
On the Fundamental Trade-offs in Learning Invariant Representations (Supplementary)

Anonymous Author(s)

Affiliation

Address

email

1 In this supplementary material, we include (1) further numerical experiments and experimental details
2 for estimating the different trade-offs in Section A (2) proofs of all theorems, lemmas and corollaries
3 in Section B.

4 A Numerical Estimation of Trade-Offs

5 A.1 Training Details of Numerical Estimation of Trade-Offs in Section 4

6 In this Section we specify all the training details of the numerical results presented in Section 4 of
7 the main paper.

8 **Trade-Offs D & L:** The optimal embedding z is learned for the trade-off **D** through the closed
9 form solution in Theorem 5 for different invariance parameter values τ in $[0, 1]$. Then this optimal
10 embedding is fed to a target task predictor which is an MLP with two hidden layers, and 4, 8 neurons
11 where we use MSE as a loss function and AdamW [1] as an optimizer. The same procedure is
12 implemented for trade-off **L**, except that the input data is v instead of x . We choose the number
13 of epochs and batch-size to be 500 and optimize the learning rate by trying six different values
14 among $\{10^{-2}, 10^{-3}, 3 \times 10^{-4}, 5 \times 10^{-4}, 10^{-4}, 10^{-5}\}$. We consider Gaussian kernel for all \mathcal{H}_x , \mathcal{H}_s ,
15 and \mathcal{H}_y and seek the band-width (i.e., σ) of Gaussian kernels using five different logarithmically
16 spaced values in $[10^{-2}, 10^2]$. Further, we optimize the regularization parameter γ in equation (10)
17 by considering three values among $\{0, 10^{-4}, 1\}$. We first set $\sigma_x = \sigma_s = \sigma_y = 1$, $\gamma = 10^{-4}$ and
18 explore the optimal learning rate by minimizing MSE in the validation set. Once the learning rate is
19 found, we explore the σ s by minimizing MSE in the validation set. In the end, we explore γ similarly.

20 **Spectral Adversarial Representation Learning (Spectral-ARL):** Spectral ARL [2] is very similar
21 to the trade-off **D** of this paper except that \mathcal{H}_s and \mathcal{H}_y are both linear RKHS. We followed the same
22 experimental setting of trade-off **D**. The results of this approach is illustrated in Figure 1 (c).

23 **Adversarial Representation Learning (ARL):** We followed the ARL formulation in (3) for different
24 invariance parameter values τ in $[0, 1]$. The embedding $z = f(x)$ is extracted via the encoder $f(\cdot)$
25 which is an MLP with two hidden layers, and 4, 2 neurons. Then, z is fed to a target task predictor
26 $g_Y(\cdot)$ and an proxy adversary $g_S(\cdot)$ network where both are MLP with two hidden layers, and 4, 8
27 neurons. We use stochastic gradient descent-ascent (SGDA) [3] with AdamW [1] as an optimizer to
28 alternately train the encoder, target predictor and proxy adversary networks. We choose the number
29 of epochs and batch-size to be 500 and optimize the learning rate among $\{10^{-2}, 10^{-3}, 3 \times 10^{-4}, 5 \times$
30 $10^{-4}, 10^{-4}, 10^{-5}\}$ by minimizing MSE in the validation set by. Since ARL can be unstable, we run
31 our experiment for five different random seeds. The mean and standard-deviation (std) of the results
32 are illustrated in Figure 1.

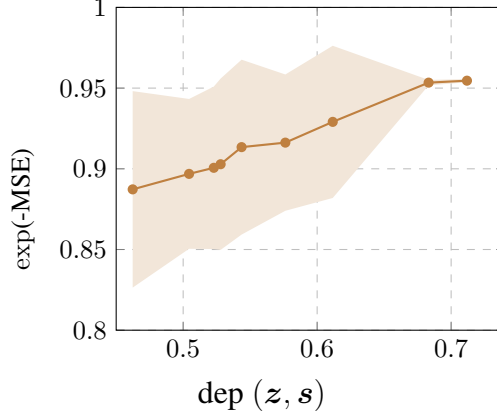


Figure 1: The mean and std of ARL method (ARL is optimized with SGDA [3]) after running five times with different random seeds on weight initialization.

33 A.2 Fair Classification

34 We consider an additional experiment, a fair classification application of IRL on Adult dataset¹. This
 35 dataset contains 45, 222 instances of different individual where each instance includes 14 attributes.
 36 The target task is a binary classification of annual income (more or less than 50K) while the sensitive
 37 attribute (i.e., semantic attribute) in which we aim to be independent of it is s =(race, gender). We
 38 randomly split the data into training (25, 222 instances), validation (10, 000 instances), and testing
 39 (10, 000 instances) and perform our experiment five times (each time with a different random seed on
 40 data split and weight initialization of involved networks).

41 We consider demographic parity (DP) [4] as the fairness criterion, where the goal is to have the
 42 prediction of target feature \hat{y} be independent of the sensitive feature s . In the context of representation
 43 learning, DP exactly falls into IRL since $\hat{y} = g_Y(z)$ is required by DP to be independent of s
 44 regardless of the target predictor $g_Y(\cdot)$. Following [5], we define DP violation (DPV) as

$$\text{DPV} = \max_{s_0, s'_0} \left| \mathbb{P}[\hat{y} | s = s_0] - \mathbb{P}[\hat{y} | s = s'_0] \right|. \quad (\text{A.1})$$

45 Following Section A.1, we learn the optimal embedding z for the trade-offs **D** and **L** using Theorem
 46 5 for different invariance parameter values $\tau \in [0, 1]$ and then feed this representation to a three-
 47 layer MLP with 64, 128, and 64 neurons, respectively. Similar RKHSs together with optimization
 48 procedure (except that the batch-size is 250) and hyperparameter tuning as Section A.1 is deployed.
 49 The mean of results are illustrated in Figure 2. Further, the std from five random splits is depicted in
 50 Figure 3 (a) and (b). Note that the baseline spectral-ARL [2] is almost similar to trade-off **D** where
 51 linear RKHS is used for both \mathcal{H}_s and \mathcal{H}_y . The mean and std of results are illustrated in Figures 2
 52 and 3 (c). For ARL method, the encoder $f(\cdot)$ is a three-layer MLP with 64, 128 and 64 neurons,
 53 respectively. Both the target task predictor $g_Y(\cdot)$ and proxy adversary $g_S(\cdot)$ are MLP with the similar
 54 architecture to encoder. We followed the same optimization procedure and hyperparameter tuning
 55 as Section A.1 except that the batch-size is 250. The mean and std of ARL results are illustrated in
 56 Figures 2 and 3 (d).

57 We observe that i) As expected, trade-off **L** dominates trade-off **D**. ii) The two baseline methods
 58 (trade-offs **F**) are dominated by trade-off **D** which is due to the suboptimality of their optimization
 59 and dependence measures.

¹The data is downloaded from the UCI ML-epository at <https://archive.ics.uci.edu/ml/datasets/adult>.

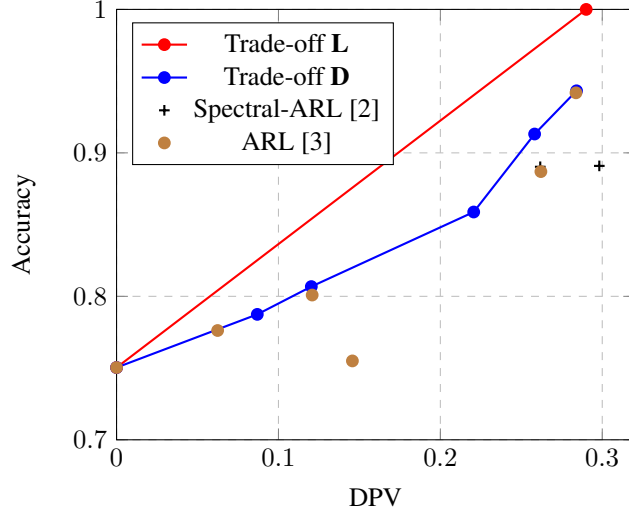


Figure 2: Fair classification: Two fundamental trade-offs, \mathbf{L} and \mathbf{D} , together with two baseline feasible trade-offs \mathbf{F} , ARL optimized with SGDA [3] and global optima of ARL with a linear RKHS [2].

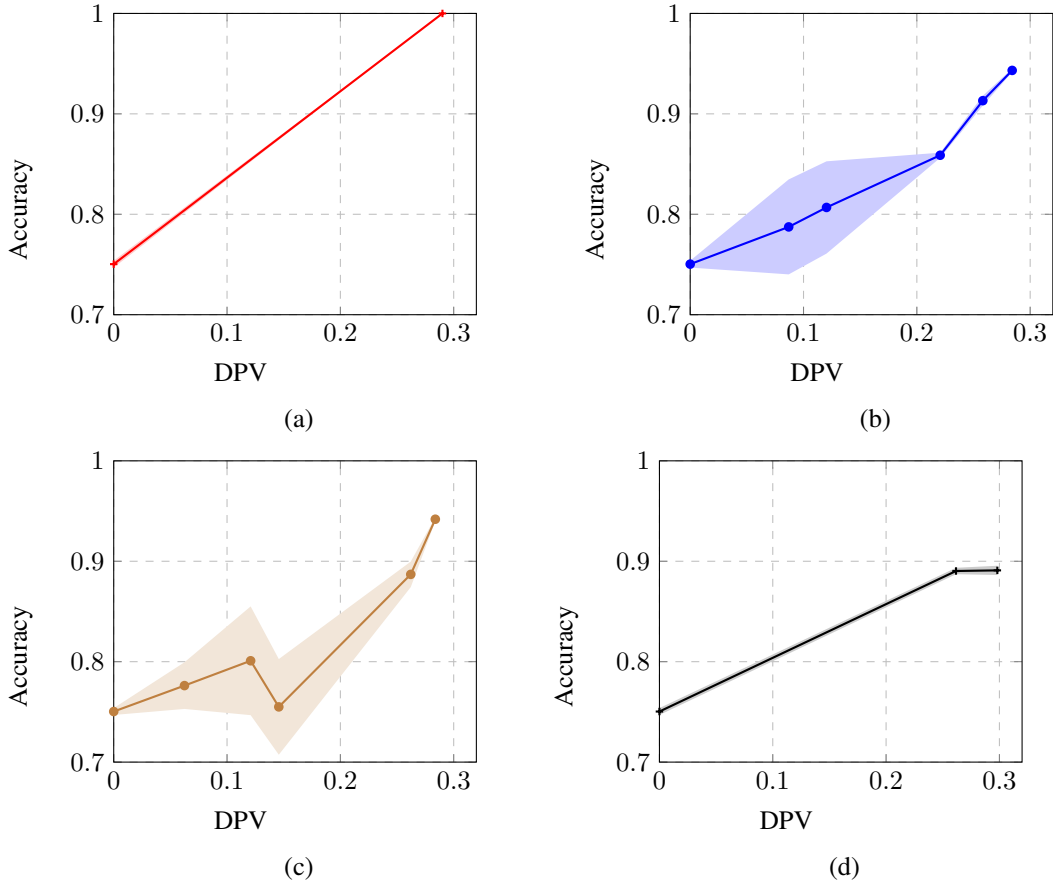


Figure 3: Fair classification: The mean and std of (a): trade-off \mathbf{L} , (b): trade-off \mathbf{D} , (c): ARL [3], and (d): Spectral-ARL [2].

60 B Proofs

61 B.1 Proof of Theorem 1

62 **Theorem 1.** Let \mathcal{H}_s contain all Borel-measurable functions and $\mathcal{L}_S(\cdot, \cdot)$ be mean squared error
63 (MSE) loss. Then,

$$z \in \arg \sup \left\{ \inf_{g_S \in \mathcal{H}_s} \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[\mathcal{L}_S(g_S(z), \mathbf{s}) \right] \right\} \Leftrightarrow \mathbb{E}[\mathbf{s} | z] = \mathbb{E}[\mathbf{s}].$$

64 *Proof.* Let s_i , $(g_S(z))_i$, and $(\mathbb{E}[\mathbf{s} | z])_i$ denote the i 'th entries of \mathbf{s} , $g_S(z)$, and $\mathbb{E}[\mathbf{s} | z]$, respectively.
65 Then, it follows that

$$\begin{aligned} \inf_{g_S \in \mathcal{H}_s} \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[\mathcal{L}_S(g_S(z), \mathbf{s}) \right] &= \inf_{g_S \in \mathcal{H}_s} \sum_{i=1}^{d_s} \left((g_S(z))_i - s_i \right)^2 \\ &= \sum_{i=1}^{d_s} \left((\mathbb{E}[\mathbf{s} | z])_i - s_i \right)^2 \\ &\leq \sum_{i=1}^{d_s} \left((\mathbb{E}[\mathbf{s}])_i - s_i \right)^2 = \sum_{i=1}^{d_s} \text{Var}[s_i], \end{aligned}$$

66 where the second step is due to the optimality of conditional mean (i.e., Bayes estimation) for MSE [6]
67 and the last step is due the fact that the independency between z and \mathbf{s} leads to an upper bound on

68 MSE. Therefore, if $z \in \arg \sup \left\{ \inf_{g_S \in \mathcal{H}_s} \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[\mathcal{L}_S(g_S(z), \mathbf{s}) \right] \right\}$ then $\mathbb{E}[\mathbf{s} | z] = \mathbb{E}[\mathbf{s}]$.

69 On the other hand, if $\mathbb{E}[\mathbf{s} | z] = \mathbb{E}[\mathbf{s}]$, then it follows immediately that $z \in$
70 $\arg \sup \left\{ \inf_{g_S \in \mathcal{H}_s} \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[\mathcal{L}_S(g_S(z), \mathbf{s}) \right] \right\}$. \square

71 B.2 Proof of Lemma 2

Lemma 2.

$$\begin{aligned} \text{dep}(z, \mathbf{s}) &= \sum_{j=1}^r \left\{ \mathbb{E}_{\mathbf{x}, \mathbf{s}, \mathbf{x}', \mathbf{s}'} \left[f_j(\mathbf{x}) f_j(\mathbf{x}') k_s(\mathbf{s}, \mathbf{s}') \right] + \mathbb{E}_{\mathbf{x}}[f_j(\mathbf{x})] \mathbb{E}_{\mathbf{x}'}[f_j(\mathbf{x}')] \mathbb{E}_{\mathbf{s}, \mathbf{s}'}[k_s(\mathbf{s}, \mathbf{s}')] \right. \\ &\quad \left. - 2 \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[f_j(\mathbf{x}) \mathbb{E}_{\mathbf{x}'}[f_j(\mathbf{x}')] \mathbb{E}_{\mathbf{s}'}[k_s(\mathbf{s}, \mathbf{s}')] \right] \right\} \end{aligned}$$

72 where (\mathbf{x}, \mathbf{s}) and $(\mathbf{x}', \mathbf{s}')$ are independently drawn from the joint distribution $\mathbf{p}_{\mathbf{x}\mathbf{s}}$.

73 *Proof.* We first note that this Lemma is inspired by HSIC [7]. In our case, $\text{dep}(z, \mathbf{s})$ is defined for a
74 fixed \mathbf{f} where HS-norm is carried only on β_s , while HSIC considers HS-norm on both β_s and \mathbf{f} .

75 Using definition (8), we get

$$\begin{aligned} \text{dep}(z, \mathbf{s}) &= \sum_{\beta_s \in \mathcal{U}_s} \sum_{j=1}^r h^2(f_j, \beta_s) \\ &= \sum_{\beta_s \in \mathcal{U}_s} \sum_{j=1}^r \langle \beta_s, \Sigma_{\mathbf{s}\mathbf{x}} f_j \rangle_{\mathcal{H}_s}^2 \\ &= \sum_{j=1}^r \sum_{\beta_s \in \mathcal{U}_s} \langle \beta_s, \Sigma_{\mathbf{s}\mathbf{x}} f_j \rangle_{\mathcal{H}_s}^2 \end{aligned}$$

$$\begin{aligned}
&\stackrel{(a)}{=} \sum_{j=1}^r \|\Sigma_{\mathbf{s}\mathbf{x}} f_j\|_{\mathcal{H}_s}^2 \\
&= \sum_{j=1}^r \langle \Sigma_{\mathbf{s}\mathbf{x}} f_j, \Sigma_{\mathbf{s}\mathbf{x}} f_j \rangle_{\mathcal{H}_s} \\
&\stackrel{(b)}{=} \sum_{j=1}^r \mathbb{Cov}_{\mathbf{x}, \mathbf{s}} \left(f_j(\mathbf{x}), (\Sigma_{\mathbf{s}\mathbf{x}} f_j)(\mathbf{s}) \right) \\
&= \sum_{j=1}^r \mathbb{Cov}_{\mathbf{x}, \mathbf{s}} \left(f_j(\mathbf{x}), \langle k_{\mathbf{s}}(\cdot, \mathbf{s}), \Sigma_{\mathbf{s}\mathbf{x}} f_j \rangle_{\mathcal{H}_s} \right) \\
&= \sum_{j=1}^r \mathbb{Cov}_{\mathbf{x}, \mathbf{s}} \left(f_j(\mathbf{x}), \mathbb{Cov}_{\mathbf{x}', \mathbf{s}'} (f_j(\mathbf{x}'), k_{\mathbf{s}}(\mathbf{s}', \mathbf{s})) \right) \\
&= \sum_{j=1}^r \mathbb{Cov}_{\mathbf{x}, \mathbf{s}} \left(f_j(\mathbf{x}), \mathbb{E}_{\mathbf{x}', \mathbf{s}'} [f_j(\mathbf{x}') k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}')] - \mathbb{E}_{\mathbf{x}'} [f_j(\mathbf{x}')] \mathbb{E}_{\mathbf{s}'} [k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}')] \right) \\
&= \sum_{j=1}^r \left\{ \mathbb{E}_{\mathbf{x}, \mathbf{s}, \mathbf{x}', \mathbf{s}'} \left[f_j(\mathbf{x}) f_j(\mathbf{x}') k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}') \right] + \mathbb{E}_{\mathbf{x}} [f_j(\mathbf{x})] \mathbb{E}_{\mathbf{x}'} [f_j(\mathbf{x}')] \mathbb{E}_{\mathbf{s}, \mathbf{s}'} [k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}')] \right. \\
&\quad \left. - 2 \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[f_j(\mathbf{x}) \mathbb{E}_{\mathbf{x}'} [f_j(\mathbf{x}')] \mathbb{E}_{\mathbf{s}'} [k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}')] \right] \right\}
\end{aligned}$$

76 where (a) is due to Parseval relation for orthonormal basis and (b) is from the definition of $\Sigma_{\mathbf{s}\mathbf{x}}$
77 in (7). \square

78 B.3 Proof of Lemma 3

79 **Lemma 3.** Let an empirical estimation of covariance be

$$\mathbb{Cov}_{\mathbf{x}, \mathbf{s}}(f_j(\mathbf{x}), \beta_s(\mathbf{s})) \approx \frac{1}{n} \sum_{i=1}^n f_j(\mathbf{x}_i) \beta_s(\mathbf{s}_i) - \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n f_j(\mathbf{x}_i) \beta_s(\mathbf{s}_k). \quad (\text{B.1})$$

80 Then, the empirical estimator of $\text{dep}(\mathbf{z}, \mathbf{s})$ is given by

$$\text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s}) := \frac{1}{n^2} \|\Theta \mathbf{K}_{\mathbf{x}} \mathbf{H} \mathbf{L}_{\mathbf{s}}\|_F^2,$$

81 where $\mathbf{K}_{\mathbf{x}}, \mathbf{K}_{\mathbf{s}} \in \mathbb{R}^{n \times n}$ are Gram matrices corresponding to $\mathcal{H}_{\mathbf{x}}$ and $\mathcal{H}_{\mathbf{s}}$, respectively, $\mathbf{H} =$
82 $\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T$, and $\mathbf{L}_{\mathbf{s}}$ is a full column-rank matrix in which $\mathbf{L}_{\mathbf{s}} \mathbf{L}_{\mathbf{s}}^T = \mathbf{K}_{\mathbf{s}}$ (Cholesky factorization).
83 This empirical estimator in (9) has a bias of $\mathcal{O}(n^{-1})$ and a convergence rate of $\mathcal{O}(n^{-1/2})$.

84 *Proof.* Firstly, let us reconstruct the orthonormal set $\mathcal{U}_{\mathbf{s}}$ through i.i.d. observations $\{\mathbf{s}_j\}_{j=1}^n$. Invoking
85 representer theorem, for two arbitrary elements β_i and β_m of $\mathcal{U}_{\mathbf{s}}$, we have

$$\begin{aligned}
\langle \beta_i, \beta_m \rangle_{\mathcal{H}_T} &= \left\langle \sum_{j=1}^n \alpha_j k_{\mathbf{s}}(\mathbf{s}_j, \cdot), \sum_{l=1}^n \eta_l k_{\mathbf{s}}(\mathbf{s}_l, \cdot) \right\rangle_{\mathcal{H}_s} \\
&= \sum_{j=1}^n \sum_{l=1}^n \alpha_j \eta_l k_T(\mathbf{y}_j, \mathbf{y}_l) \\
&= \boldsymbol{\alpha}^T \mathbf{K}_{\mathbf{s}} \boldsymbol{\eta} \\
&= \langle \mathbf{L}_{\mathbf{s}}^T \boldsymbol{\alpha}, \mathbf{L}_{\mathbf{s}}^T \boldsymbol{\eta} \rangle_{\mathbb{R}^q}
\end{aligned}$$

86 where $\mathbf{L}_{\mathbf{s}} \in \mathbb{R}^{n \times q}$ is a full column rank matrix and $\mathbf{K}_{\mathbf{s}} = \mathbf{L}_{\mathbf{s}} \mathbf{L}_{\mathbf{s}}^T$ is the Cholesky factorization. As a
87 result, $\beta_i \in \mathcal{U}_{\mathbf{s}}$ would become equivalent to $\mathbf{L}_{\mathbf{s}}^T \boldsymbol{\alpha} \in \mathcal{U}_q$ where \mathcal{U}_q is any complete orthonormal set

88 for \mathbb{R}^q . Using empirical expression for covariance in (B.1) together with equations (7) and (8), we get

$$\begin{aligned}
\text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s}) &:= \sum_{\beta_{\mathbf{s}} \in \mathcal{U}_{\mathbf{s}}} \sum_{j=1}^r \left\{ \frac{1}{n} \sum_{i=1}^n f_j(\mathbf{x}_i) \beta_{\mathbf{s}}(\mathbf{s}_i) - \frac{1}{n^2} \sum_{i=1}^n f_j(\mathbf{x}_i) \sum_{k=1}^n \beta_{\mathbf{s}}(\mathbf{s}_k) \right\}^2 \\
&= \sum_{\mathbf{L}_s^T \boldsymbol{\alpha} \in \mathcal{U}_q} \sum_{j=1}^r \left\{ \frac{1}{n} \boldsymbol{\theta}_j^T \mathbf{K}_x \mathbf{K}_y \boldsymbol{\alpha} - \frac{1}{n^2} \boldsymbol{\theta}_j^T \mathbf{K}_x \mathbf{1}_n \mathbf{1}_n^T \mathbf{K}_y \boldsymbol{\alpha} \right\}^2 \\
&= \sum_{\mathbf{L}_s^T \boldsymbol{\alpha} \in \mathcal{U}_q} \sum_{j=1}^r \left\{ \frac{1}{n} \boldsymbol{\theta}_j^T \mathbf{K}_x \mathbf{H} \mathbf{K}_s \boldsymbol{\alpha} \right\}^2 \\
&= \sum_{\mathbf{L}_s^T \boldsymbol{\alpha} \in \mathcal{U}_q} \sum_{j=1}^r \left\{ \frac{1}{n} \boldsymbol{\theta}_j^T \mathbf{K}_x \mathbf{H} \mathbf{L}_s \mathbf{L}_s^T \boldsymbol{\alpha} \right\}^2 \\
&= \sum_{\boldsymbol{\zeta} \in \mathcal{U}_q} \sum_{j=1}^r \left\{ \frac{1}{n} \boldsymbol{\theta}_j^T \mathbf{K}_x \mathbf{H} \mathbf{L}_s \boldsymbol{\zeta} \right\}^2 \\
&= \sum_{\boldsymbol{\zeta} \in \mathcal{U}_q} \frac{1}{n^2} \|\boldsymbol{\Theta} \mathbf{K}_x \mathbf{H} \mathbf{L}_s \boldsymbol{\zeta}\|_2^2 \\
&= \frac{1}{n^2} \|\boldsymbol{\Theta} \mathbf{K}_x \mathbf{H} \mathbf{L}_s\|_F^2,
\end{aligned}$$

89 where $\mathbf{f}(\mathbf{x}) = \boldsymbol{\Theta}[k_{\mathbf{x}}(\mathbf{x}_1, \mathbf{x}), \dots, k_{\mathbf{x}}(\mathbf{x}_n, \mathbf{x})]^T$ and $\boldsymbol{\Theta} := [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_r]^T$.

90 We now show that the bias of $\text{dep}^{\text{epm}}(\mathbf{z}, \mathbf{s})$ to estimate $\text{dep}(\mathbf{z}, \mathbf{s})$ in (8) is $\mathcal{O}(\frac{1}{n})$. To do this, we split
91 $\text{dep}^{\text{epm}}(\mathbf{z}, \mathbf{s})$ into three terms as

$$\begin{aligned}
\frac{1}{n^2} \|\boldsymbol{\Theta} \mathbf{K}_x \mathbf{H} \mathbf{L}_s\|_F^2 &= \frac{1}{n^2} \text{Tr} \left\{ \boldsymbol{\Theta} \mathbf{K}_x \mathbf{H} \mathbf{K}_s \mathbf{H} \mathbf{K}_x \boldsymbol{\Theta}^T \right\} \\
&= \frac{1}{n^2} \text{Tr} \left\{ \boldsymbol{\Theta} \mathbf{K}_x \left(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right) \mathbf{K}_s \left(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right) \mathbf{K}_x \boldsymbol{\Theta}^T \right\} \\
&= \frac{1}{n^2} \underbrace{\text{Tr} \left\{ \mathbf{K}_x \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{K}_x \mathbf{K}_s \right\}}_{\text{I}} - \frac{2}{n^3} \underbrace{\text{Tr} \left\{ \mathbf{1}^T \mathbf{K}_x \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{K}_x \mathbf{K}_s \mathbf{1} \right\}}_{\text{II}} \\
&\quad + \frac{1}{n^4} \underbrace{\text{Tr} \left\{ \mathbf{1}^T \mathbf{K}_x \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{K}_x \mathbf{1} \mathbf{1}^T \mathbf{K}_s \mathbf{1} \right\}}_{\text{III}} \tag{B.2}
\end{aligned}$$

92 Let \mathcal{C}_p^n denote the set of all p -tuples drawn without repetition from $\{1, \dots, n\}$. Also, let $\boldsymbol{\Theta} =$
93 $[\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_r]^T \in \mathbb{R}^{r \times n}$ and $(\mathbf{A})_{ij}$ denote the element of arbitrary matrix \mathbf{A} at i 'th row and j 'th
94 column. Then, it follows that

95 (I):

$$\begin{aligned}
\mathbb{E}\left[\text{Tr}\left\{\mathbf{K}_x \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{K}_x \mathbf{K}_s\right\}\right] &= \sum_{k=1}^r \mathbb{E}\left[\text{Tr}\left\{\underbrace{\mathbf{K}_x \boldsymbol{\theta}_k}_{\boldsymbol{\alpha}_k} \boldsymbol{\theta}_k^T \mathbf{K}_x \mathbf{K}_s\right\}\right] \\
&= \sum_{k=1}^r \mathbb{E}\left[\text{Tr}\left\{\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T \mathbf{K}_s\right\}\right] \\
&= \sum_{k=1}^r \mathbb{E}\left[\sum_i (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ii} (\mathbf{K}_s)_{ii} + \sum_{(i,j) \in \mathcal{C}_2^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ij} (\mathbf{K}_s)_{ji}\right] \\
&= n \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[f_k^2(\mathbf{x}) k_s(\mathbf{s}, \mathbf{s}) \right] \\
&\quad + \frac{n!}{(n-2)!} \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}, \mathbf{x}', \mathbf{s}'} \left[f_k(\mathbf{x}) f_k(\mathbf{x}') k_s(\mathbf{s}, \mathbf{s}') \right] \\
&= \mathcal{O}(n) + \frac{n!}{(n-2)!} \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}, \mathbf{x}', \mathbf{s}'} \left[f_k(\mathbf{x}) f_k(\mathbf{x}') k_s(\mathbf{s}, \mathbf{s}') \right] \quad (\text{B.3})
\end{aligned}$$

96 where (\mathbf{x}, \mathbf{s}) and $(\mathbf{x}', \mathbf{s}')$ are independently drawn from the joint distribution $\mathbf{p}_{\mathbf{x}\mathbf{s}}$.

97 (II):

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}^T \mathbf{K}_x \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{K}_x \mathbf{K}_s \mathbf{1}\right] &= \sum_{k=1}^r \mathbb{E}\left[\mathbf{1}^T \underbrace{\mathbf{K}_x \boldsymbol{\theta}_k}_{\boldsymbol{\alpha}_k} \boldsymbol{\theta}_k^T \mathbf{K}_x \mathbf{K}_s \mathbf{1}\right] \\
&= \sum_{k=1}^r \mathbb{E}\left[\mathbf{1}^T \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T \mathbf{K}_s \mathbf{1}\right] \\
&= \sum_{k=1}^r \mathbb{E}\left[\sum_{m=1}^n \sum_{i=1}^n \sum_{j=1}^n (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{mi} (\mathbf{K}_s)_{mj}\right] \\
&= \sum_{k=1}^r \mathbb{E}\left[\sum_i (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ii} (\mathbf{K}_s)_{ii} + \sum_{(m,j) \in \mathcal{C}_2^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{mm} (\mathbf{K}_s)_{mj}\right] \\
&\quad + \sum_{k=1}^r \mathbb{E}\left[\sum_{(m,i) \in \mathcal{C}_2^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{mi} (\mathbf{K}_s)_{mm} + \sum_{(m,j) \in \mathcal{C}_2^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{mj} (\mathbf{K}_s)_{mj}\right] \\
&\quad + \sum_{k=1}^r \mathbb{E}\left[\sum_{(m,i,j) \in \mathcal{C}_3^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{mi} (\mathbf{K}_s)_{mj}\right] \\
&= n \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[f_k^2(\mathbf{x}) k_s(\mathbf{s}, \mathbf{s}) \right] + \frac{n!}{(n-2)!} \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}, \mathbf{s}'} \left[f_k^2(\mathbf{x}) k_s(\mathbf{s}, \mathbf{s}') \right] \\
&\quad + \frac{n!}{(n-2)!} \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}, \mathbf{x}'} \left[f_k(\mathbf{x}) f_k(\mathbf{x}') k_s(\mathbf{s}, \mathbf{s}) \right] \\
&\quad + \frac{n!}{(n-2)!} \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}, \mathbf{x}', \mathbf{s}'} \left[f_k(\mathbf{x}) f_k(\mathbf{x}') k_s(\mathbf{s}, \mathbf{s}') \right] \\
&\quad + \frac{n!}{(n-3)!} \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[f_k(\mathbf{x}) \mathbb{E}_{\mathbf{x}'} [f_k(\mathbf{x}')] \mathbb{E}_{\mathbf{s}'} [k_s(\mathbf{s}, \mathbf{s}')] \right] \\
&= \mathcal{O}(n^2) + \frac{n!}{(n-3)!} \sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}} \left[f_k(\mathbf{x}) \mathbb{E}_{\mathbf{x}'} [f_k(\mathbf{x}')] \mathbb{E}_{\mathbf{s}'} [k_s(\mathbf{s}, \mathbf{s}')] \right]. \quad (\text{B.4})
\end{aligned}$$

98 (III):

$$\begin{aligned}
\mathbb{E} \left[\mathbf{1}^T \mathbf{K}_x \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mathbf{K}_x \mathbf{1} \mathbf{1}^T \mathbf{K}_s \mathbf{1} \right] &= \sum_{k=1}^r \mathbb{E} \left[\mathbf{1}^T \underbrace{\mathbf{K}_x \boldsymbol{\theta}_k}_{\boldsymbol{\alpha}_k} \boldsymbol{\theta}_k^T \mathbf{K}_x \mathbf{1} \mathbf{1}^T \mathbf{K}_s \mathbf{1} \right] \\
&= \sum_{k=1}^r \mathbb{E} \left[\mathbf{1}^T \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T \mathbf{1} \mathbf{1}^T \mathbf{K}_s \mathbf{1} \right] \\
&= \sum_{k=1}^r \mathbb{E} \left[\sum_{i,j,m,l} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ij} (\mathbf{K}_s)_{ml} \right] \\
&= \mathcal{O}(n^3) + \sum_{k=1}^r \mathbb{E} \left[\sum_{(i,j,m,l) \in \mathbf{c}_4^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ij} (\mathbf{K}_s)_{ml} \right] \\
&= \mathcal{O}(n^3) + \frac{n!}{(n-4)!} \sum_{k=1}^r \mathbb{E}_x[f_k(\mathbf{x})] E_{\mathbf{x}'}[f_k(\mathbf{x}')] \mathbb{E}_{\mathbf{s}, \mathbf{s}'}[k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}')]
\end{aligned} \tag{B.5}$$

99 Using above calculations together with Lemma 2 lead to

$$\text{dep}(\mathbf{z}, \mathbf{s}) = \mathbb{E}[\text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s})] + \mathcal{O}\left(\frac{1}{n}\right).$$

100 We now obtain the convergence of $\text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s})$. Consider the decomposition in (B.2) together with
101 (B.3), (B.4), and (B.5). Let $\boldsymbol{\alpha}_k := \mathbf{K}_x \boldsymbol{\theta}_k$, then it follows that

$$\begin{aligned}
&\mathbb{P}\left\{\text{dep}(\mathbf{z}, \mathbf{s}) - \text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s}) \geq t\right\} \\
&\leq \mathbb{P}\left\{\sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}, \mathbf{x}', \mathbf{s}'}[f_k(\mathbf{x}) f_k(\mathbf{x}') k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}')] - \frac{(n-2)!}{n!} \sum_{k=1}^r \sum_{(i,j) \in \mathbf{c}_2^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ij} (\mathbf{K}_s)_{ji} + \mathcal{O}\left(\frac{1}{n}\right) \geq at\right\} \\
&+ \mathbb{P}\left\{\sum_{k=1}^r \mathbb{E}_{\mathbf{x}, \mathbf{s}}[f_k(\mathbf{x}) \mathbb{E}_{\mathbf{x}'}[f_k(\mathbf{x}')] \mathbb{E}_{\mathbf{s}, \mathbf{s}'}[k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}')] \right. \\
&\quad \left. - \frac{(n-3)!}{n!} \sum_{k=1}^r \sum_{(i,j,m) \in \mathbf{c}_3^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{mi} (\mathbf{K}_s)_{mj} + \mathcal{O}\left(\frac{1}{n}\right) \geq bt\right\} \\
&+ \mathbb{P}\left\{\sum_{k=1}^r E_x[f_k(\mathbf{x})] E_{\mathbf{x}'}[f_k(\mathbf{x}')] \mathbb{E}_{\mathbf{s}, \mathbf{s}'}[k_{\mathbf{s}}(\mathbf{s}, \mathbf{s}')] \right. \\
&\quad \left. - \frac{(n-4)!}{n!} \sum_{k=1}^r \sum_{(i,j,m,l) \in \mathbf{c}_4^n} (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ij} (\mathbf{K}_s)_{ml} + \mathcal{O}\left(\frac{1}{n}\right) \geq (1-a-b)t\right\},
\end{aligned}$$

102 where $a, b > 0$ and $a + b < 1$. For convenience, we omit the term $\mathcal{O}\left(\frac{1}{n}\right)$ and add it back in the last
103 stage.

104 Define $\boldsymbol{\zeta} := (\mathbf{x}, \mathbf{s})$ and consider the following U-statistics [8]

$$\begin{aligned}
u_1(\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j) &= \frac{(n-2)!}{n!} \sum_{(i,j) \in \mathbf{c}_2^n} \sum_{k=1}^r (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ij} (\mathbf{K}_s)_{ij} \\
u_2(\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j, \boldsymbol{\zeta}_m) &= \frac{(n-3)!}{n!} \sum_{(i,j,m) \in \mathbf{c}_3^n} \sum_{k=1}^r (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{mi} (\mathbf{K}_s)_{mj} \\
u_3(\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j, \boldsymbol{\zeta}_m, \boldsymbol{\zeta}_l) &= \frac{(n-4)!}{n!} \sum_{(i,j,m,l) \in \mathbf{c}_4^n} \sum_{k=1}^r (\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^T)_{ij} (\mathbf{K}_s)_{ml}
\end{aligned}$$

105 Then, from Hoeffding's inequality [8] it follows that

$$\mathbb{P}\left\{\text{dep}(\mathbf{z}, \mathbf{s}) - \text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s}) \geq t\right\} \leq e^{\frac{-2a^2t^2}{2r^2M^2}n} + e^{\frac{-2b^2t^2}{3r^2M^2}n} + e^{\frac{-2(1-a-b)^2t^2}{4r^2M^2}n},$$

106 where we assumed that $k_{\mathbf{s}}(\cdot, \cdot)$ is bounded by one and $f_k^2(\mathbf{x}_i)$ is bounded by M for any $k = 1, \dots, r$
 107 and $i = 1, \dots, n$.

108 Further, if $0.22 \leq a < 1$, it holds that

$$e^{\frac{-2a^2t^2}{2r^2M^2}n} + e^{\frac{-2b^2t^2}{3r^2M^2}n} + e^{\frac{-2(1-a-b)^2t^2}{4r^2M^2}n} \leq 3e^{\frac{-a^2t^2}{r^2M^2}n}.$$

109 Consequently, we have

$$\mathbb{P}\left\{\left|\text{dep}(\mathbf{z}, \mathbf{s}) - \text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s})\right| \geq t\right\} \leq 6e^{\frac{-a^2t^2}{r^2M^2}n}.$$

110 Therefore, with probability at least $1 - \delta$, it holds

$$\left|\text{dep}(\mathbf{z}, \mathbf{s}) - \text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s})\right| \leq \sqrt{\frac{r^2M^2 \log(6/\delta)}{\alpha^2 n}} + \mathcal{O}\left(\frac{1}{n}\right). \quad (\text{B.6})$$

111 □

112 B.4 Proof of Theorem 4

113 **Theorem 4.** A solution to the optimization problem in (11) is the eigenfunctions corresponding to r
 114 largest eigenvalues of the following generalized problem

$$\left((1 - \tau)\Sigma_{\mathbf{y}\mathbf{x}}^* \Sigma_{\mathbf{y}\mathbf{x}} - \tau \Sigma_{\mathbf{s}\mathbf{x}}^* \Sigma_{\mathbf{s}\mathbf{x}}\right)f = \lambda \Sigma_{\mathbf{x}\mathbf{x}}f,$$

115 where $\Sigma_{\mathbf{s}\mathbf{x}}$ and $\Sigma_{\mathbf{y}\mathbf{x}}$ are defined in (7), and $\Sigma_{\mathbf{s}\mathbf{x}}^*$ and $\Sigma_{\mathbf{y}\mathbf{x}}^*$ are the adjoint operators of $\Sigma_{\mathbf{s}\mathbf{x}}$ and $\Sigma_{\mathbf{y}\mathbf{x}}$,
 116 respectively.

117 *Proof.* Consider $\text{dep}(\mathbf{z}, \mathbf{s})$ in (8):

$$\begin{aligned} \text{dep}(\mathbf{z}, \mathbf{s}) &= \sum_{\beta_{\mathbf{s}} \in \mathcal{U}_{\mathbf{s}}} \sum_{j=1}^r h^2(f_j, \beta_{\mathbf{s}}) \\ &= \sum_{j=1}^r \sum_{\beta_{\mathbf{s}} \in \mathcal{U}_{\mathbf{s}}} \langle \beta_{\mathbf{s}}, \Sigma_{\mathbf{s}\mathbf{x}} f_j \rangle_{\mathcal{H}_{\mathbf{s}}}^2 \\ &= \sum_{j=1}^r \|\Sigma_{\mathbf{s}\mathbf{x}} f_j\|_{\mathcal{H}_{\mathbf{s}}}^2, \end{aligned}$$

118 where the last step is due to Parseval's identity for orthonormal basis. Similarly, we have $\text{dep}(\mathbf{z}, \mathbf{y}) =$

119 $\sum_{j=1}^r \|\Sigma_{\mathbf{y}\mathbf{x}} f_j\|_{\mathcal{H}_{\mathbf{y}}}^2$. Recall that $\mathbf{z} = \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_r(\mathbf{x}))$, then, it follows that

$$\begin{aligned} J(\mathbf{f}(\mathbf{x})) &= (1 - \tau) \sum_{j=1}^r \|\Sigma_{\mathbf{y}\mathbf{x}} f_j\|_{\mathcal{H}_{\mathbf{y}}}^2 - \tau \sum_{j=1}^r \|\Sigma_{\mathbf{s}\mathbf{x}} f_j\|_{\mathcal{H}_{\mathbf{s}}}^2 \\ &= (1 - \tau) \sum_{j=1}^r \left\langle \Sigma_{\mathbf{y}\mathbf{x}} f_j, \Sigma_{\mathbf{y}\mathbf{x}} f_j \right\rangle_{\mathcal{H}_{\mathbf{y}}} - \tau \sum_{j=1}^r \left\langle \Sigma_{\mathbf{s}\mathbf{x}} f_j, \Sigma_{\mathbf{s}\mathbf{x}} f_j \right\rangle_{\mathcal{H}_{\mathbf{s}}} \\ &= \sum_{j=1}^r \left\langle f_j, ((1 - \tau)\Sigma_{\mathbf{y}\mathbf{x}}^* \Sigma_{\mathbf{y}\mathbf{x}} - \tau \Sigma_{\mathbf{s}\mathbf{x}}^* \Sigma_{\mathbf{s}\mathbf{x}}) f_j \right\rangle_{\mathcal{H}_{\mathbf{x}}}, \end{aligned}$$

120 where Σ^* is the adjoint operator of Σ . Further, note that $\text{Cov}_{\mathbf{x}}(f_i(\mathbf{x}), f_j(\mathbf{x}))$ is equal to

121 $\langle f_i, \Sigma_{\mathbf{x}\mathbf{x}} f_j \rangle_{\mathcal{H}_{\mathbf{x}}}$. As a result, the optimization problem in (11) can be restated as

$$\sup_{\langle f_i, (\Sigma_{\mathbf{x}\mathbf{x}} + \gamma I_{\mathbf{x}}) f_k \rangle_{\mathcal{H}_{\mathbf{x}}} = \delta_{i,k}} \sum_{j=1}^r \left\langle f_j, ((1 - \tau)\Sigma_{\mathbf{y}\mathbf{x}}^* \Sigma_{\mathbf{y}\mathbf{x}} - \tau \Sigma_{\mathbf{s}\mathbf{x}}^* \Sigma_{\mathbf{s}\mathbf{x}}) f_j \right\rangle_{\mathcal{H}_{\mathbf{x}}}, \quad 1 \leq i, k \leq r$$

where I_x denotes identity operator from \mathcal{H}_x to \mathcal{H}_x . This optimization problem is known as generalized Rayleigh quotient [9] and a possible solution to it is given by the eigenfunctions corresponding to the r largest eigenvalues of the following generalized problem

$$\left((1-\tau)\Sigma_{xy}\Sigma_{yx} - \tau\Sigma_{xs}\Sigma_{sx}\right)f = \lambda\left(\Sigma_{xx} + \gamma I_x\right)f.$$

□

B.5 Proofs of Theorem 5 and Corollary 5.1

Theorem 5. Consider the Cholesky factorization $K_x = L_x L_x^T$, where L_x is a full column-rank matrix. A solution to (13) is

$$\mathbf{f}^{\text{opt}} = \Theta^{\text{opt}} \left[k_x(x_1, \cdot), \dots, k_x(x_n, \cdot) \right]^T$$

where $\Theta^{\text{opt}} = U^T (L_x)^\dagger$ and the columns of U are eigenvectors corresponding to r largest eigenvalues, $\lambda_1, \dots, \lambda_r$ of generalized problem

$$\left(L_x^T((1-\tau)\tilde{K}_y - \tau\tilde{K}_s)L_x\right)u = \lambda\left(L_x^T H L_x + n\gamma I\right)u$$

where γ is the regularization parameter from (10) and the supremum value of (13) is $\sum_{j=1}^r \lambda_j$.

Proof. Consider the Cholesky factorization $K_x = L_x L_x^T$ where L_x is a full column-rank matrix. Using representer theorem, disentanglement condition in (10), can be expressed as

$$\begin{aligned} & \text{Cov}(f_i(x), f_j(x)) + \gamma \langle f_i, f_j \rangle_{\mathcal{H}_x} \\ &= \frac{1}{n} \sum_{k=1}^n f_i(x_k) f_j(x_k) - \frac{1}{n^2} \sum_{k=1}^n f_i(x_k) \sum_{m=1}^n f_j(x_m) + \gamma \langle f_i, f_j \rangle_{\mathcal{H}_x} \\ &= \frac{1}{n} \sum_{k=1}^n \sum_{t=1}^n K_x(x_k, x_t) \theta_{it} \sum_{m=1}^n K_x(x_k, x_m) \theta_{jm} - \frac{1}{n^2} \theta_i^T K_x \mathbf{1}_n \mathbf{1}_n^T K_x \theta_j + \gamma \langle f_i, f_j \rangle_{\mathcal{H}_x} \\ &= \frac{1}{n} (K_x \theta_i)^T (K_x \theta_j) - \frac{1}{n^2} \theta_i^T K_x \mathbf{1}_n \mathbf{1}_n^T K_x \theta_j + \gamma \left\langle \sum_{k=1}^n \theta_{ik} k_x(\cdot, x_k), \sum_{t=1}^n \theta_{jt} k_x(\cdot, x_t) \right\rangle_{\mathcal{H}_x} \\ &= \frac{1}{n} \theta_i^T K_x H K_x \theta_j + \gamma \theta_i^T K_x \theta_j \\ &= \frac{1}{n} \theta_i^T L_x \left(L_x^T H L_x + n\gamma I \right) L_x^T \theta_j \\ &= \delta_{i,j}. \end{aligned}$$

As a result, $\mathbf{f} \in \mathcal{A}_r$ is equivalent to

$$\frac{1}{n} \Theta L_x \underbrace{\left(L_x^T H L_x + n\gamma I \right)}_{:=C} L_x^T \Theta^T = I_r,$$

where $\Theta := [\theta_1, \dots, \theta_r]^T \in \mathbb{R}^{r \times n}$.

Let $V = \frac{1}{\sqrt{n}} L_x^T \Theta^T$ and consider the optimization problem in (13):

$$\begin{aligned} & \sup_{\mathbf{f} \in \mathcal{A}_r} \left\{ (1-\tau) \text{dep}^{\text{emp}}(\mathbf{f}(x), \mathbf{y}) - \tau \text{dep}^{\text{emp}}(\mathbf{f}(x), \mathbf{s}) \right\} \\ &= \sup_{\mathbf{f} \in \mathcal{A}_r} \frac{1}{n^2} \left\{ (1-\tau) \|\Theta K_x H L_y\|_F^2 - \tau \|\Theta K_x H L_s\|_F^2 \right\} \\ &= \sup_{\mathbf{f} \in \mathcal{A}_r} \frac{1}{n^2} \left\{ (1-\tau) \text{Tr}\{\Theta K_x H K_y H K_x \Theta^T\} - \tau \text{Tr}\{\Theta K_x H K_s H K_x \Theta^T\} \right\} \\ &= \max_{V^T C V = I_r} \frac{1}{n^2} \text{Tr}\{\Theta L_x B L_x^T \Theta^T\} \\ &= \max_{V^T C V = I_r} \frac{1}{n} \text{Tr}\{V^T B V\} \end{aligned} \tag{B.7}$$

137 where the second step is due to (9) and

$$\begin{aligned} \mathbf{B} &:= \mathbf{L}_x^T \left((1 - \tau) \mathbf{H} \mathbf{K}_y \mathbf{H} - \tau \mathbf{H} \mathbf{K}_s \mathbf{H} \right) \mathbf{L}_x \\ &= \mathbf{L}_x^T \left((1 - \tau) \tilde{\mathbf{K}}_y - \tau \tilde{\mathbf{K}}_s \right) \mathbf{L}_x. \end{aligned}$$

138 It is shown in [10] that an² optimizer of (B.7) is any matrix \mathbf{U} whose columns are eigenvectors
139 corresponding to r largest eigenvalues of generalized problem

$$\mathbf{B}\mathbf{u} = \lambda \mathbf{C}\mathbf{u}$$

140 and the maximum value is the summation of r largest eigenvalues. Once \mathbf{U} is determined, then, any
141 Θ in which $\mathbf{L}_x^T \Theta^T = \sqrt{n} \mathbf{U}$ is optimal Θ (denoted by Θ^{opt}). Note that Θ^{opt} is not unique and has a
142 general form of

$$\Theta^T = \sqrt{n} (\mathbf{L}_x^T)^\dagger \mathbf{U} + \Lambda_0, \quad \mathcal{R}(\Lambda_0) \subseteq \mathcal{N}(\mathbf{L}_x^T).$$

143 However, setting Λ_0 to zero would lead to minimum norm for Θ . Therefore, we opt $\Theta^{\text{opt}} =$
144 $\mathbf{U}^T (\mathbf{L}_x)^\dagger$, where ignoring the constant multiplier \sqrt{n} does not change the generalized eigenvalue
145 problem in (B.8). \square

146 **Corollary 5.1.** *Embedding Dimensionality:* A useful corollary of Theorem 5 is optimal embedding
147 dimensionality:

$$\arg \sup_r \left\{ \sup_{\mathbf{f} \in \mathcal{A}_r} \left\{ J^{\text{emp}}(\mathbf{f}(\mathbf{x})) := (1 - \tau) \text{dep}^{\text{emp}}(\mathbf{f}(\mathbf{x}), \mathbf{y}) - \tau \text{dep}^{\text{emp}}(\mathbf{f}(\mathbf{x}), \mathbf{s}) \right\} \right\},$$

148 which is the number of positive eigenvalues of the generalized eigenvalue problem in (14).

149 *Proof.* From the proof of Theorem 5, we know that

$$\sup_{\mathbf{f} \in \mathcal{A}_r} \left\{ (1 - \tau) \text{dep}^{\text{emp}}(\mathbf{f}(\mathbf{x}), \mathbf{y}) - \tau \text{dep}^{\text{emp}}(\mathbf{f}(\mathbf{x}), \mathbf{s}) \right\} = \sum_{j=1}^r \lambda_j,$$

150 where $\{\lambda_1, \dots, \lambda_n\}$ are eigenvalues of the generalized problem in (B.8) in decreasing order. It
151 follows immediately that

$$\arg \sup_r \left\{ \sum_{j=1}^r \lambda_j \right\} = \text{number of positive elements of } \{\lambda_1, \dots, \lambda_n\}.$$

152 \square

153 B.6 Proof of Theorem 6

154 **Theorem 1.** Assume that $k_s(\cdot, \cdot)$ and $k_y(\cdot, \cdot)$ are bounded by one and $f_k^2(\mathbf{x}_i)$ is bounded by M for
155 any $k = 1, \dots, r$ and $i = 1, \dots, n$ for which $\mathbf{f} = (f_1, \dots, f_r) \in \mathcal{A}_r$. For any $n > 1$ and $0 < \delta < 1$,
156 with probability at least $1 - \delta$, we have

$$\left| \sup_{\mathbf{f} \in \mathcal{A}_r} J(\mathbf{f}(\mathbf{x})) - \sup_{\mathbf{f} \in \mathcal{A}_r} J^{\text{emp}}(\mathbf{f}(\mathbf{x})) \right| \leq rM \sqrt{\frac{\log(6/\delta)}{a^2 n}} + \mathcal{O}\left(\frac{1}{n}\right),$$

157 where $0.22 \leq a \leq 1$ is a constant.

158 *Proof.* Recall that in the proof of Lemma 3, we have shown that with probability at least $1 - \delta$, the
159 following holds,

$$\left| \text{dep}(\mathbf{z}, \mathbf{s}) - \text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s}) \right| \leq \sqrt{\frac{r^2 M^2 \log(6/\sigma)}{\alpha^2 n}} + \mathcal{O}\left(\frac{1}{n}\right).$$

²Optimal \mathbf{V} is not unique.

160 Using the same reasoning, with probability at least $1 - \delta$, we have

$$\left| \text{dep}(\mathbf{z}, \mathbf{y}) - \text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{y}) \right| \leq \sqrt{\frac{r^2 M^2 \log(6/\sigma)}{\alpha^2 n}} + \mathcal{O}\left(\frac{1}{n}\right).$$

161 Since $J(\mathbf{f}(\mathbf{x})) = (1 - \tau) \text{dep}(\mathbf{z}, \mathbf{y}) - \tau \text{dep}(\mathbf{z}, \mathbf{s})$ and $J^{\text{emp}}(\mathbf{f}(\mathbf{x})) := (1 - \tau) \text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{y}) -$
 162 $\tau \text{dep}^{\text{emp}}(\mathbf{z}, \mathbf{s})$, it follows that with probability at least $1 - \delta$,

$$\left| J(\mathbf{f}(\mathbf{x})) - J^{\text{emp}}(\mathbf{f}(\mathbf{x})) \right| \leq rM \sqrt{\frac{\log(6/\sigma)}{\alpha^2 n}} + \mathcal{O}\left(\frac{1}{n}\right).$$

163 We complete the proof by noting that

$$\left| \sup_{\mathbf{f} \in \mathcal{A}_r} J(\mathbf{f}(\mathbf{x})) - \sup_{\mathbf{f} \in \mathcal{A}_r} J^{\text{emp}}(\mathbf{f}(\mathbf{x})) \right| \leq \sup_{\mathbf{f} \in \mathcal{A}_r} \left| J(\mathbf{f}(\mathbf{x})) - J^{\text{emp}}(\mathbf{f}(\mathbf{x})) \right|.$$

164

□

165 References

- 166 [1] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint*
 167 *arXiv:1711.05101*, 2017.
- 168 [2] B. Sadeghi, R. Yu, and V. Boddeti, “On the global optima of kernelized adversarial representation
 169 learning,” in *IEEE International Conference on Computer Vision*, pp. 7971–7979, 2019.
- 170 [3] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig, “Controllable invariance through adversarial
 171 feature learning,” in *Advances in Neural Information Processing Systems*, pp. 585–596, 2017.
- 172 [4] S. Verma and J. Rubin, “Fairness definitions explained,” in *IEEE/ACM International Workshop*
 173 *on Software Fairness (FairWare)*, pp. 1–7, IEEE, 2018.
- 174 [5] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn, “R\`enyi fair inference,” *arXiv*
 175 *preprint arXiv:1906.12005*, 2019.
- 176 [6] J. Jacod and P. Protter, *Probability essentials*. Springer Science & Business Media, 2012.
- 177 [7] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, “Measuring statistical dependence
 178 with hilbert-schmidt norms,” in *International Conference on Algorithmic Learning Theory*,
 179 pp. 63–77, Springer, 2005.
- 180 [8] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” in *The Collected*
 181 *Works of Wassily Hoeffding*, pp. 409–426, Springer, 1994.
- 182 [9] R. L. Strawderman, “The symmetric eigenvalue problem (classics in applied mathematics,
 183 number 20),” *Journal of the American Statistical Association*, vol. 94, no. 446, p. 657, 1999.
- 184 [10] E. Kokiopoulou, J. Chen, and Y. Saad, “Trace optimization and eigenproblems in dimension
 185 reduction methods,” *Numerical Linear Algebra with Applications*, vol. 18, no. 3, pp. 565–602,
 186 2011.