Université de Paris

École doctorale de Sciences Mathématiques de Paris Centre (ED 386)

Laboratoire : Mathématiques Appliqués à Paris 5, MAP5-UMR 8145 CNRS

THÈSE DE DOCTORAT

Spécialité : Mathématiques appliquées

par

Thi Thanh Yen NGUYEN

Mathematical, data-driven modeling of the dynamics of brain vulnerability and senescence in neurodegenerative diseases.

Date de soutenance : Jour Mois Année

Après avis des rapporteurs :	Prénom NOM (Institution)
	Prénom NOM (Institution)

Jury de soutenance :	Prénom NOM	(Institution)	Rapporteur
	Prénom NOM	(Institution)	Codirecteur de thèse
	Prénom NOM	(Institution)	Examinateur
	Prénom NOM	(Institution)	Examinateur
	Prénom NOM	(Institution)	Codirecteur de thèse
	Prénom NOM	(Institution)	Invité

MAP5 - Université de Paris
45, rue des Saints Pères
75006 Paris

Abstract

Optimal transport theory has found many application in diverse fields in machine learning thank to providing a powerful tool for comparing probability distributions, one of a crucial issue in machine learning. In this thesis, we leverage the optimal transport theory and statistics to deal with the problems in biology and actuary. We develop new methodology based on evaluating OT distance between empirical distributions attached on real datasets and turn it into a loss function for optimisation using a mini-batch gradient descent and Sinkhorn algorithm.

In the first part of this thesis, we present several algorithms designed to learn a pattern of correspondence between two data sets in situations where it is desirable to match elements that exhibit a relationship belonging to a known parametric model. The algorithms unfold in two stages. First, an optimal transport plan and an optimal affine transformation are learned. Second, the OT matrix is exploited to derive either several co-clusters or several sets of matched elements.

In the second part, we develop a new methodology to anticipate which cities will request a declaration of natural disaster for a drought event, a key step of the national compensation scheme. We build an inertial proximal algorithm for nonconvex optimization. The optimisation problem is designed so as to yield a sparse vector of predictions because it is known that relatively few cities will make the request.

Keywords. Optimal transport; Sinkhorn algorithm; Sinkhorn divergence; proximal algorithm; matching; Huntington's disease; omics data; natural disasters.

Résumé

Cette thèse présente les applications de la théorie du Transport Optimal et des statistiques dans deux domaines : la biologie et l'actuariat. Nous apprenons la divergence de Sinkhorn, une classe de divergences entre objets, basée sur la distance OT régularisée pour découvrir les modèles des ensembles de données. La divergence de Sinkhorn est considérée comme la fonction de perte que nous voulons minimiser en utilisant une descente de gradient en mini-batch et l'algorithme de Sinkhorn.
Dans la première partie de cette thèse, nous présentons plusieurs algorithmes conçus pour apprendre un motif de correspondance entre deux ensembles de données dans des situations où il est souhaitable de faire correspondre des éléments qui présentent une relation appartenant à un modèle paramétrique connu. Les algorithmes se déroulent en deux étapes. Premièrement, un plan de transport optimal et une transformation affine optimale sont appris. Ensuite, la matrice OT est exploitée pour dériver soit plusieurs co-clusters, soit plusieurs ensembles d'éléments appariés.
Dans la deuxième partie, nous développons une nouvelle méthodologie pour anticiper les villes qui demanderont une déclaration de catastrophe naturelle pour un événement de sécheresse, une étape clé du système d'indemnisation national. Nous construisons un algorithme proximal inertiel pour l'optimisation non convexe. Le problème d'optimisation est conçu de manière à produire un vecteur clairsemé de prédictions car on sait que relativement peu de villes feront la demande.
Mots-Clefs : Algorithme de Sinkhorn ; contraste de Sinkhorn ; co-clustering spectral ; génomique ; maladie de Huntington ; matching ; transport optimal.

Remerciements

Merci à Warith, qui m'a enseigné d'utiliser le Python et partager les idées.

Merci à toi Olivier, qui m'a fait confiance il y a maintenant de cinq ans. Ton soutien permanent, qu'il soit scientifique ou amical.

Pour finir, j'aimerais te remercier, Antoine, très sincèrement.

Un très grand merci à Anne et Fabienne pour leur soutien indéfectible, à Marie- Hélène, ..., et l'ensemble des équipes administratives et techniques pour leur aide si précieuse durant ces années passées au MAP5. Une salutation générale pour tous ceux, permanents ou passagers, que j'ai pu croiser et côtoyer, et qui contribuent à cette ambiance si particulière et chaleureuse de MAP5 que j'ai envie de nommer l'esprit terrasse de 7ème.

Merci aux tous fameux éphémères de MAP5.

Merci à Safa et Ousmane de partager les bonnes moments et de compagne longtemps.

Merci à Florian, Charlie, Antoine Monod.

En fin,

Notations and definitions

Definitions

Notations

- $[\![M]\!]$: set of integers $\{1, \ldots, M\}$.
- Ω_M : probability simplex with M bins, namely the set of probability vectors in $\mathbb{R}_+$.
- $\mathbf{1}_M$: vector of size M with all entries equal to 1.
- $\mathbf{0}_d$: vector of size d with all entries equal to 0.
- $c(x, y)$: cost function, with associated pairwise cost matrix $(C(\mathbf{x}, \mathbf{y}))_{m,n} = c(x_m, y_n)$ evaluated on $\mathbf{x}$ and $\mathbf{y}$.
- (a, b) : histograms in the simplices $\Omega_M \times \Omega_N$.
- (α, β) probability measures, defined on spaces $(\mathcal{X}, \mathcal{Y})$
- $\Pi(a, b)$: set of couplings between vectors a, b.
- $\Pi(\alpha, \beta)$: set of couplings between measures α, β.
- $(\mu^a_{\mathbf{x}} := \sum_{m \in [\![M]\!]} a_m \delta_{x_m}, \nu^b_{\mathbf{y}} := \sum_{n \in [\![N]\!]} b_n \delta_{y_n})$: the weighted empirical measure attached to $\mathbf{x} := \{x_1, \ldots, x_M\}$ and $\mathbf{y} := \{y_1, \ldots, y_N\}$, respectively.
- For $\rho \in \mathbb{R}^M$, $\operatorname{diag}(\rho)$ is the $M \times M$ matrix with diagonal ρ and zero otherwise.
- $OT_c(\alpha, \beta)$: value of optimization problem associated to the optimal transport with cost function c.
- $\langle \cdot, \cdot \rangle_F$: for the usual Euclidean dot-product between vectors; for two matrices of the same size A and B, $\langle A, B \rangle_F := \operatorname{Tr} A^\top B$ is the Frobenius dot-product.
- $K := \mathrm{e}^{-C/\gamma}$ Gibbs kernel associated to the cost matrix C.
- $a \otimes b := ab^\top \in \mathbb{R}^{M \times N}$.
- $a \odot b := (a_m b_m) \in \mathbb{R}^M$ for $(a, b) \in (\mathbb{R}^M)^2$.
- $\mathbf{f} \oplus \mathbf{g} := \mathbf{f}\mathbf{1}_M^\top + \mathbf{1}_N \mathbf{g}^\top \in \mathbb{R}^{M \times N}$ for two vectors $\mathbf{f} \in \mathbb{R}^M$, $\mathbf{g} \in \mathbb{R}^N$

Abbreviations

Conflicts in notation between chapters

We have tried to use coherent and non-conflicting notation for the mathematical objects defined in this thesis. However, for the sake of consistency with the conventions of the field, we made the choice to keep conventional notations for known quantities. ...
add more detail, where there are conflict
Theses notational conflicts have been kept to ease the understanding of the manuscript. They occur between different chapters but not inside each chapter. We stress that the potential uncertainty is removed when the context is taken into consideration.

Contents

1 Introduction 1
1.1 What this thesis is about? 1
1.2 Formalisation 2
1.3 State of the art 2
1.4 Design, programming and implementation of algorithms 2
1.5 Results 2

2 Elements of transport optimal 3
2.1 Optimal transport 3

3 Conclusion and discussion 11
3.1 Conclusion 11
3.2 Discussion 11

1

Introduction

1.1	What this thesis is about?	1
1.2	Formalisation	2
1.3	State of the art	2
1.4	Design, programming and implementation of algorithms	2
1.5	Results	2

1.1 What this thesis is about?

Optimal transport theory has found many application in diverse fields in machine learning thank to providing a powerful tool for comparing probability distributions, one of a crucial issue in machine learning. In this thesis, we leverage the optimal transport theory and statistics to deal with the problems in biology and actuary. The biological problem is to assess the possible relationships between microRNA and mRNA expression in the striatum of Huntington's disease model mice. The actuarial problem relates to anticipate the declaration of natural disaster for a drought event.

Huntington's disease

Huntington's disease, an autosomal-dominant, progressive neurodegenerative disorder, is characterized by involuntary chromatic movements with cognitive and behavioral disturbances and is caused by an expansion of a repeating CAG triplet series in the huntingtin gene Walker (2007); MacDonald et al. (1993). In normal individuals the CAG repeat length ranges from 10 to 35, while in HD individuals, it ranges from 36 to more than 120 which is inversely correlated with age of onset. Specifically, HD patients with 36-40 CAG repeats may have late onset or may not develop symptoms while repeat lengths in the 40s have symptom onset in the fourth decade and repeat lengths greater than 60 lead to juvenile onset cite . There are currently no treatments to prevent the onset or slow the progression of HD.

Huntington's disease, like several neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease and amyotrophic lateral sclerosis, relates to gene deregulation which has encouraged large studies to gene regulatory mechanisms. Gene expression is controlled by limiting the amount of mRNA produced from a particular gene at the transcription level and regulating the translation of mRNA into proteins at the post-transcriptional level. The most important instruments in the latter level are small non-coding RNAs called miRNAs. It binds to a complementary sequence in the 3'UTR of the target mRNA resulting in a rapid degradation of the mRNA or less frequently in an inhibition of its translation into protein. So the researchers are interested in studying the interaction between miRNAs and mRNAs in HD to gain a deeper understanding the disease and to develop a new treatment.

ANTICIPATE THE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT IN FRANCE

1.2 Formalisation

HUNTINGTON'S DISEASE Advanced sequencing technologies such as RNAseq produce large data sets as genome, proteome, transcriptome and metabolome. The analysis of numerous data provides insight into genetics, human biology and disease. Notably, the data of Huntington's disease in post-mortem human brains and in mouse models are increasingly available, including one of the largest omics data set of mRNA, miRNA and protein data collectively quantifying several layers of molecular regulation in the brain of HD model knock-in mice. The data promoted various studies

Encouraged by the promising findings of , our goal is to shed light on the interaction between mRNA and miRNAs based on multidimensional data which are collected at three different ages in striatum (a brain region) from an allelic series of HD model knock-in mice with increasing CAG length in the endogenous Huntingtin gene. For each combination of poly Q length and age, we have quantified miRNA and mRNA expression of 8 mice including 4 females and 4 males. After preprocessing ...we obtained the final dataset consisting of $M = 13616$ mRNA profiles and $N = 1143$ miRNA profiles. The fact that miRNA and target mRNAs are linked by a "many-to-many" mirroring relationship because a miRNA induces the degradation of a target mRNA or blocks its translation into proteins, or both and a miRNA can regulate several mRNAs. The biological hypothesis guides to identify the negatively correlative miRNA-mRNA pairs.

1.3 State of the art

HUNGTINTON DISEASE As reviewed by Nazarov and Kreis (2021),

ANTICIPATE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT

1.4 Design, programming and implementation of algorithms

1.5 Results

2

Elements of transport optimal

2.1 Optimal transport . . . 3
2.1.1 Assignment and Monge problem 3
2.1.2 Kantorovich relaxation 4
2.1.3 Entropic regularization 6
2.1.4 Sinkhorn loss 9

This thesis presents the applications of optimal transport (OT) theory along with statistics in real-world problems therefore we dedicate this chapter to provide a concise but self-contained introduction to OT and to state all the notions upon which the rest of the thesis will use. In Section 2.1, we first describes in short the basics of optimal transport by introducing the related notions of assignment and Monge problem then its generalization, namely Kantorovich problem. After that, we focus on the theoretical and numerical result of the regularized OT that has several important advantages. We finally present a family of divergences, so-called Sinkhorn divergences, interpolating between regularized OT and Maximum Mean Discrepancy (MMD) losses. In Section... ♣ continue here ♣

2.1 Optimal transport

2.1.1 Assignment and Monge problem

Optimal assignment problem Fix two integers $M, N \geq 1$, we denote two datasets by $\mathbf{x} := \{x_1, \dots, x_M\} \subset \mathcal{X}$ and $\mathbf{y} := \{y_1, \dots, y_N\} \subset \mathcal{Y}$ where $\mathcal{X}, \mathcal{Y}$ are the metric spaces. Let $[\![M]\!] := \{1, \dots, M\}$ be the set of all positive integers up to M, we consider a cost matrix $C(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+^{M \times N}$ where $(C(\mathbf{x}, \mathbf{y}))_{m,n}$ represents the cost of moving a unit of mass from x_m to y_n. Assuming $M = N$, the optimal assignment problem consists of finding a bijective $\sigma : [\![M]\!] \to [\![M]\!]$ such that the total cost $\sum_{m \in [\![M]\!]} (C(\mathbf{x}, \mathbf{y}))_{m,\sigma(m)}$ is minimized. A native solution is to evaluate the total cost of $M!$ permutations of M elements. However, $M!$ is huge even for small M so this may be very inefficient. In fact, there are several acceptable algorithms in time polynomial of the number of points M, see in (Peyré and Cuturi, 2019, Section 3.7).

MONGE PROBLEM A generalization of optimal assignment problem, known as Monge problem, was introduced by the French mathematician Gaspard Monge in Monge (1781) as follow: a worker must find the "best" way to transport a certain of soil from the ground to places where it should be use in a construction. Assume that the source and target places are known and the transportation cost to move a unit of mass between two points is as well. The goal is to determine the destination to which a source point should be transported so that the total cost is minimal. This problem can be stated equivalently as follows. Denote $\Omega_d := \{a \in (\mathbb{R}_+)^d | \sum_{i \in [\![d]\!]} a_i = 1\}$ the $(d-1)-$dimensional simplex. For any $(a,b) \in \Omega_M \times \Omega_N$, let $\alpha := \sum_{m \in [\![M]\!]} a_m \delta_{x_m}$, $\beta := \sum_{n \in [\![N]\!]} b_n \delta_{y_n}$ be two weighted empirical measure attached to $\mathbf{x}$ and $\mathbf{y}$. Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ defined the transportation cost to move a unit of mass from x_m to y_n, the Monge problem consists in solving

$$\min_{T \in \mathcal{T}} \sum_{m \in [\![M]\!]} c(x_m, T(x_m)), \tag{2.1}$$

where $\mathcal{T} := \{T : \mathbf{x} \to \mathbf{y} | b_n = \sum_{m : T(x_m) = y_n} a_m\}$, so-called the feasible set, is the set of all mappings that associates each point x_m to a single point y_n and the mass conservation constraints are meet. Note that the mapping T between two finite sets can be represented in a straightforward way by an assignment $\sigma : [\![M]\!] \to [\![N]\!]$ where $\sigma(m) = n$ iff $T(x_m) = y_n$ and the constraints are equivalent to $\sum_{m \in \sigma^{-1}(n)} a_m = b_n$. When $M = N$ and two measures are uniform, i.e. $\alpha := \frac{1}{M} \sum_{m \in [\![M]\!]} \delta_{x_m}, \beta := \frac{1}{M} \sum_{n \in [\![M]\!]} \delta_{y_n}$, then the conservation constraints induces that T is a bijection, such that $T(x_m) = y_{\sigma(n)}$ and the Monge problem corresponds to the optimal assignment problem with the cost matrix $C_{m,n} = c(x_m, y_n)$. Note that the set $\mathcal{T}$ may be empty if the two measures α and β are incompatible, for example $M < N$ or $\sum_{m \in [\![M]\!]} a_m \neq \sum_{n \in [\![N]\!]} b_n$ so the Monge problem may not have solution. In case of the existence of solution, it is very difficult and costly to solve this problem.

2.1.2 Kantorovich relaxation

The assignment problem is a special case of Monge problem when two measures are uniform attached to two sets of the same size. Monge problem allows to consider two arbitrary measures and to assign several source points to a target point. However, both problems are hard to solve in practice.
Much later Loenid Vitaliyevich Kantorovich, a Russian mathematician, rediscovered the Monge problem motivated of economic problem. Kantorovich Kantorovich (1942) proposed an excellent idea that allow to split the mass of each source point and move them to several target points. Therefore, Kantorovich formulation consists in solving, in place of a map T, a probabilistic matrix P where P_{mn} describes the amount of mass moved from x_m to y_n. This coupling matrix should be satisfied the mass conservation constraints, i. e., the sums of row and column should be equals to a and b, respectively. Formally, the set of admissible couplings is defined by

$$\Pi(a,b) := \{P \in (\mathbb{R}_+)^{M \times N} | P\mathbf{1}_N = a, P^\top \mathbf{1}_M = b\}.$$

In fact, $\Pi(a,b)$ can be expressed as the set of the joint probability matrix over $(\mathbf{x}, \mathbf{y})$ with marginal distributions w and w', respectively. Obviously, this set contains $a \times b$

so is nonempty. Another benefit is the symmetric property in the sense that P is an element of $\Pi(a, b)$ if and only if $P^\top$ is an element of $\Pi(b, a)$ as well. Given a cost matrix $C(\mathbf{x}, \mathbf{y}) \in (\mathbb{R}_+)^{M \times N}$, where $(C(\mathbf{x}, \mathbf{y}))_{mn} = c(x_m, y_n)$, Kantorovich formulation consists in solving

$$\mathrm{OT}_c(\alpha, \beta) := \min_{P \in \Pi(a,b)} \langle C(\mathbf{x}, \mathbf{y}), P \rangle_F, \tag{2.2}$$

where $\langle C(\mathbf{x}, \mathbf{y}), P \rangle_F := \sum_{(m,n) \in [\![M]\!] \times [\![N]\!]} (C(\mathbf{x}, \mathbf{y}))_{mn} P_{mn}$ is the P-specific expected cost of transport from $\mathbf{x}$ to $\mathbf{y}$. In many cases, the notation $\mathrm{OT}_c(\alpha, \beta)$ is useful to indicate explicitly the dependence on the cost function c defined the cost matrix $C(\mathbf{x}, \mathbf{y})$.

We generalize the definition (2.2) of OT_c between arbitrary measures by first introducing some useful notations of functions and probability measures. Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures over $\mathcal{X}$. Given a continuous map $f : \mathcal{X} \to \mathcal{Y}$ we denote $f_\sharp : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ its associated push-forward operator, i.e., the push-forward measure $\beta = f_\sharp(\alpha)$ of $\alpha \in \mathcal{P}(\mathcal{X})$ satisfies

$$\int_{\mathcal{Y}} h(y) d\beta(y) = \int_{\mathcal{X}} h(f(x)) d\alpha(x), \quad \forall h \in \mathcal{C}(\mathcal{Y}),$$

where $\mathcal{C}(\mathcal{Y})$ is the space of continuous and smooth functions over $\mathcal{Y}$. In the general case, we consider the joint probability distribution P over the product space $\mathcal{X} \times \mathcal{Y}$ instead of the probability matrix, but that should be satisfy the mass conservation constraints. The set of admissible couplings can be defined

$$\Pi(\alpha, \beta) := \{P \in \mathcal{P}(\mathcal{X}, \mathcal{Y}) | \pi_{X\sharp}(P) = \alpha, \pi_{Y\sharp}(P) = \beta\},$$

where $\pi_{\mathcal{X}\sharp}$ and $\pi_{\mathcal{Y}\sharp}$ are the push-forward operators of the projections $\pi_{\mathcal{X}}(x, y) = x$ and $\pi_{\mathcal{Y}}(x, y) = y$, respectively. So the Kantorovich problem in general case is

$$\mathrm{OT}_c(\alpha, \beta) := \min_{P \in \Pi(\alpha,\beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) dP(x, y) \tag{2.3}$$

This infinite-dimensional linear optimization over a space of measures have a solution under mild assumptions, for example $(\mathcal{X}, \mathcal{Y})$ are compact spaces and the cost function c is continuous. Furthermore the OT loss can be rewritten as the expectation of $c(X, Y)$

$$\mathrm{OT}_c(\alpha, \beta) = \min_{(X,Y)} \{\mathbb{E}_{X,Y}(c(X, Y)) : X \sim \alpha, Y \sim \beta\}, \tag{2.4}$$

where (X, Y) is a couple of random variables with the joint law $P \in \Pi(\alpha, \beta)$ and the marginal laws α and β, respectively.

Optimal transport loss as the distance One of advantage of OT theory is that the OT cost from one to other measure can be seen as the distance if the cost function is chosen as the distance function. Indeed, whenever $\mathcal{X}$ is equipped with a metric $d_{\mathcal{X}}$, it is natural to use it as cost function, i.e., $c(x, y) = d_{\mathcal{X}}(x, y)^p$, with $p \geq 1$. In such case, the OT cost in Equation (2.2) is called the p-Wasserstein distance, which we denote as $\mathcal{W}_p(\alpha, \beta) := \mathrm{OT}_{d^p_{\mathcal{X}}}(\alpha, \beta)$. The case $p = 1$ is also known as the Kantorovich-Tubinstein in statistics or the Earth Mover's Distance in computer vision. The Proposition below shows that these are indeed proper distances.

Proposition 2.1. *Assume $\mathcal{X} = \mathcal{Y}$, and suppose $(\mathcal{X}, d)$ is a metric space and that $(\alpha, \beta) \in$*

2.1.3 Entropic regularization

The high computational cost of solving the Kantorovich problem has led to various schemes to solve it approximately. One of the most popular such approaches is to add an entropy regularization term to the objective Cuturi (2013).

For this we define the discrete entropy of a coupling as:

$$E(P) := -\sum_{(m,n)\in[\![M]\!]\times[\![N]\!]} P_{mn}(\log P_{mn} - 1)$$

and use it to obtain a regularized version of problem (2.2) as follows

$$\mathrm{OT}_C^{\gamma}(a,b) = \min_{P\in\Pi(a,b)} \left\{\langle C_{X,Y}, P\rangle_F - \gamma E(P)\right\}. \tag{2.5}$$

Lemma 2.1. *Prove that the solution $\bar{P}$ of problem $\min_{P\in\Pi(a,b)} -E(P)$ is $a\otimes b$.*

Proof. Applying the Lagrange multiplier method yields

$$\mathcal{L}(P,\mathbf{f},\mathbf{g}) = -E(P) - \langle \mathbf{f}, P\mathbf{1}_M - a\rangle - \langle \mathbf{g}, P^\top \mathbf{1}_N - b\rangle.$$

We compute the gradient

$$\frac{\partial \mathcal{L}(P,\mathbf{f},\mathbf{g})}{\partial P_{m,n}} = \log(P_{m,n}) - \mathbf{f}_m - \mathbf{g}_n, \forall (m,n)\in[\![M]\!]\times[\![N]\!]$$

Therefore $\bar{P}_{m,n} = \mathrm{e}^{\mathbf{f}_m}\mathrm{e}^{\mathbf{g}_n}$. By substituting into the constraints $\bar{P}\mathbf{1}_M = a$ and $\bar{P}\top\mathbf{1}_N = b$, we obtain

$$\mathrm{e}^{\mathbf{g}_n}\sum_{m\in[\![M]\!]}\mathrm{e}^{\mathbf{f}_m} = b_n, \quad \mathrm{e}^{\mathbf{f}_m}\sum_{n\in[\![N]\!]}\mathrm{e}^{\mathbf{g}_n} = a_m.$$

Furthermore, we have $\sum_{m\in[\![M]\!]}\mathrm{e}^{\mathbf{f}_m}\sum_{n\in[\![N]\!]}\mathrm{e}^{\mathbf{g}_n} = 1$, hence $\bar{P}_{m,n} = \mathrm{e}^{\mathbf{f}_m}\mathrm{e}^{\mathbf{g}_n} = a_m b_n$. This concludes the proof. □

Proposition 2.2. *(adapted from Peyré and Cuturi (2019)). The solution P_γ of (2.5) converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovich problem, namely*

$$P_\gamma \xrightarrow{\gamma\to 0} \arg\min_P\{-E(P) : P\in\Pi(a,b), \langle C_{X,Y}, P\rangle_F = \mathrm{OT}_c(a,b)\} \tag{2.6}$$

so that in particular

$$\mathrm{OT}_c^{\gamma}(a,b) \xrightarrow{\gamma\to 0} \mathrm{OT}_c(a,b)$$

One also has

$$P_\gamma \xrightarrow{\gamma\to\infty} a\otimes b = a(b)^\top = (a_m b_n)_{m,n}$$

Proof. We consider a sequence (γ_ℓ) such that $\gamma_\ell \to 0$ and $\gamma_\ell > 0$. We denote P_ℓ the solution of (2.5) for $\gamma = \gamma_\ell$. Since $\Pi(a,b)$ is bounded, we can extract a sequence (that we do not relabel for the sake of simplicity) such that $P_\gamma \to P^\star$. Since $\Pi(a,b)$ is closed, $P_\star \in \Pi(a,b)$. We consider any P such that $\langle C, P\rangle_F = \mathrm{OT}_C(a,b)$. By optimality of P and P_ℓ for their respective optimization problems (for $\gamma = 0$ and $\gamma = \gamma_\ell$, one has

$$0 \le \langle C, P_\ell\rangle - \langle C, P\rangle \le \gamma_\ell\left(E(P_\ell) - E(P)\right). \tag{2.7}$$

Since E is continuous, taking the limit $\ell \to +\infty$ in this expression show that $\langle C, P^\star \rangle = \langle C, P \rangle$ so that $P^\star$ is a feasible point of (2.6). Furthermore, dividing by γ_ℓ in (2.7) and taking the limit shows that $E(P) \le E(P^\star)$, which shows that $P^\star$ is a solution of (2.7). Since the solution $P_0^\star$ to this program is unique by strict convexity of $-E$, one has $P^\star = {}_0^\star$, and the whole sequence is converging.

Similarly, considering a sequence $(\bar{\gamma}_k)$ such that $\bar{\gamma}_k \to +\infty$, we denote $\bar{P}_k$ the solution of (2.6) for $\gamma = \bar{\gamma}_k$, then $\bar{P}_k \to \bar{P}_\infty$ with $\bar{P}_\infty \in \Pi(a, b)$. Furthermore, we have the inequality

$$0 \le E(\bar{P}) - E(\bar{P}_k) \le \frac{1}{\bar{\gamma}_k} \left(\langle C, \bar{P} \rangle - \langle C, \bar{P}_k \rangle \right)$$

Taking the limit $k \to +\infty$ in this expression shows that $E(\bar{P}) = E(\bar{P}_\infty)$. By the Lemma 2.1, we imply that $\bar{P}_\infty = \bar{P} = a \otimes b$, hence this finishes the proof. $\square$

Besides computational advantages, regularizing the OT problem often leads to better empirical performance in applications where having denser correspondences is beneficial, e.g, when the support points correspond to noisy features.

The regularized version of discrete optimal transport (2.5) is a strictly convex optimization problem. Below we show that its solution has a simple analytic expression

Proposition 2.3. *(adapted from Peyré and Cuturi (2019).)*
The solution to (2.5) *is unique and has the form*

$$P^\star = \operatorname{diag}(u) K \operatorname{diag}(v) \tag{2.8}$$

where $K = \mathrm{e}^{-\frac{C}{\gamma}}$ *is Gibbs kernel associated to the cost matrix* C *and* $u \in (\mathbb{R}_+^*)^M, v \in (\mathbb{R}_+^*)^N$ *are two (unknown) scaling variables.*

Proof. The Lagrangian with respect to (2.5) is

$$\mathcal{L}(P, \mathbf{f}, \mathbf{g}) := \langle P, C \rangle - \gamma E(P) - \langle \mathbf{f}, P \mathbf{1}_N - a \rangle - \langle \mathbf{g}, P^\top \mathbf{1}_M - b \rangle,$$

where $\mathbf{f} \in \mathbb{R}_+^M$ and $\mathbf{g} \in \mathbb{R}_+^N$. Now, let us calculate the gradient

$$\frac{\partial \mathcal{L}(P, \mathbf{f}, \mathbf{g})}{\partial P_{m,n}} = C_{m,n} + \gamma \log(P_{m,n}) - (\mathbf{f}_m + \mathbf{g}_n),$$

and set it equal 0, we imply that $P_{m,n} = \mathrm{e}^{\mathbf{f}_m/\gamma} \mathrm{e}^{-C_{m,n}/\gamma} \mathrm{e}^{\mathbf{g}_n/\gamma}$. Therefore, we obtain the optimal solution as (2.8) by using the notation $u = (\mathrm{e}^{\mathbf{f}_m})_{m \in [\![M]\!]}$ and $v = (\mathrm{e}^{\mathbf{g}_n})_{n \in [\![N]\!]}$. $\square$

The factorization of the OT matrix $P^\star$ allows us to solve that problem easily by finding two nonnegative vectors (u, v). The two conservation constraints can be expressed as the following equations

$$\operatorname{diag}(u) K \operatorname{diag}(v)\, \mathbf{1}_N = a \quad \text{and} \quad \operatorname{diag}(v) K^\top \operatorname{diag}(v)\, \mathbf{1}_M = b.$$

Since $\operatorname{diag}(v)\, \mathbf{1}_M = v$ and $\operatorname{diag}(u)\, \mathbf{1}_N = u$, we simplify that equations into equivalent form

$$u \odot (Kv) = a \quad \text{ and } v \odot K^\top u = b \tag{2.9}$$

where $\odot$ denotes the component-wise multiplication of vectors. This problem, so-called the classical matrix scaling problem, can be solved through an iterative method which alternately normalizes u and v to satisfy the right-hand side and right-hand side of Equation (2.9). More specifically, initialized with any positive vector $v^{(0)} = \mathbf{1}_N$, we implement two updates in each iteration of procedure known as Sinkhorn's algorithm

$$u^{(\ell+1)} := \frac{a}{Kv^{(\ell)}} \quad \text{ and } v^{(\ell+1)} := \frac{b}{K^\top u^{(\ell+1)}}$$

where the division operator between two vectors is to be understood element-wise. Now we present an elementary proof of linear convergence of the iterations by using the Hilbert projective metric on $(\mathbb{R}_+^*)^d$.

Definition 2.1. *The Hilbert projective metric on* $(\mathbb{R}_+^*)^d$ *is defined by*

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') := \log \max \left\{ \tfrac{x_i x'_j}{x'_i x_j} : i, j \in [\![d]\!] \right\}.$$

We will use the following properties Birkhoff (1957):

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = \| \log(x) - \log(x') \|_{\text{var}}; \tag{2.10}$$

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = d_{\mathcal{H}}(x/x', \mathbf{1}_d) = d_{\mathcal{H}}(\mathbf{1}_d /x', \mathbf{1}_d /x); \tag{2.11}$$

$$\forall K \in (\mathbb{R}_+^*)^{d \times d'}, \forall x, x' \in (\mathbb{R}_+^*)^{d'}, d_{\mathcal{H}}(Kx, Kx') \leq \lambda(K) d_{\mathcal{H}}(x, x'), \tag{2.12}$$

where $\lambda(K) := \frac{\sqrt{\eta(K)}-1}{\sqrt{\eta(K)}+1} < 1$ with $\eta(K) := \max \left\{ \frac{K_{i,k} K_{j,\ell}}{K_{j,k} K_{i,\ell}} : i, j \in [\![d]\!], k, \ell \in [\![d']\!] \right\}$. We have the following convergence theorem.

Theorem 2.1. *One has* $(u^{(\ell)}, v^{(\ell)}) \to (u^\star, v^\star)$ *and*

$$d_{\mathcal{H}}(u^{(\ell)}, u^\star) = O\big(\lambda(K)^{2\ell}\big), \quad d_{\mathcal{H}}(v^{(\ell)}, v^\star) = O\big(\lambda(K)^{2\ell}\big), \tag{2.13}$$

where $u^\star, v^\star$ *are the optimal solutions. Furthermore,*

$$d_{\mathcal{H}}(u^{(\ell)}, u^\star) \leq \frac{d_{\mathcal{H}}\big(P^{(\ell)} \mathbf{1}_M, a\big)}{1 - \lambda(K)^2}, \tag{2.14}$$

$$d_{\mathcal{H}}(v^{(\ell)}, v^\star) \leq \frac{d_{\mathcal{H}}\big((P^{(\ell)})^\top \mathbf{1}_N, b\big)}{1 - \lambda(K)^2}, \tag{2.15}$$

where $P^{(\ell)} := \operatorname{diag}(u^{(\ell)}) K \operatorname{diag}(v^{(\ell)})$. *Last, one has*

$$|| \log(P^{(\ell)}) - \log(P^\star) ||_{\max} \leq d_{\mathcal{H}}(u^{(\ell)}, u^\star) + d_{\mathcal{H}}(v^{(\ell)}, v^\star), \tag{2.16}$$

where $P^\star$ *is the unique solution of* (2.5)

Proof. Using (2.11) and (2.12), we get

$$\begin{aligned} d_{\mathcal{H}}(u^{(\ell+1)}, u^\star) &= d_{\mathcal{H}} \left(\frac{a}{Kv^{(\ell)}}, \frac{a}{Kv^\star} \right) \\ &= d_{\mathcal{H}}(Kv^{(\ell)}, Kv^\star) \leq \lambda(K) d_{\mathcal{H}}(v^{(\ell)}, v^\star). \end{aligned} \tag{2.17}$$

Likewise and the fact that $\lambda(K^\top) = \lambda(K)$, we get

$$\begin{aligned} d_{\mathcal{H}}(v^{(\ell)}, v^\star) &= d_{\mathcal{H}} \left(\frac{b}{K^\top u^{(\ell)}}, \frac{b}{K^\top u^\star} \right) \\ &= d_{\mathcal{H}}(K^\top u^{(\ell)}, K^\top u^\star) \\ &\leq \lambda(K^\top) d_{\mathcal{H}}(u^{(\ell)}, u^\star) = \lambda(K) d_{\mathcal{H}}(u^{(\ell)}, u^\star). \end{aligned} \tag{2.18}$$

The inequalities (2.17) and (2.18) imply that

$$d_{\mathcal{H}}(u^{(\ell+1)}, u^\star) \leq (\lambda(K))^2 d_{\mathcal{H}}(u^{(\ell)}, u^\star).$$

That is equivalent to the left-hand side of equation (2.13), likewise to the right-hand side. By invoking in turn the triangle inequality and both (2.11) and (2.12), we get

$$\begin{aligned} d_{\mathcal{H}}(u^{(\ell)}, u^\star) &\leq d_{\mathcal{H}}(u^{(\ell+1)}, u^{(\ell)}) + d_{\mathcal{H}}(u^{(\ell+1)}, u^\star) \\ &\leq d_{\mathcal{H}}\left(\frac{a}{Kv^{(\ell)}}, u^{(\ell)}\right) + \lambda(K)^2 d_{\mathcal{H}}(u^{(\ell)}, u^\star) \\ &= d_{\mathcal{H}}\left(a, u^\ell \odot (Kv^{(\ell)})\right) + \lambda(K)^2 d_{\mathcal{H}}(u^{(\ell)}, u^\star). \end{aligned}$$

The above inequality and the fact that $u^\ell \odot (Kv^{(\ell)}) = P^{(\ell)}\mathbf{1}_M$ imply (2.14). Likewise (2.15) can be proved in an analogous way. (2.16) is trivial ♣ check again ♣. $\square$

2.1.4 Sinkhorn loss

3

Conclusion and discussion

3.1 Conclusion 11
3.2 Discussion 11

3.1 Conclusion

3.2 Discussion

Bibliography

G. Birkhoff. Extensions of Jentzsch's theorem. *Trans. Amer. Math. Soc.*, 85, 1957.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2292–2300, USA, 2013. Curran Associates Inc.

L. Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk SSSR*, pages 227–229, 1942.

M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, and et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell*, 72(6):971–983, 1993.

G. Monge. *Mémoire sur la théorie des déblais et des remblais*. Histoire de l'Académie Royale des Sciences, 1781.

P. V. Nazarov and S. Kreis. Integrative approaches for analysis of mRNA and microRNA high-throughput data. *Computational and Structural Biotechnology Journal*, 19:1154–1162, 2021.

G. Peyré and M. Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends in Machine Learning Series. Now Publishers, 2019.

F. O. Walker. Huntington's disease. *The Lancet*, 369(9557):218–228, 2007.