Information-Theoretic Bounds in Information Theory

Ayfer Özgür
Stanford University

May 18, 2021

Information Theory

The Bell System Technical Journal

Vol. XXVII *July, 1948* *No. 3*

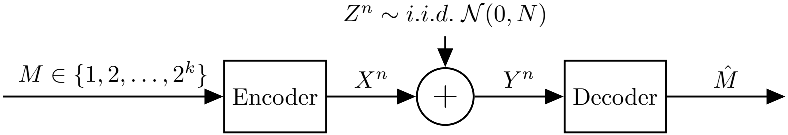
A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulati and PPM which exchange bandwidth for signal-to-no tensified the interest in a general theory of communicatio such a theory is contained in the important papers of Nyqui on this subject. In the present paper we will extend the th

Communication


$Z^n \sim i.i.d.\ \mathcal{N}(0, N)$
$M \in \{1, 2, \ldots, 2^k\}$
Encoder
X^n
$+$
Y^n
Decoder
$\hat{M}$

Encoder: $1, \ldots, 2^k \quad \rightarrow \quad X^n(1), \ldots, X^n(2^k) \in \mathbb{R}^n$ s.t. $||X^n||^2 \leq nP$

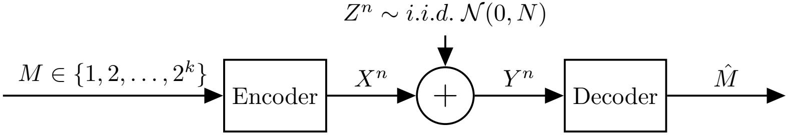
Memoryless Channel: $Y^n = X^n + Z^n$

Decoder: $\mathbb{R}^n \rightarrow 1, 2, \ldots, 2^k$

$$R = \frac{k}{n} \qquad P_e = \mathbb{P}(M \neq \hat{M})$$

Achievable R: for any $\epsilon > 0$, $\exists n, k$, encoder/decoder s.t. $R = k/n$ and $P_e \leq \epsilon$.
Capacity C: largest achievable rate R.

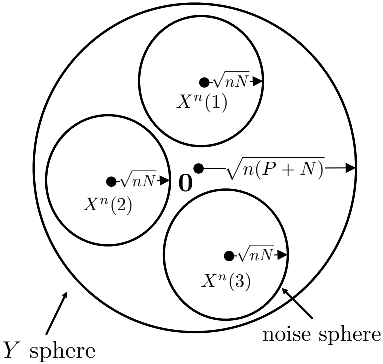
Communication


$Z^n \sim i.i.d.\ \mathcal{N}(0, N)$
$M \in \{1, 2, \ldots, 2^k\}$
Encoder
X^n
$+$
Y^n
Decoder
$\hat{M}$

Capacity of the AWGN channel

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N}\right)$$

Impossibility Bound: Sphere Packing



$$\# \text{ of } X^n \leq \frac{\left|\text{Sphere}\left(\sqrt{n(P+N)}\right)\right|}{\left|\text{Sphere}\left(\sqrt{nN}\right)\right|} \doteq \frac{2^{\frac{n}{2}\log 2\pi e(P+N)}}{2^{\frac{n}{2}\log 2\pi eN}} = 2^{\frac{n}{2}\log\left(1+\frac{P}{N}\right)}$$

Impossibility Bound: Fano's Inequality

Tools:

- Information Measure Calculus:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y), \\ &= H(X) - H(X|Y). \end{aligned}$$

- Fano's Inequality: For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = \mathbb{P}(X \neq \hat{X})$, we have

$$1 + P_e \log|\mathcal{X}| \geq H(X|Y).$$

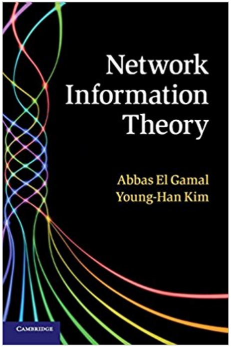
- Entropy-Power Inequality: If X^n and Y^n are independent random vectors with densities

$$2^{\frac{2}{n}h(X^n+Y^n)} \geq 2^{\frac{2}{n}h(X^n)} + 2^{\frac{2}{n}h(Y^n)}$$

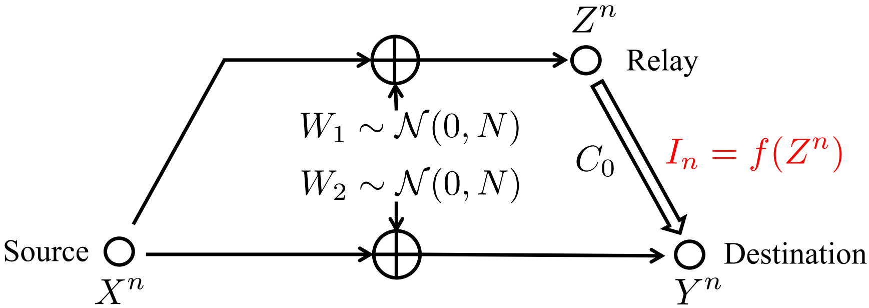
Impossibility Bound: Fano's Inequality

$$
\begin{aligned}
nR = H(M) &= I(M;\hat{M}) + H(M|\hat{M}) \\
&\leq I(X^n;Y^n) + n\epsilon_n \\
&= H(Y^n) - H(Y^n|X^n) + n\epsilon_n \\
&\leq \sum_i H(Y_i) - H(Y_i|X_i) + n\epsilon_n \\
&= \sum_i I(X_i;Y_i) + n\epsilon_n \\
&\leq n \sup_{p(X)} I(X;Y) + n\epsilon_n \\
&= \frac{n}{2}\log\left(1+\frac{P}{N}\right) + n\epsilon_n
\end{aligned}
$$

Network Information Theory



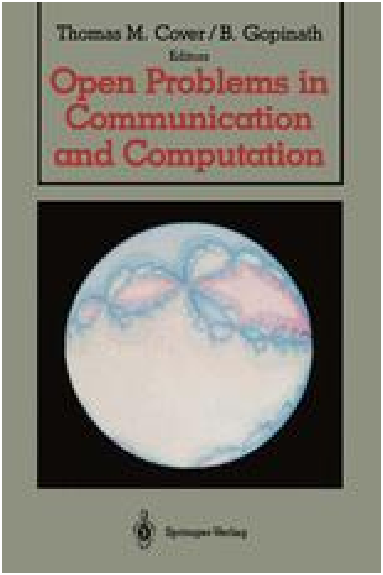
Relay Channel

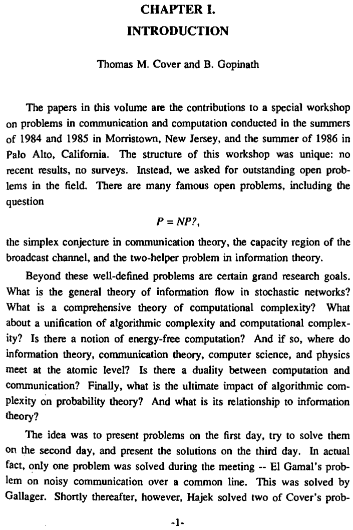


$$Z^n = X^n + W_1^n \qquad Y^n = X^n + W_2^n$$

W_1^n and W_2^n i.i.d. Gaussian $\mathcal{N}(0, N)$ independent of each other.

The story goes...


Thomas M. Cover / B. Gopinath
Editors
Open Problems in Communication and Computation
Springer-Verlag



CHAPTER I.

INTRODUCTION

Thomas M. Cover and B. Gopinath

The papers in this volume are the contributions to a special workshop on problems in communication and computation conducted in the summers of 1984 and 1985 in Morristown, New Jersey, and the summer of 1986 in Palo Alto, California. The structure of this workshop was unique: no recent results, no surveys. Instead, we asked for outstanding open problems in the field. There are many famous open problems, including the question

$$P = NP?,$$

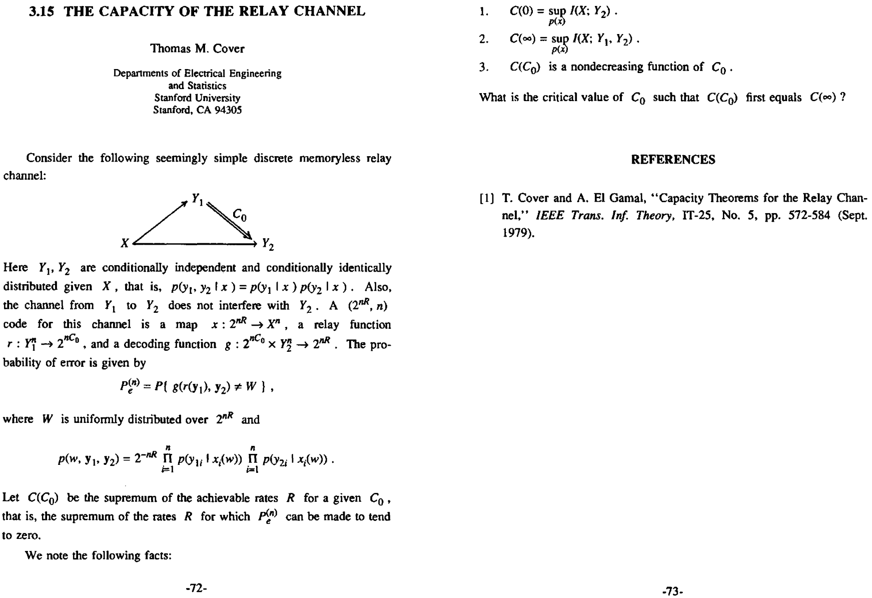
the simplex conjecture in communication theory, the capacity region of the broadcast channel, and the two-helper problem in information theory.

Beyond these well-defined problems are certain grand research goals. What is the general theory of information flow in stochastic networks? What is a comprehensive theory of computational complexity? What about a unification of algorithmic complexity and computational complexity? Is there a notion of energy-free computation? And if so, where do information theory, communication theory, computer science, and physics meet at the atomic level? Is there a duality between computation and communication? Finally, what is the ultimate impact of algorithmic complexity on probability theory? And what is its relationship to information theory?

The idea was to present problems on the first day, try to solve them on the second day, and present the solutions on the third day. In actual fact, only one problem was solved during the meeting -- El Gamal's problem on noisy communication over a common line. This was solved by Gallager. Shortly thereafter, however, Hajek solved two of Cover's prob-

-1-

Cover's Problem



3.15 THE CAPACITY OF THE RELAY CHANNEL

Thomas M. Cover

Departments of Electrical Engineering
and Statistics
Stanford University
Stanford, CA 94305

Consider the following seemingly simple discrete memoryless relay channel:

Here Y_1, Y_2 are conditionally independent and conditionally identically distributed given X, that is, $p(y_1, y_2 \mid x) = p(y_1 \mid x)\, p(y_2 \mid x)$. Also, the channel from Y_1 to Y_2 does not interfere with Y_2. A $(2^{nR}, n)$ code for this channel is a map $x : 2^{nR} \to X^n$, a relay function $r : Y_1^n \to 2^{nC_0}$, and a decoding function $g : 2^{nC_0} \times Y_2^n \to 2^{nR}$. The probability of error is given by

$$P_e^{(n)} = P\{\, g(r(y_1), y_2) \neq W \,\},$$

where W is uniformly distributed over 2^{nR} and

$$p(w, y_1, y_2) = 2^{-nR} \prod_{i=1}^{n} p(y_{1i} \mid x_i(w)) \prod_{i=1}^{n} p(y_{2i} \mid x_i(w)).$$

Let $C(C_0)$ be the supremum of the achievable rates R for a given C_0, that is, the supremum of the rates R for which $P_e^{(n)}$ can be made to tend to zero.

We note the following facts:

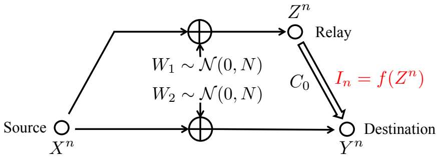
1. $C(0) = \sup_{p(x)} I(X; Y_2)$.
2. $C(\infty) = \sup_{p(x)} I(X; Y_1, Y_2)$.
3. $C(C_0)$ is a nondecreasing function of C_0.

What is the critical value of C_0 such that $C(C_0)$ first equals $C(\infty)$?

REFERENCES

[1] T. Cover and A. El Gamal, "Capacity Theorems for the Relay Channel," *IEEE Trans. Inf. Theory*, IT-25, No. 5, pp. 572-584 (Sept. 1979).

Gaussian case



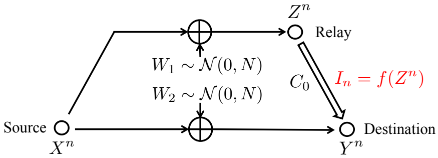
$$C(\infty) = \frac{1}{2} \log \left(1 + \frac{2P}{N}\right)$$

Cutset argument:

$$C_0^* \geq \frac{1}{2} \log \left(1 + \frac{2P}{N}\right) - \frac{1}{2} \log \left(1 + \frac{P}{N}\right).$$

Potentially, $C_0^* \to 0$ as $P/N \to 0$.

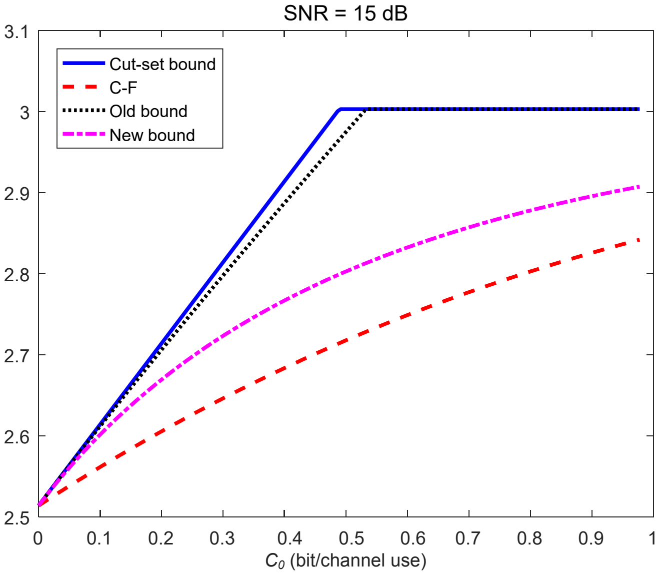
Solution to Cover's Problem



Theorem [Wu, Barnes, Özgür, 2019]

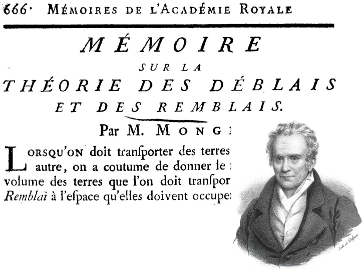
$$C_0^* = \infty$$

Upper Bound on the Capacity



Optimal Transport

Monge, 1781:



666· MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.

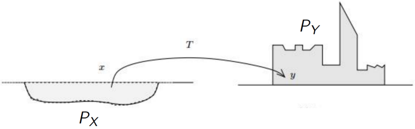
Par M. MONG

LORSQU'ON doit tranſporter des terres
autre, on a coutume de donner le
volume des terres que l'on doit tranſpor
Remblai à l'eſpace qu'elles doivent occupe

"The note on land excavation and infill"

Given two probability measures P_X and P_Y and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$

$$\inf_{T: T(X)=Y} \mathbb{E}[c(X, Y)]$$



Optimal Transport

Kantorovich, 1940:

Л. В. Канторович

О ПЕРЕМЕЩЕНИИ МАСС

Мы будем считать R метрическим компактным пространством, хотя некоторые из приведенных определений и результатов могут быть высказаны и для пространств более общего вида.

Пусть $\Phi(e)$ распределение масс, т.е. функция совокупности: 1) определенная для борелевских множеств, 2) неотрицательная: $\Phi(e) \geq 0$, 3) абсолютно-аддит
$e_i \cap e_k = \emptyset$ $(i \neq k)$, то $\Phi(e) = \Phi(e$
другое распределение масс, приче
нием масс будем называть такую
ную для пар (B)-совокупностей e, e
абсолютно-аддитивную по каждому
$\Psi(e, R) = \Phi(e)$; $\Psi(R, e') = \Phi'(e')$.

Пусть $r(x, y)$ известная непрерыв
— работа по перемещению единиц

Работой по перемещению данны
называть величину

$$W(\Psi, \Phi, \Phi') = \int_R \int_R r(x, x')\Psi(de, de'$$

где $\{e_i\}$ дизъюнктны и $\sum_1^n e_i = R$, $\{$
$x_i \in e_i$, $x'_k \in e'_k$, и λ наибольшее из
$\text{diam}\, e'_k$ $(k = 1, 2, \ldots, m)$.

Given two probability measures P_X and P_Y and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$

$$\inf_{P \in \Pi(P_X, P_Y)} \mathbb{E}_P[c(X, Y)]$$

where $\Pi(P_X, P_Y)$ is the set of couplings of P_X and P_Y.

Optimal Transport

A very fruitful area in mathematics:

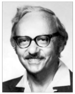
- Many famous names and awards:

Monge

Kantorovich

Koopmans

Nobel (1975)


Dantzig

. . .

Villani
Fields (2010)

Figalli
Fields (2018)

- Wide range of applications: economics, geometry, quantum mechanics, fluid dynamics, optics, mathematical statistics, and meteorology and most recently machine learning.

Transportation Cost Inequalities

Wasserstein distance:

When $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$ and $c(x^n, y^n) = ||x^n - y^n||^2$,

$$W_2^2(P_{X^n}, P_{Y^n}) = \inf_{P \in \Pi(P_{X^n}, P_{Y^n})} \mathbb{E}_P \left[||X^n - Y^n||^2 \right]$$

is called the quadratic Wasserstein distance.

Transportation Cost Inequalities

Wasserstein distance:

When $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$ and $c(x^n, y^n) = ||x^n - y^n||^2$,

$$W_2^2(P_{X^n}, P_{Y^n}) = \inf_{P \in \Pi(P_{X^n}, P_{Y^n})} \mathbb{E}_P \left[||X^n - Y^n||^2 \right]$$

is called the quadratic Wasserstein distance.

Theorem (Talagrand, 96)

For $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{X^n} \ll P_{Y^n}$,

$$W_2^2(P_{X^n}, P_{Y^n}) \leq 2D(P_{X^n} \| P_{Y^n}),$$

where the inequality is tight if and only if $P_{X^n} = \mathcal{N}(\mu, I_n)$ for some $\mu \in \mathbb{R}^n$.

Concentration of Measure

Theorem (Law of Large Numbers)

“The average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.”

Concentration of Measure

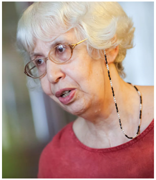
Theorem (Law of Large Numbers)

“The average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.”

Theorem (Gaussian concentration)

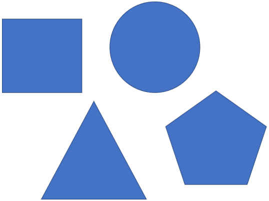
Let $X^n \sim \mathcal{N}(0, I_n)$ and f be L-Lipschitz continuous,

$$\mathbb{P}\left(|f(X^n) - \mathbb{E}[f(X^n)]| > t\right) \leq e^{-\frac{t^2}{L^2}}.$$



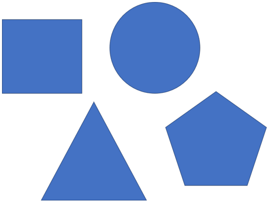
Isoperimetric Inequalities

- In $\mathbb{R}^2$:

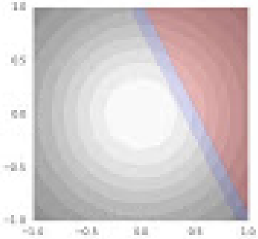


Isoperimetric Inequalities

- In $\mathbb{R}^2$:



- In Gaussian space:



Information Constrained Optimal Transport

For any $R > 0$, define the information constrained Wasserstein distance

$$W_2^2(P_X, P_Y; R) = \inf_{\substack{P \in \Pi(P_X, P_Y): \\ I(X;Y) \leq R}} \mathbb{E}_P[||X - Y||^2]$$

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

Bounding Information Constrained Optimal Transport

Theorem (Bai, Wu, Özgür, 2020)

For $P_{Y^n} = \mathcal{N}(0, I_n)$ *and* $P_{X^n} \ll P_{Y^n}$, *we have*

$$W_2^2(P_{X^n}, P_{Y^n}; R) \leq \mathbb{E}[\|X^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2}{n} h(X^n)} \left(1 - e^{-\frac{2R}{n}}\right)}$$

which is tight when $P_{X^n} = \mathcal{N}(\mu, \sigma^2 I_n)$ *for some* $\mu \in \mathbb{R}^n$ *and* $\sigma > 0$.

Corollary

Corollary

For $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{X^n} \ll P_{Y^n}$, we have

$$W_2^2(P_{X^n}, P_{Y^n}) \leq \mathbb{E}[\|X^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2}{n} h(X^n)}},$$

which is tight when $P_{X^n} = \mathcal{N}(\mu, \sigma^2 I_n)$ for some $\mu \in \mathbb{R}^n$ and $\sigma > 0$.

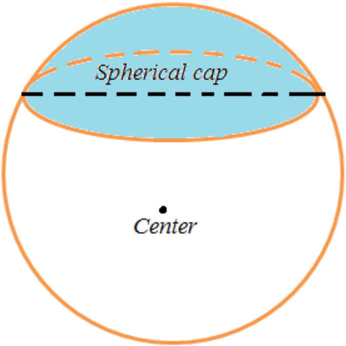
- Tighter than Talagrand's inequality for any P_{X^n}.
- Achieved with equality for a wider class of P_{X^n}.

Isoperimetry on the Sphere

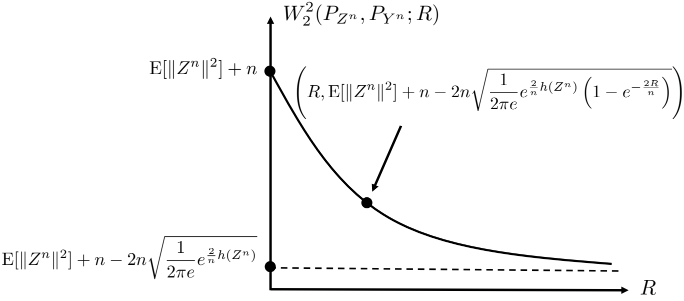
Strengthening Talagrand's Inequality:

$$W_2^2(P_{X^n}, P_{Y^n}) \leq \mathbb{E}[\|X^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2}{n} h(X^n)}},$$

$$\Downarrow$$



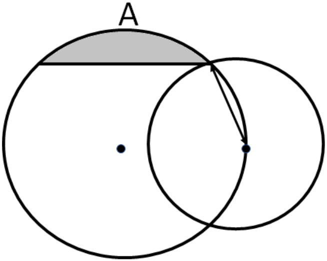
Information – Cost Trade-off



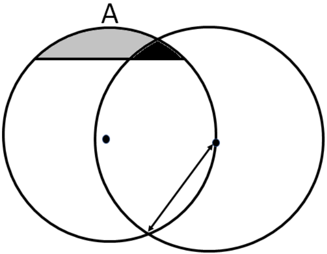
Trade-off tight when $P_{Z^n} = \mathcal{N}(\mu, \sigma^2 I_n)$

A New Measure Concentration Result on the Sphere

Classical Blowing-up Lemma:

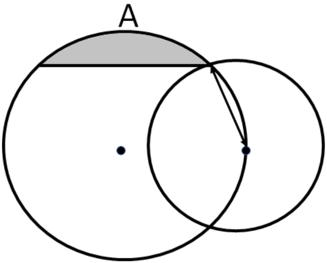


New Blowing-up Lemma:

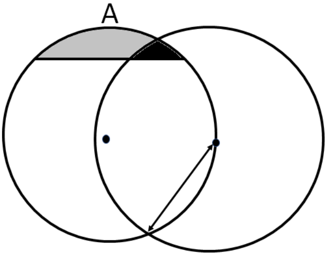


A New Measure Concentration Result on the Sphere

Classical Blowing-up Lemma:



New Blowing-up Lemma:



Born again: "A good new idea is often a reincarnation of a good old idea."

For any $R > 0$, define the information constrained Wasserstein divergence

$$W_2(P_X, P_Y; R) = \inf_{P \in \Pi(P_X, P_Y) : I(X;Y) \leq R} \{\mathbb{E}_P[||X - Y||^2]\}^{1/2}$$

Born again: "A good new idea is often a reincarnation of a good old idea."

For any $R > 0$, define the information constrained Wasserstein divergence

$$W_2(P_X, P_Y; R) = \inf_{P \in \Pi(P_X, P_Y) : I(X;Y) \leq R} \{\mathbb{E}_P[||X - Y||^2]\}^{1/2}$$

Lagrangian function:

$$\inf_{P \in \Pi(P_X, P_Y)} \mathbb{E}_P[||X - Y||^2] + \lambda I(X; Y)$$

Born again: "A good new idea is often a reincarnation of a good old idea."

For any $R > 0$, define the information constrained Wasserstein divergence

$$W_2(P_X, P_Y; R) = \inf_{P \in \Pi(P_X, P_Y) : I(X;Y) \leq R} \{\mathbb{E}_P[||X - Y||^2]\}^{1/2}$$

Lagrangian function:

$$\inf_{P \in \Pi(P_X, P_Y)} \mathbb{E}_P[||X - Y||^2] + \lambda I(X; Y)$$

Schrödinger problem,
1931:

Born again: "A good new idea is often a reincarnation of a good old idea."

For any $R > 0$, define the information constrained Wasserstein divergence

$$W_2(P_X, P_Y; R) = \inf_{P \in \Pi(P_X, P_Y): I(X;Y) \leq R} \{\mathbb{E}_P[||X - Y||^2]\}^{1/2}$$

Lagrangian function:

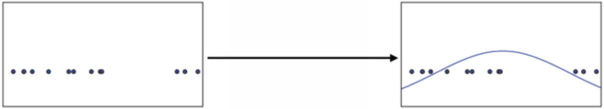
$$\inf_{P \in \Pi(P_X, P_Y)} \mathbb{E}_P[||X - Y||^2] + \lambda I(X; Y)$$

Schrödinger problem, 1931:

ML Literature: Entropy regularized OT, Sinkhorn divergence:

- Faster to compute (Sinkhorn algorithm).
- Better empirical accuracy for ML tasks.
- OT suffers the curse of dimensionality, but regularized OT does not. [Genevay, Bach, Peyre, Cuturi]

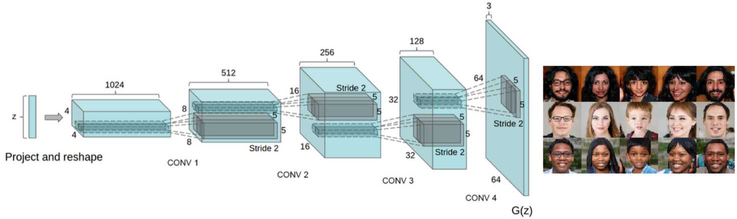
Learning Generative Models



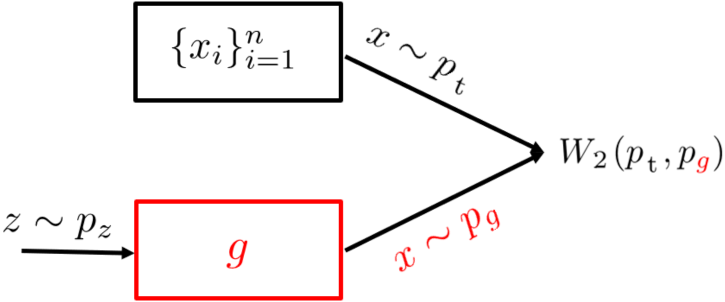
Given observed data, find a probabilistic model to generate new data, e.g. by fitting a parametric family of densities $\{p_\theta, \theta \in \Theta\}$.

Learning Generative Models

When data is high-dimensional (200x100 pixels):



Generative Adversarial Networks (GANs)



$$\min_{g \in \mathcal{G}} W_2(p_t, p_g)$$

Problems: slow convergence, mode collapse, stability issues.

Benchmark: Linear Gaussian Quadratic Case

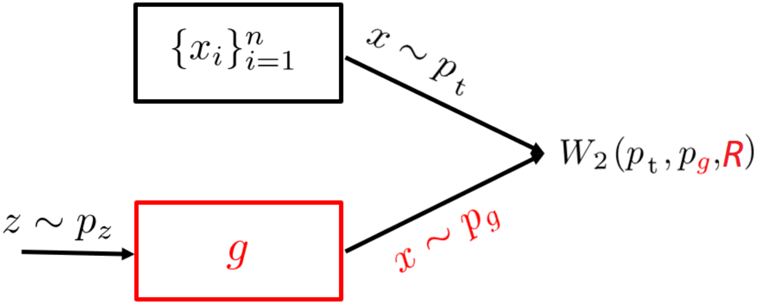
Feizi, Farnia, Ginart, Tse 2020:

- $\mathcal{G}$: set of linear functions.
- $p_t = \mathcal{N}(0, I_d)$ ($p_z = \mathcal{N}(0, I_r)$ where $r < d$).
- Quadratic: W_2 as our distance metric.

Results:

- g^* : r-PCA solution.
- Slow convergence: $g_n \to g^*$ as $n^{-2/d}$:
  e.g. if $d = 16$, for $n^{-2/d} < 0.01$ we need
  $n > 10,000,000,000,000,000$.

Generative Adversarial Networks (GANs)



$$\min_{g \in \mathcal{G}} W_2(p_t, p_g; R)$$

Benchmark: Linear Gaussian Quadratic Case

Reshetova, Bai, Wu, Özgür to be presented in ISIT'2021:

- $\mathcal{G}$: set of linear functions.
- $p_t = \mathcal{N}(0, I_d)$ ($p_z = \mathcal{N}(0, I_r)$ where $r < d$).
- Quadratic: W_2 as our distance metric.

Results:

- g^* : soft-thresholding r-PCA solution.
- Fast convergence: $g_n \to g^*$ as $\frac{K_d}{\sqrt{n}}$.