Optimal Transport: The Semismooth Newton Algorithm

Anton Conrad

Mai 2024

You can find the code at the following address: lien github.

1 Introduction

Optimal Transport focus on the Wasserstein distance. Let X and Y be two bounded domains in $\mathbb{R}^d$ and let $P(X)$ and $P(Y)$ be the sets of Borel probability measures on X and Y, respectively.

Let μ and ν two probability measure and d a distance on $X \times Y$, let

$$\Pi(\mu, \nu) = \{\pi \in \text{Prob}(X \times Y) | \pi(\cdot, Y) = \mu, \pi(X, \cdot) = \nu\}$$

The 2-Wasserstein distance between these measures is defined as:

$$W_2(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} d(x, y)^2 \, d\pi(x, y) \right)^{\frac{1}{2}}$$

So, we can chose the euclidean distance and we've got an optimisation problem that we can reformulate as:

$$\mathrm{OT}(\mu,\nu) := \frac{1}{2}\left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{X\times Y}\|x-y\|^2\,d\pi(x,y)\right)$$

This quantity has applications in many fields, including Fairness, Shape Interpolation, Generative moddeling (GANs), Classification and Clustering... but its calculation can be tricky. We'll present and implement the Semismooth Newton Algorithm which aims to tackle this calculation with interesting complexity $(\mathrm{O}(\frac{1}{\sqrt{k}}))$.

2 Reformulation: Duality and Kernel Trick

2.1 Duality

Firstly, we consider the dual problem:

$$\sup_{u,v\in C(\mathbb{R}^d)}\left(\int_X u(x)\,d\mu(x)+\int_Y v(y)\,d\nu(y)\right),\ \text{s.t.}\quad \frac{1}{2}\|x-y\|^2\geq u(x)+v(y),\quad \forall(x,y)\in X\times Y,$$

Santambrogio (2015) shows that the supremum can be attained so, the problem can be formulated as:

$$\max_{u,v\in C(\mathbb{R}^d)}\left(\int_X u(x)\,d\mu(x)+\int_Y v(y)\,d\nu(y)\right),\ \text{s.t.}\quad \frac{1}{2}\|x-y\|^2\geq u(x)+v(y),\quad \forall(x,y)\in X\times Y,$$

Futhermore, Vacher et al. transform the constraint to make it much easier to manipulate by "replacing the inequality constraints with equivalent equality constraints, and considering these constraints over n points". So, we take $\{(\tilde{x}_1,\tilde{y}_1),...,(\tilde{x}_n,\tilde{y}_n)\}\subset X\times Y$ and our problem is now:

$$\max_{u,v\in C(\mathbb{R}^d)}\left(\int_X u(x)\,d\mu(x)+\int_Y v(y)\,d\nu(y)\right),\ \text{s.t.}\quad \frac{1}{2}\|\tilde{x}_i-\tilde{y}_i\|^2 = u(\tilde{x}_i)+v(\tilde{y}_i),\quad \forall i\in[|1;n|]$$

2.2 Kernel Trick

The "Kernel Trick" is used to circumvent the curse of dimensionality by representing the u and v functions in a reproducing kernel Hilbert space (RKHS). This allows parameter calculations to be expressed in terms of kernel functions, facilitating the calculation of kernel-based OT estimators. This approach does not suffer from the exponential increase in the error rate as is the case for entropy-regularized OT estimators.

We assume that the support sets X, Y are convex, bounded, and open with Lipschitz boundaries, and that the densities of μ, ν are finite, bounded away from zero and m-times differentiable with $m > 2d+2$.

Vacher et al. (2021) show that the optimal dual fonctions $u^\star$ and $uv^\star$ belong to some explicitly defined Hillbert space H_X and H_Y respectively that are

RKHS under previous assumptions, associated with the feature maps Φ_X and Φ_Y respectively. We also introduce $H_{XY} \subset C(X \times Y)$ which also a RHKS, associated with the feature map Φ_{XY} on $X \times Y$.

Therefore, we can represent the quantity to maximize by the kernel mean embeddings $w_\mu = \int_X \phi_X(x)\, d\mu(x)$ and $w_\nu = \int_Y \phi_Y(y)\, d\nu(y)$ such that the dual optimisation problem is equivalent to the following formulation:

$$\max_{u,v,A} \langle u, w_\mu \rangle_{H_X} + \langle v, w_\nu \rangle_{H_Y} \,,$$

s.t.

$$\frac{1}{2}\|\tilde{x}_i - \tilde{y}_i\|^2 - u(\tilde{x}_i) - v(\tilde{y}_i) = \langle \phi_{XY}(\tilde{x}_i, \tilde{y}_i), A\phi_{XY}(\tilde{x}_i, \tilde{y}_i) \rangle_{H_{XY}} \,, \quad \forall i \in [|1; n|]$$

Where A is a symmetrical operator on H_{XY}.

All we have left to do is the kernel trick with:

$$k_X(x, x') = \langle \phi_X(x), \phi_X(x') \rangle_{H_X},$$

$$k_Y(y, y') = \langle \phi_Y(y), \phi_Y(y') \rangle_{H_Y},$$

$$k_{XY}((x, y), (x', y')) = \langle \phi_{XY}(x, y), \phi_{XY}(x', y') \rangle_{H_{XY}}.$$

2.3 Discretization, Regularization, Primality

Given data $(x_1, ..., x_p) \sim \mu$ an $y_1, \ldots, y_p \sim \nu$, we get the empirical measures: $\hat{\mu} = \frac{1}{p}\sum_{i=1}^p \delta_{x_i}$ and $\hat{\nu} = \frac{1}{p}\sum_{i=1}^p \delta_{y_i}$, and the empirical kernel mean embeddings: $\hat{w}_\mu = \frac{1}{p}\sum_{i=1}^p \phi_X(x_i)$ et $\hat{w}_\nu = \frac{1}{p}\sum_{i=1}^p \phi_Y(y_i)$. Futhermore, we chose to impose penalization terms for u, v, and A.

With fixed regularization parameters $\lambda_1, \lambda_2 > 0$, the optimization problem becomes:

$$\max_{u,v,A} \langle u, \hat{w}_\mu \rangle_{H_X} + \langle v, \hat{w}_\nu \rangle_{H_Y} - \lambda_1 \operatorname{Tr}(A) - \lambda_2(\|u\|^2_{H_X} + \|v\|^2_{H_Y}),$$

s.t.

$$\frac{1}{2}\|\tilde{x}_i - \tilde{y}_i\|^2 - u(\tilde{x}_i) - v(\tilde{y}_i) = \langle \phi_{XY}(\tilde{x}_i, \tilde{y}_i), A\phi_{XY}(\tilde{x}_i, \tilde{y}_i) \rangle_{H_{XY}}, \quad \forall i \in [|1; n|]$$

This problem is still hard to compute as u, and v belong to infinite-dimensional spaces. Vacher et al. (2021) tackle this issue while keeping strong duality (that is necessary to have the same optimal values for primal and dual problems). We define: $Q \in \mathbb{R}^{n \times n}$ with $Q_{ij} = k_X(\tilde{x}_i, \tilde{x}_j) + k_Y(\tilde{y}_i, \tilde{y}_j)$, and $z \in \mathbb{R}^n$ with $z_i = \hat{w}_\mu(\tilde{x}_i) + \hat{w}_\nu(\tilde{y}_i) - \lambda_2\|\tilde{x}_i - \tilde{y}_i\|^2$, and $q_2 = \|\hat{w}_\mu\|^2_{H_X} + \|\hat{w}_\nu\|^2_{H_Y}$, where:

$$\hat{w}_\mu(\tilde{x}_i) = \frac{1}{p}\sum_{j=1}^p k_X(x_j, \tilde{x}_i),$$

$$\hat{w}_\nu(\tilde{y}_i) = \frac{1}{p}\sum_{j=1}^{p} k_Y(y_j, \tilde{y}_i),$$

$$\|\hat{w}_\mu\|_{H_X}^2 = \frac{1}{\mathrm{n}_p^2}\sum_{1\leq i,j\leq p} k_X(x_i, x_j),$$

$$\|\hat{w}_\nu\|_{H_Y}^2 = \frac{1}{\mathrm{n}_p^2}\sum_{1\leq i,j\leq p} k_Y(y_i, y_j).$$

Futhermore, we define: $K \in \mathbb{R}^{n\times n}$ with $K_{ij} = k_{XY}((\tilde{x}_i, \tilde{y}_i), (\tilde{x}_j, \tilde{y}_j))$ and R as an upper triangular matrix for the Cholesky decomposition of K. Finally, we let Φ_i be the i-th column of R. Then, the previous dual problem is equivalent to:

$$\min_{\gamma\in\mathbb{R}^n} \frac{1}{4\lambda_2}\gamma^\top Q\gamma - \frac{1}{2\lambda_2}\gamma^\top z + \frac{q_2}{4\lambda_2},$$

s.t.

$$\sum_{i=1}^{n} \gamma_i \Phi_i \Phi_i^\top + \lambda_1 I \succeq 0.$$

Finally, it is the same optimization problem as:

$$\min_{\gamma\in\mathbb{R}^n} \max_{X\in S_n^+(\mathbb{R})} \left\{ \frac{1}{4\lambda_2}\gamma^\top Q\gamma - \frac{1}{2\lambda_2}\gamma^\top z + \frac{q_2}{4\lambda_2} - \langle X, \sum_{i=1}^{n}\gamma_i\Phi_i\Phi_i^\top + \lambda_1 I\rangle \right\}$$

as $\sum_{i=1}^{n}\gamma_i\Phi_i\Phi_i^\top + \lambda_1 I \succeq 0 \Leftrightarrow \max_{X\in S_n^+(\mathbb{R})} -\langle X, \sum_{i=1}^{n}\gamma_i\Phi_i\Phi_i^\top + \lambda_1 I\rangle = 0$. We call this the $(\star)$ optimization problem.

3 Resolution

We introduce: $\Phi : \mathbb{R}^{n\times n} \to \mathbb{R}^n$and its adjoint $\Phi^\star : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ as :

$$\Phi(X) = \begin{pmatrix} \langle X, \Phi_1\Phi_1^\top\rangle \\ \vdots \\ \langle X, \Phi_n\Phi_n^\top\rangle \end{pmatrix}, \quad \Phi^\star(\gamma) = \sum_{i=1}^{n}\gamma_i\Phi_i\Phi_i^\top.$$

Thus, we let $R : \mathbb{R}^n \times \mathbb{R}^{n\times n} \to \mathbb{R}^n \times \mathbb{R}^{n\times n}$ be:

$$R(\gamma, X) = \begin{pmatrix} \frac{1}{2\lambda_2}Q\gamma - \frac{1}{2\lambda_2}z - \Phi(X) \\ X - \mathrm{proj}_{S_n^+}(X - (\Phi^\star(\gamma) + \lambda_1 I)) \end{pmatrix}$$

and we arrive at the most crucial property, which is that (γ, X) is an optimal solution of $(\star)$ if and only if $R(\gamma, X) = 0$.

Then, we note $w = (\gamma, X)$ and the Semismooth Newton Algorithm is written as follows:

$$w_{k+1} = w_k + \Delta w_k,$$

where Δw_k is obtained by resolving:

$$(J_k + \mu_k I)[\Delta w_k] = -R(w_k)$$

where J_k belongs to the generalized Jacobian of R at w_k, and $I$is the identity. The parameter $\mu_k = \theta_k \|R(w_k)\|$ is chosen in order to stabilize the method. Qi and Sun (1993), Zhou and Toh (2005), Xiao et al. (2018) show that as R is strongly semismooth, the methode converges. This linear system isn't practical to solve so "we seek an approximation step Δw_k by solving":

$$\|(J_k + \mu_k I)[\Delta w_k] + R(w_k)\| \leq \tau \min\{1, \kappa \|R(w_k)\| \|\Delta w_k\|\}$$

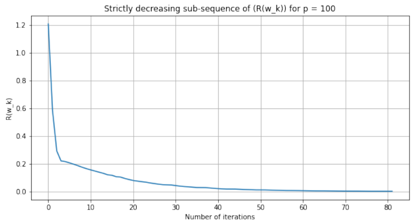
where $\|w\| = \|\gamma\|_2 + \|X\|_F$ and $0 < \tau, \kappa < 1$.

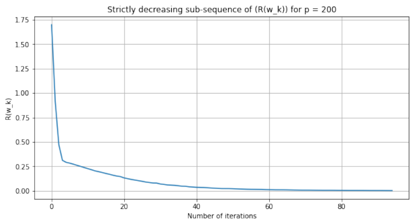
Finally, the calculation of Δw_k is technical and can be found in the code function.ipynb. Also, it is summarized in the "Algorithm 1" of the paper.

4 Tests

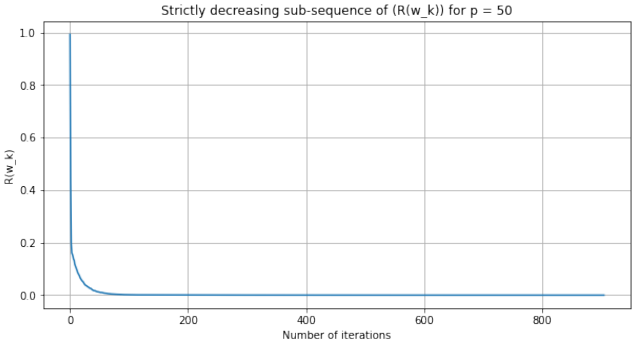
We didn't implement the extra-gradient part of the algorithm as we wanted to focus of the results of the Semismooth Newton algorithm by itself and because of "the existence of a local region where 1-step regularized Semismooth Newton algorithm can reduce the residue norm at a quadratic rate" so at some point it will not be pertinent to compute the extra-gradient.

Thus, we generated some artificial data and some filling points and chose the parameters like in the paper. We can plot the convergence of $R(w_k)$ to 0 for different values of p:

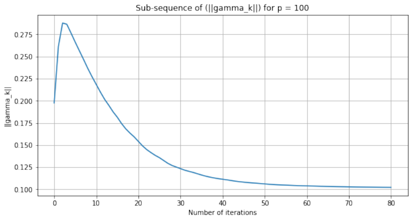


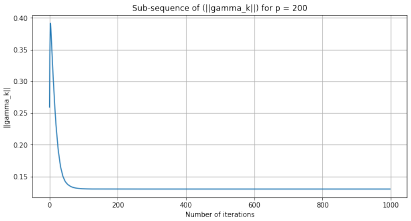


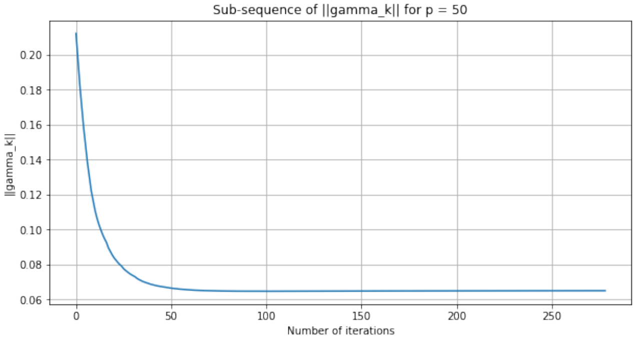
Where we can indeed observe that: $\|R(w_k)\| = O(\frac{1}{\sqrt{k}})$.



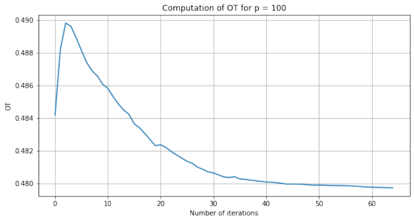
Then, we can observe the convergence of (γ_k):

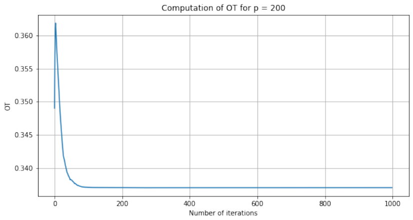


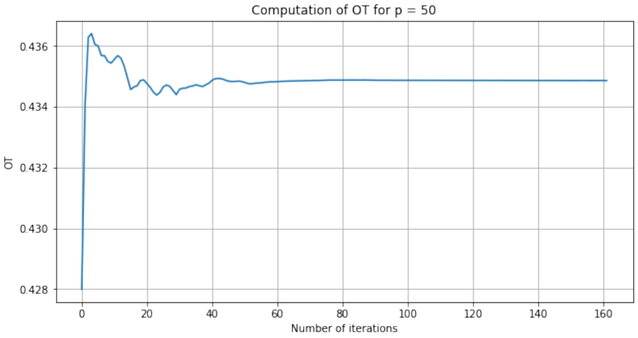




And, finally we can compute the quantity whose calculation was the first objective:







5 Conclusion

We've been provided a powerful implementation of the Optimal Transport quantity. It would be interesting to see those results on real data, and to see if the use of the extra-gradient enhance the results. We could discuss the approximation when calculating Δw_k. Futhermore, we didn't give any detail about the kernel computation. However, we are happy with our results and the complexity of our algorithm.