr}(Z^T \mathcal{D}\nabla\mathcal{L}[Z]), \tag{31}$$

$$= \left\{ \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} \left[\mathrm{Tr}(Z^T A_{i,j} Z) - \frac{1}{\sigma^2} \mathrm{Tr}(Z^T A_{i,j} W^*)^2 \right] \right\} - \mathrm{Tr}(Z^T Z \Lambda_W). \tag{32}$$

Next, let Z be such that $Z \neq 0$ and $\nabla h(W^*)^T Z = 0$, where

$$h(W^*) = {W^*}^T W^* - I. \tag{33}$$

Therefore, the constraint condition can be written on Z in (9c) can be written as

$$\begin{aligned} \nabla h(W^*)^T Z = & \lim_{t \to 0} \frac{\partial}{\partial t} \frac{(W^* + tZ)^T (W^* + tZ) - {W^*}^T W^*}{t} \\ & = Z^T W^* + {W^*}^T Z = 0. \end{aligned} \tag{34}$$

Using Equations (32) and (34) lemma 4.1 follows. □

Recall from Lemma 2 that W^* consists of the q eigenvectors of $\Phi(W^*)$ with the smallest eigenvalues. We define $\bar{W^*} \in \mathbb{R}^{d \times d-q}$ as all other eigenvectors of $\Phi(W^*)$. Because Z has the same dimension as W^*, each column of Z resides in the space of $\mathbb{R}^d$. Since the eigenvectors of $\Phi(W^*)$ span $\mathbb{R}^d$, each column of Z can be represented as a linear combination of the eigenvectors of $\Phi(W^*)$. In other words, each column z_i can therefore be written as $z_i = W^* P_W^{(i)} + \bar{W}^* P_{\bar{W}^*}^{(i)}$, where $P_{W^*}^{(i)} \in \mathbb{R}^{q \times 1}$ and $P_{\bar{W}^*}^{(i)} \in \mathbb{R}^{d-q \times 1}$ represents the coordinates for the two sets of eigenvectors. Using the same notation, we also define $\Lambda^* \in \mathbb{R}^{q \times q}$ as the eigenvalues corresponding to W^* and $\bar{\Lambda}^* \in \mathbb{R}^{d-q \times d-q}$ as the eigenvalues corresponding to $\bar{W}^*$. The entire matrix Z can therefore be represented as

$$Z = \bar{W}^* P_{\bar{W}^*} + W^* P_{W^*}. \tag{35}$$

Furthermore, it can be easily shown that P_{W^*} is a skew symmetric matrix, or $-P_{W^*} = P_{W^*}^T$. By setting Z from Equation (20) into (35), the constraint can be rewritten as

$$[P_{\bar{W}^*}^T \bar{W^*}^T + {P_W^*}^T {W^*}^T] W^* + {W^*}^T [\bar{W}^* P_{\bar{W}^*} + W^* P_{W^*}] = 0. \tag{36}$$

Simplifying the equation yields the relationship

$${P_W^*}^T + P_{W^*} = 0. \tag{37}$$

Using these definitions, we define the following sublemma.

Lemma 4.2. *Given a fixed point W^* and a Z satisfying condition (20), the condition* $\mathrm{Tr}(Z^T \mathcal{D}\nabla\mathcal{L}[Z]) \geq 0$ *is equivalent to*

$$\mathrm{Tr}(P_{\bar{W}^*}^T \bar{\Lambda}^* P_{\bar{W}^*}) - \mathrm{Tr}(P_{\bar{W}^*} \Lambda^* P_{\bar{W}^*}^T) \geq C_2, \tag{38}$$

where

$$C_2 = \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^4} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} \mathrm{Tr}(Z^T A_{i,j} W^*)^2, \tag{39}$$

$P_{W^*}, P_{\bar{W}^*}$ are given by Equation (35), and $\Lambda^*, \bar{\Lambda}^*$ are the diagonal matrices containing the bottom and top eigenvalues of $\Phi(W^*)$ respectively.

Proof. By condition (19),

$$\mathrm{Tr}(Z^T \mathcal{D}\nabla\mathcal{L}[Z]) = C_1 - C_2 + C_3, \tag{40}$$

where

$$\begin{aligned}
C_1 &= \mathrm{Tr}\left(Z^T \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} A_{i,j} Z\right), \\
C_2 &= \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^4} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} \mathrm{Tr}(Z^T A_{i,j} W^*)^2, \\
C_3 &= -\mathrm{Tr}(Z^T Z \Lambda^*).
\end{aligned}$$

C_1 can be written as

$$\begin{aligned}
C_1 &= \mathrm{Tr}\left(Z^T \sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}((W^{*T} A_{i,j} W^*)}{2\sigma^2}} A_{i,j} Z\right) \\
&= \mathrm{Tr}(Z^T \Phi(W^*)[\bar{W}^* P_{\bar{W}^*} + W^* P_{W^*}]) \\
&= \mathrm{Tr}(Z^T [\Phi(W^*)\bar{W}^* P_{\bar{W}^*} + \Phi(W^*) W^* P_{W^*}]) \\
&= \mathrm{Tr}(Z^T [\bar{W}^* \bar{\Lambda} P_{\bar{W}^*} + W^* \Lambda P_{W^*}]) && \text{By definition of eigenvalues.} \\
&= \mathrm{Tr}([P_{\bar{W}^*}^T \bar{W}^{*T} + P_W^{*T} W^{*T}][\bar{W}^* \bar{\Lambda} P_{\bar{W}^*} + W^* \Lambda P_{W^*}]) && \text{Substitute for } Z \\
&= \mathrm{Tr}(P_{\bar{W}^*}^T \bar{\Lambda} P_{\bar{W}^*}) + \mathrm{Tr}(P_{W^*}^T \Lambda P_W) && \text{Given } W^{*T} W^* = I,\ \bar{W}^{*T} W^* = 0.
\end{aligned}$$

Similarly

$$\begin{aligned}
C_3 &= -\mathrm{Tr}(Z^T Z \Lambda) \\
&= -\mathrm{Tr}([P_{\bar{W}^*}^T \bar{W}^{*T} + P_{W^*}^T W^{*T}][\bar{W}^* P_{\bar{W}^*} + W^* P_{W^*}]\Lambda) \\
&= -\mathrm{Tr}([P_{\bar{W}^*}^T P_{\bar{W}^*} + P_{W^*}^T P_{W^*}]\Lambda) \\
&= -\mathrm{Tr}(P_{\bar{W}^*}^T P_{\bar{W}^*} \Lambda) - \mathrm{Tr}(P_{W^*}^T P_{W^*} \Lambda).
\end{aligned}$$

Because P_{W^*} is a square skew symmetric matrix, the diagonal elements of $P_{W^*} P_{W^*}^T$ is the same as the diagonal of $P_{W^*} P_{W^*}^T$. From this observation, we conclude that $\mathrm{Tr}(P_{W^*} P_{W^*}^T \Lambda) = \mathrm{Tr}(P_{W^*}^T P_{W^*} \Lambda)$. Hence,

$$C_3 = -\mathrm{Tr}(P_{\bar{W}^*} \Lambda P_{\bar{W}^*}^T) - \mathrm{Tr}(P_{W^*}^T \Lambda P_{W^*}).$$

Putting all 3 parts together yields

$$\begin{aligned}
\mathrm{Tr}(Z^T \mathcal{D}\nabla\mathcal{L}[Z]) &= \mathrm{Tr}(P_{\bar{W}^*}^T \bar{\Lambda} P_{\bar{W}^*}) + \mathrm{Tr}(P_{W^*}^T \Lambda P_{W^*}) - C_2 - \mathrm{Tr}(P_{\bar{W}^*} \Lambda P_{\bar{W}^*}^T) - \mathrm{Tr}(P_{W^*}^T \Lambda P_{W^*}) \\
&= \mathrm{Tr}(P_{\bar{W}^*}^T \bar{\Lambda} P_{\bar{W}^*}) - \mathrm{Tr}(P_{\bar{W}^*} \Lambda P_{\bar{W}^*}^T) - C_2.
\end{aligned} \tag{41}$$

The 2nd order condition (9c) is, therefore, satisfied, when

$$\mathrm{Tr}(P_{\bar{W}^*}^T \bar{\Lambda} P_{\bar{W}^*}) - \mathrm{Tr}(P_{\bar{W}^*} \Lambda P_{\bar{W}^*}^T) \geq C_2. \tag{42}$$

□

Lemma 4.3. *Given* W^*,$\bar{W}^*$,$\bar{\Lambda}^*$, *and* Λ^* *as defined in Equation (35), if the corresponding smallest eigenvalue of* $\bar{\Lambda}^*$ *is larger than the largest eigenvalue of* Λ^*, *then given a large enough* σ *the condition (9c) of Lemma 1 is satisfied.*

Proof. To proof sublemma (4.3), we provide bounds on each of the terms in (42). Starting with C_2 defined at (39). It has a trace term, $(\mathrm{Tr}(Z^T A_{ij} W^*))^2$ that can be rewritten as

$$(\mathrm{Tr}(A_{ij} W^* Z^T))^2 = (\mathrm{Tr}(A_{ij} W^* P_{W^*}^T {W^*}^T + A_{ij} W^* P_{\bar{W}^*}^T {\bar{W}^*}^T))^2. \tag{43}$$

Since A_{ij} is symmetric and $W^* P_{W^*}^T {W^*}^T$ is skew-symmetric, then $\mathrm{Tr}(A_{ij} W^* P_{W^*}^T {W^*}^T) = 0$. Hence

$$(\mathrm{Tr}(Z^T A_{ij} W^*))^2 = (\mathrm{Tr}(A_{ij} W^* Z^T))^2 = (\mathrm{Tr}(A_{ij} W^* P_{\bar{W}^*}^T {\bar{W}^*}^T))^2 \tag{44}$$

$$\leq \mathrm{Tr}(A_{i,j}^T A_{ij}) \, \mathrm{Tr}(P_{\bar{W}^*}^T P_{\bar{W}^*}) \tag{45}$$

where the last inequality follows from Cauchy-Schwartz inequality and that fact that ${W^*}^T W^* = I$ and ${\bar{W}^*}^T \bar{W}^* = I$. Thus, C_2 in (41) is bounded by

$$C_2 \leq \sum_{i,j} \frac{|\gamma_{i,j}|}{\sigma^4} e^{-\frac{\mathrm{Tr}(({W^*}^T A_{i,j} W^*)}{2\sigma^2}} \mathrm{Tr}(A_{i,j}^T A_{ij}) \, \mathrm{Tr}(P_{\bar{W}^*}^T P_{\bar{W}^*}) \tag{46}$$

Similarly, the remaining terms in (40) can be bounded by

$$C_1 = \mathrm{Tr}(P_{\bar{W}^*}^T \bar{\Lambda}^* P_{\bar{W}^*}) \geq \min_i (\bar{\Lambda}^*{}_i) \, \mathrm{Tr}(P_{\bar{W}^*} P_{\bar{W}^*}^T) \tag{47}$$

$$C_3 = -\mathrm{Tr}(P_{\bar{W}^*} \Lambda^* P_{\bar{W}^*}^T) \geq -\max_i (\Lambda_i^*) \, \mathrm{Tr}(P_{\bar{W}^*}^T P_{\bar{W}^*}). \tag{48}$$

Using the bounds for each term, the Equation (42) can be rewritten as

$$[\min_i (\bar{\Lambda}^*{}_i) - \max_j (\Lambda_j^*)] \, \mathrm{Tr}(P_{\bar{W}^*}^T P_{\bar{W}^*}) \geq \sum_{i,j} \frac{|\gamma_{i,j}|}{\sigma^4} e^{-\frac{\mathrm{Tr}(({W^*}^T A_{i,j} W^*)}{2\sigma^2}} \mathrm{Tr}(A_{i,j}^T A_{ij}) \, \mathrm{Tr}(P_{\bar{W}^*}^T P_{\bar{W}^*}) \tag{49}$$

$$[\min_i (\bar{\Lambda}^*{}_i) - \max_j (\Lambda_j^*)] \geq \sum_{i,j} \frac{|\gamma_{i,j}|}{\sigma^4} e^{-\frac{\mathrm{Tr}(({W^*}^T A_{i,j} W^*)}{2\sigma^2}} \mathrm{Tr}(A_{i,j}^T A_{ij}) \tag{50}$$

It should be noted that Λ^* is a function of $\frac{1}{\sigma^2}$. This relationship could be removed by multiplying both sides of the inequality by σ^* to yield

$$\sigma^2 [\min_i (\bar{\Lambda}^*{}_i) - \max_j (\Lambda_j^*)] \geq \sum_{i,j} \frac{|\gamma_{i,j}|}{\sigma^2} e^{-\frac{\mathrm{Tr}(({W^*}^T A_{i,j} W^*)}{2\sigma^2}} \mathrm{Tr}(A_{i,j}^T A_{ij}). \tag{51}$$

Since σ^2 is always a positive value, as long as all the eigenvalues from $\bar{\Lambda}^*$ is larger than all the eigenvalues from Λ^*, the left hand side of the equation will always be greater than 0. As $\sigma \to \infty$, the right hand side approaches 0, and the condition (9c) of Lemma 1 is satisfied. □

As a side note, the eigen gap between $\min(\bar{\Lambda}^*)$ and $\max(\Lambda^*)$ controls the range of potential σ values i.e. the larger the eigen gap the easier for σ to satisfy (51). Therefore, the ideal cutoff point should have a large eigen gap.

Appendix E Convergence Plot from Experiments

Figure 4 summarizes the convergence activity of various experiments. For each experiment, the top figure provides the magnitude of the objective function. It can be seen that the values converges towards a fixed point. The middle plot provide updates of the gradient of the Lagrangian. It can be seen that the gradient converges towards 0. The bottom plot shows the changes in W during each iteration. The change in W converge towards 0.

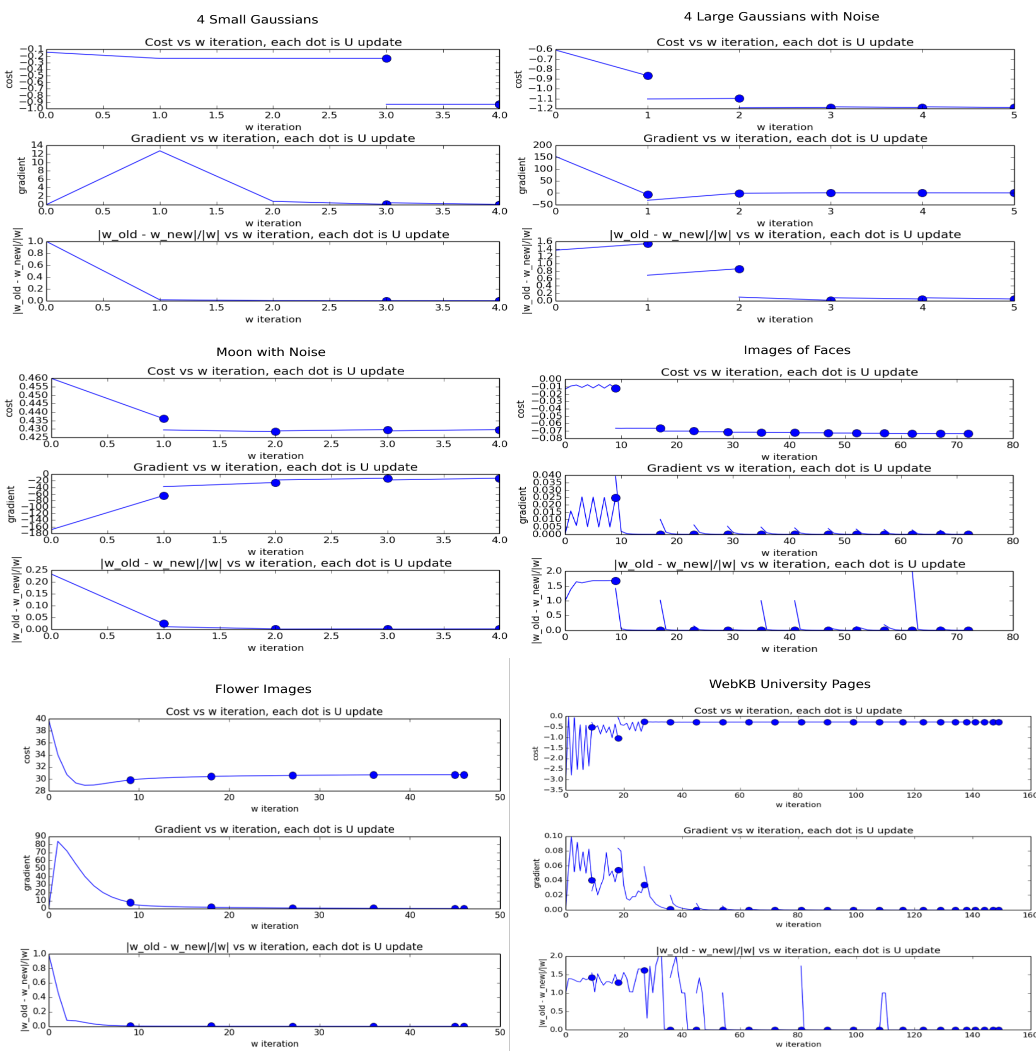

Figure 4: Convergence Results from the Experiments.

Appendix F Proof of Convergence

The convergence property of ISM has been analyzed and yields the following theorem.

Theorem 2. *A sequence $\{W_k\}_{k\in\mathbb{N}}$ generated by Algorithm 2 contains a convergent subsequence.*

Proof. According to Bolzano-Weierstrass theorem, if we can show that the sequences generated from the 1st order relaxation is bounded, it has a convergent subsequence. If we study the Equation $\Phi(W)$ more closely, the key driver of the sequence of W_k is the matrix Φ, therefore, if we can show that if this matrix is bounded, the sequence itself is also bounded. We look inside the construction of the matrix itself.

$$\Phi_{n+1} = \left[\sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}(W_n^T A_{i,j} W_n)}{2\sigma^2}} A_{i,j}\right]$$

From this equation, start with the matrix $A_{i,j} = (x_i - x_j)(x_i - x_j)^T$. Since x_i, x_j are data points that are always centered and scaled to a variance of 1, the size of this matrix is always constrained. It also implies that $A_{i,j}$ is a PSD matrix. From this, the exponential term is always limited between the value of 0 and 1. The value of σ is a constant given from the initialization stage. Lastly, we have the $\gamma_{i,j}$ term. Since $\gamma = D^{-1/2}H(UU^T - \lambda YY^T)HD^{-1/2}$. The degree matrix came from the exponential kernel. Since the kernels are bounded, D is also bounded. The centering matrix H and the previous clustering result Y can be considered as bounded constants. Since the spectral embedding U is a orthonormal matrix, it is always bounded. From this, given that the components of Φ is bounded, the infinity norm of the Φ is always bounded. The eigenvalue matrix of Λ is therefore also bounded. Using the Bolzano-Weierstrass Theorem, the sequence contains a convergent sub-sequence. Given that Φ is a continuous function of W, by continuity, W also has a convergent sub-sequence. □

Appendix G Proof for the initialization

Although the proof was originally shown through the usage of the 2nd order Taylor Approximation. A simpler approach was later discovered to arrive to the same formulation faster. We first note that Taylor's Expansion around 0 of an exponential is

$$e^x = 1 + x + \frac{x^2}{2!} + ...$$

Given the objective Lagrangian in eq (6), we simplify the Lagrangian by using the Taylor approximation only on the problematic exponential term. The approximation is expanded up to the 1st order centering around 0 to yield

$$\mathcal{L} \approx -\sum_{i,j} \gamma_{i,j} \left(1 - \frac{\mathrm{Tr}(W^T A_{i,j} W)}{2\sigma^2}\right) + \frac{1}{2}\mathrm{Tr}(\Lambda(I - W^T W)).$$

By taking the derivative of the approximated Lagrangian and setting the derivative to zero, an eigenvalue/eigenvector relationship emerges as

$$\Phi W = \left[\sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} A_{i,j}\right] W_0 = W_0 \Lambda.$$

From this, we see that Φ_0 is no longer a function of W. Using this Φ_0 we can then calculate a closed form solution for W_0

Appendix H Proof for the computational complexity

For ISM, DG and SM, the bottleneck resides in the computation of the gradient.

$$f(W) = \sum_{i,j} \gamma_{i,j} e^{-\frac{\mathrm{Tr}(W^T A_{i,j} W)}{2\sigma^2}}$$

$$\frac{\partial f}{\partial W} = \left[\sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}(W^T A_{i,j} W)}{2\sigma^2}} A_{i,j} \right] W$$

$$\frac{\partial f}{\partial W} = \left[\sum_{i,j} \frac{\gamma_{i,j}}{\sigma^2} e^{-\frac{\mathrm{Tr}(W^T \Delta x_{i,j} \Delta x_{i,j}^T W)}{2\sigma^2}} A_{i,j} \right] W$$

Where $A_{i,j} = \Delta x_{i,j} \Delta x_{i,j}^T$. The variables have the following dimensions.

$$x_{i,j} \in \mathbb{R}^{d \times 1}$$
$$W \in \mathbb{R}^{d \times q}$$

To compute a new W with DG, we first mulitply $\Delta x_{i,j}^T W$, which is $O(d)$. Note that W in DG is always 1 single column. Next, it multiplies with its own transpose to yied $O(d+q^2)$. Then we compute $A_{i,j}$ to get $O(d+q^2+d^2)$. Since this operation needs to be added n^2 times, we get, $O(n^2(d+q^2+d^2))$. Since $d \gg q$, this notation reduces down to $O(n^2 d^2)$. Let T_1 be the number of iterations until convergence, then it becomes $O(T_1 n^2 d^2)$. Lastly, in DG, this operation needs to be repeated q times, hence, $O(T_1 n^2 d^2 q)$.

To compute a new W with SM, we first mulitply $\Delta x_{i,j}^T W$, which is $O(dq)$. Next, it multiplies with its own transpose to yied $O(dq+q^2)$. Then we compute $A_{i,j}$ to get $O(dq+q^2+d^2)$. Since this operation needs to be added n^2 times, we get, $O(n^2(dq+q^2+d^2))$. Since $d \gg q$, this notation reduces down to $O(n^2 d^2)$. The SM method requires the computation of the inverse of $d \times d$ matrix. Since inverses is cubic, it becomes $O(n^2 d^2 + d^3)$. Lastly, let T_2 be the number of iterations until convergence, then it becomes $O(T_2(n^2 d^2 + d^3))$.

To compute a new W with ISM, we first mulitply $\Delta x_{i,j}^T W$, which is $O(dq)$. Next, it multiplies with its own transpose to yied $O(dq+q^2)$. Then we compute $A_{i,j}$ to get $O(dq+q^2+d^2)$. Since this operation needs to be added n^2 times, we get, $O(n^2(dq+q^2+d^2))$. Since $d \gg q$, this notation reduces down to $O(n^2 d^2)$. The ISM method requires the computation of the eigen decomposition of $d \times d$ matrix. Since inverses is cubic, it becomes $O(n^2 d^2 + d^3)$. Lastly, let T_3 be the number of iterations until convergence, then it becomes $O(T_3(n^2 d^2 + d^3))$.

Appendix I Measure of Non-linear Relationship by HSIC Versus Correlation

The figure below demonstrates a visual comparison of HSIC and correlation. It can be seen that HSIC measures non-linear relationships, while correlation does not.

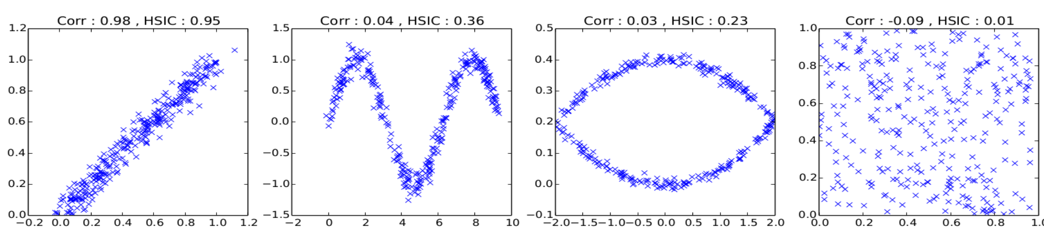

Figure 5: Showing that HSIC captures non-linear information.

Appendix J Hyperparameters Used in Each Experiment

	σ	λ	q
Gauss A	1	0.04	1
Gauss B 200	5	2	3
Moon 400	0.1	1	3
Moon+N 200	0.2	0.1	6
Flower	2	10	2
Face	3.1	1	17
Web KB	18.7	0.057	4