

# Sur la relation entre les flux rectifiés et le transport optimal

Nathan Bouvier

Thomas Wininger

May 2026

## 1 Introduction

Dans le domaine de l'IA générative, les méthodes à base de flux permet de transporter une distribution  $\mu_0$  vers une autre distribution  $\mu_1$ , en se basant sur des notions de transport optimal. Si la première distribution est par exemple une gaussienne, et la seconde une distribution d'image de chats, ces techniques vont permettre d'apprendre un couplage entre les deux distributions, en se basant sur le fait qu'un tel couplage peut être défini comme un champ vectoriel de vitesse  $v_t$ , et tel que l'intégration de  $\dot{z}_t = v_t(z_t)$  permet de passer d'une distribution à l'autre de manière inversible.

Cependant, même si cela permet de trouver un couplage entre deux distributions, ce couplage n'est pas forcément optimal, ce qui empêche son utilisation pour calculer des plans de transport optimaux. Pour résoudre ce problème, des travaux récents [1], [2] introduisent un algorithme itératif sensé rapprocher le flux du transport optimal. Plus précisément, étant donné un couplage initial  $(X_0, X_1)$  entre deux distributions  $\mu_0, \mu_1$ , on considère l'interpolation  $X_t = (1-t)X_0 + tX_1$  de distribution  $\mu_t$ , et on cherche à trouver le champ de vitesse  $v_t$  satisfaisant:

$$v_t \in \operatorname{argmin}_{w_t \in L^2(\mu_t)} \mathcal{L}(w_t | X_0, X_1) := \int_0^1 \mathbb{E}[\|w_t(X_t) - X_1 + X_0\|^2] \quad (1.1)$$

Ce champ de vitesse existe, est unique, et satisfait l'équation de Fokker-Planck. Sous des conditions de régularité,  $v_t$  est un transport de  $\mu_0$  à  $\mu_1$  dans le sens  $\mu_1 = z_{1\#}\mu_0$ , où  $z_t(x)$  est la solution de l'équation différentielle ordinaire  $\dot{z}_t(x) = v_t(z_t(x))$ , avec  $z_0(x) = x$ . Ce couplage peut ensuite être rectifié en résolvant l'équation différentielle pour trouver  $z_t$ , et en fixant  $Z_0 = X_0$  et  $Z_1$  la solution de  $\dot{Z}_t = v_t(Z_t)$  à  $t = 1$ .

Néanmoins, même si cette méthode permet de réduire la distance de transport, le papier que nous présentons ici : « *On the Relation between Rectified Flows and Optimal Transport* » [3], contredit les affirmations que cette itération converge vers

le transport optimal. À l'aide de contre-exemples, Hertrich et al. montrent qu'il faut des conditions supplémentaires pour faire le lien entre flux rectifié et transport optimal. Ils analysent les propriétés de cette itération, et proposent une amélioration pour garantir qu'elle reste applicable.

Le sujet est abordé dans la partie « *Conditionnal flow matching* » du cours 8 sur la diffusion. Sont présentés les quantités classiques pour le *flow matching*, à savoir, pour  $X_t = tX_0 + (1-t)X_1$  et  $T \sim \mathcal{U}([0, 1])$ :

$$\mathcal{L}(v) = \mathbb{E}[\|v_T(X_T) - (X_1 - X_0)\|^2] \quad (1.2)$$

ainsi que l'équation de la continuité pour  $X_t \sim \mu_t$ :

$$\partial_x \mu + \operatorname{div}_x(\mu v) = 0 \quad (1.3)$$

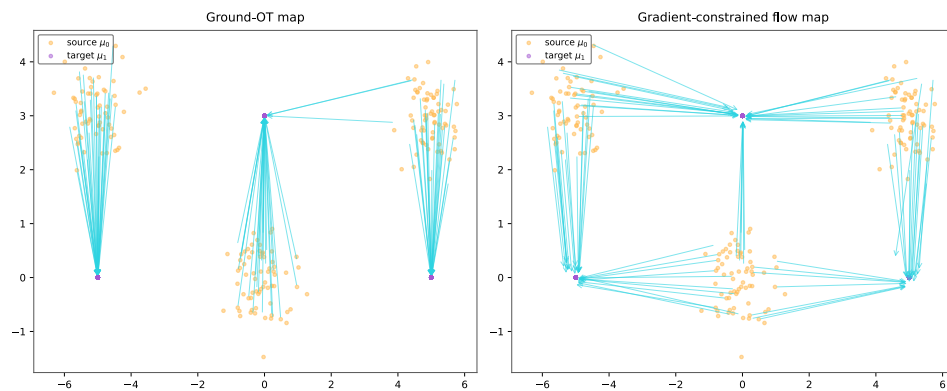
Pour un flot associé à un transport optimal, l'équation de la continuité doit être vérifiée en tout point de temps et d'espace. Par des théorèmes classiques, cela implique en particulier  $v = \nabla_x \varphi$  pour tout  $t$ .

Certains chercheurs se sont basés sur cette propriété pour définir un problème de minimisation contraint, pour lequel on cherche  $\operatorname{argmin}_{v=\nabla\varphi} \mathcal{L}(v)$ . C'est dans ce contexte que [3] présentent un contre exemple dans le cas où le support des distributions n'est pas connecté, ainsi qu'une garantie théorique dans le cas contraire (**Théorème 11**), à savoir qu'une perte nulle équivaut à un transport optimal, pour des distributions rectifiables.

## 2 Expériences

### 2.1 Transport optimal de mixtures de gaussiennes

Le papier montre l'existence d'un contre exemple pour lequel l'algorithme original est censé pouvoir échouer, c'est-à-dire, l'existence d'un champ  $v$  de perte nulle tel que  $v \neq v^*$  le champ de vitesse du transport optimal. Nous nous sommes donc inspiré de cet exemple pour créer deux distributions dont le calcul du transport optimal poserait problème, partant d'une mixture de trois gaussiennes, vers trois diracs.

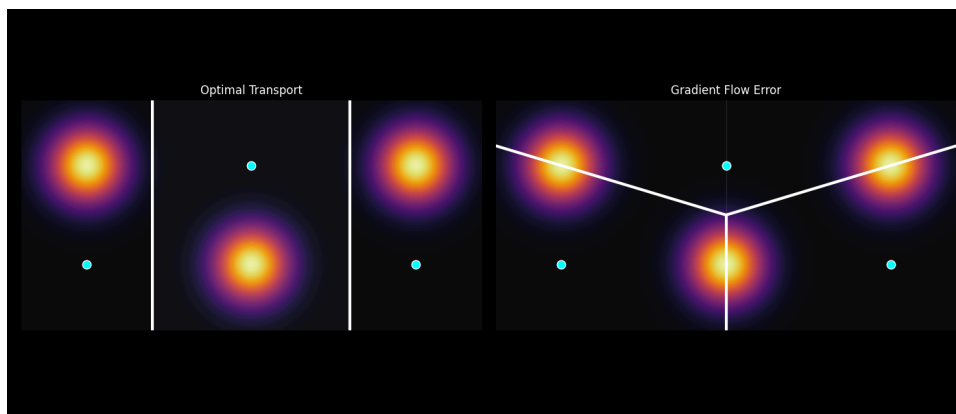


**Fig. 1:**  $\mathcal{N}(0, \text{Id})$  GMM transport.

Ici, une question se pose : Comment vérifier la qualité du transport optimal, si cette quantité est si dure à calculer, et que la perte ne la représente pas en pratique ? Une réponse simple est de calculer le transport optimal non pas sur les distributions, mais directement sur les échantillons obtenu depuis les distributions. Ainsi, il ne faut plus que calculer le transport optimal discret, ce qui se calcule assez facilement de manière exacte tant que le nombre d'échantillons n'est pas trop élevé.

Ainsi, on peut facilement vérifier si l'algorithme génère effectivement un transport approprié. Et assez logiquement, la position des diracs et leur support disjoint génère la aussi un contre exemple à la validité de l'algorithme proposé.

Plutôt que de montrer rigoureusement la validité de ce contre exemple, voyons le plutôt sous la forme d'un dessin. Sont représenté autour de chaque dirac les cellules approximatives auquel les densités sont assignées. Clairement, il est possible pour chaque point de l'espace de le relier au dirac qui se trouve dans la même cellule, sans ambiguïté, ce qui est équivalent à ce que  $\mathcal{L}(v) = 0$ .

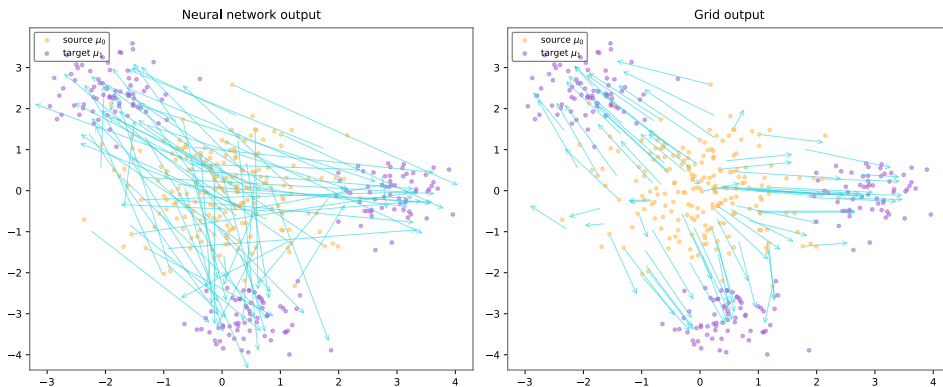


## 2.2 Approximations à l'aide de réseaux de neurones

Le papier affirme que, « dans ces cas, c'est évident que notre but n'est pas de calculer le champ de vitesses exact, mais plutôt de s'appuyer sur l'approximation apportée par les réseaux de neurones, un élément clef qui donne au modèle ses capacités de généralisations ». L'explication immédiate est que le calcul du transport optimal sur des valeurs discrètes (échantillons) retourne exactement les résultats du transport optimal discret. C'est d'ailleurs cette propriété que l'on utilise dans la partie précédente pour vérifier l'exactitude des résultats.

Si on veut pouvoir parler de distributions, il faut pouvoir généraliser à  $v$  valable sur un ensemble continu autour de chaque point, ce que les chercheurs espèrent en utilisant un réseau de neurone pour approximer la fonction de vitesse.

Nous avons donc mis en place un autre modèle à même de généraliser: Une grille de potentiel. Ainsi, on découpe l'espace temps en hypercube dont seuls les points obtiennent un potentiel assigné. Ensuite, le potentiel devient simplement une combinaison linéaire des sommets de l'hypercube. Finalement, on réutilise la même méthode que pour entraîner le perceptron multicouche (la fonction de perte est tout à fait dérivable). Les résultats sont parlants: La fonction de perte est tout à fait similaire à celle d'un réseau de neurone classique, pour une généralisation directe si on fait attention à avoir suffisamment d'échantillons par hypercube.



**Fig. 3:**  $\mathcal{N}(0, Id)$ , GMM transport.

Évidemment, cette méthode est difficilement applicable en très grande dimension, pour laquelle le nombre de cube devient exponentiel, mais pour une dimension relativement faible, l'utilisation d'un réseau de neurone nuis directement à l'interprétabilité de l'algorithme.

### 2.3 Écart entre les méthodes classiques et la rectification régularisée

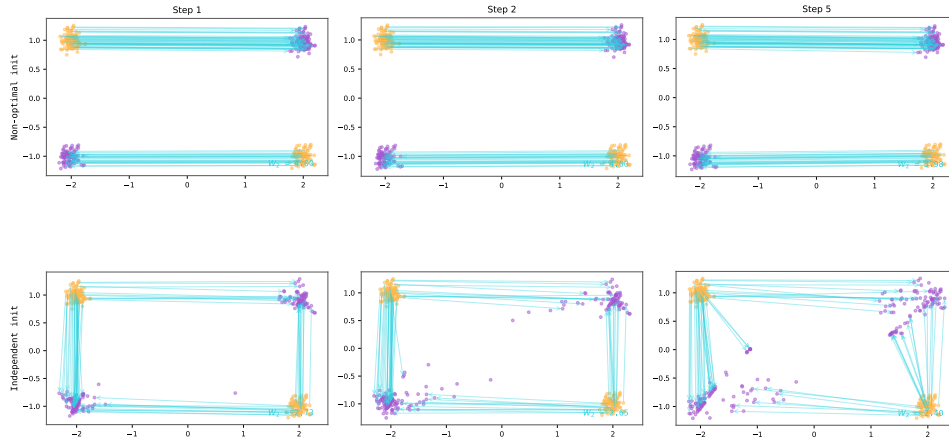
Le papier introduit un contre-exemple à support continu (Section 4.1), où itérer l'algorithme de rectification en utilisant une contrainte de gradient sur la vitesse :  $w_t = \nabla \varphi_t$  pour  $\varphi_t : \mathbb{R}^d \rightarrow \mathbb{R}$ , noté  $\mathcal{R}_p$ , ne converge pas vers le transport optimal. En effet, soit une distribution  $\eta \in \mathcal{P}(\mathbb{R}^2)$ , avec le support  $\{x \in \mathbb{R}^2, \|x\| \leq 0.3\}$ , ainsi que ses versions translatées  $\eta^b = (\cdot + b)_\# \eta$ , les auteurs définissent  $\mu_0$  et  $\mu_1$  de la manière suivante :

$$\mu_0 = \frac{1}{2}(\eta^{-2,1} + \eta^{2,-1}), \quad \mu_1 = \frac{1}{2}(\eta^{-2,-1} + \eta^{2,1}) \quad (2.1)$$

Ainsi que  $X_0 \sim \mu_0$ , et

$$X_1 = \begin{cases} X_0 - (4, 0) & \text{si } (X_0)_2 < -0.5 \\ X_0 + (4, 0) & \text{si } (X_0)_2 > 0.5 \end{cases} \quad (2.2)$$

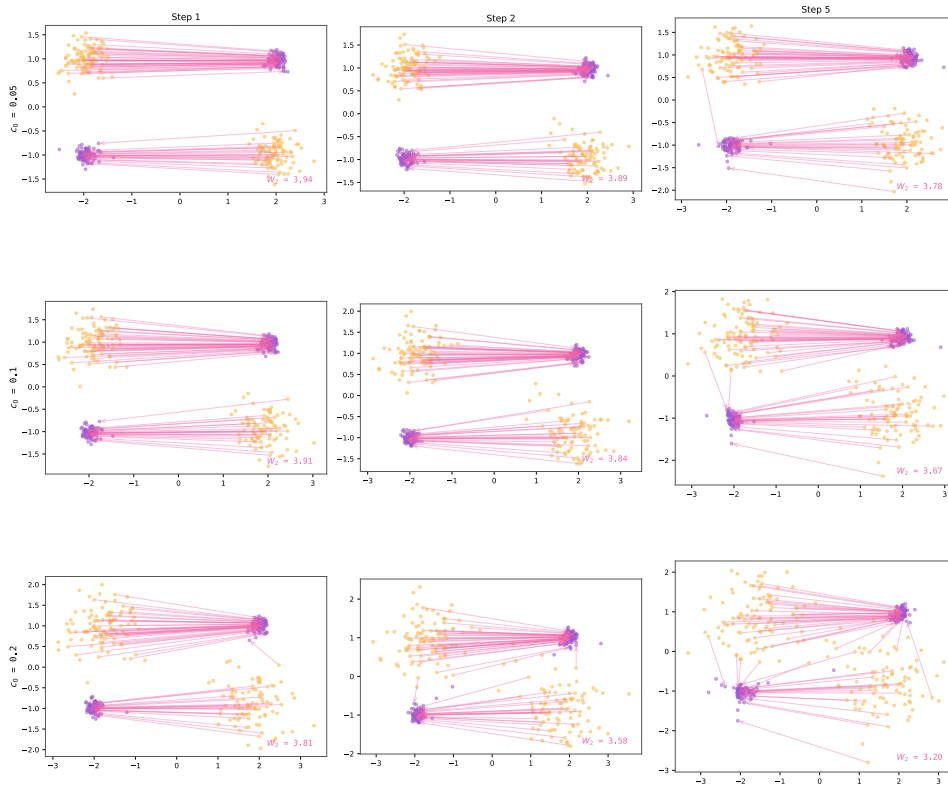
(notés  $\tilde{X}_0$  et  $\tilde{X}_1$  dans le papier). En initialisant la procédure de rectification avec ce couple  $(X_0, X_1)$  on atteint un point fixe, qui n'est cependant pas optimal, contredisant le théorème. Théorème qui reste vrai en supposant un support sur tout  $\mathbb{R}^d$ . Cependant, en partant d'une distribution uniforme pour  $X_0$  et  $X_1$ , le point fixe n'est pas atteint pour le cas non-optimal, et le coût diminue (Fig. 4).



**Fig. 4:** Algorithme de rectification en partant de  $(X_0, X_1)$  définis par Équation (2.2) **en haut**, et en partant d'une distribution uniforme **en bas**.

Pour éviter ce cas, les auteurs introduisent une régularisation avant d'appliquer  $\mathcal{R}$ . Soit un couplage initial  $(Z_0^{(0)}, Z_1^{(0)})$ , ils définissent la séquence de couplages  $(Z_0^{(i)}, Z_1^{(i)})$  comme  $(Z_0^{(i+1)}, Z_1^{(i+1)}) = \mathcal{R}(X_0^{(i)}, X_1^{(i)})$ , avec  $X_0^{(i)} = \sqrt{1 - c_i} Z_0^{(i)} + \sqrt{c_i} W^{(i)}$ , où  $W^{(i)} \sim \mathcal{N}(0, \text{Id})$  indépendant de  $(Z_0^{(i)}, Z_1^{(i)})$ ,  $X_1^{(i)} = Z_1^{(i)}$ , et  $c_i \in ]0, 1[$ . Ils montrent que cette méthode permet d'assurer que les couples  $(X_0^{(i)}, X_1^{(i)})$  restent rectifiables, et proposent même un résultat de convergence, pour  $L^{(i)} = \inf_{w_t} \mathcal{L}(w_t | X_0^{(i)}, X_1^{(i)})$ , et  $\bar{c}_K = \frac{1}{K} \sum_{i=0}^{K-1} c_i$ , on a  $\min_{i=0, \dots, K-1} L^{(i)} \in 0(\frac{1}{K} + \bar{c}_K)$ . C'est à dire, pour un bruit constant :  $c_i = c$ , la valeur minimale tend vers zéro, à une constante près, qui dépend linéairement du bruit  $c$  injecté (ce qui ne garantit pas la convergence vers le plan optimal).

En pratique, on observe que cette méthode de rectification régulariser permet effectivement de sortir du cas non-optimal (Fig. 5), et de faire décroître le coût (Fig. 6).



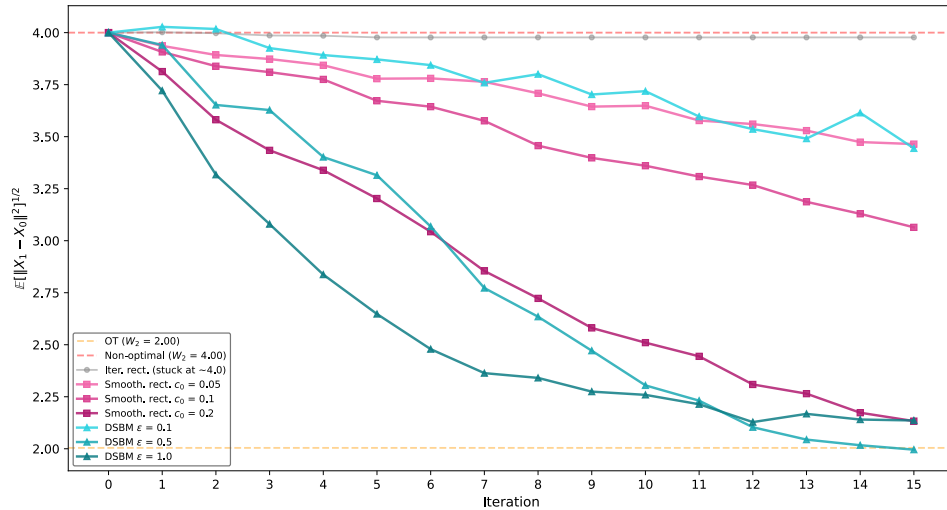
**Fig. 5:** Rectification régularisée en partant de l'initialisation non-optimale, pour différentes valeurs de bruit.

Au lieu de rajouter du bruit dans  $X_0^{(i)}$ , une autre approche introduite par [4], [5], appelée « *Diffusion Schrödinger Bridge Matching* » (DSBM), vise à rajouter du bruit dans l'interpolation entre  $X_0$  et  $X_1$  :

$$X_t = (1-t)X_0 + tX_1 + \sqrt{\varepsilon t(1-t)}Z, \quad Z \sim \mathcal{N}(0, \text{Id}) \quad (2.3)$$

De cette manière, le champ de vitesse  $v_t = \frac{1}{1-t}\mathbb{E}[X_1 - X_t | X_t = x]$  satisfait bien l'équation de Fokker-Planck, et l'on peut générer des points  $X_1$  en résolvant  $dY_t = v_t(Y_t)dt + \sqrt{\varepsilon}dW_t$ , avec  $W_t$  un mouvement brownien. On obtient la rectification  $\mathcal{R}_\varepsilon(X_0, X_1) = (Y_0, Y_1)$ , équivalente à  $\mathcal{R}$  quand  $\varepsilon$  est nul.

Alors que l'on ne sait pas si l'algorithme de rectification converge vers le transport optimal, c'est une question ouverte, on sait que DSBM lui converge, vers le transport optimal régularisé. En pratique, dans nos exemples, le coût diminue même plus vite que la rectification (Fig. 6).



**Fig. 6:** Vitesses de convergence des algorithmes de rectification comparées au SB.

Sur la Fig. 6, on voit que l'algorithme de rectification sans régularisation reste bloqué à la valeur non-optimale 4.00, et que les autres procédures réussissent à faire diminuer le coût, plus ou moins rapidement en fonction du bruit  $\varepsilon$ . Avec un plus grand bruit, la décroissance est plus rapide, mais cela vient à un coup : la convergence étant en  $\mathcal{O}(\frac{1}{K} + \bar{c}_K)$ .

Pour des valeurs d' $\varepsilon$  faibles, les méthodes sont toutes lentes. Il est aussi très intéressant de voir que dans cet exemple, DSBM descend en dessous de la valeur du transport optimal 2.00. En effet, comme l'objectif n'est pas exactement le

transport optimal, mais une version régularisée, DSBM peut atteindre des valeurs inférieures, sans pour autant avoir trouvé le plan optimal.

### 3 Ouverture

Une question majeure posée par le papier est la question de la convergence de l'algorithme de rectification. [2] a prouvé qu'il permet de diminuer le coût à chaque étape, et la régularisation permet de rajouter des garanties, mais il reste la question du point de convergence. Il est possible que ce soit le transport optimal, car, contrairement à DSBM qui converge vers la version régularisée du transport optimal, l'algorithme de rectification ne rajoute pas de terme supplémentaire à la fonction de perte, et pourrait donc converger vers le plan optimal de transport. Si c'est le cas, il serait intéressant d'étudier la vitesse de convergence, le défaut principal de DSBM étant sa vitesse de convergence arbitrairement lente.

Un deuxième point intéressant soulevé par l'article, est d'essayer de généraliser certains résultats, notamment de regarder le cas non-Gaussien, dans l'algorithme de rectification régularisé. De manière similaire, des articles récents sortent du cas gaussien avec DSBM en utilisant un décroissement polynomial avec des processus de Lévy  $\alpha$ -stables. Il serait intéressant de voir les liens entre les différentes méthodes, par exemple, en les reliant toutes à l'équation de Fokker-Planck, et à la façon dont chacune traite le problème.

Finalement, nous pensons intéressant de regarder ce qu'il se passe dans les régions de vide en pratique. Étant donné que l'on utilise des réseaux de neurones, les valeurs ne sont jamais à 0, ce qui peut potentiellement amener des problèmes, ou au contraire, rendre certaines régularisations inutiles, car présentes par défaut. Ces régions peuvent aussi amener des irrégularités topologiques, notamment avec les méthodes de Langevin, qui utilisent du bruit décroissant. À chaque étape, le bruit décroît de manière discrète, ce qui peut faire perdre des propriétés topologiques intéressantes.

### References

- [1] X. Liu, C. Gong, et Q. Liu, « Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow ». [En ligne]. Disponible sur: <https://arxiv.org/abs/2209.03003>
- [2] Q. Liu, « Rectified Flow: A Marginal Preserving Approach to Optimal Transport ». [En ligne]. Disponible sur: <https://arxiv.org/abs/2209.14577>
- [3] J. Hertrich, A. Chambolle, et J. Delon, « On the Relation between Rectified Flows and Optimal Transport ». [En ligne]. Disponible sur: <https://arxiv.org/abs/2505.19712>
- [4] M. S. Albergo, N. M. Boffi, et E. Vanden-Eijnden, « Stochastic Interpolants: A Unifying Framework for Flows and Diffusions ». [En ligne]. Disponible sur: <https://arxiv.org/abs/2303.08797>

- [5] Y. Shi, V. D. Bortoli, A. Campbell, et A. Doucet, « Diffusion Schrödinger Bridge Matching ». [En ligne]. Disponible sur: <https://arxiv.org/abs/2303.16852>

## Appendix

### A Utilisation de l'IA générative

Nous avons utilisé Claude Sonnet pour générer le code des figures, afin d'avoir une mise en forme agréable, ainsi que pour générer le code lié à la résolution des EDO, et du transport optimal avec la librairie POT.

Les prompts étaient basiques (liste de couleurs, demande d'utiliser du typage), et n'ont pas eu besoin d'itération, ou de correction particulière.

Nous n'avons cependant pas utilisé d'IA pour écrire le rapport, car d'expérience, ça amène plus de problèmes que ça n'en résout.