Towards a Geometric Theory of Information

Jose Gallego

Predoc III

Outline

- Information Theory, Machine Learning and a geometric gap
- Recap on IT and convexity
- GAIT: A Geometric Approach to Information Theory
- Applications
- Research Agenda

Information Theory, Machine Learning and a geometric gap

“Two sides of the same coin”

Induction of Decision Trees

J.R. QUINLAN
Centre for Advanced Computing Sciences, New South Wales Institute Australia

(Received August 1, 1985)

The information bottleneck method

Naftali Tishby,[1,2] Fernando C. Pereira,[3] and William Bialek[1]

30 September 1999

Natural Gradient Works Efficiently in Learning

Shun-ichi Amari
RIKEN Frontier Research Program, Saitama 351-01, Japan

Neural Computation **10, 251–276** (1998)

Reinforcement Learning with Deep Energy-Based Policies

Tuomas Haarnoja [*1] **Haoran Tang** [*2] **Pieter Abbeel** [1 3 4] **Sergey Levine** [1]

Fixing a Broken ELBO

Alexander A. Alemi [1] **Ben Poole** [2 *] **Ian Fischer** [1] **Joshua V. Dillon** [1] **Rif A. Saurous** [1] **Kevin Murphy** [1]

Mutual Information Neural Estimation

Mohamed Ishmael Belghazi [1] **Aristide Baratin** [1 2] **Sai Rajeswar** [1] **Sherjil Ozair** [1] **Yoshua Bengio** [1 3 4]
Aaron Courville [1 3] **R Devon Hjelm** [1 4]

Where is the Information in a Deep Neural Network?

Alessandro Achille [1 2] **Giovanni Paolini** [1 3] **Stefano Soatto** [1 2]

Published as a conference paper at ICLR 2018

ON THE INFORMATION BOTTLENECK THEORY OF DEEP LEARNING

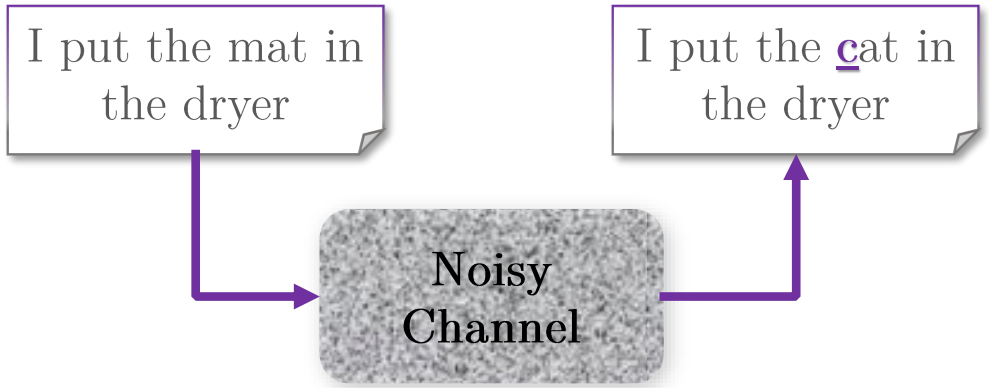
Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani
Harvard University

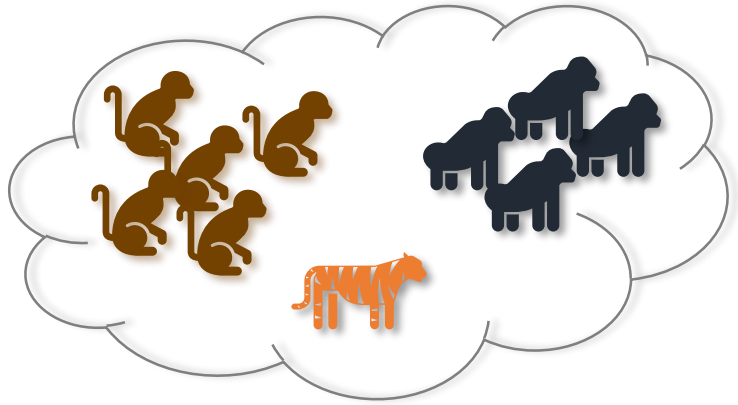
Artemy Kolchinsky, Brendan D. Tracey
Santa Fe Institute

David D. Cox
Harvard University
MIT-IBM Watson AI Lab

(MacKay, 2003)

Symbols carry structure





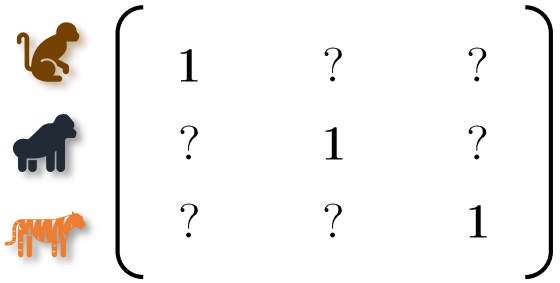
$$\mathcal{X} = \{a, b, c, \dots\}$$

$$\mathcal{X} = \{ \text{🐒}, \text{🦍}, \text{🐅} \}$$

$$\begin{array}{c} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{array} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\begin{array}{c} \text{🐒} \\ \text{🦍} \\ \text{🐅} \end{array} \begin{pmatrix} 1 & 0.7 & 0.1 \\ 0.7 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix}$$

Geometry in ML?



- Structured action spaces in reinforcement learning
- Utility-sensitive prediction
- Reconstruction losses for generative models/autoencoders
- Statistical distances between probabilistic models
- ...

Research question

How can we extend information-theoretic notions to account for the geometric structure of the space?

Recap on IT and convexity

Setting

- $\mathcal{X} = \{x_1, \dots, x_{|\mathcal{X}|}\}$ is a discrete alphabet of symbols
- X is a random variable over $\mathcal{X}$ with probability distribution $\mathbb{P}$
- $p(x) = \mathbb{P}(X = x)$ is its PMF

Entropy

Theorem (Shannon, 1948)

There is only one* function $\mathbb{H}[X]$ satisfying:

(E1) $\mathbb{H}$ is a continuous function of $p(x)$.

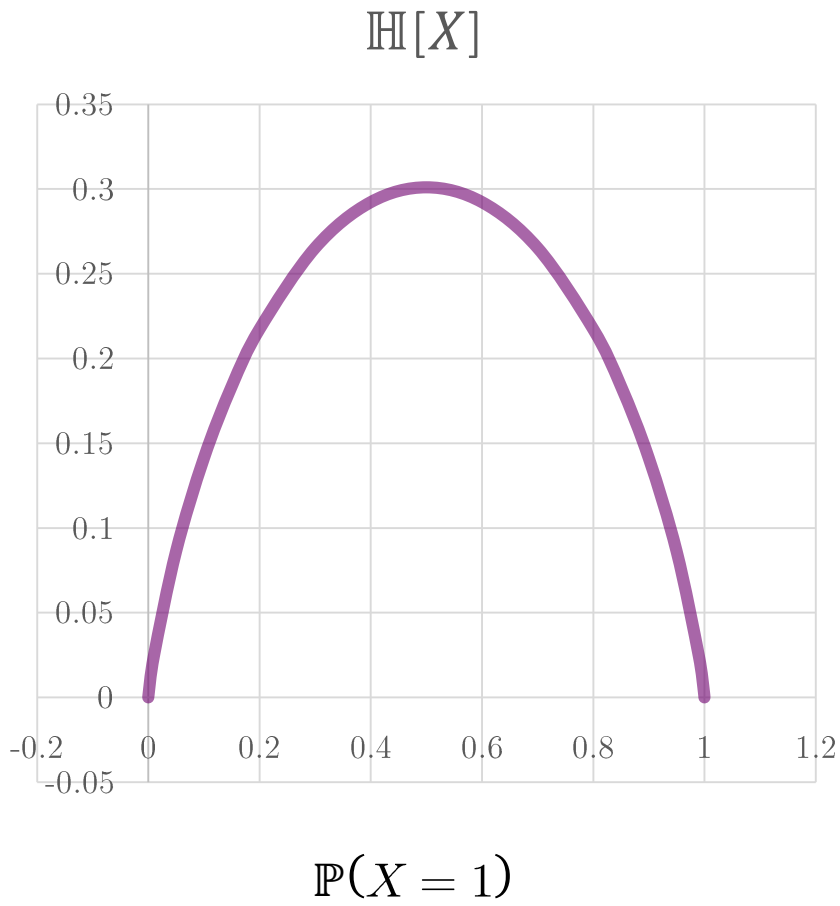
(E2) If X is uniformly distributed over an alphabet $\mathcal{X}$, then $\mathbb{H}$ should be a monotonic increasing function of $|\mathcal{X}|$.

(E3) If a choice is broken down into two successive choices, the original entropy should be the weighted sum of the individual entropies.

The entropy of X is defined as $\mathbb{H}[X] = -\sum_{x \in \mathcal{X}} p(x) \log(p(x))$.

Entropy - Properties

- Invariant with respect to relabelling (!)
  - "Intrinsic" property of the random variable
  - Oblivious of geometric structure
- (Strictly) Concave function of the PMF
- Maximum attained at uniform distribution

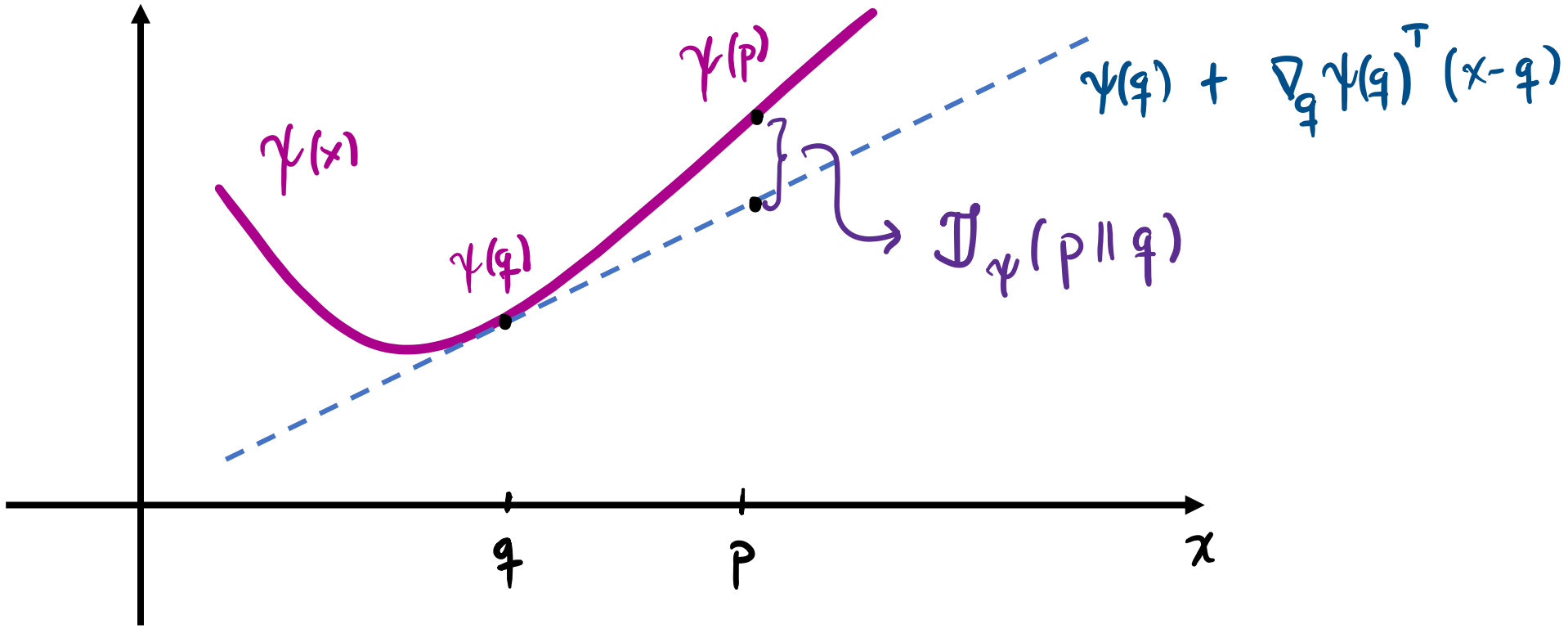


Entropy - Associated concepts

Entropy definition works *nicely* with jointly distributed random variables and "operations"

- Joint entropy of several random variables $\mathbb{H}[X, Y, ...]$
- Conditional entropy of Y given X: $\mathbb{H}[Y \mid X] = -\sum_{x \in \mathcal{X}} p(x)\mathbb{H}[Y \mid X = x]$
- **Chain rule** $\mathbb{H}[X, Y] = \mathbb{H}[X] + \mathbb{H}[Y|X] = \mathbb{H}[Y] + \mathbb{H}[X|Y]$

ConVexity and Bregman Divergence



Kullback-Leibler Divergence

Definition (Kullback-Leibler Divergence)

The Kullback-Leibler divergence is the Bregman divergence induced by the negative Shannon entropy.

$$\mathbb{KL}[\mathbb{P} \,||\, \mathbb{Q}] = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$



GAIT: A Geometric Approach to Information Theory

with Ankit **Vani**, Max **Schwarzer** and Simon **Lacoste-Julien**

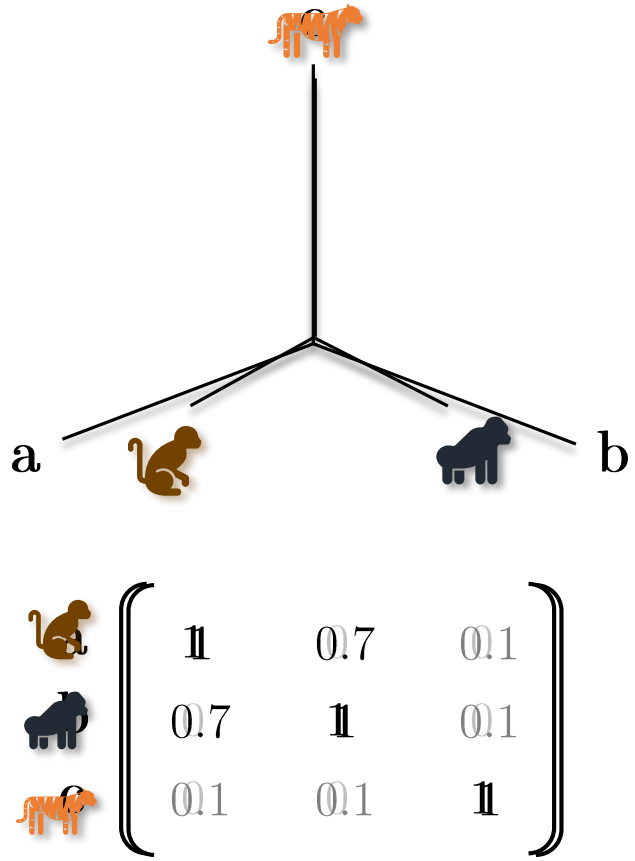
Contributions

How can we extend information-theoretic notions to account for the geometric structure of the space?

- Import notion of entropy on a similarity space, originally from ecology, to ML
- Extend IT concepts to incorporate geometry
- Define new divergence – comparable to Wasserstein but with closed-form expression
- Show usefulness in common ML tasks

Similarity spaces

- $(\mathcal{X}, \kappa)$ is a similarity space
- $\kappa(x, y)$ denotes the similarity between x and y
  - $0 \leq \kappa(x, y) \leq 1$
  - $\kappa(x, x) = 1$
- $\mathbb{P}$ is a distribution over $\mathcal{X}$

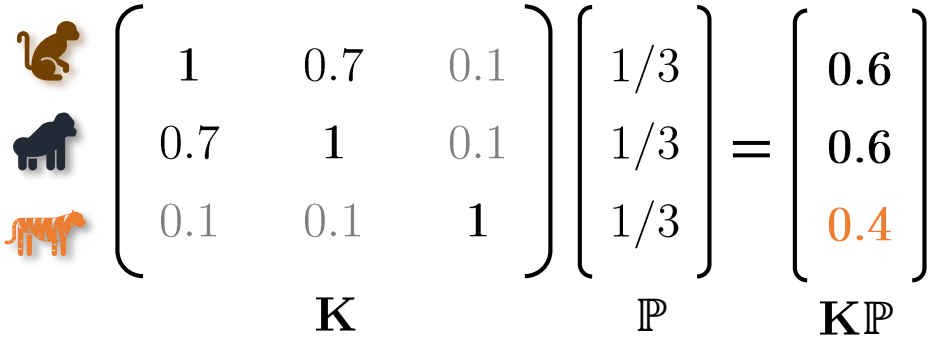


Similarity profile

Definition

$$\mathbf{K}\mathbb{P}(x) \triangleq \mathbb{E}_{y\sim\mathbb{P}}\, \kappa(x, y)$$

- Measures the **ordinariness** of x
- $1/\mathbf{K}\mathbb{P}(x)$ quantifies the **rarity** of x
- C.f. kernel mean embedding of distributions



Similarity sensitive entropy

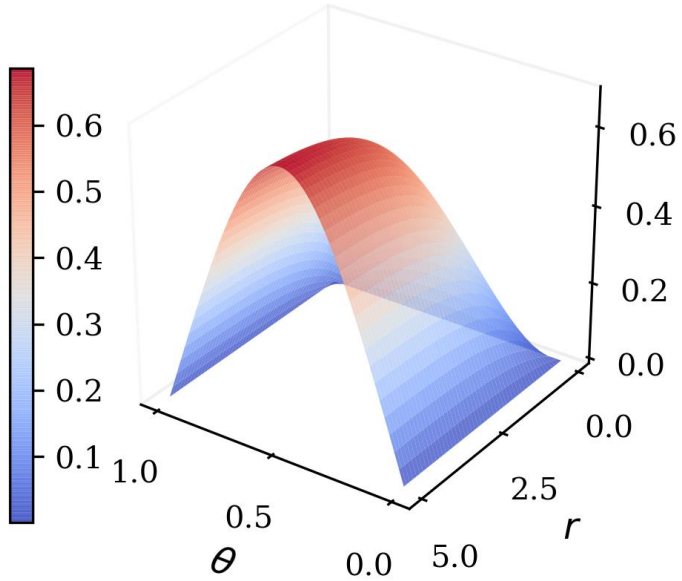
Definition (Leinster and Cobbold, 2012)

$$\mathbb{H}_1^{\mathbf{K}}[\mathbb{P}] \triangleq \mathbb{E}_{x\sim\mathbb{P}}\left[\log \frac{1}{\mathbf{K}\mathbb{P}(x)}\right]$$

- Actually defined and studied an indexed family of entropies in the sense of Rényi
- We concentrate in "Shannon-equivalent" case for convenience
- We extend conditional entropy & mutual information

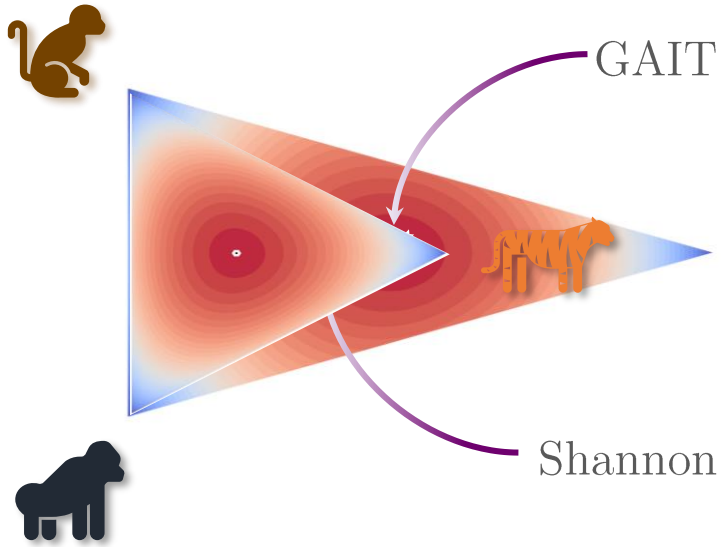
Properties

- **Absence:** The entropy remains **unchanged when we restrict** the similarity space **to the support** of the random variable.
- **Modularity:** If the space is partitioned into fully dissimilar groups, then the entropy of a distribution is a **weighted average of the block-wise entropies**.
- **Identical elements:** If two elements are identical, then **combining them into one** and adding their probabilities **leaves the entropy unchanged**.
- **K-Monotonicity:** $0 = \mathbb{H}_1^{\mathbf{J}}[\mathbb{P}] \leq \mathbb{H}_1^{\mathbf{K}}[\mathbb{P}] \leq \mathbb{H}_1^{\mathbf{I}}[\mathbb{P}] = \mathbb{H}[\mathbb{P}] \leq \log(|\mathcal{X}|)$



(Leinster and Cobbold, 2012)

Maximum entropy



$$\text{🎯} = \operatorname*{argmax}_{\mathbb{P}} \mathbb{H}_1^{\mathbf{K}}[\mathbb{P}] = \begin{bmatrix} 0.27 \\ 0.27 \\ \mathbf{0.46} \end{bmatrix} \quad \mathbf{K}\,\text{🎯} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

Similarity-sensitive divergence

Definition (GAIT Divergence)

$$\mathfrak{D}^{\mathbf{K}}\left[\mathbb{P}||\mathbb{Q}\right] \triangleq 1 + \mathbb{E}_{x\sim\mathbb{P}}\left[\log \frac{\mathbf{K}\mathbb{P}(x)}{\mathbf{K}\mathbb{Q}(x)}\right] - \mathbb{E}_{x\sim\mathbb{Q}}\left[\frac{\mathbf{K}\mathbb{P}(x)}{\mathbf{K}\mathbb{Q}(x)}\right]$$

- Bregman divergence induced by $-\mathbb{H}_1^{\mathbf{K}}$
- $\mathfrak{D}^{\mathbf{I}} = \mathbb{KL}$
- Local-global comparisons (c.f. f-divergence)

Concavity of $\mathbb{H}_1^{\mathbf{K}}$

Conjecture

If κ is a (strictly) positive definite kernel such that $\kappa(x, z) \geq \kappa(x, y)\, \kappa(y, z)$.

Then $\mathbb{H}_1^{\mathbf{K}}$ is a (strictly) concave function.

- Extensive empirical verification of this conjecture
- Proof for $|\mathcal{X}| = 2$
- For $|\mathcal{X}| > 2$, triangle inequality is apparently a sufficient extra condition (not necessary)

Advantages of $\mathfrak{D}^{\mathbf{K}}$

I. Similarity-sensitive

- Prediction with related classes
- Qualitative evaluations of generative models

II. Possibly disjoint supports

- Empirical measures
- Like Wasserstein

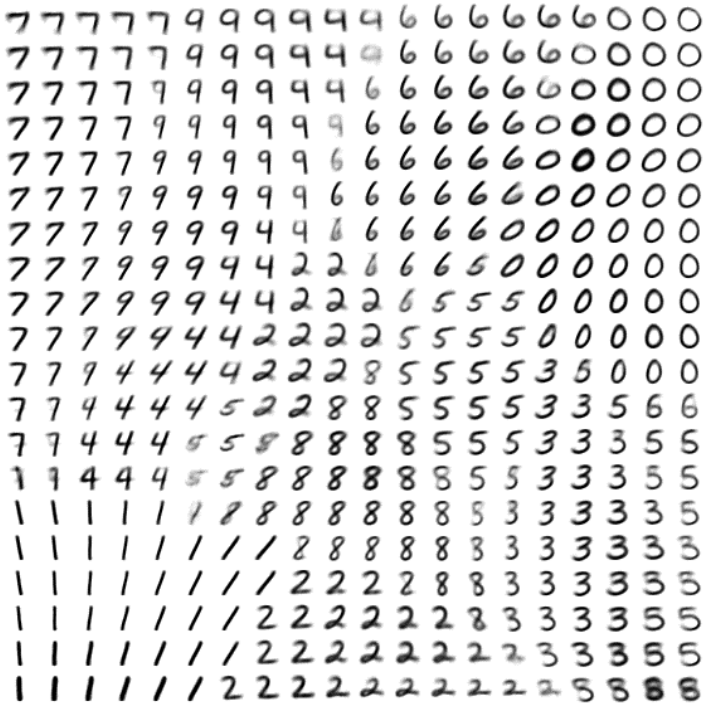
III. Closed-form expression

- Unlike Wasserstein
- No need for LP, do MC estimation

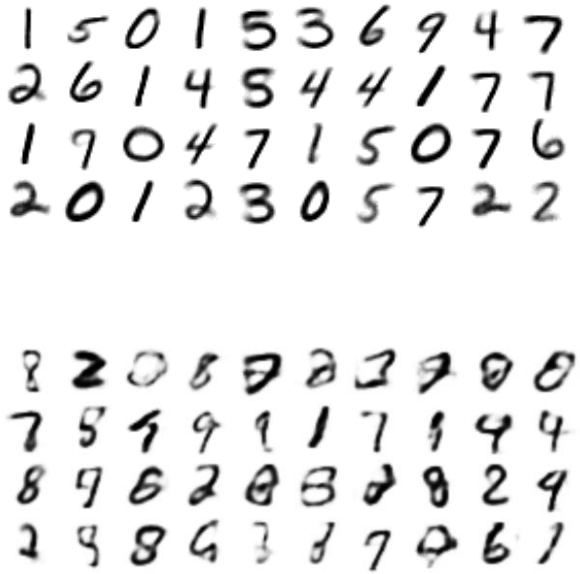
Applications

Learning generative models

$$\underset{\theta}{\operatorname{argmin}}\, \mathfrak{D}^{\mathbf{K}}[\, \mathbb{P}_{real} \,||\, g_\theta \# \mathbb{P}_z]$$



Learned 2D manifold

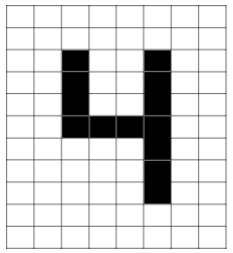


Generated samples with 20D latent space

(Kingma and Welling, 2014)

Computing image barycenters

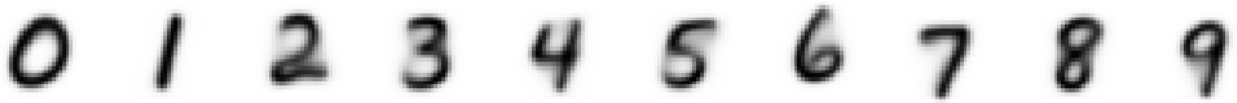
$$\underset{\mathbb{Q}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n} \mathfrak{D}^{\mathbf{K}}[\mathbb{P}_i||\mathbb{Q}]$$



GAIT

5 s/class

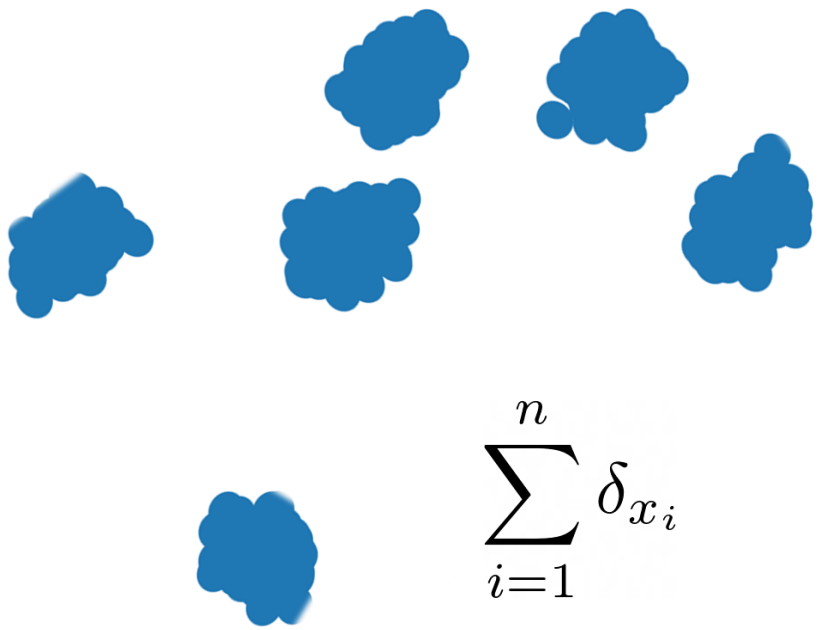
Sinkhorn
(convolutional)



90 s/class

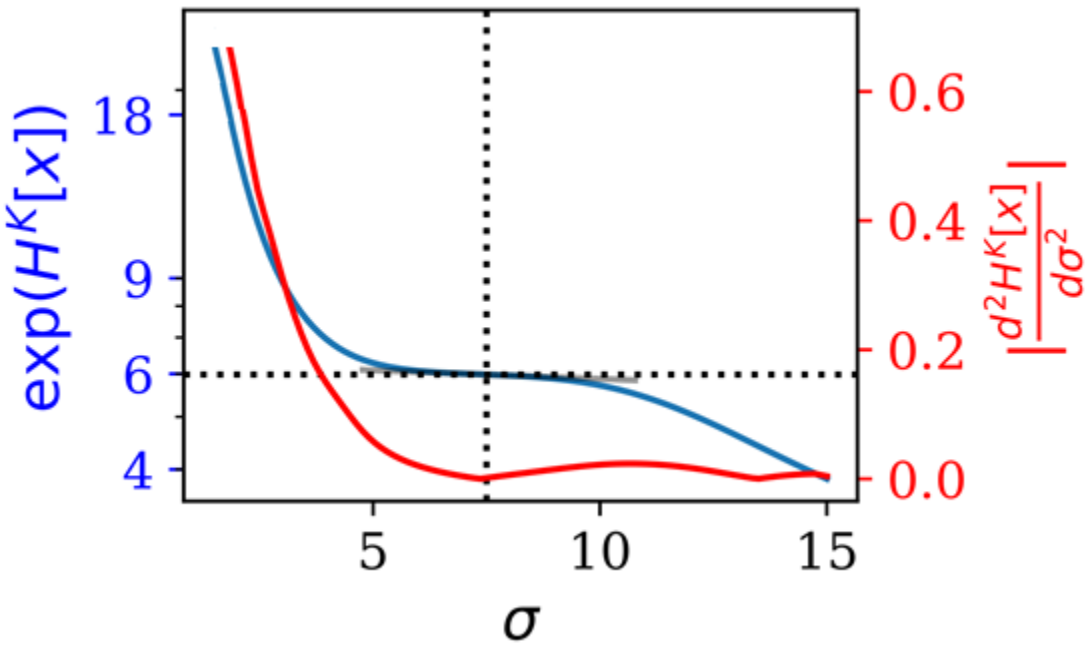
(Cuturi et al., 2014), (Solomon et al., 2015)

Counting modes



$$\sum_{i=1}^{n} \delta_{x_i}$$

$$\kappa_\sigma(x, y) = \exp\left(\frac{||x - y||_2}{\sigma}\right)$$



$$\mathbb{H}[X] = \log(n) \neq \#\text{ of modes}$$

(Arora et al., 2018)

Mutual Information

Definition (Mutual Information)

The mutual information between two jointly distributed random variables X and Y is the KL divergence between the joint distribution and its independent factorization.

$$\mathbb{I}[X;Y] \triangleq \mathbb{KL}[p(X,Y) \,||\, p(X) \otimes p(Y)]$$

Theorem (Data Processing Inequality)

Let $X \to Y \to Z$ form a Markov chain. Then, $\mathbb{I}[X;Y] \geq \mathbb{I}[Y;Z]$

Similarity sensitive DPI

Theorem (S.S. Data Processing Inequality)

If $X \to Y \to Z$ is a Markov chain, we have that

$$\mathbb{I}^{\mathbf{K},\mathbf{\Psi}}[X;Z] \leq \mathbb{I}^{\mathbf{K},\mathbf{\Lambda}}[X;Y] + \mathbb{I}^{\mathbf{K},\mathbf{\Psi},\mathbf{\Lambda}}[X;Z|Y].$$

For $\mathbf{\Lambda} = \mathbf{I}$, the original DPI is recovered.

Imperfect identifiability of Y

Research Agenda

Research agenda: Theory

- **Concavity**
  - Validity of proposed definitions
  - Call for (new?) mathematical tools
- **Operational representation**
  - $\mathbb{H}_2[X] = -\sum_{x \in \mathcal{X}} p(x) \log_2(p(x)) \leq \inf_{C \text{ inject.}} \mathbb{L}_X(C)$
  - Analogy for GAIT entropy?
- **Axiomatization**
  - Leinster and Cobbold's definition *ad hoc*
  - Underlying desirable properties?
- **Information Theory results**
  - Following "geometric" DPI
  - Extension of core results in IT
- **Links with kernel theory**
  - KT: Connections are evident, but unexplored
  - Functional analysis vs IT approaches

Research agenda: Applications

- **Representation learning**
  - Tschannen et al. (2020) call for "a new notion of information [which] should account for **both the amount of information** stored in a representation **and the geometry** of the induced space necessary for good performance on downstream tasks"
  - Wasserstein Dependency Measure; Hilbert Schmidt Information Criterion

- **Entropy regularization for Reinforcement Learning**
  - Encouraging exploration in *structured* action spaces
  - Naïve entropy may over-estimate the actual exploration level

(Tschannen et al., 2020), (Ozair et al., 2019), (Gretton et al., 2005)

Timeline

- **S20:** Completion of internship at Qualcomm Research
- **F20:** Applications in representation learning. Supervision of undergraduate student's thesis
- **W20-S21:** Desiderata and axiomatic characterization of geometric information theory concepts
- **F21:** Applications in data compression or entropy regularization in reinforcement learning
- **W21-S22:** Further research and thesis writing period
- **S22:** Estimated graduation date

References

- S. Arora, A. Risteski, and Y. Zhang. Do GANs Learn the Distribution? Some Theory and Empirics. In *ICLR*, 2018.
- L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 7(3):200-217, 1967.
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2005.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. 2004.
- M. Cuturi and A. Doucet. Fast Computation of Wasserstein Barycenters. In *ICML*, 2014.
- A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 2005.
- T. Leinster and C. A. Cobbold. Measuring diversity: The importance of species similarity. Ecology, 93(3):477-489, 2012.
- D. J. MacKay. *Information Theory, Inference and Learning Algorithms*. 2003.
- D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.
- S. Ozair, C. Lynch, Y. Bengio, A. Van den Oord, S. Levine, and P. Sermanet. Wasserstein dependency measure for representation learning. In *NeurIPS*, 2019.
- A. Rényi. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1961.
- C. Shannon. A mathematical theory of communication. The Bell System Technical Journal, 27(3):379-423, 1948.
- A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *ALT*, 2007.
- J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains. ACM Transactions on Graphics (TOG), 34(4):1-11, 2015.
- M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On Mutual Information Maximization for Representation Learning. In *ICLR*, 2020.