Optimal Transport in Machine Learning

Yoann Coudert–Osmont Némo Fournier Jérémy Petithomme
Victor Mercklé Charles Gassot

28 avril 2019

Table des matières

Chapitre A

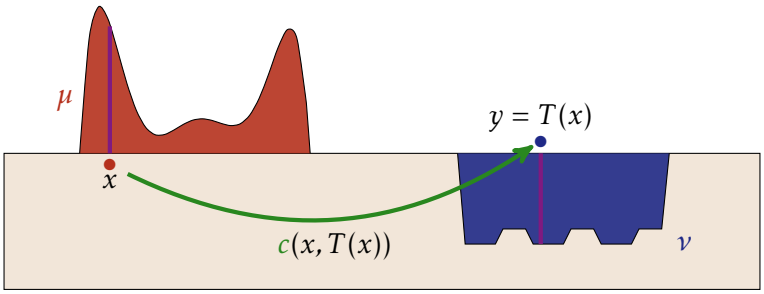
Overview of Optimal Transport

Table des matières

Le transport optimal permet de définir une géométrie et une distance sur l'ensemble des mesures de probabilités sur un espace.

I - Introduction

1/ Problème de Monge



Monge On se place dans un espace de probabilité Ω. Puis on se donne μ et ν, deux mesures de probabilités sur $\mathcal{P}(\Omega)$. De plus on se donne une fonction de coût $c : \Omega \times \Omega \to \mathbb{R}$. Cette fonction peut être la distance euclidienne par exemple. Le problème de Monge est alors le suivant :

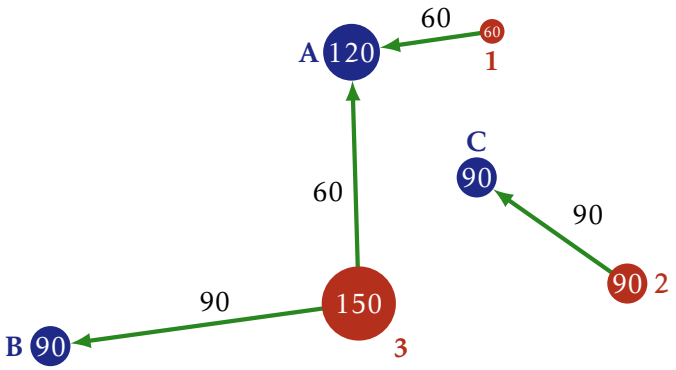
$$\int_{\Omega} c\left(x, T(x)\right) \mu(x)$$

$$\text{s.t.} \quad \forall A \in \mathcal{P}(\Omega),\ \nu(A) = \mu\left(T^{-1}(A)\right)$$

La fonction T est appelé un plan de transport. Elle indique où est déplacé chaque morceau de masse. La masse qui se trouve en x est déplacé à la position $T(x)$. La fonction c permet alors de quantifier le coût du déplacement d'une unité de masse. On cherche alors à minimiser la distance moyenne de déplacement d'une unité de masse en transport optimal.

Limitations Le problème de cette formulation est qu'elle n'est pas utilisable sur des mesures discrètes. En effet si μ possède un Dirac plus grand que tous les Diracs de ν alors la construction d'un plan de transport devient impossible.

2/ Problème de Kantorovich



Changement de plan de transport L'idée de Kantorovich a été de changer la fonction $T : \Omega \to \Omega$ en une probabilité P sur l'espace produit $\mathcal{P}(\Omega \times \Omega)$. Sur l'exemple ci-dessus, la formulation de Monge nous interdit de séparé la masse de taille 150 en deux. Avec la formulation de Kantorovich on peut faire cela et ici une masse de taille 90 va a un endroit et une masse de taille 60 à un autre endroit. Dans le cas discret P peut être représenté par une matrice :

Matrice de Transport

150	p_{1A}	p_{1B}	p_{1C}
90	p_{2A}	p_{2B}	p_{2C}
60	p_{3A}	p_{3B}	p_{3C}
	120	90	90

Matrice des Distances

d_{1A}	d_{1B}	d_{1C}
d_{2A}	d_{2B}	d_{2C}
d_{3A}	d_{3B}	d_{3C}

Cas discret En notant D notre matrice des distances, on obtient la formulation suivante dans le cas discret :

$$\min_P \langle P, D \rangle = \sum_i \sum_j p_{ij} d_{ij}$$
$$\text{s.t.} \quad \begin{cases} \forall i & \sum_j p_{ij} = \mu_i \\ \forall j & \sum_i p_{ij} = \nu_j \end{cases}$$

Cas continue Dans le cas continue on introduit l'ensemble des probabilités conjointes suivant :

$$\Pi(\mu,\nu) = \{P \in \mathcal{P}(\Omega \times \Omega) \mid \forall A, B \in \Omega,\, P(A \times \Omega) = \mu(A),\, P(\Omega \times B) = \nu(B)\}$$

Cet ensemble traduit les conditions du cas discret dans le cas continue. On reprend notre fonction de coût c définie dans le problème de Monge. La formulation de Kantorovich devient le problème de minimisation suivant :

$$\inf_{P \in \Pi(\mu,\nu)} \mathbb{E}_P[c(X,Y)] = \int\int c(x,y)P(dx,dy)$$

3/ Distance de Wasserstein

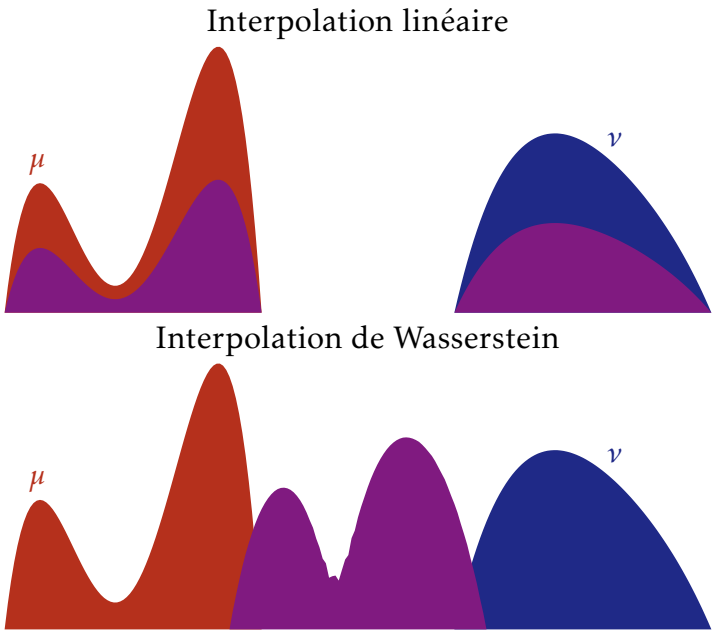
Ce problème d'optimisation permet de définir une distance entre les mesures de probabilités.

Définition

Étant donnée un espace de probabilité Ω, une fonction de coût $c : \Omega \times \Omega \to \mathbb{R}$ et deux mesures de probabilités μ et ν dans $\mathcal{P}(\Omega)$, la **p-distance de Wasserstein** entre μ et ν est définie par :

$$W_p(\mu,\nu) = \left(\inf_{P \in \Pi(\mu,\nu)} \int\int c(x,y)^p P(dx,dy)\right)^{1/p}$$

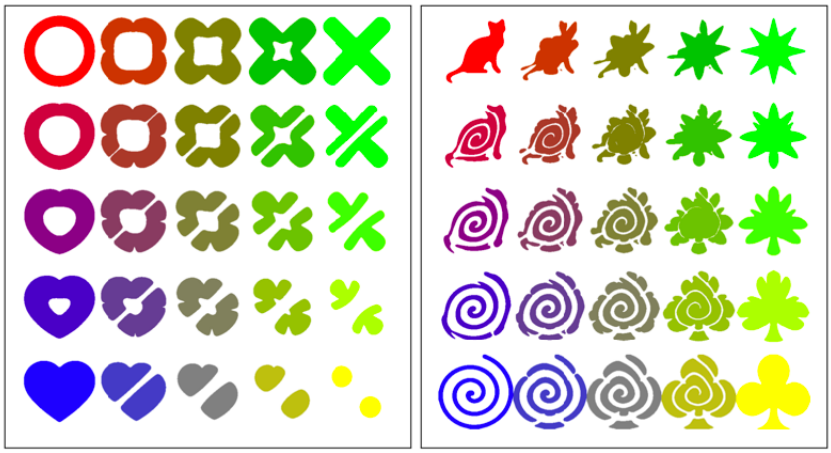
Cette distance peut permettre de calculer des barycentres entres des distributions de probabilité beaucoup plus naturels que des barycentres obtenus avec la norme l_2. La figure suivante nous montre la différence entre une interpolation (barycentre de deux distributions) obtenue via la distance de Wasserstein et une interpolation linéaire.



On rappelle que le calcul du barycentre μ des des N distributions ν_i associé aux poids λ_i revient au problème de minimisation suivant :

$$\min_{\mu \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_p^p(\mu, \nu_i)$$

Et voici maintenant ce que donne des barycentres avec 4 distributions :



II - Comment calculer le transport optimal ?

1/ Dual

Pour les calculs on revient vers les espaces discrets. Nos mesures de probabilité sont des sommes de diracs :

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \qquad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

Notre fonction de coût est une matrice $D \in \mathbb{R}_+^{n \times m}$ tel que $D_{ij} = d(x_i, y_j)^p$. Notre distribution conjointe P est aussi une matrice de $\mathbb{R}_+^{n \times m}$ qui doit vérifier certaines contraintes pouvant se représenter de manière matricielle. L'ensemble dans lequel se trouve P est le suivant :

$$U(a,b) = \left\{ P \in \mathbb{R}_+^{n \times m} \mid P \mathbb{1}_m = a,\ P^\top \mathbb{1}_n = b \right\}$$

Définition

La définition de la distance de Wasserstein dans un espace discret est la suivante :

$$W_p^p(\mu, \nu) = \min_{P \in U(a,b)} \langle P, D \rangle$$

Dual On s'intéresse maintenant au problème dual. On introduit le lagrangien :

$$L(P, \alpha, \beta) = \langle P, D \rangle + \alpha^\top (a - P \mathbb{1}_m) + \beta^\top (b - P^\top \mathbb{1}_n)$$

La fonction objective du dual est :

$$g(\alpha,\beta) = \min_{P\geqslant 0} L(P,\alpha,\beta) = \min_{P\geqslant 0} \alpha^\mathsf{T} a + \beta^\mathsf{T} b + \langle P, D - \alpha \mathbb{1}_m^\mathsf{T} - \mathbb{1}_n \beta^\mathsf{T}\rangle$$

Si la matrice $D - \alpha \mathbb{1}_m^\mathsf{T} - \mathbb{1}_n \beta^\mathsf{T}$ possède un coefficient négatif alors en faisant tendre le même coefficient de la matrice P vers $+\infty$ on obtient pour minimum $-\infty$. En revanche si tous les coefficients sont positifs alors $\langle P, D - \alpha \mathbb{1}_m^\mathsf{T} - \mathbb{1}_n \beta^\mathsf{T}\rangle$ est positif et il suffit de prendre $P = 0$ pour se retrouver avec une valeur nulle. On obtient alors :

$$g(\alpha,\beta) = \begin{cases} \alpha^\mathsf{T} a + \beta^\mathsf{T} b & \text{si } D - \alpha \mathbb{1}_m^\mathsf{T} - \mathbb{1}_n \beta^\mathsf{T} \geqslant 0 \\ -\infty & \text{sinon} \end{cases}$$

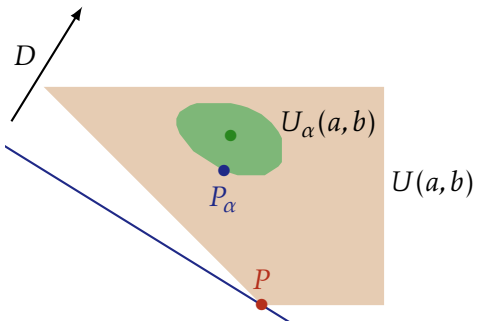
Finalement le dual est la maximisation de cette fonction g.

Définition

Le **dual** du transport optimal a la formulation suivante :

$$W_p^p(\mu,\nu) = \max_{\substack{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^m \\ \alpha_i+\beta_j\leqslant D(x_i,y_j)^p}} \alpha^\mathsf{T} a + \beta^\mathsf{T} b$$

Avec un solveur de flot de coût minimum on peut résoudre cette optimisation linéaire en temps $\mathcal{O}(n^3 \log(n))$. Mais la solution est instable. Comme illustré sur la figure ci-dessous. Si on change un peu notre distance D ou si on modifie un peu nos mesures source et destination μ et ν la valeur de P peut subir un changement brutal puisque P est un sommet de notre simplex. Une solution consiste à faire une régularisation de notre problème pour se retrouver avec un ensemble $U_\alpha(a,b)$ strictement convexe. Cet ensemble est strictement inclus dans $U(a,b)$ et la solution optimale de ce nouveau problème ne sera plus optimal pour le problème initial mais sera tout de même correcte. Finalement la régularisation que l'on va voire permet aussi d'obtenir un algorithme de résolution itératif de moindre complexité.



2/ Régularisation

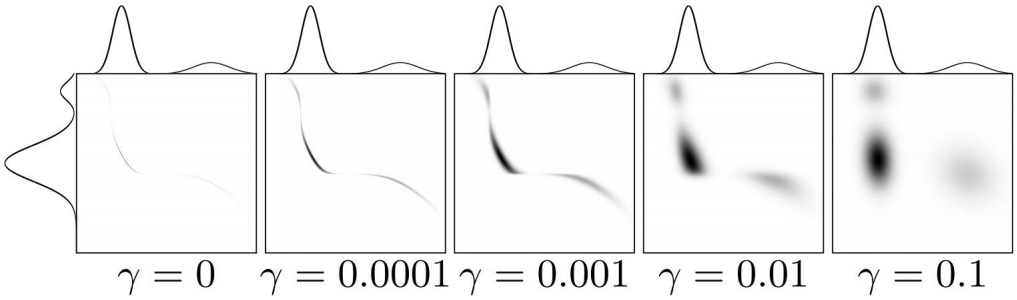
La régularisation que l'on effectue consiste à maximiser l'entropie de probabilité conjointe P. On fait apparaître un nouveau paramètre γ qui contrôle cette régularisation et on obtient la distance suivante :

Définition

La **distance de Wasserstein régularisée** est définie de la manière suivante :

$$W_\gamma(\mu,\nu) = \min_{P\in U(a,b)} \langle P, D\rangle - \gamma H(P)$$

Remarque On rappelle que l'on peut considérer deux variables aléatoires X et Y tel que $X \sim \mu$, $Y \sim \nu$ et $X, Y \sim P$. L'entropie de P est maximale lorsque X et Y sont indépendants, c'est à dire lorsque $P = ab^\mathsf{T}$ et l'entropie maximal est alors $H(\mu) + H(\nu)$. La régularisation donne donc une matrice avec des valeurs strictement positive plus réparties.



Par stricte convexité, il existe une unique matrice P_γ qui minimise la distance :

$$P_\gamma = \arg\min_{P\in U(a,b)} \langle P, D\rangle - \gamma H(P)$$

Proposition 1

Il existe un unique couple de vecteurs u et v appartenant à $\mathbb{R}_+^n$ et $\mathbb{R}_+^m$ tel que :

$$P_\gamma = \mathrm{diag}(u) K \mathrm{diag}(v), \qquad K = e^{-D/\gamma}$$

Démonstration On écrit le nouveau laplacien :

$$L(P,\alpha,\beta) = \sum_{ij} \Big(P_{ij} D_{ij} + \gamma P_{ij} \big(\log_2 P_{ij} - 1 \big) \Big) + \alpha^\mathsf{T}(a - P\mathbb{1}_m) + \beta^\mathsf{T}(b - P^\mathsf{T}\mathbb{1}_n)$$

On calcule alors la dérivée partielle par rapport à P_{ij} :

$$\frac{\partial L}{\partial P_{ij}} = M_{ij} + \gamma \log_2 P_{ij} - \alpha_i - \beta_j$$

Le dual est la maximisation du minimum du laplacien par rapport à P. Donc on doit avoir P tel que la dérivée partielle est nulle :

$$\frac{\partial L}{\partial P_{ij}} = 0 \Rightarrow P_{ij} = e^{\frac{\alpha_i}{\gamma}} e^{-\frac{D_{ij}}{\gamma}} e^{\frac{\beta_j}{\gamma}} = u_i K_{ij} v_j$$

■

Sinkhorn On dispose alors d'une méthode itérative pour trouver cette matrice P_γ. On traduit la condition $P_\gamma \in U(a,b)$:

$$P_\gamma \in U(a,b) \Leftrightarrow \left\{ \begin{array}{rcl} \mathrm{diag}(u)K\mathrm{diag}(v)\mathbb{1}_m & = & a \\ \mathrm{diag}(v)K^\mathsf{T}\mathrm{diag}(u)\mathbb{1}_n & = & b \end{array} \right.$$

$$P_\gamma \in U(a,b) \Leftrightarrow \left\{ \begin{array}{rcl} \mathrm{diag}(u)Kv & = & a \\ \mathrm{diag}(v)K^\mathsf{T}u & = & b \end{array} \right.$$

On introduit alors $\odot$ le produit coefficient par coefficient :

$$P_\gamma \in U(a,b) \Leftrightarrow \left\{ \begin{array}{rcl} u \odot Kv & = & a \\ v \odot K^\mathsf{T}u & = & b \end{array} \right.$$

$$P_\gamma \in U(a,b) \Leftrightarrow \left\{ \begin{array}{rcl} u & = & a/Kv \\ v & = & b/K^\mathsf{T}u \end{array} \right.$$

L'algorithme de Sinkhorn est alors le suivant :

Algorithm 1: Sinkhorn

repeat
  $u \leftarrow a/Kv$;
  $v \leftarrow b/K^\mathsf{T}u$;
until *convergence de* u, v;

Complexité Il a été prouvé que l'algorithme converge en temps linéaire. De plus une étape de calcul se fait en $\mathcal{O}(nm)$ mais les calculs peuvent être parallélisés. Il est aussi possible d'utiliser des convolutions sur une grille pour obtenir une complexité en $\mathcal{O}(n \log n)$.

Dual On peut aussi relever la formulation du dual avec régularisation :

$$W_\gamma(\mu,\nu) = \max_{\alpha,\beta} \alpha^\mathsf{T}a + \beta^\mathsf{T}b - \gamma\left(e^{\alpha/\gamma}\right)^\mathsf{T} K \left(e^{\beta/\gamma}\right)$$

Lien avec la divergence KL En posant le noyau :

$$K_\gamma(i,j) = \exp\left(-\frac{d_{i,j}}{\gamma}\right)$$

On obtient une reformulation de la régularisation entropique avec la divergence KL :

$$W_\gamma(\mu,\nu) = \min_{P \in U(a,b)} \gamma KL(P|K_\gamma)$$

Chapitre B

Review of papers

Table des matières

I - Sinkhorn Distances : Lightspeed Computation of Optimal Transport

[2]

II - Wasserstein GAN

[1]

III - Convolutional Wasserstein Distances : Efficient Optimal Transportation on Geometric Domains

[3]

Chapitre C

Implementation

Bibliographie

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv e-prints*, page arXiv :1701.07875, Jan 2017.

[2] Marco Cuturi. Sinkhorn distances : Lightspeed computation of optimal transport. pages 2292–2300, 2013.

[3] Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas J. Guibas. Convolutional wasserstein distances : efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4) :66 :1–66 :11, 2015.