The Statistics of Evaluating Generative Models with Divergence Frontiers

Krishna Pillutla

University of Washington

October 8, 2021 @ Science of DL

Joint Work With

Lang Liu

Sean Welleck

Sewoong Oh

Yejin Choi

Zaid Harchaoui

Outline

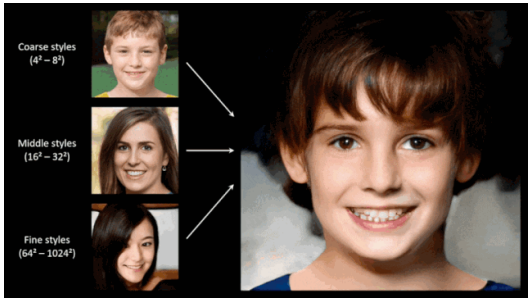
1 Introduction

2 Divergence Frontiers: Review

3 Statistics of Divergence Frontiers: Main Results

4 Experiments

Deep Generative Models

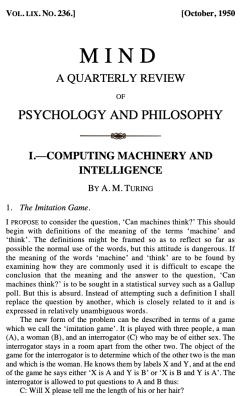


Prompt: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Machine Completion: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.
Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains · · ·

Ultimate Goal of Generative Models

- Generated images/text are indistinguishable from real-world images or human-written text
- Turing's Imitation Game a.k.a. The Turing Test
- "Can machines think?"



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Statistical Evaluation of Generative Models

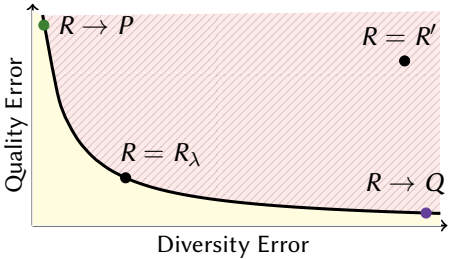
- Compare the *distribution* generated by the model with the target distribution
- **Quality**: Are the generated images or text good?
- **Diversity**: Is the model able to capture all of the target distribution?

Statistical Evaluation of Generative Models

- Compare the *distribution* generated by the model with the target distribution
- **Quality**: Are the generated images or text good?
- **Diversity**: Is the model able to capture all of the target distribution?
- *Divergence Frontiers* are one such a framework

Divergence Frontiers

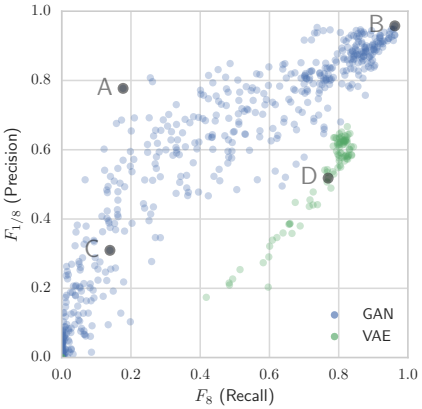
- ROC-like operating characteristics of a generative model Q w.r.t. target distribution P
- *Softly* measure quality and diversity
- R is an auxiliary distribution used to define the frontier (details later)



Introduced by Sajjadi et. al. (NeurIPS 2018), formalized by Djolonga et. al. (AISTATS 2020)

Divergence Frontiers in Vision

- Can quantify mode-dropping and mode-inventing in GANs
- Can quantify that GANs tend to produce higher quality and less diverse images than VAEs
- Quality $\equiv$ Precision, Diversity $\equiv$ Recall

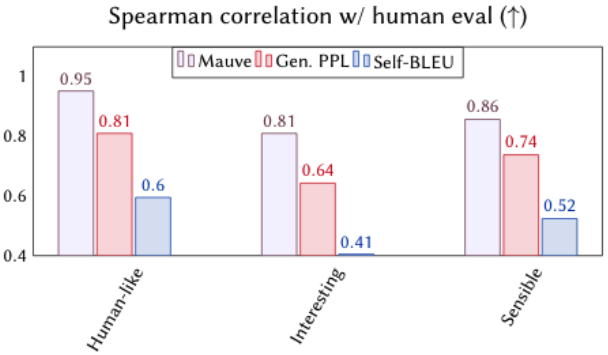


Sajjadi et. al. (NeurIPS 2018)

Divergence Frontiers in NLP

Mauve: compare open-ended text generation models:

- Strong correlation with human judgements
- Can quantify the effect of
  - model size
  - decoding algorithms
  - generation length

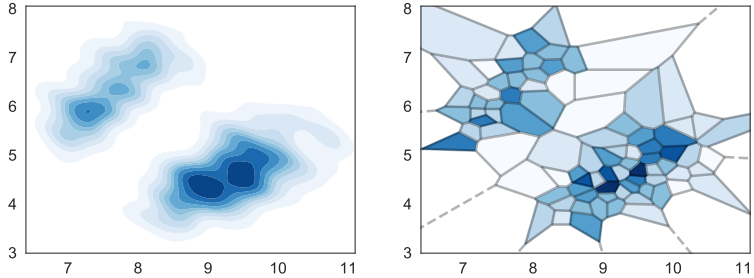


Pillutla et. al. (NeurIPS 2021)

Computation of Divergence Frontiers in Practice

Computing KL divergences between high-dimensional distributions is challenging. Use two approximations in practice:

- **Quantization**: Quantize high dimensional distributions into k-dimensional multinomial distributions: $P, Q \mapsto P_S, Q_S$ where $|S| = k$
- **Estimation**: Estimate using n samples each from P_S and Q_S using the plug-in estimate



Our Contributions

We analyze the error of this procedure:

- **Quantization Error** is $O(1/k)$
- **Estimation Error** is $O(\sqrt{k/n})$

Empirical insights from the theory:

- Use smoothed estimators instead of the usual plug-in estimator
- Theoretical guidance on quantization size $k = n^{-1/3}$

Outline

1 Introduction

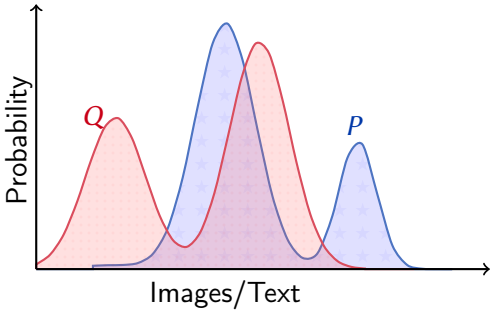
2 Divergence Frontiers: Review

3 Statistics of Divergence Frontiers: Main Results

4 Experiments

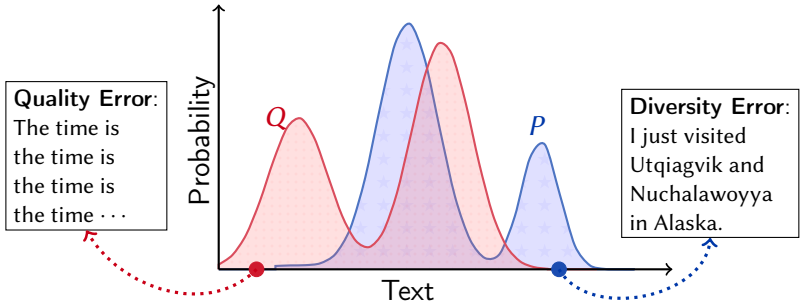
Modeling P with Q

- Denote P for the target distribution and Q for the model distribution



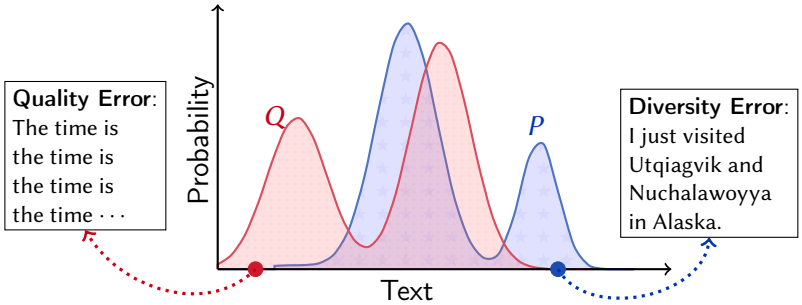
Modeling P with Q: Evaluation

- Denote P for the target distribution and Q for the model distribution
- **Quality error**: Q places high mass on regions unlikely under P
- **Diversity error**: Q cannot produce text/images plausible under P



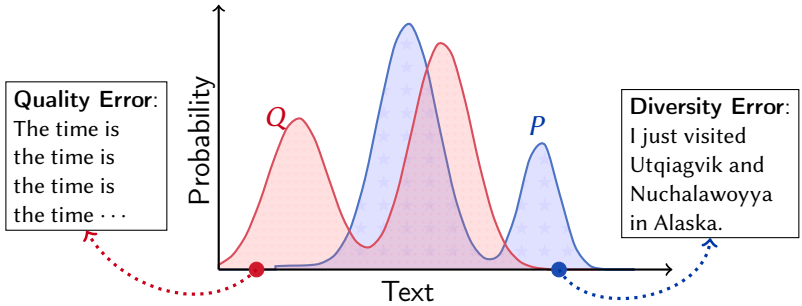
Modeling P with Q: Evaluation

- Denote P for the target distribution and Q for the model distribution
- **Quality error**: Q places high mass on regions unlikely under P: $\mathrm{KL}(Q|P)$?
- **Diversity error**: Q cannot produce text/images plausible under P: $\mathrm{KL}(P|Q)$?



Modeling P with Q: Mixture Distribution and Interpolated KL

- Use the mixture distribution $R_\lambda = \lambda P + (1-\lambda)Q$ for some λ between 0 and 1
- Diversity error is $\mathrm{KL}_\lambda(P|Q) := \mathrm{KL}(P|\lambda P + (1-\lambda)Q)$
- Quality error is $\mathrm{KL}_{1-\lambda}(Q|P) := \mathrm{KL}(Q|\lambda P + (1-\lambda)Q)$



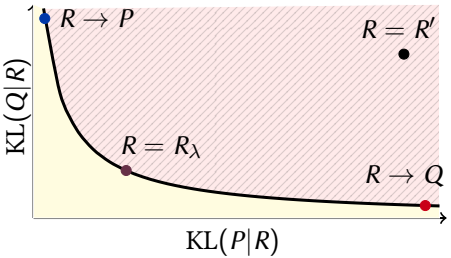
Divergence Frontiers: Definition

- Since $\lambda \in (0, 1)$ is not fixed, varying it gives a parametric curve $\mathcal{F}(P, Q)$ with

$$x(\lambda) = \mathrm{KL}_\lambda(P|Q), \quad y(\lambda) = \mathrm{KL}_{1-\lambda}(Q|P)$$

- It is called the *divergence frontier* because it is the Pareto frontier of the multi-objective optimization

$$\min_R \Big(\mathrm{KL}(P|R), \mathrm{KL}(Q|R)\Big)$$



Outline

1 Introduction

2 Divergence Frontiers: Review

3 Statistics of Divergence Frontiers: Main Results

4 Experiments

Our Goal

- **Estimation Error**: What is the error in estimating the divergence frontier $\mathcal{F}(P, Q)$ given n samples from *mulitnomial distributions* P, Q?
- **Quantization Error**: We quantize P, Q to get P_S, Q_S by some means. What is the closest $\mathcal{F}(P_S, Q_S)$ can be to $\mathcal{F}(P, Q)$?

Estimation Error of the Frontier with the Plug-In Estimator

Denote

- True population frontier: $\mathcal{F}(P,Q) = \left\{ \big(x(\lambda), y(\lambda)\big) \ : \ \lambda \in (0,1) \right\}$
- Plug-in estimate of the frontier: $\mathcal{F}(\hat{P}_n, \hat{Q}_n) = \left\{ \big(\hat{x}_n(\lambda), \hat{y}_n(\lambda)\big) \ : \ \lambda \in (0,1) \right\}$

Theorem

If the support size of P and Q is k, then,

$$\mathbb{E}\left[\sup_{\lambda \in [\lambda_n, 1-\lambda_n]} \left\| \big(\hat{x}_n(\lambda), \hat{y}_n(\lambda)\big) - \big(x(\lambda), y(\lambda)\big) \right\|_1 \right] \lesssim \frac{\log n}{\lambda_n} \left(\sqrt{\frac{k}{n}} + \frac{k}{n} \right) .$$

Estimation Error of the Frontier with the Plug-In Estimator

Denote

- True population frontier: $\mathcal{F}(P, Q) = \left\{ \big(x(\lambda), y(\lambda)\big) \ : \ \lambda \in (0,1) \right\}$
- Plug-in estimate of the frontier: $\mathcal{F}(\hat{P}_n, \hat{Q}_n) = \left\{ \big(\hat{x}_n(\lambda), \hat{y}_n(\lambda)\big) \ : \ \lambda \in (0,1) \right\}$

Theorem

If the support size of P and Q is k, then,

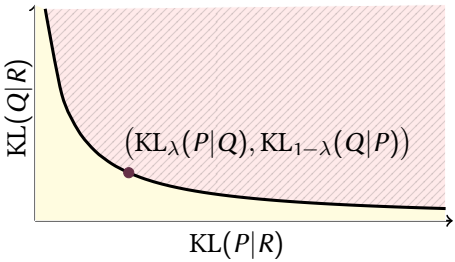
$$\mathbb{E}\left[\sup_{\lambda \in [\lambda_n, 1-\lambda_n]} \left\| \big(\hat{x}_n(\lambda), \hat{y}_n(\lambda)\big) - \big(x(\lambda), y(\lambda)\big) \right\|_1 \right] \lesssim \frac{\log n}{\lambda_n} \left(\sqrt{\frac{k}{n}} + \frac{k}{n} \right) .$$

Truncation is necessary because $\mathrm{KL}_\lambda \to \mathrm{KL}$ as $\lambda \to 0$ and minimax error of KL estimation is ∞ without boundedness assumptions (Bu, Zou, Liang & Veeravalli, IEEE Trans. Inf. Theory, 2018)

Integral Summary of the Divergence Frontier

Summarize the entire divergence curve in a scalar called the **Frontier Integral**:

$$\mathrm{FI}(P, Q) = 2 \int_0^1 \left(\lambda \, \mathrm{KL}_\lambda(P|Q) + (1-\lambda)\, \mathrm{KL}_{1-\lambda}(Q|P)\right) d\lambda$$



Integral Summary of the Divergence Frontier

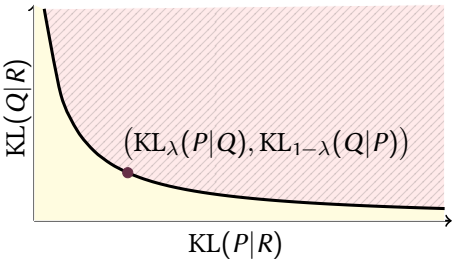
Summarize the entire divergence curve in a scalar called the **Frontier Integral**:

$$\mathrm{FI}(P, Q) = 2 \int_0^1 \left(\lambda \,\mathrm{KL}_\lambda(P|Q) + (1-\lambda)\,\mathrm{KL}_{1-\lambda}(Q|P)\right) d\lambda$$

- Linear combination of Quality ($\equiv$ Type-I) error and Diversity ($\equiv$ Type-II) error
- Integrand is

$$\min_R \left\{\lambda \,\mathrm{KL}(P|R) + (1-\lambda)\,\mathrm{KL}(Q|R)\right\}$$

- FI is a symmetric f-divergence



Improving Estimation with Smoothing

When the support size k is large, the statistical performance can be improved with add-constant or Good-Turing estimators.

Theorem

Let P, Q have a support size of $k < \infty$. We have,

$$\mathbb{E}\left|\mathrm{FI}(\hat{P}, \hat{Q}) - \mathrm{FI}(P, Q)\right| \lesssim \begin{cases} \sqrt{\frac{k}{n}} \log n\,, & \text{with the plug-in estimator} \\ \frac{\sqrt{nk}+bk}{n+bk} \log\left(n/b + k\right), & \text{with the add-b estimator} \end{cases}$$

- Also: distribution dependent bounds, independent of k for the plug-in
- High probability bounds

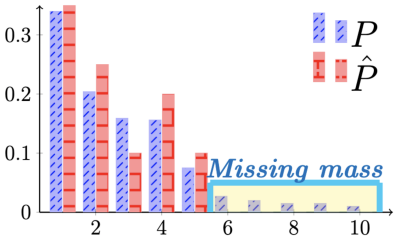
Missing Mass + Benefit of Smoothing

The bounds apply to long-tailed distributions

- no dependence on $\min_a P(a)$
- requires careful analysis of the missing mass

When k/n is large, smoothing is better:

- Plug-in: $O(k \log n/n)$.
- Add-b: $O(\log n + \log(k/n))$



Quantization Bound

Proposition

Let $\mathcal{X}$ be an arbitrary measurable space. There exists a distribution-dependent partition S_k of $\mathcal{X}$ with level $|S_k| = k$ with

$$\left|\mathrm{FI}(P, Q) - \mathrm{FI}(P_{S_k}, Q_{S_k})\right| \leq C\, k^{-1}.$$

- Overall error of estimation + quantization: $\tilde{O}(\sqrt{k/n} + 1/k)$. Balance errors at $k \asymp n^{1/3}$ so that the total error is $\tilde{O}(n^{-1/3})$

Quantization Bound

Proposition

Let $\mathcal{X}$ be an arbitrary measurable space. There exists a distribution-dependent partition S_k of $\mathcal{X}$ with level $|S_k| = k$ with

$$\left|\mathrm{FI}(P, Q) - \mathrm{FI}(P_{S_k}, Q_{S_k})\right| \leq C\, k^{-1}.$$

- Overall error of estimation + quantization: $\tilde{O}(\sqrt{k/n} + 1/k)$. Balance errors at $k \asymp n^{1/3}$ so that the total error is $\tilde{O}(n^{-1/3})$
- In practice, use data-dependent quantization with deep networks but theory out of reach
- Non-parametric density estimation can give data-dependent quantization schemes with theoretical guarantees. Do not work well empirically due to curse of dimensionality

Extensions

Extensions of f-divergences (with conjugate $f^\star(x) = xf(1/x)$). Assume:

- Boundedness: $f(0) + f^\star(0) < \infty$
- Slow growth: $f'(t) \propto -\log t^{-1}$ as $t \to 0$ and same for $(f^\star)'$
- Technical condition on 2nd derivative

Assumptions are satisfied by KL_λ, FI, interpolated χ^2, etc.

Extensions

Extensions of f-divergences (with conjugate $f^\star(x) = xf(1/x)$). Assume:

- Boundedness: $f(0) + f^\star(0) < \infty$
- Slow growth: $f'(t) \propto -\log t^{-1}$ as $t \to 0$ and same for $(f^\star)'$
- Technical condition on 2nd derivative

Assumptions are satisfied by KL_λ, FI, interpolated χ^2, etc. Then, we have bounds on:

- Estimation error $|D_f(\hat{P}|\hat{Q}) - D_f(P|Q)|$ for plug-in and add-b estimators
- + High probability bounds
- Quantization error $|D_f(P_S|Q_S) - D_f(P|Q)|$

Extensions

Extensions of f-divergences (with conjugate $f^\star(x) = xf(1/x)$). Assume:

- Boundedness: $f(0) + f^\star(0) < \infty$
- Slow growth: $f'(t) \propto -\log t^{-1}$ as $t \to 0$ and same for $(f^\star)'$
- Technical condition on 2nd derivative

Assumptions are satisfied by KL_λ, FI, interpolated χ^2, etc. Then, we have bounds on:

- Estimation error $|D_f(\hat{P}|\hat{Q}) - D_f(P|Q)|$ for plug-in and add-b estimators
- + High probability bounds
- Quantization error $|D_f(P_S|Q_S) - D_f(P|Q)|$

Proofs are elementary once we have these assumptions! Based on Taylor expansions

Outline

1 Introduction

2 Divergence Frontiers: Review

3 Statistics of Divergence Frontiers: Main Results

4 Experiments

Experiments

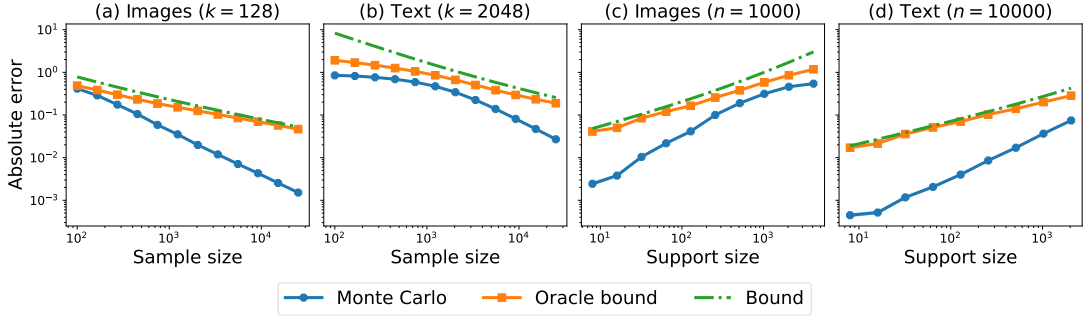
Goals: Are the bounds tight? What practical insights can we extract from the theory?
Gold standard: Measure absolute error with the ground-truth for the synthetic case, and with Monte-Carlo estimate otherwise

Setting

- Synthetic data (discrete): Zipf(r) distribution with $P(i) = i^{-r}$
- Real data:
  - Image: CIFAR-10 + StyleGAN
  - Text: GPT-2 + WikiText-103

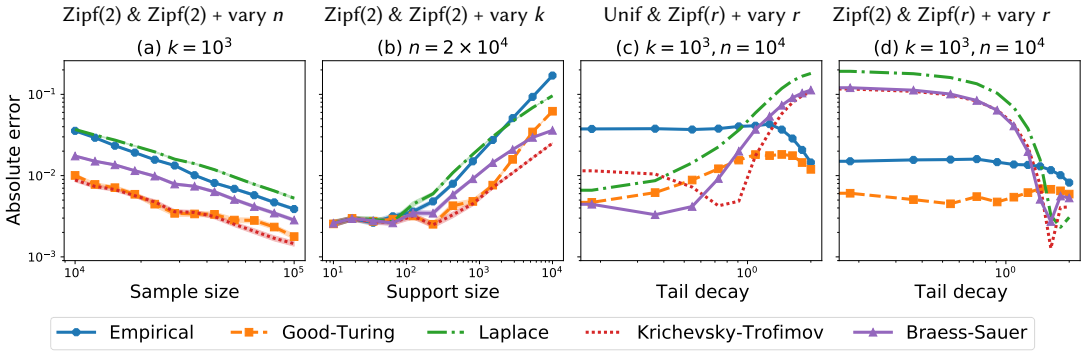
Estimation Bounds are Tight

Real Data



Smoothed Estimators Help

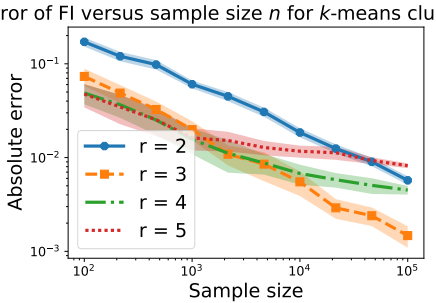
Synthetic Data



Krichevsky-Trofimov (add-1/2) is a good default choice

Theoretical Guidance on Quantization Level is Correct

Synthetic Data: $P = N(0, I)$ and $Q = N(1, I)$ in $\mathbb{R}^2$



Theoretical guidance of $k = n^{1/3}$ works the best empirically as well

Conclusion

- Statistical analysis of divergence frontiers: bounds on both estimation and quantization errors
- Empirically:
  - ▷ Bounds capture empirical behavior (real and synthetic)
  - ▷ Smoothed estimators work better
  - ▷ Theoretical guidance on quantization level is correct

Thank you

Thank you!

Please email questions to pillutla@cs.washington.edu.

Estimation Error of the Frontier Integral with the Plug-In Estimate

Theorem

Let the support of P, Q have size $k < \infty$. We have,

$$\mathbb{E}\left|\mathrm{FI}(\hat{P}_n, \hat{Q}_n) - \mathrm{FI}(P, Q)\right| \lesssim \left(\sqrt{\frac{k}{n}} + \frac{k}{n}\right) \log n$$

Estimation Error of the Frontier Integral with the Plug-In Estimate

Theorem

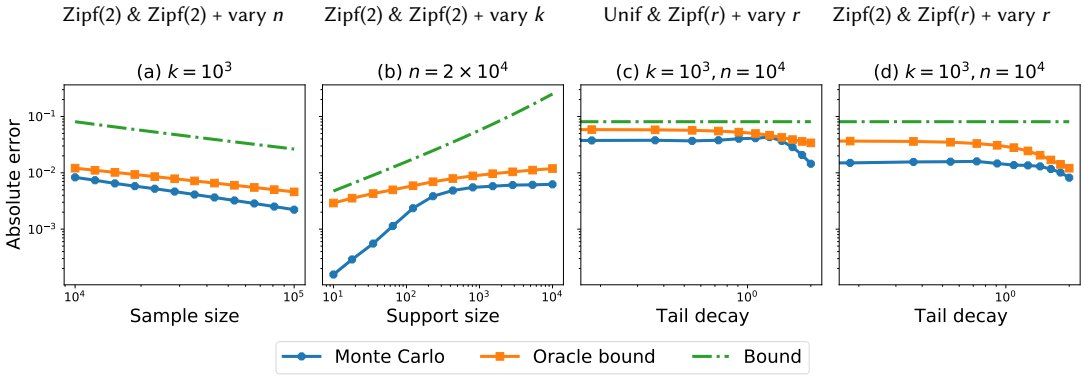
Let the support of P, Q have size $k < \infty$. We have,

$$\mathbb{E}\left|\mathrm{FI}(\hat{P}_n, \hat{Q}_n) - \mathrm{FI}(P, Q)\right| \lesssim \left(\sqrt{\frac{k}{n}} + \frac{k}{n}\right) \log n$$

- Bound does not depend on $\min_a P(a)$, so it is good for long-tailed distributions; account for missing mass to achieve this
- Can give a distribution-dependent bound, applicable for $k = \infty$
- Parametric rate $\tilde{O}(1/\sqrt{n})$, tight for KL estimation w/ bounded distributions.

Estimation Bounds are Tight

Synthetic Data



Smoothed Estimators Help

Real Data