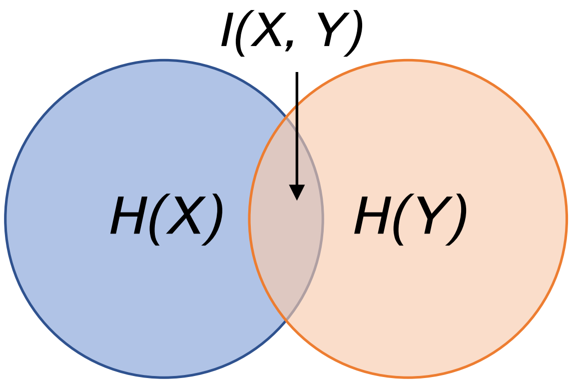
Understanding the Limitations of Variational Mutual Information Estimators

Jiaming Song, Stefano Ermon

Stanford University

Mutual Information (MI)

$$I(X;Y) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right]$$

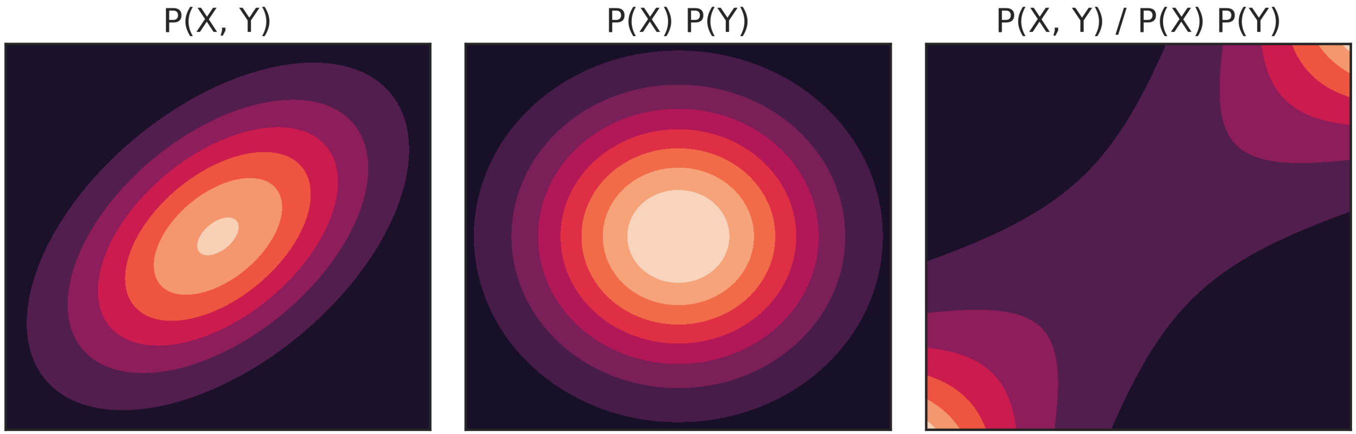


- Important in representation learning
  - e.g. X = data, Y = representation
  - Applied in InfoMax, MoCo, SimCLR…
- However, hard to estimate from samples

Variational MI Estimation

1. Estimate ratio $\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$

2. Obtain $I(X;Y) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right]$



The Generative Approach

Estimate via likelihood-based generative models:

- *P(X, Y)*
- *P(X), P(Y)*
- *P(X | Y), P(Y | X)*

Example: Barber-Agakov

$$I(X;Y) \geq \mathbb{E}_{P(X,Y)}\left[\log \frac{q(\boldsymbol{x}|\boldsymbol{y})}{p(\boldsymbol{x})}\right]$$

The Discriminative Approach

Estimate the ratio directly by discriminating
- *P(X, Y) samples*
- *P(X) P(Y) samples*

Example: MINE

$$I(X, Y) = \sup_{f} \mathbb{E}_{P(X,Y)}[f(\boldsymbol{x}, \boldsymbol{y})] - \log \mathbb{E}_{P(X)P(Y)}[e^{f(\boldsymbol{x},\boldsymbol{y})}]$$

$$f^* = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$$

A Unified View

$$I(X,Y) = \sup_{r\in\Delta} \mathbb{E}_{P(X,Y)}[\log r(\boldsymbol{x},\boldsymbol{y})]$$

Δ Is the set of valid density ratios over $P(X)P(Y)$

Different parametrizations over Δ

- Generative: *Barber-Agakov*
- Discriminative: *MINE, CPC, NWJ*

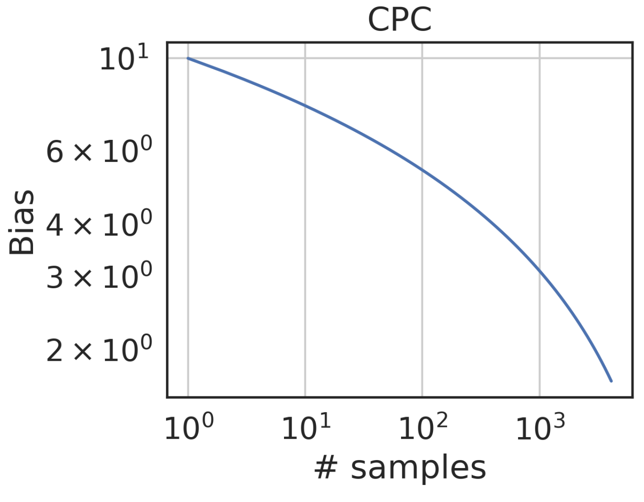
[BA] The IM algorithm: a variational approach to information maximization.

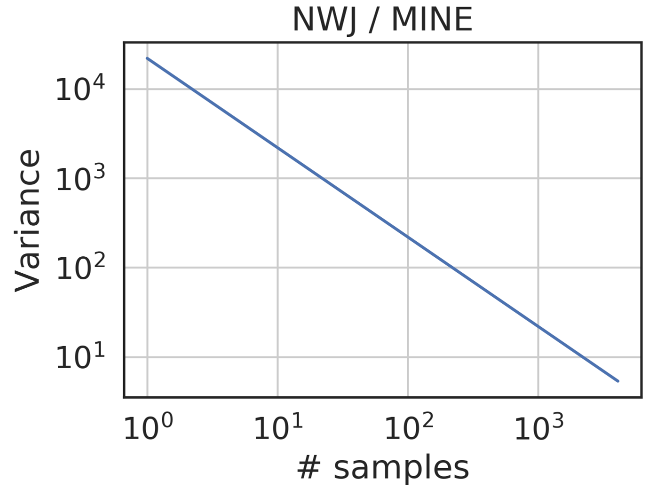
[MINE] Mutual Information Neural Estimation.

[CPC] Representation learning with contrastive predictive coding.

Limitation 1: Poor Sample Efficiency

- CPC: need $O\left(e^{I(X,Y)}\right)$ samples for low bias
- MINE: need $O\left(e^{I(X,Y)}\right)$ samples for low variance





(Assuming ground truth MI = 10)

Solution: SMILE

High variance in MINE comes from a sumexp term

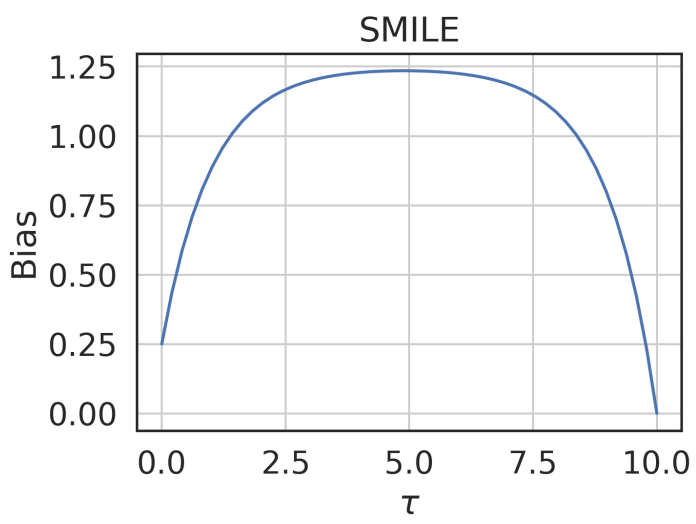
$$\mathbb{E}_{P(X)P(Y)}[e^{f(\boldsymbol{x},\boldsymbol{y})}]$$

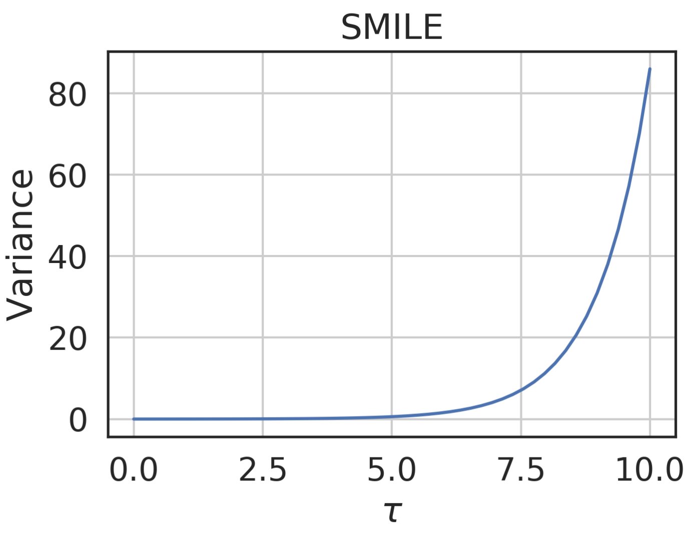
Smoothed Mutual Information Lower-bound Estimator

$$f(\boldsymbol{x},\boldsymbol{y}) \mapsto \mathrm{clip}(f(\boldsymbol{x},\boldsymbol{y}), -\tau, \tau)$$

Bias-Variance trade-off of SMILE

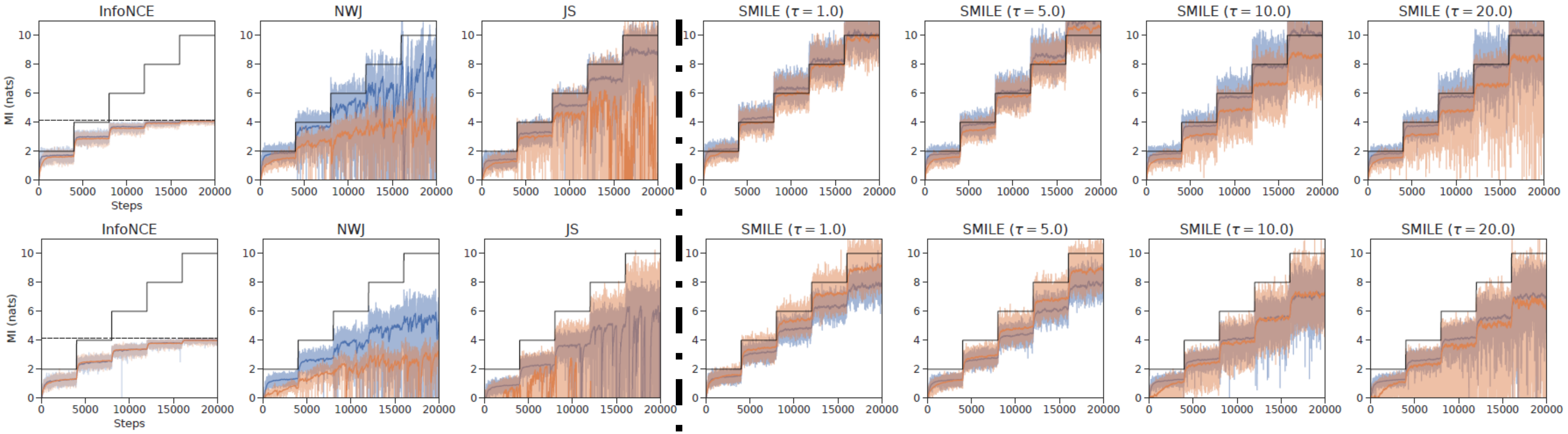
Significant decrease in variance without much bias!





(Assuming sumexp value = 1.25, batch size = 64, abs(f) < 10)

Experiment 1: MI estimation



Existing: high bias / variance

SMILE: better bias / variance tradeoff

(x = iterations, y = MI, black = ground truth, colored = architectures)

Limitation 2: Self-Consistency Issues

Why is MI appealing for representation learning?

1. MI = 0 for independent random variables (rvs).

$$I(X;Y) = 0 \quad \text{if} \quad X \perp Y$$

2. Data-processing does not increase information

$$I(X;Y) \geq I(X;Z) \quad \text{if} \quad X \to Y \to Z$$

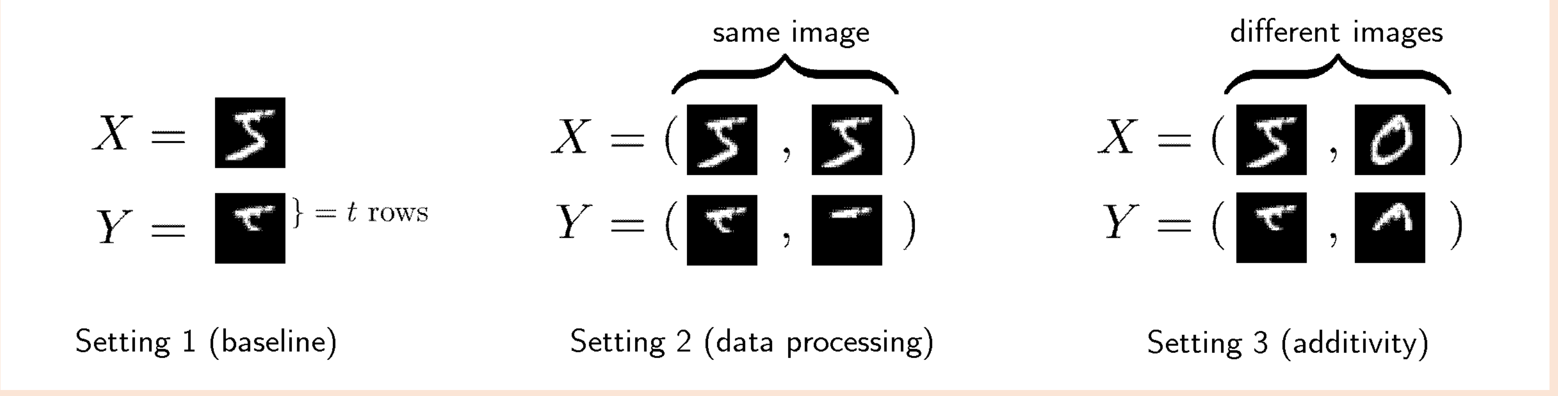
3. MI multiplies when we "concat" independent rvs.

$$I([X_1, X_2]; [Y_1, Y_2]) = 2I(X;Y)$$

Experiment 2: Consistency

Ideally, these properties should hold for MI estimators!

1. X and Y are independent -> estimate = 0
2. Estimate with more rows >= estimate with fewer rows
3. Estimate should add with more independent copies



Results

	Generative	Discriminative
Independence	✗	✓
Data-processing	✗	✓
Additivity	✓	✗

None of the approaches satisfy all the properties of MI!

Takeaway

Two limitations in variational MI estimators

- Need large batch sizes to obtain good estimates
- Does not satisfy consistency constraints of MI

Are variational MI estimators doing what we want?

Paper: https://arxiv.org/abs/1910.06222

Code: https://github.com/ermongroup/smile-mi-estimator

Contact: tsong@cs.stanford.edu