# Sliced Mutual Information: A Scalable Measure of Statistical Dependence

Ziv Goldfeld

Cornell University

Joint work with Kristjan Greenewald, Theshani Nuradha, and Galen Reeves

2022 ITA Workshop

May 24th, 2022

# Mutual Information: Virtues and Challenges

**Definition (Shannon'48)**

The mutual information (MI) between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{I}(X;Y) := \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \log\left(\frac{dP_{XY}}{dP_X \otimes P_Y}\right) dP_{X,Y} = \mathsf{D}_{\mathsf{KL}}(P_{XY} \| P_X \otimes P_Y)$$

# Mutual Information: Virtues and Challenges

**Definition (Shannon'48)**

The mutual information (MI) between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{I}(X;Y) := \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \log\left(\frac{dP_{XY}}{dP_X \otimes P_Y}\right) dP_{X,Y} = \mathsf{D}_{\mathsf{KL}}(P_{XY} \| P_X \otimes P_Y)$$

**Fundamental measure of statistical dependence:**

# Mutual Information: Virtues and Challenges

**Definition (Shannon'48)**

The mutual information (MI) between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{I}(X;Y) := \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \log\left(\frac{dP_{XY}}{dP_X \otimes P_Y}\right) dP_{X,Y} = \mathsf{D}_{\mathsf{KL}}(P_{XY} \| P_X \otimes P_Y)$$

**Fundamental measure of statistical dependence:**

- Meaningful units & structural properties

# Mutual Information: Virtues and Challenges

**Definition (Shannon'48)**

The mutual information (MI) between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{I}(X;Y) := \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \log\left(\frac{dP_{XY}}{dP_X \otimes P_Y}\right) dP_{X,Y} = \mathsf{D}_{\mathsf{KL}}(P_{XY} \| P_X \otimes P_Y)$$

**Fundamental measure of statistical dependence:**

- Meaningful units & structural properties
- Emerges as solution to operational problems

# Mutual Information: Virtues and Challenges

**Definition (Shannon'48)**

The mutual information (MI) between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{I}(X;Y) := \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \log\left(\frac{dP_{XY}}{dP_X \otimes P_Y}\right) dP_{X,Y} = \mathsf{D}_{\mathsf{KL}}(P_{XY} \| P_X \otimes P_Y)$$

**Fundamental measure of statistical dependence:**

- Meaningful units & structural properties
- Emerges as solution to operational problems
- Applications in information theory, statistics, machine learning.

# Mutual Information: Virtues and Challenges

**Definition (Shannon'48)**

The mutual information (MI) between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{I}(X;Y) := \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \log\left(\frac{dP_{XY}}{dP_X \otimes P_Y}\right) dP_{X,Y} = \mathsf{D}_{\mathsf{KL}}(P_{XY} \| P_X \otimes P_Y)$$

**Fundamental measure of statistical dependence:**

- Meaningful units **&** structural properties
- Emerges as solution to operational problems
- Applications in information theory, **statistics, machine learning**.

# Mutual Information: Virtues and Challenges

**Definition (Shannon'48)**

The mutual information (MI) between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{I}(X;Y) := \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \log\left(\frac{dP_{XY}}{dP_X \otimes P_Y}\right) dP_{X,Y} = \mathsf{D}_{\mathsf{KL}}(P_{XY} \| P_X \otimes P_Y)$$

**Fundamental measure of statistical dependence:**

- Meaningful units **&** structural properties
- Emerges as solution to operational problems
- Applications in information theory, **statistics, machine learning**.

**Challenge:** MI estimation in high dim. sample complexity $n^\star(\epsilon, d) \asymp \epsilon^{-d}$

# Mutual Information: Virtues and Challenges

**Definition (Shannon'48)**

The mutual information (MI) between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{I}(X;Y) := \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_y}} \log\left(\frac{dP_{XY}}{dP_X \otimes P_Y}\right) dP_{X,Y} = \mathsf{D}_{\mathsf{KL}}(P_{XY} \| P_X \otimes P_Y)$$

**Fundamental measure of statistical dependence:**

- Meaningful units **&** structural properties
- Emerges as solution to operational problems
- Applications in information theory, **statistics, machine learning**.

**Challenge:** MI estimation in high dim. sample complexity $n^\star(\epsilon, d) \asymp \epsilon^{-d}$

⊛ **Goal:** Scalable MI surrogate that preserves its structure

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI btw. $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is (here $\sigma_d = \mathrm{Unif}(\mathbb{S}^{d-1})$)

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_{d_x}(\theta) d\sigma_{d_y}(\phi).$$

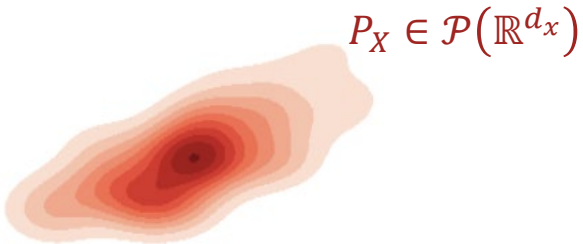
# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI btw. $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is (here $\sigma_d = \mathrm{Unif}(\mathbb{S}^{d-1})$)

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_{d_x}(\theta) d\sigma_{d_y}(\phi).$$

**Illustration:**



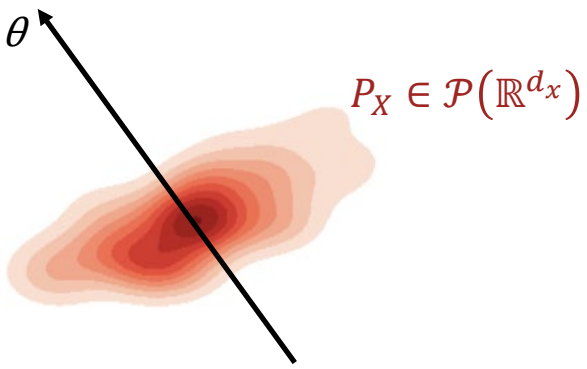
# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI btw. $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is (here $\sigma_d = \mathrm{Unif}(\mathbb{S}^{d-1})$)

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_{d_x}(\theta) d\sigma_{d_y}(\phi).$$

**Illustration:**



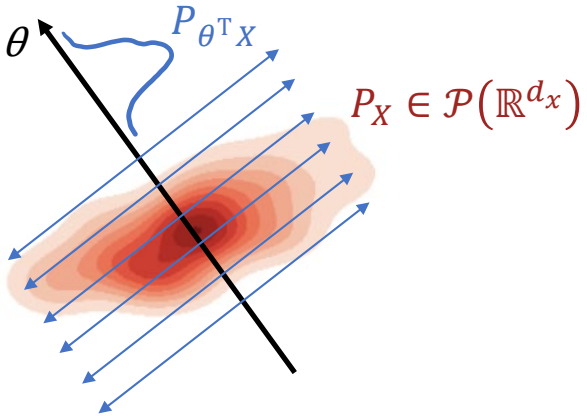
# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI btw. $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is (here $\sigma_d = \mathrm{Unif}(\mathbb{S}^{d-1})$)

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_{d_x}(\theta) d\sigma_{d_y}(\phi).$$

**Illustration:**



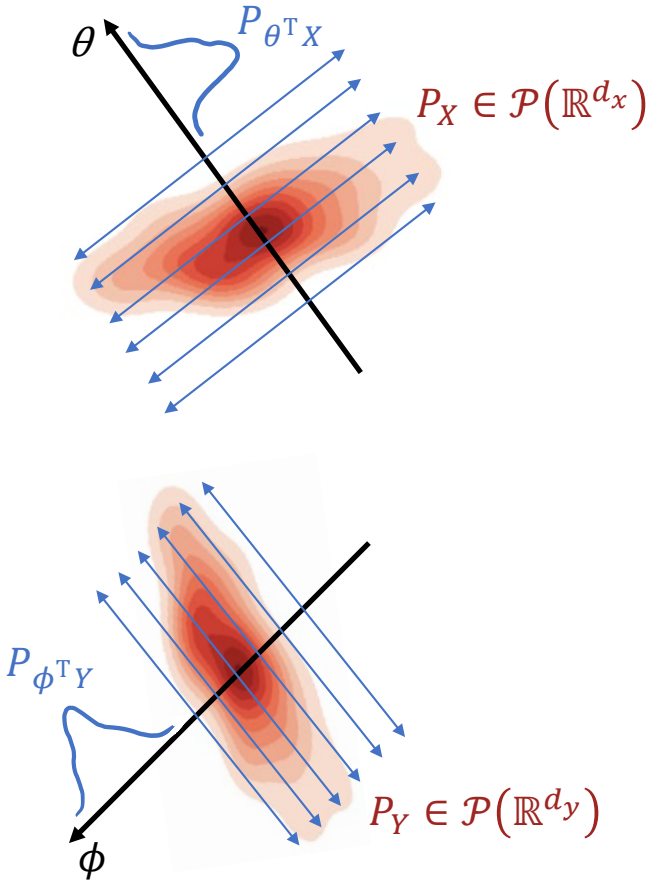
# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI btw. $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is (here $\sigma_d = \mathrm{Unif}(\mathbb{S}^{d-1})$)

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_{d_x}(\theta) d\sigma_{d_y}(\phi).$$

**Illustration:**



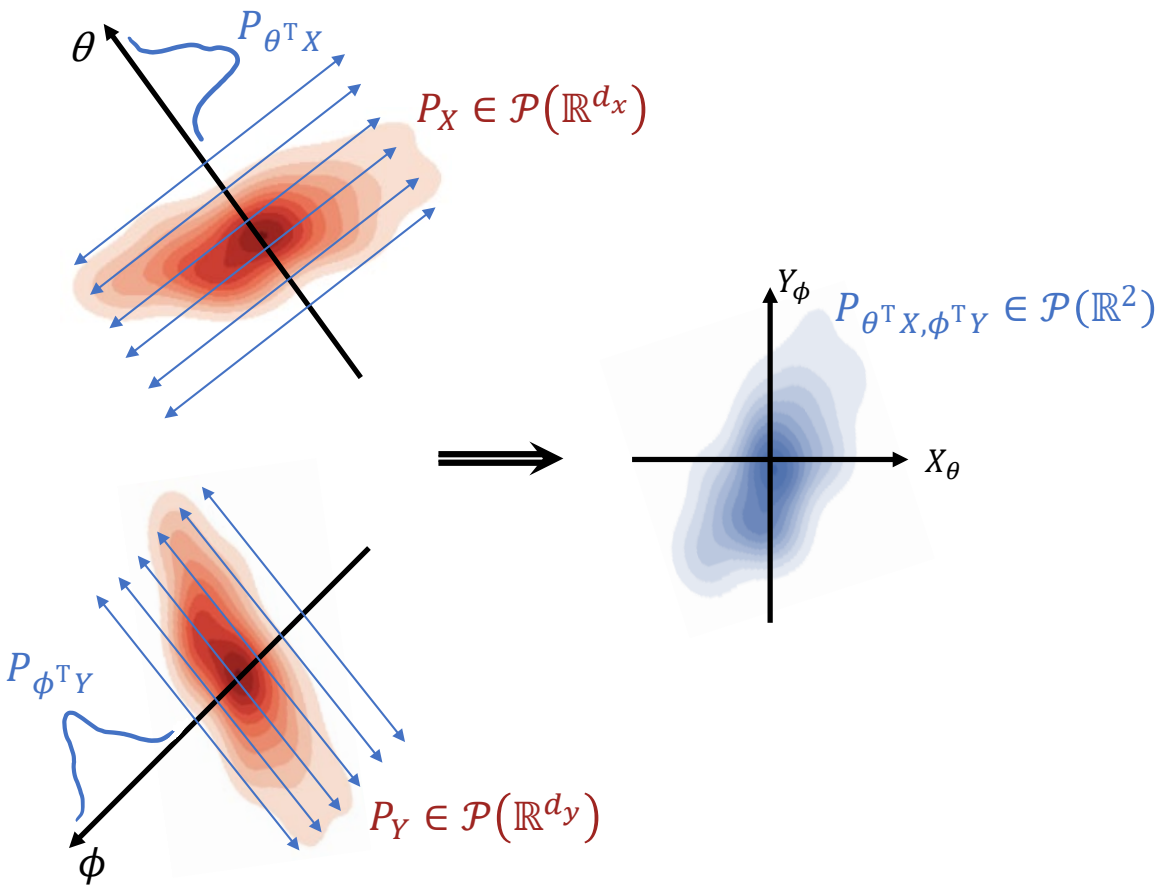
# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI btw. $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is (here $\sigma_d = \mathrm{Unif}(\mathbb{S}^{d-1})$)

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_{d_x}(\theta) d\sigma_{d_y}(\phi).$$

**Illustration:**



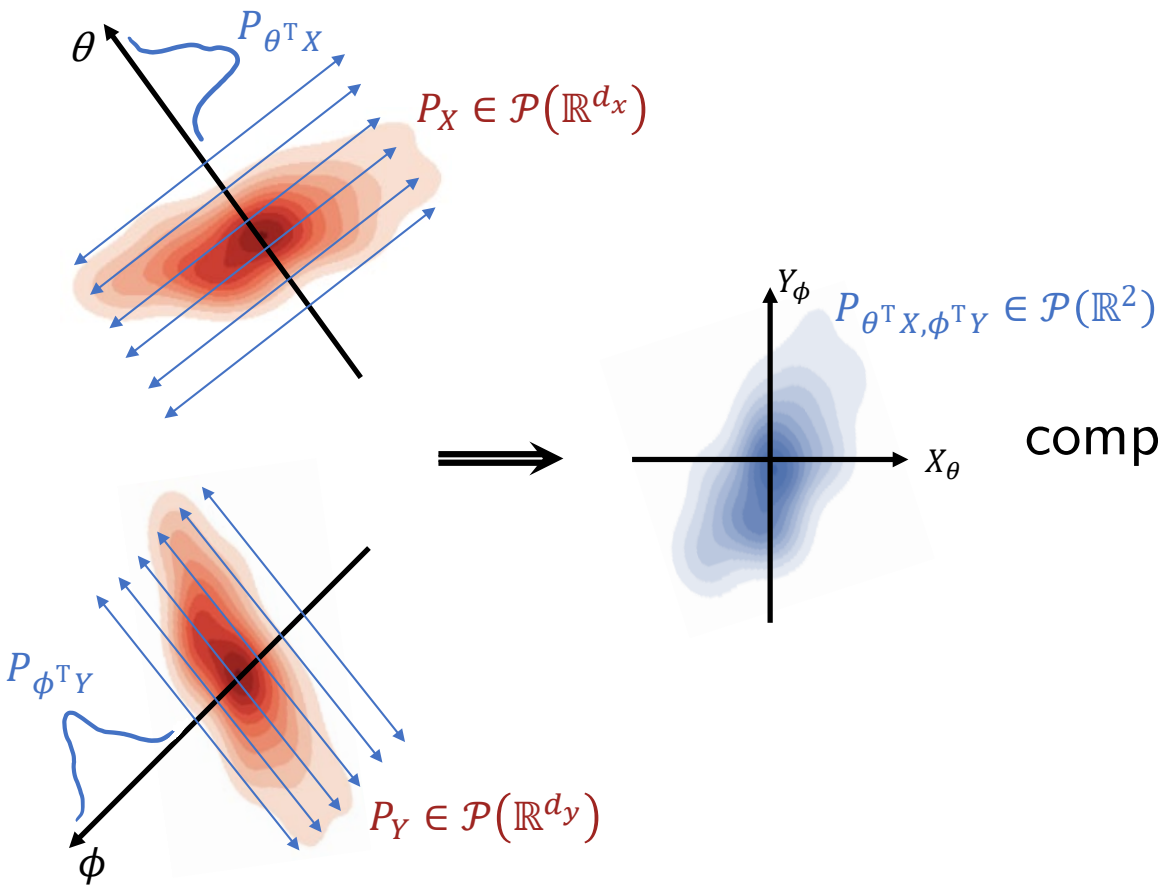
# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI btw. $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is (here $\sigma_d = \mathrm{Unif}(\mathbb{S}^{d-1})$)

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_{d_x}(\theta) d\sigma_{d_y}(\phi).$$

**Illustration:**




compute $\mathsf{I}(\theta^\intercal X; \phi^\intercal Y)$

**&**

average

# Properties of Sliced Mutual Information

**Theorem (ZG-Greenewald'21)**

*SMI preserves many of the properties of MI, including:*

# Properties of Sliced Mutual Information

**Theorem (ZG-Greenewald'21)**

*SMI preserves many of the properties of MI, including:*

1. ***Independence:*** $\mathsf{SI}(X;Y) = 0 \iff X \perp Y$

# Properties of Sliced Mutual Information

## Theorem (ZG-Greenewald'21)

*SMI preserves many of the properties of MI, including:*

1. ***Independence:*** $\mathsf{SI}(X;Y) = 0 \iff X \perp Y$

$$\boxed{\theta^\intercal X \perp \phi^\intercal Y, \quad \forall(\theta,\phi) \in \mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1} \iff X \perp Y}$$

# Properties of Sliced Mutual Information

## Theorem (ZG-Greenewald'21)

*SMI preserves many of the properties of MI, including:*

1. ***Independence:*** $\mathsf{SI}(X;Y) = 0 \iff X \perp Y$

2. ***Sliced entropy:*** $\mathsf{SI}(X;Y) = \mathsf{sh}(X) - \mathsf{sh}(X|Y) = \dots$

$$\textit{where} \quad \mathsf{sh}(X) := \mathsf{h}(\Theta^\intercal X|\Theta)$$

# Properties of Sliced Mutual Information

## Theorem (ZG-Greenewald'21)

*SMI preserves many of the properties of MI, including:*

1. ***Independence:*** $\mathsf{SI}(X;Y) = 0 \iff X \perp Y$

2. ***Sliced entropy:*** $\mathsf{SI}(X;Y) = \mathsf{sh}(X) - \mathsf{sh}(X|Y) = \dots$

   *where* $\mathsf{sh}(X) := \mathsf{h}(\Theta^\intercal X|\Theta)$

3. ***Max sliced entropy:***
   - $\Sigma_P \preccurlyeq \Sigma \implies P^\star = \mathcal{N}(0, \Sigma)$
   - $\mathrm{spt}(P) \subseteq \mathbb{B}_d(r) \implies P^\star = \mathrm{Unif}(r\mathbb{S}^{d-1})$

# Properties of Sliced Mutual Information

## Theorem (ZG-Greenewald'21)

*SMI preserves many of the properties of MI, including:*

1. ***Independence:*** $\mathsf{SI}(X;Y) = 0 \iff X \perp Y$

2. ***Sliced entropy:*** $\mathsf{SI}(X;Y) = \mathsf{sh}(X) - \mathsf{sh}(X|Y) = \dots$

   *where* $\mathsf{sh}(X) := \mathsf{h}(\Theta^\intercal X|\Theta)$

3. ***Max sliced entropy:***
   - $\Sigma_P \preccurlyeq \Sigma \implies P^\star = \mathcal{N}(0,\Sigma)$
   - $\mathrm{spt}(P) \subseteq \mathbb{B}_d(r) \implies P^\star = \mathrm{Unif}(r\mathbb{S}^{d-1})$

4. ***Tensorization:*** $(X_1,Y_1),\dots,(X_n,Y_n)$ *are mutually independent*

   $\implies \mathsf{SI}(X_1,\dots,X_n;Y_1,\dots,Y_n) = \sum_{i=1}^n \mathsf{SI}(X_i;Y_i)$

# Properties of Sliced Mutual Information

## Theorem (ZG-Greenewald'21)

*SMI preserves many of the properties of MI, including:*

1. ***Independence:*** $\mathsf{SI}(X;Y) = 0 \iff X \perp Y$

2. ***Sliced entropy:*** $\mathsf{SI}(X;Y) = \mathsf{sh}(X) - \mathsf{sh}(X|Y) = \dots$

   *where* $\mathsf{sh}(X) := \mathsf{h}(\Theta^\intercal X|\Theta)$

3. ***Max sliced entropy:***
   - $\Sigma_P \preccurlyeq \Sigma \implies P^\star = \mathcal{N}(0,\Sigma)$
   - $\mathrm{spt}(P) \subseteq \mathbb{B}_d(r) \implies P^\star = \mathrm{Unif}(r\mathbb{S}^{d-1})$

4. ***Tensorization:*** $(X_1,Y_1),\dots,(X_n,Y_n)$ *are mutually independent*

   $\implies \mathsf{SI}(X_1,\dots,X_n;Y_1,\dots,Y_n) = \sum_{i=1}^n \mathsf{SI}(X_i;Y_i)$

5. ***DV:*** $\mathsf{SI}(X;Y) = \sup_f \mathbb{E}\big[f(\Theta,\Phi,\Theta^\intercal X,\Phi^\intercal Y)\big] - \log\Big(\mathbb{E}\Big[e^{f(\Theta,\Phi,\Theta^\intercal \tilde{X},\Phi^\intercal \tilde{Y})}\Big]\Big)$

# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

# Sliced Mutual Information & Processing

<u>**Mutual information:**</u> Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\Longrightarrow$ MI **cannot grow** via processing

# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\Longrightarrow$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\Longrightarrow$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs

# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\implies$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs
- If $f(X)$ has more informative projections: $\mathsf{SI}(X;Y) < \mathsf{SI}\big(f(X);Y\big)$

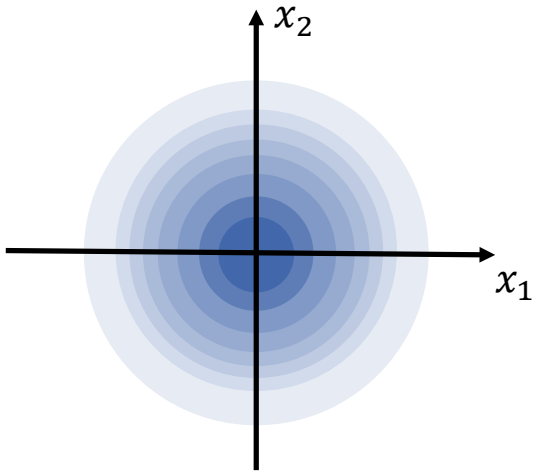
# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\Longrightarrow$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs
- If $f(X)$ has more informative projections: $\mathsf{SI}(X;Y) < \mathsf{SI}\big(f(X);Y\big)$
- **Example:** $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$



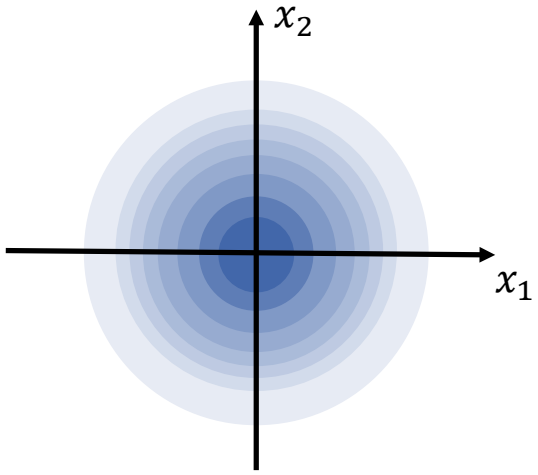
# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\Longrightarrow$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs
- If $f(X)$ has more informative projections: $\mathsf{SI}(X;Y) < \mathsf{SI}\big(f(X);Y\big)$
- **Example:** $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = X_1$



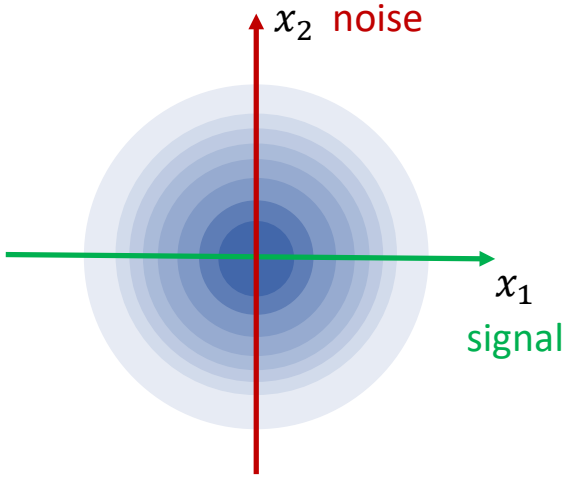
# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\Longrightarrow$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs
- If $f(X)$ has more informative projections: $\mathsf{SI}(X;Y) < \mathsf{SI}\big(f(X);Y\big)$
- **Example:** $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = X_1$



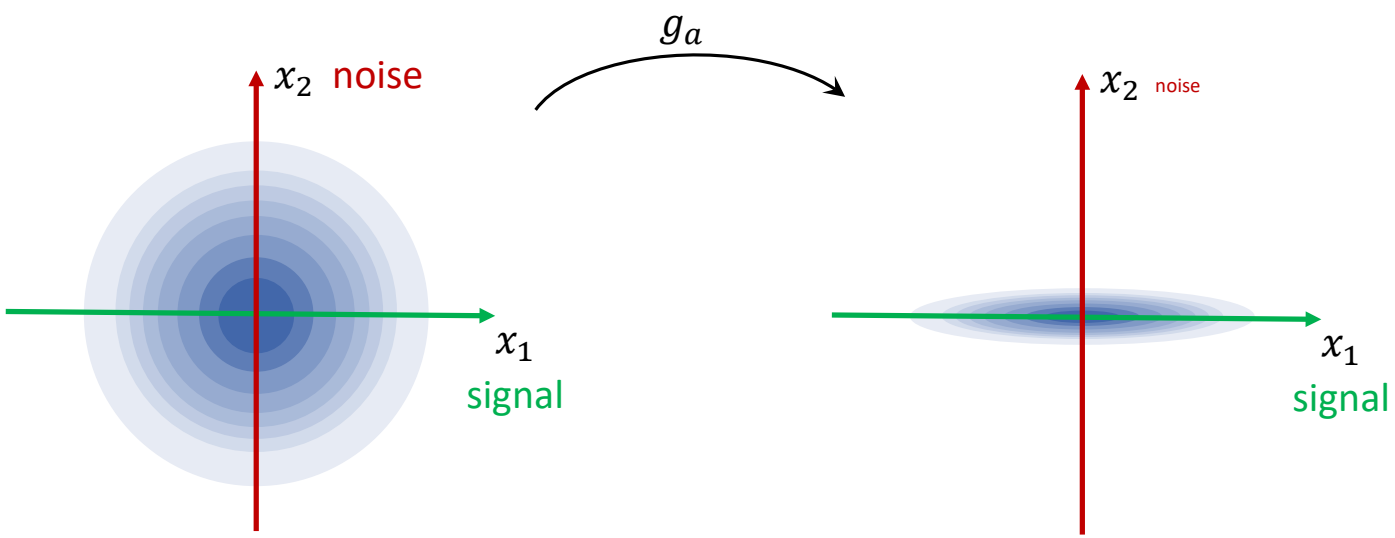
# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}(f(X);Y)$

$\Longrightarrow$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs
- If $f(X)$ has more informative projections: $\mathsf{SI}(X;Y) < \mathsf{SI}(f(X);Y)$
- **Example:** $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = X_1$, $g_a(x_1, x_2) = (x_1 \; a x_2)^\intercal$



# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\implies$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs
- If $f(X)$ has more informative projections: $\mathsf{SI}(X;Y) < \mathsf{SI}\big(f(X);Y\big)$
- **Example:** $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = X_1$, $g_a(x_1, x_2) = (x_1 \; ax_2)^\intercal$

$$\implies \mathsf{SI}(X;Y) < \mathsf{SI}\big(g_a(X);Y\big), \quad \forall a < 1$$

# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}\big(f(X);Y\big)$

$\Longrightarrow$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs
- If $f(X)$ has more informative projections: $\mathsf{SI}(X;Y) < \mathsf{SI}\big(f(X);Y\big)$
- **Example:** $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = X_1$, $g_a(x_1, x_2) = (x_1 \; ax_2)^\intercal$

$$\Longrightarrow \mathsf{SI}(X;Y) < \mathsf{SI}\big(g_a(X);Y\big), \quad \forall a < 1$$

$\Longrightarrow$ SMI **can increase** via processing (violates DPI)

# Sliced Mutual Information & Processing

**Mutual information:** Satisfies DPI $\quad \mathsf{I}(X;Y) \geq \mathsf{I}(f(X);Y)$

$\Longrightarrow$ MI **cannot grow** via processing

**Sliced mutual information:** Follows the DPI?

- Only considers projections of RVs
- If $f(X)$ has more informative projections: $\mathsf{SI}(X;Y) < \mathsf{SI}(f(X);Y)$
- **Example:** $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = X_1$, $g_a(x_1, x_2) = (x_1 \; a x_2)^\intercal$

$$\Longrightarrow \mathsf{SI}(X;Y) < \mathsf{SI}(g_a(X);Y), \quad \forall a < 1$$

$\Longrightarrow$ SMI **can increase** via processing (violates DPI)

**Can be used for feature extraction via SMI maximization**

# Scalable Estimation from Samples

**Estimator:** Given samples $(X^n, Y^n)$ i.i.d. from $P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$

# Scalable Estimation from Samples

**Estimator:** Given samples $(X^n, Y^n)$ i.i.d. from $P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$

1. Take scalar MI estimator $\hat{\mathsf{I}}(A^n, B^n)$ w/ error $\delta(n)$ over some class

# Scalable Estimation from Samples

**Estimator:** Given samples $(X^n, Y^n)$ i.i.d. from $P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$

1. Take scalar MI estimator $\hat{\mathsf{I}}(A^n, B^n)$ w/ error $\delta(n)$ over some class
2. Sample $m$ random directions $\Theta^m$ and $\Phi^m$ from corresponding spheres

# Scalable Estimation from Samples

**Estimator:** Given samples $(X^n, Y^n)$ i.i.d. from $P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$

1. Take scalar MI estimator $\hat{\mathsf{I}}(A^n, B^n)$ w/ error $\delta(n)$ over some class
2. Sample $m$ random directions $\Theta^m$ and $\Phi^m$ from corresponding spheres
3. Compute $(\Theta_i^\intercal X)^n := (\Theta_i^\intercal X_1, \ldots, \Theta_i^\intercal X_n)$, $i \in [m]$, and $(\Phi_i^\intercal Y)^n$.

# Scalable Estimation from Samples

**Estimator:** Given samples $(X^n, Y^n)$ i.i.d. from $P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$

1. Take scalar MI estimator $\hat{\mathsf{I}}(A^n, B^n)$ w/ error $\delta(n)$ over some class
2. Sample $m$ random directions $\Theta^m$ and $\Phi^m$ from corresponding spheres
3. Compute $(\Theta_i^\intercal X)^n := (\Theta_i^\intercal X_1, \ldots, \Theta_i^\intercal X_n)$, $i \in [m]$, and $(\Phi_i^\intercal Y)^n$.
4. **Estimate SMI via MC:** $$\boxed{\widehat{\mathsf{SI}}_{m,n} := \frac{1}{m} \sum_{i=1}^{m} \hat{\mathsf{I}}((\Theta_i^\intercal X)^n, (\Phi_i^\intercal Y)^n)}$$

# Scalable Estimation from Samples

**Estimator:** Given samples $(X^n, Y^n)$ i.i.d. from $P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$

1. Take scalar MI estimator $\hat{\mathsf{I}}(A^n, B^n)$ w/ error $\delta(n)$ over some class
2. Sample $m$ random directions $\Theta^m$ and $\Phi^m$ from corresponding spheres
3. Compute $(\Theta_i^\intercal X)^n := (\Theta_i^\intercal X_1, \ldots, \Theta_i^\intercal X_n)$, $i \in [m]$, and $(\Phi_i^\intercal Y)^n$.
4. **Estimate SMI via MC:** $$\boxed{\widehat{\mathsf{SI}}_{m,n} := \frac{1}{m} \sum_{i=1}^{m} \hat{\mathsf{I}}((\Theta_i^\intercal X)^n, (\Phi_i^\intercal Y)^n)}$$

**Theorem (ZG-Greenewald-Reeves'22)**

*If $P_{XY}$ has finite 2nd moments & Fisher information $\mathsf{J}(P_{XY}) < \infty$, then*

$$\mathbb{E}\left[\left|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}_{m,n}\right|\right] \leq C(P_{XY})\sqrt{\frac{d_x + d_y}{d_x d_y}} m^{-\frac{1}{2}} + \delta(n),$$

*where* $C(P_{XY}) = 21\sqrt{\|\mathsf{J}_F(P_{XY})\|_{\mathrm{op}}(\|\Sigma_X\|_{\mathrm{op}} \vee \|\Sigma_Y\|_{\mathrm{op}})}$.

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta, \phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^{m} \mathsf{I}_{XY}(\Theta_i, \Phi_i)$

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta, \phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^{m} \mathsf{I}_{XY}(\Theta_i, \Phi_i)$

**Decompose:** $$\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \leq \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$$

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta, \phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^{m} \mathsf{I}_{XY}(\Theta_i, \Phi_i)$

**Decompose:** $$\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \leq \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$$

**Bound:** For each term

$$\textbf{(I)} \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i, \Phi_i) - \mathsf{I}_{XY}(\Theta_i, \Phi_i)\right|\right] \leq \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta, \phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^{m} \mathsf{I}_{XY}(\Theta_i, \Phi_i)$

**Decompose:** $$\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \leq \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$$

**Bound:** For each term

$$\textbf{(I)} \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i, \Phi_i) - \mathsf{I}_{XY}(\Theta_i, \Phi_i)\right|\right] \leq \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

$$\textbf{(II)} \leq m^{-1/2}\sqrt{\mathrm{Var}\big(\mathsf{I}_{XY}(\Theta, \Phi)\big)} \qquad \Big(\mathbb{E}\left[\circledast\right] = \mathsf{SI}(X;Y) \ \textbf{\&}\ L^1(P) \leq L^2(P)\Big)$$

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta,\phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^{m} \mathsf{I}_{XY}(\Theta_i,\Phi_i)$

**Decompose:** $\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \le \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$

**Bound:** For each term

$$\textbf{(I)} \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i,\Phi_i) - \mathsf{I}_{XY}(\Theta_i,\Phi_i)\right|\right] \le \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

$$\textbf{(II)} \le m^{-1/2}\sqrt{\mathrm{Var}\big(\mathsf{I}_{XY}(\Theta,\Phi)\big)} \qquad \Big(\mathbb{E}\left[\circledast\right] = \mathsf{SI}(X;Y) \ \textbf{\&}\ L^1(P) \le L^2(P)\Big)$$

**Variance:** New continuity result $\mathsf{h}(P) - \mathsf{h}(Q) \le \sqrt{\mathsf{J}(Q)}\,\mathsf{W}_2(P,Q)$ (via HWI)

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta,\phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^{m} \mathsf{I}_{XY}(\Theta_i,\Phi_i)$

**Decompose:** $\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \le \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$

**Bound:** For each term

$$\textbf{(I)} \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i,\Phi_i) - \mathsf{I}_{XY}(\Theta_i,\Phi_i)\right|\right] \le \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

$$\textbf{(II)} \le m^{-1/2}\sqrt{\mathrm{Var}\big(\mathsf{I}_{XY}(\Theta,\Phi)\big)} \qquad \Big(\mathbb{E}\left[\circledast\right] = \mathsf{SI}(X;Y) \ \textbf{\&}\ L^1(P) \le L^2(P)\Big)$$

**Variance:** New continuity result $\mathsf{h}(P) - \mathsf{h}(Q) \le \sqrt{\mathsf{J}(Q)}\,\mathsf{W}_2(P,Q)$ (via HWI)

Compare to [Polyanskiy-Wu'16]

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta,\phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^{m} \mathsf{I}_{XY}(\Theta_i,\Phi_i)$

**Decompose:** $\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \le \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$

**Bound:** For each term

$$\textbf{(I)} \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i,\Phi_i) - \mathsf{I}_{XY}(\Theta_i,\Phi_i)\right|\right] \le \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

$$\textbf{(II)} \le m^{-1/2}\sqrt{\mathrm{Var}\big(\mathsf{I}_{XY}(\Theta,\Phi)\big)} \qquad \Big(\mathbb{E}\left[\circledast\right] = \mathsf{SI}(X;Y) \ \textbf{\&}\ L^1(P) \le L^2(P)\Big)$$

**Variance:** New continuity result $\mathsf{h}(P) - \mathsf{h}(Q) \le \sqrt{\mathsf{J}(Q)}\,\mathsf{W}_2(P,Q)$ (via HWI)

Compare to [Polyanskiy-Wu'16]

- Relax $(c_1, c_2)$-regularity

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta,\phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^m \mathsf{I}_{XY}(\Theta_i,\Phi_i)$

**Decompose:** $$\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \le \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$$

**Bound:** For each term

$$\textbf{(I)} \le \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i,\Phi_i) - \mathsf{I}_{XY}(\Theta_i,\Phi_i)\right|\right] \le \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

$$\textbf{(II)} \le m^{-1/2}\sqrt{\mathrm{Var}\big(\mathsf{I}_{XY}(\Theta,\Phi)\big)} \qquad \Big(\mathbb{E}\left[\circledast\right] = \mathsf{SI}(X;Y) \ \textbf{\&}\ L^1(P) \le L^2(P)\Big)$$

**Variance:** New continuity result $\mathsf{h}(P) - \mathsf{h}(Q) \le \sqrt{\mathsf{J}(Q)}\,\mathsf{W}_2(P,Q)$ (via HWI)

Compare to [Polyanskiy-Wu'16]

- Relax $(c_1, c_2)$-regularity
- Sharp constant (optimal)

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta,\phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^{m} \mathsf{I}_{XY}(\Theta_i,\Phi_i)$

**Decompose:** $\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \le \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$

**Bound:** For each term

$$\textbf{(I)} \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i,\Phi_i) - \mathsf{I}_{XY}(\Theta_i,\Phi_i)\right|\right] \le \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

$$\textbf{(II)} \le m^{-1/2}\sqrt{\mathrm{Var}\big(\mathsf{I}_{XY}(\Theta,\Phi)\big)} \qquad \Big(\mathbb{E}\left[\circledast\right] = \mathsf{SI}(X;Y) \ \textbf{\&}\ L^1(P) \le L^2(P)\Big)$$

**Variance:** New continuity result $\mathsf{h}(P) - \mathsf{h}(Q) \le \sqrt{\mathsf{J}(Q)}\,\mathsf{W}_2(P,Q)$ (via HWI)

1. $\mathsf{I}_{XY}$ is Lipschitz on $\mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta,\phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^m \mathsf{I}_{XY}(\Theta_i,\Phi_i)$

**Decompose:** $\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \le \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$

**Bound:** For each term

$$\textbf{(I)} \le \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i,\Phi_i) - \mathsf{I}_{XY}(\Theta_i,\Phi_i)\right|\right] \le \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

$$\textbf{(II)} \le m^{-1/2}\sqrt{\mathrm{Var}\big(\mathsf{I}_{XY}(\Theta,\Phi)\big)} \qquad \Big(\mathbb{E}\left[\circledast\right] = \mathsf{SI}(X;Y) \ \textbf{\&}\ L^1(P) \le L^2(P)\Big)$$

**Variance:** New continuity result $\mathsf{h}(P) - \mathsf{h}(Q) \le \sqrt{\mathsf{J}(Q)}\mathsf{W}_2(P,Q)$ (via HWI)

1. $\mathsf{I}_{XY}$ is Lipschitz on $\mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$
2. Concentration of Lip. functions on $\mathbb{S}^{d-1}$ **&** Efron-Stein-Steele ineq.

# Estimation Error Bound: Proof Outline

**Define:** $\mathsf{I}_{XY}(\theta,\phi) := \mathsf{I}(\theta^\intercal X, \phi^\intercal Y)$ **&** $\circledast := \frac{1}{m}\sum_{i=1}^m \mathsf{I}_{XY}(\Theta_i,\Phi_i)$

**Decompose:** $\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \mathsf{SI}(X;Y)\big|\Big] \le \underbrace{\mathbb{E}\Big[\big|\widehat{\mathsf{SI}}_{m,n} - \circledast\big|\Big]}_{\textbf{(I)}} + \underbrace{\mathbb{E}\Big[\big|\circledast - \mathsf{SI}(X;Y)\big|\Big]}_{\textbf{(II)}}$

**Bound:** For each term

$$\textbf{(I)} \le \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left|\hat{\mathsf{I}}_{XY}(\Theta_i,\Phi_i) - \mathsf{I}_{XY}(\Theta_i,\Phi_i)\right|\right] \le \delta(n) \qquad \Big(\text{assumption on } \hat{\mathsf{I}}\Big)$$

$$\textbf{(II)} \le m^{-1/2}\sqrt{\mathrm{Var}\big(\mathsf{I}_{XY}(\Theta,\Phi)\big)} \qquad \Big(\mathbb{E}\left[\circledast\right] = \mathsf{SI}(X;Y) \ \textbf{\&}\ L^1(P) \le L^2(P)\Big)$$

**Variance:** New continuity result $\mathsf{h}(P) - \mathsf{h}(Q) \le \sqrt{\mathsf{J}(Q)}\mathsf{W}_2(P,Q)$ (via HWI)

1. $\mathsf{I}_{XY}$ is Lipschitz on $\mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$
2. Concentration of Lip. functions on $\mathbb{S}^{d-1}$ **&** Efron-Stein-Steele ineq.

$$\Longrightarrow \quad \mathrm{Var}\big(\mathsf{I}_{XY}(\Theta,\Phi)\big) \le C(P_{XY})^2\big(d_x^{-1} + d_y^{-1}\big)$$

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

**KDE & BPA:** Minimax optimal $\mathsf{h}(X)$ est. over Lipschitz balls [Han et al.'20]

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

**KDE & BPA:** Minimax optimal $\mathsf{h}(X)$ est. over Lipschitz balls [Han et al.'20]

- Plug-in KDE of $f_X$ into best polynomial approximation of $\mathsf{h}(X)$

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

**KDE & BPA:** Minimax optimal $\mathsf{h}(X)$ est. over Lipschitz balls [Han et al.'20]

- Plug-in KDE of $f_X$ into best polynomial approximation of $\mathsf{h}(X)$

$$\Longrightarrow \quad \boxed{\mathbb{E}\Big[\Big|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}_{m,n}^{(\mathsf{Lip})}\Big|\Big] \lesssim m^{-1/2} + n^{-1/2}\big(1 + (\log n)^{1/4}\big)}$$

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

**KDE & BPA:** Minimax optimal $\mathsf{h}(X)$ est. over Lipschitz balls [Han et al.'20]

- Plug-in KDE of $f_X$ into best polynomial approximation of $\mathsf{h}(X)$

$$\Longrightarrow \quad \boxed{\mathbb{E}\Big[\Big|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}_{m,n}^{(\mathsf{Lip})}\Big|\Big] \lesssim m^{-1/2} + n^{-1/2}\big(1 + (\log n)^{1/4}\big)}$$

**Neural est.:** Minimax optimal $\mathsf{D}_{\mathsf{KL}}$ est. over Barron class [Sreekumar-ZG'22]

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

**KDE & BPA:** Minimax optimal $\mathsf{h}(X)$ est. over Lipschitz balls [Han et al.'20]

- Plug-in KDE of $f_X$ into best polynomial approximation of $\mathsf{h}(X)$

$$\Longrightarrow \quad \boxed{\mathbb{E}\Big[\Big|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}_{m,n}^{(\mathsf{Lip})}\Big|\Big] \lesssim m^{-1/2} + n^{-1/2}\big(1 + (\log n)^{1/4}\big)}$$

**Neural est.:** Minimax optimal $\mathsf{D}_{\mathsf{KL}}$ est. over Barron class [Sreekumar-ZG'22]

- $\mathsf{I}(X;Y) = \sup_f \mathbb{E}[f(X,Y)] - \log\left(\mathbb{E}[e^{f(\tilde{X},\tilde{Y})}]\right)$

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

**KDE & BPA:** Minimax optimal $\mathsf{h}(X)$ est. over Lipschitz balls [Han et al.'20]

- Plug-in KDE of $f_X$ into best polynomial approximation of $\mathsf{h}(X)$

$$\Longrightarrow \quad \boxed{\mathbb{E}\Big[\Big|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}_{m,n}^{(\mathsf{Lip})}\Big|\Big] \lesssim m^{-1/2} + n^{-1/2}\big(1 + (\log n)^{1/4}\big)}$$

**Neural est.:** Minimax optimal $\mathsf{D}_{\mathsf{KL}}$ est. over Barron class [Sreekumar-ZG'22]

- $$\mathsf{I}(X;Y) = \sup_f \mathbb{E}\big[f(X,Y)\big] - \log\big(\mathbb{E}[e^{f(\tilde{X},\tilde{Y})}]\big)$$
$$\approx \sup_{g \in \mathcal{F}_{\mathsf{nn}}^k} \frac{1}{n}\sum_{i=1}^n g(X_i, Y_i) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g(X_{\sigma(i)}, Y_i)}\right)$$

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

**KDE & BPA:** Minimax optimal $\mathsf{h}(X)$ est. over Lipschitz balls [Han et al.'20]

- Plug-in KDE of $f_X$ into best polynomial approximation of $\mathsf{h}(X)$

$$\Longrightarrow \quad \boxed{\mathbb{E}\Big[\Big|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}^{(\mathsf{Lip})}_{m,n}\Big|\Big] \lesssim m^{-1/2} + n^{-1/2}\big(1 + (\log n)^{1/4}\big)}$$

**Neural est.:** Minimax optimal $\mathsf{D}_{\mathsf{KL}}$ est. over Barron class [Sreekumar-ZG'22]

- $$\mathsf{I}(X;Y) = \sup_f \mathbb{E}\big[f(X,Y)\big] - \log\big(\mathbb{E}[e^{f(\tilde{X},\tilde{Y})}]\big)$$
$$\approx \sup_{g \in \mathcal{F}^k_{\mathsf{nn}}} \frac{1}{n}\sum_{i=1}^n g(X_i, Y_i) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g(X_{\sigma(i)}, Y_i)}\right)$$

$$\Longrightarrow \quad \boxed{\mathbb{E}\Big[\Big|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}^{(\mathsf{NE})}_{k,m,n}\Big|\Big] \lesssim m^{-1/2} + k^{-1/2} + n^{-1/2}}$$

# SMI Estimation: Examples

**Upshot:** The SMI estimator attains error $\lesssim m^{-1/2} + \delta(n)$

**KDE & BPA:** Minimax optimal $\mathsf{h}(X)$ est. over Lipschitz balls [Han et al.'20]

- Plug-in KDE of $f_X$ into best polynomial approximation of $\mathsf{h}(X)$

$$\Longrightarrow \boxed{\mathbb{E}\Big[\Big|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}^{(\mathsf{Lip})}_{m,n}\Big|\Big] \lesssim m^{-1/2} + n^{-1/2}\big(1 + (\log n)^{1/4}\big)}$$

**Neural est.:** Minimax optimal $\mathsf{D}_{\mathsf{KL}}$ est. over Barron class [Sreekumar-ZG'22]

- $$\mathsf{I}(X;Y) = \sup_f \mathbb{E}\big[f(X,Y)\big] - \log\big(\mathbb{E}[e^{f(\tilde{X},\tilde{Y})}]\big)$$
$$\approx \sup_{g \in \mathcal{F}^k_{\mathsf{nn}}} \frac{1}{n}\sum_{i=1}^n g(X_i, Y_i) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g(X_{\sigma(i)}, Y_i)}\right)$$

$$\Longrightarrow \boxed{\mathbb{E}\Big[\Big|\mathsf{SI}(X;Y) - \widehat{\mathsf{SI}}^{(\mathsf{NE})}_{k,m,n}\Big|\Big] \lesssim m^{-1/2} + k^{-1/2} + n^{-1/2}}$$

⊛ **No curse of dimensionality:** Compare to classic MI rate $n^{-1/(d_x+d_y)}$

# Experiments: Independence Testing

**Recall:** $\mathsf{SI}(X;Y) = 0 \iff (X,Y)$ independent

# Experiments: Independence Testing

**Recall:** $\mathsf{SI}(X;Y) = 0 \iff (X,Y)$ independent

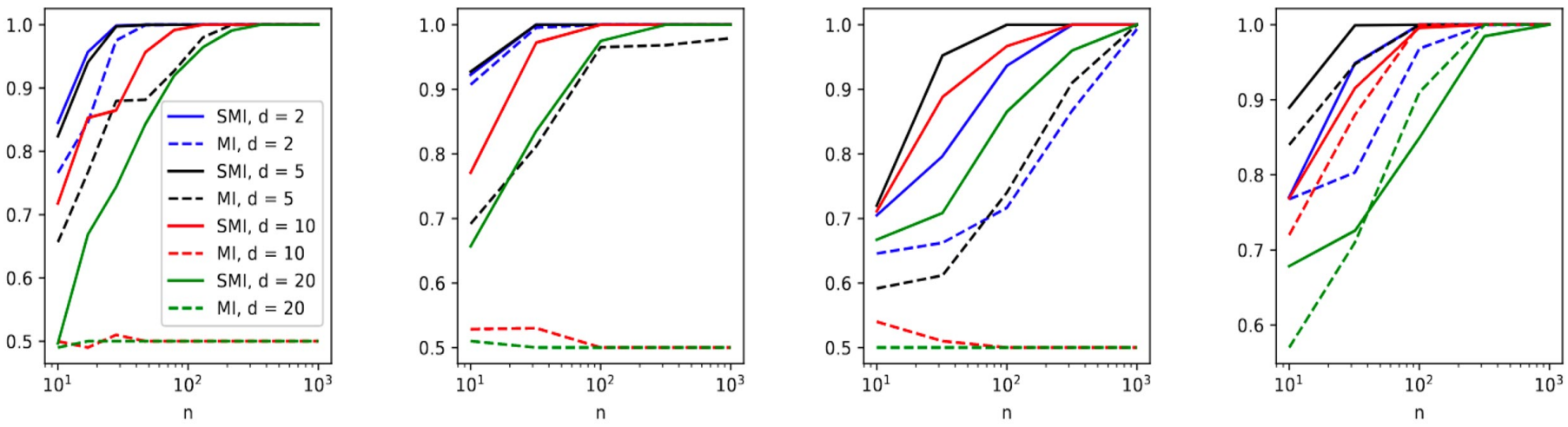
$\implies$ Compute SMI **&** threshold for independence testing

# Experiments: Independence Testing

**Recall:** $\mathsf{SI}(X;Y) = 0 \iff (X,Y)$ independent

$\implies$ Compute SMI **&** threshold for independence testing

**Figure: Area under the ROC curve**




(a) $Y$ encodes a single feature.

$$Y = \frac{1}{\sqrt{d}}(\mathbf{1}^\intercal X)\mathbf{1} + Z$$

(b) $Y$ encodes two features.

$$Y_i = \begin{cases} \frac{1}{d}(\mathbf{1}_{\lfloor d/2 \rfloor}0\ldots0)^\intercal X + Z_i, & i \le \frac{d}{2} \\ \frac{1}{d}(0\ldots0\mathbf{1}_{\lceil d/2 \rceil})^\intercal X + Z_i, & i > \frac{d}{2} \end{cases}$$

(c) Low rank common signal.

$$X = \mathrm{P}_1 V + Z_1$$
$$Y = \mathrm{P}_2 V + Z_2$$

(d) Independent coordinates.

$$Y = X + Z$$

# Experiments: Feature Extraction

**Goal:** Maximize SI$(\mathrm{A}X; \mathrm{A}Y)$, $X, Y$ i.i.d. from same MNIST class (0 or 1)

# Experiments: Feature Extraction

<u>**Goal:**</u> Maximize $\mathsf{SI}(\mathrm{A}X; \mathrm{A}Y)$, $X, Y$ i.i.d. from same MNIST class (0 or 1)

- **Upper bound:** $\mathsf{SI}(\mathrm{A}X; \mathrm{A}Y) \leq \mathsf{I}(X; Y) = 1$ bit

# Experiments: Feature Extraction

<u>**Goal:**</u> Maximize $\mathsf{SI}(\mathrm{A}X; \mathrm{A}Y)$, $X, Y$ i.i.d. from same MNIST class (0 or 1)

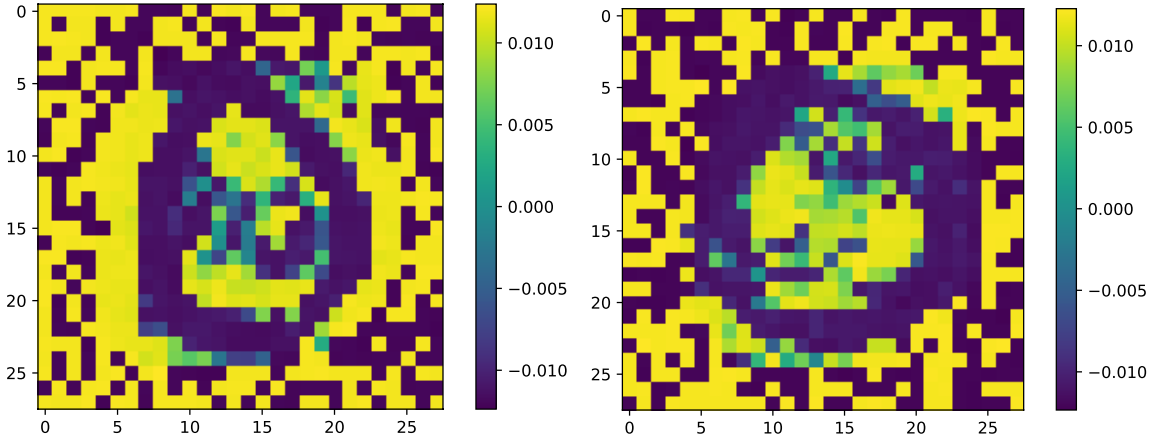
- **Upper bound:** $\mathsf{SI}(\mathrm{A}X; \mathrm{A}Y) \leq \mathsf{I}(X; Y) = 1$ bit
- **Interpretation:** $\mathrm{A}$ is linear features most useful for label classification

# Experiments: Feature Extraction

**Goal:** Maximize $\mathsf{SI}(\mathrm{A}X;\mathrm{A}Y)$, $X, Y$ i.i.d. from same MNIST class (0 or 1)

- **Upper bound:** $\mathsf{SI}(\mathrm{A}X;\mathrm{A}Y) \leq \mathsf{I}(X;Y) = 1$ bit
- **Interpretation:** $\mathrm{A}$ is linear features most useful for label classification

**Figure: Feature extraction for MNIST**



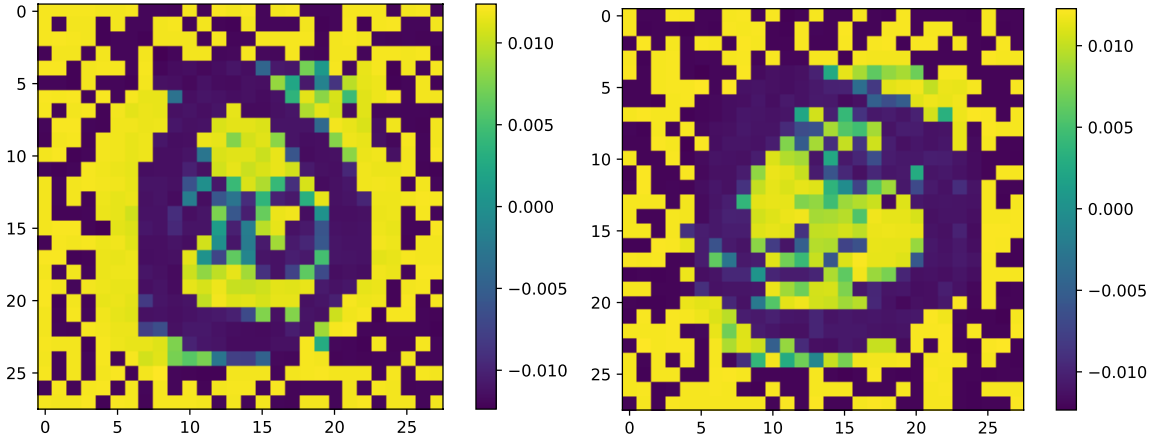

Rows 0 and 1 of optimized $\mathrm{A}$ (rearranged as MNIST image)

# Experiments: Feature Extraction

<u>Goal:</u> Maximize $\mathsf{SI}(\mathrm{A}X;\mathrm{A}Y)$, $X, Y$ i.i.d. from same MNIST class (0 or 1)

- **Upper bound:** $\mathsf{SI}(\mathrm{A}X;\mathrm{A}Y) \leq \mathsf{I}(X;Y) = 1$ bit
- **Interpretation:** $\mathrm{A}$ is linear features most useful for label classification

**Figure: Feature extraction for MNIST**




Rows 0 and 1 of optimized $\mathrm{A}$ (rearranged as MNIST image)

$\Longrightarrow$ $\mathsf{SI}(\mathrm{A}^\star X;\mathrm{A}^\star Y) = 0.68$ (compare to 0.0752 for random $\mathrm{A}$)

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI
- **Processing:** Violates DPI and can increase from processing (useful?)

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI
- **Processing:** Violates DPI and can increase from processing (useful?)
- **Estimation:** Efficiently computable **&** fast to estimate

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI
- **Processing:** Violates DPI and can increase from processing (useful?)
- **Estimation:** Efficiently computable & fast to estimate
- **Applications:** Independence testing & feature extraction

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI
- **Processing:** Violates DPI and can increase from processing (useful?)
- **Estimation:** Efficiently computable & fast to estimate
- **Applications:** Independence testing & feature extraction

**Future directions:** Theoretical and applied

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI
- **Processing:** Violates DPI and can increase from processing (useful?)
- **Estimation:** Efficiently computable & fast to estimate
- **Applications:** Independence testing & feature extraction

**Future directions:** Theoretical and applied

- Extensions to $k$-dimensional projections

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI
- **Processing:** Violates DPI and can increase from processing (useful?)
- **Estimation:** Efficiently computable & fast to estimate
- **Applications:** Independence testing & feature extraction

**Future directions:** Theoretical and applied

- Extensions to $k$-dimensional projections
- Decomposition to Gaussian SMI plus residual (negligible?)

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI
- **Processing:** Violates DPI and can increase from processing (useful?)
- **Estimation:** Efficiently computable & fast to estimate
- **Applications:** Independence testing & feature extraction

**Future directions:** Theoretical and applied

- Extensions to $k$-dimensional projections
- Decomposition to Gaussian SMI plus residual (negligible?)
- Applications to more complex learning tasks (InfoGAN, InfoMAX, ...)

# Summary

**Sliced mutual information:** Avg. scalar MI terms btw. 1D projections

- **Structure:** Preserves many properties of classic MI
- **Processing:** Violates DPI and can increase from processing (useful?)
- **Estimation:** Efficiently computable & fast to estimate
- **Applications:** Independence testing & feature extraction

**Future directions:** Theoretical and applied

- Extensions to $k$-dimensional projections
- Decomposition to Gaussian SMI plus residual (negligible?)
- Applications to more complex learning tasks (InfoGAN, InfoMAX, ...)

**Thank you!**

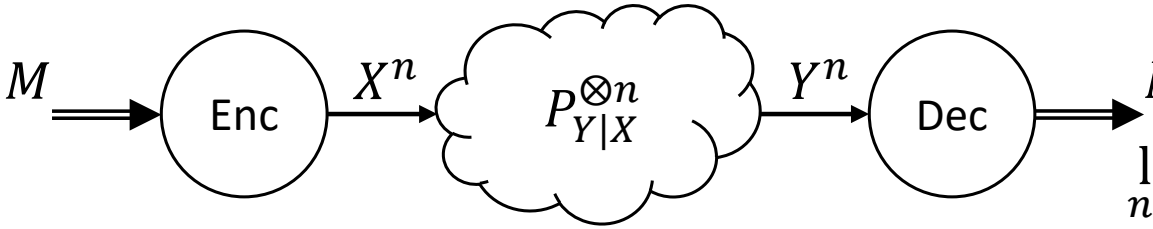
# Mutual Information: Applications

**Information Theory:** Emerges as solution to operation problems

# Mutual Information: Applications

**Information Theory:** Emerges as solution to operation problems

- **Noisy communication:** Channel capacity $\mathsf{C}(P_{Y|X}) = \max_{P_X} \mathsf{I}(X;Y)$

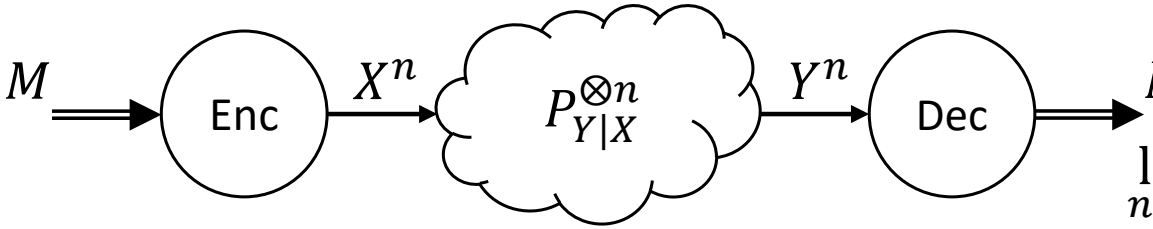



$$\lim_{n\to\infty} \mathbb{P}\big(\widehat{M} \neq M\big) = 0$$

# Mutual Information: Applications

**Information Theory:** Emerges as solution to operation problems

- **Noisy communication:** Channel capacity $\mathsf{C}(P_{Y|X}) = \max_{P_X} \mathsf{I}(X;Y)$



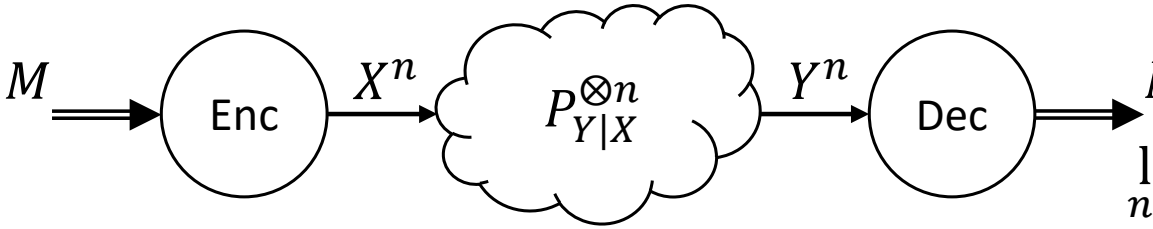

$$\lim_{n\to\infty} \mathbb{P}\big(\widehat{M} \neq M\big) = 0$$

- **Compression:** Rate-distortion $\mathsf{R}(D, P_X) = \min_{P_{\hat{X}|X}:\ \mathbb{E}[d(X,\hat{X})]\leq D} \mathsf{I}(X;\hat{X})$

# Mutual Information: Applications

**Information Theory:** Emerges as solution to operation problems

- **Noisy communication:** Channel capacity $\mathsf{C}(P_{Y|X}) = \max_{P_X} \mathsf{I}(X;Y)$



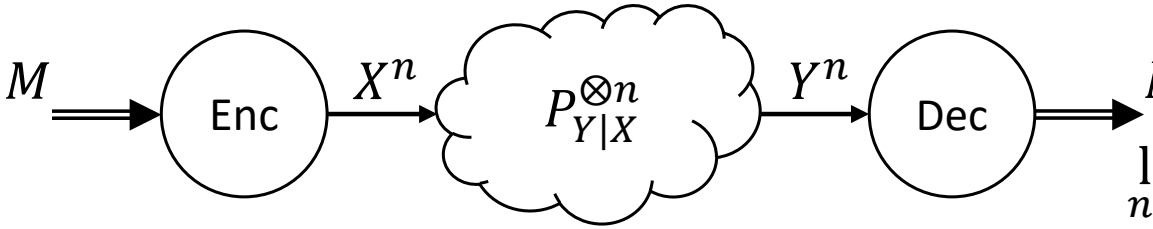

$$\lim_{n\to\infty} \mathbb{P}\big(\widehat{M} \neq M\big) = 0$$

- **Compression:** Rate-distortion $\mathsf{R}(D, P_X) = \min_{P_{\hat{X}|X}:\ \mathbb{E}[d(X,\hat{X})]\leq D} \mathsf{I}(X;\hat{X})$
- **More:** Distribution simulation, privacy & security, common info...

# Mutual Information: Applications

**Information Theory:** Emerges as solution to operation problems

- **Noisy communication:** Channel capacity $\mathsf{C}(P_{Y|X}) = \max_{P_X} \mathsf{I}(X;Y)$




$$\lim_{n\to\infty} \mathbb{P}\big(\widehat{M} \neq M\big) = 0$$

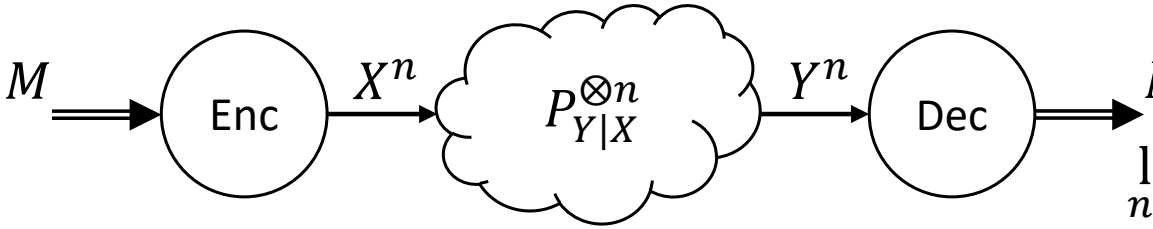
- **Compression:** Rate-distortion $\mathsf{R}(D, P_X) = \min_{P_{\hat{X}|X}:\ \mathbb{E}[d(X,\hat{X})]\leq D} \mathsf{I}(X;\hat{X})$
- **More:** Distribution simulation, privacy & security, common info...

**Statistics:** Ind. testing, impossibility results, dependence quantification...

# Mutual Information: Applications

**Information Theory:** Emerges as solution to operation problems

- **Noisy communication:** Channel capacity $\mathsf{C}(P_{Y|X}) = \max_{P_X} \mathsf{I}(X;Y)$




$$\lim_{n\to\infty} \mathbb{P}\big(\widehat{M} \neq M\big) = 0$$

- **Compression:** Rate-distortion $\mathsf{R}(D, P_X) = \min_{P_{\hat{X}|X}:\ \mathbb{E}[d(X,\hat{X})]\leq D} \mathsf{I}(X;\hat{X})$
- **More:** Distribution simulation, privacy & security, common info...

**Statistics:** Ind. testing, impossibility results, dependence quantification...

**Machine Learning:** Hosts of modern applications

# Mutual Information in Machine Learning

**Representation learning:** InfoMax principle

# Mutual Information in Machine Learning

**Representation learning:** InfoMax principle

- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$

# Mutual Information in Machine Learning

**Representation learning:** InfoMax principle

- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$
- **Encoder:** $e_\theta : \mathbb{R}^d \to \mathbb{R}^m, \; m \ll d$

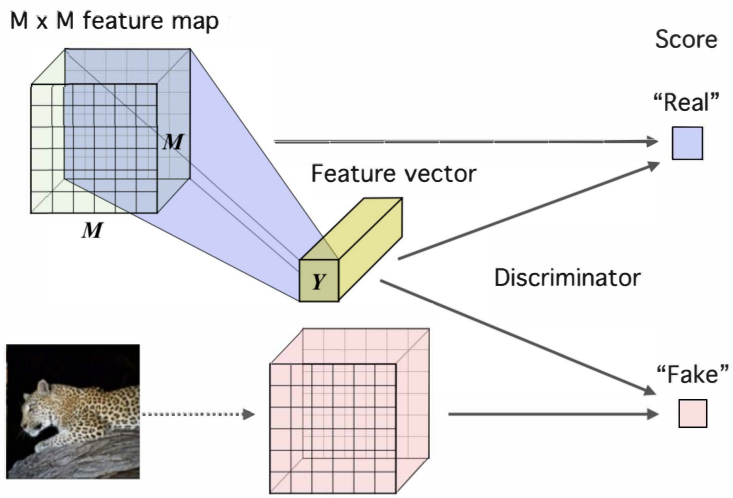
# Mutual Information in Machine Learning

**Representation learning:** InfoMax principle

- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$
- **Encoder:** $e_\theta : \mathbb{R}^d \to \mathbb{R}^m,\ m \ll d$
- **Objective:** $\sup_{\theta \in \Theta} \mathsf{I}(e_\theta(X); X)$

# Mutual Information in Machine Learning

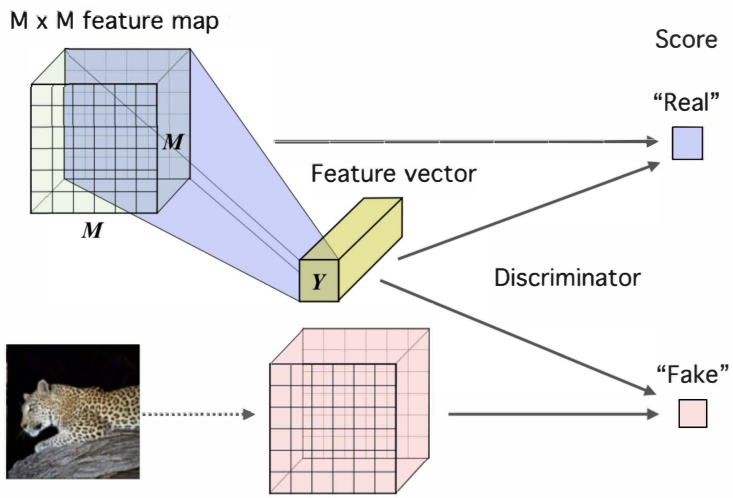
**Representation learning:** InfoMax principle




- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$
- **Encoder:** $e_\theta : \mathbb{R}^d \to \mathbb{R}^m, \; m \ll d$
- **Objective:** $\sup_{\theta \in \Theta} \mathsf{I}(e_\theta(X); X)$

1. **DIM [Hjelm et al.'18]**: Parameterized DV lower bound

# Mutual Information in Machine Learning

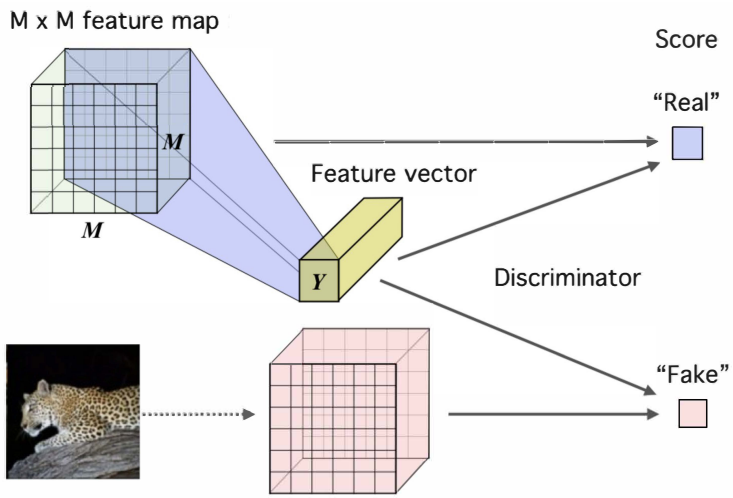



**Representation learning:** InfoMax principle

- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$
- **Encoder:** $e_\theta : \mathbb{R}^d \to \mathbb{R}^m, \; m \ll d$
- **Objective:** $\sup_{\theta \in \Theta} \mathsf{I}(e_\theta(X); X)$

1. **DIM [Hjelm et al.'18]**: Parameterized DV lower bound
2. **CPC [Oord et al.'19]**: Contrastive $\mathcal{L}_{\mathsf{InfoNCE}}$ lower bound

# Mutual Information in Machine Learning



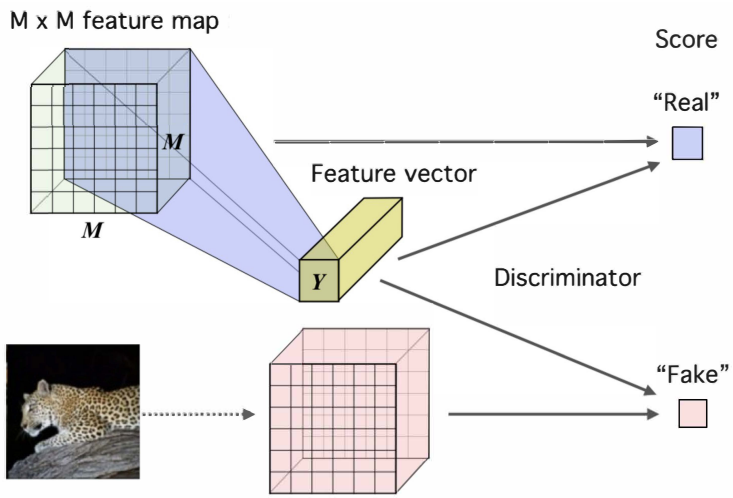

**Representation learning:** InfoMax principle

- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$
- **Encoder:** $e_\theta : \mathbb{R}^d \to \mathbb{R}^m, \ m \ll d$
- **Objective:** $\sup_{\theta \in \Theta} \mathsf{I}(e_\theta(X); X)$

1. **DIM [Hjelm et al.'18]**: Parameterized DV lower bound
2. **CPC [Oord et al.'19]**: Contrastive $\mathcal{L}_{\mathsf{InfoNCE}}$ lower bound

**Generative modeling :** Disentangled latent space

# Mutual Information in Machine Learning




**Representation learning:** InfoMax principle

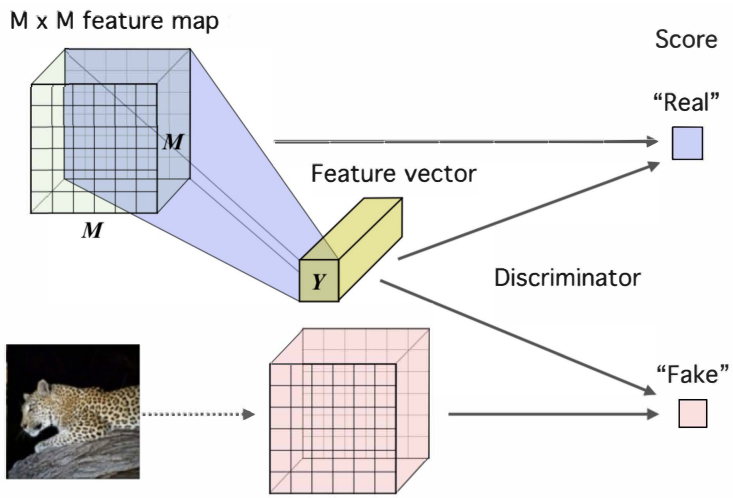
- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$
- **Encoder:** $e_\theta : \mathbb{R}^d \to \mathbb{R}^m, \ m \ll d$
- **Objective:** $\sup_{\theta \in \Theta} \mathsf{I}(e_\theta(X); X)$

1. **DIM [Hjelm et al.'18]**: Parameterized DV lower bound
2. **CPC [Oord et al.'19]**: Contrastive $\mathcal{L}_{\mathsf{InfoNCE}}$ lower bound

**Generative modeling :** Disentangled latent space

- **Goal:** Learn $g_\theta(Z) \sim Q_\theta$ that mimics $X \sim P_X$

# Mutual Information in Machine Learning




**Representation learning:** InfoMax principle

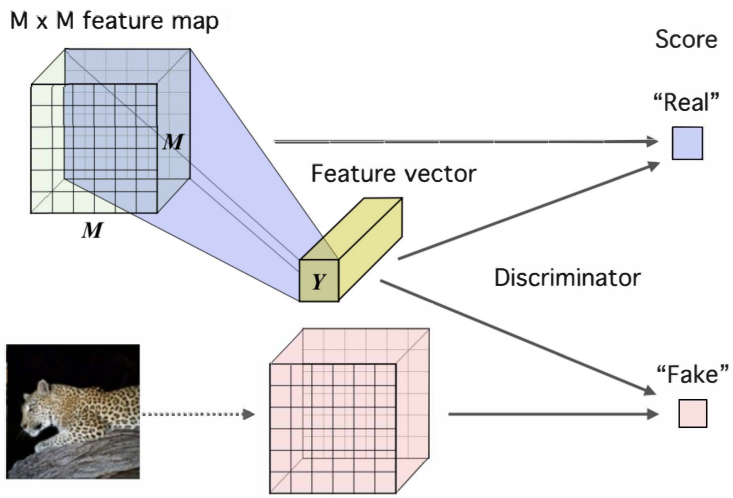
- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$
- **Encoder:** $e_\theta : \mathbb{R}^d \to \mathbb{R}^m, \; m \ll d$
- **Objective:** $\sup_{\theta \in \Theta} \mathsf{I}(e_\theta(X); X)$

1. **DIM [Hjelm et al.'18]**: Parameterized DV lower bound
2. **CPC [Oord et al.'19]**: Contrastive $\mathcal{L}_{\mathsf{InfoNCE}}$ lower bound

**Generative modeling :** Disentangled latent space

- **Goal:** Learn $g_\theta(Z) \sim Q_\theta$ that mimics $X \sim P_X$
- **GAN:** $\min_{\theta \in \Theta} \max_{\phi \in \Phi} \mathcal{L}_{\mathsf{GAN}}(g_\theta, d_\phi)$

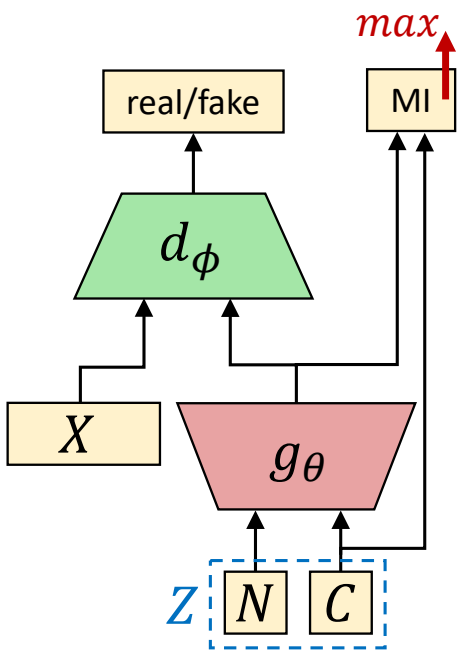
# Mutual Information in Machine Learning




**Representation learning:** InfoMax principle

- **Data:** $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$
- **Encoder:** $e_\theta : \mathbb{R}^d \to \mathbb{R}^m, \; m \ll d$
- **Objective:** $\sup_{\theta \in \Theta} \mathsf{I}(e_\theta(X); X)$

1. **DIM [Hjelm et al.'18]**: Parameterized DV lower bound
2. **CPC [Oord et al.'19]**: Contrastive $\mathcal{L}_{\mathsf{InfoNCE}}$ lower bound

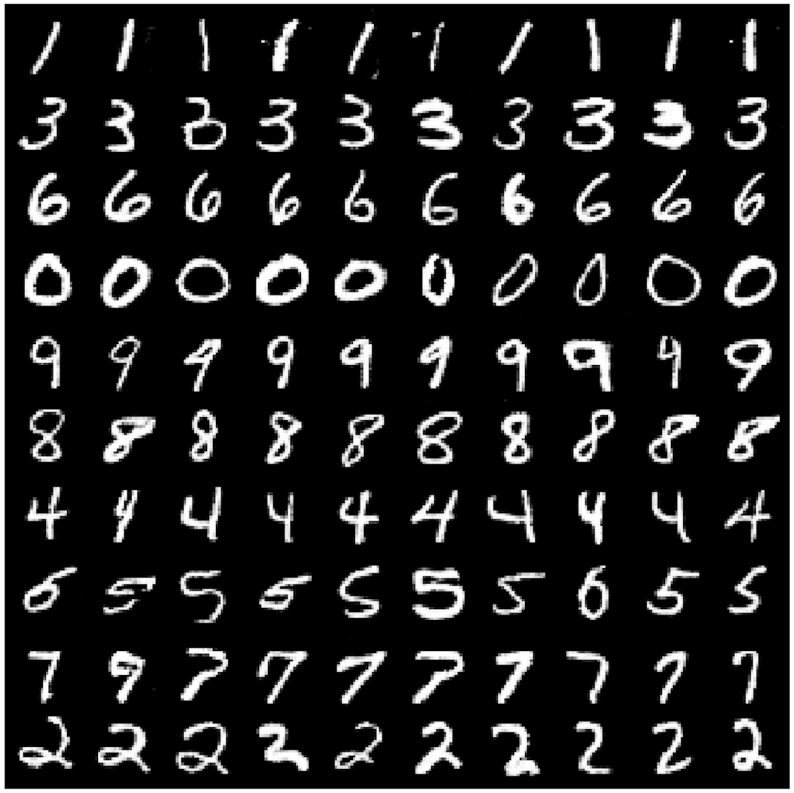



**Generative modeling :** Disentangled latent space

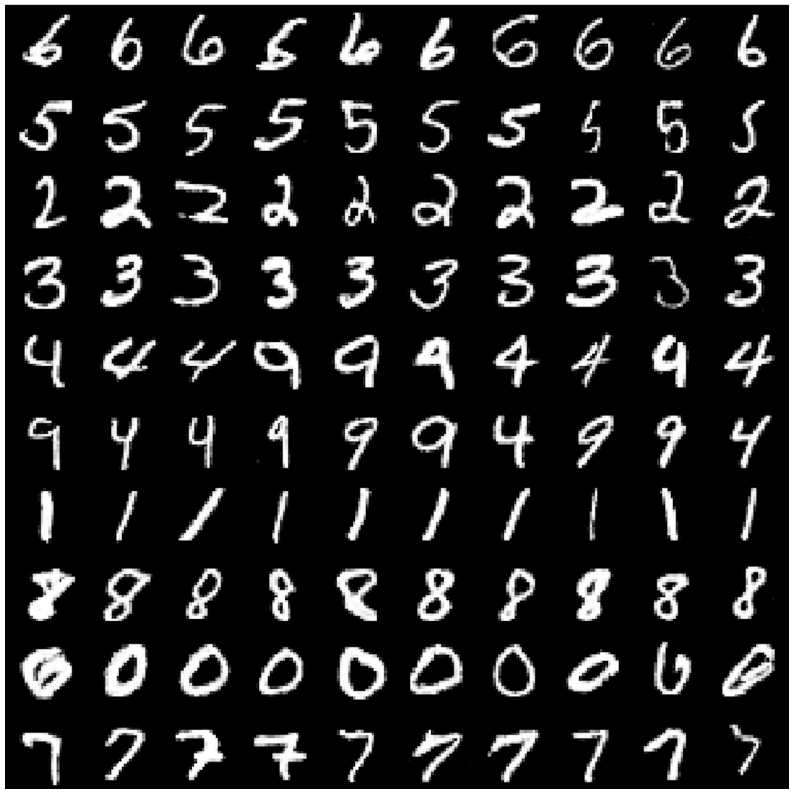
- **Goal:** Learn $g_\theta(Z) \sim Q_\theta$ that mimics $X \sim P_X$
- **GAN:** $\min_{\theta \in \Theta} \max_{\phi \in \Phi} \mathcal{L}_{\mathsf{GAN}}(g_\theta, d_\phi)$
- **InfoGAN [Chen et al.'16], [Belghazi et al.'18]:**
  - Model $Z = (N \; C)$ **&** regularize $\mathcal{L}_{\mathsf{GAN}}(g_\theta, d_\phi)$ by $\beta \mathsf{I}(g_\theta(N, C); C)$

# Sliced InfoGAN: MNIST Results

**Regular InfoGAN (MI)**



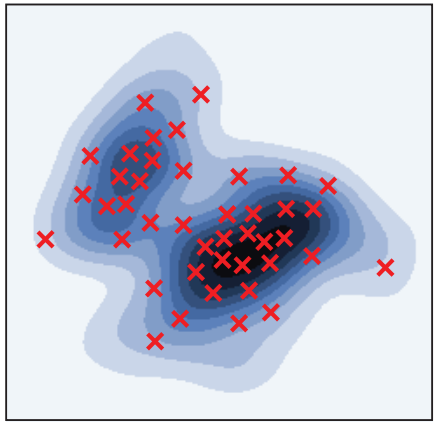
**Sliced InfoGAN (SMI)**



**Codes:** $C_1 \in [0:9]$ (digits), $C_2 \in [-2, 2]$ (rotation), $C_3 \in [-2, 2]$ (width)

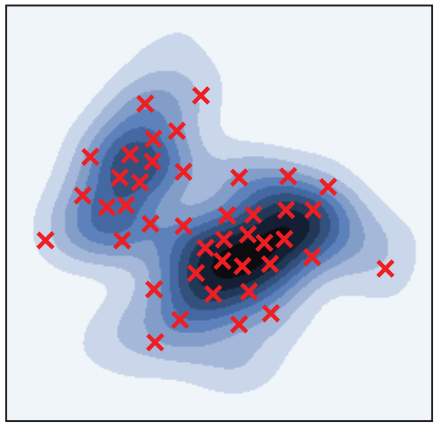
# Mutual Information Estimation

**In practice:** Don't have $P_{XY}$ but samples $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{XY}$



# Mutual Information Estimation

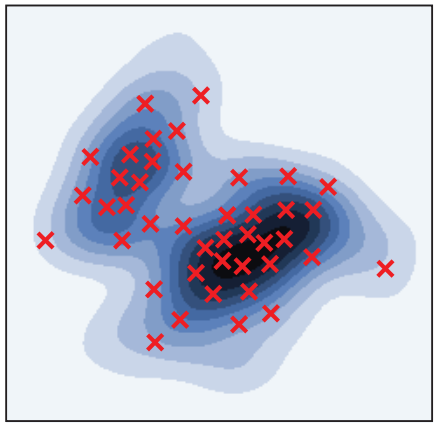
**In practice:** Don't have $P_{XY}$ but samples $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{XY}$



- **Estimation:** $\hat{\mathsf{I}}(X^n, Y^n)$ via $k$-NN, KDE, etc.

# Mutual Information Estimation

**In practice:** Don't have $P_{XY}$ but samples $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{XY}$

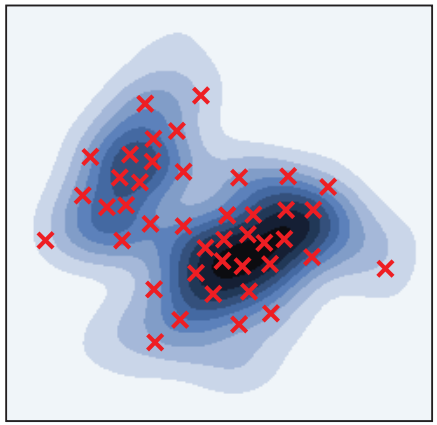


- **Estimation:** $\hat{\mathsf{I}}(X^n, Y^n)$ via $k$-NN, KDE, etc.

$\Longrightarrow$ Can we approximate $\mathsf{I}(X;Y) \approx \hat{\mathsf{I}}(X^n, Y^n)$?

# Mutual Information Estimation

**In practice:** Don't have $P_{XY}$ but samples $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{XY}$



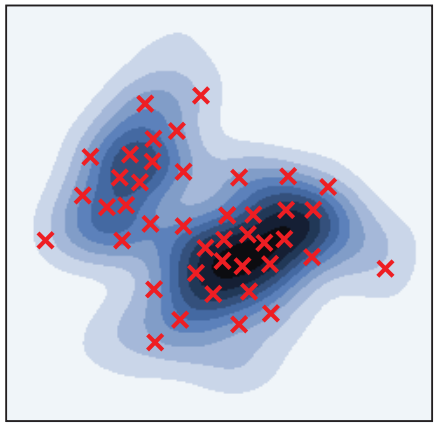
- **Estimation:** $\hat{\mathsf{I}}(X^n, Y^n)$ via $k$-NN, KDE, etc.

$\Longrightarrow$ Can we approximate $\mathsf{I}(X;Y) \approx \hat{\mathsf{I}}(X^n, Y^n)$?

**Main challenge:** Estimation in high dim. is infeasible

# Mutual Information Estimation

**In practice:** Don't have $P_{XY}$ but samples $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{XY}$



- **Estimation:** $\hat{\mathsf{I}}(X^n, Y^n)$ via $k$-NN, KDE, etc.

$\Longrightarrow$ Can we approximate $\mathsf{I}(X;Y) \approx \hat{\mathsf{I}}(X^n, Y^n)$?

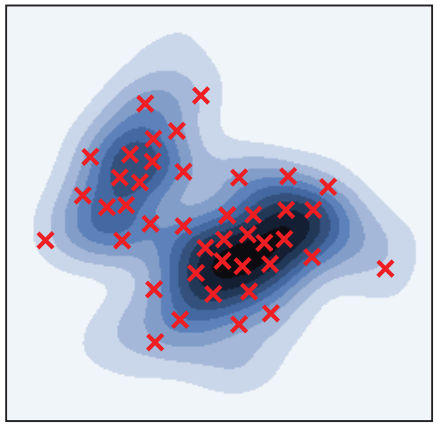
**Main challenge:** Estimation in high dim. is infeasible

- **Sample complexity:** $n^\star(\epsilon, d) \asymp \epsilon^{-d}$ (under regularity)
  - Hölder smooth [Jiao-Gao-Han'18]
  - (Gen.) Lipschitz smooth [Han-Jiao-Weissman-Wu'20]

$\vdots$

# Mutual Information Estimation

**In practice:** Don't have $P_{XY}$ but samples $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{XY}$



- **Estimation:** $\hat{\mathsf{I}}(X^n, Y^n)$ via $k$-NN, KDE, etc.

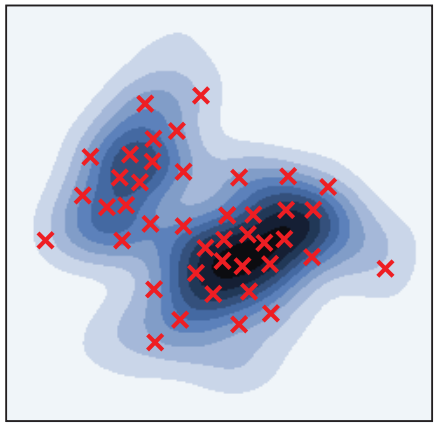
$\Longrightarrow$ Can we approximate $\mathsf{I}(X;Y) \approx \hat{\mathsf{I}}(X^n, Y^n)$?

**Main challenge:** Estimation in high dim. is infeasible

- **Sample complexity:** $n^\star(\epsilon, d) \asymp \epsilon^{-d}$ (under regularity)
  - Hölder smooth [Jiao-Gao-Han'18]
  - (Gen.) Lipschitz smooth [Han-Jiao-Weissman-Wu'20]

$\vdots$

# Mutual Information Estimation



**In practice:** Don't have $P_{XY}$ but samples $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{XY}$

- **Estimation:** $\hat{\mathsf{I}}(X^n, Y^n)$ via $k$-NN, KDE, etc.

$\Longrightarrow$ Can we approximate $\mathsf{I}(X;Y) \approx \hat{\mathsf{I}}(X^n, Y^n)$?

**Main challenge:** Estimation in high dim. is infeasible

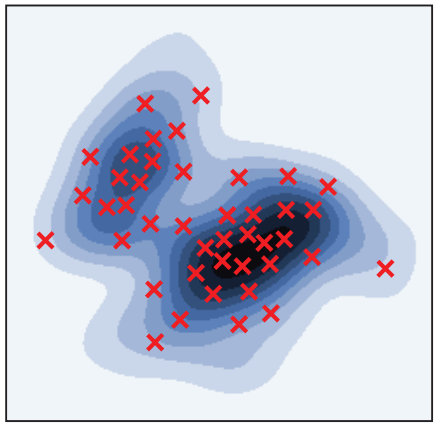
- **Sample complexity:** $n^\star(\epsilon, d) \asymp \epsilon^{-d}$ (under regularity)
  - Hölder smooth [Jiao-Gao-Han'18]
  - (Gen.) Lipschitz smooth [Han-Jiao-Weissman-Wu'20]

$$\vdots$$

- **Formal limitations:** $n^\star(\epsilon, d) \gtrsim \exp\left(\mathsf{I}(X;Y)\right)$ [McAllester-Stratos'20]

# Mutual Information Estimation

**In practice:** Don't have $P_{XY}$ but samples $(X_i, Y_i) \overset{\text{iid}}{\sim} P_{XY}$



- **Estimation:** $\hat{\mathsf{I}}(X^n, Y^n)$ via $k$-NN, KDE, etc.

$\Longrightarrow$ Can we approximate $\mathsf{I}(X;Y) \approx \hat{\mathsf{I}}(X^n, Y^n)$?

**Main challenge:** Estimation in high dim. is infeasible

- **Sample complexity:** $n^\star(\epsilon, d) \asymp \epsilon^{-d}$ (under regularity)
  - Hölder smooth [Jiao-Gao-Han'18]
  - (Gen.) Lipschitz smooth [Han-Jiao-Weissman-Wu'20]

$$\vdots$$

- **Formal limitations:** $n^\star(\epsilon, d) \gtrsim \exp\left(\mathsf{I}(X;Y)\right)$ [McAllester-Stratos'20]

⊛ **Goal:** Scalable MI surrogate that preserves its structure

# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

$\Longrightarrow$ Scalable **&** preserves properties

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

  $\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P, Q)$

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

  $\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P, Q)$

- 1-proj. SMI can nullify btw dependent $(X, Y)$

# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

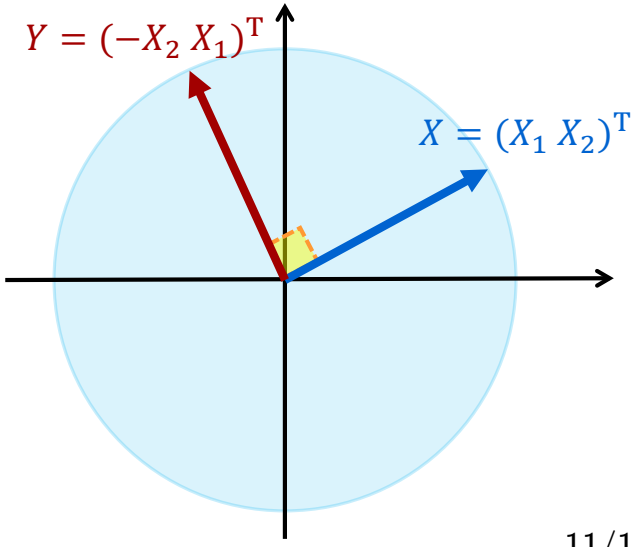
**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

$\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P, Q)$

- 1-proj. SMI can nullify btw dependent $(X, Y)$
- $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = (-X_2 \; X_1)^\intercal$



# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

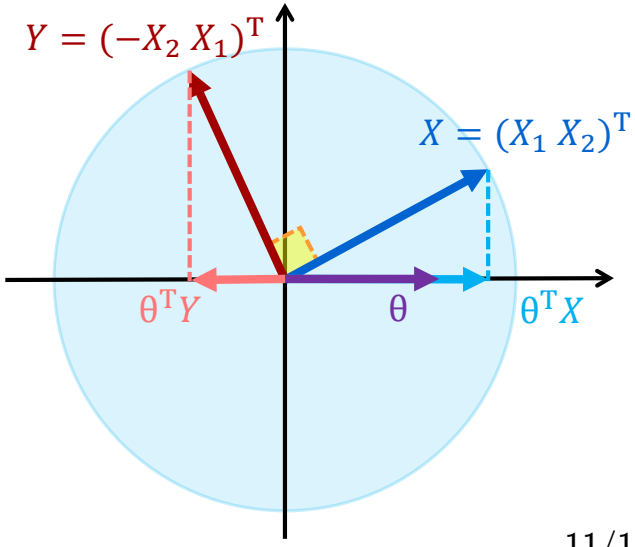
**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

$\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P, Q)$

- 1-proj. SMI can nullify btw dependent $(X, Y)$
- $X = (X_1\ X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = (-X_2\ X_1)^\intercal$



# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

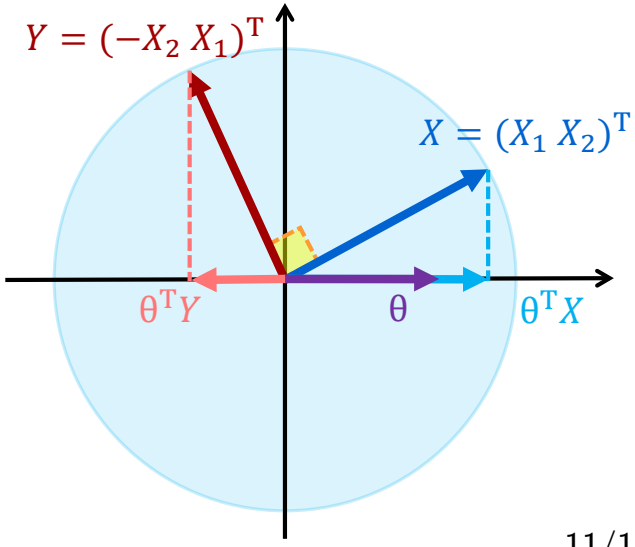
**Sliced divergences:** $\delta(P,Q)$ btw $P,Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P,Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

$\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P,Q)$

- 1-proj. SMI can nullify btw dependent $(X, Y)$
- $X = (X_1\ X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = (-X_2\ X_1)^\intercal$
  - $\mathrm{cov}(\theta^\intercal X, \theta^\intercal Y) = 0 \Longrightarrow \widetilde{\mathsf{SI}}(X;Y) = 0$



# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

$\Longrightarrow$ Scalable **&** preserves properties

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

  $\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P, Q)$

# Sliced Mutual Information (SMI)

**Definition (ZG-Greenewald'21)**

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

  $\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P, Q)$

- 1-proj. SMI can nullify btw dependent $(X, Y)$

# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

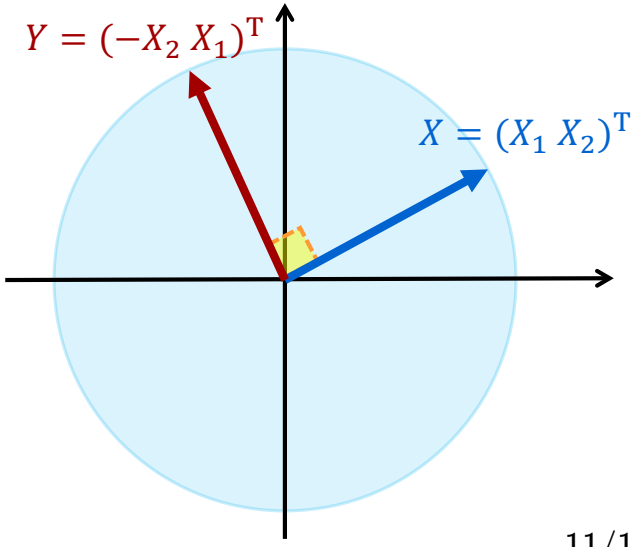
**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

$\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P, Q)$

- 1-proj. SMI can nullify btw dependent $(X, Y)$
- $X = (X_1 \; X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = (-X_2 \; X_1)^\intercal$



# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

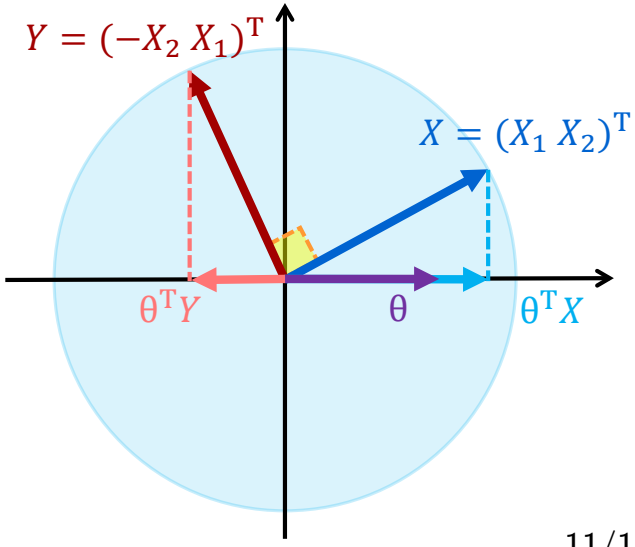
**Sliced divergences:** $\delta(P, Q)$ btw $P, Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P, Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

$\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P, Q)$

- 1-proj. SMI can nullify btw dependent $(X, Y)$
- $X = (X_1\ X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = (-X_2\ X_1)^\intercal$



# Sliced Mutual Information (SMI)

## Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_x(\theta) d\sigma_y(\phi),$$

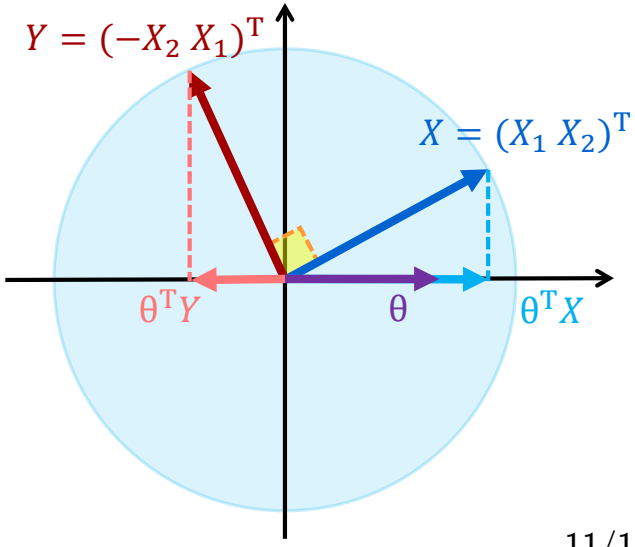
**Sliced divergences:** $\delta(P,Q)$ btw $P,Q \in \mathcal{P}(\mathbb{R}^d)$ (Wasserstein, $f$-div., IPM)

- **Slicing:** $\bar{\delta}(P,Q) := \int_{\mathbb{S}^{d-1}} \delta(\mathfrak{p}^\theta_\sharp P, \mathfrak{p}^\theta_\sharp Q) d\sigma(\theta)$, where $\mathfrak{p}^\theta(x) = \theta^\intercal x$

$\Longrightarrow$ Scalable **&** preserves properties

**Difference:** **2** proj. for $\mathsf{SI}(X;Y)$ **vs.** **1** for $\bar{\delta}(P,Q)$

- 1-proj. SMI can nullify btw dependent $(X, Y)$
- $X = (X_1\ X_2)^\intercal \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = (-X_2\ X_1)^\intercal$
  - $\mathrm{cov}(\theta^\intercal X, \theta^\intercal Y) = 0 \Longrightarrow \widetilde{\mathsf{SI}}(X;Y) = 0$



# Independence Identification: Proof Outline

**Char. functions:** $\varphi_{X,Y}(t,s) := \mathbb{E}\big[e^{it^\intercal X + is^\intercal Y}\big]$ **&** $\varphi_X(t) := \varphi_{X,Y}(t,0)$

# Independence Identification: Proof Outline

**Char. functions:** $\varphi_{X,Y}(t,s) := \mathbb{E}\big[e^{it^\intercal X + is^\intercal Y}\big]$ **&** $\varphi_X(t) := \varphi_{X,Y}(t,0)$

**Recall:** $X \perp Y \iff \varphi_{X,Y}(t,s) = \varphi_X(t)\varphi_Y(s), \quad \forall\, t \in \mathbb{R}^{d_x},\, s \in \mathbb{R}^{d_y}$

# Independence Identification: Proof Outline

**Char. functions:** $\varphi_{X,Y}(t,s) := \mathbb{E}\big[e^{it^\intercal X + is^\intercal Y}\big]$ **&** $\varphi_X(t) := \varphi_{X,Y}(t,0)$

**Recall:** $X \perp Y \iff \varphi_{X,Y}(t,s) = \varphi_X(t)\varphi_Y(s), \quad \forall\, t \in \mathbb{R}^{d_x},\, s \in \mathbb{R}^{d_y}$

**Consider:** $\mathsf{SI}(X;Y) = \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)\, \mathrm{d}\sigma_{d_x}(\theta)\, \mathrm{d}\sigma_{d_y}(\phi) = 0$

# Independence Identification: Proof Outline

**Char. functions:** $\varphi_{X,Y}(t,s) := \mathbb{E}\big[e^{it^\intercal X + is^\intercal Y}\big]$ **&** $\varphi_X(t) := \varphi_{X,Y}(t,0)$

**Recall:** $X \perp Y \iff \varphi_{X,Y}(t,s) = \varphi_X(t)\varphi_Y(s), \quad \forall\, t \in \mathbb{R}^{d_x},\, s \in \mathbb{R}^{d_y}$

**Consider:** $\mathsf{SI}(X;Y) = \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)\, \mathrm{d}\sigma_{d_x}(\theta)\, \mathrm{d}\sigma_{d_y}(\phi) = 0$

$$\iff$$

$$\mathsf{I}(\theta^\intercal X; \phi^\intercal Y) = 0, \quad \forall\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

# Independence Identification: Proof Outline

**Char. functions:** $\varphi_{X,Y}(t,s) := \mathbb{E}\big[e^{it^\intercal X + is^\intercal Y}\big]$ **&** $\varphi_X(t) := \varphi_{X,Y}(t,0)$

**Recall:** $X \perp Y \iff \varphi_{X,Y}(t,s) = \varphi_X(t)\varphi_Y(s), \quad \forall\, t \in \mathbb{R}^{d_x},\, s \in \mathbb{R}^{d_y}$

**Consider:** $\mathsf{SI}(X;Y) = \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)\, \mathrm{d}\sigma_{d_x}(\theta)\, \mathrm{d}\sigma_{d_y}(\phi) = 0$

$$\iff$$

$$\mathsf{I}(\theta^\intercal X; \phi^\intercal Y) = 0, \quad \forall\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

$$\iff$$

$$\varphi_{\theta^\intercal X, \phi^\intercal Y}(u,v) = \varphi_{\theta^\intercal X}(u)\varphi_{\phi^\intercal Y}(v), \quad \forall\, u,v \in \mathbb{R},\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

# Independence Identification: Proof Outline

**Char. functions:** $\varphi_{X,Y}(t,s) := \mathbb{E}\big[e^{it^\intercal X + is^\intercal Y}\big]$ **&** $\varphi_X(t) := \varphi_{X,Y}(t,0)$

**Recall:** $X \perp Y \iff \varphi_{X,Y}(t,s) = \varphi_X(t)\varphi_Y(s), \quad \forall\, t \in \mathbb{R}^{d_x},\, s \in \mathbb{R}^{d_y}$

**Consider:** $\mathsf{SI}(X;Y) = \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)\, \mathrm{d}\sigma_{d_x}(\theta)\, \mathrm{d}\sigma_{d_y}(\phi) = 0$

$$\iff$$

$$\mathsf{I}(\theta^\intercal X; \phi^\intercal Y) = 0, \quad \forall\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

$$\iff$$

$$\varphi_{\theta^\intercal X, \phi^\intercal Y}(u,v) = \varphi_{\theta^\intercal X}(u)\varphi_{\phi^\intercal Y}(v), \quad \forall\, u,v \in \mathbb{R},\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

$$\iff$$

$$\varphi_{X,Y}(u\theta, v\phi) = \varphi_X(u\theta)\varphi_Y(v\phi), \quad \forall\, u,v \in \mathbb{R},\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

# Independence Identification: Proof Outline

**Char. functions:** $\varphi_{X,Y}(t,s) := \mathbb{E}\big[e^{it^\intercal X + is^\intercal Y}\big]$ **&** $\varphi_X(t) := \varphi_{X,Y}(t,0)$

**Recall:** $X \perp Y \iff \varphi_{X,Y}(t,s) = \varphi_X(t)\varphi_Y(s), \quad \forall\, t \in \mathbb{R}^{d_x},\, s \in \mathbb{R}^{d_y}$

**Consider:** $\mathsf{SI}(X;Y) = \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)\, \mathrm{d}\sigma_{d_x}(\theta)\, \mathrm{d}\sigma_{d_y}(\phi) = 0$

$$\iff$$

$$\mathsf{I}(\theta^\intercal X; \phi^\intercal Y) = 0, \quad \forall\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

$$\iff$$

$$\varphi_{\theta^\intercal X, \phi^\intercal Y}(u,v) = \varphi_{\theta^\intercal X}(u)\varphi_{\phi^\intercal Y}(v), \quad \forall\, u,v \in \mathbb{R},\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

$$\iff$$

$$\varphi_{X,Y}(u\theta, v\phi) = \varphi_X(u\theta)\varphi_Y(v\phi), \quad \forall\, u,v \in \mathbb{R},\, \theta \in \mathbb{S}^{d_x-1},\, \phi \in \mathbb{S}^{d_y-1}$$

$$\iff$$

$$\varphi_{X,Y}(t,s) = \varphi_X(t)\varphi_Y(s), \quad \forall\, t \in \mathbb{R}^{d_x},\, s \in \mathbb{R}^{d_y}$$

# Sliced Mutual Information & Processing (Cont.)

**Proposition (ZG-Greenewald'21)**

*Extracting maximum-SMI linear feature:*

$$\sup_{\mathrm{A}_x, \mathrm{A}_y, b_x, b_y} \mathsf{SI}(\mathrm{A}_x X + b_x; \mathrm{A}_y Y + b_y) = \sup_{\theta, \phi} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y).$$

*Also, if* $(\theta_\star, \phi_\star) \in \operatorname{argmax} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)$, *then* $[\mathrm{A}_x^\star]_{1:} = \theta_\star^\intercal$, $[\mathrm{A}_y^\star]_{1:} = \phi_\star^\intercal$.

# Sliced Mutual Information & Processing (Cont.)

**Proposition (ZG-Greenewald'21)**

*Extracting maximum-SMI linear feature:*

$$\sup_{\mathrm{A}_x, \mathrm{A}_y, b_x, b_y} \mathsf{SI}(\mathrm{A}_x X + b_x; \mathrm{A}_y Y + b_y) = \sup_{\theta, \phi} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y).$$

*Also, if* $(\theta_\star, \phi_\star) \in \operatorname{argmax} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)$, *then* $[\mathrm{A}_x^\star]_{1:} = \theta_\star^\intercal$, $[\mathrm{A}_y^\star]_{1:} = \phi_\star^\intercal$.

**Extensions:** Similar results for

# Sliced Mutual Information & Processing (Cont.)

**Proposition (ZG-Greenewald'21)**

*Extracting maximum-SMI linear feature:*

$$\sup_{\mathrm{A}_x, \mathrm{A}_y, b_x, b_y} \mathsf{SI}(\mathrm{A}_x X + b_x; \mathrm{A}_y Y + b_y) = \sup_{\theta, \phi} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y).$$

*Also, if* $(\theta_\star, \phi_\star) \in \operatorname{argmax} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)$, *then* $[\mathrm{A}_x^\star]_{1:} = \theta_\star^\intercal$, $[\mathrm{A}_y^\star]_{1:} = \phi_\star^\intercal$.

**Extensions:** Similar results for

- Rank-constrained matrices

# Sliced Mutual Information & Processing (Cont.)

**Proposition (ZG-Greenewald'21)**

*Extracting maximum-SMI linear feature:*

$$\sup_{\mathrm{A}_x, \mathrm{A}_y, b_x, b_y} \mathsf{SI}(\mathrm{A}_x X + b_x; \mathrm{A}_y Y + b_y) = \sup_{\theta, \phi} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y).$$

*Also, if* $(\theta_\star, \phi_\star) \in \operatorname{argmax} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)$, *then* $[\mathrm{A}_x^\star]_{1:} = \theta_\star^\intercal$, $[\mathrm{A}_y^\star]_{1:} = \phi_\star^\intercal$.

**Extensions:** Similar results for

- Rank-constrained matrices
- Shallow NNs

# Sliced Mutual Information & Processing (Cont.)

**Proposition (ZG-Greenewald'21)**

*Extracting maximum-SMI linear feature:*

$$\sup_{\mathrm{A}_x, \mathrm{A}_y, b_x, b_y} \mathsf{SI}(\mathrm{A}_x X + b_x; \mathrm{A}_y Y + b_y) = \sup_{\theta, \phi} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y).$$

*Also, if* $(\theta_\star, \phi_\star) \in \operatorname{argmax} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y)$, *then* $[\mathrm{A}_x^\star]_{1:} = \theta_\star^\intercal$, $[\mathrm{A}_y^\star]_{1:} = \phi_\star^\intercal$.

**Extensions:** Similar results for

- Rank-constrained matrices
- Shallow NNs
- Analysis extends to other non-linear settings

# Experiments: Empirical Convergence

**Estimator:** $\widehat{\mathsf{SI}}_{m,n}$ with $k$-NN MI estimator [Kozachenko-Leonenko'87]

# Experiments: Empirical Convergence

**Estimator:** $\widehat{\mathsf{SI}}_{m,n}$ with $k$-NN MI estimator [Kozachenko-Leonenko'87]

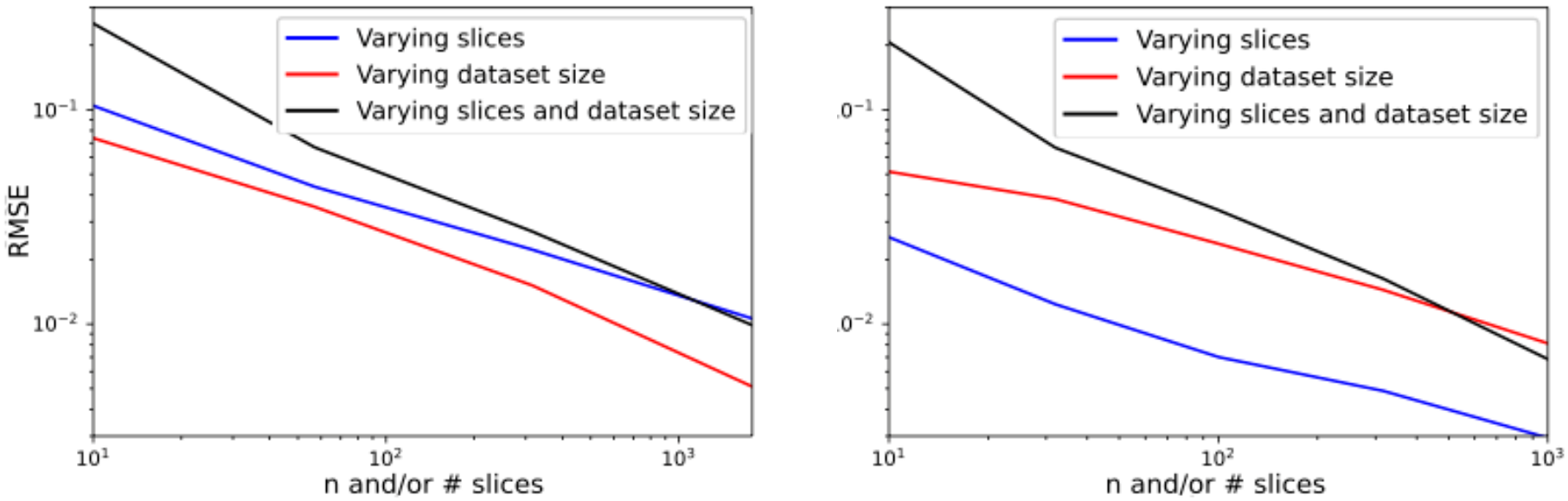
- Let $Z \sim \mathcal{N}(0, \mathrm{I}_{15})$ and define:
  - $\boldsymbol{d = 3}$**:** $X = Z_{[1:3]}$ **&** $Y = Z_{[2:4]}$
  - $\boldsymbol{d = 10}$**:** $X = Z_{[1:10]}$ **&** $Y = Z_{[5:15]}$

# Experiments: Empirical Convergence

**Estimator:** $\widehat{\mathsf{SI}}_{m,n}$ with $k$-NN MI estimator [Kozachenko-Leonenko'87]

- Let $Z \sim \mathcal{N}(0, \mathrm{I}_{15})$ and define:
  - $\boldsymbol{d = 3}$**:** $X = Z_{[1:3]}$ **&** $Y = Z_{[2:4]}$
  - $\boldsymbol{d = 10}$**:** $X = Z_{[1:10]}$ **&** $Y = Z_{[5:15]}$

**Figure: Empirical convergence rates**




$\boldsymbol{d = 3}$ $\qquad$ $\boldsymbol{d = 10}$