TP 2 - TRANSPORT OPTIMAL ENTROPIQUE ET INFÉRENCE DE TRAJECTOIRES

Elias Ventre
contact: elias.ventre@inria.fr

Résumé

Dans ce travail pratique, nous implémentons l'algorithme de Sinkhorn pour calculer le transport optimal entropique entre deux distributions empiriques. Nous utilisons ensuite ces couplages optimaux pour reconstruire des trajectoires cellulaires à partir de mesures temporelles discrètes, et évaluons la qualité de reconstruction via l'interpolation de McCann. Cette approche s'inscrit dans le contexte de l'analyse de données de RNA-seq sur cellule unique, où l'on cherche à inférer les trajectoires de différenciation cellulaire.

1 Introduction

1.1 Contexte biologique

La différenciation cellulaire est un processus dynamique où des cellules changent progressivement leur expression génique pour adopter des rôles spécifiques. Comprendre ces trajectoires de différenciation est fondamental en biologie du développement et en médecine régénérative.

Les technologies modernes de séquençage d'ARN sur cellule unique (scRNA-seq) permettent de mesurer l'expression de milliers de gènes dans des cellules individuelles. Cependant, la mesure détruit les cellules : on ne peut observer que des *snapshots* de populations cellulaires à différents temps, sans pouvoir suivre les trajectoires individuelles.

Problème : Étant données des distributions empiriques de cellules $\hat{P}_{t_1}, \dots, \hat{P}_{t_N}$ à différents temps, comment reconstruire les trajectoires de différenciation ?

1.2 Modélisation mathématique

On modélise la différenciation cellulaire par une équation différentielle stochastique (EDS) avec dérive gradient :

$$dX_t = -\nabla\Psi(t, X_t)dt + \sqrt{2\sigma}dB_t \tag{1}$$

où $\nabla\Psi$ représente le paysage épigénétique (dérive), σ est le coefficient de diffusion, et B_t est un mouvement brownien.

1.3 Objectifs du TP

Ce travail pratique vise à :

— Comprendre le lien entre transport optimal entropique et problème de Schrödinger

— Implémenter l'algorithme de Sinkhorn avec la bibliothèque POT

— Reconstruire des trajectoires par chaînage de couplages optimaux

— Calculer l'interpolation de McCann (géodésiques de Wasserstein)

— Évaluer quantitativement la qualité de reconstruction

— Observer l'effet du paramètre de régularisation entropique ε

2 Transport optimal entropique

2.1 Rappels théoriques

2.1.1 Transport optimal classique

Étant données deux distributions de probabilité μ et ν sur $\mathbb{R}^d$, le problème de transport optimal de Kantorovich cherche le couplage π qui minimise le coût de transport :

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi[c(X,Y)] \tag{2}$$

où $\Pi(\mu,\nu)$ est l'ensemble des couplages ayant μ et ν comme marginales, et $c(x,y) = |x-y|^2$ est le coût quadratique.

Pour des distributions empiriques $\mu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ et $\nu = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$, le calcul du transport optimal exact a une complexité $O(m^3 \log m)$ avec l'algorithme du réseau de flots.

2.1.2 Régularisation entropique

Pour améliorer la stabilité numérique et réduire la complexité, on ajoute un terme d'entropie de Shannon :

$$\gamma_\varepsilon^* = \arg\min_{\pi \in \Pi(\mu,\nu)} \{\mathbb{E}_\pi[c(X,Y)] + \varepsilon H(\pi)\} \tag{3}$$

où $H(\pi) = -\int \pi(x,y)\log \pi(x,y)\, dxdy$ est l'entropie.

L'algorithme de Sinkhorn [2] résout ce problème avec une complexité $O(m^2 K)$ où K est le nombre d'itérations (typiquement $K \ll m$).

2.1.3 Lien avec le problème de Schrödinger

Le transport optimal entropique peut se reformuler comme un problème d'entropie relative :

$$\min_{\pi \in \Pi(\mu,\nu)} KL\left(\pi \,\middle|\, e^{-c/\varepsilon}\right) = \min_{\pi \in \Pi(\mu,\nu)} H(\pi) + \frac{1}{\varepsilon}\mathbb{E}_\pi[c] \tag{4}$$

Cela correspond au *problème de Schrödinger* : trouver le processus stochastique le plus proche (au sens de l'entropie relative) du mouvement brownien, qui relie μ à ν en temps Δt [3].

Résultat clé : Pour $\varepsilon = \sigma \Delta t$, le couplage optimal $\gamma_\varepsilon^*(x,\cdot)$ correspond au noyau de transition d'une EDS à dérive gradient reliant μ à ν !

2.2 Algorithme de Sinkhorn

L'algorithme de Sinkhorn exploite la structure du problème régularisé. La solution s'écrit sous la forme :

$$\gamma^* = \mathrm{diag}(u) K \mathrm{diag}(v) \tag{5}$$

où $K_{ij} = e^{-c_{ij}/\varepsilon}$ est le noyau de Gibbs, et $(u,v) \in \mathbb{R}_+^m \times \mathbb{R}_+^m$ sont les vecteurs de scaling obtenus par itération :

$$u^{(k+1)} = \frac{a}{K v^{(k)}}, \quad v^{(k+1)} = \frac{b}{K^\top u^{(k+1)}} \tag{6}$$

où les divisions sont élément par élément. On initialise typiquement avec $v^{(0)} = \mathbf{1}_m$ (vecteur de uns), et on itère jusqu'à convergence. L'algorithme converge vers la solution unique du problème régularisé.

Algorithm 1 Algorithme de Sinkhorn

Entrée : Distributions $a, b \in \mathbb{R}_+^m$, matrice de coût $C \in \mathbb{R}^{m \times m}$, $\varepsilon > 0$
Initialisation : $v \leftarrow \mathbf{1}_m$
while non convergence **do**
    $u \leftarrow a \oslash (Kv)$ ▷ $\oslash$ = division élément par élément
    $v \leftarrow b \oslash (K^\top u)$
end while
Retour : $\gamma = \mathrm{diag}(u) K \mathrm{diag}(v)$

3 Inférence de trajectoires

3.1 Méthode Waddington-OT

La méthode Waddington-OT [1] utilise le transport optimal entropique pour reconstruire des trajectoires cellulaires. Entre deux snapshots consécutifs $\hat{P}_{t_i}$ et $\hat{P}_{t_{i+1}}$, on calcule :

$$\gamma^*_{t_i,t_{i+1}} = \arg \min_{\pi \in \Pi(\hat{P}_{t_i}, \hat{P}_{t_{i+1}})} \left\{ \mathbb{E}_\pi[|X - Y|^2] + \sigma \Delta t_i \, H(\pi) \right\} \tag{7}$$

Ce couplage optimal peut ensuite être utilisé de deux façons :

1. **Chaînage stochastique** : Construire des trajectoires en échantillonnant successivement selon les couplages
2. **Interpolation de McCann** : Prédire les distributions intermédiaires

3.2 Interpolation de McCann

Étant données deux distributions μ_0 et μ_1, et un plan de transport optimal π^*, l'*interpolation de McCann* (ou *géodésique de Wasserstein*) au temps $t \in [0, 1]$ est définie par :

$$\mu_t := [(1-t)X + tY]_{\#} \, \pi^* \tag{8}$$

où $(X, Y) \sim \pi^*$ et $f_{\#}\mu$ désigne la mesure image par f.

Propriété géodésique : Cette interpolation est une géodésique à vitesse constante pour la distance de Wasserstein-2 :

$$W_2(\mu_s, \mu_t) = |t - s| \cdot W_2(\mu_0, \mu_1) \tag{9}$$

Pour des distributions empiriques, l'implémentation consiste à :

1. Échantillonner N paires (x_i, y_j) selon le couplage optimal γ^*
2. Calculer les points interpolés $z_k = (1-t)x_{i_k} + t y_{j_k}$ pour $k = 1, \ldots, N$
3. La distribution empirique $\hat{\mu}_t = \frac{1}{N} \sum_{k=1}^N \delta_{z_k}$ est l'interpolation de McCann

3.3 Évaluation de la reconstruction

Pour évaluer la qualité de l'interpolation, on compare les distributions interpolées avec les vraies distributions intermédiaires (obtenues par simulation) en utilisant la distance de Wasserstein-2 :

$$W_2(\mu, \nu) = \left(\min_{\pi \in \Pi(\mu,\nu)} \int |x - y|^2 d\pi(x, y) \right)^{1/2} \tag{10}$$

Cette distance fournit une mesure quantitative de l'erreur de reconstruction.

4 Travail attendu

Le travail pratique à réaliser est structuré en deux notebooks Jupyter :

4.1 Notebook 1 : Transport optimal entropique (45mns)

1. **Simulation des données** : Générer des snapshots de particules suivant une EDS avec potentiel à branchement
2. **Implémentation de Sinkhorn** : Calculer le couplage optimal entre deux snapshots consécutifs
3. **Visualisation** : Afficher la matrice de couplage et le plan de transport
4. **Étude paramétrique** : Analyser l'effet du paramètre ε sur :
   — La structure du couplage (sparsité)
   — Le coût de transport
   — L'entropie du couplage

4.2 Notebook 2 : Inférence de trajectoires (1h20)

1. **Chaînage de couplages** : Reconstruire des trajectoires en enchaînant les couplages optimaux
2. **Interpolation de McCann** : Implémenter l'interpolation géodésique
3. **Comparaison avec vérité terrain** :
   — Calculer des snapshots denses par simulation
   — Mesurer l'erreur W_2 entre interpolation et vérité terrain
   — Comparer pour différentes valeurs de ε
4. **Analyse des résultats** :
   — Identifier les régions temporelles difficiles (branchements)
   — Déterminer le ε optimal
   — Discuter les limitations de l'approche

5 Détails d'implémentation

5.1 Outils recommandés

— **Python 3.8**+ avec les bibliothèques :
   — `numpy` : Calcul numérique
   — `matplotlib` : Visualisation
   — `scipy` : Distances et optimisation
   — **POT** (Python Optimal Transport) : Implémentation de Sinkhorn
— **Jupyter Notebook** pour l'environnement interactif

Installation :

```
pip install numpy matplotlib scipy POT jupyter
```

5.2 Structure du code

Le repository contient :

— `src/simulation.py` : Module de simulation des EDS
— `notebooks/01_entropic_ot.ipynb` : Notebook partie 1
— `notebooks/02_trajectory_inference.ipynb` : Notebook partie 2
— `README.md` : Documentation complète

5.3 Fonctions clés à implémenter

Calcul du couplage OT :

```
def compute_ot_coupling(X_source, X_target, epsilon):
    """
    Calcule le couplage OT entropique entre deux distributions.

    Parameters
    ----------
    X_source, X_target : ndarray
        Particules sources et cibles
    epsilon : float
        Paramètre de régularisation

    Returns
    -------
    gamma : ndarray
        Plan de transport optimal
    """
```

Interpolation de McCann :

```
def mccann_interpolation(X_source, X_target, gamma, t):
    """
    Calcule l'interpolation de McCann au temps t.

    Parameters
    ----------
    X_source, X_target : ndarray
        Particules sources et cibles
    gamma : ndarray
        Plan de transport optimal
    t : float
        Temps d'interpolation (entre 0 et 1)

    Returns
    -------
    X_interp : ndarray
        Particules interpolées
    """
```

5.4 Choix du paramètre ε

D'après la théorie du problème de Schrödinger, le choix optimal est :

$$\varepsilon = \sigma \Delta t \tag{11}$$

où σ est le coefficient de diffusion de l'EDS et Δt le pas de temps entre snapshots.

Dans le TP, vous comparerez ce choix théorique avec d'autres valeurs ($\varepsilon/5$, $\varepsilon/2$, 2ε, 5ε) pour observer :

— ε trop petit $\Rightarrow$ couplage trop sparse, instabilité numérique

— ε trop grand $\Rightarrow$ couplage trop diffus, perte d'information

6 Évaluation

Vous serez évalué sur :

1. **Code fonctionnel** (40%) :
   — Notebooks exécutables sans erreur
   — Fonctions correctement implémentées
   — Visualisations claires et informatives
2. **Analyse des résultats** (40%) :
   — Réponses aux questions de réflexion
   — Interprétation des graphiques
   — Compréhension des phénomènes observés
3. **Rapport** (20%) :
   — Synthèse des résultats (2-3 pages)
   — Choix justifiés de paramètres
   — Discussion des limitations
   — Suggestions d'amélioration

Format du rendu :

— Notebooks complétés (`.ipynb`)
— Rapport de synthèse (préférablement LaTeX, HTML ou Markdown)
— À rendre avant le **[DATE À COMPLÉTER]**

7 Extensions possibles

Pour les étudiants intéressés, voici quelques pistes d'approfondissement :

7.1 Optimisation sur les marginales (gWOT)

L'approche présentée suppose que les marginales $\hat{P}_{t_i}$ sont fixées. En réalité, ces distributions empiriques sont bruitées (peu de cellules). La méthode gWOT [4] propose d'optimiser simultanément sur les marginales elles-mêmes :

$$\min_{P} \sigma H(P|W_\sigma) + \frac{1}{\lambda} \sum_{i=1}^{N} \Delta t_i H(\hat{P}_{t_i}|P_{t_i}) \tag{12}$$

où le terme de régularisation empêche l'overfitting.

7.2 Fused Gromov-Wasserstein

Pour prendre en compte la structure spatiale des tissus, on peut utiliser le transport de Gromov-Wasserstein fusionné [5] qui compare non seulement les expressions géniques mais aussi les relations de proximité entre cellules.

7.3 Données réelles

Appliquer la méthode à des datasets publics :

— **iPSC reprogramming** [1]
— **Mouse embryo development** [5]

8 Ressources

8.1 Articles de référence

Références

[1] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. *Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming.* Cell, 176(4) :928–943, 2019.

[2] M. Cuturi. *Sinkhorn distances : Lightspeed computation of optimal transport.* Advances in Neural Information Processing Systems, 26, 2013.

[3] C. Léonard. *A survey of the Schrödinger problem and some of its connections with optimal transport.* Discrete & Continuous Dynamical Systems-A, 34(4) :1533, 2013.

[4] L. Chizat, S. Zhang, M. Heitz, G. Schiebinger. *Trajectory inference via mean-field Langevin in path space.* Advances in Neural Information Processing Systems, 35 :27276–27289, 2022.

[5] D. Klein, G. Palla, M. Lange, C.-Y. Lin, L. Hetzel, S. Anchang, et al. *Mapping cells through time and space with moscot.* Nature, 2025.

[6] G. Peyré, M. Cuturi. *Computational optimal transport : With applications to data science.* Foundations and Trends in Machine Learning, 11(5-6) :355–607, 2019.

8.2 Documentation en ligne

— **POT Documentation** : https://pythonot.github.io/

— **Waddington-OT** : https://github.com/broadinstitute/wot

— **MOSCOT** : https://github.com/theislab/moscot

— **Cours en ligne** : https://optimaltransport.github.io/

Bon travail !

N'hésitez pas à me contacter pour toute question : elias.ventre@inria.fr