Formal Limitations on the Measurement of Mutual Information

Karl Stratos[1]

Joint work with David McAllester[2]

[1]Rutgers University

[2]Toyota Technological Institute at Chicago

Variational Estimation of Mutual Information

- Many recent works estimate mutual information by estimating a **lower bound** from samples
  - MINE (Belghazi et al., 2018), CPC (van den Oord et al., 2018), DIM (Hjelm et al., 2019), ...
  - Avoids direct estimation (difficult)

- **Q.** Is the estimate of a lower bound a faithful estimate of the underlying mutual information?

This Work

We show for *all* distribution-free high-confidence lower bounds B on mutual information $I(X, Y)$,

$$B((x_1, y_1) \ldots (x_N, y_N)) \;=\; O(\log N)$$

Fundamental statistical limitation

- **Implication.** Impossible to guarantee meaningfully large mutual information by estimating a lower bound
- We propose a **new estimator** that dodges this limitation.

Outline

Theorem

Proof sketch

- Entropy lemma
- Birthday paradox

Difference-of-entropies (DoE) estimator

Smallness of Lower Bounds on Mutual Information

Theorem. Let B be a bound such that for any $\mathbf{pop}_{XY}$,

$$I(X,Y) \geq B(S_N)$$

for $S_N \sim \mathbf{pop}_{XY}^N$ with probability at least 0.99. Then for any $\mathbf{pop}_{XY}$,

$$B(S_N) \leq 2\log N + 5$$

for $S_N \sim \mathbf{pop}_{XY}^N$ with probability at least 0.96, assuming $N \geq 50$.

Outline

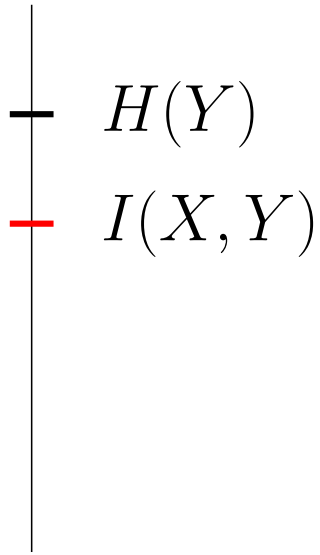
Theorem

Proof sketch

- Entropy lemma
- Birthday paradox

Difference-of-entropies (DoE) estimator

Reduction to Lower Bounding Entropy



$$I(X,Y) = H(Y) - H(Y|X)$$
$$\leq H(Y)$$

If large lower bounds on entropy are impossible, then large lower bounds on mutual information are impossible.

We will prove the stronger result!

Smallness of Lower Bounds on Entropy

Lemma. Let B be a bound such that for any $\mathbf{pop}_X$,

$$H(X) \geq B(S_N)$$

for $S_N \sim \mathbf{pop}_X^N$ with probability at least 0.99. Then for any $\mathbf{pop}_X$,

$$B(S_N) \leq 2 \log N + 5$$

for $S_N \sim \mathbf{pop}_X^N$ with probability at least 0.96, assuming $N \geq 50$.

Game Plan

If the support of P is small, entropy is small and the bound is already small by the premise.

$$B(S_N) \leq \underbrace{H(P)}_{\text{small}}$$

Thus assume the support of P is large.

- We will construct an **adversarial distribution** $\widetilde{P}$ with small entropy that is impossible to distinguish from P. Then again

$$B(S_N) \leq \underbrace{H\left(\widetilde{P}\right)}_{\text{small}}$$

Outline

Theorem

Proof sketch

- Entropy lemma
- Birthday paradox

Difference-of-entropies (DoE) estimator

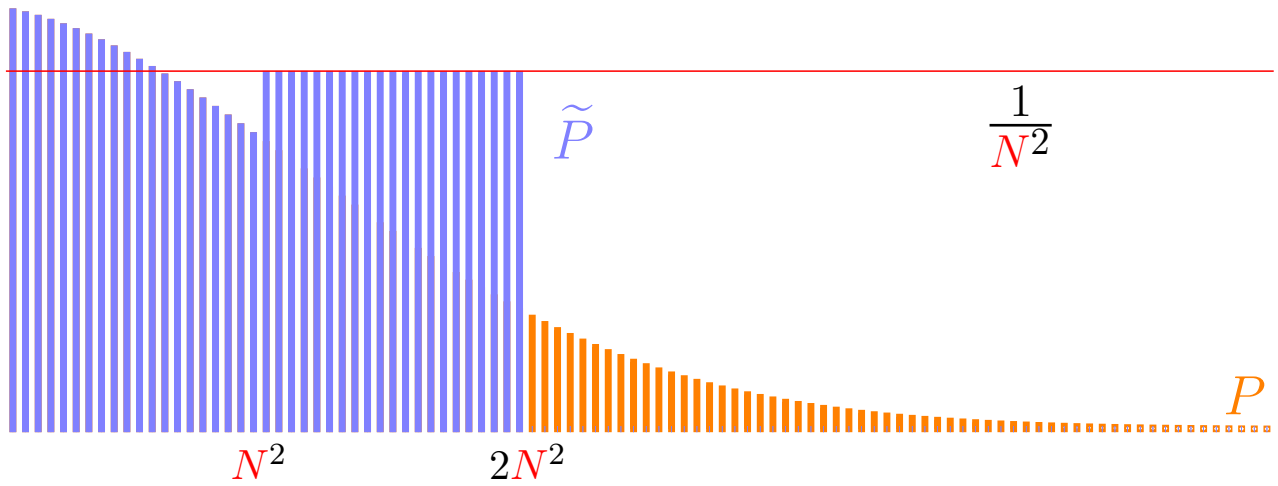
Birthday Paradox

If we draw

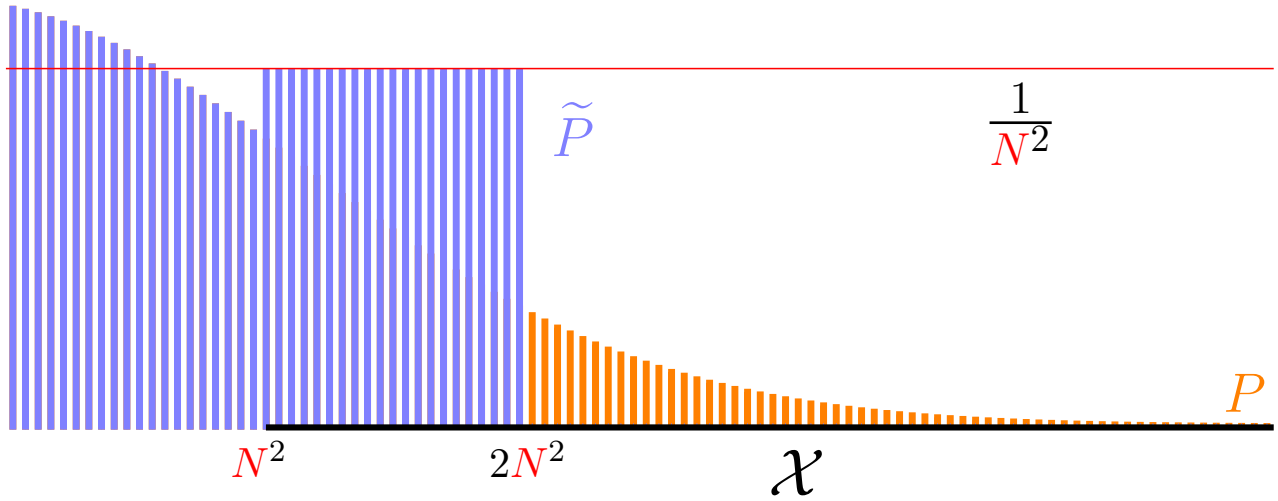
$$N < |\mathcal{X}|^2$$

samples from P, we probably **don't** have duplicates from $\mathcal{X}$, assuming each $x \in \mathcal{X}$ is sufficiently unlikely under P

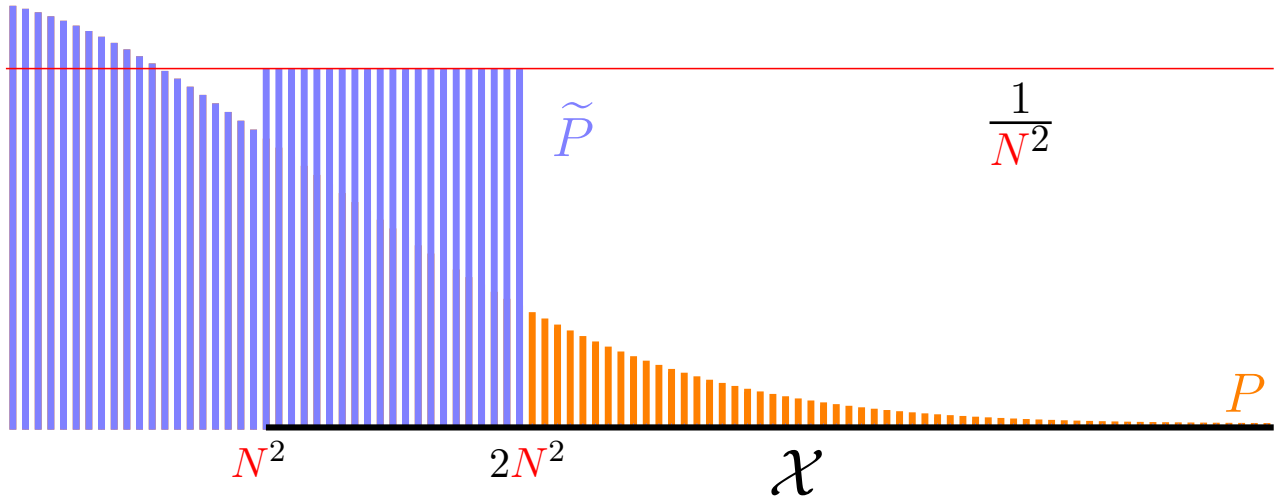
Construction of an Adversarial Distribution



Construction of an Adversarial Distribution



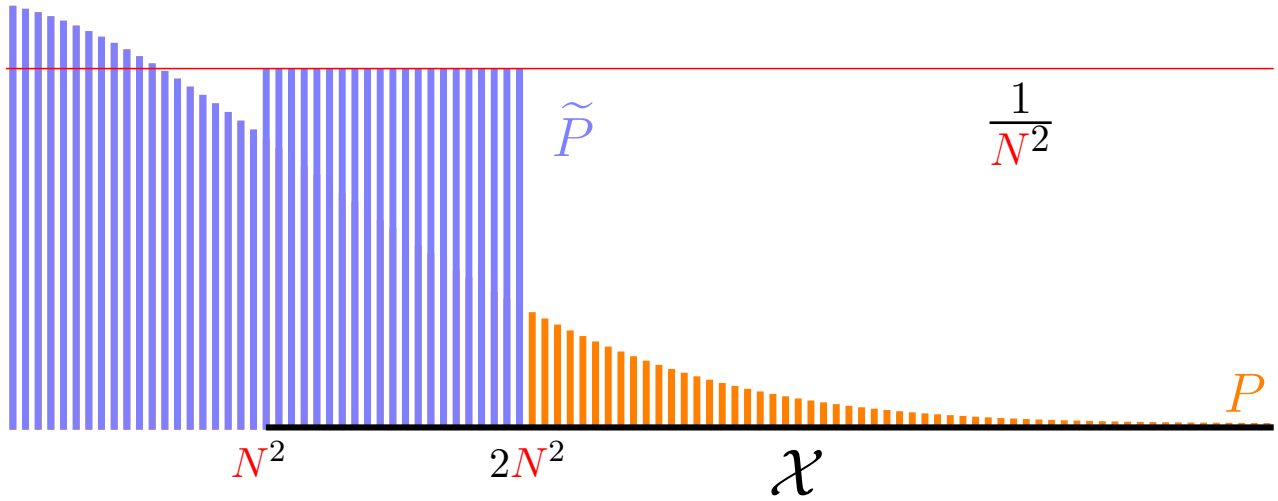
Construction of an Adversarial Distribution



For <u>either</u> $S \sim P^N$ <u>or</u> $\widetilde{S} \sim \widetilde{P}^N$

- We don't have duplicates from $\mathcal{X}$ (birthday paradox)

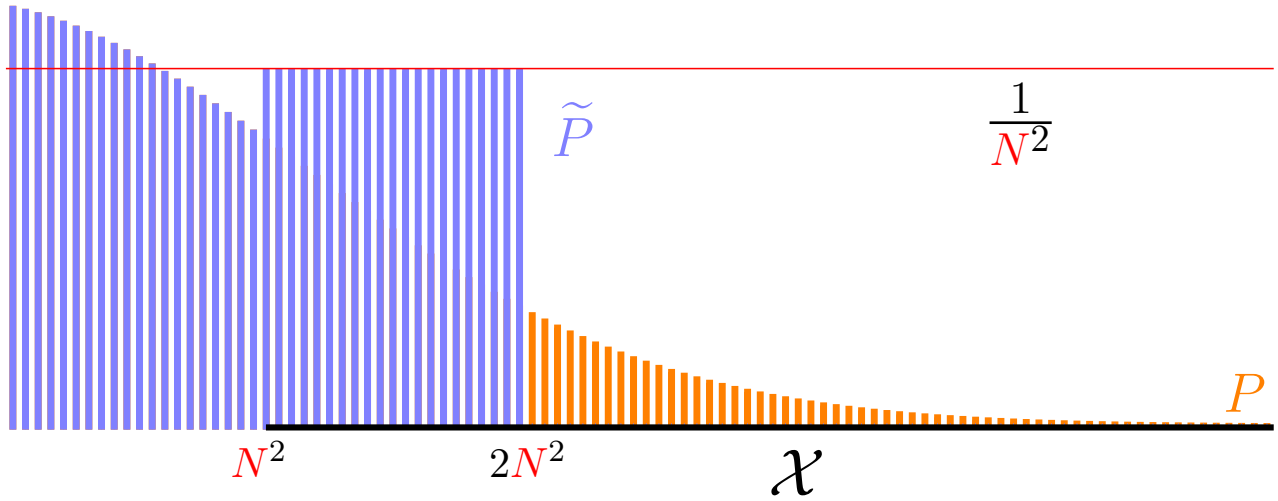
Construction of an Adversarial Distribution



For either $S \sim P^N$ or $\widetilde{S} \sim \widetilde{P}^N$

- We don't have duplicates from $\mathcal{X}$ (birthday paradox)
- In that case, we can't distinguish S from $\widetilde{S}$!

Construction of an Adversarial Distribution



For either $S \sim P^N$ or $\widetilde{S} \sim \widetilde{P}^N$

- We don't have duplicates from $\mathcal{X}$ (birthday paradox)
- In that case, we can't distinguish S from $\widetilde{S}$!
- **Premise.**

$$B(S_N) \leq H(\widetilde{P}) = O(\log N)$$

Importance of the Distribution-Free Assumption

- The lower bound is distribution-free, allowing us to construct an indistinguishable adversarial distribution with small entropy.

- Then we turn the premise (the lower bound cannot be larger than the true entropy) on its head!

Outline

Theorem

Proof sketch

- Entropy lemma
- Birthday paradox

Difference-of-entropies (DoE) estimator

DoE: Difference-of-Entropies Estimator

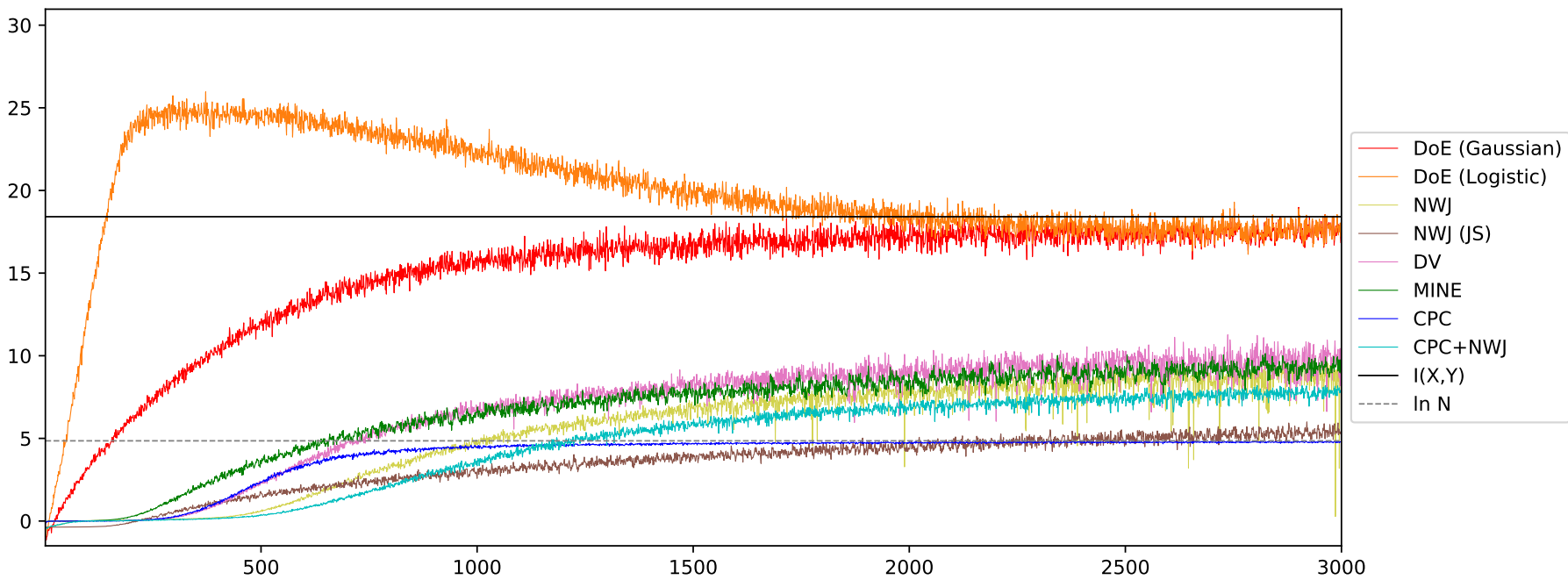
$$
\begin{aligned}
I(X,Y) &= H(P_Y) - H(P_{Y|X}) \\
&= \min_{Q_Y} H(P_Y, Q_Y) - \min_{T_{Y|X}} H(P_{Y|X}, T_{Y|X})
\end{aligned}
$$

- Gives neither an upper bound nor a lower bound
- But each entropy term can be measured effectively

Comparison with Lower Bounds on Mutual Information

x-**axis**: number of updates (128 samples each)

y-**axis**: estimate of mutual information (nats)



Takeaway

- Estimates of distribution-free high-confidence lower bounds on mutual information cannot be larger than the log of the sample size.

- We propose **DoE**: measuring mutual information as a difference of cross entropy upper bounds.

  Empirically much more accurate than lower bound estimators.

- Concurrent work: DoE for representation learning

  Learning Discrete Structured Representations by Adversarially Maximizing Mutual Information (ICML 2020)