NATIXIS
BEYOND BANKING

The Prediction of Geostrategic Event

by O. Croissant

7th of jan 2024

Plan

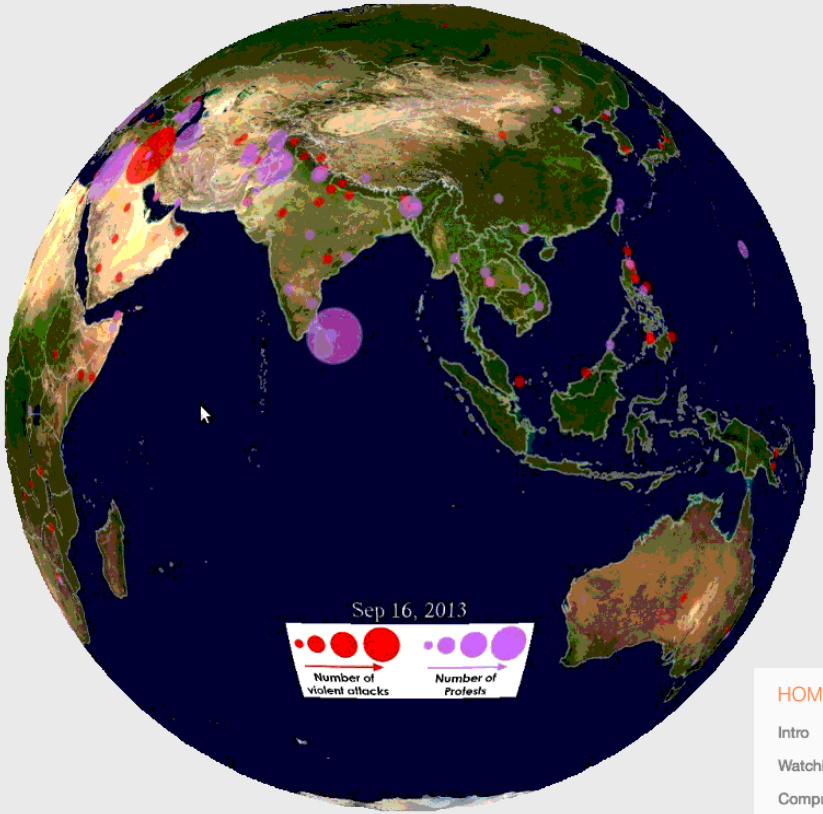
- 2) EVENT PREDICTION DATASET
- 3) TRANSFORMER PRESENTATION
- 4) LANGUAGE MODELS AND SENTIENT AI
- 5) GRAM MATRIX AND GRADIENT FLOWS
- 6) KL DIVERGENCE
- 7) THE BENAMOU BRENIER APPROACH
- 8) GRADIENT FLOWS: THE DEMONSTRATIONS
- 9) GRADIENT FLOWS: THE THEORY
- 11) MEAN-FIELD THEORIES
- 12) RELATIONSHIPS WITH QUANTUM STATISTICAL PHYSICS
- 15) MULTIVARIATE RENYI INFORMATION
- 16) INFORMATION THEORY
- 17) OPTIMAL TRANSPORT AND NLP LOSSES

To Summary

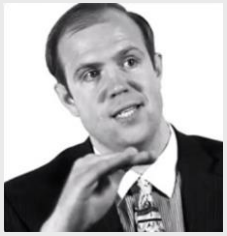
2) EVENT PREDICTION DATASET

To Summary

GDELT Project



Kalev H. Leetaru



One of Foreign Policy Magazine's Top 100 Global Thinkers of 2013,

300 million event records in over 300 categories covering the entire world from 1979 to present

Hundreds of thousands of broadcast, print, and online news sources from every corner of the globe in more than 100 languages

HOME	BLOG	DATA	SOLUTIONS
Intro	The GDELT Blog	Intro	Intro
Watching		GDELT Analysis Service	Situational Awareness
Computing		Google BigQuery	Influencer Networks
Downloading		Raw Data Files	Risk Assessment & Global Trends
Blogging		Documentation	Policy Reaction
			Humanitarian & Crisis Response

To Summary

GDELT Data

GDELT Analysis Service

Free cloud-based service that offers a variety of tools and services to allow you to visualize, explore, and export GDELT - a great way to get started using GDELT for the first time.

Google BigQuery

Query, export, and even conduct sophisticated analyses and modeling of the entire dataset using standard SQL, with even the most complex queries returning in near-realtime.

Raw Data Files

Advanced users and those with unique use cases can download the entire underlying event and graph datasets in CSV format - over 2.5TB for last year alone.

GDELT 1.0 Event Database

Daily Updates

All GDELT Event Files

- md5sums
- filesizes
- GDELT.MASTERREDUCEDV2.1979-2013.zip (1.1GB) (MD5: f6fcb7e955e35f93c9dae427c07b545d)
- 20221205.export.CSV.zip (5.7MB) (MD5: 73dcc6556 5e3ac010495c3a75d16ed51)
- 20221204.export.CSV.zip (3.1MB) (MD5: 3711b97dac626e39fb1c866eb1e2616d)
- 20221203.export.CSV.zip (3.9MB) (MD5: d425df05eeceb6cf925c73c6317c5e51)
- 20221202.export.CSV.zip (6.6MB) (MD5: ac387488b14515a0925574f9f4c8378f)
- 20221201.export.CSV.zip (6.9MB) (MD5: 8c30a650f08f6b2c03dc704c15cb4ffe)
- 20221130.export.CSV.zip (7.0MB) (MD5: 836da08caf3b581f05f39c196c554509)
- 20221129.export.CSV.zip (6.9MB) (MD5: 2e2f164883e6dd99868316ba584ff7f8)
- 20221128.export.CSV.zip (5.4MB) (MD5: 272608df805117bd8c96342811435eab)
- 20221127.export.CSV.zip (2.8MB) (MD5: 789fee593ec9ffa60332f59f689326fd)
- 20221126.export.CSV.zip (3.1MB) (MD5: 7dc39f89d4f7ab090d2d33337360f651)
- 20221125.export.CSV.zip (4.7MB) (MD5: 2b8bd6163f7d7d4436edb5372b0ce0db)
- 20221124.export.CSV.zip (5.3MB) (MD5: 0a405df3a590d8e455d581437f62ac52)
- 20221123.export.CSV.zip (6.7MB) (MD5: d425fd85d3fa2e25ed9e4a6952016b4b)
- 20221122.export.CSV.zip (6.8MB) (MD5: 91c812a921d6e93c33ceb9905d1b9f8e)
- 20221121.export.CSV.zip (5.9MB) (MD5: ba83b678a382da3f4129276e7cc1545b)
- 20221120.export.CSV.zip (3.3MB) (MD5: 111515d7de868f1919e189e13b427bc3)

GDELT 2.0 Event Database

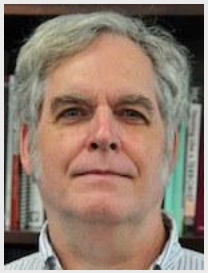
15 Minute Updates

From realtime translation of the world's news in 65 languages, to measurement of more than 2,300 emotions and themes from every article

To Summary

CAMEO

Philip A. Schrodt (Project Director):



Pennsylvania State University

CAMEO
Conflict and Mediation Event Observations
Event and Actor Codebook

Event Data Project
Department of Political Science
Pennsylvania State University
Pond Laboratory
University Park, PA 16802

Version: 1.1b3 : March 2012

CAMEO	073
Name	**Provide humanitarian aid**
Description	Extend, provide humanitarian aid, mainly in the form of emergency assistance.
Usage Notes	This code refers to events such as provisions of shelter, food, medicine, and evacuation of victims. The lead must report the delivery of such aid; promises to provide aid should be coded under category 033. Note that provisions of peacekeeping or other military forces are coded as 074 instead.
Example	Swiss doctors handed over 700 kg of medicine to the Red Crescent in Bam, Iran, according to the Swiss Agency for Development and Cooperation.
Example	Benin opened its borders today to most West Africans ordered out of Nigeria as illegal aliens, but was still refusing admittance to Ghanaians, by far the biggest group involved, Benin police said.
Example	U.N. helicopters evacuated the wounded from the besieged Bosnian town of Gorazde on Friday.

http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

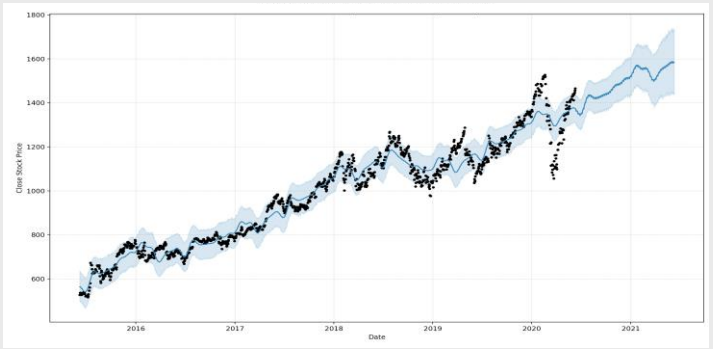
To Summary

Fusion of NLP and Prediction

Textual information

TimeSeries

Prediction



To Summary

History as sequences of events

Sequence of events = phrases of a language

A people

A region

A culture

To Summary

Decategorification

58 CAMEO Columns worldwide X 300 Millions rows

```
['GLOBALEVENTID','SQLDATE','MonthYear','Year','FractionDate','Actor1Code',
 'Actor1Name','Actor1CountryCode','Actor1KnownGroupCode','Actor1EthnicCode',
 'Actor1Religion1Code','Actor1Religion2Code','Actor1Type1Code',
 'Actor1Type2Code','Actor1Type3Code','Actor2Code','Actor2Name',
 'Actor2CountryCode','Actor2KnownGroupCode','Actor2EthnicCode',
 'Actor2Religion1Code','Actor2Religion2Code','Actor2Type1Code',
 'Actor2Type2Code','Actor2Type3Code','IsRootEvent','EventCode','EventBaseCode',
 'EventRootCode','QuadClass','GoldsteinScale','NumMentions','NumSources',
 'NumArticles','AvgTone','Actor1Geo_Type','Actor1Geo_FullName',
 'Actor1Geo_CountryCode','Actor1Geo_ADM1Code','Actor1Geo_Lat','Actor1Geo_Long',
 'Actor1Geo_FeatureID','Actor2Geo_Type','Actor2Geo_FullName',
 'Actor2Geo_CountryCode','Actor2Geo_ADM1Code','Actor2Geo_Lat','Actor2Geo_Long',
 'Actor2Geo_FeatureID','ActionGeo_Type','ActionGeo_FullName',
 'ActionGeo_CountryCode','ActionGeo_ADM1Code','ActionGeo_Lat','ActionGeo_Long',
 'ActionGeo_FeatureID','DATEADDED']
```

Country_linked 14 columns dataset

'ActionGeo_CountryCode'

```
['GLOBALEVENTID','SQLDATE','MonthYear','Year','FractionDate',
 'IsRootEvent','EventCode','EventBaseCode',
 'EventRootCode','QuadClass','GoldsteinScale','Actor1Geo_CountryCode',
 'Actor2Geo_CountryCode']
```

+ Multi_Country_Flag

Transformer

Embeddings space

Maximization of the Log-Volume of Translation roots

Textual Translation

optional

Addition of synonymity noise

Transformer

Historical Sequence of textual events-sets

To Summary

Building of the dataset

Historical Sequence of textual events-sets

Keeping :
- Ponctual Statistical Structure
- Set Statistical Structure

Source : 100-10000 pts

Random Generation

Target: 3 – 10 pts

Distribution of points
on the unitary hyper-sphere
In the embedding space

Distribution of points
on the unitary hyper-sphere
In the embedding space

3-10 vectors embedding space

Small text : 3-10 simple phrases

Decoder - Transformer

Daily descriptor

To Summary

Learning

https://shivanandroy.com/fine-tune-t5-transformer-with-pytorch/

https://www.youtube.com/watch?v=U6x8_BP69DM

https://lilianweng.github.io/posts/2021-01-02-controllable-text-generation/

Dataset

10 000 Sequences of 50 daily descriptors

Metric

Rouge =

$$\frac{\sum_r \sum_s \text{match}(\text{gram}_{s,c})}{\sum_r \sum_s \text{count}(\text{gram}_s)}$$

To Summary

Event Prediction Principle

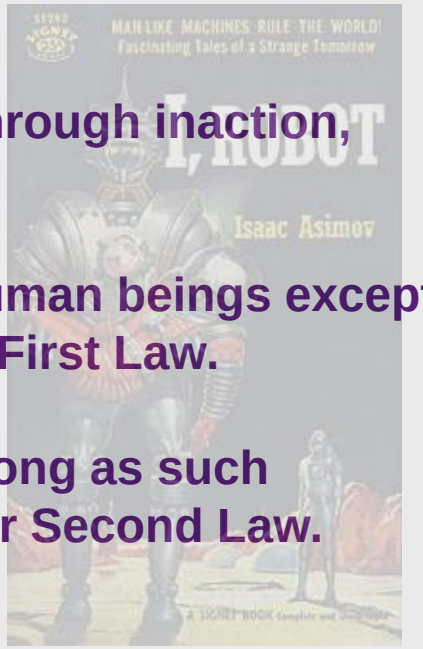
1) Generation of 100 possible future event-set

2) Projection of the cloud of points on the basis of standard event to determine probabilities.

3) Reconstruction of the standard events from the projections most likely arguments for the detailled actors and countries

To Summary

Isaac Asimov : 3 laws of robotics (+1)



First Law

A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law

A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law

A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

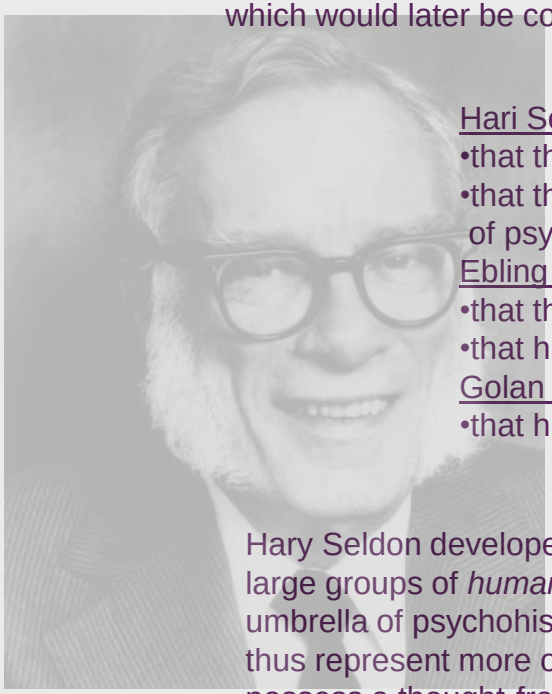
Zeroth Law

A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

To Summary

Isaac Asimov : Psychohistory

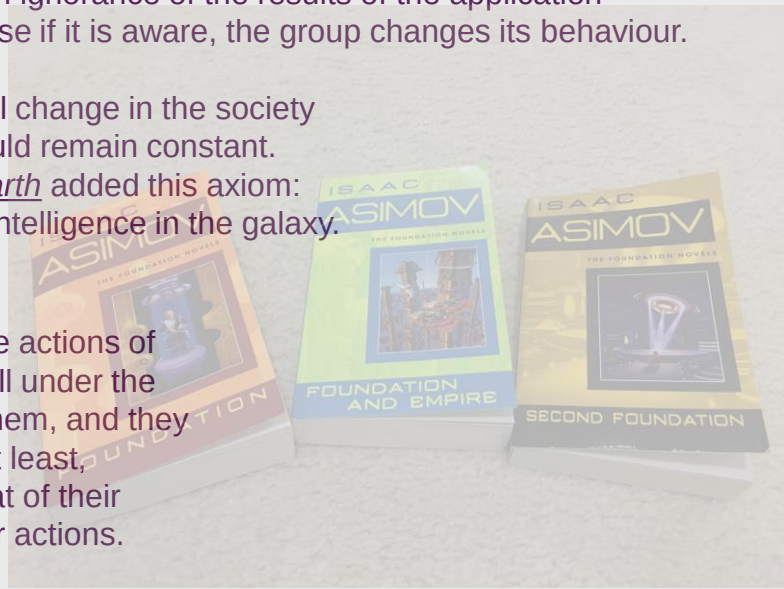
Psychohistory is a fictional science in Isaac Asimov's *Foundation* universe which combines history, sociology, and mathematical statistics to make general predictions about the future behavior of very large groups of people, such as the Galactic Empire. It was first introduced in the four short stories (1942–1944) which would later be collected as the 1951 novel *Foundation*.



Hari Seldon, the hero in Asimov's romans established two axioms:
- that the population whose behavior was modeled should be sufficiently large
- that the population should remain in ignorance of the results of the application of psychohistorical analyses because if it is aware, the group changes its behaviour.

Ebling Mis added these axioms:
- that there would be no fundamental change in the society
- that human reactions to stimuli would remain constant.

Golan Trevize in *Foundation and Earth* added this axiom:
- that humans are the only sentient intelligence in the galaxy.

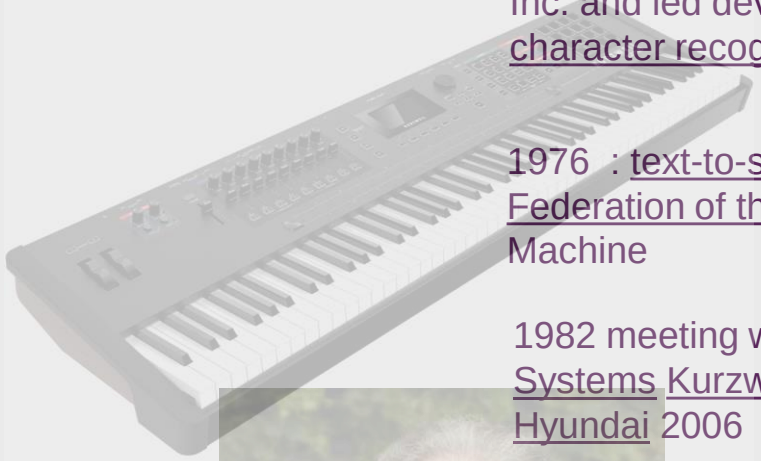


Hary Seldon developed psychohistory to predict the actions of large groups of *humans*. Even robots technically fall under the umbrella of psychohistory, because humans built them, and they thus represent more or less a human "action", or at least, possess a thought-framework similar enough to that of their human creators that psychohistory can predict their actions.

(from wikipedia)

To Summary

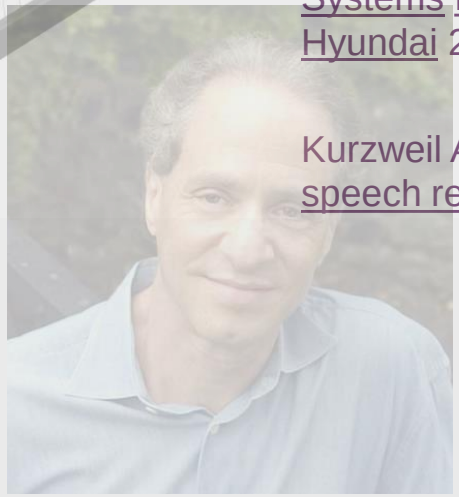
Ray Kurzweil:Genius Inventor



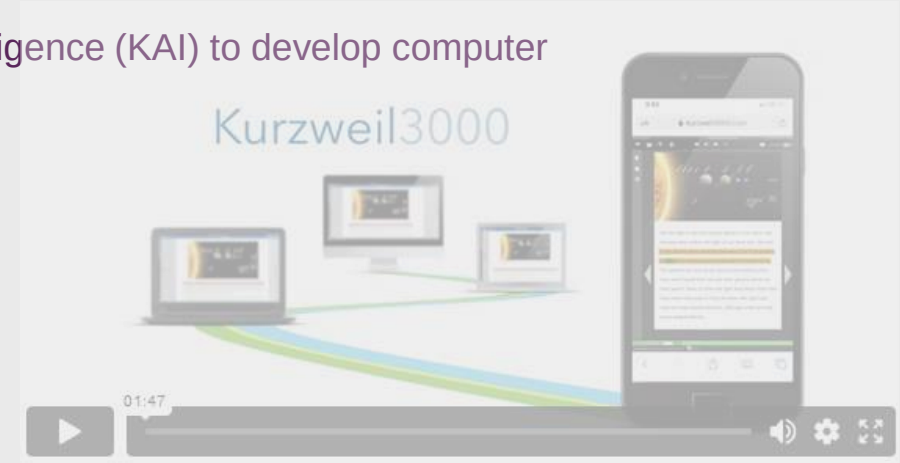
1974: Kurzweil founded Kurzweil Computer Products, Inc. and led development of the first omni-font optical character recognition

1976 : text-to-speech synthesis : leader of the National Federation of the Blind. Built the Kurzweil Reading Machine

1982 meeting with Stevie Wonder : Kurzweil Music Systems Kurzweil K250 -> Young Chang in 1990 -> . Hyundai 2006

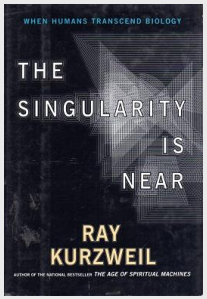


Kurzweil Applied Intelligence (KAI) to develop computer speech recognition



To Summary

Ray Kurzweil: The Singularity



1990 :, *The Age of Intelligent Machines*

1999 : *The Age of Spiritual Machines*

2005 : *The Singularity Is Near*

2009: Kurzweil, in collaboration with Google and the NASA Ames Research Center, announced the creation of the Singularity University

2012 : *How to Create a Mind*

Pattern Recognition Theory of Mind, the theory that the neocortex is a hierarchical system of pattern recognizers, and argues that emulating this architecture in machines could lead to an artificial superintelligence

2009 : Once the Singularity has been reached, Kurzweil says that machine intelligence will be infinitely more powerful than all human intelligence combined. Afterwards he predicts intelligence will radiate outward from the planet until it saturates the universe. The Singularity is also the point at which machines' intelligence and humans would merge. Kurzweil spells out the date very clearly: "I set the date for the Singularity—representing a profound and disruptive transformation in human capability—as 2045"

To Summary

5) GRAM MATRIX AND GRADIENT FLOWS: AN OVERVIEW

To Summary

Gram Matrix == Density Operator == Neural Tangent Kernel == Normalized Activation Batch Product == Generator Information Gradient Flow

Quantum Statistical physics

$$\hat{\rho} = \sum_k \frac{e^{-\beta E_k^{\mathrm{M}}}}{Z} |\Psi_k\rangle\langle\Psi_k|$$

Reny's Information

$$\mathbf{S}_\alpha(A_1, A_2, \cdots, A_k) = \mathbf{S}_\alpha\left(\frac{A_1 \circ A_2 \circ \cdots \circ A_k}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_k)}\right)$$

$$G(x_1, \ldots, x_n) = \begin{vmatrix} \langle x_1, x_1\rangle & \langle x_1, x_2\rangle & \ldots & \langle x_1, x_n\rangle \\ \langle x_2, x_1\rangle & \langle x_2, x_2\rangle & \ldots & \langle x_2, x_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_n, x_1\rangle & \langle x_n, x_2\rangle & \ldots & \langle x_n, x_n\rangle \end{vmatrix}$$

RKHS

$$k(x, y) = \langle k_x, k_y\rangle = \langle \phi(x), \phi(y)\rangle$$

Neural Tangent Kernel

$$K_{\theta_t}^L(x, x') = \nabla_\theta f(x, \theta_t) \nabla_\theta f(x', \theta_t)^T$$

Gradient Flows

$$\frac{\partial \mu_t}{\partial t} - \mathit{div}\left(\mu_t \nabla \frac{\partial \mathcal{G}(\mu_t)}{\partial \mu_t}\right) = 0$$

$$\partial_t \theta_p(t) = \left\langle \partial_{\theta_p} F^{(L)}(\theta(t)), d_t \right\rangle_{p^{in}}$$

To Summary

Gradient Flow

Statistical Physics

Entropy, Free Energy, Focker-Planck, Langevin, Boltzmann, Gibbs

Information Theory

Variational Inference, Regularization, Mutual Information, Information Bottleneck

$$\partial_t \theta_p(t) = \left\langle \partial_{\theta_p} F^{(L)}(\theta(t)), d_t \right\rangle_{p^{in}}$$

Stochastic Gradient Descent (SGD)

Differential Geometry

Riemannian manifolds, Geodesics, Connection,Curvature

Diffusion

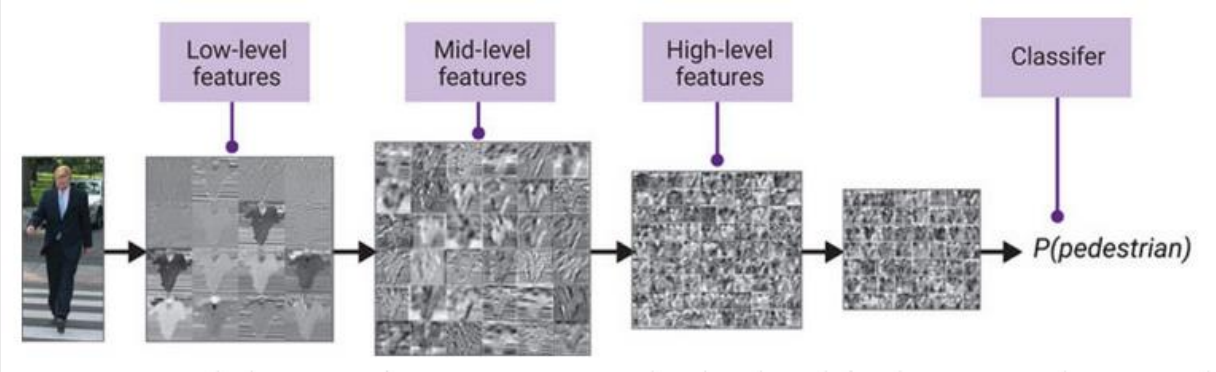
Domain adaptation,Clustering, Sampling, Metropolis-adjusted Langevin algorithm, Hamiltonian Monte Carlo

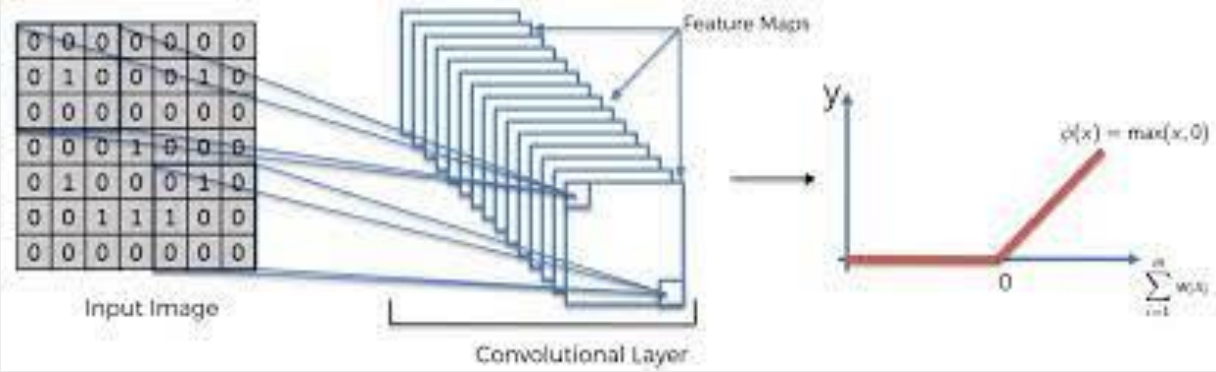
Optimal Transport

Wasserstein space , Sinkhorn algorithm

To Summary

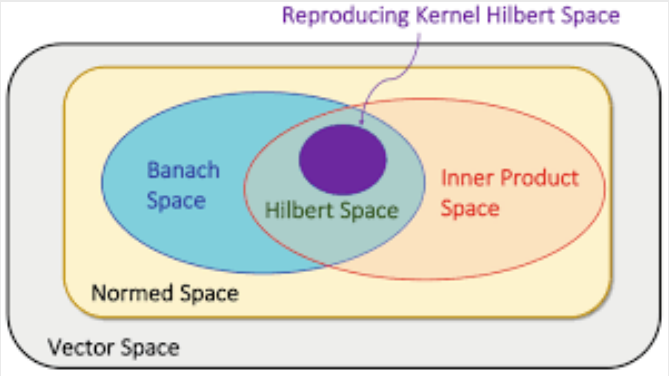
Feature Map





To Summary

RKHS



Riesz representation theorem

Let $\mathcal{H}$ be a Hilbert space over $\mathbb{R}$. If $T \in \mathcal{H}^*$, then there exists a unique vector u in $\mathcal{H}$ such that

$$T(v) = \langle v, u \rangle_{\mathcal{H}} \text{ for all } v \in \mathcal{H}$$

Inverting the Riesz Theorem

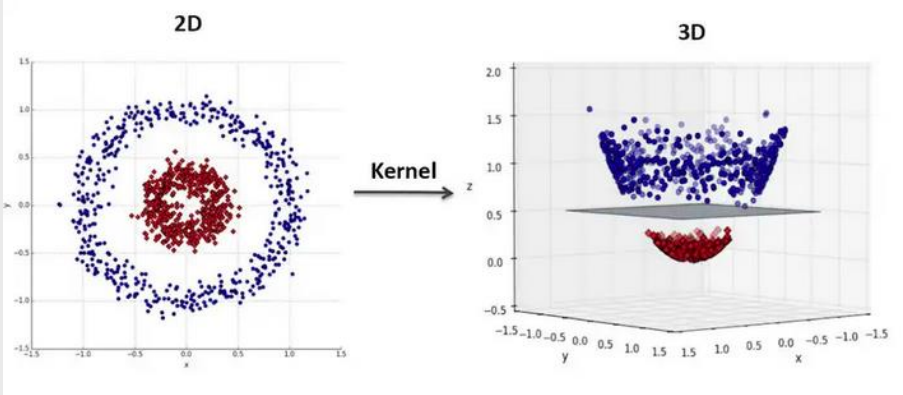
We say $\mathcal{H}$ is a *Reproducing Kernel Hilbert Space* if there exists a $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$, such that

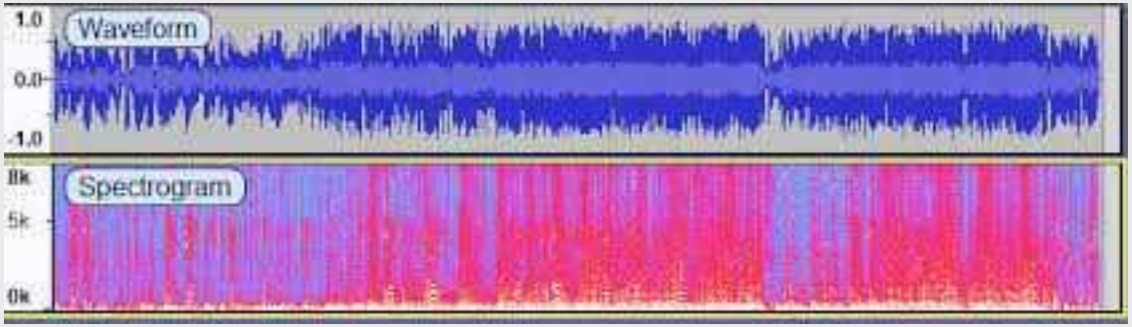
1. k has the reproducing property, i.e., $f(x) = \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}}$
2. k spans $\mathcal{H}$, that is, $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$

x -> k(y,x) is called a Feature Map

To Summary

Example of Feature Map





To Summary

Gram Matrix To RKHS :

A smart way to extend features and introduce natural codependance

Mercer Kernel K(x, y) :

Function of two variables that is symmetric and positive definite

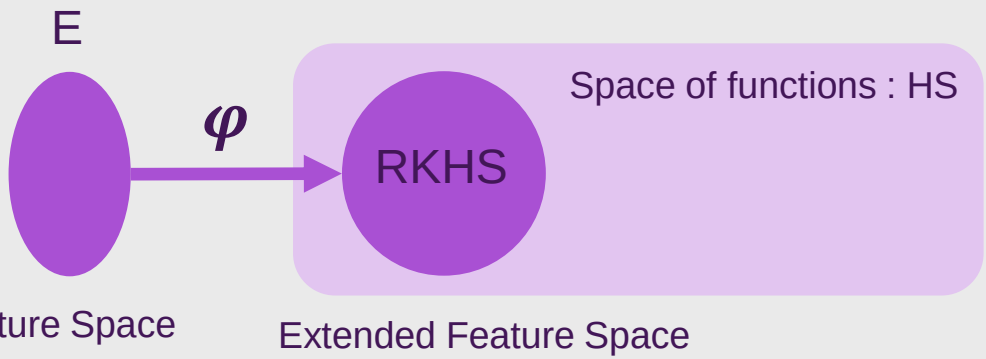
$$\int \int K(x,y)f(x)f(y)dx\,dy \geq 0 \iff \langle f\,, f\rangle \geq 0$$

Riez theorem

Every linear **continuous** operator is implemented as a scalar product.
The reciprocal is false (think about valuation at x)

RKHS

A closed subpace of the hilbert space of functions on E , where the reciprocal of Riez theorem is true



Reproducing Property

$$\langle \varphi(x), \varphi(y)\rangle = \langle x\,, y\rangle$$

To Summary

Neural Tangent kernel

$$f(W,X) \approx f(W_0,X) + \nabla_W f(W_0,X)^T (W - W_0) + \dots$$

$$\varphi(X) = \nabla_W f(W_0,X)$$

$$K(X_i,X_j) = \langle \varphi(X_i), \varphi(X_j) \rangle$$ **Neural Tangent Kernel**

NTK can be viewed as a specific type of Gram matrix that captures the local geometry of the parameter space during training

To Summary

Renyi alpha entropy

$$\mathbf{H}_\alpha(f) = \frac{1}{1-\alpha}\log\int_{\mathcal{X}} f^\alpha(x)dx$$

$$\alpha \to 1 \; H_1(X) = -\sum_x p(x)\log p(x)$$

$$\alpha \to \infty \; H_\infty(X) = -\log\max_x p(x)$$

Shannon Entropy

Min-Entropy

Shannon Mutual Information

$$I_1(X;Y) \stackrel{\text{def}}{=} H_1(X) - H_1(X|Y) = H_1(X) + H_1(Y) - H_1(X,Y)$$

Feature selection in machine learning: Renyi min-entropy vs Shannon entropy
arXiv:2001.09654v1 [cs.LG] 27 Jan 2020

To Summary

Why Reny's Information

Sensitivity to rare or frequent events

$0 < \alpha < 1$ => it is more sensitive to rare events,
$\alpha > 1$ => it is more sensitive to frequent events.

Information bottleneck method

Better trade-off between compression and relevance, making it easier to find the optimal solution

Time series data in finance

In this context, the adjustable parameter α allows for different perspectives on the complexity of the patterns.

Non-asymptotic or non-ergodic settings

When the law of large numbers does not readily apply, other entropy measures typically take over, for example the min-, the max-, or the collision entropy. The Renyi entropies nicely unify these different and isolated measures.

On quantum Renyi entropies: a new generalization and some properties
Martin Muller-Lennert, Frederic Dupuis, Oleg Szehr, Serge Fehr, and Marco Tomamichel

The world according to Renyi: Thermodynamics of multifractal systems Petr Jizba and Toshihico Arimitsu

To Summary

6) KL DIVERGENCE

To Summary

Kullback-Leibler Divergence

$$D_{KL}[p||q] = \int_{-\infty}^{\infty} p(x) Log \frac{p(x)}{q(x)} dx$$

$D_{KL}[p||p] \geq 0$ et $D_{KL}(p||p) = 0$

But it is not a distance but triangular inégality does not work

Is not symmetric

let $q(x)$ a prob. distribution defined on E et $\mathrm{F} \subset E$

Let $q_F(x)$ a prob. distribution restricted to F (conditionned by F)

Then $D_{KL}[q||q_F)] = \mathrm{Log}\ \frac{1}{Prob[F]}$

- This is the supplementary information: *conditionned by F* (in bits)
- This is the Renyi divergence for $\alpha = 1, D_\alpha[p||q] = \frac{1}{1-\alpha}\int_{-\infty}^{\infty} p(x)^\alpha q(x)^{1-\alpha}\, dx$

Hellinger ↓

The topology generated by D_{KL} and D_α are not the same : $D_{1/2} \rightarrow Distance$

$\lim_{\theta_1 \to \theta} \frac{1}{(\theta_1-\theta)^2} D_\alpha\left[p_{\theta_1}||p_\theta\right] = \frac{\alpha}{2} J(\theta)$ <---------- Fisher Information

To Summary

KL-Divergence and Crossed Entropy..

Choosing
The right tool

Claude Shannon

$$D_{KL}[p||q]=\int p(x)Logp(x)dx-\int p(x)Logq(x)dx$$

Measure a spread

$$D_{KL}[p||q] = \mathrm{H}(p) - H(p,q)$$

Entropy

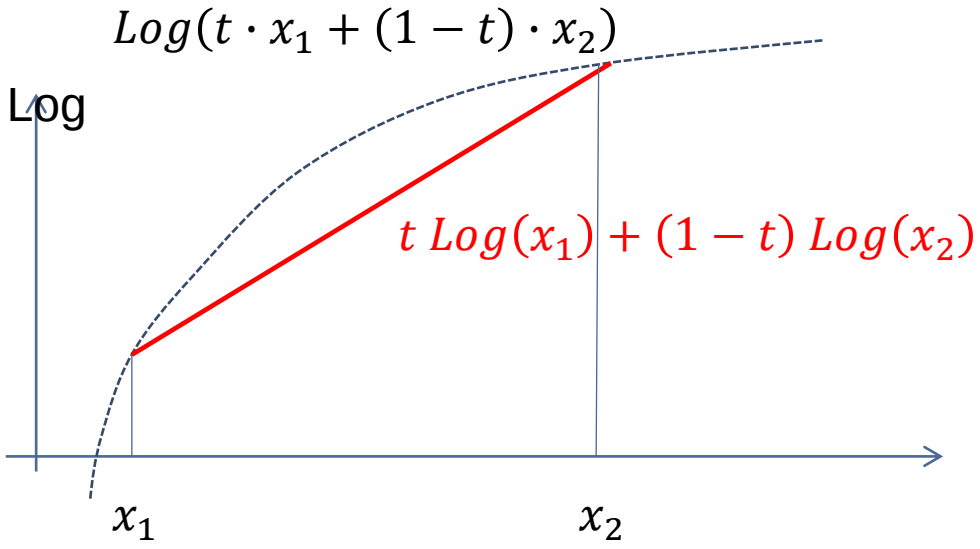
Crossed Entropy

- An Efficient criteria
- Classification problems
- Far Gradient still active

To Summary

Jensen Inequality

2 points



N points >2

Jensen :

F concave => $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$

To Summary

Mutual Information Properties

Definition

$$I(X;Y) = D_{KL}[p(x,y)||p(x)p(y)]$$

The information improvment compared to an independant case

Data processing Inequality (DPI)

For any markov chains $X \rightarrow Y \rightarrow Z$

$$I(X;Y) \geq I(X;Z)$$

There is less information on X in Z than in Y

Invariance by reparametrisation

$I(X;Y) = I(\varphi(X);\psi(Y))$ For any invertible φ and ψ

To Summary

7) BENAMOU BRENIER THEORY

To Summary

Benamou Brenier Theory 1

Optimal Transport - Benamou-Brenier Formulation (Brittany Hamfeldt)
https://www.youtube.com/watch?v=xxwalfOJ4sg&list=PLJ6garKOlK2qKVhRm6UwvcQ46wK-ciHbl&index=24

$$\rho(0,x) = f(x) \qquad \rho(1,x) = g(x)$$

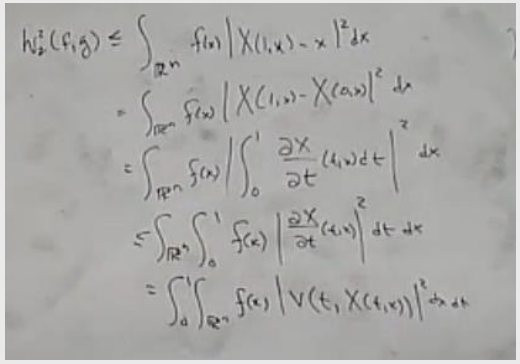
Continuity equation $\rho_t + \nabla(V\rho) = 0$

Lagrangian Coordinantes : $X(t,x)$:

$$\frac{\partial X}{\partial t}(t,x) = V(t, X(t,x)) \qquad X(0,x) = X$$

Push forward : $X(t,.) \# f = \rho(t,.)$ $X(1,.) \# f = g$

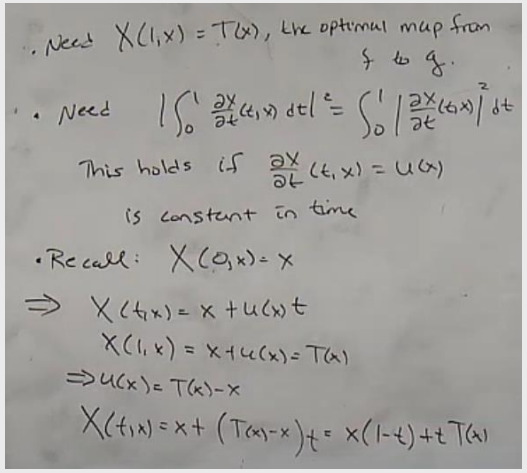
$W_2^2(f,g)$ Verifies

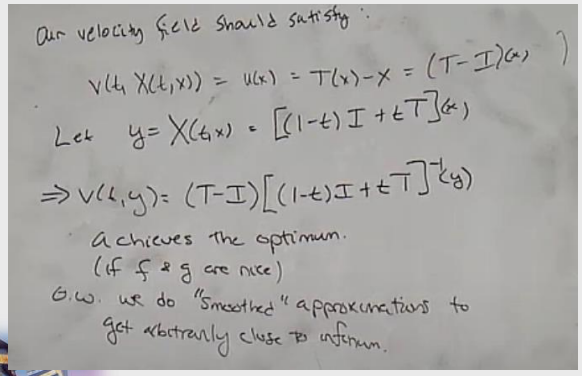


NATIXIS BEYOND BANKING

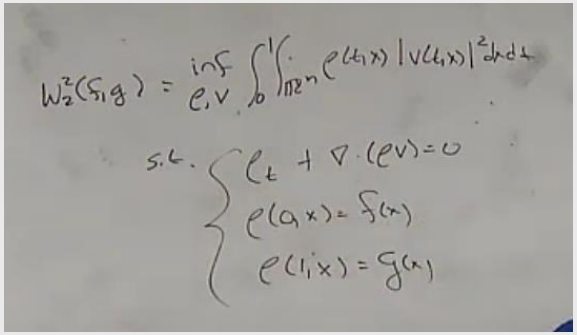
To Summary

Benamou Brenier Theory 2



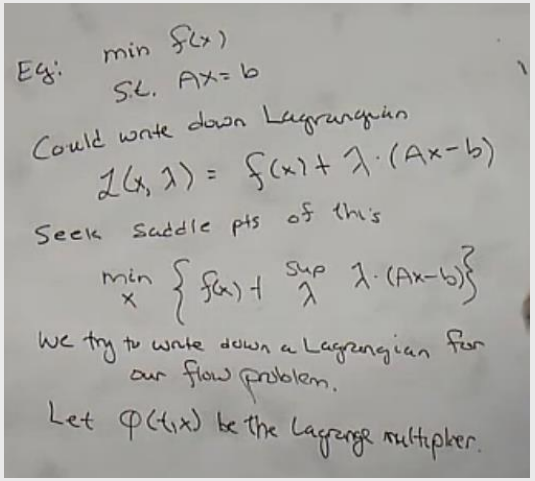


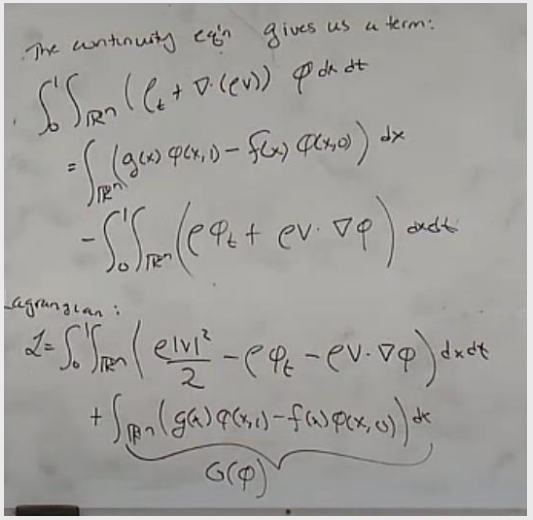
Benamou Brenier formulation

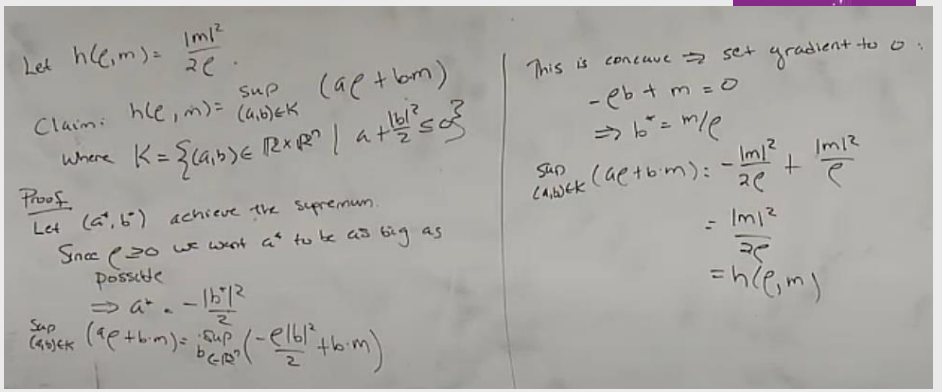


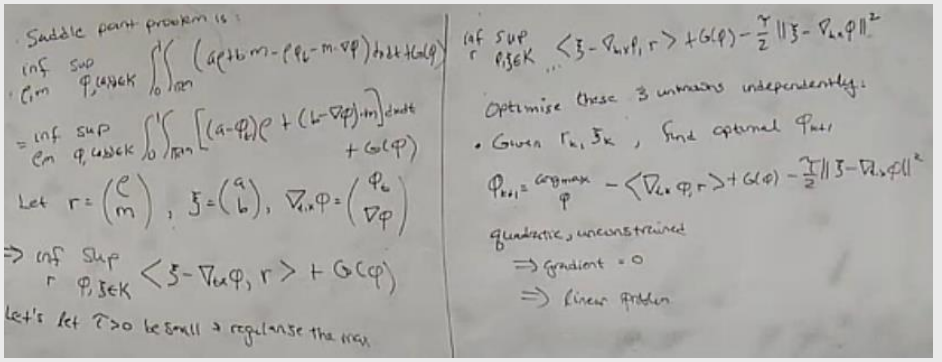
To Summary

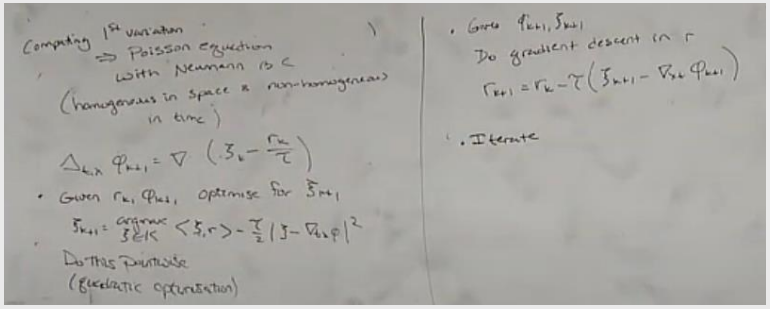
Benamou Brenier Theory 3











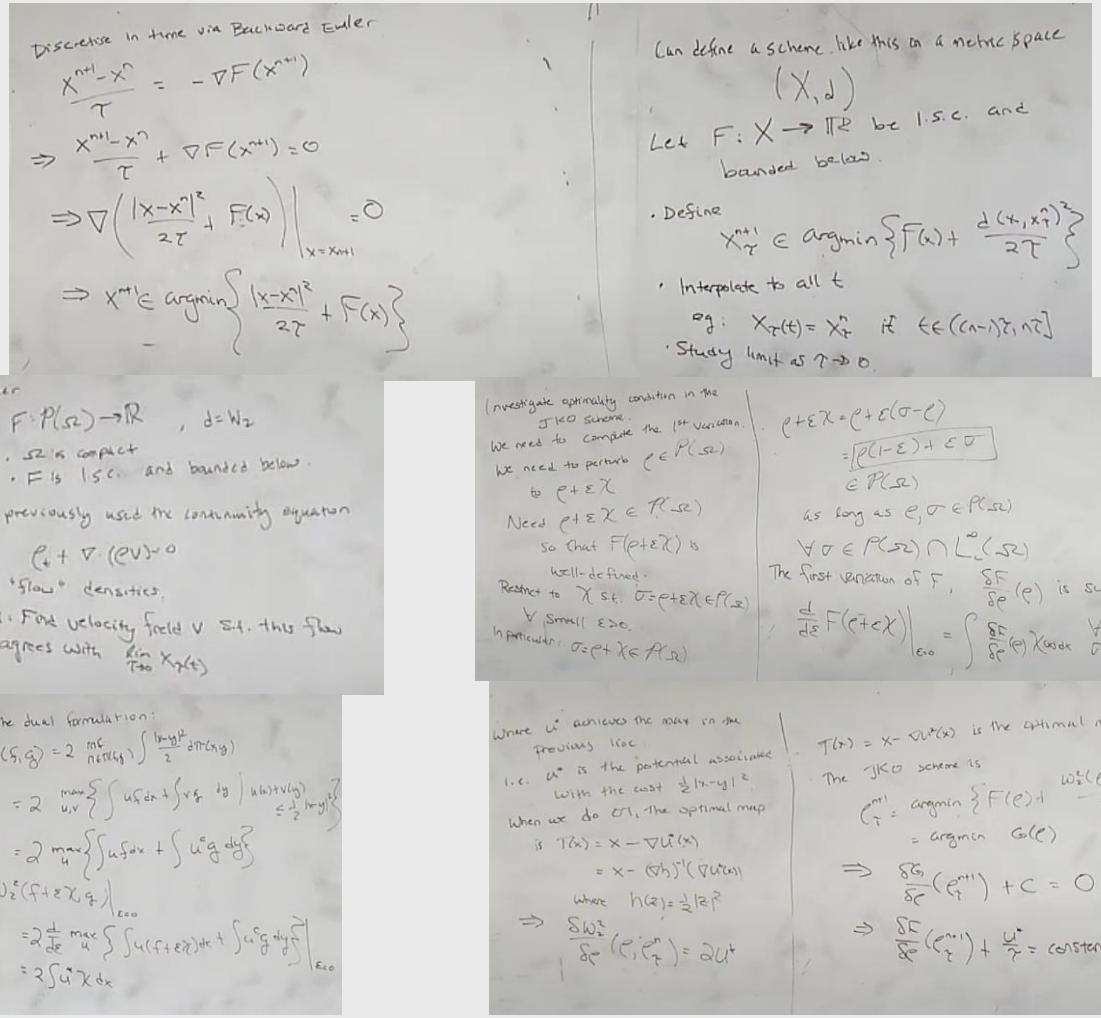
To Summary

8) GRADIENT FLOWS: THE DEMONSTRATIONS

To Summary

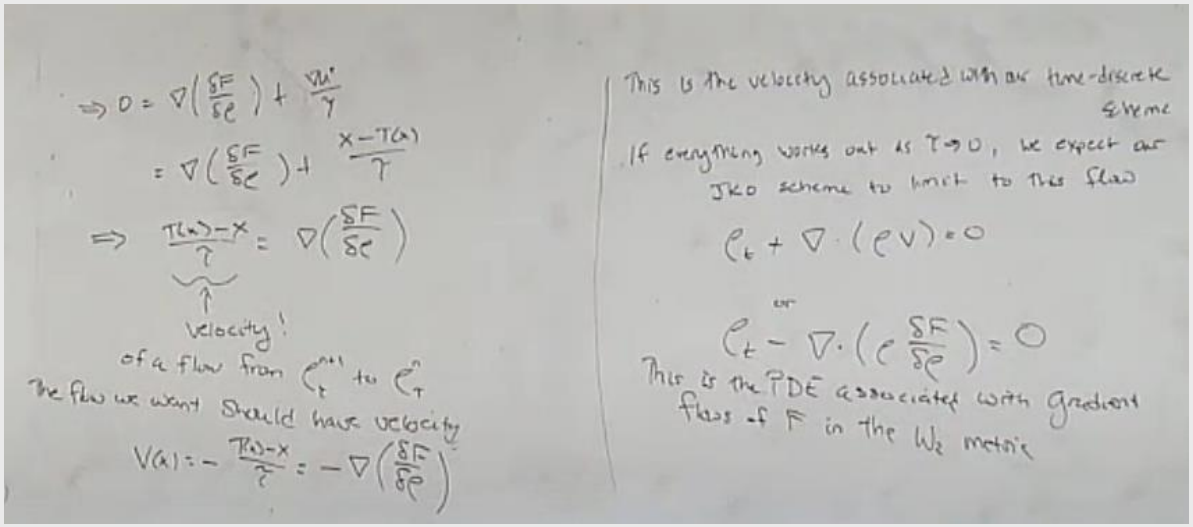
Gradient Flows 1

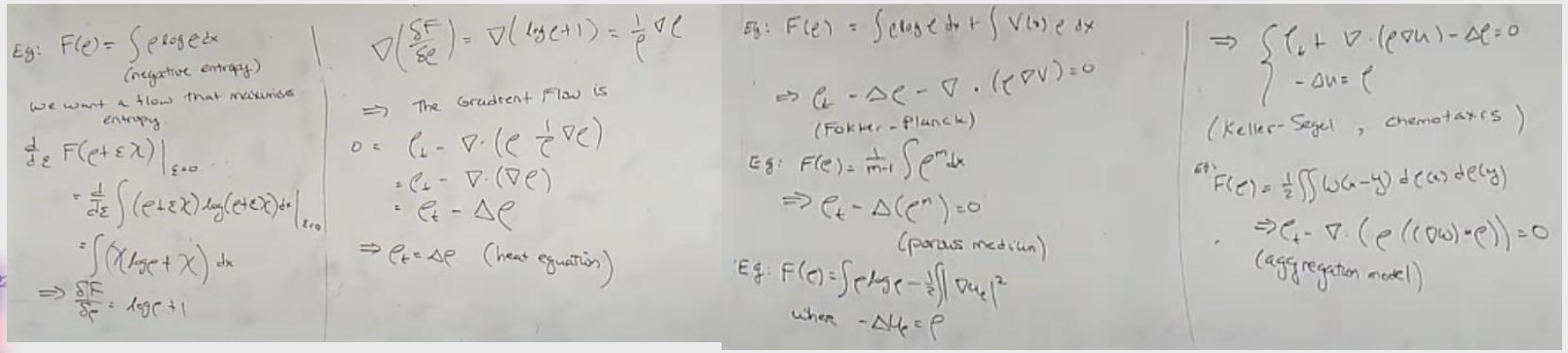
https://www.youtube.com/watch?v=zzGBxAqJV0Q&list=PLJ6garKOlK2qKVhRm6UwvcQ46wK-ciHbl&index=25



To Summary

Gradient Flows 2





To Summary

9) GRADIENT FLOWS THEORY

To Summary

Diffusion Equation via Optimal Transport

Jordan –Kinderlehrer-Otto (98)

- the 2-Kantorovich metric on the space of probability measures

$$W_2(\mu,\nu) = \inf_{\gamma \in \mathrm{Cpl}(\mu,\nu)} \sqrt{\int_{\mathbf{R}^n \times \mathbf{R}^n} |x-y|^2 \, \mathrm{d}\gamma(x,y)}$$

- the (negative of the) Boltzmann-Shannon entropy

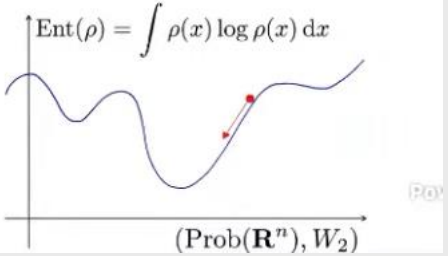
$$\mathrm{Ent}(\mu) = \int_{\mathbf{R}^n} \rho(x) \log \rho(x) \, \mathrm{d}x, \qquad \text{if} \quad \mathrm{d}\mu(x) = \rho(x)\,\mathrm{d}x$$

- the heat equation

$$\partial_t \mu = \Delta \mu$$

Theorem (J-K-O '98)

The heat flow is the gradient flow of the entropy w.r.t W_2.



$$\partial_t \mu = \Delta \mu \iff \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} W_2(\mu_t,\nu)^2 \le \mathrm{Ent}(\nu) - \mathrm{Ent}(\mu_t) \qquad \forall \nu$$

To Summary

Gradient Flow Definition

Gradient flows in $\mathbf{R}^n$

Let $\varphi : \mathbf{R}^n \to \mathbf{R}$ smooth and convex. For $u : \mathbf{R}_+ \to \mathbf{R}^n$ TFAE:

1. u solves the gradient flow equation $u'(t) = -\nabla\varphi(u(t))$.
2. u satisfies the evolution variational inequality

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u(t) - y|^2 = (u(t) - y) \cdot u'(t) \leq \varphi(y) - \varphi(u(t)) \qquad \forall y$$

(De Giorgi '93, Ambrosio–Gigli–Savaré '05)

To Summary

Brenier's Theorem

Def (pushforward) : Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T : \mathbb{R}^d \to \mathbb{R}^d$. The pushforward measure $T_{\#}\mu$ is characterized by:

- $\forall$ B meas. set, $T_{\#}\mu(B) = \mu(T^{-1}(B))$
- $x \sim \mu$, $T(x) \sim T_{\#}\mu$

Brenier's theorem : Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\mu \ll Leb$. Then,

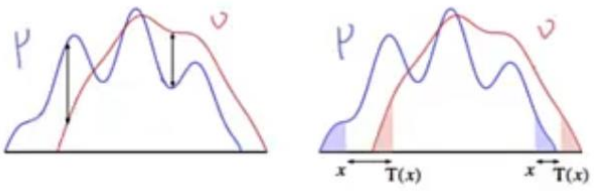
- Then $\exists!$ $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $T_{\mu\#}^\nu \mu = \nu$, and a convex function g s.t. $T_\mu^\nu = \nabla g$ μ-a.e.
- $W_2^2(\mu, \nu) = \|I - T_\mu^\nu\|_{L_2(\mu)}^2 = \inf_{T \in L_2(\mu)} \int (x - T(x))^2 d\mu(x)$ where $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int \|f(x)\|^2 d\mu(x) < \infty\}$

W_2 geodesics?

$\rho(0) = \mu, \rho(1) = \nu.$

$\rho(t) = ((1-t)I + tT_\mu^\nu)_{\#}\mu$

$\neq \rho(t) = \underbrace{(1-t)\mu + t\nu}_{\text{mixture}}$



To Summary

Continuity Equation

Let $T > 0$. Consider a family $\mu : [0, T] \to \mathcal{P}_2(\mathbb{R}^d), t \mapsto \mu_t$. It satisfies a continuity equation if there exists $(V_t)_{t\in[0,T]}$ such that $V_t \in L^2(\mu_t)$ and distributionnally:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0$$

rules density μ_t of particles $x_t \in \mathbb{R}^d$ driven by a vector field V_t:

$$\frac{dx_t}{dt} = V_t(x_t)$$

Riemannian interpretation [Otto, 2001] :
The tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at μ_t verifies $\mathcal{T}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu_t)$.

To Summary

Wasserstein Gradient Flow

Wasserstein gradient $\nabla_W F(\mu)$ **of functional** $F(\mu)$

$F(\mu)$	$\nabla_W F(\mu)$
$\int V \mathrm{d}\mu$	$\nabla V(\cdot)$
$\mathrm{KL}(\mu\|\nu)$	$\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\cdot)$
$\chi^2(\mu\|\nu)$	$2\nabla \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\cdot)$

Interpret as

$$\dot{X}_t = -\nabla_W F(\mu_t)(X_t)$$

Gradient flow: $\dot{\mu}_t = -\nabla_W F(\mu_t)$

https://www.youtube.com/watch?v=E0aK81IH6tM

To Summary

Wasserstein Gradient Flows(2)

Let $\mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}, \mu \mapsto \mathcal{G}(\mu)$ a regular functional. The differential of $\mathcal{G}$ evaluated at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is the unique function $\frac{\partial \mathcal{G}(\mu)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$, s.t. $\mu' - \mu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\mathcal{G}(\mu + \epsilon(\mu' - \mu)) - \mathcal{G}(\mu)\right] = \int_{\mathbb{R}^d} \frac{\partial \mathcal{G}(\mu)}{\partial \mu}(x) \, (d\mu' - d\mu)(x)$$

Then $\mu : [0, T] \to \mathcal{P}_2(\mathbb{R}^d), t \mapsto \mu_t$ satisfies a Wasserstein gradient flow of $\mathcal{G}$ if distributionally:

$$\frac{\partial \mu_t}{\partial t} - div\left(\mu_t \nabla \frac{\partial \mathcal{G}(\mu_t)}{\partial \mu_t}\right) = 0, \text{ i.e. } V_t = -\nabla_W \mathcal{G}(\mu)$$

where $\nabla_W \mathcal{G}(\mu) := \nabla \frac{\partial \mathcal{G}(\mu)}{\partial \mu} \in L^2(\mu)$ is called the Wasserstein gradient of $\mathcal{G}$.

To Summary

Wasserstein Gradient Flows of Free Energies

In particular, if the functional $\mathcal{G}$ is a free energy:

$$\mathcal{G}(\mu) = \underbrace{\int H(\mu(x))dx}_{\text{internal energy } \mathcal{H}(\mu)} + \underbrace{\int V(x)d\mu(x)}_{\text{potential energy } \mathcal{E}_V(\mu)} + \underbrace{\int W(x,y)d\mu(x)d\mu(y)}_{\text{interaction energy } \mathcal{W}(\mu)}$$

$$\text{Then : } \frac{\partial \mu_t}{\partial t} = div(\mu_t \underbrace{\nabla(H'(\mu_t) + V + W * \mu_t)}_{\nabla_W \mathcal{G}(\mu)}). \qquad (1)$$

For instance, if $\mathcal{H}(\mu)$ is the negative entropy ($H(s) = s\log(s)$), then (1) rules the density μ_t of particles $x_t \in \mathbb{R}^d$ driven by :

$$\frac{dx_t}{dt} = -\nabla V(x_t) - \int_{\mathbb{R}^d} \nabla W(x, x_t)d\mu_t(x) + \sqrt{2}dB_t,$$

$\mu_t = Law(x_t)$, B_t is a Brownian motion.

To Summary

Wasserstein Gradient Flows Space Discretization

If the vector field depends on the density of the particles at time t, replace μ_t by the empirical measure of a system of N interacting particles:

$$X_0^1, \ldots, X_0^N \sim \mu_0$$

and for $j = 1, \ldots, N$:

$$\frac{d\hat{x}_t^j}{dt} = -\nabla V(\hat{x}_t^j) - \frac{1}{N}\sum_{i=1}^{N} \nabla W(\hat{x}_t^i, \hat{x}_t^j) + \sqrt{2}dB_t.$$

To Summary

Wasserstein Gradient Flows Time Discretization

1. Forward :

$$\mu_{n+1} = exp_{\mu_n}(-\gamma \nabla_W \mathcal{G}(\mu_n)) = (I - \gamma \nabla_W \mathcal{G}(\mu_n))_{\#}\mu_n$$

where $exp_\mu : L^2(\mu) \to \mathcal{P}, \phi \mapsto (I + \phi)_{\#}\mu$,

and which corresponds in $\mathbb{R}^d$ to:

$$X_{n+1} = X_n - \gamma \nabla_W \mathcal{G}(\mu_n)(X_n) \sim \mu_{n+1}, \text{ if } X_n \sim \mu_n.$$

2. Backward :

$$\mu_{n+1} = JKO_{\gamma\mathcal{G}}(\mu_n)$$

$$\text{where } JKO_{\gamma\mathcal{G}}(\mu_n) = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\text{argmin}} \left\{ \mathcal{G}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \mu_n) \right\}.$$

3. Splitting schemes : if $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2$, e.g. Forward/Backward:

$$\nu_{n+1} = (I - \gamma \nabla_W \mathcal{G}_1)_{\#}\mu_n^1$$
$$\mu_{n+1} = JKO_{\gamma\mathcal{G}_2}(\nu_{n+1})$$

To Summary

Relative Entropy/KL Divergence

For any $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, the Kullback-Leibler divergence of μ w.r.t. π is defined by

$$\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi$$

and is $+\infty$ otherwise.

We consider the functional $\mathrm{KL}(\cdot|\pi) : \mathcal{P}_2(\mathbb{R}^d) \to [0, +\infty]$.

For any $\mu \in \mathcal{P}_2(\mathbb{R}^d), \mu \ll \pi$, the differential of $\mathrm{KL}(\cdot|\pi)$ evaluated at μ, $\frac{\partial \mathrm{KL}(\mu|\pi)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ is the function

$$\log\left(\frac{\mu}{\pi}\right)(.) + 1 : \mathbb{R}^d \to \mathbb{R}.$$

Hence, for μ regular enough, $\nabla_W \mathrm{KL}(\cdot|\pi)$ is:

$$\nabla \log\left(\frac{\mu}{\pi}\right)(.) : \mathbb{R}^d \to \mathbb{R}.$$

To Summary

Wasserstein Bayesian Statistics

- Let $\mathcal{D} = (w_i, y_i)_{i=1,\dots,N}$ observed data.
- Assume an underlying model parametrized by $\theta \in \mathbb{R}^d$ (e.g. $p(y|w,\theta)$ gaussian)

  $\Longrightarrow$ Likelihood: $p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(y_i|\theta, w_i)$.
- The parameter $\theta \sim p$ the prior distribution.

Bayes' rule : $\pi(\theta) := p(\theta|\mathcal{D}) = \dfrac{p(\mathcal{D}|\theta)p(\theta)}{Z}$, $Z = \displaystyle\int_{\mathbb{R}^d} p(\mathcal{D}|\theta)p(\theta)d\theta$.

π **is known up to a constant** since Z is untractable.
How to sample from π then?

1. **MCMC methods** (Langevin Monte Carlo [Roberts and Tweedie, 1996], Hamiltonian Monte Carlo [Neal et al., 2011]...)
2. **Sampling as optimization of the KL** [Wibisono, 2018]

$$\pi = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \ \mathrm{KL}(\mu|\pi)$$

15/

To Summary

EM Algorithm

$$\max_{\phi,\theta} \mathcal{F}(y, q_\phi) = \mathbb{E}_{q_\phi(z)} \left[\log p_\theta(y|z)\right] - KL\left[q_\phi(z)||p(z)\right]$$

Alternative optimization for the variational parameters and then model parameters (VEM).

Repeat:

E-Step: (inference)

For i=1,…,N

$$\phi_n \propto \nabla_\phi \mathbb{E}_{q_\phi(z)}[\log p_\theta(y_n|z_n)] - \nabla_\phi KL[q(z_n)||p(z_n)]$$

M-Step: (Parameter Learning)

$$\theta \propto \frac{1}{N} \sum_n \mathbb{E}_{q_\phi(z)} \left[\nabla_\theta \log p_\theta(y_n|z_n)\right]$$

To Summary

Overlap Matrix vs Gram Matrix

The invariant hamiltonian with respect to orthogonal transformation makes the distances also invariant that implies the dynamic entirely determined by the overlap matrix.

	Overlap matrix	Gram matrix
Statistical set	Disorder	Batch
Time of Computation	End of Convergence	Every N Iterations
Use	Compute Expectations	Estimate Information
Size of the statistical Set	Large	Moderate
perspective	Bayesian	Frequentist

$$|G(\{v_1,\ldots,v_n\})| = \begin{vmatrix} \langle v_1,v_1\rangle & \langle v_1,v_2\rangle & \ldots & \langle v_1,v_n\rangle \\ \langle v_2,v_1\rangle & \langle v_2,v_2\rangle & \ldots & \langle v_2,v_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_n,v_1\rangle & \langle v_n,v_2\rangle & \ldots & \langle v_n,v_n\rangle \end{vmatrix}$$

To Summary

12) POSSIBLE RELATIONSHIP WITH QUANTUM STATISTICAL PHYSICS

To Summary

Gram Matrix and Quantum Statistical Physics

Density Matrix $\rho = \sum_i \alpha_i |\psi_i >< \psi_j|$ where $|\psi_i > = \sum_j \beta_{i,j} |\phi_j >$

Statistical Mix

Intrication

Batch Average

Neural Codependance

Gram Matrix $g = \sum_{i,j} \gamma_{i,j} \langle \bar{\varphi}(x_i) | \bar{\varphi}(x_j) \rangle = \sum_{i,j} \gamma_{i,j} |\bar{\varphi}(x_j) >< \bar{\varphi}(x_i)|$

where $|\bar{\varphi}(x_j) > = \frac{1}{n} \sum_{k=1}^{n} \varphi(x_{k,j})$

kernel

To Summary

Comparaison of Gram Matrix and Density Operator

Both are two dimensional tensors

Density Operator is used to encode linearly the **statistic mixing** and the **intrication property** of quantum object.

The kernel theory for features space allow to encode the **statistical behaviour** under a batch of input and the **co-dependance** of all the neurons in the layer.

To Summary

Hamiltonian associated with the Wasserstein Gradient Flow

$$\frac{dG}{dt} = -i\hbar[H, G]$$

Where G is the density operator

The heat equation can be seen as the gradient flow of the entropy functional with respect to the Wasserstein distance

$$\frac{d\rho}{dW} = \Delta\rho$$

Wigner equation

$$\frac{\partial f_w(x,p)}{\partial t} = -\frac{p \cdot \nabla_x}{m} f_w(x,p) + \int_{-\infty}^{+\infty} dq\, f_w(x, p+q) V_w(x,p)$$

Time independant:

$$H(x,p) * f_w(x,p) = E \cdot f_w(x,p)$$

To Summary

Area Law of Mutual Information

Holographic Principle

Fundamental level:
Information content of a region should depend on its surface area rather than on its volume
except for critical systems

Mutual Information $I(A:B) = H(\rho_A) + H(\rho_B) - H(\rho_{AB})$

Entropy $H(\rho) = -Tr(\rho\, Ln(\rho))$

Motivations
(i) Coincides with the entanglement entropy at zero temperature
(ii) Total amount of information of one system about another without 'overlooking' hidden correlations
(iii) Area Law can be rigourously proven at any finite temperature

Area Law $I(A:B) = I(\partial A:\partial B) \leq H(\partial A) \leq |\partial A| \log d$

Heuristic Explanation
Existence of a characteristic length scale, the correlation length, on which two-point correlations decay

To Summary

Verification of the Area law for a quantum system

nature physics

Article

https://doi.org/10.1038/s41567-023-02027-1

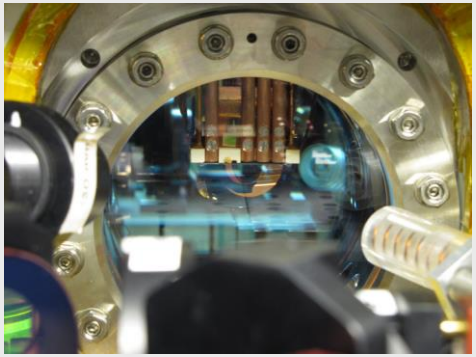
Verification of the area law of mutual information in a quantum field simulator

Received: 11 July 2022

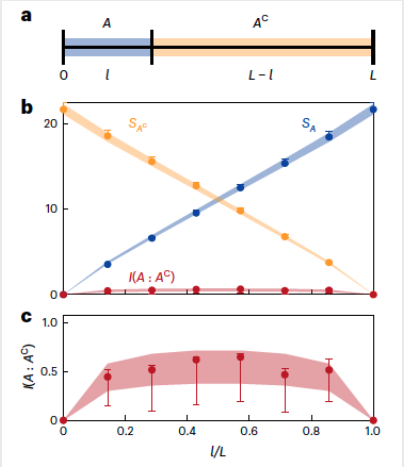
Accepted: 20 March 2023

Published online: 24 April 2023

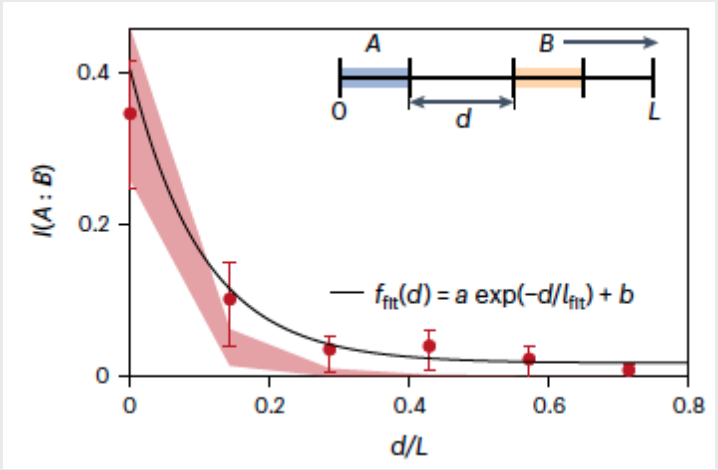
Mohammadamin Tajik[1], Ivan Kukuljan[2,3], Spyros Sotiriadis[4], Bernhard Rauer[1,5], Thomas Schweigler[1], Federica Cataldini[1], João Sabino[1,6], Frederik Møller[1], Philipp Schüttelkopf[1], Si-Cong Ji[1], Dries Sels[7,8], Eugene Demler[9] & Jörg Schmiedmayer[1]



Vacuum chamber containing the atom chip



Area law of MI and volume law of vN entropy



Shared information content between two spatially separated subsystems

To Summary

Machine learning

Mean-Field Theories

Replicas Theories

Integra -Differential Equation

memory effect
Laplace and Fourier transforms,
Characteristics,
Variational Methods

Learning

$$\delta W_i = -\eta * \frac{\delta F}{\delta W_i}$$

Foliated Spaces

Non Commutative Differential Theories (Alain Connes)

Statistical Physics

Wigner Weyl Quantization
Hamiltonian Formulation
Von Neuman Entropy

Wasserstein Flows

Continuity Equation
Free Energy

To Summary

Getting the Hamiltonian

Layer by layer approach:
- Compute Mutual information I(input, l) and I(l, output) → bottleneck
- Every neuron is exchangeable with another neuron of same layer

Mean Field Approach: representing the effect of learning iteration :
- Every layer backward propagation (Average activation)
- Integra-differential evolution equation of the Average activations

Fokker-Planck equation : Continuity equation for the density weights and biases
- Layer by layer
- Embedding of the weights in a kernel

Wigner-Weyl Quantization
- Doubling the number of weights
- First order dynamic equation

$$Learning.: \quad \frac{dG}{dt} = -i\hbar[H, G]$$

To Summary

Wigner-Weyl transform

$$\hat{A}(\hat{q},\hat{p}) \mapsto V_{\mathrm{W}}(\hat{A}) = A(x,p) = \\ = \frac{\hbar}{2\pi}\int \mathrm{d}\xi \int \mathrm{d}\eta\, \mathrm{Tr}\left\{\hat{A}(\hat{q},\hat{p})\mathrm{e}^{\mathrm{i}\xi\hat{q}+\mathrm{i}\eta\hat{p}}\right\}\mathrm{e}^{-\mathrm{i}\xi x-\mathrm{i}\eta p}$$

$$\begin{cases} A * B & \equiv V_{\mathrm{W}}\left(\hat{A}\cdot\hat{B}\right) \\ [A,B]_M & \equiv \frac{1}{\mathrm{i}\hbar}(A*B - B*A) \end{cases}$$

N.C. Dias, J.N. Prata / Annals of Physics 313 (2004) 110–146

To Summary

15) MULTIVARIATE RENYI INFORMATION

To Summary

Multivariate Renyi Entropy

$$\mathbf{S}_\alpha(A) = \frac{1}{1-\alpha}\log_2\left(\mathrm{tr}(A^\alpha)\right) = \frac{1}{1-\alpha}\log_2\left[\sum_{i=1}^{n}\lambda_i(A)^\alpha\right]$$

where $A_{ij} = \frac{1}{n}\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$ and $\lambda_i(A)$ denotes the i-th eigenvalue of A

To Summary

Bisupport Renyi Mutual Information

$$\kappa_1 : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \text{ and } \kappa_2 : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}.$$
$$A_{ij} = \kappa_1(x_i, x_j),\ B_{ij} = \kappa_2(x_i, x_j)$$

$$\mathbf{S}_\alpha(A, B) = \mathbf{S}_\alpha\left(\frac{A \circ B}{\mathrm{tr}(A \circ B)}\right)$$

$A \circ B$ denotes the Hadamard product

$$\mathbf{S}_\alpha\left(\tfrac{A\circ B}{\mathrm{tr}(A\circ B)}\right) \leq \mathbf{S}_\alpha(A) + \mathbf{S}_\alpha(B),$$
$$\mathbf{S}_\alpha\left(\tfrac{A\circ B}{\mathrm{tr}(A\circ B)}\right) \geq \max[\mathbf{S}_\alpha(A), \mathbf{S}_\alpha(B)]$$

$$\mathbf{S}_\alpha(A|B) = \mathbf{S}_\alpha(A, B) - \mathbf{S}_\alpha(B),$$

$$\mathbf{I}_\alpha(A; B) = \mathbf{S}_\alpha(A) + \mathbf{S}_\alpha(B) - \mathbf{S}_\alpha(A, B)$$

To Summary

Multisupport Renyi Mutual Information

$$\mathbf{S}_\alpha(A_1, A_2, \cdots, A_k) = \mathbf{S}_\alpha\left(\frac{A_1 \circ A_2 \circ \cdots \circ A_k}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_k)}\right)$$

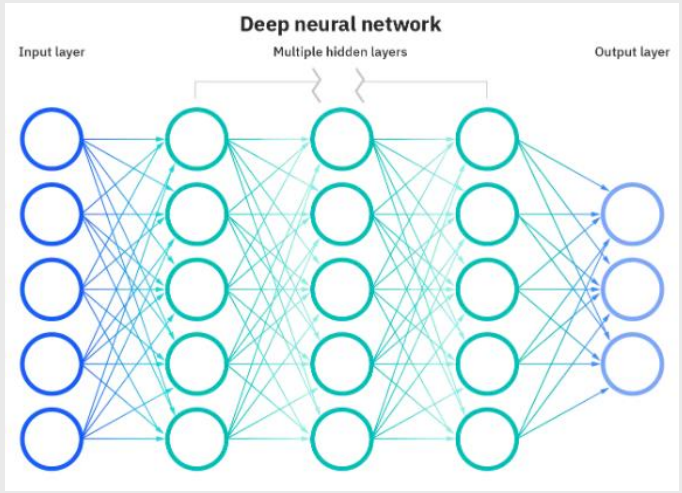
$$\mathbf{S}_\alpha\left(\frac{A_1 \circ A_2 \circ \cdots \circ A_k}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_k)}\right) \leq \mathbf{S}_\alpha(A_1) + \mathbf{S}_\alpha(A_2) + \cdots + \mathbf{S}_\alpha(A_k)$$

$$\mathbf{S}_\alpha\left(\frac{A_1 \circ A_2 \circ \cdots \circ A_k}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_k)}\right) \geq \max[\mathbf{S}_\alpha(A_1), \mathbf{S}_\alpha(A_2), \cdots, \mathbf{S}_\alpha(A_k)]$$

$$\mathbf{I}_\alpha(B; \{A_1, A_2, \cdots, A_k\}) = \mathbf{S}_\alpha(B) + \mathbf{S}_\alpha\left(\frac{A_1 \circ A_2 \circ \cdots \circ A_k}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_k)}\right) - \mathbf{S}_\alpha\left(\frac{A_1 \circ A_2 \circ \cdots \circ A_k \circ B}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_k \circ B)}\right)$$

To Summary

Computation of Mutual Information



Input : x_i

output layer Y_k

Intermediate layer $T_{l,j}$

i : sample of the training set

j : single neuron of the l-the layer

k: final label probability

Mutual information I(x_i, $T_{l,j}$)

Mutual information I($T_{l,j}$, Y_k)

To Summary

16) INFORMATION THEORY

To Summary

Kullback-Leibler Divergence

$$D_{KL}[p||q] = \int_{-\infty}^{\infty} p(x) Log \frac{p(x)}{q(x)} dx$$

$D_{KL}[p||p] \geq 0$ et $D_{KL}(p||p) = 0$

But it is not a distance but triangular inégality does not work

Is not symmetric

let $q(x)$ a prob. distribution defined on E et $\mathrm{F} \subset E$

Let $q_F(x)$ a prob. distribution restricted to F (conditionned by F)

Then $D_{KL}[q||q_F)] = \mathrm{Log}\ \frac{1}{Prob[F]}$

- This is the supplementary information: *conditionned by F* (in bits)
- This is the Renyi divergence for $\alpha = 1, D_\alpha[p||q] = \frac{1}{1-\alpha}\int_{-\infty}^{\infty} p(x)^\alpha q(x)^{1-\alpha}\, dx$

Hellinger ↓

The topology generated by D_{KL} and D_α are not the same : $D_{1/2} \rightarrow Distance$

$\lim_{\theta_1 \to \theta} \frac{1}{(\theta_1-\theta)^2} D_\alpha\left[p_{\theta_1}||p_\theta\right] = \frac{\alpha}{2} J(\theta)$ ⟵ Fisher Information

To Summary

KL-Divergence and Crossed Entropy..

Choosing
The right tool

Claude Shannon

$$D_{KL}[p||q] = \int p(x) Log p(x) dx - \int p(x) Log q(x) dx$$

Measures a spread

$$D_{KL}[p||q] = \mathrm{H}(p) - H(p,q)$$

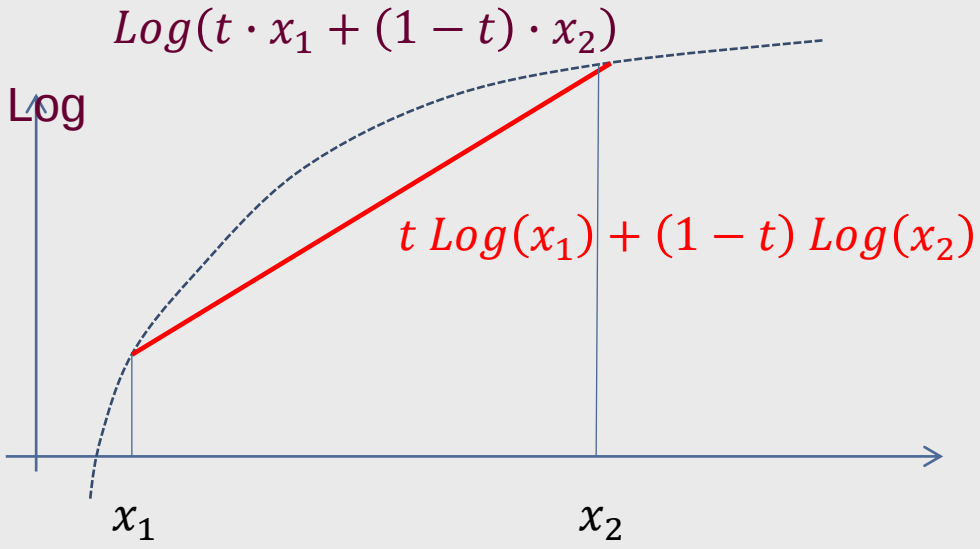
Entropy

Crossed Entropy

- An Efficient criteria
- Classification problems
- Far Gradient still active

To Summary

Jensen Inequality



2 points

N points >2

Jensen :

F concave => $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$

To Summary

Variationnal Inférence (mécanics)

$$Log\ p(x)\ - D_{KL}[q(z)||p(z|x)] = (\mathbb{E}_q[Log\ p(z,x)] - \mathbb{E}_q[Log\ q(z)])$$

But $p(z,x) = p(x|z)\ p(z)$

$$Log\ p(x)\ - D_{KL}[q(z)||p(z|x)] = (\mathbb{E}_q[Log\ p(x|z) + Log\ p(z)] - \mathbb{E}_q[Log\ q(z)])$$

But $D_{KL}[q(z)||p(z)] = \mathbb{E}_q[Log\ p(z)] - \mathbb{E}_q[Log\ q(z)]$

$$Log\ p(x)\ - D_{KL}[q(z)||p(z|x)] = \mathbb{E}_q[Log\ p(x|z)] - D_{KL}[q(z)||p(z)]$$

Should depend only on x

<0 <0

$$Log\ p(x)\ - D_{KL}[q(z|x)||p(z|x)] = \mathbb{E}_q[Log\ p(x|z)] - D_{KL}[q(z|x)||p(z)]$$

Total Infos

Spurious Infos during the construction of z

Necessary Infos to decode

Necessary Infos to build z

To Summary

Variationnal Inference (Boundary)

X observed
Z hidden

Z is not observed

$$Log\, p(x) = Log\left[\int_z p(x,z)dz\right]$$

q(z) arbitrary

$$= Log\left[\int_z p(x,z)\frac{q(z)}{q(z)}dz\right]$$

$$= Log\left[\mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]\right]$$

Jensen

Entropy of z

$$\geq \mathbb{E}_q\left[Log[p(x,z)]\right] - \mathbb{E}_q\left[Log[q(z)]\right] \equiv \mathrm{ELBO(p,q)}$$

Then maximizing expectation of the lower bound gives a lower bound on the likelihood which is a good approximation

To Summary

Relationship with KL divergence

We want to approximate $p(z|x) = \frac{p(z,x)}{p(x)}$ by $q(z)$

Definition

$$D_{KL}[q(z)||p(z|x)] = \mathbb{E}_q\left[Log\ \frac{q(z)}{p(z|x)}\right]$$

$$= \mathbb{E}_q[Log\ q(z)] - \mathbb{E}_q[Log\ p(z|x)]$$

$$= \mathbb{E}_q[Log\ q(z)] - \mathbb{E}_q[Log\ p(z,x)] + Log\ p(x)$$

Does'nt depend on q

Equation K

$$= -(\mathbb{E}_q[Log\ p(z,x)] - \mathbb{E}_q[Log\ q(z)]) + Log\ p(x)$$

Only varying q(z)
Then Minimizing ELBO ⇔ Maximizing KL divergence

To Summary

Mutual Information Properties

Definition

$$I(X;Y) = D_{KL}[p(x,y)||p(x)p(y)]$$

The information improvement compared to an independent case

Data processing Inequality (DPI)

For any markov chains $X \rightarrow Y \rightarrow Z$

$$I(X;Y) \geq I(X;Z)$$

There is less information on X in Z than in Y

Invariance by reparametrisation

$I(X;Y) = I(\varphi(X);\psi(Y))$ For any invertible φ and ψ

To Summary

Relationship with KL divergence

We want to approximate $p(z|x) = \frac{p(z,x)}{p(x)}$ by $q(z)$

Definition

$$D_{KL}[q(z)||p(z|x)] = \mathbb{E}_q\left[Log\ \frac{q(z)}{p(z|x)}\right]$$

$$= \mathbb{E}_q[Log\ q(z)] - \mathbb{E}_q[Log\ p(z|x)]$$

$$= \mathbb{E}_q[Log\ q(z)] - \mathbb{E}_q[Log\ p(z,x)] + Log\ p(x)$$

Does'nt depend on q

Equation K

$$= -(\mathbb{E}_q[Log\ p(z,x)] - \mathbb{E}_q[Log\ q(z)]) + Log\ p(x)$$

Only varying q(z)
Then Minimizing ELBO ⇔ Maximizing KL divergence

To Summary

17) OPTIMAL TRANSPORT AND NLP LOSSES

To Summary

Optimal Transport

Ω a measurable space, $c : \Omega \times \Omega \to \mathbb{R}$.
μ, ν two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $T : \Omega \to \Omega$

$$\inf_{T_\sharp \mu = \nu} \int_\Omega c(x, T(x)) \mu(dx)$$

[Brenier'87] If $\Omega = \mathbb{R}^d$, $c = \| \cdot - \cdot \|^2$, μ, ν a.c., then $T = \nabla u$, u **convex.**

[Brenier'87]: For any u **convex**, ∇u is the OT Monge map between μ and $\nabla u_\sharp \mu$.

To Summary

OT links with PDEs

If $\Omega = \mathbb{R}^d$, $c = \|\cdot - \cdot\|^2$, μ, ν have densities p, q, then $T_\sharp \mu = \nu$ is equivalent to

$$p(x) = q(T(x))|\det J_T(x)|$$

Monge-Ampère: find convex f such that

$$|\nabla^2 f(x)| = \frac{p(x)}{q(\nabla f(x))}$$

To Summary

Kantorovich Approach

Def. Given μ, ν in $\mathcal{P}(\Omega)$; a cost function c on $\Omega \times \Omega$, the Kantorovich problem is

$$\inf_{P \in \Pi(\mu,\nu)} \iint c(x,y) P(dx,dy).$$

PRIMAL

$$W_p^p(\mu,\nu) = \sup_{\substack{\varphi \in L_1(\mu), \psi \in L_1(\nu) \\ \varphi(x)+\psi(y) \leq D^p(x,y)}} \int \varphi d\mu + \int \psi d\nu.$$

DUAL

To Summary

D-Concavity

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x D^p(x, y) - \varphi(x).$$

φ is D^p-concave if $\exists \phi : \varphi = \overline{\phi}$

φ is D^p-concave $\Rightarrow \overline{\overline{\varphi}} = \varphi$

Prop. If $c = D$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

$$W_1(\mu, \nu) = \sup_{\varphi \text{ 1-Lipschitz}} \int \varphi(d\mu - d\nu).$$

W1

To Summary

Divergences

Remark. If $\Omega = \mathbb{R}$, $c(x, y) = c(|x - y|)$, c convex, F_μ^{-1}, F_ν^{-1} quantile functions,

$$W(\mu, \nu) = \int_0^1 c(|F_\mu^{-1}(x) - F_\nu^{-1}(x)|)dx$$

$$T = F_\mu^{-1} \circ F_\nu$$

Sliced Wasserstein Distance [Rabin+'11]

$$SW(\mu, \nu) = \mathbb{E}_{\theta \sim \mathcal{S}^{d-1}} \left[\int_0^1 c(|F^{-1}_{\theta^T_\sharp \mu}(x) - F^{-1}_{\theta^T_\sharp \nu}(x)|)dx \right]$$

To Summary

Gaussian case

Remark. If $\Omega = \mathbb{R}^d$, $c(x,y) = \|x-y\|^2$, and $\mu = \mathcal{N}(\mathbf{m}_\mu, \mathbf{\Sigma}_\mu), \nu = \mathcal{N}(\mathbf{m}_\nu, \mathbf{\Sigma}_\nu)$ then

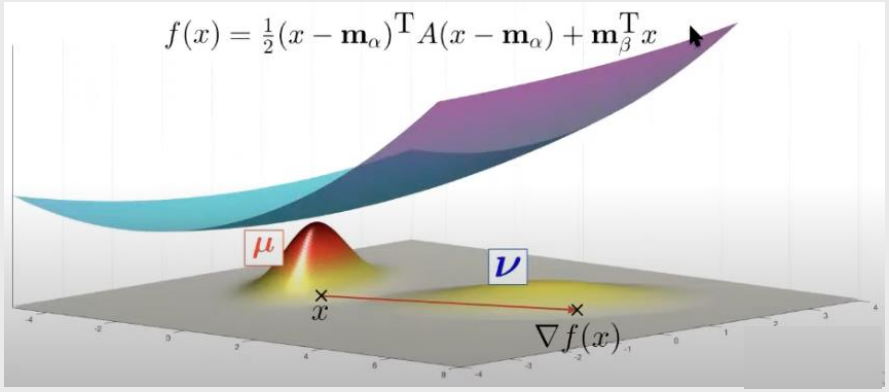
$$W_2^2(\mu,\nu) = \|\mathbf{m}_\mu - \mathbf{m}_\nu\|^2 + B(\mathbf{\Sigma}_\mu, \mathbf{\Sigma}_\nu)^2$$

where B is the Bures metric

$$B(\mathbf{\Sigma}_\mu, \mathbf{\Sigma}_\nu)^2 = \mathrm{trace}(\mathbf{\Sigma}_\mu + \mathbf{\Sigma}_\nu - 2(\mathbf{\Sigma}_\mu^{1/2}\mathbf{\Sigma}_\nu\mathbf{\Sigma}_\mu^{1/2})^{1/2})$$

The map $T : x \mapsto \mathbf{m}_\nu + A(x - \mathbf{m}_\mu)$ is **optimal**,

where $A = \mathbf{\Sigma}_\mu^{-\frac{1}{2}} \left(\mathbf{\Sigma}_\mu^{\frac{1}{2}} \mathbf{\Sigma}_\nu \mathbf{\Sigma}_\mu^{\frac{1}{2}} \right)^{\frac{1}{2}} \mathbf{\Sigma}_\mu^{-\frac{1}{2}}$.



To Summary

Entropic Regularization (Wilson 69)

Def. Regularized Wasserstein, $\gamma \geq 0$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$

$$E(P) \stackrel{\text{def}}{=} -\sum_{i,j=1}^{nm} P_{ij}(\log P_{ij} - 1)$$

Note: Unique optimal solution because of strong concavity of entropy

To Summary

Sinkhorn Algorithm

$$L(P,\alpha,\beta) = \sum_{ij} P_{ij}M_{ij} + \gamma P_{ij}(\log P_{ij} - 1) + \alpha^T(P\mathbf{1} - a) + \beta^T(P^T\mathbf{1} - b)$$

$$\partial L/\partial P_{ij} = M_{ij} + \gamma \log P_{ij} + \alpha_i + \beta_j$$

$$(\partial L/\partial P_{ij} = 0) \Rightarrow P_{ij} = e^{\frac{\alpha_i}{\gamma}}\, e^{-\frac{M_{ij}}{\gamma}}\, e^{\frac{\beta_j}{\gamma}} = u_i\, K_{ij} v_j$$

Prop. If $P_\gamma \stackrel{\text{def}}{=} \underset{P \in U(a,b)}{\operatorname{argmin}} \langle P, M_{XY} \rangle - \gamma E(P)$

then $\exists! u \in \mathbb{R}_+^n, v \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(u) K \mathbf{diag}(v), \quad K \stackrel{\text{def}}{=} e^{-M_{XY}/\gamma}$$

$$P_\gamma \in U(a,b) \Leftrightarrow \begin{cases} u = a/Kv \\ v = b/K^T u \end{cases}$$

To Summary

Sinkhorn Complexity

Sinkhorn's Algorithm : Repeat

1. $u = a / Kv$
2. $v = b / K^T u$

- [**Sinkhorn'64**] proved first convergence result [**Franklin+'89**] characterised linear convergence
- Recent wave of great results by [**Altschuler+'17**] [**Dvurechensky+18**][**Lin+19**]
- $O(nm)$ complexity, GPGPU parallel [**C'13**] .
- $O(n \log n)$ on gridded spaces using convolutions. [**Solomon'+15**]

To Summary

Divergences

Definition (φ-divergence)

Let φ convex l.s.c. function such that $\varphi(1) = 0$, the φ-divergence D_φ between two measures α and β is defined by :

$$D_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha(x)}{\mathrm{d}\beta(x)}\right)\mathrm{d}\beta(x).$$

Example (Kullback Leibler Divergence)

$$D_{KL}(\alpha|\beta) = \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x)\right)\mathrm{d}\alpha(x) \quad \leftrightarrow \quad \varphi(x) = x\log(x)$$

To Summary

The Wasserstein Distance

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$W_c(\alpha, \beta) = \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) \qquad (\mathcal{P})$$

For $c(x, y) = \|x - y\|_2^p$, $W_c(\alpha, \beta)^{1/p}$ is the **p-Wasserstein distance**.

To Summary

Wasserstein with entropic regularization

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$W_{c,\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon H(\pi|\alpha\otimes\beta), \quad (\mathcal{P}_\varepsilon)$$

where

$$H(\pi|\alpha\otimes\beta) \overset{\text{def.}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\alpha(x)\mathrm{d}\beta(y)}\right)\mathrm{d}\pi(x,y).$$

relative entropy of the transport plan π with respect to the product measure $\alpha\otimes\beta$.

To Summary

Sinkhorn Algorithm

Computation of the Optimal Distance as a machine learning layer: Differentiable

Sinkhorn's Algorithm

$$\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) \; ; \; T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\star)$$

Prop. One has $T = \mathrm{diag}(a) K \, \mathrm{diag}(b)$, where $K = e^{-\frac{d^p}{\varepsilon}}$.

Row constraint: $T \mathbb{1}_{N_2} = \mu \iff a \odot (Kb) = \mu$

Col. constraint: $T^\top \mathbb{1}_{N_2} = \nu \iff b \odot (K^\top a) = \nu$

Sinkhorn iterations: $a \leftarrow \frac{\mu}{Kb}$ and $b \leftarrow \frac{\nu}{K^\top a}$

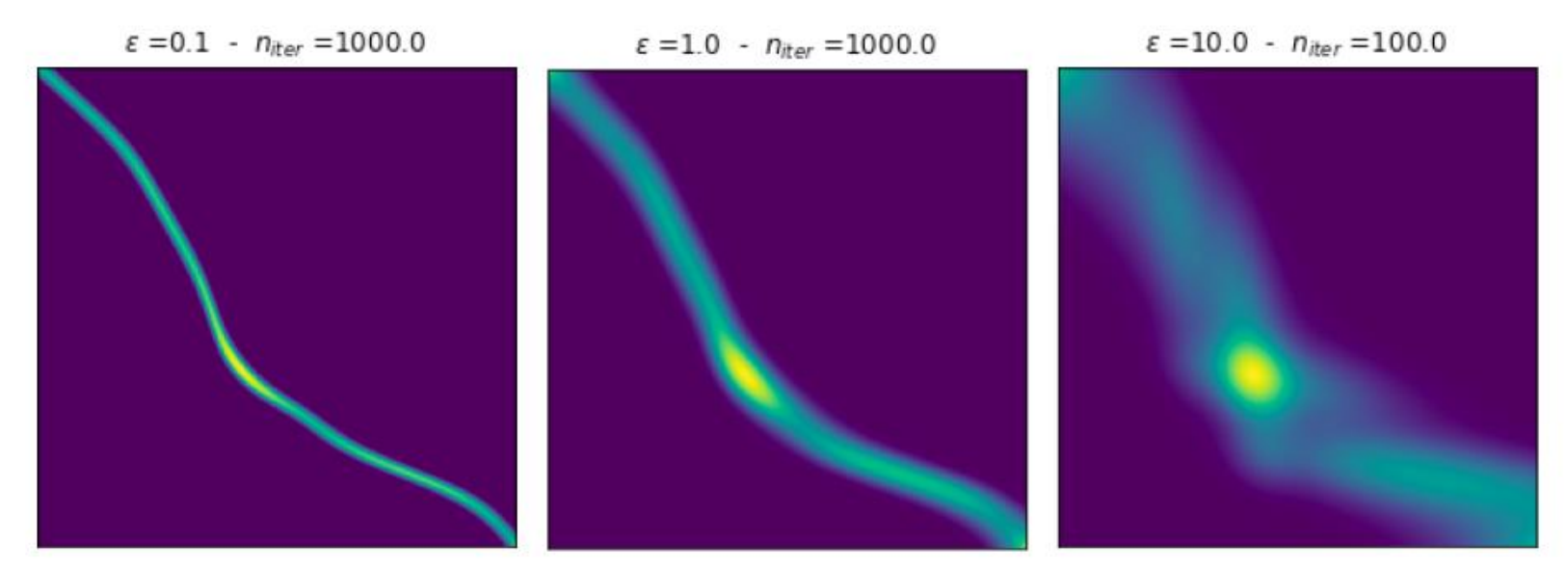
Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances

Marco Cuturi : Advances in Neural Information Processing Systems 26, pages 2292--2300, 2013 : arXiv:1306.0895v1

To Summary

Influence of the regularization

Influence of the regularization parameter on the transport plan .



To Summary

Sinkhorn Divergences

Issue of regularized Wass. Distance : $W_{c,\varepsilon}(\alpha, \alpha) \neq 0$
Proposed Solution : introduce corrective terms to 'debias' regularized Wasserstein distance

Definition (Sinkhorn Divergences)

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

$$SD_{c,\varepsilon}(\alpha, \beta) \stackrel{\text{def.}}{=} W_{c,\varepsilon}(\alpha, \beta) - \frac{1}{2} W_{c,\varepsilon}(\alpha, \alpha) - \frac{1}{2} W_{c,\varepsilon}(\beta, \beta),$$

To Summary

Conclusion

1) Artificial Intelligence will change every aspect of the society

2) NLP is the field where the progress in AI may lead to the building to a synthetic mind

3) The tools that are needed (numerical methods, mathematical concept, hardware) are there

To Summary

References

- Deepcast: universal Time-Series Forcaster, Nikolay Laptev, Jiafan Yu, Ram Rajagopal, ICLR 2018
- Building Transformer-Based Natural Language Processing Applications – Workshop – Deep Learning Institute Nvidia
- Improving Stock Market Prediction via Heterogeneous Information Fusion, Xi Zhanga, Yunjia Zhanga, Senzhang Wangb, Yuntao Yaoa, Binxing Fanga,Philip S. Yuc, arXiv:1801.00588v1 [cs.SI] 2 Jan 2018
- Financial Time Series Forecasting with Deep Learning : A Systematic Literature Review: 2005-2019, Omer Berat Sezera, M. Ugur Gudeleka, Ahmet Murat Ozbayoglua
- An Attempt to Understand Natural Language Processing And Illustration On A Financial Dataset, Charles-Albert Lehalle, Cornell-Citi Financial Data Science Webinar | February 2021
- Time Series Forecasting and Classification Models Based on Recurrent with Attention Mechanism and Generative Adversarial Networks, Kun Zhou , Wenyong Wang , Teng Hu and Kai Deng, December 2020, Sensors
- Transfert Learning for Time Series Forecasting and Classification, January 12, 2021 by Isaac Godfried , TOPBOTS
- Transfer learning for time series classification, Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar and Pierre-Alain Muller, arXiv:1811.01533v1 [cs.LG] 5 Nov 2018
- Visualizing and Measuring the Geometry of BERT, Andy Coenen, Emily Reif, Ann Yuan Been Kim, Adam Pearce, Fernanda Viégas, Martin Wattenberg, arXiv:1906.02715v2 [cs.LG] 28 Oct 2019