APPENDIX A
PROOF OF LEMMA 1

To show the first property, we define the $|\mathcal{X}_i| \times (|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdots |\mathcal{X}_d|)$ matrix $\mathbf{B}_i$, for $i = 1, \ldots, d$, as

$$B_i(x_i'; x^d) = \begin{cases} \frac{\sqrt{P_{X^d}(x^d)}}{\sqrt{P_{X_i}(x_i')}} & \text{if } x_i' = x_i \\ 0 & \text{otherwise.} \end{cases}$$

Then, one can verify that

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_d \end{bmatrix} \begin{bmatrix} \mathbf{B}_1^{\mathrm{T}} & \cdots & \mathbf{B}_d^{\mathrm{T}} \end{bmatrix},$$

which implies that $\mathbf{B}$ is positive semidefinite.

To establish the second property, note that $\boldsymbol{\psi}^{(0)}$ is an eigenvector of $\mathbf{B}$ with eigenvalue d, and thus $\left(\boldsymbol{\psi}^{(0)}\right)^{\mathrm{T}} \mathbf{B}\boldsymbol{\psi}^{(0)} = d$. Moreover, it is shown in [23] that the largest singular value of $\mathbf{B}_{ij}$ is 1, i.e., $\|\mathbf{B}_{ij}\|_{\mathrm{s}} = 1$, where $\|\cdot\|_{\mathrm{s}}$ denotes the spectral norm of its matrix argument.

Therefore, for $\boldsymbol{\psi} = \left[\boldsymbol{\psi}_1^{\mathrm{T}}, \ldots, \boldsymbol{\psi}_d^{\mathrm{T}}\right]^{\mathrm{T}}$ with each $\boldsymbol{\psi}_i$ being an $|\mathcal{X}_i|$-dimensional vector, we have

$$\begin{aligned} \boldsymbol{\psi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\psi} = \sum_{i=1}^{d}\sum_{j=1}^{d} \boldsymbol{\psi}_i^{\mathrm{T}}\mathbf{B}_{ij}\boldsymbol{\psi}_j &\leq \sum_{i=1}^{d}\sum_{j=1}^{d} \|\boldsymbol{\psi}_i\| \cdot \|\mathbf{B}_{ij}\|_{\mathrm{s}} \cdot \|\boldsymbol{\psi}_j\| \\ &= \left(\sum_{i=1}^{d} \|\boldsymbol{\psi}_i\|\right)^2 \\ &\leq d \sum_{i=1}^{d} \|\boldsymbol{\psi}_i\|^2 = d\|\boldsymbol{\psi}\|^2, \end{aligned}$$

where the second inequality follows from the fact that the arithmetic mean is no greater than the quadratic mean. Hence, we have

$$\max_{\boldsymbol{\psi}\colon \|\boldsymbol{\psi}\|=1} \boldsymbol{\psi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\psi} = d,$$

i.e., the largest eigenvalue of $\mathbf{B}$ is d.

To verify the third property, we construct $\boldsymbol{\psi}'$ as

$$\boldsymbol{\psi}' = \begin{bmatrix} \boldsymbol{\psi}_1' \\ \mathbf{0} \end{bmatrix},$$

where $\boldsymbol{\psi}_1' \in \mathbb{R}^{|\mathcal{X}_1|}$ is chosen such that $\langle \boldsymbol{\psi}_1', \boldsymbol{v}_1 \rangle = 0$ and $\|\boldsymbol{\psi}_1'\|^2 = 1$, and where $\mathbf{0}$ is the $(m - |\mathcal{X}_1|)$-dimensional zero vector. Therefore, we have $\left\langle \boldsymbol{\psi}', \boldsymbol{\psi}^{(0)} \right\rangle = 0$ and $\|\boldsymbol{\psi}'\|^2 = 1$. Note that the second eigenvalue $\lambda^{(1)}$ of $\mathbf{B}$ can be written as

$$\lambda^{(1)} = \max_{\boldsymbol{\psi}\colon \|\boldsymbol{\psi}\|=1, \left\langle \boldsymbol{\psi}, \boldsymbol{\psi}^{(0)} \right\rangle = 0} \boldsymbol{\psi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\psi},$$

which implies that $\lambda^{(1)} \geq (\boldsymbol{\psi}')^{\mathrm{T}}\mathbf{B}\boldsymbol{\psi}' = \|\boldsymbol{\psi}_1'\|^2 = 1$.

To verify the fourth property, we define the $(d-1)$-dimensional subspace $\mathcal{S}_{\mathrm{eig}}$ as

$$\mathcal{S}_{\mathrm{eig}} \triangleq \left\{ \boldsymbol{\psi} = \left[\alpha_1\boldsymbol{v}_1^{\mathrm{T}}, \ldots, \alpha_d\boldsymbol{v}_d^{\mathrm{T}}\right]^{\mathrm{T}} : \sum_{i=1}^{d} \alpha_i = 0 \right\}. \tag{A-1}$$

Then, for all $\boldsymbol{\psi} \in \mathcal{S}_{\mathrm{eig}}$, from $\mathbf{B}_{ij}\boldsymbol{v}_j = \boldsymbol{v}_i$, it is straightforward to verify that $\mathbf{B}\boldsymbol{\psi} = \mathbf{0}_m$, where $\mathbf{0}_m$ is the zero vector in $\mathbb{R}^m$. Therefore, $\mathcal{S}_{\mathrm{eig}}$ is an eigenspace of $\mathbf{B}$ associated with $d-1$ zero eigenvalues. Since $\mathbf{B}$ is positive semidefinite, without loss of generality we can assume that $\mathcal{S}_{\mathrm{eig}}$ is spanned by $\boldsymbol{\psi}^{(m-d+1)}, \ldots, \boldsymbol{\psi}^{(m-1)}$, which correspond to eigenvalues $\lambda^{(m-d+1)} = \cdots = \lambda^{(m-1)} = 0$.

Finally, to establish the last property, for each $\ell = 1, \ldots, m-d$, from $\left\langle \boldsymbol{\psi}^{(\ell)}, \boldsymbol{\psi}^{(0)} \right\rangle = 0$ we have

$$\sum_{i=1}^{d} \left\langle \boldsymbol{\psi}_i^{(\ell)}, \boldsymbol{v}_i \right\rangle = 0.$$

Therefore, from the third property, we have

$$\boldsymbol{\psi}' = \begin{bmatrix} \left\langle \boldsymbol{\psi}_1^{(\ell)}, \boldsymbol{v}_1 \right\rangle \boldsymbol{v}_1 \\ \vdots \\ \left\langle \boldsymbol{\psi}_d^{(\ell)}, \boldsymbol{v}_d \right\rangle \boldsymbol{v}_d \end{bmatrix} \in \mathcal{S}_{\mathrm{eig}}.$$

Hence, we obtain $\langle \boldsymbol{\psi}', \boldsymbol{\psi}^{(\ell)} \rangle = 0$, i.e.,

$$\sum_{i=1}^{d} \left\langle \boldsymbol{\psi}_i^{(\ell)}, \boldsymbol{v}_i \right\rangle^2 = 0,$$

which implies that $\left\langle \boldsymbol{\psi}_i^{(\ell)}, \boldsymbol{v}_i \right\rangle = 0$ for $i = 1, \ldots, d$.

APPENDIX B
PROOF OF THEOREM 1

First, we replace δ by $\frac{1}{2}\epsilon^2$ for the convenience of presentation when applying the local geometric approach. With this notation, the constraint (2) becomes

$$I(U; X^d) \leq \frac{1}{2}\epsilon^2, \tag{A-2}$$

with ϵ assumed to be small. Then, it follows from (3) and (A-2) that for all u, the conditional distribution $P_{X^d|U=u}$ can be written as a perturbation to the marginal distribution:

$$P_{X^d|U}(x^d|u) = P_{X^d}(x^d) + \epsilon\sqrt{P_{X^d}(x^d)}\phi_u(x^d) \tag{A-3}$$

where ϕ_u can be viewed as an $(|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdots |\mathcal{X}_d|)$-dimensional vector. Moreover, it follows from the second order Taylor's expansion for the K-L divergence that

$$\begin{aligned} I(U; X^d) &= \mathbb{E}_U\left[D(P_{X^d|U}\|P_{X^d})\right] \\ &= \frac{1}{2}\epsilon^2\mathbb{E}_U\left[\|\boldsymbol{\phi}_U\|^2\right] + o(\epsilon^2), \end{aligned}$$

where $\|\cdot\|$ denotes the l_2-norm. Thus, by ignoring the higher order term of ϵ as we assume ϵ to be small, the constraint $I(U; X^d) \leq \frac{1}{2}\epsilon^2$ can be reduced to

$$\mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2] \leq 1. \tag{A-4}$$

In addition, the objective function $\ell(X^d|U)$ can also be expressed in terms of mutual information:

$$\begin{aligned} &D(P_{X^d}\|P_{X_1}\cdots P_{X_d}) - D(P_{X^d}\|P_{X_1}\cdots P_{X_d}|U) \\ &= \sum_{i=1}^{d} I(U; X_i) - I(U; X^d) \end{aligned} \tag{A-5}$$

and for each i, the mutual information $I(U;X_i)$ can be again approximated as the l_2-norm square

$$I(U;X_i) = \frac{1}{2}\epsilon^2 \mathbb{E}_U[\|\boldsymbol{\psi}_{i,U}\|^2] + o(\epsilon^2),$$

where for $U = u$, the vector $\boldsymbol{\psi}_{i,u}$ is the $|\mathcal{X}_i|$-dimensional perturbation vector defined as

$$\psi_{i,u}(x_i) = \frac{P_{X_i|U}(x_i|u) - P_{X_i}(x_i)}{\epsilon\sqrt{P_{X_i}(x_i)}} \tag{A-6}$$

Then, by ignoring the higher order terms of ϵ, the optimization problem we want to solve can be transferred to a linear algebraic problem

$$\max_{\mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2]\le 1} \sum_{i=1}^{d} \mathbb{E}_U[\|\boldsymbol{\psi}_{i,U}\|^2] - \mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2]. \tag{A-7}$$

To solve (A-7), observe that P_{X_i} and $P_{X_i|U}$ are marginal distributions of P_{X^d} and $P_{X^d|U}$, thus there is a correlation between $\boldsymbol{\phi}_U$ and $\boldsymbol{\psi}_{i,U}$:

$$\psi_{i,u}(x_i) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d} \frac{\sqrt{P_{X^d}(x^d)}}{\sqrt{P_{X_i}(x_i)}} \phi_u(x^d)$$

which can be represented in matrix form as $\boldsymbol{\psi}_{i,u} = \mathbf{B}_i \cdot \boldsymbol{\phi}_u$, where $\mathbf{B}_i$ is an $|\mathcal{X}_i| \times (|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdots |\mathcal{X}_d|)$ matrix with entries

$$B_i(x_i'; (x_1,\ldots,x_d)) = \begin{cases} \frac{\sqrt{P_{X^d}(x^d)}}{\sqrt{P_{X_i}(x_i')}} & \text{if } x_i' = x_i \\ 0 & \text{otherwise.} \end{cases} \tag{A-8}$$

Therefore, if we define an $(|\mathcal{X}_1| + \cdots + |\mathcal{X}_m|) \times (|\mathcal{X}_1| \cdots |\mathcal{X}_m|)$-dimensional matrix

$$\mathbf{B}_0 \triangleq \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_d \end{bmatrix}, \tag{A-9}$$

then since

$$\sum_{i=1}^{d} \mathbb{E}_U[\|\boldsymbol{\psi}_{i,U}\|^2] = \sum_{i=1}^{d} \mathbb{E}_U[\|\mathbf{B}_i \cdot \boldsymbol{\phi}_U\|^2] = \mathbb{E}_U[\|\mathbf{B}_0 \cdot \boldsymbol{\phi}_U\|^2],$$

we can rewrite (A-7) as

$$\max_{\mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2]\le 1} \mathbb{E}_U[\|\mathbf{B}_0 \cdot \boldsymbol{\phi}_U\|^2] - \mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2]. \tag{A-10}$$

Moreover, since $\boldsymbol{\phi}_U$ is a perturbation vector of probability distributions, by summing over all x^d for both sides of (A-3), it has to satisfy an extra constraint

$$\sum_{x^d} \sqrt{P_{X^d}(x^d)} \phi_u(x^d) = 0, \tag{A-11}$$

which implies that $\boldsymbol{\phi}_U$ is orthogonal to an $(|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdots |\mathcal{X}_d|)$-dimensional vector $\boldsymbol{\phi}^{(0)}$, whose entries are $\sqrt{P_{X^d}(x^d)}$. In particular, it is shown in [23] that $\boldsymbol{\phi}^{(0)}$ is the right singular vector of $\mathbf{B}_0$ with the largest singular value $\sigma_0 = \sqrt{d}$, and the corresponding left singular vector is $\boldsymbol{\psi}^{(0)}$. In addition, it can be verified that $\mathbf{B}_0$ satisfies $\mathbf{B}_0\mathbf{B}_0^{\mathrm{T}} = \mathbf{B}$ with $\mathbf{B}$ as defined in (4). Therefore, the second largest singular value of $\mathbf{B}_0$ is $\sigma_1 = \sqrt{\lambda^{(1)}} \ge 1$, and the optimal solution of (A-10) is to align the vectors $\boldsymbol{\phi}_{U=u}$, for all u, along the second largest right singular vector of $\mathbf{B}_0$.

It turns out that it is easier to compute the second largest left singular vector of $\mathbf{B}_0$ instead of the right one. This is equivalent to computing the second largest eigenvector of the matrix $\mathbf{B}_0\mathbf{B}_0^{\mathrm{T}} = \mathbf{B}$.

Now, the second largest right singular vector $\boldsymbol{\phi}^{(1)}$ of $\mathbf{B}_0$ can be computed as

$$\begin{aligned} \phi^{(1)}(x^d) &= \frac{1}{\sqrt{\lambda^{(1)}}} (\mathbf{B}_0^{\mathrm{T}} \boldsymbol{\psi}^{(1)})(x^d) \\ &= \frac{1}{\sqrt{\lambda^{(1)}}} \cdot \left(\sqrt{P_{X^d}(x^d)} \sum_{i=1}^{d} \frac{\psi_i^{(1)}(x_i)}{\sqrt{P_{X_i}(x_i)}} \right) \\ &= \sqrt{P_{X^d}(x^d)} \cdot \left(\frac{1}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right), \end{aligned} \tag{A-12}$$

where $\mathbf{B}_0^{\mathrm{T}}\boldsymbol{\psi}^{(1)}$ is a vector and $(\mathbf{B}_0^{\mathrm{T}}\boldsymbol{\psi}^{(1)})(x^d)$ is the x^d-th entry of this vector. Since all the $\boldsymbol{\phi}_{U=u}$ should be aligned to $\boldsymbol{\phi}^{(1)}$, there exists a function $h\colon \mathcal{U} \mapsto \mathbb{R}$, such that

$$P_{X^d|U}(x^d|u) = P_{X^d}(x^d) \left(1 + \frac{\epsilon h(u)}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right) + o(\epsilon),$$

where the term $o(\epsilon)$ comes from the local approximation we made for (A-4). Therefore, the optimal joint distributions for our optimization problem can be written as

$$\begin{aligned} &P_{UX^d}(u,x^d) \\ &= P_U(u)P_{X^d}(x^d) \left(1 + \frac{\epsilon h(u)}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right) + o(\epsilon). \end{aligned} \tag{A-13}$$

Note that if we sum both sides of (A-13) over all $u \in \mathcal{U}$, then we have $\sum_{u\in\mathcal{U}} P_U(u)h(u) = 0$, which implies that $h(U)$ is a zero-mean function. Moreover, it is easy to compute from (A-4) that the variance $\mathbb{E}[h^2(U)] = 1$. Finally, note that the exponential family $\mathcal{P}_{\exp}^{(\delta)}$, when δ is small, can be written as

$$\begin{aligned} \mathcal{P}_{\exp}^{(\delta)} = \Bigg\{ & P_U(u)P_{X^d}(x^d) \\ & \cdot \left(1 + \frac{\sqrt{2\delta} h(u)}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right) + o\left(\sqrt{\delta}\right) : h \in \mathcal{H}_\delta \Bigg\}. \end{aligned} \tag{A-14}$$

Since $\delta = \frac{1}{2}\epsilon^2$ the proof is completed by comparing (A-14) and (A-13).

Appendix C
Proof of Theorem 2

First, we introduce a useful lemma (see, e.g., [24, Corollary 4.3.39, p. 248]).

Lemma A-1. *Given an arbitrary $k_1 \times k_2$ matrix $\mathbf{A}$ and any $k \in \{1,\ldots,\min\{k_1,k_2\}\}$, we have*

$$\max_{\mathbf{M}\in\mathbb{R}^{k_2\times k}} \|\mathbf{A}\mathbf{M}\|_{\mathrm{F}}^2 = \sum_{i=1}^{k} \sigma_i^2, \tag{A-15}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, and where $\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}}$ denotes the singular values of $\mathbf{A}$. Moreover, the maximum in (A-15) *can be achieved by $\mathbf{M} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_k \end{bmatrix} \mathbf{Q}$, where $\boldsymbol{v}_i$ denotes the right singular vector of $\mathbf{A}$ corresponding to σ_i, for $i = 1, \ldots, \min\{m,n\}$, and $\mathbf{Q} \in \mathbb{R}^{k \times k}$ is an (arbitrary) orthogonal matrix.*

To begin the proof, similar to Theorem 1, we replace δ by $\frac{1}{2}\epsilon^2$ and write the conditional distribution $P_{X^d|U^k=u^k}$ as a perturbation to the joint distribution P_{X^d}:

$$P_{X^d|U^k}(x^d|u^k) = P_{X^d}(x^d) + \epsilon\sqrt{P_{X^d}(x^d)}\boldsymbol{\phi}_{u^k}(x^d), \quad \text{(A-16)}$$

where $\boldsymbol{\phi}_{u^k}$ is an $(|\mathcal{X}_1|\cdot|\mathcal{X}_2|\cdots|\mathcal{X}_d|)$-dimensional vector. Again, it follows from the second-order Taylor series expansion of the K-L divergence that

$$\begin{aligned} I(U^k; X^d) &= \mathbb{E}_{U^k}[D(P_{X^d|U^k}\|P_{X^d})] \\ &= \frac{1}{2}\epsilon^2 \mathbb{E}_{U^k}\left[\|\boldsymbol{\phi}_{U^k}\|^2\right] + o(\epsilon^2). \end{aligned}$$

Similarly, for all $i = 1, \ldots, k$, the conditional distribution $P_{X^d|U_i=u_i}$ can be written as

$$P_{X^d|U_i}(x^d|u_i) = P_{X^d}(x^d) + \epsilon\sqrt{P_{X^d}(x^d)}\boldsymbol{\phi}_{u_i}(x^d), \quad \text{(A-17)}$$

and we have

$$I(U_i; X^d) = \frac{1}{2}\epsilon^2 \mathbb{E}_{U_i}\left[\|\boldsymbol{\phi}_{U_i}\|^2\right] + o(\epsilon^2).$$

Therefore, by ignoring the higher order terms of ϵ, the first constraint can be reduced to

$$1 \geq \mathbb{E}_{U_1}\left[\|\boldsymbol{\phi}_{U_i}\|^2\right] \geq \cdots \geq \mathbb{E}_{U_k}\left[\|\boldsymbol{\phi}_{U_i}\|^2\right].$$

Moreover, due to the independence and conditional independence among the U^k, $\boldsymbol{\phi}_{u^k}$ and $\boldsymbol{\phi}_{u_i}$ satisfy

$$\boldsymbol{\phi}_{u^k} = \sum_{i=1}^{k} \boldsymbol{\phi}_{u_i} + o(1) \quad \text{(A-18)}$$

and

$$\langle \boldsymbol{\phi}_{u_i}, \boldsymbol{\phi}_{u_j} \rangle = 0, \quad \text{for all } i \neq j, u_i \in \mathcal{U}_i, u_j \in \mathcal{U}_j. \quad \text{(A-19)}$$

Indeed, we have

$$\begin{aligned} P_{X_d|U^k}(x^d|u^k) &= \frac{P_{X^d}(x^d)P_{U^k|X^d}(u^k|x^d)}{P_{U^k}(u^k)} \\ &= P_{X^d}(x^d)\prod_{i=1}^{k}\frac{P_{U_i|X^d}(u_i|x^d)}{P_{U_i}(u_i)}, \end{aligned}$$

which implies

$$\frac{P_{X_d|U^k}(x^d|u^k)}{P_{X^d}(x^d)} = \prod_{i=1}^{k}\frac{P_{U_i|X^d}(u_i|x^d)}{P_{U_i}(u_i)} = \prod_{i=1}^{k}\frac{P_{X^d|U_i}(x^d|u_i)}{P_{X^d}(x^d)}. \quad \text{(A-20)}$$

Substituting (A-16) and (A-17) into (A-20) then yields

$$1 + \epsilon\frac{\boldsymbol{\phi}_{u^k}(x_1^d)}{\sqrt{P_{X^d}(x^d)}} = \prod_{i=1}^{k}\left[1 + \epsilon\frac{\boldsymbol{\phi}_{u_i}(x_1^d)}{\sqrt{P_{X^d}(x^d)}}\right],$$

and via comparing the ϵ-order terms for both sides we obtain (A-18).

To obtain (A-19), note that from (A-17), for all $i \neq j$, $u_i \in \mathcal{U}_i$ and $u_j \in \mathcal{U}_j$, we have

$$\begin{aligned} &\epsilon^2 \langle \boldsymbol{\phi}_{u_i}, \boldsymbol{\phi}_{u_j} \rangle \\ &= \epsilon^2 \sum_{x^d} \phi_{u_i}(x^d)\phi_{u_j}(x^d) && \text{(A-21)} \\ &= \sum_{x_1^d}\bigg(\frac{1}{P_{X^d}(x^d)}\cdot[P_{X^d|U_i}(x^d|u_i) - P_{X^d}(x^d)] \\ &\qquad \cdot[P_{X^d|U_j}(x^d|u_i) - P_{X^d}(x^d)]\bigg) && \text{(A-22)} \\ &= \sum_{x_1^d}\frac{P_{X^d|U_i}(x^d|u_i)P_{X^d|U_j}(x^d|u_j)}{P_{X^d}(x^d)} - 1 && \text{(A-23)} \\ &= \sum_{x_1^d} P_{X^d}(x^d)\cdot\frac{P_{U_i|X^d}(u_i|x^d)}{P_{U_i}(u_i)}\cdot\frac{P_{U_j|X^d}(u_j|x^d)}{P_{U_j}(u_j)} - 1 \\ &= \frac{1}{P_{U_iU_j}(u_i,u_j)}\sum_{x_1^d} P_{X^d}(x^d)P_{U_iU_j|X^d}(u_i,u_j|x^d) - 1 && \text{(A-24)} \\ &= 0, \end{aligned}$$

where to obtain (A-24) we have again exploited the independence and conditional independence of U_i and U_j.

In addition, the objective function $\mathcal{L}(X^d|U^k)$ can be expressed as [cf. (A-5)]

$$\begin{aligned} &D(P_{X^d}\|P_{X_1}\ldots P_{X_d}) - D(P_{X^d}\|P_{X_1}\ldots P_{X_d}|U^k) \\ &= \sum_{i=1}^{d} I(U^k; X_i) - I(U^k; X^d) \\ &= \sum_{i=1}^{d} I(U^k; X_i) - \sum_{j=1}^{k} I(U_j; X^d), && \text{(A-25)} \end{aligned}$$

where to obtain the last equality we have used the fact that

$$\begin{aligned} I(U^k; X^d) &= \mathbb{E}_{U^kX^d}\left[\log\frac{P_{U^k|X^d}(U^k|X^d)}{P_{U^k}(U^k)}\right] && \text{(A-26)} \\ &= \mathbb{E}_{U^kX^d}\left[\sum_{j=1}^{k}\log\frac{P_{U_j|X^d}(U_j|X^d)}{P_{U_j}(U_j)}\right] && \text{(A-27)} \\ &= \sum_{j=1}^{k} I(U_j; X^d), && \text{(A-28)} \end{aligned}$$

and where (A-27) follows from the facts that $U_1, \ldots, U_k$ are mutually independent and are conditionally independent given X^d.

For each i, the mutual information $I(U^k; X_i)$ can be approximated as

$$I(U^k; X_i) = \frac{1}{2}\epsilon^2 \mathbb{E}_{U^k}\left[\left\|\boldsymbol{\psi}_{i,U^k}\right\|^2\right] + o(\epsilon^2),$$

where for $U^k = u^k$, the vector $\boldsymbol{\psi}_{i,u^k}$ is an $|\mathcal{X}_i|$-dimensional perturbation vector defined as

$$\psi_{i,u^k}(x_i) = \frac{P_{X_i|U^k}(x_i|u^k) - P_{X_i}(x_i)}{\epsilon\sqrt{P_{X_i}(x_i)}}.$$

Therefore, by ignoring the higher order terms of ϵ, the maximization of total correlation can be rewritten as

$$\max_{\boldsymbol{\phi}_{u^k}} \quad \sum_{i=1}^{d} \mathbb{E}_{U^k}\left[\left\|\boldsymbol{\psi}_{i,U^k}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] \tag{A-29a}$$

$$\text{subject to: } 1 \geq \mathbb{E}_{U_1}\left[\left\|\boldsymbol{\phi}_{U_1}\right\|^2\right] \geq \cdots \geq \mathbb{E}_{U_k}\left[\left\|\boldsymbol{\phi}_{U_k}\right\|^2\right] \tag{A-29b}$$

$$\langle \boldsymbol{\phi}_{u_i}, \boldsymbol{\phi}_{u_j}\rangle = 0, \quad i \neq j, u_i \in \mathcal{U}_i, u_j \in \mathcal{U}_j \tag{A-29c}$$

$$\left\langle \boldsymbol{\phi}_{u_j}, \boldsymbol{\phi}^{(0)} \right\rangle = 0, \forall\, u_j \in \mathcal{U}_j, j = 1, \dots, k \tag{A-29d}$$

$$\boldsymbol{\phi}_{u^k} = \sum_{j=1}^{k} \boldsymbol{\phi}_{u_j}, \forall\, u^k \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_k. \tag{A-29e}$$

To solve (A-29), first observe that we have $\boldsymbol{\psi}_{i,U^k} = \mathbf{B}_i \boldsymbol{\phi}_{U^k}$, where $\mathbf{B}_i$ is as defined in (A-8). Then, the objective function (A-29a) can be rewritten as

$$\begin{aligned}
&\sum_{i=1}^{d} \mathbb{E}_{U^k}\left[\left\|\boldsymbol{\psi}_{i,U^k}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}_{U^k}\left[\left\|\mathbf{B}_i \boldsymbol{\phi}_{U^k}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(A-30)} \\
&= \mathbb{E}_{U^k}\left[\left\|\mathbf{B}_0 \boldsymbol{\phi}_{U^k}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(A-31)} \\
&= \mathbb{E}_{U^k}\left[\left\|\sum_{j=1}^{k} \mathbf{B}_0 \boldsymbol{\phi}_{U_j}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(A-32)} \\
&= \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\mathbf{B}_0 \boldsymbol{\phi}_{U_j}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(A-33)} \\
&= \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\mathbf{B}_0 \boldsymbol{\phi}_{U_j}\right\|^2 - \left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(A-34)}
\end{aligned}$$

where $\mathbf{B}_0$ is as defined in (A-9). To obtain (A-33), we have used the fact that, for $i \neq j$,

$$\begin{aligned}
\mathbb{E}_{U^k}\left[\boldsymbol{\phi}_{U_i}^{\mathrm{T}} \mathbf{B}_0^{\mathrm{T}} \mathbf{B}_0 \boldsymbol{\phi}_{U_j}\right] &= \left(\mathbb{E}_{U_i}\left[\boldsymbol{\phi}_{U_i}\right]\right)^{\mathrm{T}} \mathbf{B}_0^{\mathrm{T}} \mathbf{B}_0 \left(\mathbb{E}_{U_j}\left[\boldsymbol{\phi}_{U_j}\right]\right) \\
&= 0,
\end{aligned}$$

where the first equality follows from the fact that U_i and U_j are independent, and the second equality follows from that $\mathbb{E}_{U_i}\left[\boldsymbol{\phi}_{U_i}\right] = 0$.

To maximize (A-34), $\boldsymbol{\phi}_{u_i}$ should be aligned to the same direction for all $u_i \in \mathcal{U}_i$. Otherwise, we can align all $\boldsymbol{\phi}_{u_i}$ to

$$\operatorname*{arg\,max}_{\boldsymbol{\phi}_{u_i} \colon u_i \in \mathcal{U}_i} \frac{\|\mathbf{B}_0 \boldsymbol{\phi}_{u_i}\|^2}{\|\boldsymbol{\phi}_{u_i}\|^2}$$

while keeping $\mathbb{E}_{U_i}[\|\boldsymbol{\phi}_{U_i}\|^2]$ fixed, which yields a larger value for the objective function.

Therefore, for each i and $u_i \in \mathcal{U}_i$, we can write $\boldsymbol{\phi}_{u_i}$ as

$$\boldsymbol{\phi}_{u_i} = h_i(u_i)\boldsymbol{\phi}_i, \tag{A-35}$$

where $h_i \colon \mathcal{U}_i \mapsto \mathbb{R}$ and $\boldsymbol{\phi}_i$ is a unit-norm vector. Then, we have $\mathbb{E}_{U_i}[\boldsymbol{\phi}_{U_i}] = \mathbb{E}_{U_i}[h_i(U_i)]\boldsymbol{\phi}_i = 0$ and

$$\mathbb{E}_{U_i}[\|\boldsymbol{\phi}_{U_i}\|^2] = \mathbb{E}_{U_i}[h_i^2(U_i)], \tag{A-36a}$$

$$\mathbb{E}_{U_i}[\|\mathbf{B}_0\boldsymbol{\phi}_{U_i}\|^2] = \mathbb{E}_{U_i}[h_i^2(U_i)]\|\mathbf{B}_0\boldsymbol{\phi}_i\|^2. \tag{A-36b}$$

Now, the constraint (A-29b) can be reduced to

$$1 \geq \mathbb{E}_{U_1}[h_1^2(U_1)] \geq \cdots \geq \mathbb{E}_{U_k}[h_k^2(U_k)]. \tag{A-37}$$

In addition, it follows from (A-36) that $\mathbb{E}_{U_i}[\|\mathbf{B}_0\boldsymbol{\phi}_{U_i}\|^2] - \mathbb{E}_{U_i}[\|\boldsymbol{\phi}_{U_i}\|^2] = \mathbb{E}_{U_i}[h_i^2(U_i)]\left[\|\mathbf{B}_0\boldsymbol{\phi}_i\|^2 - 1\right]$. As a result, to maximize (A-34), h_i should be chosen such that

$$\mathbb{E}_{U_i}[h_i^2(U_i)] = \begin{cases} 1 & \text{if } \|\mathbf{B}_0\boldsymbol{\phi}_i\|^2 > 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then, from (A-37) there exists $k_0 \in \{1, \dots, k\}$ such that

$$\mathbb{E}_{U_i}[h_i^2(U_i)] = \begin{cases} 1 & i = 1, \dots, k_0 \\ 0 & i > k_0, \end{cases} \tag{A-38}$$

and the objective function (A-34) can be reduced to

$$\begin{aligned}
\sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\mathbf{B}_0\boldsymbol{\phi}_{U_j}\right\|^2 - \left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] &= \sum_{j=1}^{k_0} \|\mathbf{B}_0\boldsymbol{\phi}_i\|^2 - k_0 \\
&= \|\mathbf{B}_0\boldsymbol{\Phi}_0\|_{\mathrm{F}}^2 - k_0,
\end{aligned}$$

where we have defined $\boldsymbol{\Phi}_0 \triangleq \begin{bmatrix}\boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_{k_0}\end{bmatrix}$.

As a result, the optimization problem (A-29) is equivalent to

$$\max_{\boldsymbol{\Phi}_0} \quad \|\mathbf{B}_0\boldsymbol{\Phi}_0\|_{\mathrm{F}}^2 - k_0 \tag{A-39a}$$

$$\text{subject to: } \boldsymbol{\Phi}_0^{\mathrm{T}}\boldsymbol{\Phi}_0 = \mathbf{I}_{k_0} \tag{A-39b}$$

$$\boldsymbol{\Phi}_0^{\mathrm{T}}\boldsymbol{\phi}^{(0)} = \mathbf{0}_{k_0}, \tag{A-39c}$$

where $\mathbf{I}_{k_0}$ is the identity matrix of order k_0, and $\mathbf{0}_{k_0}$ is the zero vector in $\mathbb{R}^{k_0}$. In addition, since $\boldsymbol{\phi}^{(0)}$ is the first right singular vector of $\mathbf{B}_0$, (A-39) can be further reduced to

$$\max_{\boldsymbol{\Phi}_0} \quad \left\|\tilde{\mathbf{B}}_0\boldsymbol{\Phi}_0\right\|_{\mathrm{F}}^2 - k_0 \tag{A-40a}$$

$$\text{subject to: } \boldsymbol{\Phi}_0^{\mathrm{T}}\boldsymbol{\Phi}_0 = \mathbf{I}_{k_0}, \tag{A-40b}$$

where $\tilde{\mathbf{B}}_0 \triangleq \mathbf{B}_0 - \sqrt{\lambda^{(0)}}\boldsymbol{\psi}^0\left(\boldsymbol{\phi}^{(0)}\right)^{\mathrm{T}}$.

From Lemma A-1, the optimal value of (A-40) is

$$\sum_{i=1}^{k_0} \lambda^{(i)} - k_0 = \sum_{i=1}^{k_0}\left[\lambda^{(i)} - 1\right]. \tag{A-41}$$

To maximize (A-41), k_0 should be chosen as the largest i such that $\lambda^{(i)} > 1$, i.e., $k_0 = \min\{k, k^*\}$. In addition, the optimal $\boldsymbol{\Phi}_0$ is $\boldsymbol{\Phi}_0 = \begin{bmatrix}\boldsymbol{\phi}^{(1)} & \cdots & \boldsymbol{\phi}^{(k_0)}\end{bmatrix}\mathbf{Q}$ for $\mathbf{Q} \in \mathbb{R}^{k_0 \times k_0}$ with $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_{k_0}$. Hence, we have

$$\boldsymbol{\phi}_\ell = \sum_{j=1}^{k_0} q_{j\ell}\boldsymbol{\phi}^{(j)}.$$

Following the same derivation as that for (A-12), we can express $\phi^{(j)}$ as

$$\frac{\phi^{(j)}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \frac{1}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i).$$

and thus

$$\frac{\phi_\ell(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{j=1}^{k_0} q_{j\ell} \cdot \frac{\phi^{(j)}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i).$$

Then, it follows from (A-35) that

$$\frac{\phi_{u_\ell}(x^d)}{\sqrt{P_{X^d}(x^d)}} = h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i) \tag{A-42}$$

for $\ell = 1, \dots, k_0$. Moreover, from (A-29e), we have

$$\boldsymbol{\phi}_{u^k} = \sum_{\ell=1}^{k} \boldsymbol{\phi}_{u_\ell} = \sum_{\ell=1}^{k_0} \boldsymbol{\phi}_{u_\ell},$$

where the second equality follows from the consequence of (A-35) and (A-38) that $\boldsymbol{\phi}_{u_\ell} = \mathbf{0}$ for $\ell > k_0$.

Therefore,

$$\begin{aligned}\frac{\phi_{u^k}(x^d)}{\sqrt{P_{X^d}(x^d)}} &= \sum_{\ell=1}^{k_0} \frac{\phi_{u_\ell}(x^d)}{\sqrt{P_{X^d}(x^d)}} \\ &= \sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i),\end{aligned}$$

which implies

$$\begin{aligned}&P_{X^d|U^k}(x^d|u^k) \\ &= P_{X^d}(x^d) \left[1 + \epsilon \frac{\phi_{u^k}(x^d)}{\sqrt{P_{X^d}(x^d)}}\right] + o(\epsilon) \\ &= P_{X^d}(x^d) \left[1 + \epsilon \sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i)\right] + o(\epsilon)\end{aligned}$$

and

$$\begin{aligned}&P_{X^d U^k}(x^d, u^k) \\ &\quad = P_{X^d}(x^d) \left[\prod_{j=1}^{k} P_{U_j}(u_j)\right] \\ &\qquad \cdot \left[1 + \epsilon \sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i)\right] + o(\epsilon).\end{aligned} \tag{A-43}$$

Finally, note that the exponential family $\mathcal{P}_{\exp,k}^{(\delta)}$, when δ is small, can be written as

$$\begin{aligned}\mathcal{P}_{\exp,k}^{(\delta)} = \Bigg\{ &P_{X^d}(x^d) \left[\prod_{j=1}^{k} P_{U_j}(u_j)\right] \\ &\cdot \left[1 + \sqrt{2\delta} \sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i)\right] \\ &: \ h_\ell \in \mathcal{H}_\ell, \mathbf{Q} = [q_{ij}]_{k_0 \times k_0}, \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_{k_0} \Bigg\}.\end{aligned} \tag{A-44}$$

Since $\delta = \frac{1}{2}\epsilon^2$ the proof is completed by comparing (A-44) and (A-43).

Appendix D
Joint Correlation Maximization

For functions $\underline{f}_i : \mathcal{X}_i \mapsto \mathbb{R}^k$, $i = 1, \dots, d$, we define $\boldsymbol{\Psi}_i \in \mathbb{R}^{|\mathcal{X}_i| \times k}$ such that the row vectors of $\boldsymbol{\Psi}_i$ are $\sqrt{P_{X_i}(x_i)}\underline{f}_i^{\mathrm{T}}(x_i)$, for all $x_i \in \mathcal{X}_i$. Furthermore, we define the $m \times k$ matrix $\boldsymbol{\Psi}$ as $\boldsymbol{\Psi} = \begin{bmatrix}\boldsymbol{\Psi}_1^{\mathrm{T}} & \cdots & \boldsymbol{\Psi}_d\end{bmatrix}$. Then the optimization problem (12) can be rewritten as

$$\max_{\boldsymbol{\Psi}:\ \boldsymbol{\Psi} \in \mathbb{R}^{m \times k}} \ \mathrm{tr}\left\{\boldsymbol{\Psi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\Psi}\right\} \tag{A-45a}$$

$$\text{subject to:} \quad \boldsymbol{\Psi}_i^{\mathrm{T}}\boldsymbol{v}_i = \mathbf{0}_k, \ \text{for all } i \tag{A-45b}$$

$$\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} = \mathbf{I}_k, \tag{A-45c}$$

where $\mathbf{0}_k$ is the zero vector in $\mathbb{R}^k$, and $\mathbf{I}_k$ is the $k \times k$ identity matrix. To establish the equivalence of (12) and (A-45), note that we have

$$\begin{aligned}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} = \sum_{i=1}^{d} \boldsymbol{\Psi}_i^{\mathrm{T}}\boldsymbol{\Psi}_i &= \sum_{i=1}^{d} \sum_{x_i \in \mathcal{X}_i} P_{X_i}(x_i)\underline{f}_i(x_i)\underline{f}_i^{\mathrm{T}}(x_i) \\ &= \sum_{i=1}^{d} \mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_i^{\mathrm{T}}(X_i)\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{d} \underline{f}_i(X_i)\underline{f}_i^{\mathrm{T}}(X_i)\right]\end{aligned}$$

and

$$\begin{aligned}\mathrm{tr}\left\{\boldsymbol{\Psi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\Psi}\right\} &= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{tr}\left\{\boldsymbol{\Psi}_i^{\mathrm{T}}\mathbf{B}_{ij}\boldsymbol{\Psi}_j\right\} \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{tr}\left\{\mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_j^{\mathrm{T}}(X_j)\right]\right\} \\ &= \sum_{i=1}^{d} \mathrm{tr}\left\{\mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_i^{\mathrm{T}}(X_i)\right]\right\} \\ &\qquad + \sum_{i \neq j} \mathrm{tr}\left\{\mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_j^{\mathrm{T}}(X_j)\right]\right\} \\ &= \mathrm{tr}\left\{\sum_{i=1}^{d} \mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_i^{\mathrm{T}}(X_i)\right]\right\} \\ &\qquad + \sum_{i \neq j} \mathrm{tr}\left\{\mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_j^{\mathrm{T}}(X_j)\right]\right\} \\ &= k + \mathbb{E}\left[\sum_{i \neq j} \underline{f}_i^{\mathrm{T}}(X_i)\underline{f}_j(X_j)\right].\end{aligned}$$

From Lemma 1, for $k < m - d$, the solution of (A-45) can be represented as $\boldsymbol{\Psi}^* = \begin{bmatrix}\boldsymbol{\psi}^{(1)} & \cdots & \boldsymbol{\psi}^{(k)}\end{bmatrix}\mathbf{Q}$, where $\mathbf{Q} \in \mathbb{R}^{k \times k}$ is an orthogonal matrix. Therefore, the optimal solution of (12) corresponds to $f_i^{(\ell)}$ with $i = 1, \dots, d$ and $\ell = 1, \dots, k$.

Appendix E
Common Bits Patterns Extraction

First, we define $\ell_{\max}$ as the largest ℓ such that $w(\mathcal{J}_\ell) > 0$, i.e., $\ell_{\max} \triangleq \max\{\ell\colon 0 \le \ell \le 2^r - 1, w(\mathcal{J}_\ell) > 0\}$. Then, $w(\mathcal{J}_\ell) > 0$ is equivalent to $\ell \le \ell_{\max}$, and (14) can be equivalently expressed as

$$\lambda^{(\ell)} = w(\mathcal{J}_\ell), \quad \ell \le \ell_{\max}, \tag{A-46}$$

and

$$\lambda^{(\ell)} = 0, \quad \ell > \ell_{\max}. \tag{A-47}$$

Note that (7) establishes a one-to-one correspondence between the functions $f_i^{(\ell)}$ $(i = 1, \dots, d)$ and the vector $\boldsymbol{\psi}^{(\ell)}$. With this correspondence, we use $\tilde{\boldsymbol{\psi}}^{(\ell)}$ to denote the vector corresponding to the functions $f_i^{(\ell)}$ as defined in (15). Then the proof can be accomplished in two steps. First, we show that $\tilde{\boldsymbol{\psi}}^{(\ell)}$ $(\ell = 0, \dots, \ell_{\max})$ are $(\ell_{\max} + 1)$ orthogonal eigenvectors of $\mathbf{B}$ associated with eigenvalues $w(\mathcal{J}_\ell)$ $(\ell = 0, \dots, \ell_{\max})$, i.e., for all $0 \le \ell \le \ell_{\max}$ and $0 \le \ell' \le \ell_{\max}$, the $\tilde{\boldsymbol{\psi}}^{(\ell)}$'s satisfy

$$\mathbf{B}\tilde{\boldsymbol{\psi}}^{(\ell)} = w(\mathcal{J}_\ell)\tilde{\boldsymbol{\psi}}^{(\ell)} \quad \text{and} \quad \left\langle \tilde{\boldsymbol{\psi}}^{(\ell)}, \tilde{\boldsymbol{\psi}}^{(\ell')} \right\rangle = \delta_{\ell\ell'}, \tag{A-48}$$

where $\delta_{\ell\ell'}$ is the Kronecker delta. Then, it suffices to verify that all other eigenvalues of $\mathbf{B}$ are zeros [cf. (A-47)].

To begin, we equivalently express (A-48) using $f_i^{(\ell)}$ as

$$\sum_{j=1}^{d} \mathbb{E}\left[f_j^{(\ell)}(X_j)\middle|X_i\right] = w(\mathcal{J}_\ell) f_i^{(\ell)}(X_i), \quad 1 \le i \le d, \tag{A-49}$$

and

$$\sum_{i=1}^{d} \mathbb{E}\left[f_i^{(\ell)}(X_i) f_i^{(\ell')}(X_i)\right] = \delta_{\ell\ell'}. \tag{A-50}$$

Then, since we have [cf. (13)]

$$\sum_{i=1}^{d} \mathbb{1}_{\{\mathcal{J}_\ell \subset \mathcal{I}_i\}} = \sum_{j=1}^{d} \mathbb{1}_{\{\mathcal{J}_\ell \subset \mathcal{I}_j\}} = w(\mathcal{J}_\ell),$$

it suffices to show that

$$\mathbb{E}\left[f_j^{(\ell)}(X_j)\middle|X_i\right] = f_i^{(\ell)}(X_i) \cdot \mathbb{1}_{\{\mathcal{J}_\ell \subset \mathcal{I}_j\}}, \quad 1 \le i, j \le d, \tag{A-51}$$

and

$$\mathbb{E}\left[f_i^{(\ell)}(X_i) f_i^{(\ell')}(X_i)\right] = \frac{\mathbb{1}_{\{\mathcal{J}_\ell \subset \mathcal{I}_i\}}}{w(\mathcal{J}_\ell)} \cdot \delta_{\ell\ell'}, \quad 1 \le i \le d. \tag{A-52}$$

To obtain (A-51), note that if $\mathcal{J}_\ell \not\subset \mathcal{I}_j$, it follows from (15) that $f_j(X_j) = 0$, and thus (A-51) holds. Otherwise, we have $\mathcal{J}_\ell \subset \mathcal{I}_j$ and

$$\mathbb{E}\left[f_j^{(\ell)}(X_j)\middle|X_i\right] = \frac{1}{\sqrt{w(\mathcal{J}_\ell)}} \mathbb{E}\left[\prod_{s \in \mathcal{J}_\ell} b_s \middle| X_i\right]. \tag{A-53}$$

Since $X_i = b_{\mathcal{I}_i}$ is composed of all the b_s's with indices in $\mathcal{I}_i$, we have

$$\mathbb{E}\left[\prod_{s \in \mathcal{J}_\ell} b_s \middle| X_i\right] = \begin{cases} \prod_{s \in \mathcal{J}_\ell} b_s & \text{if } \mathcal{J}_\ell \subset \mathcal{I}_i \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we obtain

$$\begin{aligned} \mathbb{E}\left[f_j^{(\ell)}(X_j)\middle|X_i\right] &= \frac{1}{\sqrt{w(\mathcal{J}_\ell)}} \mathbb{E}\left[\prod_{s \in \mathcal{J}_\ell} b_s \middle| X_i\right] \\ &= \begin{cases} \frac{1}{\sqrt{w(\mathcal{J}_\ell)}} \prod_{s \in \mathcal{J}_\ell} b_s & \text{if } \mathcal{J}_\ell \subset \mathcal{I}_i \\ 0 & \text{otherwise} \end{cases} \\ &= f_i^{(\ell)}(X_i) = f_i^{(\ell)}(X_i) \cdot \mathbb{1}_{\{\mathcal{J}_\ell \subset \mathcal{I}_j\}}. \end{aligned}$$

Likewise, (A-52) follows immediately from (15) when $\ell = \ell'$, and it suffices to consider the case $\ell \ne \ell'$ and prove that

$$\mathbb{E}\left[f_i^{(\ell)}(X_i) f_i^{(\ell')}(X_i)\right] = 0. \tag{A-54}$$

Indeed, when $\mathcal{J}_\ell \not\subset \mathcal{I}_i$ or $\mathcal{J}_{\ell'} \not\subset \mathcal{I}_i$, (A-54) is trivially true. Otherwise, we have $\mathcal{J}_\ell \subset \mathcal{I}_i$ and $\mathcal{J}_{\ell'} \subset \mathcal{I}_i$, and it follows from (15) that

$$f_i^{(\ell)}(X_i) f_i^{(\ell')}(X_i) = \frac{1}{\sqrt{w(\mathcal{J}_\ell) w(\mathcal{J}_{\ell'})}} \prod_{j \in \mathcal{J}_\ell \,\triangle\, \mathcal{J}_{\ell'}} b_j,$$

where "$\triangle$" denotes the symmetric difference of two sets, i.e., $A \triangle B = (A \setminus B) \cup (B \setminus A)$. Therefore, we have

$$\mathbb{E}\left[f_i^{(\ell)}(X_i) f_i^{(\ell')}(X_i)\right] = \frac{1}{\sqrt{w(\mathcal{J}_\ell) w(\mathcal{J}_{\ell'})}} \prod_{j \in \mathcal{J}_\ell \,\triangle\, \mathcal{J}_{\ell'}} \mathbb{E}[b_j] = 0,$$

where we have used the fact that the set $(\mathcal{J}_\ell \triangle \mathcal{J}_{\ell'})$ is non-empty, since $\mathcal{J}_\ell \ne \mathcal{J}_{\ell'}$.

Finally, to prove (A-47), i.e., eigenvalues other that $w(\mathcal{J}_\ell)$ $(\ell = 0, \dots, \ell_{\max})$ are all zeros, note that

$$\begin{aligned} \sum_{\ell=0}^{\ell_{\max}} w(\mathcal{J}_\ell) = \sum_{\ell=0}^{2^r - 1} w(\mathcal{J}_\ell) &= \sum_{\mathcal{I} \subset [r]} w(\mathcal{I}) \\ &= \sum_{\mathcal{I} \subset [r]} \sum_{i=1}^{d} \mathbb{1}_{\{\mathcal{I} \subset \mathcal{I}_i\}} \\ &= \sum_{i=1}^{d} \sum_{\mathcal{I} \subset [r]} \mathbb{1}_{\{\mathcal{I} \subset \mathcal{I}_i\}} \\ &= \sum_{i=1}^{d} 2^{|\mathcal{I}_i|} = \sum_{i=1}^{d} |\mathcal{X}_i| = m. \end{aligned}$$

On the other hand, we have the sum of all eigenvalues

$$\sum_{\ell=0}^{m-1} \lambda^{(\ell)} = \operatorname{tr}\{\mathbf{B}\} = m.$$

From Lemma 1, all eigenvalues of $\mathbf{B}$ are non-negative, which implies (A-47).

APPENDIX F
PROOF OF PROPOSITION 1

To begin, we write the matrix $\tilde{\mathbf{B}}$ of (6) as a block matrix

$$\tilde{\mathbf{B}} = \begin{bmatrix} \tilde{\mathbf{B}}_{11} & \tilde{\mathbf{B}}_{12} & \cdots & \tilde{\mathbf{B}}_{1d} \\ \tilde{\mathbf{B}}_{21} & \tilde{\mathbf{B}}_{22} & \cdots & \tilde{\mathbf{B}}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{B}}_{d1} & \tilde{\mathbf{B}}_{d2} & \cdots & \tilde{\mathbf{B}}_{dd} \end{bmatrix}, \tag{A-55}$$

where each block $\tilde{\mathbf{B}}_{ij}$ is an $(|\mathcal{X}_i| \times |\mathcal{X}_j|)$-dimensional matrix. Then, we can rewrite $\left\|\tilde{\mathbf{B}} - \mathbf{\Psi}\mathbf{\Psi}^{\mathrm{T}}\right\|_{\mathrm{F}}^2$ as

$$\begin{aligned} &\left\|\tilde{\mathbf{B}} - \mathbf{\Psi}\mathbf{\Psi}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d}\left\|\tilde{\mathbf{B}}_{ij} - \mathbf{\Psi}_i\mathbf{\Psi}_j^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d}\left[\left\|\tilde{\mathbf{B}}_{ij}\right\|_{\mathrm{F}}^2 - 2\,\mathrm{tr}\left\{\mathbf{\Psi}_i^{\mathrm{T}}\tilde{\mathbf{B}}_{ij}\mathbf{\Psi}_j\right\} + \left\|\mathbf{\Psi}_i\mathbf{\Psi}_j\right\|_{\mathrm{F}}^2\right] \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d}\left[\left\|\tilde{\mathbf{B}}_{ij}\right\|_{\mathrm{F}}^2 - 2H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right)\right] \\ &= \left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 - 2H\left(\underline{f}_1(X_1), \ldots, \underline{f}_d(X_d)\right), \end{aligned} \tag{A-56}$$

where we have used the fact that

$$\begin{aligned} &\mathrm{tr}\left\{\mathbf{\Psi}_i^{\mathrm{T}}\tilde{\mathbf{B}}_{ij}\mathbf{\Psi}_j\right\} - \frac{1}{2}\left\|\mathbf{\Psi}_i\mathbf{\Psi}_j\right\|_{\mathrm{F}}^2 \\ &= \mathbb{E}\left[\underline{f}_i^{\mathrm{T}}(X_i)\underline{f}_j(X_j)\right] - \left(\mathbb{E}\left[\underline{f}_i(X_i)\right]\right)^{\mathrm{T}}\mathbb{E}\left[\underline{f}_j(X_j)\right] \\ &\quad - \frac{1}{2}\,\mathrm{tr}\left\{\mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_i^{\mathrm{T}}(X_i)\right]\mathbb{E}\left[\underline{f}_j(X_i)\underline{f}_j^{\mathrm{T}}(X_j)\right]\right\} \\ &= H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right). \end{aligned}$$

REFERENCES

[1] Y. LeCun, C. Cortes, and C. J. Burges, "Mnist handwritten digit database," *AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[2] I. T. Jolliffe, "Principal component analysis. 1986," *Spring-verlag, New York*, 1986.

[3] Q. Huang, F. Wang, and L. Guibas, "Functional map networks for analyzing and exploring large shape collections," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 36:1–36:11, Jul. 2014.

[4] F. Wang, Q. Huang, and L. J. Guibas, "Image co-segmentation via consistent functional maps," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, Washington, DC, USA, 2013, pp. 849–856.

[5] F. Wang, Q. Huang, M. Ovsjanikov, and L. J. Guibas, "Unsupervised multi-class joint image segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2014, pp. 3142–3149.

[6] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of research and development*, vol. 4, no. 1, pp. 66–82, 1960.

[7] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American statistical Association*, vol. 80, no. 391, pp. 580–598, 1985.

[8] A. Rényi, "On measures of dependence," *Acta mathematica hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[9] G. J. Op't Veld and M. C. Gastpar, "Caching gaussians: Minimizing total correlation on the Gray-Wyner network," in *2016 Annual Conference on Information Science and Systems (CISS)*. IEEE, 2016, pp. 478–483.

[10] ——, "Total correlation of Gaussian vector sources on the Gray-Wyner network," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 385–392.

[11] G. Ver Steeg and A. Galstyan, "Discovering structure in high-dimensional data through correlation explanation," in *Advances in Neural Information Processing Systems*, 2014, pp. 577–585.

[12] G. H. Golub and C. Van Loan, "Matrix computations 4th edition," 2013.

[13] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "On universal features for high-dimensional learning and inference," *submitted to IEEE Trans. Inform. Theory*, 2019. Preprint.

[14] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[15] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," *arXiv preprint arXiv:1905.06600*, 2019.

[16] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang, "An efficient approach to informative feature extraction from multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5281–5288.

[17] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*. IEEE, 2015, pp. 972–979.

[18] M. Razaviyayn, F. Farnia, and D. Tse, "Discrete Rényi classifiers," in *Advances in Neural Information Processing Systems*, 2015, pp. 3276–3284.

[19] S. Feizi, A. Makhdoumi, K. Duffy, M. Medard, and M. Kellis, "Network maximal correlation," *arXiv preprint arXiv:1606.04789*, 2016.

[20] S. Feizi and D. Tse, "Maximally correlated principal component analysis," *arXiv preprint arXiv:1702.05471*, 2017.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] Y. B. Y. LeCun, L. Bottou and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278 – 2324, Nov. 1998.

[23] S.-L. Huang and L. Zheng, "Linear information coupling problems," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 1029–1033.

[24] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.