Supplementary Document
Maximal Correlation-Based Post-Nonlinear Learning for Bivariate Causal Discovery

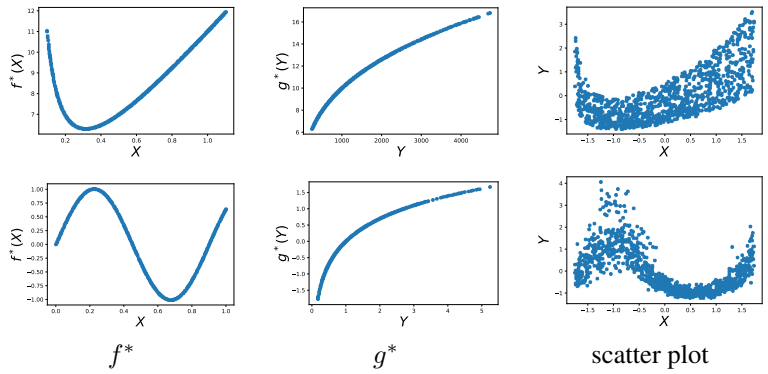
A Discussion on MI minimization

It was shown that minimizing the MI in (5) is equivalent to maximizing $\mathbb{E}\log p\left(r_{(\to)}\right) + \mathbb{E}\log\left|\frac{d}{dy}g_{(\to)}\left(y\right)\right|$ (Zhang & Hyvärinen, 2009), where p is the assumed noise density. We find this objective interpretable, since the first term, $\mathbb{E}\log p\left(r_{(\to)}\right)$, can be understood as the data fitting term, and the second term, $\mathbb{E}\log\left|\frac{d}{dy}g_{(\to)}\left(y\right)\right|$, can be understood from an information-geometric perspective (Daniušis et al., 2010). However, such equivalent form requires a known noise distribution to calculate the log-likelihood. Some works (Ma et al., 2020; Uemura & Shimizu, 2020) have been proposed to avoid this difficulty by using HSIC instead of MI.

B Experiments on Minimizing HSIC

In this section, we show the PNL model learning result by minimizing (12). We generated two synthetic datasets from PNL model, $Y = f_2\left(f_1(X) + \epsilon\right)$, and each contains 1000 data samples. The data generation mechanisms are as follows (see Figure 4),

- Syn-1: $f_1(X) = X^{-1} + 10X, f_2(Z) = Z^3, X \sim U(0.1, 1.1), \epsilon \sim U(0, 5)$,
- Syn-2: $f_1(X) = \sin(7X), f_2(Z) = \exp(Z), X \sim U(0, 1), \epsilon \sim N(0, 0.3^2)$.



f^* g^* scatter plot

Figure 4: The ground truth transformations of f^* and g^* of Syn-1 (top) and Syn-2 (bottom).

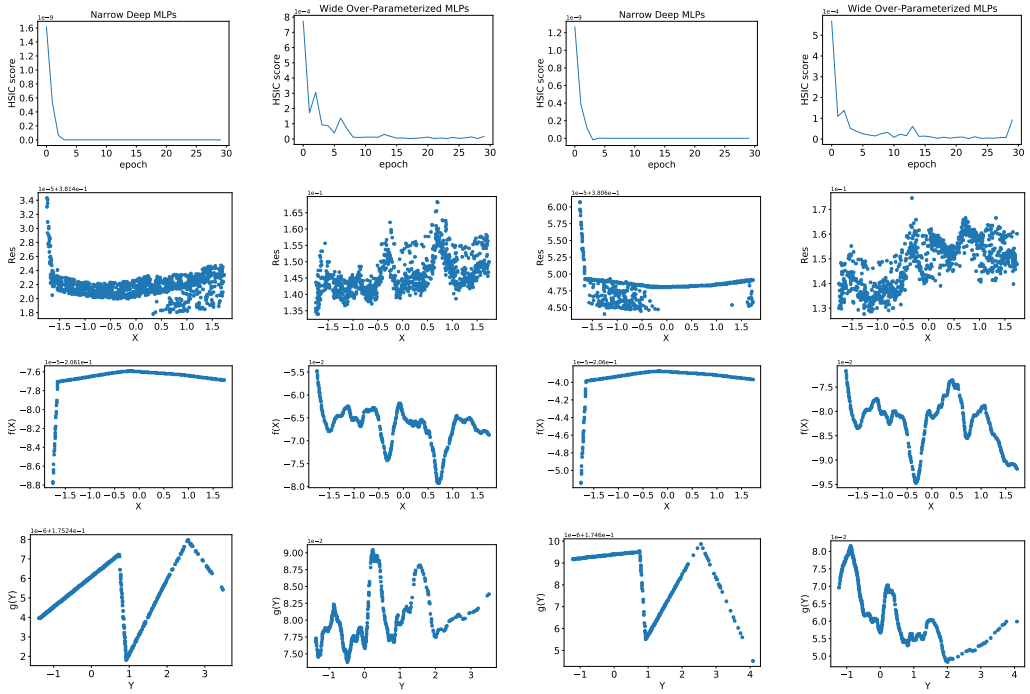
We build different MLPs with the following configurations.

- Narrow deep MLP: the input and output are both one-dimensional; there are 9 hidden layers, each with 5 neurons. The activation function is `Leaky-ReLU`.
- Wide over-parameterized MLP: the input and output are both one-dimensional; there is only one single hidden layer with 9000 neurons. The activation function is `Leaky-ReLU`.

We use the default initialization method in PyTorch (Paszke et al., 2019), and make sure the exact same initial weights for narrow/wide MLPs are used (i.e., the initializations for different datasets are the same).

Optimization Setup: We set the batch size to be 32. We use `Adam` (Kingma & Ba, 2015) for the optimization (the learning rates are 10^{-3} and 10^{-6} for narrow deep and wide over-parameterized MLPs, respectively, while all other parameters are set by default).

We report the learning results in Figure 5. The learned transformations (see row 3 and row 4 in Figure 5) deviates far away from the underlying functions, and are quite similar across datasets. The



Syn-1, narrow MLPs | Syn-1, wide MLPs | Syn-2, narrow MLPs | Syn-2, wide MLPs

Figure 5: Visualization of the learned nonlinearities (trained solely with HSIC, under different datasets and MLP configurations). From top to bottom, the **convergence results**, **residual plot**, **learned** f, **learned** g, are plotted. Each column shows one specific configuration. None of them learns meaningful nonlinearities, and the learned transformations are quite similar across datasets.

possible reason is that, the solutions were started from the same initialization and trapped at the local minima near the initializations.

To verify whether such HSIC-based PNL learning algorithm is stable for causal discovery, we further evaluate the AbPNL on the following dataset. We build 100 data pairs with different random seeds, following the same mechanism, Syn-1, and each contains 1000 data samples. And we applied the AbPNL (Uemura & Shimizu, 2020) with different initializations on each of those data pairs. The results in Table 3 show that the causal discovery stableness for AbPNL is not satisfactory.

Table 3: Comparison of bivaraite causal discovery AUC on 100 realizations of Syn-1

Dataset	ANM	CDS	IGCI	RECI	CDCI	AbPNL	ACE	MC-PNL
Syn-1	0.495	1	0.528	1	1	0.281	1	1

C Synthetic Datasets for Independence Test

In this section, we describe the synthetic data generation from PNL model for the independent test. The data were generated from the following model, $Y = f_2\left(f_1(X) + \epsilon\right), X \sim \text{GMM}, \epsilon \sim N(0, \sigma_\epsilon^2)$, where f_1, f_2 are randomly initialized monotonic neural networks (Wehenkel & Louppe, 2019) with 3 layers and 100 integration steps, and each layer contains 100 units. The cause term X is sampled from a Gaussian mixture model as described in Lopez-Paz et al. (2017). The datasets were configured with various noise levels and sample sizes. There are three different injected noise levels, $\sigma_\epsilon \in \{0.1, 1, 10\}$, and three different sample sizes, $N \in \{1000, 2000, 5000\}$. And under each configuration, we generated 100 data pairs for evaluating the independence test accuracy.

D A UNIVERSAL VIEW OF DEPENDENCE MEASURES

Actually the discussed dependence measures in Section 3.2 are all closely related to the *mean squared contingency* introduced by (Rényi, 1959) and rediscovered due to its squared version called *squared-loss mutual information* (SMI) (Suzuki et al., 2009),

$$\mathrm{SMI} := \iint p(x)p(y)\left(\frac{p(x,y)}{p(x)p(y)} - 1\right)^2 \mathrm{d}x\mathrm{d}y = \iint \frac{p(x,y)}{p(x)p(y)} p(x,y)\mathrm{d}x\mathrm{d}y - 1. \tag{16}$$

When the density ratio $\mathrm{DR}(x,y) := \frac{p(x,y)}{p(x)p(y)}$ is constant 1 (namely X and Y are independent), the SMI should be zero. To estimate the SMI, one can first approximate $\mathrm{DR}(x,y)$ by a surrogate function $\mathrm{DR}_{\boldsymbol{\theta}}(x,y)$ parameterized by $\boldsymbol{\theta}$. The optimal parameter $\hat{\boldsymbol{\theta}}$ can be obtained via minimizing the following squared-error loss J^{DR},

$$\begin{aligned} J^{\mathrm{DR}}(\boldsymbol{\theta}) &:= \iint \left(\mathrm{DR}_{\boldsymbol{\theta}}(x,y) - \mathrm{DR}(x,y)\right)^2 p(x)p(y)\mathrm{d}x\mathrm{d}y \\ &= \iint \mathrm{DR}_{\boldsymbol{\theta}}(x,y)^2 p(x)p(y)\mathrm{d}x\mathrm{d}y - 2\iint \mathrm{DR}_{\boldsymbol{\theta}}(x,y)p(x,y)\mathrm{d}x\mathrm{d}y + \text{Const.} \end{aligned} \tag{17}$$

Then the empirical SMI can be calculated as, $\widehat{\mathrm{SMI}} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{DR}_{\hat{\boldsymbol{\theta}}}(x_j, y_j) - 1$.

We show that, with different parameterizations of the density ratio, the resulting SMI will be equivalent to different dependence measures, see Table 4.

Table 4: Connections between DR parameterization and dependence measure

Density ratio surrogate function $\mathrm{DR}_{\boldsymbol{\theta}}(x,y)$	Corresponding dependence measure
$\mathrm{DR}_{\boldsymbol{\theta}}(x,y) = 1 + \sum_{i=1}^{n} \theta_i K(x,x_i) L(y,y_i)$	variant of LSMI (Sugiyama & Yamada, 2012)
$\mathrm{DR}_{\boldsymbol{\theta}}(x,y) = 1 + \sum_{i=1}^{n} \frac{1}{n} K(x,x_i) L(y,y_i)$	HSIC (Gretton et al., 2005)
$\mathrm{DR}_{\boldsymbol{\theta}}(x,y) = 1 + \sum_{i=1}^{m} f_i(x)g_i(y)$	m-mode HGR correlation (Wang et al., 2019)
$\mathrm{DR}_{\boldsymbol{\theta}}(x,y) = 1 + f(x)g(y)$ [1]	HGR correlation (Rényi, 1959)

[1] When f, g are the linear combinations of random features, $f(x) = \boldsymbol{\alpha}^T\boldsymbol{\phi}(x), g(y) = \boldsymbol{\beta}^T\boldsymbol{\psi}(y)$, the corresponding dependence measure will be RDC (López-Paz et al., 2013),

Sugiyama & Yamada (2012) proposed to approximate the density ratio by $\mathrm{DR}_{\hat{\boldsymbol{\theta}}}(x,y) = \sum_{i=1}^{n} \hat{\theta}_i K(x,x_i) L(y,y_i)$, where $\hat{\boldsymbol{\theta}}$ has a closed-form solution via minimizing (17). After then, they approximated the SMI using the empirical average of Equation (16), $\frac{1}{n}\sum_{j=1}^{n} \mathrm{DR}_{\hat{\boldsymbol{\theta}}}(x_j,y_j) - 1 = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} \hat{\theta}_i K(x,x_i) L(y,y_i) - 1$. It is shown that, the first term is actually the empirical HSIC, when $\{\hat{\theta}_i\}_{i=1}^{n} = \frac{1}{n}$. We argue that there is a flaw above, as when X and Y are independent, both the SMI and HSIC score should be zero. A simple modification is to model the density ratio by $\mathrm{DR}_{\boldsymbol{\theta}}(x,y) = 1 + \sum_{i=1}^{n} \theta_i K(x,x_i) L(y,y_i)$. The constant 1 here is to exclude all the independence terms, and the rest ones should model the dependency only. This modification will not hurt the quadratic form of $J^{\mathrm{DR}}(\boldsymbol{\theta})$, and maintains good interpretation. The SMI reduced to HSIC score, when $\{\theta_i\}_{i=1}^{n} = \frac{1}{n}$,

We extend this idea to approximate the density ratio by $\mathrm{DR}_{\boldsymbol{\theta}}(x,y) = 1 + f(x)g(y)$, where f, g are zero mean and unit variance functions parameterized by $\boldsymbol{\theta}$, the resulting SMI will be equal to the HGR maximal correlation. Similarly, the constant 1 will capture the independence part, and $f(x)g(y)$ will capture the dependencies.

Proposition 1. *The density ratio estimation problem (17) is equivalent to the maximal HGR correlation problem (7), when the density ratio is modeled in the form of* $\mathrm{DR}_{\boldsymbol{\theta}}(x,y) = 1 + f(x)g(y)$*, and* f, g *are restricted to zero mean and unit variance functions.*

Proof. We substitute $\mathrm{DR}_{\hat{\boldsymbol{\theta}}}(x, y)$ into Equation (17),

$$\begin{aligned} J^{\mathrm{DR}}(f, g) &= \iint (1 + f(x)g(y))^2 p(x)p(y)\mathrm{d}x\mathrm{d}y - 2\iint (1 + f(x)g(y))p(x, y)\mathrm{d}x\mathrm{d}y + \text{Const.} \\ &= 1 + 2\mathbb{E}(f(X))\mathbb{E}(g(Y)) + \mathrm{var}(f(X))\mathrm{var}(g(Y)) - 2 - 2\mathbb{E}(f(X)g(Y)) + \text{Const.} \end{aligned}$$

Then it is not hard to see, $\min_{f,g} J^{\mathrm{DR}}(f, g)$, subject to $\mathbb{E}(f) = \mathbb{E}(g) = 0, \mathrm{var}(f) = \mathrm{var}(g) = 1$, is equivalent to the maximal HGR correlation problem (7) . □

Proposition 2. *The density ratio estimation problem (17) is equivalent to the Soft-HGR problem (9), when the density ratio is modeled in the form of* $\mathrm{DR}_{\boldsymbol{\theta}}(x, y) = 1 + f(x)g(y)$*, and* f, g *are restricted to zero mean functions.*

We further note that the above density ratio estimation can be regard as a truncated singular value decomposition $\mathrm{DR}_{\hat{\boldsymbol{\theta}}}(x, y) = 1 + \sum_{i=1}^{m} f_i(x)g_i(y)$, where $m = 1$. When letting $m > 1$ and imposing zero mean and unit variance constraints on all f_i and g_i, the corresponding J^{DR} minimization problem is equivalent to solving the m-mode HGR maximal correlation (Wang et al., 2019; Lee, 2021).

Definition 3 (m-mode HGR maximal correlation)**.** *Given* $1 \leq m \leq \min\{|\mathcal{X}|, |\mathcal{Y}|\}$*, the* m*-mode maximal correlation problem for random variables* $X \in \mathcal{X}, Y \in \mathcal{Y}$ *is,*

$$(\mathbf{f}^*, \mathbf{g}^*) \triangleq \underset{\substack{\mathbf{f}:\mathcal{X}\to\mathbb{R}^m, \mathbf{g}:\mathcal{Y}\to\mathbb{R}^m \\ \mathbb{E}[\mathbf{f}(X)]=\mathbb{E}[\mathbf{g}(Y)]=\mathbf{0}, \\ \mathbb{E}\left[\mathbf{f}(X)\mathbf{f}^{\mathrm{T}}(X)\right]=\mathbb{E}\left[\mathbf{g}(Y)\mathbf{g}^{\mathrm{T}}(Y)\right]=\mathbf{I}}}{\arg\max} \mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(X)\mathbf{g}(Y)\right], \tag{18}$$

where $\mathbf{f} = [f_1, f_2, \ldots, f_m]^{\mathrm{T}}, \mathbf{g} = [g_1, g_2, \ldots, g_m]^{\mathrm{T}}$ *are referred as the maximal correlation functions.*

E Random Feature Generation

We generate the random features as described in López-Paz et al. (2013). The generation process has the following two steps: *copula transformation* (optional) and *random nonlinear projection*.

Step 1. *Copula transformation.* We first estimate the empirical cumulative distribution of both X and Y by,

$$P_n^X(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \leq x), P_n^Y(y) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i \leq y).$$

Then we can apply the empirical copula transformation to data samples $\{(x_i, y_i)\}_{i=1}^n$, $u_i^X = P_n^X(x_i)$ and $u_i^Y = P_n^Y(y_i)$, where the marginals U^X and U^Y follow uniform distribution $U(0, 1)$.

Step 2. *Random nonlinear projection.* We design a k-dimensional random feature vector $\boldsymbol{\phi}(x) = [\sin(w_1 x + b_1), \cdots, \sin(w_k x + b_k)]^T$, where $w_i, b_i \sim N(0, s^2)$. The random feature matrix $\Phi \in \mathbb{R}^{k\times n}$ is stacked as,

$$\Phi(\mathbf{x}; k, s) := \begin{pmatrix} \sin(w_1 x_1 + b_1) & \cdots & \sin(w_1 x_n + b_1) \\ \vdots & \vdots & \vdots \\ \sin(w_k x_1 + b_k) & \cdots & \sin(w_k x_n + b_k) \end{pmatrix}.$$

One can replace the x_i here by u_i^X from the first step to form the random feature matrix. Similar procedures can be applied to $\mathbf{y}$ as well to generate Ψ. The number of random Fourier features k is user-defined, which is typically chosen from a few tens to a few thousands (Rahimi & Recht, 2008; Theodoridis, 2015). In our experiments, we set $k = 30$ and $s = 2$.

F ON THE OPTIMIZATION OF PROBLEM (14)

F.1 SUBPROBLEM: EQUALITY CONSTRAINED QUADRATIC PROGRAMMING

To simplify the notation, we rewrite the sub-problem into the following form,

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) := \frac{1}{2}x^T Ax - b^T x, \\ s.t. & v^T x = c. \end{array} \tag{19}$$

With the KKT conditions, one can find the unique optimal solution x^* by solving the following linear system,

$$\underbrace{\begin{pmatrix} A & v \\ v^T & 0 \end{pmatrix}}_{=:\mathrm{KKT}} \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}, \tag{20}$$

when the KKT matrix is non-singular. In our setting, we can choose Φ and Ψ properly to make $\Phi\Phi^T$ and $\Psi\Psi^T$ positive definite, or add a small positive definite perturbation matrix ϵI, such that the unique optimum would be obtained. Besides, the sub-problem is of smaller size and easy to solve.

F.2 LANDSCAPE STUDY WITH HESSIAN

To simplify the notation, we rewrite

$$J(\boldsymbol{\alpha}, \boldsymbol{\beta}; A, B, C, D, E) = \boldsymbol{\alpha}^T A \boldsymbol{\alpha} \boldsymbol{\beta}^T B \boldsymbol{\beta} - \boldsymbol{\alpha}^T C \boldsymbol{\beta} + \boldsymbol{\alpha}^T D \boldsymbol{\alpha} + \boldsymbol{\beta}^T E \boldsymbol{\beta}, \tag{21}$$

where,

$$\begin{array}{l} A = \frac{1}{2n^2}\Phi\Phi^T, \\ B = \Psi\Psi^T, \\ C = \frac{1}{n}\Phi\Psi^T + \frac{\lambda}{(n-1)^2}\Phi H K_{\mathbf{xx}} H \Psi^T, \\ D = \frac{\lambda}{(n-1)^2}\Phi H K_{\mathbf{xx}} H \Phi^T, \\ E = \frac{\lambda}{(n-1)^2}\Psi H K_{\mathbf{xx}} H \Psi^T. \end{array} \tag{22}$$

And the corresponding Hessian is

$$\nabla^2 J(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{pmatrix} 2A\boldsymbol{\beta}^T B \boldsymbol{\beta} + 2D & A\boldsymbol{\alpha}\boldsymbol{\beta}^T B - C \\ B^T \boldsymbol{\beta}\boldsymbol{\alpha}^T A - C^T & 2B\boldsymbol{\alpha}^T A \boldsymbol{\alpha} + 2E \end{pmatrix}. \tag{23}$$

Now we are able to verify the property of the critical points via checking their Hessians numerically.

One obvious critical point is the all zero vector $\mathbf{0}$. From our experiments, the Hessian at $\mathbf{0}$ is mostly indefinite, as long as the convex regularization term λ is not too huge, which means $\mathbf{0}$ is a saddle point. In practice, the algorithm rarely converges to $\mathbf{0}$.

G FINE-TUNE WITH BANDED LOSS / UNIVERSAL HSIC

In the PNL model, the injected noise are assumed to be independently and identically distributed. Thus, the residual plot should forms a "horizontal band". We design a **banded residual loss** to fine-tune the models as follows. The data samples are separated into b bins $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^b$ according to the ordering of X, and we expect the residuals in those bins $\mathrm{Res}_i = f(\mathbf{x}^{(i)}) - g(\mathbf{y}^{(i)})$ to have the same distribution, see Figure 6. To this end, we adopt the empirical maximum mean discrepancy (MMD) (Gretton et al., 2012) as a measure of distribution discrepancy. The **banded residual loss** is defined as $\mathrm{band}^{(\mathrm{MMD})} := \sum_{i=1}^b \widehat{\mathrm{MMD}}(\mathrm{Res}_i, \mathrm{Res}_{all})$, where $\mathrm{Res}_{all} = f(\mathbf{x}) - g(\mathbf{y})$. Then we append this μ-penalized banded loss to Problem (14) as,

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad J(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \mu \cdot \mathrm{band}^{(\mathrm{MMD})}, \quad \text{s.t.} \quad \boldsymbol{\alpha}^T \Phi \mathbf{1} = \boldsymbol{\beta}^T \Psi \mathbf{1} = 0. \tag{24}$$

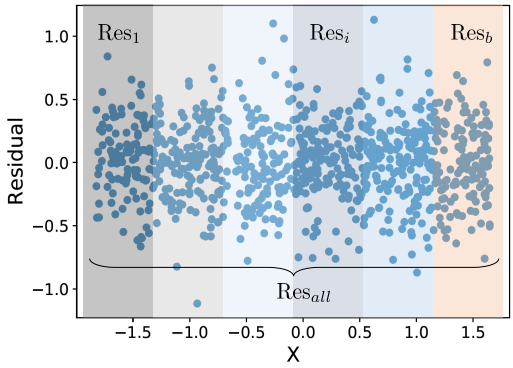
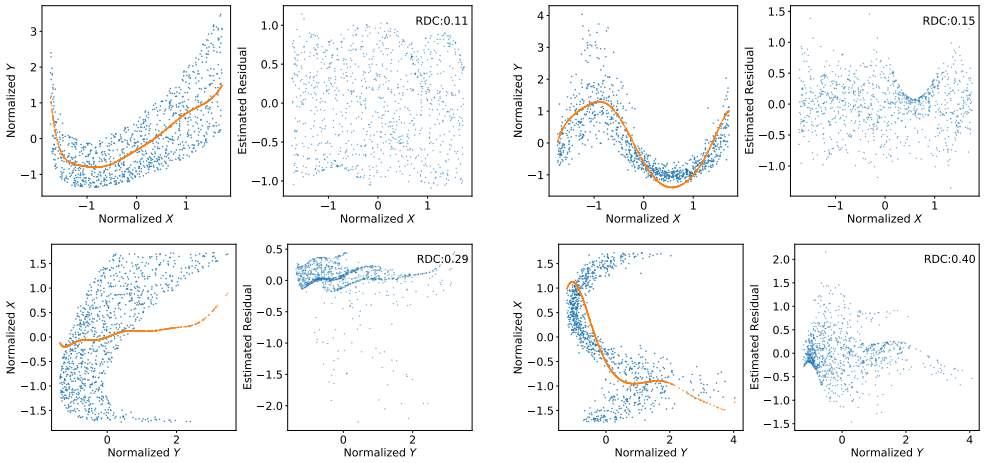


Figure 6: The construction of banded residual loss.

The above banded residual loss involves MMD, which is highly non-convex and brings difficulties to the optimization. We used the projected gradient descent with momentum to optimize the loss function. The residual plot shows a band shape, see top row in Figure 7.

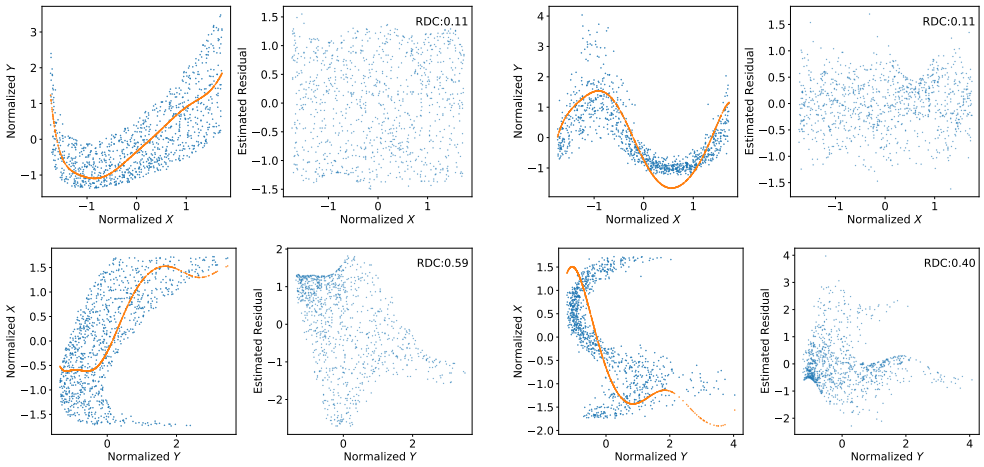


(a) Syn-1 (b) Syn-2

Figure 7: Fine-tuning with the banded residual loss.

We also show the results of fine-tuning by enlarging the penalty (to $\lambda = 10000$) HSIC term with universal Gaussian RBF kernel in Figure 8.

Definition 4 (Universal Kernel (Gretton et al., 2005)). *A continuous kernel $k(\cdot, \cdot)$ on a compact metric space $(\mathcal{X}, d)$ is called universal if and only if the RKHS $\mathcal{F}$ induced by the kernel is dense in $C(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$, with respect to the infinity norm $\|f - g\|_\infty$.*



(a) Syn-1 (b) Syn-2

Figure 8: Fine-tuning with the HSIC-RBF loss.

H ADDITIONAL CONVERGENCE RESULTS

In this section, we show the convergence results on Syn-1 as well.

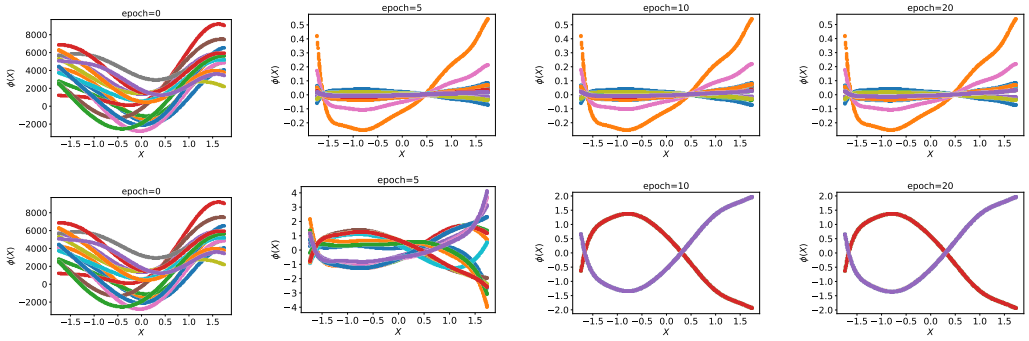


Figure 9: The Algorithm 1 converges on Syn-1. We plot the snapshots of the feature transformations f at training epochs [0, 5, 10, 20], under 15 random initializations (indicated by colors). **Upper:** $\lambda = 0$, most initializations converge to local minimizers (symmetry: $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto (a\boldsymbol{\alpha}, a^{-1}\boldsymbol{\beta})$). **Lower:** $\lambda = 5$, most initializations converge to two local minimizers (symmetry: $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto -(\boldsymbol{\alpha}, \boldsymbol{\beta})$).

I ON THE CHOICE OF λ

We tried seven different values for λ, and report the AUC scores on the PNL-A-unif dataset with different noise levels. We found that the MC-PNL is suitable to use in the small noise regime. We also found that for the data with small noise, smaller λ is preferred; and for the data with large injected noise, larger λ is preferred.

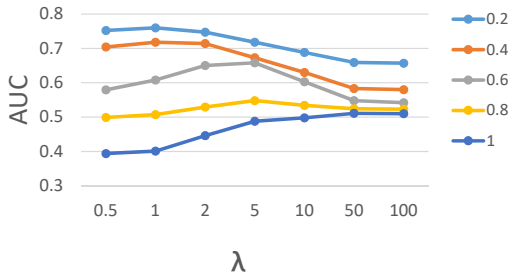


Figure 10: The detailed AUC scores vs. λ under five noise levels on PNL-A-unif data.

J Detailed Data Descriptions

In this section, we describe the datasets in detail.

Gene Datasets:

For D4-S1, D4-S2A, D4-S2B, D4-S2C, we used the preprocessed data in Duong & Nguyen (2022) [1]. D4-S1 contains 36 variable pairs with 105 samples in each pair; D4-S2A, D4-S2B, D4-S2C contains 528, 747, and 579 variable pairs respectively, and each pair contains 210 samples.

The GSE57872 dataset is built on Patel et al. (2014), in which the data has continuous values. Following Choi et al. (2020), we first screen out 657 gene pairs that have corresponding labels in the TRRUST database (Han et al., 2017). The gene contains many repeated values. we examined each gene pair and deleted those repeated expression values.

[1] https://github.com/baosws/CDCI