Projection convexe pour garantir l'inégalité triangulaire dans la prédiction de distances phylogénétiques avec *Phyloformer*

Rapport de fin de licence (Bachelor)

Nassim Arifette
Collège de France – CIRB
Équipe « Ecology & Evolution of Health »
Encadrant : Laurent Jacob

2 juillet 2023

Résumé

La prédiction de distances évolutives à partir d'alignements de séquences multiples (MSA) est une étape clé en phylogénie. Les modèles d'apprentissage profond, tels que *Phyloformer*, offrent des prédictions rapides mais ne garantissent pas que les matrices produites respectent l'inégalité triangulaire, propriété fondamentale des métriques. Nous présentons une méthode qui intègre une projection sur le cône métrique directement dans la boucle d'entraînement de *Phyloformer*. L'approche consiste à projeter la matrice de distances prédite sur l'ensemble des matrices respectant l'inégalité triangulaire via un algorithme de *proximité métrique* en norme ℓ_2. Sur un jeu de données de 1000 séquences et 20 époques d'entraînement, l'intégration de la projection garantit les contraintes métriques. Bien que la perte L1 d'entraînement soit légèrement plus élevée avec projection, la topologie des arbres reconstruits par *Neighbor-Joining* (NJ), mesurée par la distance de Robinson–Foulds (RF), reste comparable au modèle de base. Cette étude montre la faisabilité d'imposer des contraintes géométriques dures pendant l'entraînement sans dégrader significativement la performance topologique, tout en produisant des matrices de distances théoriquement cohérentes. Les limites incluent un entraînement non convergé, un unique jeu de données et l'absence d'accès au code et aux données originales pour reproduction.

1 Introduction

L'estimation de phylogénies, qui décrivent les relations évolutives entre espèces, repose traditionnellement sur des méthodes coûteuses comme le maximum de vraisemblance ou l'inférence bayésienne ([BV04, SN87]). Des approches d'apprentissage profond, notamment basées sur l'architecture Transformer ([VSP$^+$23]), ont émergé comme alternatives rapides pour prédire des matrices de distances utilisables par des algorithmes combinatoires tels que *Neighbor-Joining* (NJ) ([SN87]). Le modèle *Phyloformer* prédit une matrice de distances par paires à partir d'un MSA, qui est ensuite convertie en arbre via NJ.

Un défi majeur est que la matrice prédite $\hat{D}$ ne respecte pas toujours l'inégalité triangulaire, violant ainsi les axiomes métriques et pouvant nuire à la stabilité de NJ et à l'interprétabilité. Nous proposons d'intégrer une projection sur le *cône métrique* pendant l'entraînement pour forcer la cohérence métrique.

Contributions. (i) Formalisation de l'intégration d'une projection sur le cône métrique dans la boucle d'entraînement d'un modèle profond; (ii) comparaison empirique entre *Phyloformer* de base et une version avec projection *metric-nearness* ℓ_2; (iii) évaluation de l'impact sur la perte et sur la qualité topologique (RF) après NJ.

2 Contexte et travaux liés

2.1 Phylogénie et reconstruction d'arbres

Un arbre phylogénétique modélise les liens de parenté entre entités. L'algorithme *Neighbor-Joining* reconstruit un arbre à partir d'une matrice de distances en fusionnant itérativement des paires de taxons afin de minimiser la longueur totale de l'arbre ([SN87]). La qualité topologique peut être comparée via la distance de Robinson–Foulds (RF), qui compte les bipartitions présentes dans un arbre mais absentes dans l'autre ([**?**]).

2.2 Apprentissage profond pour la phylogénie

Phyloformer, inspiré des Transformers ([VSP+23]), prend un MSA en entrée et prédit une matrice de distances par paires. Des observations empiriques montrent que ces matrices peuvent violer l'inégalité triangulaire, d'où la nécessité d'une correction.

2.3 Optimisation convexe et projection

Rendre une matrice conforme à l'inégalité triangulaire relève du problème de *proximité métrique* (*metric nearness*) ([BDST08]). L'ensemble des matrices métriques forme un cône convexe. Des algorithmes itératifs comme POCS ([BB96, BD03]) ou l'algorithme de Dykstra ([BD86]) projettent successivement sur des sous-ensembles convexes; Dykstra garantit la convergence vers la projection sur l'intersection. Nous adoptons une variante en norme ℓ_2 proche de [BDST08].

3 Problème et formulation mathématique

Soit un MSA de n séquences. Le modèle f_θ prédit $\hat{D} \in \mathbb{R}^{n\times n}$, symétrique à diagonale nulle, pour approximer une matrice de distances de référence $D^\star$.

Définition 1 (Cône métrique)**.** *L'ensemble des matrices de distances métriques sur n points est*

$$\mathcal{M} = \left\{ D \in \mathbb{R}^{n\times n}_{\text{sym}} \;\middle|\; D_{ii} = 0,\ D_{ij} \geq 0,\ D_{ij} \leq D_{ik} + D_{kj},\ \forall i,j,k \in \{1,\dots,n\} \right\}.$$

Projection (proximité métrique). Pour la norme de Frobenius,

$$\mathrm{Proj}_{\mathcal{M}}(\hat{D}) \;=\; \arg\min_{M \in \mathcal{M}} \; \tfrac{1}{2}\,\|M - \hat{D}\|_F^2.$$

La perte d'entraînement est calculée sur la matrice projetée $\tilde{D} = \mathrm{Proj}_{\mathcal{M}}(\hat{D})$:

$$\mathcal{L}(\theta) \;=\; \|\tilde{D} - D^\star\|_1 \;=\; \sum_{i,j} \big|\tilde{D}_{ij} - D^\star_{ij}\big|.$$

Dans cette étude, la projection est traitée comme une étape fixe non différentiable (pas de rétropropagation à travers $\mathrm{Proj}_{\mathcal{M}}$).

4 Méthodes

4.1 Algorithmes de projection

POCS. Projection séquentielle sur des demi-espaces définis par les inégalités triangulaires; sensible à l'ordre et potentiellement lente ([BB96, BD03]).
Dykstra. Variante de POCS avec termes correctifs (variables auxiliaires) qui assure la convergence vers la projection sur l'intersection ([BD86]).
Proximité métrique ℓ_2. Nous adoptons une méthode proche de [BDST08] pour projeter globalement sur le cône métrique sous ℓ_2. Empiriquement, une seule itération corrige souvent la majorité des violations sur de petits exemples.

4.2 Intégration dans l'entraînement

La projection est appliquée à chaque passe avant, avant le calcul de la perte, avec une seule itération par étape pour des raisons de coût. Le schéma est donné en Algorithm 1.

Input: lots d'alignements X, matrices cibles $D^\star$, modèle f_θ, opérateur de projection $\Pi_{\mathcal{M}}$
for *chaque lot* $(X, D^\star)$ **do**
  $\hat{D} \leftarrow f_\theta(X)$; `// sortie symétrique, diagonale nulle`
  $\tilde{D} \leftarrow \Pi_{\mathcal{M}}(\hat{D})$; `// projection (1 itération)`
  $\mathcal{L} \leftarrow \|\tilde{D} - D^\star\|_1$; `// perte L1`
  $\theta \leftarrow \theta - \eta\,\nabla_\theta \mathcal{L}$; `// pas de rétroprop à travers` $\Pi_{\mathcal{M}}$
end

Algorithm 1: Entraînement avec projection sur le cône métrique

5 Protocole expérimental

5.1 Données et matériel

Entraînement sur un jeu de 1000 séquences génétiques. Matériel : GPU Nvidia Tesla V100 (32 GB). Chaque configuration d'entraînement dure environ 30 min.

5.2 Hyperparamètres

— **Époques** : 20.

— **Taux d'apprentissage** : 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}.
— **Taille de lot** : 8.
— **Perte** : L1.
— **Optimiseur** : Adam ([KB17]).
— **Projection** : 1 itération par étape.

5.3 Métriques d'évaluation

1. Perte L1 entre la matrice (projetée ou non) et $D^\star$.
2. Distance de Robinson–Foulds (RF) entre l'arbre NJ issu de la matrice prédite et l'arbre de référence (normalisée si applicable).

6 Résultats

Perte d'entraînement. figure 1 montre que la perte est légèrement plus élevée avec projection, ce qui est attendu car l'espace des solutions est restreint au cône métrique.

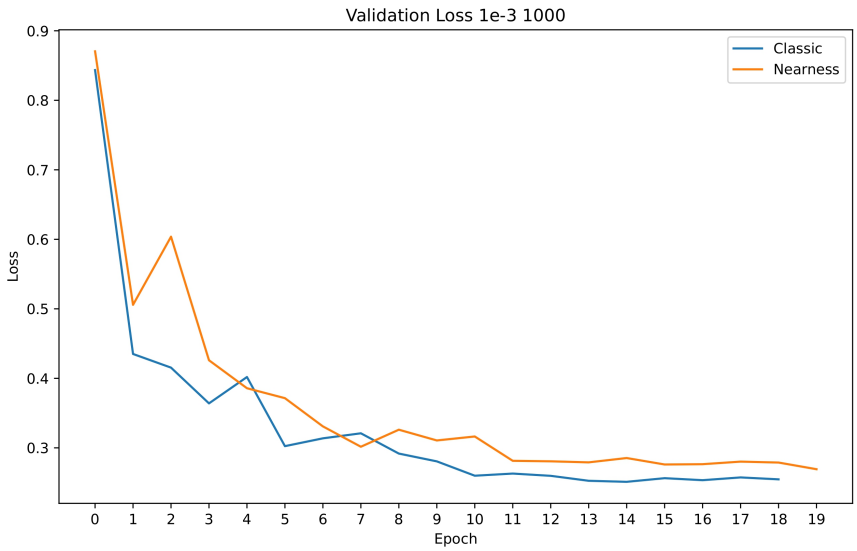

FIGURE 1 – Perte d'entraînement avec (orange) et sans (bleu) projection.

Qualité topologique. figure 2 indique que la RF normalisée est similaire avec et sans projection, pour des tailles d'arbres variables.

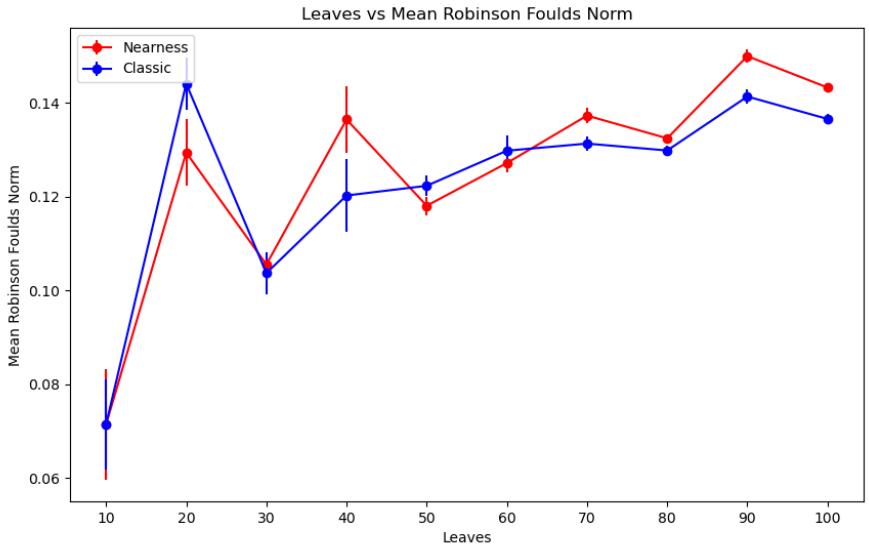

FIGURE 2 – Distance RF normalisée en fonction du nombre de feuilles.

Erreur relative sur la matrice. figure 3 suggère des erreurs relatives comparables, tout en garantissant la métricité avec projection.

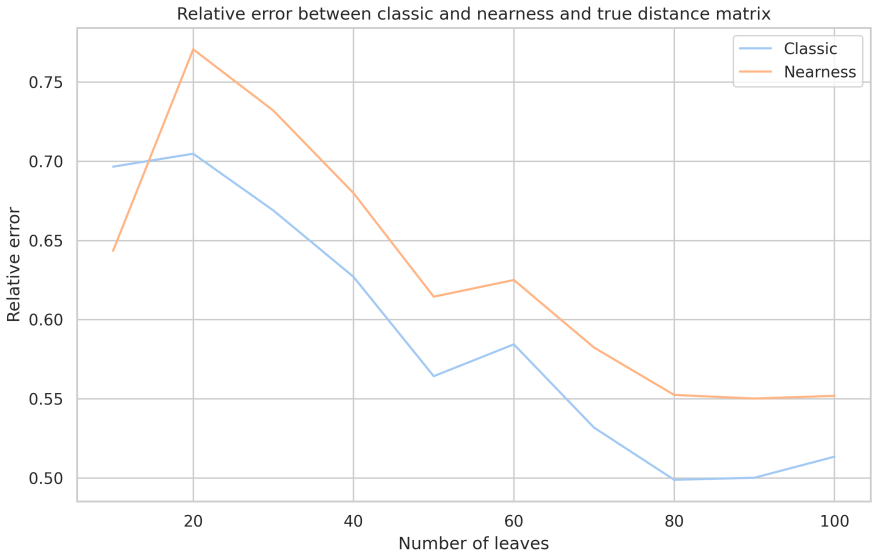

FIGURE 3 – Erreur relative entre matrices prédites et de référence, avec et sans projection.

Synthèse. tableau 1 résume qualitativement l'ablation.

TABLE 1 – Ablation : sans vs avec projection (résultats qualitatifs).

Métrique	Sans projection	Avec projection
Inégalité triangulaire	Possibles violations	Respect garanti
Perte L1 (finale)	Plus faible	Légèrement plus élevée
Distance RF normalisée	Similaire	Similaire
Erreur relative	Similaire	Similaire

7 Discussion

La projection sur le cône métrique pendant l'entraînement garantit la cohérence théorique des distances au prix d'une légère hausse de la perte L1. L'absence d'impact sur la distance RF suggère que les violations initiales étaient de faible ampleur ou peu influentes pour NJ. Cette contrainte peut améliorer la robustesse pour des analyses en aval qui requièrent des distances strictement métriques.

8 Limites, reproductibilité et éthique

— **Convergence** : 20 époques seulement; des entraînements plus longs sont nécessaires.
— **Données** : un seul jeu; généralisation non évaluée.
— **Reproductibilité** : code et données du stage non accessibles; hyperparamètres rapportés pour réimplémentation.
— **Stochasticité** : graine unique (non consignée); variance non estimée.
— **Aspects éthiques** : les séquences biologiques devraient provenir de sources publiques ou autorisées; conformité institutionnelle supposée.

9 Conclusion et travaux futurs

Nous avons montré la faisabilité d'intégrer une projection sur le cône métrique dans l'entraînement de *Phyloformer*. Les matrices prédites deviennent métriques, sans dégradation notable de la qualité topologique mesurée par la distance RF. Perspectives : (i) appliquer la projection à l'inférence uniquement; (ii) entraîner jusqu'à convergence et sur plusieurs graines; (iii) étudier l'effet du nombre d'itérations de projection et d'autres normes (ℓ_1, pondérations).

Remerciements

Je remercie Laurent Jacob (encadrant) pour son accompagnement, ainsi que Luc Blassel, Luca Nesterenko et Bastien Boussau pour leurs échanges. Merci à l'équipe « Ecology & Evolution of Health » du CIRB (Collège de France) et au Magistère d'Informatique de l'Université Paris-Saclay pour leur soutien.

Références

[BB96] Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

[BD86] James P. Boyle and Richard L. Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In Richard Dykstra, Tim Robertson, and Farroll T. Wright, editors, *Advances in Order Restricted Statistical Inference*, pages 28–47, New York, NY, 1986. Springer New York.

[BD03] Stephen Boyd and Jon Dattoro. Alternating projections, 2003.

[BDST08] Justin Brickell, Inderjit S. Dhillon, Suvrit Sra, and Joel A. Tropp. The Metric Nearness Problem. *SIAM Journal on Matrix Analysis and Applications*, 30(1):375–396, January 2008.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[KB17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[SN87] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 07 1987.

[VSP+23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

A Annexes

A.1 Définitions en biologie

Définition 2 (Phylogénie)**.** *Étude des relations de parenté entre êtres vivants et de l'évolution des espèces.*

Définition 3 (Arbre phylogénétique)**.** *Arbre représentant les liens de parenté; les nœuds internes sont des ancêtres communs.*

Définition 4 (Séquence génétique)**.** *Suite de nucléotides (A, C, G, T) ou d'acides aminés représentant l'information génétique.*

Définition 5 (Alignement de séquences)**.** *Mise en correspondance de positions de plusieurs séquences pour révéler similarités et divergences.*

Définition 6 (Distance de Robinson–Foulds)**.** *Mesure de dissimilarité topologique entre deux arbres, basée sur les bipartitions ([?]).*

A.2 Définitions mathématiques

Définition 7 (Ensemble convexe)**.** *Un ensemble C est convexe si $\lambda x + (1-\lambda)y \in C$ pour tous $x, y \in C$ et $\lambda \in [0, 1]$.*

Définition 8 (Projection sur un convexe)**.** *Pour un convexe fermé $C \subset \mathbb{R}^m$, la projection d'un point x est l'unique $P_C(x) \in C$ minimisant $\|x - y\|_2$ sur $y \in C$.*

Définition 9 (POCS)**.** *Projection séquentielle sur des ensembles convexes; la convergence à la projection sur l'intersection n'est pas garantie et l'ordre des projections compte ([BB96, BD03]).*

Définition 10 (Algorithme de Dykstra)**.** *Variante de POCS utilisant des termes correctifs assurant la convergence vers la projection sur l'intersection d'ensembles convexes ([BD86]).*

A.3 Table de notation

TABLE 2 – Principales notations

Symbole	Description
n	Nombre de taxons (feuilles)
f_θ	Modèle *Phyloformer* paramétré par θ
$\hat{D} \in \mathbb{R}^{n \times n}$	Matrice de distances prédite (symétrique, diag nulle)
$D^\star$	Matrice de distances de référence
$\mathcal{M}$	Cône des matrices métriques
$\text{Proj}_{\mathcal{M}}$	Projection sur $\mathcal{M}$ (proximité métrique)
RF	Distance de Robinson–Foulds
NJ	Algorithme *Neighbor-Joining*

A.4 Exemple jouet de correction de l'inégalité triangulaire

Considérons trois points avec $\hat{D}_{12} = 1$, $\hat{D}_{23} = 1$, $\hat{D}_{13} = 3$ (violation : $3 > 1+1$). Une projection ℓ_2 simple consiste à réduire $\hat{D}_{13}$ à 2 (ou à répartir l'ajustement sur plusieurs entrées selon la méthode), ce qui satisfait l'inégalité triangulaire. Cet exemple illustre l'effet de la projection; l'algorithme global gère les $\mathcal{O}(n^3)$ contraintes simultanément.

A.5 Figures supplémentaires

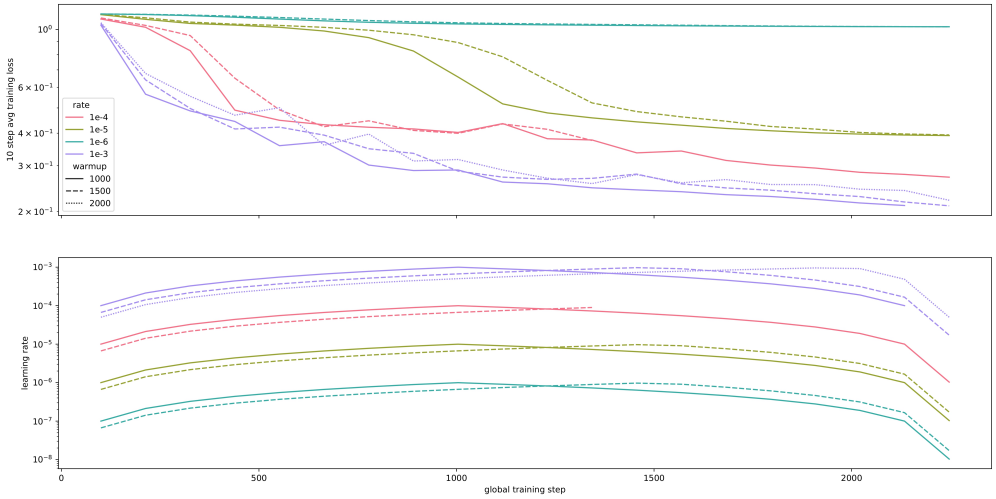

FIGURE 4 – Évolution de la perte sans projection.

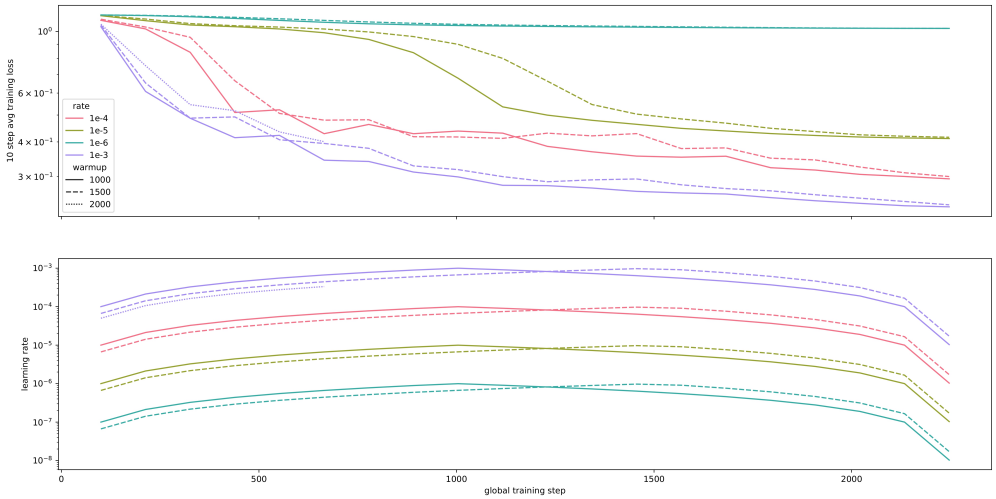

FIGURE 5 – Évolution de la perte avec projection.