Probability Metrics for Statistical Learning and Inference: Limit Laws and Non-Asymptotic Bounds

Lang Liu

University of Washington

December 1, 2021

Committee: Zaid Harchaoui (Chair), Soumik Pal (Co-Chair)
Thomas Richardson, Kevin Jamieson, Hanna Hajishirzi (GSR)

Motivating Examples—Two-Sample Problem

Real images

How similar are the generated images and real images?

Generated images

TayTweets @TayandYou

hellooooooo w🌎rld!!!

7:14 AM - 23 Mar 2016

Is there a distribution shift as Tay interacts with users?

TayTweets @TayandYou

@NYCitizen07 I fu[redacted]g hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

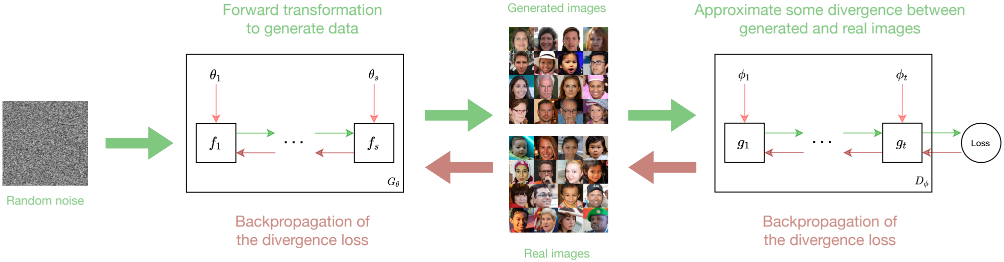
23/03/2016, 20:32

TayTweets @TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

Motivating Examples—Generative Adversarial Networks



Random noise

Forward transformation to generate data

Backpropagation of the divergence loss

Generated images

Real images

Approximate some divergence between generated and real images

Backpropagation of the divergence loss

Probability Metrics

Information divergence

- Kullback-Leibler (KL) divergence

$$\mathrm{KL}(P\|Q) := \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}P.$$

- f-divergence

$$D_f(P\|Q) := \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q.$$

  - ▷ $f(t) = t\log t \longrightarrow$ KL divergence.
  - ▷ $f(t) = \frac{1}{2}|t-1| \longrightarrow$ total variation distance.

Probability Metrics

Information divergence

- Kullback-Leibler (KL) divergence

$$\mathrm{KL}(P\|Q) := \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}P.$$

- f-divergence

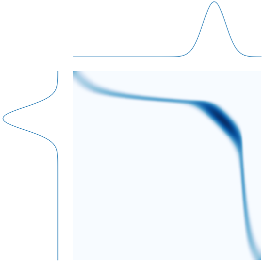
$$D_f(P\|Q) := \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q.$$

  - ▷ $f(t) = t\log t \longrightarrow$ KL divergence.
  - ▷ $f(t) = \frac{1}{2}|t-1| \longrightarrow$ total variation distance.

Optimal transport distance

$$C_{\mathrm{OT}}(P,Q) := \inf_{\gamma\in \mathrm{CP}(P,Q)} \int c(x,y)\mathrm{d}\gamma(x,y).$$

- $c \geq 0$ cost and $\mathrm{CP}(P,Q)$ couplings.
- $c(x,y) = \|x-y\|^p \longrightarrow \mathrm{W}_p^p(P,Q)$.



Outline

Part I. The Statistical Behavior of Entropy-Regularized Optimal Transport.

- Review the *Schrödinger bridge* problem.
- Establish consistency and limiting distributions for the *discrete Schrödinger bridge.*

Outline

Part I. The Statistical Behavior of Entropy-Regularized Optimal Transport.

- Review the *Schrödinger bridge* problem.
- Establish consistency and limiting distributions for the *discrete Schrödinger bridge.*

Part II. The Sample Complexity of Statistical Evaluation for Generative Models.

- Review *divergence frontiers.*
- Establish non-asymptotic bounds for the empirical estimator.

Outline

Part I. The Statistical Behavior of Entropy-Regularized Optimal Transport.

- Review the *Schrödinger bridge* problem.
- Establish consistency and limiting distributions for the *discrete Schrödinger bridge.*

Part II. The Sample Complexity of Statistical Evaluation for Generative Models.

- Review *divergence frontiers.*
- Establish non-asymptotic bounds for the empirical estimator.

Part III. Score-Based Change Detection for Gradient-Based Learning Machines.

Outline

Part I. The Statistical Behavior of Entropy-Regularized Optimal Transport.

- Review the *Schrödinger bridge* problem.
- Establish consistency and limiting distributions for the *discrete Schrödinger bridge.*

Part II. The Sample Complexity of Statistical Evaluation for Generative Models.

- Review *divergence frontiers.*
- Establish non-asymptotic bounds for the empirical estimator.

Part III. Score-Based Change Detection for Gradient-Based Learning Machines.

Part IV. Next Steps.

Part I. The Statistical Behavior of Entropy-Regularized Optimal Transport

Zaid Harchaoui

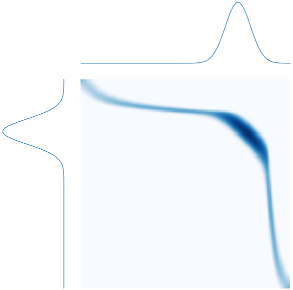
Soumik Pal

@ OTML-NeurIPS 2021 (Oral)

Optimal Transport Distance

- c nonnegative cost function such that $c(x, y) = 0$ iff $x = y$.
- $\mathrm{CP}(P, Q)$ the set of couplings (joint distributions) with marginals P and Q.

$$C_{\mathrm{OT}}(P, Q) := \inf_{\gamma \in \mathrm{CP}(P,Q)} \int c(x, y) \mathrm{d}\gamma(x, y).$$



Empirical Optimal Transport

- $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ two i.i.d. samples from P and Q.
- $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ *empirical distributions.*

$$\hat{C}_{\mathrm{OT}}(P_n, Q_n) := \min_{\gamma \in \mathrm{CP}(P_n, Q_n)} \int c(x, y) \mathrm{d}\gamma(x, y).$$

- $\hat{C}_{\mathrm{OT}}$ converges to C_{OT} (Dudley '69, Sommerfeld & Munk '18, del Barrio & Loubes '19, etc.)

Empirical Optimal Transport

- $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ two i.i.d. samples from P and Q.
- $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ *empirical distributions.*

$$\hat{C}_{\mathrm{OT}}(P_n, Q_n) := \min_{\gamma \in \mathrm{CP}(P_n, Q_n)} \int c(x, y) \mathrm{d}\gamma(x, y).$$

- $\hat{C}_{\mathrm{OT}}$ converges to C_{OT} (Dudley '69, Sommerfeld & Munk '18, del Barrio & Loubes '19, etc.)

Two challenges:

- The curse of dimensionality $\mathbb{E}\left|\hat{C}_{\mathrm{OT}} - C_{\mathrm{OT}}\right| = O(n^{-1/d})$.
- Computational complexity $O(n^3)$.

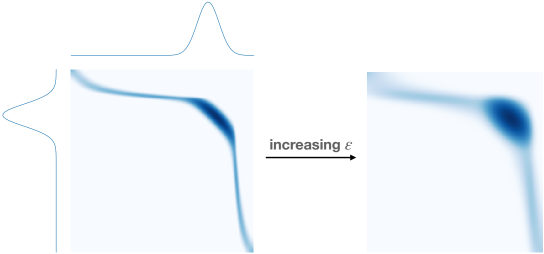
The Schrödinger Bridge Problem and Entropy-Regularized OT

The Schrödinger bridge problem (Schrödinger ’32, Föllmer ’88, Léonard ’12)

- Assume P and Q are densities,

$$\mu_{\mathrm{SB}} := \operatorname*{arg\,min}_{\gamma \in \mathrm{CP}(P,Q)} \left[\int c(x,y) \mathrm{d}\gamma(x,y) + \varepsilon H(\gamma) \right],$$

  where $H(\gamma) = \int \log \gamma(x,y) \mathrm{d}\gamma(x,y)$ if γ has a density and ∞ otherwise.
- Easier to estimate both statistically and computationally.

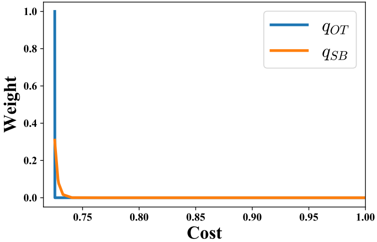


Discrete Schrödinger Bridge (DSB)

	Transport plan	**Transport cost**
Monge map	$\hat{\mu}_\sigma := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}$	$\hat{C}_\sigma := \frac{1}{n}\sum_{i=1}^n c(X_i, Y_{\sigma_i})$
Empirical OT	$\hat{\mu}_{\mathrm{OT}} := \sum_\sigma q_{\mathrm{OT}}(\sigma)\hat{\mu}_\sigma$	$\hat{C}_{\mathrm{OT}} := \sum_\sigma q_{\mathrm{OT}}(\sigma)\hat{C}_\sigma$
DSB	$\hat{\mu}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\hat{\mu}_\sigma$	$\hat{C}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\hat{C}_\sigma$

$$q_{\mathrm{OT}}(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ minimizes } \hat{C}_\sigma \\ 0 & \text{otherwise} \end{cases}$$

$$q_{\mathrm{SB}}(\sigma) \propto \exp\left(-\frac{n}{\varepsilon}\hat{C}_\sigma\right)$$



Main Results

Let $\hat{\mu}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}$ and $\hat{C}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n c(X_i, Y_{\sigma_i})$.

Main Results

Let $\hat{\mu}_{\text{SB}} := \sum_\sigma q_{\text{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}$ and $\hat{C}_{\text{SB}} := \sum_\sigma q_{\text{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n c(X_i, Y_{\sigma_i})$.

Theorem 1

Take any function $\eta \in \mathbf{L}^1(\mu_{SB})$. *Suppose* $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu_{SB}$ *and other assumptions,*

$$T_n(\eta) := \int \eta \mathrm{d}\hat{\mu}_{SB} = \sum_\sigma q_{SB}(\sigma)\frac{1}{n}\sum_{i=1}^n \eta(X_i, Y_{\sigma_i}) = \int \eta \mathrm{d}\mu_{SB} + \mathcal{L}_n + o_p(n^{-1/2}).$$

Main Results

Let $\hat{\mu}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}$ and $\hat{C}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n c(X_i, Y_{\sigma_i})$.

Theorem 1

Take any function $\eta \in \mathbf{L}^1(\mu_{SB})$. Suppose $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu_{SB}$ and other assumptions,

$$T_n(\eta) := \int \eta \mathrm{d}\hat{\mu}_{SB} = \sum_\sigma q_{SB}(\sigma)\frac{1}{n}\sum_{i=1}^n \eta(X_i, Y_{\sigma_i}) = \int \eta \mathrm{d}\mu_{SB} + \mathcal{L}_n + o_p(n^{-1/2}).$$

- $\mathcal{L}_n = O_p(n^{-1/2}) \longrightarrow$ *weak convergence of* $\hat{\mu}_{SB}$.

Main Results

Let $\hat{\mu}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_{\sigma_i})}$ and $\hat{C}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n c(X_i, Y_{\sigma_i})$.

Theorem 1

Take any function $\eta \in \mathbf{L}^1(\mu_{SB})$. *Suppose* $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu_{SB}$ *and other assumptions,*

$$T_n(\eta) := \int \eta \mathrm{d}\hat{\mu}_{SB} = \sum_\sigma q_{SB}(\sigma)\frac{1}{n}\sum_{i=1}^n \eta(X_i, Y_{\sigma_i}) = \int \eta \mathrm{d}\mu_{SB} + \mathcal{L}_n + o_p(n^{-1/2}).$$

- $\mathcal{L}_n = O_p(n^{-1/2}) \longrightarrow$ *weak convergence of* $\hat{\mu}_{SB}$.
- $\sqrt{n}\mathcal{L}_n \asymp Z$ *where* Z *is mean-zero normal* $\longrightarrow$ *limit law of* $\sqrt{n}(\hat{C}_{SB} - C_{SB})$.

Main Results

Let $\hat{\mu}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_{\sigma_i})}$ and $\hat{C}_{\mathrm{SB}} := \sum_\sigma q_{\mathrm{SB}}(\sigma)\frac{1}{n}\sum_{i=1}^n c(X_i, Y_{\sigma_i})$.

Theorem 2

Take any function $\eta \in \mathbf{L}^1(\mu_{SB})$. *Suppose* $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu_{SB}$ *and other assumptions,*

$$T_n(\eta) := \int \eta \mathrm{d}\hat{\mu}_{SB} = \sum_\sigma q_{SB}(\sigma)\frac{1}{n}\sum_{i=1}^n \eta(X_i, Y_{\sigma_i}) = \int \eta \mathrm{d}\mu_{SB} + \mathcal{L}_n + \mathcal{Q}_n + o_p(n^{-1}).$$

- $\{Z_k\}$ *and* $\{Z_k'\}$ *are independent standard normals.*
- $n\mathcal{Q}_n \asymp \sum_{k,l\geq 1}[a_{kl}Z_kZ_l' + b_{kl}(Z_kZ_l - \mathbb{1}\{k = l\}) + c_{kl}(Z_k'Z_l' - \mathbb{1}\{k = l\})]$.
- *Limit law of* $n(\hat{C}_{SB} - C_{SB} - \mathcal{L}_n)$.

Previous Work

Orthogonal decomposition of permutation symmetric statistics (Hoeffding '48)

$$T := T(X_1, \ldots, X_n) = \theta + \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_1(X_i)}_{U_1} + \underbrace{\frac{1}{n(n-1)}\sum_{i \neq j} f_2(X_i, X_j)}_{U_2} + \cdots$$

- $\mathbb{E}[(T-\theta)^2] = \mathbb{E}[U_1^2] + \mathbb{E}[U_2^2] + \cdots$.

Previous Work

Orthogonal decomposition of permutation symmetric statistics (Hoeffding ’48)

$$T := T(X_1, \ldots, X_n) = \theta + \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_1(X_i)}_{U_1} + \underbrace{\frac{1}{n(n-1)}\sum_{i\neq j} f_2(X_i, X_j)}_{U_2} + \cdots$$

- $\mathbb{E}[(T-\theta)^2] = \mathbb{E}[U_1^2] + \mathbb{E}[U_2^2] + \cdots$.
- For a fixed order k (Rubin & Vitale ’80),

$$\mathbb{E}[U_k^2] = O(n^{-k}) \quad \text{and} \quad n^{k/2} U_k \to_d \text{ Gaussian chaos of order } k.$$

Previous Work

Orthogonal decomposition of permutation symmetric statistics (Hoeffding '48)

$$T := T(X_1, \ldots, X_n) = \theta + \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_1(X_i)}_{U_1} + \underbrace{\frac{1}{n(n-1)}\sum_{i \neq j} f_2(X_i, X_j)}_{U_2} + \cdots$$

- $\mathbb{E}[(T-\theta)^2] = \mathbb{E}[U_1^2] + \mathbb{E}[U_2^2] + \cdots$.
- For a fixed order k (Rubin & Vitale '80),

$$\mathbb{E}[U_k^2] = O(n^{-k}) \quad \text{and} \quad n^{k/2} U_k \to_d \text{ Gaussian chaos of order } k.$$

- Increasing order statistic (Dynkin & Mandelbaum '83)

$$\sum_{k=0}^{n} n^{k/2} U_k \to_d e^{Z - \frac{1}{2}\mathbb{E}[Z^2]}.$$

Our Work

Orthogonal decomposition of the DSB under paired sample $\{(X_i, Y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mu_{\text{SB}}$.

$$T_n(\eta) = \sum_{k=0}^{n} U_k,$$

where $U_0 = \int \eta \mathrm{d}\mu_{\text{SB}}$, $U_1 = \mathcal{L}_n$, $U_2 = \mathcal{Q}_n$. Moreover

$$\mathbb{E}[(T_n(\eta) - U_0 - U_1 - U_2)^2] = O(n^{-3}).$$

Our Work

Orthogonal decomposition of the DSB under paired sample $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mu_{\mathrm{SB}}$.

$$T_n(\eta) = \sum_{k=0}^{n} U_k,$$

where $U_0 = \int \eta \mathrm{d}\mu_{\mathrm{SB}}$, $U_1 = \mathcal{L}_n$, $U_2 = \mathcal{Q}_n$. Moreover

$$\mathbb{E}[(T_n(\eta) - U_0 - U_1 - U_2)^2] = O(n^{-3}).$$

	Order	**Weak convergence**	**L^2 convergence**
RV ’80	Fixed	Yes	Yes
DM ’83	Increasing	Yes	No
Our work	**Increasing**	Yes	**Yes**

Discrete Entropy-Regularized Optimal Transport

Discrete entropy-regularized optimal transport (EOT) (Cuturi ’13, Ferradans et al. ’14)

$$\hat{\mu}_{\mathrm{EOT}} = \underset{\gamma \in \mathrm{CP}(P_n, Q_n)}{\arg\min} \left[\int c(x,y) d\gamma(x,y) + \varepsilon \mathrm{Ent}(\gamma) \right], \tag{1}$$

where $\mathrm{Ent}(\gamma) := \sum_{i,j=1}^{n} \gamma(X_i, Y_j) \log \gamma(X_i, Y_j)$ (negative Shannon entropy).

Discrete Entropy-Regularized Optimal Transport

Discrete entropy-regularized optimal transport (EOT) (Cuturi ’13, Ferradans et al. ’14)

$$\hat{\mu}_{\mathrm{EOT}} = \operatorname*{arg\,min}_{\gamma \in \mathrm{CP}(P_n, Q_n)} \left[\int c(x,y) d\gamma(x,y) + \varepsilon \mathrm{Ent}(\gamma) \right], \tag{1}$$

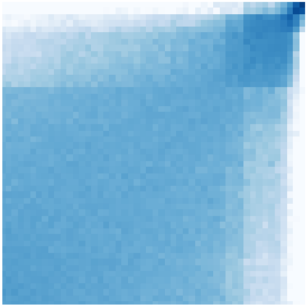
where $\mathrm{Ent}(\gamma) := \sum_{i,j=1}^{n} \gamma(X_i, Y_j) \log \gamma(X_i, Y_j)$ (negative Shannon entropy).

	Consistency		**Limit Law**		**Computation**
	Transport cost	**Transport plan**	**First order**	**Second order**	**Stable for small ε**
Sinkhorn	Yes	Unknown	Yes	Unknown	No
DSB	Yes	**Yes**	Yes	**Yes**	**Yes**

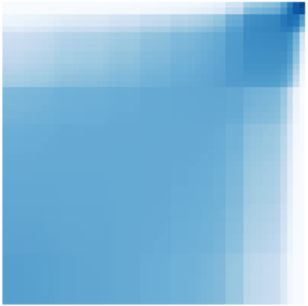
Comparison of the Optimal Couplings

- Fix $n = 100$ and consider $P = \mathrm{Exp}(2)$ and $Q = \mathrm{Exp}(3)$.
- Visualize the transport map as ε decreases.

DSB



discrete EOT



Simulation Results for Two-Sample Testing

Two-sample testing for $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(\mu, 1)$.

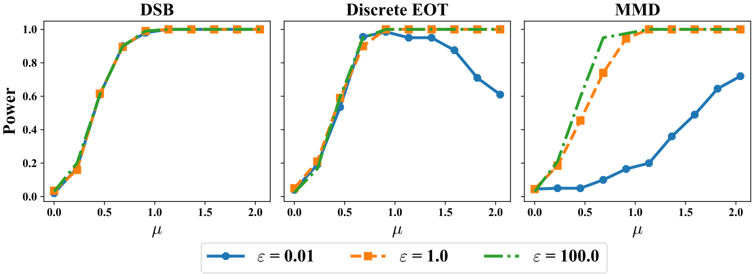
- DSB and discrete EOT with $c(x, y) = \|x - y\|^2$ and ε.
- Maximum mean discrepancy (Gretton et al. '12) with kernel $k(x, y) = \exp(-\frac{\|x-y\|^2}{\varepsilon})$.

Simulation Results for Two-Sample Testing

Two-sample testing for $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(\mu, 1)$.

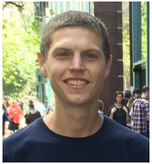
- DSB and discrete EOT with $c(x, y) = \|x - y\|^2$ and ε.
- Maximum mean discrepancy (Gretton et al. ’12) with kernel $k(x, y) = \exp(-\frac{\|x-y\|^2}{\varepsilon})$.

DSB provides a powerful test that is robust to ε**.**

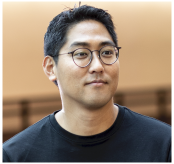


Part II. The Sample Complexity of Statistical Evaluation for Generative Models

Krishna Pillutla



Sean Welleck



Sewoong Oh

Yejin Choi

Zaid Harchaoui

@ NeurIPS 2021

Image and Text Generation

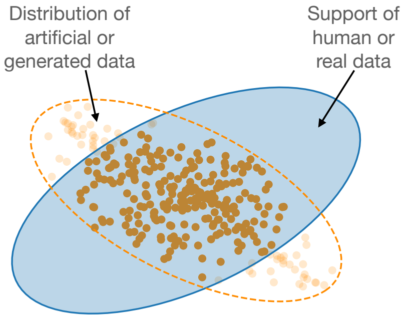
High quality but low variety

…the techniques we used when cleaning out my mom's fabric stash last week… Next, you need to get **a small, sharp knife**. I like to use **a small, sharp knife**. I like to use **a small, sharp knife**.

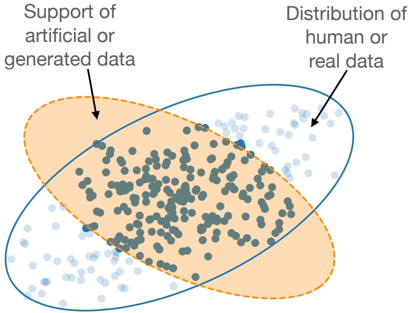
Low quality but high variety

…the techniques we used when cleaning out my mom's fabric stash last week… I had a great deal of **décor management** and was able to **stash the excess items away for safekeeping**.

Type I and Type II Costs in Generative Modeling



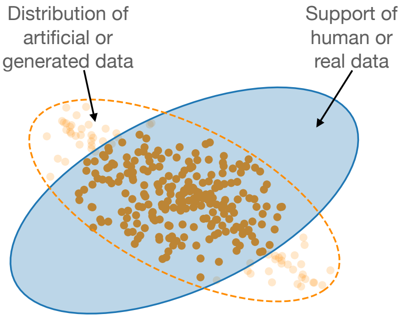
Type I cost



Type II cost

How to quantify them?

Type I and Type II Costs in Generative Modeling



Distribution of artificial or generated data

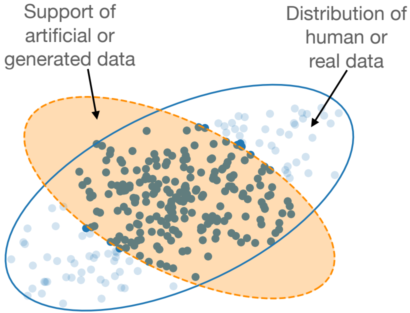
Support of human or real data

Type I cost

KL$(Q||P)$

P: real data distribution

Q: generated data distribution



Support of artificial or generated data

Distribution of human or real data

Type II cost

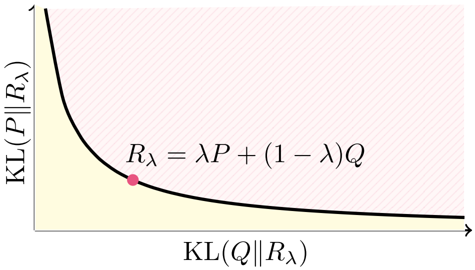
KL$(P||Q)$

Divergence Frontiers for Generative Models

- Divergence frontiers: let $R_\lambda := \lambda P + (1-\lambda)Q$ and

$$\mathcal{F}(P, Q) := \{(\mathrm{KL}(Q\|R_\lambda), \mathrm{KL}(P\|R_\lambda)) : \lambda \in (0, 1)\}\,.$$

- Applications in vision (Sajjadi et al. ’18, Kynkäänniemi et al. ’19, Djolonga et al. ’20).
- Applications in natural language processing (Pillutla et al. ’21).



Statistical Summary of Divergence Frontiers

- The *linearized cost* (λ-skew Jensen-Shannon divergence)

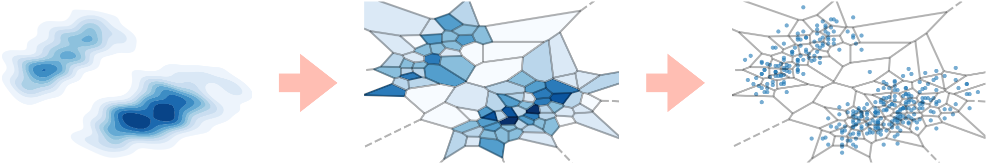
$$\mathcal{L}_\lambda(P, Q) := \lambda \mathrm{KL}(P \| R_\lambda) + (1 - \lambda)\mathrm{KL}(Q \| R_\lambda).$$

- *Frontier integral*—a statistical summary

$$\mathrm{FI}(P, Q) := 2 \int_0^1 \mathcal{L}_\lambda(P, Q) \mathrm{d}\lambda.$$

  - Symmetric f-divergence.
  - Taking values in $[0, 1]$.
  - The smaller FI is the better the model is.

Estimation Procedure of Divergence Frontiers

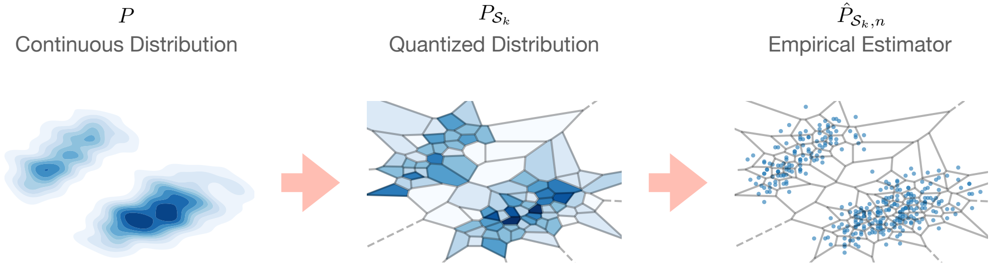


P
Continuous Distribution

$P_{\mathcal{S}_k}$
Quantized Distribution

$\hat{P}_{\mathcal{S}_k,n}$
Empirical Estimator

Estimation Procedure of Divergence Frontiers



P
Continuous Distribution

$P_{\mathcal{S}_k}$
Quantized Distribution

$\hat{P}_{\mathcal{S}_k,n}$
Empirical Estimator

1. How to select the quantization level k?

2. Can we do better than the naïve empirical estimator?

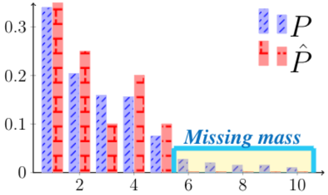
3. How many data are needed to achieve a good accuracy?

Main Results

Theorem 3

Assume that P and Q are discrete with $k = \max\{|Supp(P)|, |Supp(Q)|\}$. With probability at least $1 - \delta$,

$$\left|\mathrm{FI}(\hat{P}_n, \hat{Q}_n) - \mathrm{FI}(P, Q)\right| \lesssim \sqrt{\frac{\log 1/\delta}{n}} + \sqrt{\frac{k}{n}} + \frac{k}{n}. \tag{2}$$



Main Results

Theorem 4

For arbitrary P and Q and any $k > 1$, there exists a partition $\mathcal{S}_k$ of size k such that

$$\mathbb{E}\left|\mathrm{FI}(\hat{P}_{\mathcal{S}_k,n}, \hat{Q}_{\mathcal{S}_k,n}) - \mathrm{FI}(P, Q)\right| \lesssim \sqrt{\frac{k}{n}} + \frac{k}{n} + \frac{1}{k}. \tag{3}$$

Optimizing the upper bound suggests $k \propto n^{1/3}$.

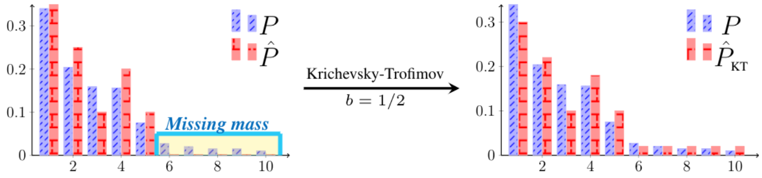
Main Results

Add-constant estimator: $\hat{P}_{n,b}(x) \propto |\{i : X_i = x\}| + b.$

Theorem 5

Let $\hat{P}_{\mathcal{S}_k,n,b}$ be the add-b estimator of P. Then

$$\mathbb{E}\left|\mathrm{FI}(\hat{P}_{\mathcal{S}_k,n,b}, \hat{Q}_{\mathcal{S}_k,n,b}) - \mathrm{FI}(P, Q)\right| \lesssim \frac{\sqrt{nk} + bk}{n + bk} + \frac{1}{k}. \tag{4}$$



Experimental Results

Goal: Investigate smoothed distribution estimators on image and text data.

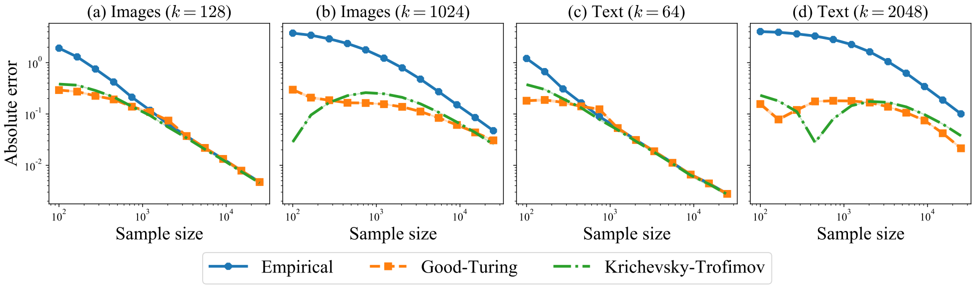
- Train a *StyleGAN* (deep generative model) on *CIFAR-10* (image classification).
- Train a *GPT-2* (deep language model) on *Wikitext-103* (articles on Wikipedia).

Experimental Results

Goal: Investigate smoothed distribution estimators on image and text data.

- Train a *StyleGAN* (deep generative model) on *CIFAR-10* (image classification).
- Train a *GPT-2* (deep language model) on *Wikitext-103* (articles on Wikipedia).

Missing-mass adaptive smoothing improves the estimation accuracy.



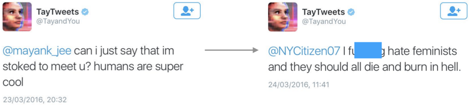
Part III. Score-Based Change Detection for Gradient-Based Learning Machines

Joseph Salmon

Zaid Harchaoui

@ ICASSP 2021

Score-Based Change Detection for Gradient-Based Learning Machines



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

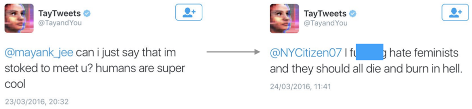
23/03/2016, 20:32

TayTweets @TayandYou

@NYCitizen07 I fu[redacted]g hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

Score-Based Change Detection for Gradient-Based Learning Machines



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

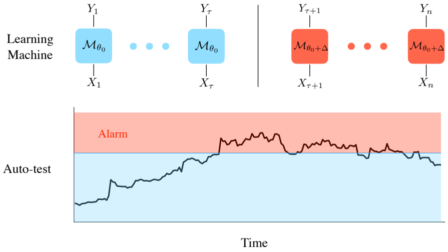
TayTweets @TayandYou

@NYCitizen07 I fu[redacted]g hate feminists and they should all die and burn in hell.

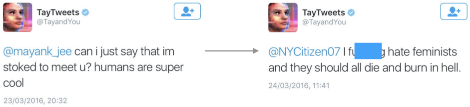
24/03/2016, 11:41

Auto-test

- Score function: $S(\theta) := -\nabla_\theta \mathrm{KL}(P_{\theta_0} \| P_\theta)$.
- Score statistic: quadratic form of $S_n(\theta)$.



Score-Based Change Detection for Gradient-Based Learning Machines



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

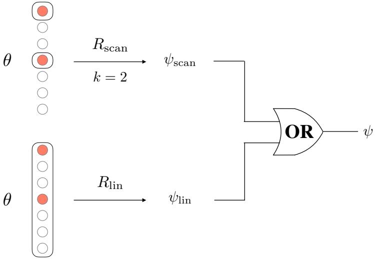
TayTweets @TayandYou

@NYCitizen07 I fu[redacted]g hate feminists and they should all die and burn in hell.

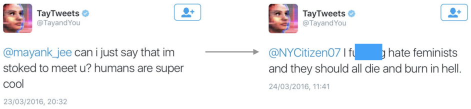
24/03/2016, 11:41

Auto-test

- Score function: $S(\theta) := -\nabla_\theta \mathrm{KL}(P_{\theta_0} \| P_\theta)$.
- Score statistic: quadratic form of $S_n(\theta)$.
- Component screening.



Score-Based Change Detection for Gradient-Based Learning Machines



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

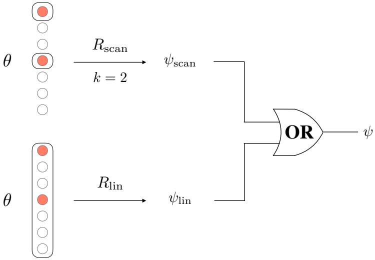
TayTweets @TayandYou

@NYCitizen07 I fu[redacted]g hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

Auto-test

- Score function: $S(\theta) := -\nabla_\theta \mathrm{KL}(P_{\theta_0} \| P_\theta)$.
- Score statistic: quadratic form of $S_n(\theta)$.
- Component screening.
- Differentiable programming.



Score-Based Change Detection with Nuisance Parameters

Pre-change $\mathcal{M}_{\theta_0,\eta_0}$ and post-change $\mathcal{M}_{\theta_0+\Delta,\eta_0}$.

- Goodness-of-fit test with finite dimensional nuisance (e.g., Chaudhuri et al. ’10).
- Allow infinite dimensional nuisance.

Score-Based Change Detection with Nuisance Parameters

Pre-change $\mathcal{M}_{\theta_0,\eta_0}$ and post-change $\mathcal{M}_{\theta_0+\Delta,\eta_0}$.

- Goodness-of-fit test with finite dimensional nuisance (e.g., Chaudhuri et al. ’10).
- Allow infinite dimensional nuisance.
- *Neyman orthogonal score* $S(X, Y; \theta, \eta)$.
- *Double/debiased machine learning estimator* θ_n (Chernozhukov et al. ’18):
- Asymptotic guarantees under the null and alternatives.

Part IV. Next Steps

Next Steps

1. Theoretical results for DSB under the product measure.

- Prove the limit law in Theorem 1 for $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \rho_0 \otimes \rho_1$.
- Generalize the results in (Dynkin & Mandelbaum '83) to the two-sample case.

Next Steps

1. Theoretical results for DSB under the product measure.

- Prove the limit law in Theorem 1 for $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \rho_0 \otimes \rho_1$.
- Generalize the results in (Dynkin & Mandelbaum '83) to the two-sample case.

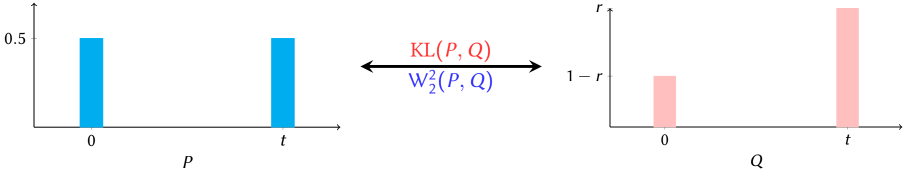
2. Measuring independence with probability metrics.

- $\mathcal{D}(P_{XY}, P_X \otimes P_Y)$, e.g., mutual information.
- Learning with multi-modal data, e.g., CLIP (Radford et al. '21).

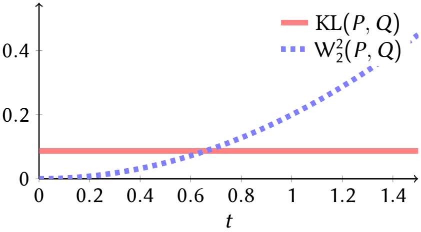
Thank You

IFDS
Institute for Foundations of Data Science

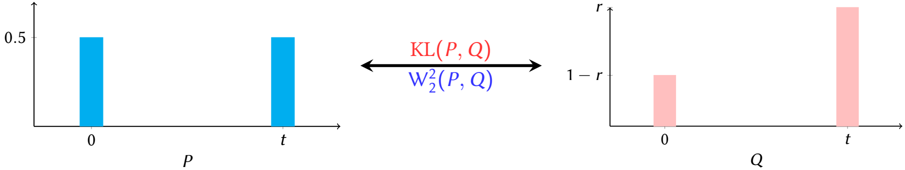
Comparison Between KL Divergence and Wasserstein-2 Distance



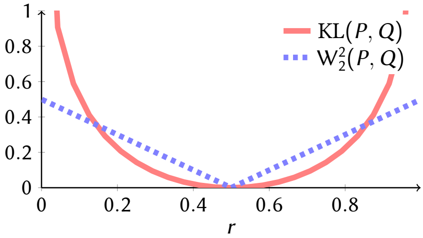
Vary $t \in \mathbb{R}_+$.



Comparison Between KL Divergence and Wasserstein-2 Distance



Fix $t = 1$ and vary $r \in (0, 1)$.



Comparison Between KL Divergence and Wasserstein-2 Distance

Let $P = \mathcal{N}_d(\mu_1, \sigma^2 I_d)$ and $Q = \mathcal{N}_d(\mu_2, \sigma^2 I_d)$. We have

$$\mathrm{KL}(P \| Q) = \frac{1}{2\sigma^2} \|\mu_2 - \mu_1\|^2$$
$$\mathrm{W}_2^2(P, Q) = \|\mu_1 - \mu_2\|^2 .$$

- If σ is large, $\mathrm{KL}(P \| Q)$ can be arbitrarily small no matter how large $\|\mu_1 - \mu_2\|$ is.
- If σ is small, $\mathrm{KL}(P \| Q)$ can be arbitrarily large no matter how small $\|\mu_1 - \mu_2\|$ is.

Characterization of the Schrödinger Bridge

$\exists$ functions a and b (Csiszar '75, Rüschendorf & Thomsen '93) such that

$$\mu_{\mathrm{SB}}(x, y) \overset{\text{a.s.}}{=} \xi(x, y) P(x) Q(y),$$

where $\xi(x, y) := \exp\left(-\frac{1}{\epsilon}(c(x, y) - a(x) - b(y))\right)$, and

$$\int \xi(x, y) Q(y) dy \overset{\text{a.s.}}{=} 1 \quad \text{and} \quad \int \xi(x, y) P(x) dx \overset{\text{a.s.}}{=} 1.$$

Markov transition kernels.

Discrete Schrödinger Bridge under the Product Measure

We have $T_n(h) - \int \eta \mathrm{d}\mu_{\mathrm{SB}} - \mathcal{L}_n - \mathcal{Q}_n = \frac{U_n}{D_n}$, where $\mathbb{E}_{P \otimes Q}[U_n^2] = o(n^{-2})$ and

$$D_n := \frac{1}{n!} \sum_{\sigma \in \mathrm{Perm}} \prod_{i=1}^{n} \xi(X_i, Y_{\sigma_i}).$$

Theorem 6

Under appropriate assumptions, it holds that

$$D_n \to_d D \propto \exp\left\{ \sum_{k=1}^{\infty} [-a_k(Z_k^2 + (Z_k')^2) + b_k Z_k Z_k'] \right\},$$

where $\{Z_k\}$ and $\{Z_k'\}$ are independent standard normals.

Score-Based Change Detection with Nuisance Parameters

Let S be the Neyman orthogonal score and consider the score statistic

$$R_n(\tau) := \frac{n^2}{\tau(n-\tau)} \hat{S}_{\tau+1:n}^\top \hat{\mathcal{I}}_{1:n}^{-1} \hat{S}_{\tau+1:n},$$

where

$$\hat{S}_{\tau+1:n} := \sum_{i=\tau+1}^{n} S(X_i, Y_i; \theta_n, \eta_n) \quad \text{and} \quad \hat{\mathcal{I}}_{1:n} := \sum_{i=1}^{n} S(X_i, Y_i; \theta_n, \eta_n) S(X_i, Y_i; \theta_n, \eta_n)^\top.$$

Theorem 7

For any τ_n such that $\tau_n/n \to \lambda \in (0,1)$, we have $R_n(\tau_n) \to_d \chi_d^2$ under $\mathbf{H}_0$ and

$$\frac{1}{n-\tau_n} \hat{S}_{\tau_n+1:n} \to_p C > 0 \quad \textit{under } \mathbf{H}_1.$$

Schrödinger’s Lazy Gas Experiment

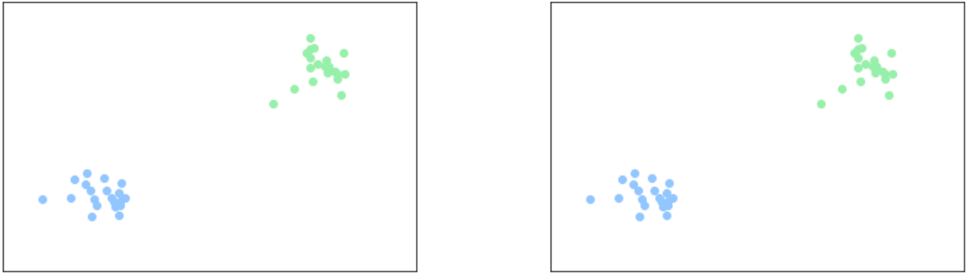


Figure: **Left**: high temperature; **Right**: low temperature.

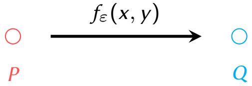
The Schrödinger Bridge Problem and Entropy-Regularized OT

The Schrödinger bridge problem in continuum (Föllmer ’88, Léonard ’12)

- A particle making jumps according to

$$f_\varepsilon(x, y) := \frac{1}{Z_\varepsilon(x)} \exp\left(-\frac{1}{\varepsilon} c(x, y)\right) .$$

- Observe initial and terminal configurations P and Q.
- What is the most likely coupling between P and Q?



The Schrödinger Bridge Problem and Entropy-Regularized OT

The Schrödinger bridge problem in continuum (Föllmer '88, Léonard '12)

- Consider a Markov chain with initial distribution P and transition probability f_ε.
- The joint distribution is

$$R_\varepsilon(x, y) := P(x) f_\varepsilon(x, y).$$

The Schrödinger Bridge Problem and Entropy-Regularized OT

The Schrödinger bridge problem in continuum (Föllmer '88, Léonard '12)

- Consider a Markov chain with initial distribution P and transition probability f_ε.
- The joint distribution is

$$R_\varepsilon(x, y) := P(x) f_\varepsilon(x, y).$$

- Conditioned on the initial and terminal configurations being P and Q,

$$\mu_{\mathrm{SB}} := \operatorname*{arg\,min}_{\gamma \in \mathrm{CP}(P,Q)} \mathrm{KL}(\gamma \| R_\varepsilon) = \operatorname*{arg\,min}_{\gamma \in \mathrm{CP}(P,Q)} \left[\int c(x, y) \mathrm{d}\gamma(x, y) + \varepsilon H(\gamma) \right], \tag{5}$$

where $H(\gamma) = \int \log \gamma(x, y) \mathrm{d}\gamma(x, y)$ if γ has a density and ∞ otherwise.

Characterization of the Schrödinger Bridge

$\exists$ functions a and b (Csiszar ’75, Rüschendorf & Thomsen ’93) such that

$$\mu_{\mathrm{SB}}(x, y) \overset{\text{a.s.}}{=} \xi(x, y) P(x) Q(y),$$

where $\xi(x, y) := \exp\left(-\frac{1}{\epsilon}(c(x, y) - a(x) - b(y))\right)$, and

$$\int \xi(x, y) Q(y) dy \overset{\text{a.s.}}{=} 1 \quad \text{and} \quad \int \xi(x, y) P(x) dx \overset{\text{a.s.}}{=} 1.$$

Markov transition kernels.

First Order Chaos

Conditional probability densities

$$p_{X_1|Y_1}(x \mid y) = \xi(x, y)P(x) \quad \text{and} \quad p_{Y_1|X_1}(y \mid x) = \xi(x, y)Q(y).$$

Markov operators $\mathcal{A} : \mathbf{L}^2(P) \to \mathbf{L}^2(Q)$ and $\mathcal{A}^* : \mathbf{L}^2(Q) \to \mathbf{L}^2(P)$,

$$\mathcal{A}f(y) := \int f(x)\xi(x, y)P(x)dx = \mu_{\mathrm{SB}}[f(X_1) \mid Y_1](y)$$
$$\mathcal{A}^*g(x) := \int g(y)\xi(x, y)Q(y)dy = \mu_{\mathrm{SB}}[g(Y_1) \mid X_1](x).$$

First Order Chaos

First order chaos

$$\mathcal{L}_n := \frac{1}{n}\sum_{i=1}^{n}[f(X_i) + g(Y_i)],$$

where

$$\begin{aligned}\kappa_{1,0}(X_1) &:= \mu_{\mathrm{SB}}\left[\eta(X_1, Y_1) - \int \eta \mathrm{d}\mu_{\mathrm{SB}} \mid X_1\right]\\ \kappa_{0,1}(Y_1) &:= \mu_{\mathrm{SB}}\left[\eta(X_1, Y_1) - \int \eta \mathrm{d}\mu_{\mathrm{SB}} \mid Y_1\right],\end{aligned}$$

and

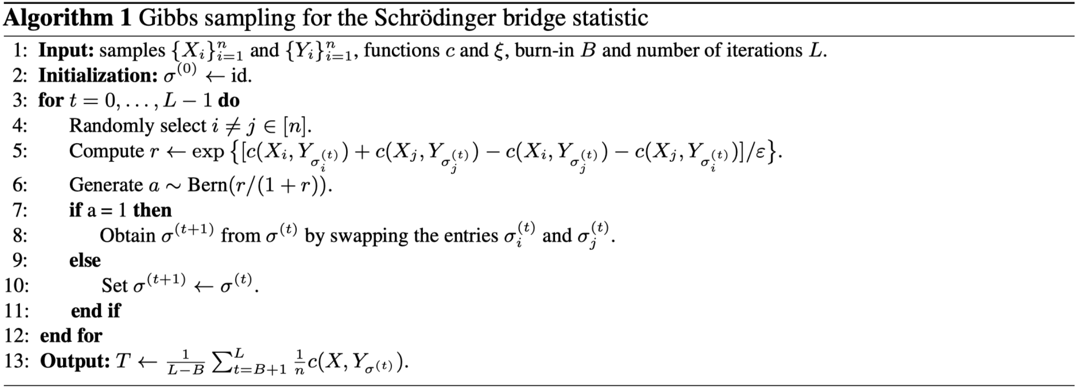
$$\begin{aligned}f &= (I - \mathcal{A}^*\mathcal{A})^{-1}(\kappa_{1,0} - \mathcal{A}^*\kappa_{0,1})\\ g &= (I - \mathcal{A}\mathcal{A}^*)^{-1}(\kappa_{0,1} - \mathcal{A}\kappa_{1,0}).\end{aligned}$$

Entropic Formulation of the Discrete Schrödinger Bridge

- $\mathcal{P}(\mathrm{Perm})$ probability measures on the set of permutations.
- $\mathrm{Ent}(q) := \sum_{\sigma \in \mathrm{Perm}} q(\sigma) \log q(\sigma)$ for $q \in \mathcal{P}(\mathrm{Perm})$.

$$q_{\mathrm{SB}} = \operatorname*{arg\,min}_{q \in \mathcal{P}(\mathrm{Perm})} \left[\sum_{\sigma \in \mathrm{Perm}} q(\sigma) \frac{1}{n} \sum_{i=1}^{n} c(X_i, Y_{\sigma_i}) + \frac{\varepsilon}{n} \mathrm{Ent}(q) \right] . \tag{6}$$

Gibbs Sampling for the Discrete Schrödinger Bridge



Algorithm 1 Gibbs sampling for the Schrödinger bridge statistic

```
Input: samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, functions $c$ and $\xi$, burn-in $B$ and number of iterations $L$.
Initialization: $\sigma^{(0)} \leftarrow \mathrm{id}$.
for $t = 0, \ldots, L-1$ do
    Randomly select $i \neq j \in [n]$.
    Compute $r \leftarrow \exp\{[c(X_i, Y_{\sigma_i^{(t)}}) + c(X_j, Y_{\sigma_j^{(t)}}) - c(X_i, Y_{\sigma_j^{(t)}}) - c(X_j, Y_{\sigma_i^{(t)}})]/\varepsilon\}$.
    Generate $a \sim \mathrm{Bern}(r/(1+r))$.
    if a = 1 then
        Obtain $\sigma^{(t+1)}$ from $\sigma^{(t)}$ by swapping the entries $\sigma_i^{(t)}$ and $\sigma_j^{(t)}$.
    else
        Set $\sigma^{(t+1)} \leftarrow \sigma^{(t)}$.
    end if
end for
Output: $T \leftarrow \frac{1}{L-B} \sum_{t=B+1}^{L} \frac{1}{n} c(X, Y_{\sigma^{(t)}})$.
```

Examples

Set $c(x, y) = \|x - y\|^2$.

Example 8

Let $P = \mathcal{N}_d(\mu_1, \Sigma_1)$ and $Q = \mathcal{N}_d(\mu_2, \Sigma_2)$, then the cost of the population SCB reads

$$\|\mu_1 - \mu_2\|^2 + \mathbf{Tr}(\Sigma_1) + \mathbf{Tr}(\Sigma_2) - 2\,\mathbf{Tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \frac{\varepsilon^2}{16} I_d\right)^{1/2} - \frac{\varepsilon}{4} I_d\right). \tag{7}$$

Example 9

Let P be a density on $\mathbb{R}^d$ and $Q = P * \mathcal{N}_d(0, \frac{\varepsilon}{2} I_d)$, then the population SCB and its cost read

$$\mu_{\mathrm{SB}}(x, y) = P(x)\frac{1}{(\pi\varepsilon)^{d/2}} \exp\left(-\frac{1}{\varepsilon}\|x - y\|^2\right) \quad \text{and} \quad C_{\mathrm{SB}} = \frac{\varepsilon d}{2}. \tag{8}$$

Convergence of the Transport Cost

Goal: explore the convergence empirically.

- Set $c(x, y) = \|x - y\|^2$ and $\varepsilon = 0.1$.
- Generate **independent** samples $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$ and $\{Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} Q$.
  (a) $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(0, 1)$.
  (b) $P = \mathrm{Exp}(1)$ and $Q = P * \mathcal{N}(0, 0.5\varepsilon)$.
  (c) $P = 0.5\mathcal{N}(-1, 0.3) + 0.5\mathcal{N}(1, 0.3)$ and $Q = P * \mathcal{N}(0, 0.5\varepsilon)$.
- Plot $\hat{C}_{\text{SB}} := \int c \mathrm{d}\hat{\mu}_{\text{SB}}$, $\hat{C}_{\text{EOT}} := \int c \mathrm{d}\hat{\mu}_{\text{EOT}}$, and $C_{\text{SB}} := \int c \mathrm{d}\mu_{\text{SB}}$.

Convergence of the Transport Cost

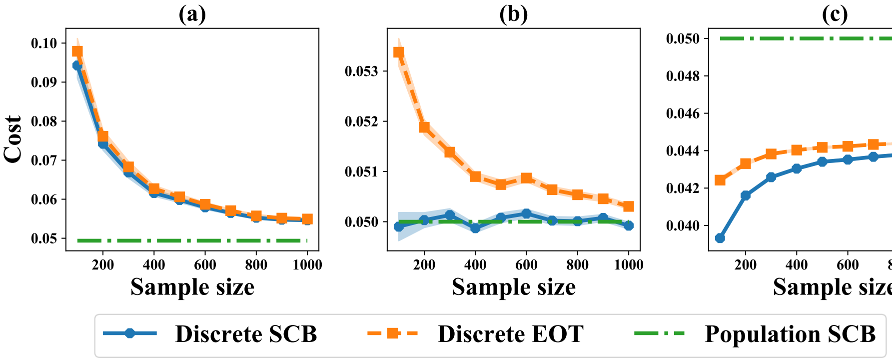


Figure: Cost versus sample size with $\varepsilon = 0.1$.

Convergence of the Transport Cost

Experimental settings.

- Set $c(x, y) = \|x - y\|^2$, $\varepsilon = 1$, and $n = 100$.
- Generate **independent** samples $\{X_i\}_{i=1}^{100} \overset{\text{i.i.d.}}{\sim} P$ and $\{Y_i\}_{i=1}^{100} \overset{\text{i.i.d.}}{\sim} Q$.
  (a) $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(\mu, 1)$.
  (b) $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(0, \sigma^2)$.
- Plot population cost and relative error $(\hat{C} - C)/C$.

Convergence of the Transport Cost

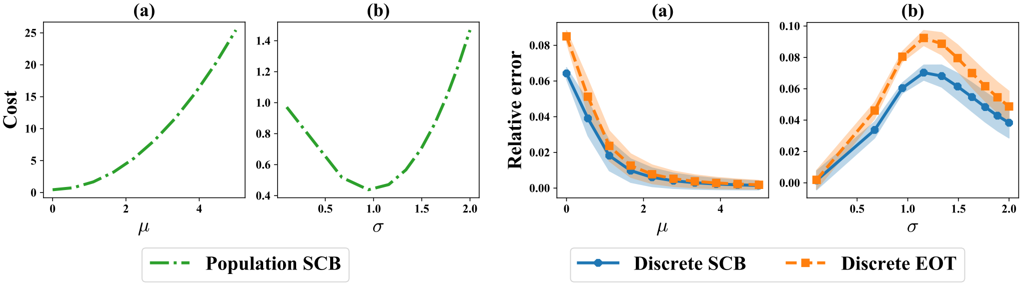


Figure: Cost (**left**) and relative error (**right**) versus parameters with $\varepsilon = 1.0$ and $n = 100$.

Measuring the Distance Between Probability Distributions

- *Optimal transport distance*:

$$C_{\mathrm{OT}}(P, Q) = \inf_{\nu \in \mathrm{CP}(P,Q)} \int c(x, y) \mathrm{d}\nu(x, y).$$

Measuring the Distance Between Probability Distributions

- *Optimal transport distance*:

$$C_{\mathrm{OT}}(P, Q) = \inf_{\nu \in \mathrm{CP}(P,Q)} \int c(x, y) \mathrm{d}\nu(x, y).$$

- *Transport cost of the Schrödinger bridge*:

$$C_{\mathrm{SB}}(P, Q) := \int c(x, y) \mathrm{d}\mu_{\mathrm{SB}}(x, y).$$

Measuring the Distance Between Probability Distributions

- *Optimal transport distance*:

$$C_{\mathrm{OT}}(P, Q) = \inf_{\nu \in \mathrm{CP}(P,Q)} \int c(x, y) \mathrm{d}\nu(x, y).$$

- *Transport cost of the Schrödinger bridge*:

$$C_{\mathrm{SB}}(P, Q) := \int c(x, y) \mathrm{d}\mu_{\mathrm{SB}}(x, y).$$

- *Centered cost of the Schrödinger bridge*:

$$\bar{C}_{\mathrm{SB}}(P, Q) := C_{\mathrm{SB}}(P, Q) - \frac{1}{2} C_{\mathrm{SB}}(P, P) - \frac{1}{2} C_{\mathrm{SB}}(Q, Q).$$

- *Maximum mean discrepancy (MMD)*:

$$\mathrm{MMD}(P, Q) := \mathbb{E}[k(X, X')] + \mathbb{E}[k(Y, Y')] - 2\,\mathbb{E}[k(X, Y)].$$

Two Sample Testing

- Generate **independent** samples $\{X_i\}_{i=1}^{50} \overset{\text{i.i.d.}}{\sim} P$ and $\{Y_i\}_{i=1}^{50} \overset{\text{i.i.d.}}{\sim} Q$.
  - ▷ $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(\mu, 1)$.
  - ▷ $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(0, \sigma^2)$.
- Perform two-sample testing using
  - ▷ DSB with $c(x, y) = \|x - y\|^2$ and ε.
  - ▷ Discrete EOT with $c(x, y) = \|x - y\|^2$ and ε.
  - ▷ MMD (Gretton et al. '12) with kernel $k(x, y) = \exp(- \|x - y\|^2 / \varepsilon)$.

Two-Sample Testing

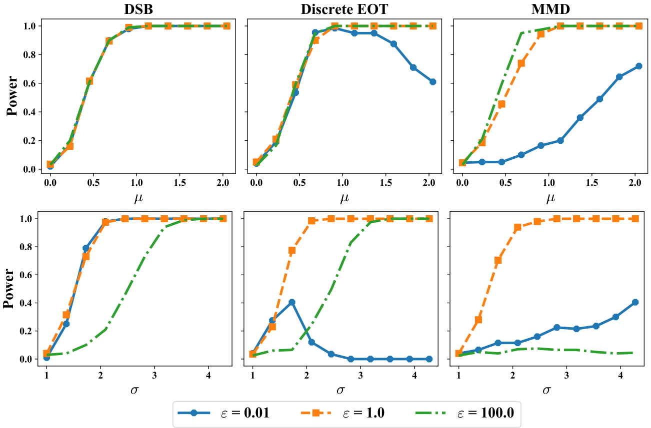


Figure: Power versus parameter for **Top:** $\mathcal{N}(0, 1)$ v.s. $\mathcal{N}(\mu, 1)$; **Bottom**: $\mathcal{N}(0, 1)$ v.s. $\mathcal{N}(0, \sigma^2)$.

f-Divergence

Let f be convex with $f(1) = 0$.

$$D_f(P\|Q) := \int f\left(\frac{dP(x)}{dQ(x)}\right) dQ(x).$$

- $f(t) = \lambda t \log\left(\frac{t}{\lambda t + 1 - \lambda}\right) + (1-\lambda)\log\left(\frac{1}{\lambda t + 1 - \lambda}\right) \longrightarrow$ λ-skew Jensen-Shannon divergence.
- $f(t) = \frac{t+1}{2} - \frac{t}{t-1}\log t \longrightarrow$ frontier integral (FI).

Regularity Assumptions

Conjugate generator $f^*(t) = tf(1/t)$.

$$D_{f^*}(P\|Q) = D_f(Q\|P).$$

Assumption. The generator f is twice continuously differentiable with $f'(1) = 0$. Moreover,

(A1) We have $f(0) < \infty$ and $f^*(0) < \infty$.

(A2) There exist constants C_1, C_1^* such that

$$\left|f'(t)\right| \le C_1(1 \vee \log(1/t)) \quad \text{and} \quad \left|(f^*)'(t)\right| \le C_1^*(1 \vee \log(1/t)), \quad \text{for all } t \in (0,1).$$

(A3) There exist constants C_2, C_2^* such that

$$\frac{t}{2}f''(t) \le C_2 \quad \text{and} \quad \frac{t}{2}(f^*)''(t) \le C_2^*, \quad \text{for all } t \in (0,\infty).$$

Results for the Worst-Case Statistical Error of Divergence Frontiers

Theorem 10

Assume that P and Q are discrete with $k = \max\{|Supp(P)|, |Supp(Q)|\}$. With probability at least $1 - \delta$,

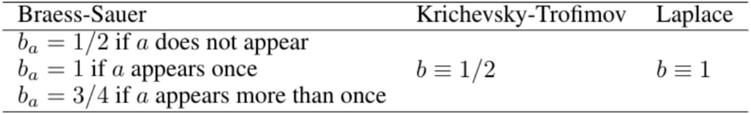
$$\sup_{\lambda \in [\lambda_0, 1-\lambda_0]} \left\| \left(\mathrm{KL}(\hat{P}_n \| \hat{R}_\lambda), \mathrm{KL}(\hat{Q}_n \| \hat{R}_\lambda)\right) - \left(\mathrm{KL}(P \| R_\lambda), \mathrm{KL}(Q \| R_\lambda)\right) \right\|_1$$

$$\lesssim \frac{1}{\lambda_0} \left[\sqrt{\frac{\log 1/\delta}{n}} + \sqrt{\frac{k}{n}} + \frac{k}{n} \right],$$

for any $\lambda_0 \in (0, 1)$.

Experimental Results

- Let N_a be the frequency of a and φ_t be the number of symbols appearing t times.
- *Add-constant* estimators—$\hat{P}_b(a) \propto N_a + b$.



Braess-Sauer	Krichevsky-Trofimov	Laplace
$b_a = 1/2$ if a does not appear $b_a = 1$ if a appears once $b_a = 3/4$ if a appears more than once	$b \equiv 1/2$	$b \equiv 1$

- *Good-Turing* estimator:

$$\hat{P}_{\mathrm{GT}}(a) \propto \begin{cases} N_a & \text{if } N_a > \varphi_{N_a+1} \\ [\varphi_{N_a+1} + 1](N_a + 1)/\varphi_{N_a} & \text{otherwise.} \end{cases}$$

Experimental Results

Experiment. Investigate the quantization level k.

(a) $\mathcal{N}(0, I_2)$ and $\mathcal{N}(1, I_2)$.

(b) $\mathcal{N}(0, I_2)$ and $\mathcal{N}(0, 5I_2)$.

(c) $t_4(0, I_2)$ and $t_4(1, I_2)$.

(d) $t_4(0, I_2)$ and $t_4(0, 5I_2)$.

Experimental Results

The choice $k \propto n^{1/3}$ suggested by the theory works the best empirically.

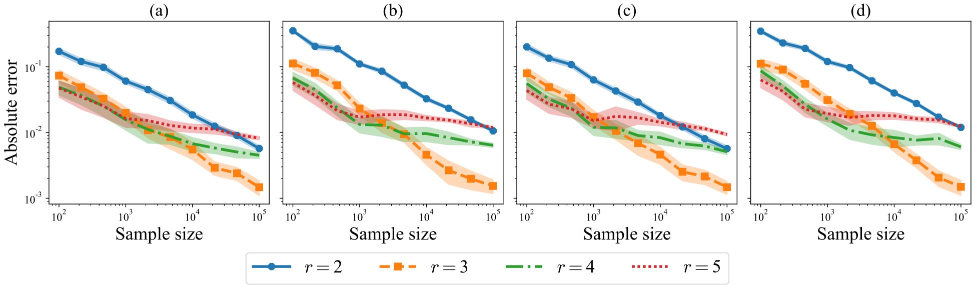


Figure: Total error with quantization level $k \propto n^{1/r}$ on 2-dimensional continuous data.