# Deep Nonlinear Sufficient Dimension Reduction

Zhou Yu

School of Statistics

East China Normal University

November 6, 2025

# Outline



# Outline



# Review of Sufficient Dimension Reduction

Ker-Chau Li (1991) introduced the following model:

$$Y = h(\beta_1^\top X, \beta_2^\top X, \ldots, \beta_d^\top X, \varepsilon). \tag{1}$$

- $Y$ is an univariate output variable;
- The dimension of $X$ is $p$;
- The random error $\varepsilon$ is independent of $X$;
- The space $\mathcal{B}$ generated by $\beta_1, \ldots, \beta_d$ is called the sufficient dimension reduction subspace.
- Model (1) is equivalent to:
  1. The conditional distribution of $Y$ given $X$ depends on $X$ only through the $d$ dimensional variable $\beta_1^\top X, \ldots, \beta_d^\top X$.
  2. Conditioning on $\beta_1^\top X, \ldots, \beta_d^\top X$, $Y$ and $X$ are independent.

# Review of Sufficient Dimension Reduction

(1) can be characterized via conditional independence

$$Y \perp\!\!\!\perp X | \beta_1^\top X, \beta_2^\top X, \ldots, \beta_d^\top X, \tag{2}$$

which is defined as the linear sufficient dimension reduction (SDR).

- Approaching the SDR problem as an inverse regression problem is the most mainstream. But it requires making assumptions on the probability distribution of $X$, which is difficult to justify.
- Rather than the inverse regression framework, [Fukumizu et al., 2009] proposed the kernel dimension reduction, which involves the use of conditional covariance operators on reproducing kernel Hilbert spaces (RKHSs) and does not impose strong assumptions on the probability distribution of $X$.

# Review of Sufficient Dimension Reduction

Nonlinear sufficient dimension reduction can be modeled as

$$Y \perp\!\!\!\perp X | f_1(X), f_2(X) \ldots, f_d(X) \tag{3}$$

- From the perspective of RKHS $\mathcal{H}_\kappa$ generated by the kernel $\kappa(\cdot, X)$, the reproducing property can reformulate $f(X)$ as $\langle f, \kappa(\cdot, X) \rangle_{\mathcal{H}_\kappa}$, which is essentially a direct generalization of the inner product $\beta^\top X$ in (2).
- (3) is equivalent to:
  - The conditional distribution of $Y$ given $X$ depends on $X$ only through the $d$ nonlinear functions $f_1(X), \ldots, f_d(X)$.
  - Conditioning on $f_1(X), \ldots, f_d(X)$, $Y$ and $X$ are independent.

# Review of Sufficient Dimension Reduction

- Consider the model

$$Y = (X_1 + X_2)^2 \log(X_1^2 + X_2^2 + 0.001) + \varepsilon$$

.

- For linear SDR, $d = 2$, $\beta_1 = (1, 0, \ldots, 0)^\top$ and $\beta_1 = (0, 1, 0, \ldots, 0)^\top$. $(\beta_1, \beta_2)$ is not identifiable.
- The concern is $\mathrm{Span}(\beta_1, \beta_2)$ rather than $(\beta_1, \beta_2)$ itself.
- For nonlinear SDR, $d = 1$, $f_0^{(1)}(X) = (X_1+X_2)^2 \log(X_1^2+X_2^2+0.001)$. $f_0^{(1)}(X)$ is not identifiable.
- The concern of nonlinear SDR is the $\sigma-$field generated by $f_0^{(1)}(X)$.

# Review of Sufficient Dimension Reduction

Further, [Lee et al., 2013] developed a general theory for SDR at the level of $\sigma$-fields. A sub $\sigma$-field $\mathcal{G}$ of $\sigma(X)$ is called the sufficient dimension reduction $\sigma$-field if

$$Y \perp\!\!\!\perp X | \mathcal{G}. \tag{4}$$

The minimal sufficient dimension reduction $\sigma$-field is called central $\sigma$-field, denoted by $\mathcal{G}_{Y|X}$, and the set of all $\mathcal{G}_{Y|X}$ -measurable, square-integrable functions is referred to as the central class $\mathfrak{M}_{Y|X}$.

- The definition in (2) and (3) can be incorporated into this framework by regarding $\mathcal{G}$ as $\sigma(\beta^\top X)$ and $\sigma\{\boldsymbol{f}(X)\}$ respectively.
- GSIR (Generalized Sliced Inverse Regression) uses $\overline{\text{ran}}\left(E^*_{X|Y}E_{X|Y}\right)$ to recover the central class. The $E_{X|Y}$ is a conditional expectation operator from $L_2(P_X)$ to $L_2(P_Y)$.

# Motivation

- A common strategy to achieve nonlinear SDR is to combine the kernel trick with the existing linear SDR methods. The drawback of these methods is that they require the computation of the eigenvectors or inverse of a $n \times n$ matrix.
- As the field of machine learning flourishes, many scholars have embarked on the application of machine learning to other statistical fields.
- The nonlinear representation of the neural network function class is more intuitive and powerful compared to that of RKHS.
- The dependence measures of variables play a crucial role in variable screening and SDR.

# Outline



# Nonlinar SDR at the Level of $\sigma$-fields

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be two random vectors. Consider the following nonlinear SDR problem

$$Y \perp\!\!\!\perp X | \boldsymbol{f}_0(X), \tag{5}$$

where $\boldsymbol{f}_0(\cdot)$ is a $\mathbb{R}^d$-valued function ($d \leq p$).

## Assumption 1

*Assume that the family of probability measures $\{P_{X|Y}(\cdot|y) : y \in \Omega\}$ is dominated by a $\sigma$-finite measure. Then there exists a unique minimal sufficient $\sigma$-field (or central $\sigma$-field) $\mathscr{G}_{Y|X}$ such that*

$$Y \perp\!\!\!\perp X | \mathscr{G}_{Y|X}.$$

*Moreover, $\boldsymbol{f}_0$ in (5) satisfies $E\{\boldsymbol{f}_0(X)\} = 0$, $\mathrm{Var}\{\boldsymbol{f}_0(X)\} = I_d$ and*

$$\sigma\{\boldsymbol{f}_0(X)\} = \mathscr{G}_{Y|X}.$$

# Another Look of Nonlinar SDR

Let $\mathfrak{M}_{\sigma\{\boldsymbol{f}_0(X)\}}$ be the central class corresponding to the central $\sigma$-field $\sigma\{\boldsymbol{f}_0(X)\}$. And let $L_2(P_X)$ be the function class of all square-integrable functions with respect to the measure $P_X$. It can be expressed as the orthogonal direct sum

$$L_2(P_X) = \mathfrak{M}_{\sigma\{\boldsymbol{f}_0(X)\}} \oplus \mathfrak{M}^{\perp}_{\sigma\{\boldsymbol{f}_0(X)\}},$$

where $\mathfrak{M}^{\perp}_{\sigma\{\boldsymbol{f}_0(X)\}}$ is orthogonal complement of $\mathfrak{M}_{\sigma\{\boldsymbol{f}_0(X)\}}$. Without loss of generality, consider $g \in \mathfrak{M}^{\perp}_{\sigma\{\boldsymbol{f}_0(X)\}}$ and $E(g(X)) = 0$, it holds that

$$\begin{aligned} E\{g(X)|Y\} &= E\big[E\{g(X)\big|\boldsymbol{f}_0(X), Y\}\big|Y\big] \\ &= E\big[E\{g(X)\big|\boldsymbol{f}_0(X)\}\big|Y\big] \\ &= 0 \\ &\overset{a.s}{=} E\{g(X)\}. \end{aligned}$$

# Generalized Martingale Difference Divergence

- $\mu$ is a Lévy measure on $\mathbb{R}^p\backslash\{0\}$ satisfying $\int_{\mathbb{R}^p\backslash\{0\}}\left(1\wedge\|s\|^2\right)\mu(ds)<\infty$.
- Based on a Lévy measure $\mu$, GMDD of $U\in\mathbb{R}^q$ given $V\in\mathbb{R}^p$ is defined as

$$\mathrm{GMDD}(U|V)=\int_{\mathbb{R}^p}\|\boldsymbol{g}_{U,V}(s)-\boldsymbol{g}_U g_V(s)\|^2\mu(\mathrm{d}s),$$

  where $\boldsymbol{g}_{U,V}(s)=EUe^{is^\top V}$, $\boldsymbol{g}_U=EU$, $g_V(s)=Ee^{is^\top V}$.
- $\mathrm{GMDD}(U|V)\geq 0$. Moreover,

$$\mathrm{GMDD}(U|V)=0\Longleftrightarrow E(U|V)\overset{a.s}{=}E(U)$$

- Let $\Theta(x)=\int\left\{1-\cos\left(s^\top x\right)\right\}\mu(\mathrm{d}s)$ and $\kappa(x_1,x_2)=\Theta(x_1-x_2)$.

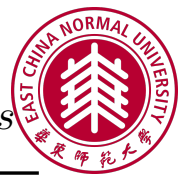
$$\mathrm{GMDD}(U|V)=-E\left\{(U-EU)^\top(\widetilde{U}-E\widetilde{U})\kappa(V,\widetilde{V})\right\},$$

  where $(\widetilde{U},\widetilde{V})$ is an independent copy of $(U,V)$.

# Generalized Martingale Difference Divergence

Table: Some real-valued cndfs on $\mathbb{R}^p$ and the corresponding Lévy measures.

| cndf $\Theta(x)$ | Lévy measure $\mu(\mathrm{d}s)$ |
|---|---|
| $\int \left(1-\cos\left(s^\top x\right)\right)\mu(\mathrm{d}s)$ | $\mu$ finite measure |
| $\frac{\lvert x\rvert_p^2}{\lambda^2+\lvert x\rvert_p^2}, \lambda>0, x\in\mathbb{R}$ | $2\lambda^{-1}e^{-\lambda\lvert s\rvert}\mathrm{d}s$ |
| $\frac{\lvert x\rvert_p^2}{\lambda^2+\lvert x\rvert_p^2}, \lambda>0, x\in\mathbb{R}^p$ | $\int_0^\infty e^{\lambda^2\nu-\frac{\lvert s\rvert^2}{4\nu}}\frac{\lambda^2}{(4\pi\nu)^{p/2}}\,\mathrm{d}\nu\,\mathrm{d}s$ |
| $1-e^{-\frac{1}{2}\lvert x\rvert_p^2}$ | $(2\pi)^{-p/2}e^{-\frac{1}{2}\lvert s\rvert^2}\,\mathrm{d}s$ |
| $\lvert x\rvert_p$ | $\frac{\Gamma\left(\frac{1+p}{2}\right)}{\pi^{\frac{1+p}{2}}}\frac{\mathrm{d}s}{\lvert s\rvert^{p+1}}$ |
| $\frac{1}{2}\lvert x\rvert_p^2$ | no Lévy measure |
| $\lvert x\rvert_p^\alpha, \alpha\in(0,2)$ | $\frac{\alpha 2^{\alpha-1}\Gamma\left(\frac{\alpha+p}{2}\right)}{\pi^{p/2}\Gamma\left(1-\frac{\alpha}{2}\right)}\frac{\mathrm{d}s}{\lvert s\rvert^{p+\alpha}}$ |
| $\ln\left(1+\frac{x^2}{2}\right), x\in\mathbb{R}$ | $\frac{1}{\lvert s\rvert}e^{-\sqrt{\frac{1}{2}}\lvert s\rvert}\mathrm{d}s$ |
| $\ln\left(1+\frac{\lvert x\rvert_p^2}{2}\right), x\in\mathbb{R}^p$ | $\int_0^\infty\int_0^1\frac{1}{\lambda}(2\pi\lambda\rho)^{-p/2}e^{-\lvert s\rvert^2/(2\lambda\rho)}e^{-\rho}d\lambda d\rho ds$ |
| $\ln\cosh(x), x\in\mathbb{R}$ | $\frac{ds}{2s\sinh(\pi s/2)}$ |
| $\ln\cosh\left(\lvert x\rvert_p\right), x\in\mathbb{R}^p$ | $\frac{\Gamma\left(\frac{1+p}{2}\right)}{\pi^{\frac{1+p}{2}}}\frac{\mathrm{d}s}{\lvert s\rvert^{p+1}}+\iint_0^1\int_0^\infty\frac{1}{\lambda}e^{\frac{\lambda\rho}{2}}e^{-\rho}\sum_{n=1}^\infty\frac{\left(-\frac{\lambda\rho}{2}\right)^n}{n!(2\pi)^p}e^{-2n\lvert r\rvert}e^{ir^\top s}d\rho d\lambda dr ds$ |



# Another Look of Nonlinar SDR

## Lemma 1

*Let $g$ be any square-integrable function that lies in the orthogonal complement space of $\mathfrak{M}_{\sigma\{\boldsymbol{f}_0(X)\}}$, then* $\mathrm{GMDD}\left(g(X)|Y\right) = 0$.

For any $f \in L_2(P_X)$, the property of GMDD tells

$$-E\Big(\big[f(X) - E\{f(X)\}\big]\big[f(\widetilde{X}) - E\{f(\widetilde{X})\}\big]\kappa(Y,\widetilde{Y})\Big) \geq 0.$$

Inspired by this relation and Lemma 1, we speculate that the function from $\mathfrak{M}_{\sigma\{\boldsymbol{f}_0(X)\}}$ may maximize the GMDD index. The following theorem gives an affirmative answer.

# Nonlinear SDR via GMDD

## Theorem 2

*Provided assumption 1 holds, let* $m \in \mathbb{N}^+$ *be any positive integer and* $\boldsymbol{f}^* = (f_1^*, \cdots, f_m^*)$ *be an optimal solution of the following objective function*

$$\begin{aligned} \max_{\boldsymbol{f} \in \{L_2(P_X)\}^m} \quad & -E\Big(\big[\boldsymbol{f}(X) - E\{\boldsymbol{f}(X)\}\big]^\top \big[\boldsymbol{f}(\widetilde{X}) - E\{\boldsymbol{f}(\widetilde{X})\}\big]\kappa(Y, \widetilde{Y})\Big) \\ \text{subject to} \quad & \operatorname{Var}\{\boldsymbol{f}(X)\} = I_m. \end{aligned} \tag{6}$$

*Then* $f_j^* \in \mathfrak{M}_{\sigma\{\boldsymbol{f}_0(X)\}}$ *for all* $1 \le j \le m$.

Note that the dimensionality of $\mathfrak{M}_{\sigma\{\boldsymbol{f}_0(X)\}}$ is infinite. However, $\boldsymbol{f}_0 : \mathbb{R}^p \mapsto \mathbb{R}^d$ in (5) indicates the most suitable choice $m = d$. Theorem 2 establishes the unbiasedness of nonlinear SDR method (6) since each component of any optimal solution $\boldsymbol{f}^*$ is $\sigma\{\boldsymbol{f}_0(X)\}$ measurable. In other words, $\sigma\{\boldsymbol{f}^*(X)\} \subseteq \sigma\{\boldsymbol{f}_0(X)\}$.

# The uniform framework of linear and nonlinear SDR

- Consider the linear SDR case of $\boldsymbol{f}(X) = \beta^{\mathsf{T}} X$, then (6) becomes

$$\max_{\beta \in \mathbb{R}^{p\times d}} \; -\operatorname{tr}\Big(\beta^{\mathsf{T}} E\Big[\{X - E(X)\}\{\widetilde{X} - E(\widetilde{X})\}^{\mathsf{T}} \kappa(Y, \widetilde{Y})\Big]\beta\Big)$$
$$\text{subject to} \quad \beta^{\mathsf{T}} \Sigma \beta = I_d,$$

  where $\Sigma = \operatorname{Var}(X)$.

- Classical Sliced Inverse Regression can be reformulated similarly as

$$\max_{\beta \in \mathbb{R}^{p\times d}} \; \operatorname{tr}\Big[\beta^{\mathsf{T}} \cdot \operatorname{Var}(E(X|Y)) \cdot \beta\Big]$$
$$\text{subject to} \quad \beta^{\mathsf{T}} \Sigma \beta = I_d.$$

# Two Unconstrained Objectives

- To solve the optimization problem (6) with constraints, we propose the successive direction extraction method.
- Let $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$ with $\lambda_j > 0$, for $1 \le j \le d$ and first consider

$$L_1(\lambda_1, f_1) = E\Big(\big[f_1(X) - E\{f_1(X)\}\big]\big[f_1(\widetilde{X}) - E\{f_1(\widetilde{X})\}\big]\kappa(Y, \widetilde{Y})\Big) + \lambda_1\big[\operatorname{Var}\{f_1(X)\} - 1\big]^2,$$

which gives

$$f^*_{1,\lambda_1} = \underset{f_1 \in L_2(P_X)}{\operatorname{argmin}} \; L_1(\lambda_1, f_1).$$

# Two Unconstrained Objectives

- For $j \geq 2$, let

$$
\begin{aligned}
& L_j\big(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, f_j)\big) \\
=& E\Big(\big[f_j(X) - E\{f_j(X)\}\big]\big[f_j(\widetilde{X}) - E\{f_j(\widetilde{X})\}\big]\kappa(Y, \widetilde{Y})\Big) \\
& + \lambda_j\big[\operatorname{Var}\{f_j(X)\} - 1\big]^2 + \tilde{\lambda}\left\{\sum_{i=1}^{j-1} \operatorname{Cov}\left(f^*_{i,\lambda_i}(X), f_j(X)\right)\right\}^2,
\end{aligned} \tag{7}
$$

where $\boldsymbol{f}^*_{[j-1]} = (f^*_{1,\lambda_1}, \cdots, f^*_{j-1,\lambda_{j-1}})$, $\tilde{\lambda} > 0$ is predefined constant. Then it follows that

$$
f^*_{j,\lambda_j} = \underset{f_j \in L_2(P_X)}{\operatorname{argmin}} \; L_j\big(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, f_j)\big). \tag{8}
$$

# Two Unconstrained Objectives

- A1 There is a strict gap among $\lambda_i^*$ in Theorem 2 for $m = d + 1$, *i.e.*,

$$\lambda_i^* = E\bar{f}_i^*(X)\bar{f}_i^*(\tilde{X})\kappa(Y,\tilde{Y}) < E\bar{f}_j^*(X)\bar{f}_j^*(\tilde{X})\kappa(Y,\tilde{Y}) = \lambda_j^* < 0,$$

  for $1 \le i < j \le d + 1$. Additionally, we assume $\tilde{\lambda} > -\lambda_1^*$, where $\tilde{\lambda}$ is the constant in (7).

For linear SDR, this assumption necessitates the distinctness of the first $d$ eigenvalues of generalized eigendecomposition problem:

$$\mathrm{GEV}(E\Big[\{X - E(X)\}\{\widetilde{X} - E(\widetilde{X})\}^\top \kappa(Y,\widetilde{Y})\Big], \Sigma).$$

Assumption A1 guarantees that $f_1^*, \cdots, f_d^*$ in Theorem 2 is unique.

# Two Unconstrained Objectives

## Theorem 3

*Under the assumption A1, the solution of the successive direction extraction method is proportional to the optimal direction $f_j^*$ given by (6) in Theorem 2, i.e.,*

$$f_{j,\lambda_j}^* = \pm\sqrt{1-\frac{\lambda_j^*}{2\lambda_j}}f_j^*,\ 1 \le j \le d.$$

# Two Unconstrained Objectives

To alleviate the computational cost, we use a slack variable $\lambda$ of scalar type called the norm multiplier instead of a matrix, regardless of the dimension $d$. Now consider another lagrangian function

$$\begin{aligned} L_F(\lambda, \boldsymbol{f}) =& E\Big(\big[\boldsymbol{f}(X) - E\{\boldsymbol{f}(X)\}\big]^\top \big[\boldsymbol{f}(\widetilde{X}) - E\{\boldsymbol{f}(\widetilde{X})\}\big]\kappa(Y, \widetilde{Y})\Big) \\ &+ \lambda \| \operatorname{Var}\{\boldsymbol{f}(X)\} - I_d \|_F. \end{aligned} \tag{9}$$

## Theorem 4

*Let $\lambda > 0$ and $\boldsymbol{f}^*_{\lambda,F}$ be an optimal solution of the optimization target*

$$\min_{\boldsymbol{f} \in \{L_2(P_X)\}^d} L_F(\lambda, \boldsymbol{f}),$$

*then $f^*_{\lambda,F,j} \in \mathfrak{M}_{\sigma\{\boldsymbol{f}_0(X)\}}$ for all $1 \le j \le d$. If we further assume $\lambda > \left|\lambda^*_1 + \cdots + \lambda^*_d\right|$, the minimization of $L_F(\lambda, \boldsymbol{f})$ is equivalent to (6) in Theorem 2.*

# Estimation

Define two sets

$$\mathcal{I}_{n,k} = \left\{(\pi_1, \ldots, \pi_k) \in \{1, \ldots, i\}^n : \pi_j \neq \pi_l, \text{ for } j \neq l\right\};$$
$$\mathcal{J}_{n,k} = \left\{(\pi_1, \ldots, \pi_k) \in \{1, \ldots, i\}^n : \pi_j < \pi_l, \text{ for } j < l\right\}.$$

Let $Z_i = (X_i, Y_i)$. Given the data

$$\mathscr{D}_n = \left\{Z_1, Z_2, \ldots, Z_n\right\},$$

the most intuitive empirical form of the successive direction extraction method follows as

# Estimation

$$L_{1,n}(\lambda_1, f_1) = \binom{n}{4}^{-1} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\mathcal{J}_{n,4}} h_{\lambda_1,f_1}(Z_{\pi_1}, Z_{\pi_2}, Z_{\pi_3}, Z_{\pi_4}),$$

$$\hat{f}_{1,\lambda_1} = \operatorname*{argmin}_{f_1\in\mathcal{F}_n} L_{1,n}(\lambda_1, f_1),$$

$$L_{j,n}(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, f_j)) = \binom{n}{4}^{-1} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\mathcal{J}_{n,4}} \Big\{h_{\lambda_j,f_j}(Z_{\pi_1}, Z_{\pi_2}, Z_{\pi_3}, Z_{\pi_4}) + \sum_{s=1}^{j-1} h_{\tilde{\lambda},\hat{f}_{s,\lambda_s},f_j}(Z_{\pi_1}, Z_{\pi_2}, Z_{\pi_3}, Z_{\pi_4})\Big\},$$

$$\hat{f}_{j,\lambda_j} = \operatorname*{argmin}_{f_j\in\mathcal{F}_n} L_{j,n}(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, f_j)),$$

# Estimation

where $\hat{\boldsymbol{f}}_{[j-1]} = (\hat{f}_{1,\lambda_1}, \cdots, \hat{f}_{j-1,\lambda_{j-1}})$ is the collection of previous estimated directions, and the two symmetric kernel functions are

$$
\begin{aligned}
& h_{\lambda,f}(Z_1, Z_2, Z_3, Z_4) \\
=& \frac{1}{12} \sum_{(i,j)\in\mathcal{I}_{4,2}} f(X_i)f(X_j)\kappa(Y_i, Y_j) - \frac{1}{12} \sum_{(i,j,k)\in\mathcal{I}_{4,3}} f(X_i)f(X_j)\kappa(Y_j, Y_k) \\
& + \frac{1}{24} \sum_{(i,j,k,l)\in\mathcal{I}_{4,4}} f(X_i)f(X_j)\kappa(Y_k, Y_l) + \lambda h_{1,f,f} \\
& - 2\lambda \left\{ \frac{1}{4} \sum_{(i)\in\mathcal{I}_{4,1}} f(X_i)f(X_i) + \frac{1}{12} \sum_{(i,j)\in\mathcal{I}_{4,2}} f(X_i)f(X_j) \right\} + \lambda
\end{aligned}
$$

# Estimation

$$
\begin{aligned}
&h_{\tilde{\lambda},f_s,f}(Z_1,Z_2,Z_3,Z_4)\\
=&\tilde{\lambda}\Bigg\{\frac{1}{12}\sum_{(i,j)\in\mathcal{I}_{4,2}} f(X_i)f_s(X_i)f(X_j)f_s(X_j)\\
&-\frac{1}{12}\sum_{(i,j,k)\in\mathcal{I}_{4,3}} f(X_i)f_s(X_i)f(X_j)f_s(X_k)\\
&+\frac{1}{24}\sum_{(i,j,k,l)\in\mathcal{I}_{4,4}} f(X_i)f_s(X_j)f(X_k)f_s(X_l)\Bigg\}.
\end{aligned}
$$

# Estimation

- Choose a suitable class $\mathcal{F}_n$ to approximate $L_2(P_X)$.
- we optimize $L_{j,n}\big(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, f_j)\big)$ sequentially from 1 to $d$ (when $j = 1$, $L_{j,n}\big(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, f_j)\big)$ degenerates to $L_{1,n}\big(\lambda_1, f_1\big)$), and all estimations from previous steps should fill into $\hat{\boldsymbol{f}}_{[j-1]}$ in preparation for the next direction of estimation.
- Under the guidance of Theorem 3, we divide the sample $\mathscr{D}_n$ into a learning sample $\mathscr{D}_{n_l}$ of size $n_{n_l}$ and a testing sample $\mathscr{D}_{n_t}$ of $n_t$, where $n = n_l + n_t$. For any $\lambda \in \Lambda$, a pre-given set of the optional hyperparameter, define

$$\hat{f}_{j,\lambda_j,n_t} \in \underset{f_j \in \mathcal{F}_{nt}}{\operatorname{argmin}} L_{j,n_t}\big(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, f_j)\big)$$

based on the learning sample $D_{n_l}$. Then we choose $\lambda^* \in \Lambda$ maximizing the empirical distance correlation ([Böttcher et al., 2018]) of $\hat{f}_{\lambda,n_l}(X)$ and $Y$ on the testing sample $D_{n_t}$.

# The determination of dimension $d$

1. Calculate the estimator

$$\hat{\boldsymbol{f}}_\lambda \in \underset{\boldsymbol{f}\in\mathcal{F}_n^r}{\operatorname{argmin}} L_F(\lambda, \boldsymbol{f})$$

   with $\lambda = 1$ and a relatively large $r$ such as $r = \min(p, 32)$.
2. Give an estimator of GMDD matrix by

$$\hat{\Delta} = \frac{2}{n(n-1)} \sum_{j<k} \left\{ \hat{\boldsymbol{f}}_\lambda(X_j) - \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{f}}_\lambda(X_i) \right\} \left\{ \hat{\boldsymbol{f}}_\lambda(X_k) - \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{f}}_\lambda(X_i) \right\}$$

3. Calculate eigenvalues $\{\hat{\lambda}_i\}_{i=1}^r$ of $\hat{\Delta}$ and sort them in descending order.
4. Define the eigenvalue ratio as $\frac{\hat{\lambda}_i}{\sum_{i=1}^r \hat{\lambda}_i}$. And we determine the ultimate dimension $d$ for SDR to be the minimum integer $k$ such that the cumulative eigenvalue ratio

$$\sum_{j=1}^k \hat{\lambda}_j \Big/ \sum_{i=1}^r \hat{\lambda}_i \geq 0.9.$$

# Deep Neural Networks

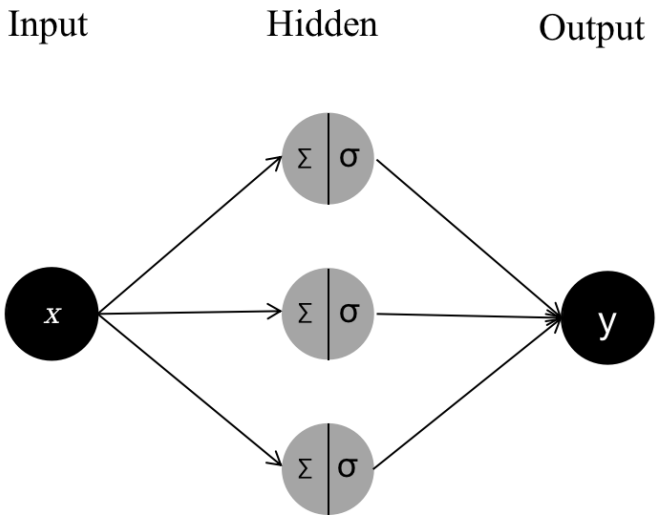



Figure: Single layer ReLU neural network.

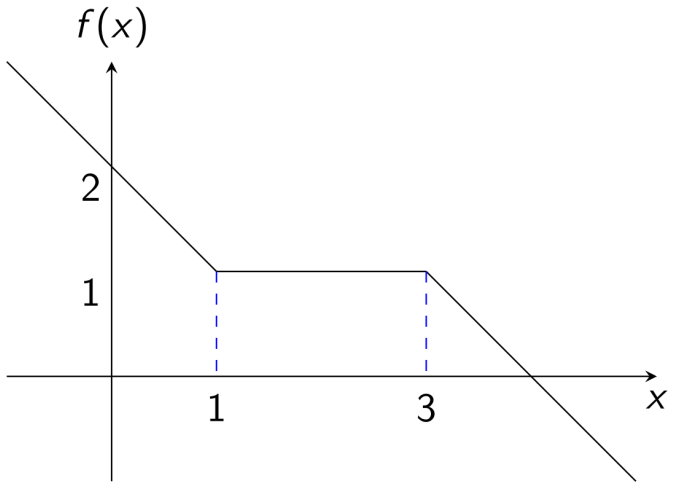
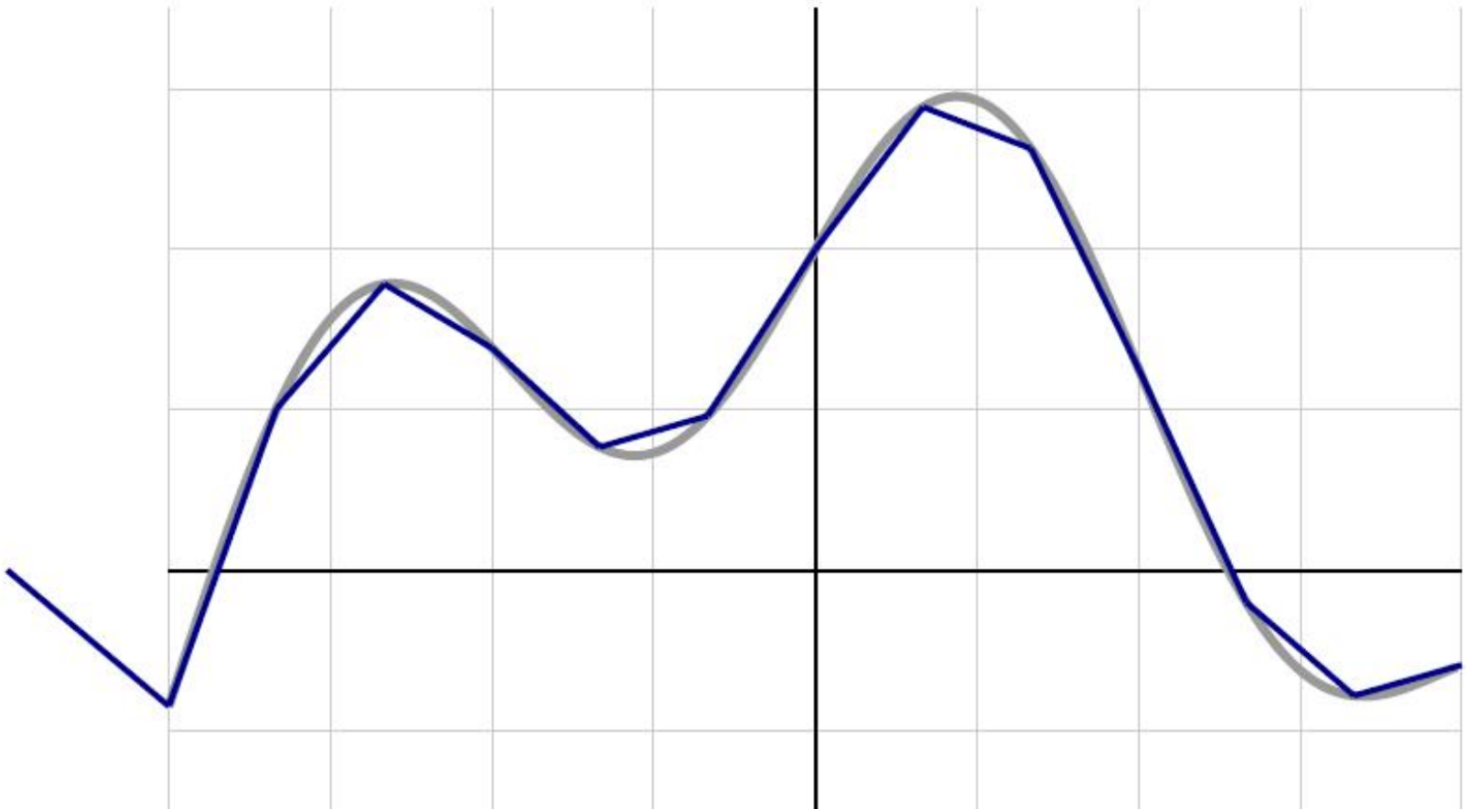



Figure: Piecewise linear approximation, $f(x) = \sigma(1)+\sigma(-x+1)-\sigma(x-3)$.

# Deep Neural Networks



# Deep Neural Networks

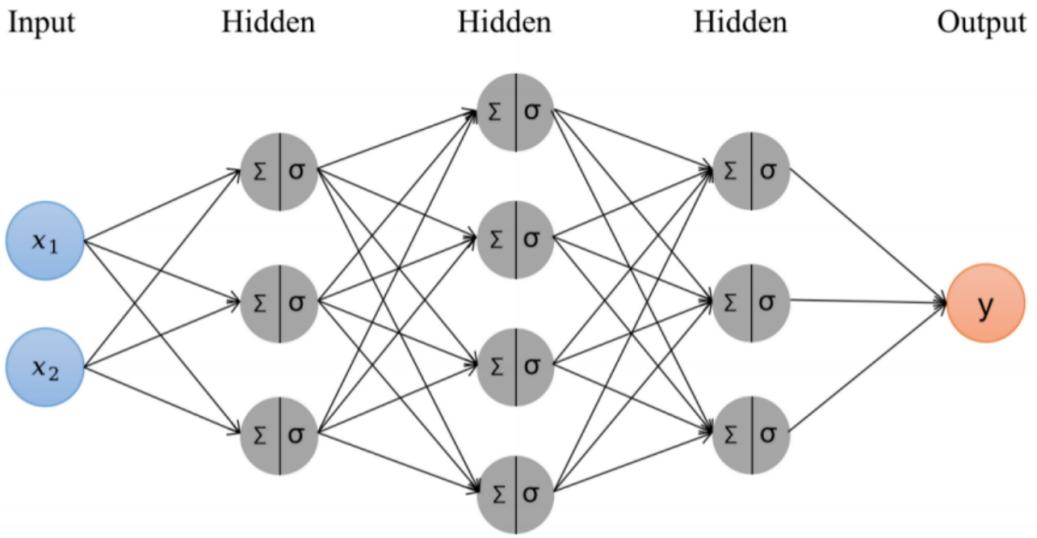



Figure: A feedforward neural network with width $N = 4$, depth $L = 3$, and size $S = 44$.

- Depth $\mathcal{D}$: the number of all hidden layers;
- Width $\mathcal{W}$: the maximum width of all hidden layers;
- Size $\mathcal{S}$: the total number of parameters in the network;
- Number of neurons $\mathcal{U}$: the number of nodes in hidden layers.

# Deep Neural Networks

There is a lot of literature investigating the convergence properties of nonparametric regression using feedforward neural networks.

- [Györfi et al., 2002] gave a discussion about consistency and prediction error bound of deep neural networks based on the least squares objective function.
- [Farrell et al., 2021] established non-asymptotic bounds for deep feed-forward architecture net for general Lipschitz loss functions with the demand that the regression function lies in a Sobolev ball.
- [Schmidt-Hieber, 2020] proved sparsely connected and well-designed neural networks with ReLU activation function can achieve the nearly minimax rates of convergence under a general composition assumption on the regression function.

# Deep Neural Networks

- [Bauer and Kohler, 2019] derived the prediction error bound when considering to use specially structured sigmoid activated multilayer neural networks to fit a similar structured regression function.
- [Kohler et al., 2022] gave similar results but took into account a different form of regression function and a simpler structured network.
- [Jiao et al., 2021] further improved the prefactor of optimal bounds on the prediction error to make it more meaningful in the high-dimensional setting.

# Outline



# Excess Risk

We define the excess risk for the $j$th step estimation

$$R_j(\lambda_j, \hat{f}_{j,\lambda_j}) = L_j(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \hat{f}_{j,\lambda_j})) - L_j(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, f^*_{j,\lambda_j}))$$

and the total excess risk

$$R(\boldsymbol{\lambda}, \hat{\boldsymbol{f}}) = \sum_{j=1}^{d} R_j(\lambda_j, \hat{f}_{j,\lambda_j}),$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$, $\hat{\boldsymbol{f}} = \hat{\boldsymbol{f}}_{[d]} = (\hat{f}_{1,\lambda_1}, \cdots, \hat{f}_{d,\lambda_d})$ is obtained from (10) and $\boldsymbol{f}^*_{[j-1]} = (f^*_{1,\lambda_1}, \cdots, f^*_{j-1,\lambda_{j-1}})$.

# Excess Risk

- Consider the simplest case where $d = 1$.

$$
\begin{aligned}
R(\boldsymbol{\lambda}, \hat{\boldsymbol{f}}) =& R_1(\lambda_1, \hat{f}_{1,\lambda_1}) = L_1(\lambda_1, \hat{f}_{1,\lambda_1}) - L_1\left(\lambda_1, f^*_{1,\lambda_1}\right) \\
&= \underbrace{\left\{ L_1(\lambda_1, \hat{f}_{1,\lambda_1}) - \inf_{f_1 \in \mathcal{F}_n} L_1(\lambda_1, f_1) \right\}}_{\textit{the statistical error}} \\
&+ \underbrace{\left\{ \inf_{f_1 \in \mathcal{F}_n} L_1(\lambda_1, f_1) - L_1\left(\lambda_1, f^*_{1,\lambda_1}\right) \right\}}_{\textit{the approximation error}}.
\end{aligned}
$$

- A common strategy for tackling the first term is to reduce the U - process related problem to an empirical process problem by introducing Rademacher random variables.
- The approximation error can be solved by the functional approximation theory.

# Assumptions

We make the following assumptions.

- (A2) There exists an absolute constant $B > 1$ such that $\|f^*_{j,\lambda_j}\|_\infty \le B$, $\|f\|_\infty \le B$, and $\kappa(y, \tilde{y}) \le B$ for any $1 \le j \le d$, $f \in \mathcal{F}_n$, and $y, \tilde{y} \in \mathbb{R}^q$.
- (A3) $f^*_{j,\lambda_j}$ is $\beta$-Hölder continuous for all $1 \le j \le d$, *i.e.*, $f^*_{j,\lambda_j} \in C^\beta([0,1]^p)$.
- (A4) There exist an universal constant $C$ and a parameter $V$ depending on $\mathcal{F}_n$ such that

$$\log N(\epsilon, \mathcal{F}_n, \|\cdot\|_{L_2(Q)}) \le CV\{1 + \log(1/\epsilon)\},$$

where $Q$ is the probability measure $P_X$ or $P_n$, the empirical probability measure of $X$.

# Assumptions

- By Theorem 2.6.7 in [Vaart and Wellner, 1996] and Theorem 7 in [Bartlett the log covering number of the scaler-valued ReLU neural network class $\mathcal{F}_n$ with respect to $L_2$-norm for any probability measure $Q$ can be bounded by

$$\begin{aligned}\log N\left(\epsilon, \mathcal{F}_n, \|\cdot\|_{L_2(Q)}\right) &\leq K_1 \cdot \mathrm{VC}(\mathcal{F}_n)\{1+\log(1/\epsilon)\} \\ &\leq K_2 \cdot \mathcal{S}\mathcal{L}\log(\mathcal{S})\{1+\log(1/\epsilon)\}.\end{aligned}$$

  Assumption (A4) is satisfied with $V = \mathrm{VC}(\mathcal{F}_n)$ for ReLU neural networks.

- Regarding the approximation error, the $\beta$-Hölder continuous function class $C^\beta([0,1]^p)$ can be approximated by the ReLU networks with error (see Theorem 1.1 in [Shen, 2020])

$$\sup_{f\in\mathcal{F}_n; h\in C^\beta([0,1]^p)} \|f-h\|_\infty = \mathcal{O}\left((\mathcal{N}\mathcal{L})^{-\frac{2\beta}{p}}\right).$$

# Slow Rate

Define the Rademacher average

$$R_n = \sup_{f_1 \in \mathcal{F}_n} \frac{1}{\lfloor n/4 \rfloor} \left| \sum_{i=1}^{\lfloor n/4 \rfloor} \epsilon_i h_{\lambda_1, f_1}(Z_i, Z_{\lfloor n/4 \rfloor + i}, Z_{2\lfloor n/4 \rfloor + i}, Z_{3\lfloor n/4 \rfloor + i}) \right|,$$

where $\epsilon_1, ..., \epsilon_n$ are i.i.d. Rademacher random variables (i.e., random symmetric sign variables) and $Z = (X, Y)$.

## Lemma 5

*Assume $\hat{f}_{1,\lambda_1}$ is a minimizer of $L_{1,n}(\lambda_1, f_1)$, then for any convex nondecreasing function $\psi$,*

$$E\psi\left( L_1(\lambda_1, \hat{f}_{1,\lambda_1}) - \inf_{f_1 \in \mathcal{F}_n} L_1(\lambda_1, f_1) \right) \leq E\psi(4R_n).$$

# Slow Rate

By setting $\psi(x) = e^x$, we can obtain the moment generating function of $L_1(\lambda_1, \hat{f}_{1,\lambda_1}) - \inf_{f_1 \in \mathcal{F}_n} L_1(\lambda_1, f_1)$, which is the essential component to deriving the non-asymptotic bound.

## Theorem 6

*Under the assumption A1, A2 and A4, if $n \geq V$, for any $\delta > 0$ with probability at least $1 - \delta$, it holds that*

$$
\begin{aligned}
& L_1(\lambda_1, \hat{f}_{1,\lambda_1}) - L_1\left(\lambda_1, f^*_{1,\lambda_1}\right) \\
=& \mathcal{O}\left( \sqrt{\frac{V}{n}} + \sqrt{\frac{\log(1/\delta)}{n-3}} + \inf_{f_1 \in \mathcal{F}_n} \| f_1 - f^*_{1,\lambda_1} \|^2_\infty \right).
\end{aligned}
$$

# Slow Rate For Neural Networks

## Corollary 7

*For any arbitrary* $N \in \mathbb{N}^+$, *consider* $\mathcal{F}_n$ *as scalar-valued ReLu neural network classes with the width* $\mathcal{N} = 3^{p+1}\max\left(p\lfloor N^{1/p}\rfloor, N+1\right)$ *and depth* $\mathcal{L} = 12n^{\frac{p}{2(p+4\beta)}}+14+2p$. *Under the assumption A1, A2 and A3, if* $n \geq \mathcal{S}\mathcal{L}\log(\mathcal{S})$, *the following result holds for any* $\delta > 0$ *with probability at least* $1-\delta$,

$$L_1(\lambda_1, \hat{f}_{1,\lambda_1}) - L_1(\lambda_1, f^*_{1,\lambda_1}) = \mathcal{O}\left(n^{-\frac{2\beta}{p+4\beta}}\log^{\frac{1}{2}} n + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

- [Shen, 2020] provided a quantitative and non-asymptotic approximation rate of deep ReLU neural networks in terms of width and depth.
- [Bartlett et al., 2019] established the nearly tight VC dimension bound of the function class constructed by feedforward neural networks with the piecewise-linear activation function.

# Fast Rate

## Theorem 8

*Under the assumption A1, A2 and A4, the following holds for any $\delta > 0$ with probability at least $1 - \delta$,*

$$L_1\big(\lambda_1, \hat{f}_{1,\lambda_1}\big) - L_1\big(\lambda_1, f^*_{1,\lambda_1}\big) = \mathcal{O}\left(\frac{V}{n} + \frac{V\log(1/\delta)}{n} + \inf_{f_1 \in \mathcal{F}_n} \|f_1 - f^*_{1,\lambda_1}\|^2_\infty\right)$$

$$\rho^2\big(\hat{f}_{1,\lambda_1}, f^*_{1,\lambda_1}\big) = \mathcal{O}\left(\frac{V}{n} + \frac{V\log(1/\delta)}{n} + \inf_{f_1 \in \mathcal{F}_n} \|f_1 - f^*_{1,\lambda_1}\|^2_\infty\right).$$

# Fast Rate

## Corollary 9

*For any arbitrary* $N \in \mathbb{N}^+$, *consider* $\mathcal{F}_n^d$ *as the ReLu neural network classes with the width* $\mathcal{N} = 3^{p+1}\max\left(p\lfloor N^{1/p}\rfloor, N+1\right)$ *and depth* $\mathcal{L} = 12n^{\frac{p}{2(p+2\beta)}} + 14 + 2p$. *Under the assumption A1, A2 and A3, if* $n \geq \mathcal{S}\mathcal{L}\log(\mathcal{S})$, *the following holds for any* $\delta > 0$ *with probability at least* $1-\delta$,

$$L_1\big(\lambda_1, \hat{f}_{1,\lambda_1}\big) - L_1\big(\lambda_1, f^*_{1,\lambda_1}\big) = \mathcal{O}\left(n^{-\frac{2\beta}{p+2\beta}}\left\{1+\log(1/\delta)\right\}\log n\right),$$

$$\rho^2\big(\hat{f}_{1,\lambda_1}, f^*_{1,\lambda_1}\big) = \mathcal{O}\left(n^{-\frac{2\beta}{p+2\beta}}\left\{1+\log(1/\delta)\right\}\log n\right).$$

# Fast Rate

- The implication of Corollary 9 is

$$E\rho^2(\hat{f}_{1,\lambda_1}, f^*_{1,\lambda_1}) = \mathcal{O}\left(n^{-\frac{2\beta}{p+2\beta}} \log n\right).$$

- The implication of Corollary 9 is

$$E\rho^2(\hat{f}_{1,\lambda_1}, f^*_{1,\lambda_1}) = \mathcal{O}\left(n^{-\frac{2\beta}{p+2\beta}} \log n\right).$$

- Remarkably, our nonlinear SDR estimator, leveraging deep neural networks, achieves this minimax optimal rate up to $\log n$.
- Minimax optimal rate rate up to $\log n$ also is obtained in [Schmidt-Hieber, for nonparametric regression probelm with ReLU networks.

## Theorem 10

*Under the same assumptions of Theorem 8, the following holds for any $\delta > 0$ with probability at least $1 - (4j-3)\delta$,*

$$L_j\big(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \hat{f}_{j,\lambda_j})\big) - L_j\big(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, f^*_{j,\lambda_j})\big) = \mathcal{O}\big(A(n,V,\delta)\big),$$
$$\rho^2(\hat{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) = \mathcal{O}\big(A(n,V,\delta)\big),$$

*where*

$$A(n,V,\delta) = \frac{V}{n} + \frac{V\log(1/\delta)}{n} + \sup_{1\le j\le d}\inf_{f_j\in\mathcal{F}_n} \|f_j - f^*_{j,\lambda_j}\|^2_\infty.$$

*Moreover, if the assumptions in Corollary 9 hold true, then for $1 \le j \le d$,*

$$E\rho^2(\hat{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) = \mathcal{O}\left(n^{-\frac{2\beta}{p+2\beta}}\log n\right).$$

# Fast Rate: Hoeffding Decomposition

Set

$$q_{\lambda_1,f_1}(Z_1,Z_2,Z_3,Z_4) = h_{\lambda_1,f_1}(Z_1,Z_2,Z_3,Z_4) - h_{\lambda_1,f^*_{1,\lambda_1}}(Z_1,Z_2,Z_3,Z_4),$$

then $R_1(\lambda_1,f_1) = L_1(\lambda_1,f_1) - L_1(\lambda_1,f^*_{1,\lambda_1}) = E\{q_{\lambda_1,f_1}(Z_1,Z_2,Z_3,Z_4)\}$ and the corresponding sample estimate of $R_1(\lambda_1,f_1)$ is

$$R_{1,n}(\lambda_1,f_1) = \frac{(n-4)!}{n!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\mathcal{I}_{n,4}} q_{\lambda_1,f_1}(Z_{\pi_1},Z_{\pi_2},Z_{\pi_3},Z_{\pi_4}).$$

Define the conditional kernel as

$$\begin{aligned} q^k_{\lambda_1,f_1}(z_1,\ldots,z_k) &= E\{q_{\lambda_1,f_1}(Z_1,\ldots,Z_4)\big|Z_1=z_1,\ldots,Z_k=z_k\} \\ &= \int q_{\lambda_1,f_1}(z_1,\ldots,z_4)\,dP(z_{k+1})\ldots dP(z_4), \quad k=0,\ldots, \end{aligned}$$

and an operator $U_{n,i}$ which maps a $i$th order kernel $q$ to a $U$-statistic as

$$U_{n,i}(q) = \frac{(n-i)!}{n!} \sum_{(\pi_1,\ldots,\pi_i)\in\mathcal{I}_{n,i}} q(Z_{\pi_1},\ldots,Z_{\pi_i}).$$

# Fast Rate: Hoeffding Decomposition

Then the Hoeffding decomposition of $R_{1,n}(\lambda_1, f_1)$ is

$$\begin{aligned} R_{1,n}(\lambda_1, f_1) =& A_{1,1,n}(\lambda_1, f_1) + A_{2,1,n}(\lambda_1, f_1) \\ =& \sum_{i=0}^{1} \binom{4}{i} U_{n,i}(q^{(i)}_{\lambda_1,f_1}) + \sum_{i=2}^{4} \binom{4}{i} U_{n,i}(q^{(i)}_{\lambda_1,f_1}), \end{aligned}$$

where $\binom{4}{i} = \frac{4!}{i!(4-i)!}$ and
$q^{(i)}_{\lambda_1,f_1}(x_1, \ldots, x_i) = \sum_{k=0}^{i} (-1)^{i-k} \sum_{\mathcal{J}_{i,k}} q^{k}_{\lambda_1,f_1}(x_{\pi_1}, \ldots, x_{\pi_k})$.

# Fast Rate: Main Steps

Let

$$\ell_1(\lambda_1, f_1, z) = 4Eh_{\lambda_1,f_1}(z, Z_2, Z_3, Z_4) - 3L_1(\lambda_1, f_1).$$

Consider a new target function and its centered version as

$$v_{1,n}(\lambda_1, f_1) = \frac{1}{n}\sum_{i=1}^{n} \ell_1(\lambda_1, f_1, Z_i),$$

$$\bar{v}_{1,n}(\lambda_1, f_1) = \frac{1}{n}\sum_{i=1}^{n} \ell_1(\lambda_1, f_1, Z_i) - L_1(\lambda_1, f_1).$$

Then $A_{1,1,n}(\lambda_1, f_1)$ can be expressed as

$$A_{1,1,n}(\lambda_1, f_1) = \sum_{i=1}^{n} \ell_1(\lambda_1, f_1, z_i) - \sum_{i=1}^{n} \ell_1(\lambda_1, f^*_{1,\lambda_1}, z_i),$$

# Fast Rate: Main Steps

- Define a distance on $L_2(P_X) \times L_2(P_X)$ as

$$\rho^2(f,g) = C_1 \min\left\{\|f-g\|^2_{L_2(P_X)}, \|f+g\|^2_{L_2(P_X)}\right\},$$

where $C_1$ is appropriate constant.

- By the empirical process theory, we should find a function $\phi : [0,\infty) \mapsto [0,\infty)$ with property:

$$W(\sigma) = E\left[\sup_{f_1 \in \mathcal{F}_n, d^2(f^*_{1,\mathcal{F}_n}, f_1) \le \sigma^2} \left\{\bar{v}_n(\lambda, f^*_{1,\mathcal{F}_n}) - \bar{v}_n(\lambda, f_1)\right\}\right] \le \phi(\sigma)/\sqrt{n}$$

where $f^*_{1,\mathcal{F}_n}$ is the minimizer of $L_1(\lambda_1, f_1)$ over $\mathcal{F}_n$.

# Fast Rate: Generalization to $d > 1$

- Define

$$A(n, V, \delta) = \frac{V}{n} + \frac{V \log(1/\delta)}{n} + \sup_{1 \le j \le d} \inf_{f_j \in \mathcal{F}_n} \| f_j - f^*_{j,\lambda_j} \|^2_\infty.$$

- Assume that with probability $1 - (4j-7)\delta$ for any $0 < i \le j - 1 < d$, it holds that

$$\begin{aligned} L_i\big(\lambda_i, (\boldsymbol{f}^*_{[i-1]}, \hat{f}_{i,\lambda_i})\big) - L_i\big(\lambda_i, (\boldsymbol{f}^*_{[i-1]}, f^*_{i,\lambda_i})\big) &= \mathcal{O}\big(A(n, V, \delta)\big), \\ \rho^2(\hat{f}_{i,\lambda_i}, f^*_{i,\lambda_i}) &= \mathcal{O}\big(A(n, V, \delta)\big). \end{aligned} \tag{11}$$

- Introduce the intermediate function $\tilde{f}_{j,\lambda_j}$ as the minimizer of

$$L_{j,n}\big(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, f_j)\big)$$

over $f_j \in \mathcal{F}_n$.

# Fast Rate: Generalization to $d > 1$

- All the relationship in $d = 1$ can transfer to $d > 1$ with different universal constant, except the inequality

$$\left| A_{1,1,n}(\lambda_1, \hat{f}_{1,\lambda_1}) - \inf_{f_1} A_{1,1,n}(\lambda_1, f_1) \right| = \mathcal{O}\left( \sup_{f_1} \left| A_{2,1,n}(\lambda_1, f_1) \right| \right).$$

Such gap also holds for $A_{1,j,n}$ with $\tilde{f}_{j,\lambda_j}$ rather than $\hat{f}_{j,\lambda_j}$, that is,

$$\begin{aligned} &\left| A_{1,j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \tilde{f}_{j,\lambda_j})) - \inf_{f_j} A_{1,j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, f_j)) \right| \\ =&\mathcal{O}\left( \sup_{f_j} \left| A_{2,j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, f_j)) \right| \right), \end{aligned}$$

which implies $\rho^2(\tilde{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) = \mathcal{O}(A(n, V, \delta))$.

# Fast Rate: Generalization to $d > 1$

- the bound of $R_j(\lambda_j, \hat{f}_{j,\lambda_j})$ can be derived once we know

$$|A_{1,j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \hat{f}_{j,\lambda_j})) - A_{1,j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \tilde{f}_{j,\lambda_j}))|.$$

- By the same relationship $R_{j,n} = A_{1,j,n} + A_{2,j,n}$ as the case of $d = 1$, we can turn to bound

$$\begin{aligned}
&R_{j,n}(\lambda_j, \hat{f}_{j,\lambda_j}) - R_{j,n}(\lambda_j, \tilde{f}_{j,\lambda_j}) \\
=&L_{j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \hat{f}_{j,\lambda_j})) - L_{j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \tilde{f}_{j,\lambda_j})) \\
=&\left\{L_{j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \hat{f}_{j,\lambda_j})) - L_{j,n}(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, \hat{f}_{j,\lambda_j}))\right\} \\
+&\left\{L_{j,n}(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, \hat{f}_{j,\lambda_j})) - L_{j,n}(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, \tilde{f}_{j,\lambda_j}))\right\} \\
+&\left\{L_{j,n}(\lambda_j, (\hat{\boldsymbol{f}}_{[j-1]}, \tilde{f}_{j,\lambda_j})) - L_{j,n}(\lambda_j, (\boldsymbol{f}^*_{[j-1]}, \tilde{f}_{j,\lambda_j}))\right\} \\
\triangleq&s_1 + s_2 + s_3,
\end{aligned}$$

# Fast Rate: Generalization to $d > 1$

- by the inequality $(\sum_{i=1}^{k} x_i)^2 \leq k \sum_{i=1}^{k} x_i^2$, the orthogonality of $f^*_{s,\lambda_s}, f^*_{k,\lambda_k}$ with $s \neq k$ and (11), we have

$$
\begin{aligned}
Es_1 =& \mathcal{O}\left( \sum_{i=1}^{j-1} \left\{ \mathrm{Cov}^2(f^*_{i,\lambda_i}, \hat{f}_{j,\lambda_j}) - \mathrm{Cov}^2(\hat{f}_{i,\lambda_i}, \hat{f}_{j,\lambda_j}) \right\} \right) \\
=& \mathcal{O}\left( \sum_{i=1}^{j-1} \left\{ \rho^2(\hat{f}_{i,\lambda_i}, f^*_{i,\lambda_i}) + \rho(\hat{f}_{i,\lambda_i}, f^*_{i,\lambda_i}) \rho(\hat{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) \right\} \right) \\
=& \mathcal{O}\Big( A(n,V,\delta) + \sqrt{A(n,V,\delta)} \rho(\hat{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) \Big), \\
Es_3 =& \mathcal{O}\left( \sum_{i=1}^{j-1} \left\{ \rho^2(\hat{f}_{i,\lambda_i}, f^*_{i,\lambda_i}) + \rho(\hat{f}_{i,\lambda_i}, f^*_{i,\lambda_i}) \rho(\tilde{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) \right\} \right) \\
=& \mathcal{O}\Big( A(n,V,\delta) + \sqrt{A(n,V,\delta)} \rho(\tilde{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) \Big) \\
=& \mathcal{O}\Big( A(n,V,\delta) \Big).
\end{aligned}
$$

# Fast Rate: Generalization to $d > 1$

- Bounded difference inequality entails

$$s_1 = \mathcal{O}\Big( A(n, V, \delta) + \sqrt{A(n, V, \delta)} \rho(\hat{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) \Big),$$

$$s_3 = \mathcal{O}\Big( A(n, V, \delta) \Big)$$

with high probability, which gives

$$\sum_{i=1}^{3} s_i = \mathcal{O}\Big( A(n, V, \delta) + \sqrt{A(n, V, \delta)} \rho(\hat{f}_{j,\lambda_j}, f^*_{j,\lambda_j}) \Big).$$

# Fast Rate: Generalization to $d > 1$

- Finally, we can bound

$$\begin{aligned}&\left|A_{1,j,n}\big(\lambda_j,(\boldsymbol{f}^*_{[j-1]},\hat{f}_{j,\lambda_j})\big)-\inf_{f_j}A_{1,j,n}\big(\lambda_j,(\boldsymbol{f}^*_{[j-1]},f_j)\big)\right|\\&=\mathcal{O}\Big(A(n,V,\delta)+\sqrt{A(n,V,\delta)}\rho(\hat{f}_{j,\lambda_j},f^*_{j,\lambda_j})\Big)\end{aligned}$$

- Therefore, we obtain

$$\begin{aligned}\rho^2(\hat{f}_{j,\lambda_j},f^*_{j,\lambda_j})=&\mathcal{O}(R_j(\lambda_j,\hat{\boldsymbol{f}}_{j,\lambda_j}))\\=&\mathcal{O}\Big(A(n,V,\delta)+\sqrt{A(n,V,\delta)}\rho(\hat{f}_{j,\lambda_j},f^*_{j,\lambda_j})\Big).\end{aligned}$$

It concludes

$$\rho^2(\hat{f}_{j,\lambda_j},f^*_{j,\lambda_j})=\mathcal{O}\Big(A(n,V,\delta)\Big).$$

# Outline



# Simulation of $d = 1$

We compare our method against GSIR. Six different variants of our method with six different distance metric $\kappa$ for $Y$ is denoted by the numbers $1 - 6$.
**Setting**:

$$\begin{cases} A: & Y = \frac{X_1}{1+(1+X_2)^2} + (1+X_2)^2 + \epsilon_1, \\ B: & Y = \sin\big((X_1+X_2)\pi/10\big) + X_1^2 + \epsilon_2, \\ C: & Y = (X_1+X_2)^2 \log(X_1^2 + X_2^2 + 0.001) + \epsilon_3, \end{cases}$$

with $\epsilon \sim N(0, 0.25)$ independent of $X$, as well as three different distributional scenarios for the predictor vector $X = (X_1, \ldots, X_p)^\top$:

$$\begin{cases} \text{I}: & X_i \sim N(0, 1/2) \text{ where } 1 \le i \le p \text{ and } X_i, X_j \text{ are mutually indepen} \\ \text{II}: & X_i \sim -1 + \text{Pois}(1) \text{ where } 1 \le i \le p \text{ and } X_i, X_j \text{ are mutually inde} \\ \text{III}: & X \sim t_4(\mathbf{0}_p, 0.6 \cdot \mathbf{I}_p + 0.4 \cdot \mathbf{1}_{p\times p}), \end{cases}$$

# Simulation of $d = 1$

The quality of estimated sufficient predictors is assessed by their distance correlation with the true sufficient predictors.

Table: Distance correlation with true predictor when $p = 50, d = 1$ for GMDDNet-S and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| A-I | 1k | 0.88(0.01) | 0.89(0.01) | 0.76(0.03) | 0.74(0.03) | 0.86(0.01) | 0.89(0.01) | 0.92(0.00) |
| | 2k | 0.89(0.01) | 0.92(0.01) | 0.79(0.01) | 0.78(0.02) | 0.88(0.01) | 0.91(0.01) | 0.92(0.01) |
| | 5k | 0.93(0.00) | 0.95(0.00) | 0.81(0.01) | 0.81(0.01) | 0.91(0.01) | 0.94(0.00) | 0.93(0.00) |
| A-II | 1k | 0.93(0.01) | 0.94(0.01) | 0.74(0.02) | 0.74(0.02) | 0.86(0.01) | 0.95(0.00) | 0.92(0.00) |
| | 2k | 0.95(0.01) | 0.97(0.00) | 0.78(0.02) | 0.76(0.01) | 0.88(0.01) | 0.97(0.00) | 0.92(0.00) |
| | 5k | 0.97(0.00) | 0.98(0.00) | 0.82(0.01) | 0.80(0.01) | 0.92(0.01) | 0.98(0.00) | 0.92(0.00) |
| A-III | 1k | 0.86(0.06) | 0.75(0.11) | 0.61(0.06) | 0.60(0.06) | 0.82(0.04) | 0.80(0.11) | 0.61(0.05) |
| | 2k | 0.93(0.03) | 0.80(0.13) | 0.66(0.03) | 0.65(0.03) | 0.88(0.02) | 0.88(0.08) | 0.59(0.12) |
| | 5k | 0.96(0.01) | 0.89(0.06) | 0.69(0.02) | 0.67(0.02) | 0.88(0.02) | 0.95(0.03) | 0.61(0.02) |

# Simulation of $d = 1$

Table: Distance correlation with true predictor when $p = 50, d = 1$ for GMDDNet-S and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| B-I | 1k | 0.50(0.05) | 0.50(0.05) | 0.35(0.04) | 0.37(0.04) | 0.50(0.04) | 0.50(0.07) | 0.38(0.04) |
| | 2k | 0.66(0.03) | 0.67(0.03) | 0.56(0.04) | 0.58(0.04) | 0.67(0.02) | 0.67(0.03) | 0.42(0.03) |
| | 5k | 0.74(0.01) | 0.77(0.02) | 0.64(0.02) | 0.67(0.02) | 0.72(0.01) | 0.76(0.02) | 0.43(0.02) |
| B-II | 1k | 0.81(0.03) | 0.80(0.04) | 0.38(0.07) | 0.31(0.09) | 0.74(0.04) | 0.83(0.03) | 0.70(0.02) |
| | 2k | 0.87(0.02) | 0.90(0.01) | 0.53(0.02) | 0.50(0.07) | 0.82(0.02) | 0.91(0.01) | 0.72(0.01) |
| | 5k | 0.90(0.02) | 0.94(0.01) | 0.58(0.02) | 0.59(0.02) | 0.86(0.01) | 0.94(0.01) | 0.72(0.02) |
| B-III | 1k | 0.88(0.09) | 0.72(0.13) | 0.66(0.04) | 0.66(0.04) | 0.86(0.02) | 0.80(0.13) | 0.57(0.03) |
| | 2k | 0.92(0.08) | 0.80(0.14) | 0.66(0.04) | 0.67(0.04) | 0.88(0.02) | 0.86(0.15) | 0.55(0.02) |
| | 5k | 0.97(0.01) | 0.84(0.14) | 0.70(0.02) | 0.71(0.02) | 0.90(0.01) | 0.91(0.06) | 0.57(0.03) |

# Simulation of $d = 1$

Table: Distance correlation with true predictor when $p = 50, d = 1$ for GMDDNet-S and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| C-I | 1k | 0.34(0.08) | 0.31(0.08) | 0.29(0.05) | 0.29(0.06) | 0.38(0.05) | 0.34(0.08) | 0.11(0.04) |
| | 2k | 0.72(0.05) | 0.67(0.10) | 0.50(0.05) | 0.54(0.04) | 0.70(0.06) | 0.70(0.10) | 0.14(0.02) |
| | 5k | 0.88(0.04) | 0.90(0.03) | 0.61(0.03) | 0.67(0.05) | 0.81(0.04) | 0.88(0.02) | 0.19(0.02) |
| C-II | 1k | 0.70(0.04) | 0.60(0.06) | 0.22(0.06) | 0.23(0.08) | 0.65(0.04) | 0.65(0.05) | 0.61(0.02) |
| | 2k | 0.81(0.03) | 0.69(0.03) | 0.41(0.08) | 0.42(0.09) | 0.74(0.02) | 0.74(0.04) | 0.61(0.02) |
| | 5k | 0.95(0.00) | 0.82(0.12) | 0.58(0.02) | 0.56(0.02) | 0.91(0.01) | 0.92(0.04) | 0.62(0.01) |
| C-III | 1k | 0.73(0.08) | 0.60(0.10) | 0.54(0.07) | 0.53(0.08) | 0.77(0.07) | 0.64(0.13) | 0.62(0.08) |
| | 2k | 0.79(0.06) | 0.56(0.11) | 0.55(0.03) | 0.54(0.03) | 0.82(0.02) | 0.66(0.11) | 0.64(0.03) |
| | 5k | 0.85(0.12) | 0.66(0.10) | 0.54(0.08) | 0.53(0.09) | 0.73(0.12) | 0.73(0.13) | 0.59(0.09) |

# Simulation of $d = 1$

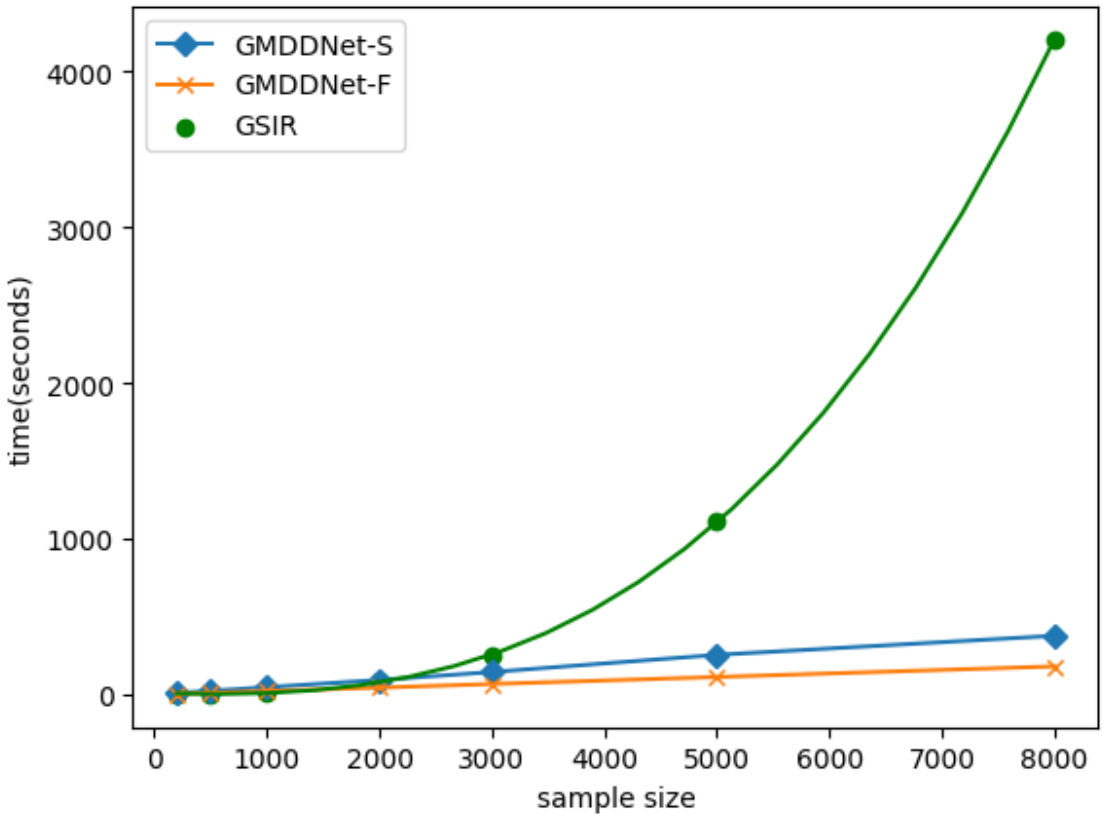



Figure: Time for different methods

# Simulation of $d = 2$

Now consider the case of $d = 2$:

$$\begin{cases} D: & Y = \left((1+X_1)^2 + \epsilon_1, \sin\left((X_4+X_5)\pi/10\right) + X_4^2 + \epsilon_2\right)^\top, \\ E: & Y = \mathbb{I}\{f_1(X) + \epsilon_3 > 0\} + 2 \cdot \mathbb{I}\{f_2(X) + \epsilon_4 > 0\}, \\ F: & Y = \left(f_3(X) + \epsilon_5, f_4(X) + \epsilon_6, \epsilon_7, \cdots, \epsilon_{14}\right)^\top \in \mathbb{R}^{10}, \end{cases}$$

where

$$f_1(X) = \log \frac{(2X_1+1)^2 + (2X_2+1)^2}{(2X_6+1)^2 + (2X_7+1)^2},$$

$$f_2(X) = \log \frac{(2X_4+1)^2 + (2X_5+1)^2}{(2X_6+1)^2 + (2X_7+1)^2},$$

$$f_3(X) = \sqrt{(X_1+1)^2 + (X_2+1)^2 + 1} + \sqrt{X_1^2 + X_2^2 + 1},$$

$$f_4(X) = (1+X_4)^2 + (1+X_5)^2,$$

the distributions for the predictor vector $X$ still adhere to the I, II, III aforementioned and the mutually independent $\epsilon_i \sim N(0, 0.25)$.

# Simulation of $d = 2$

Table: Distance correlation with true predictor at $p = 50, d = 2$ for GMDDNet-S and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| D-I | 1k | 0.80(0.02) | 0.79(0.03) | 0.69(0.02) | 0.68(0.02) | 0.79(0.02) | 0.80(0.02) | 0.79(0.01) |
| | 2k | 0.83(0.01) | 0.83(0.01) | 0.73(0.02) | 0.72(0.01) | 0.82(0.01) | 0.83(0.01) | 0.80(0.01) |
| | 5k | 0.85(0.01) | 0.87(0.01) | 0.78(0.00) | 0.77(0.01) | 0.84(0.00) | 0.88(0.01) | 0.81(0.00) |
| D-II | 1k | 0.91(0.01) | 0.92(0.02) | 0.63(0.02) | 0.62(0.02) | 0.80(0.01) | 0.93(0.01) | 0.84(0.01) |
| | 2k | 0.92(0.01) | 0.94(0.01) | 0.70(0.02) | 0.70(0.02) | 0.84(0.01) | 0.94(0.01) | 0.84(0.01) |
| | 5k | 0.94(0.00) | 0.97(0.01) | 0.74(0.02) | 0.74(0.01) | 0.89(0.01) | 0.97(0.01) | 0.85(0.00) |
| D-III | 1k | 0.84(0.04) | 0.72(0.08) | 0.58(0.03) | 0.54(0.03) | 0.79(0.04) | 0.79(0.05) | 0.58(0.02) |
| | 2k | 0.88(0.05) | 0.73(0.11) | 0.65(0.01) | 0.61(0.02) | 0.83(0.02) | 0.82(0.07) | 0.59(0.02) |
| | 5k | 0.93(0.01) | 0.83(0.04) | 0.68(0.03) | 0.65(0.02) | 0.85(0.02) | 0.91(0.02) | 0.59(0.02) |

# Simulation of $d = 2$

Table: Distance correlation with true predictor at $p = 50, d = 2$ for GMDDNet-S and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| E-I | 1k | —— | —— | —— | —— | —— | 0.46(0.03) | 0.61(0.02) |
| | 2k | —— | —— | —— | —— | —— | 0.55(0.02) | 0.62(0.01) |
| | 5k | —— | —— | —— | —— | —— | 0.64(0.02) | 0.65(0.02) |
| E-II | 1k | —— | —— | —— | —— | —— | 0.59(0.02) | 0.79(0.01) |
| | 2k | —— | —— | —— | —— | —— | 0.61(0.02) | 0.81(0.01) |
| | 5k | —— | —— | —— | —— | —— | 0.68(0.01) | 0.82(0.01) |
| E-III | 1k | —— | —— | —— | —— | —— | 0.42(0.02) | 0.24(0.03) |
| | 2k | —— | —— | —— | —— | —— | 0.55(0.04) | 0.28(0.03) |
| | 5k | —— | —— | —— | —— | —— | 0.65(0.01) | 0.34(0.04) |

# Simulation of $d = 2$

Table: Distance correlation with true predictor at $p = 50, d = 2$ for GMDDNet-S and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| F-I | 1k | 0.76(0.02) | 0.75(0.02) | 0.70(0.02) | 0.60(0.03) | 0.75(0.02) | 0.76(0.02) | 0.79(0.01) |
| | 2k | 0.81(0.01) | 0.81(0.02) | 0.75(0.02) | 0.65(0.02) | 0.80(0.01) | 0.81(0.02) | 0.80(0.01) |
| | 5k | 0.85(0.01) | 0.84(0.01) | 0.82(0.00) | 0.74(0.00) | 0.86(0.01) | 0.85(0.00) | 0.80(0.01) |
| F-II | 1k | 0.80(0.02) | 0.80(0.02) | 0.70(0.02) | 0.55(0.02) | 0.79(0.02) | 0.81(0.02) | 0.83(0.01) |
| | 2k | 0.87(0.01) | 0.84(0.02) | 0.75(0.01) | 0.59(0.02) | 0.84(0.01) | 0.86(0.01) | 0.84(0.01) |
| | 5k | 0.91(0.00) | 0.87(0.01) | 0.83(0.01) | 0.67(0.01) | 0.90(0.00) | 0.89(0.01) | 0.84(0.01) |
| F-III | 1k | 0.85(0.03) | 0.77(0.05) | 0.64(0.03) | 0.52(0.03) | 0.77(0.02) | 0.84(0.03) | 0.64(0.03) |
| | 2k | 0.88(0.02) | 0.80(0.09) | 0.69(0.03) | 0.56(0.02) | 0.81(0.01) | 0.88(0.04) | 0.65(0.02) |
| | 5k | 0.91(0.01) | 0.89(0.01) | 0.74(0.02) | 0.62(0.01) | 0.86(0.00) | 0.92(0.02) | 0.63(0.01) |

# Simulation of $d = 2$

Table: Distance correlation with true predictor when $p = 50, d = 2$ for GMDDNet-F and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| D-I | 1k | 0.74(0.03) | 0.73(0.03) | 0.70(0.04) | 0.70(0.03) | 0.75(0.03) | 0.74(0.03) | 0.79(0.01) |
| | 2k | 0.81(0.03) | 0.78(0.04) | 0.78(0.02) | 0.78(0.01) | 0.83(0.02) | 0.80(0.03) | 0.80(0.01) |
| | 5k | 0.88(0.02) | 0.90(0.01) | 0.82(0.01) | 0.81(0.01) | 0.86(0.01) | 0.89(0.02) | 0.81(0.00) |
| D-II | 1k | 0.84(0.02) | 0.83(0.05) | 0.70(0.03) | 0.68(0.03) | 0.82(0.01) | 0.84(0.04) | 0.84(0.01) |
| | 2k | 0.89(0.02) | 0.87(0.05) | 0.74(0.02) | 0.72(0.02) | 0.87(0.01) | 0.91(0.02) | 0.84(0.01) |
| | 5k | 0.94(0.01) | 0.92(0.04) | 0.77(0.01) | 0.75(0.02) | 0.90(0.00) | 0.94(0.02) | 0.85(0.00) |
| D-III | 1k | 0.78(0.05) | 0.62(0.08) | 0.65(0.02) | 0.62(0.02) | 0.76(0.03) | 0.68(0.07) | 0.58(0.02) |
| | 2k | 0.75(0.18) | 0.51(0.09) | 0.68(0.02) | 0.66(0.02) | 0.82(0.02) | 0.63(0.12) | 0.59(0.02) |
| | 5k | 0.86(0.03) | 0.53(0.09) | 0.70(0.03) | 0.68(0.02) | 0.84(0.02) | 0.69(0.05) | 0.59(0.02) |

# Simulation of $d = 2$

Table: Distance correlation with true predictor when $p = 50, d = 2$ for GMDDNet-F and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| E-I | 1k | —— | —— | —— | —— | —— | 0.48(0.03) | 0.61(0.02) |
| | 2k | —— | —— | —— | —— | —— | 0.56(0.02) | 0.62(0.01) |
| | 5k | —— | —— | —— | —— | —— | 0.70(0.02) | 0.65(0.02) |
| E-II | 1k | —— | —— | —— | —— | —— | 0.61(0.03) | 0.79(0.01) |
| | 2k | —— | —— | —— | —— | —— | 0.65(0.02) | 0.81(0.01) |
| | 5k | —— | —— | —— | —— | —— | 0.70(0.01) | 0.82(0.01) |
| E-III | 1k | —— | —— | —— | —— | —— | 0.29(0.07) | 0.24(0.03) |
| | 2k | —— | —— | —— | —— | —— | 0.45(0.03) | 0.28(0.03) |
| | 5k | —— | —— | —— | —— | —— | 0.62(0.02) | 0.34(0.04) |

# Simulation of $d = 2$

Table: Distance correlation with true predictor when $p = 50, d = 2$ for GMDDNet-F and GSIR.

| $(X, Y)$ | sample size | 1 | 2 | 3 | 4 | 5 | 6 | GSIR |
|---|---|---|---|---|---|---|---|---|
| F-I | 1k | 0.76(0.03) | 0.76(0.03) | 0.72(0.03) | 0.63(0.05) | 0.77(0.03) | 0.77(0.02) | 0.79(0.01) |
| | 2k | 0.80(0.02) | 0.77(0.03) | 0.81(0.01) | 0.75(0.04) | 0.82(0.02) | 0.79(0.03) | 0.80(0.01) |
| | 5k | 0.86(0.01) | 0.85(0.02) | 0.87(0.01) | 0.79(0.02) | 0.89(0.01) | 0.86(0.02) | 0.80(0.01) |
| F-II | 1k | 0.81(0.02) | 0.77(0.03) | 0.72(0.03) | 0.57(0.03) | 0.80(0.02) | 0.79(0.04) | 0.83(0.01) |
| | 2k | 0.84(0.02) | 0.79(0.03) | 0.79(0.02) | 0.65(0.02) | 0.85(0.02) | 0.82(0.02) | 0.84(0.01) |
| | 5k | 0.91(0.01) | 0.85(0.04) | 0.85(0.01) | 0.71(0.03) | 0.90(0.01) | 0.86(0.02) | 0.84(0.01) |
| F-III | 1k | 0.77(0.04) | 0.66(0.06) | 0.72(0.03) | 0.62(0.03) | 0.76(0.03) | 0.69(0.05) | 0.64(0.03) |
| | 2k | 0.79(0.06) | 0.61(0.11) | 0.73(0.03) | 0.64(0.09) | 0.80(0.02) | 0.68(0.12) | 0.65(0.02) |
| | 5k | 0.85(0.02) | 0.65(0.04) | 0.79(0.02) | 0.68(0.03) | 0.85(0.01) | 0.72(0.03) | 0.63(0.01) |

# Dimension determination

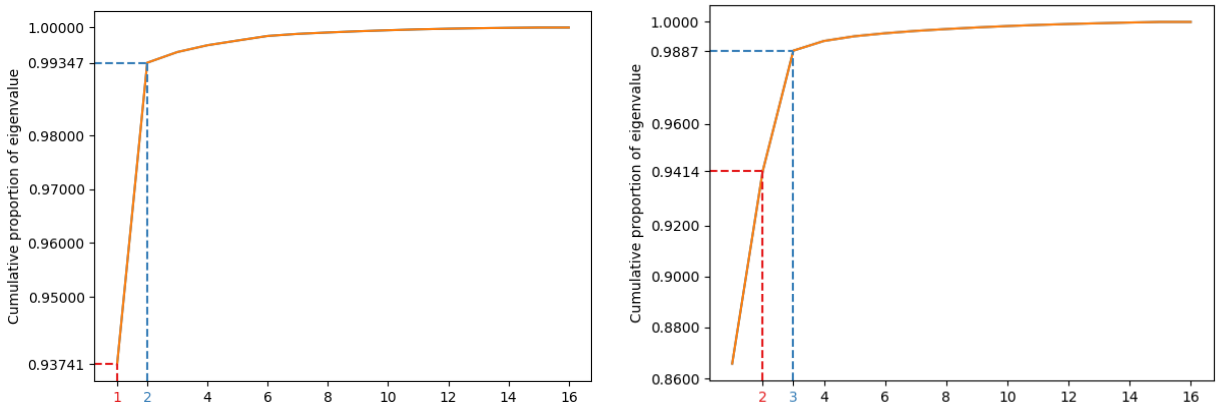



Figure: The eigenvalue ratio plot for for setting A-I with $n = 2000, p = 50, d = 1$ and setting E-I with $n = 2000, p = 50, d = 2$.

# Outline



# Regression

The dataset called "Communities and Crime " consists of $1994$ instances with $127$ features and one response variable, "violent crimes per capita".

Table: Mean square error of different methods on the regression task.

| Method | Indirect methods | | | Direct methods | |
|---|---|---|---|---|---|
| | GMDDNet-S | GMDDNet-F | GSIR | DNN | Linear regression |
| MSE | 0.0170 | 0.0161 | 0.0174 | 0.0195 | $8.9 \times 10^{17}$ |

# Classification

The MNIST database and the Fashion-MNIST both comprise $70,000$ $28 \times 28$ images belonging to ten categories, with $60,000$ images in the training set and $10,000$ in the test set. Target dimension $d = 9$ can be observed below.

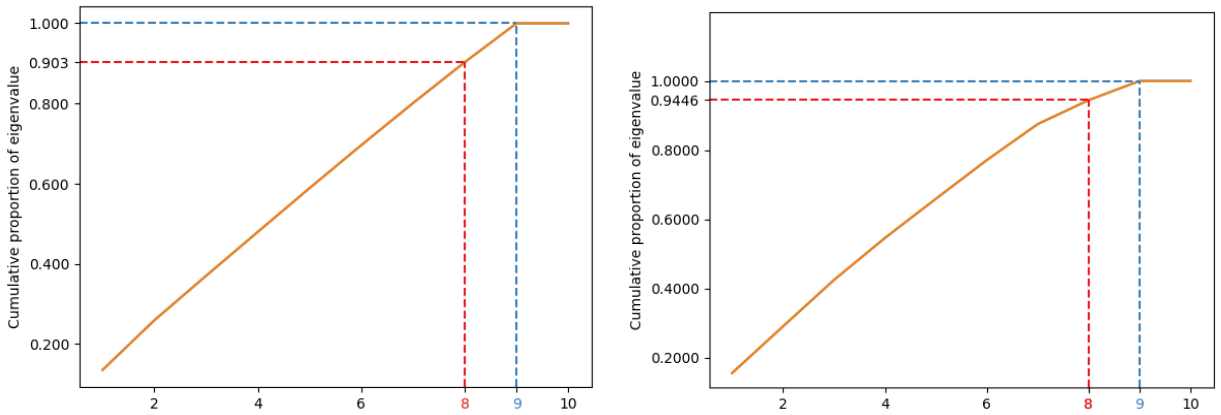



Figure: The eigenvalue ratio plot for MNIST and FashionMNIST

# Classification

We apply the logistic regression to the reduced features of all SDR methods. Accuracy and visualization are shown for comparison.

Table: Accuracy of different methods on MNIST with sample size $5000$.

| Method | DNN | | | LeNet | | | |
|---|---|---|---|---|---|---|---|
| | GMDDNet-S | GMDDNet-F | Direct | GMDDNet-S | GMDDNet-F | Direct | GSIR |
| accuracy | 0.9583 | 0.949 | 0.9534 | 0.9731 | 0.9719 | 0.9708 | 0.8313 |

# 2D visualization

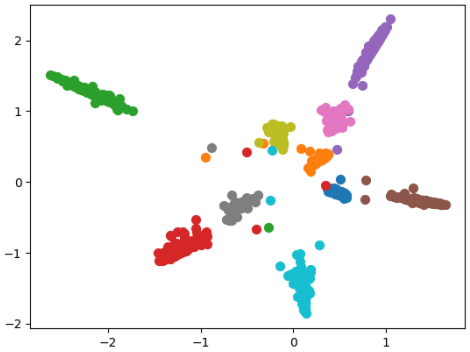



Figure: DNN with norm multiplier

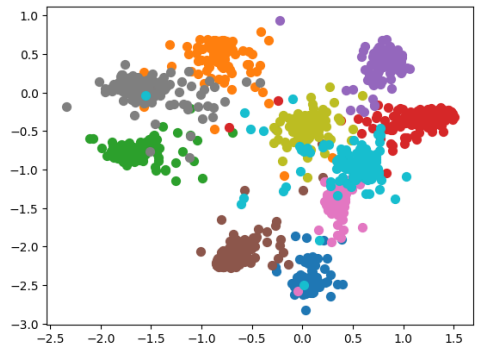



Figure: LeNet with norm multiplier

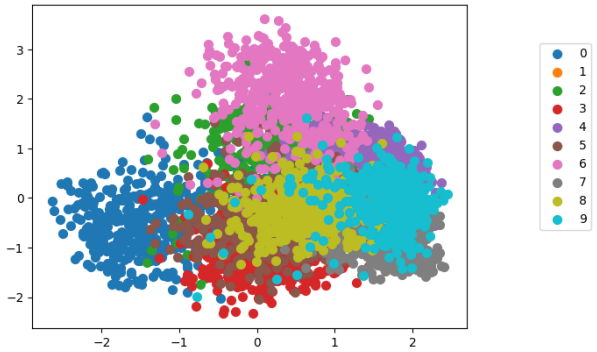



Figure: GSIR

Figure: 2D visualization of MNIST using DNN, LeNet, GSIR with sample size 5k

# 2D visualization

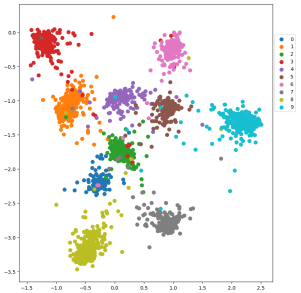


Figure: LeNet

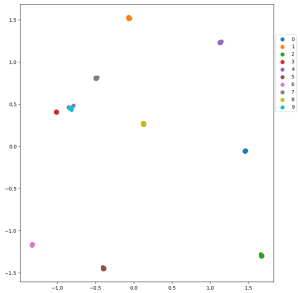


Figure: ResNet

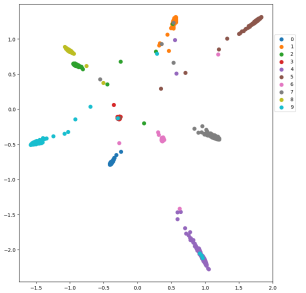


Figure: DenseNet

Figure: 2D-visualization of MNIST using LeNet, ResNet and DenseNet with sample size 60k (Left panel: GMDDNet-F with convolutional neural networks; Right panel: direct convolutional neural networks).

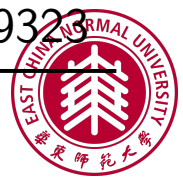
# Other CNNs

Table: Accuracy of different CNNs on MNIST with sample size 60000.

| Method | LeNet | | ResNet18 | | DenseNet | |
|---|---|---|---|---|---|---|
| | GMDDNet-F | Direct | GMDDNet-F | Direct | GMDDNet-F | Direct |
| Accuracy | 0.9909 | 0.988 | 0.9948 | 0.9939 | 0.996 | 0.996 |

Table: Accuracy of different CNNs on FashionMNIST with sample size 60000.

| Method | LeNet | | ResNet18 | | DenseNet | |
|---|---|---|---|---|---|---|
| | GMDDNet-F | Direct | GMDDNet-F | Direct | GMDDNet-F | Direct |
| Accuracy | 0.9143 | 0.9087 | 0.9157 | 0.9148 | 0.9345 | 0.9323 |



# Outline



# Summary

- We propose a novel nonlinear SDR optimization objective from the perspective of GMDD, which is suitable for multidimensional $Y$ and frees us from the restrictive step of slicing response variables in the classical SDR process.
- The nonlinear dimensional reduction function is estimated by a deep ReLU neural network.
- Our method is developed under the classical theoretical framework of SDR and the most basic unbiasedness can be theoretically guaranteed without any stringent assumptions.
- The excess risk nearly attains the minimax rate.
- The validity of our deep nonlinear SDR is demonstrated through simulations and real data.

*Thank You!*

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019).
Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks.
*Journal of Machine Learning Research*, 20(63):1–17.

Bauer, B. and Kohler, M. (2019).
On deep learning as a remedy for the curse of dimensionality in non-parametric regression.
*The Annals of Statistics*, 47(4):2261–2285.

Böttcher, B., Keller-Ressel, M., and Schilling, R. L. (2018).
Detecting independence of random vectors: generalized distance covariance and gaussian covariance.
*Modern Stochastics: Theory and Applications*, 5(3):353–383.

Farrell, M. H., Liang, T., and Misra, S. (2021).
Deep neural networks for estimation and inference.
*Econometrica*, 89(1):181–213.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009).
Kernel dimension reduction in regression.


Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al. (2002).
*A distribution-free theory of nonparametric regression*, volume 1.
Springer.


Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2021).
Deep nonparametric regression on approximately low-dimensional manifolds.
*arXiv preprint arXiv:2104.06708*.


Kohler, M., Krzyżak, A., and Langer, S. (2022).
Estimation of a function of low local dimensionality by deep neural networks.
*IEEE Transactions on Information Theory*.


Lee, K.-Y., Li, B., and Chiaromonte, F. (2013).
A general theory for nonlinear sufficient dimension reduction: Formulation and estimation.
*The Annals of Statistics*, 41(1):221 – 249.


Schmidt-Hieber, J. (2020).


Nonparametric regression using deep neural networks with relu activation function.
*The Annals of Statistics*, 48(4):1875–1897.


Shen, Z. (2020).
Deep network approximation characterized by number of neurons.
*Communications in Computational Physics*, 28(5):1768–1811.


Vaart, A. W. and Wellner, J. A. (1996).
Weak convergence.
In *Weak convergence and empirical processes*, pages 16–28. Springer.