Introduction

Le projet "Raoul", depuis sa création, vise à explorer en profondeur les mécanismes internes des modèles de traitement du langage, en particulier les « Large Language Models » (LLM), dont les performances remarquables posent néanmoins des défis en termes d'interprétabilité et de compréhension des processus qui génèrent leurs résultats.

Pour l'année 2024-2025, le projet mettra l'accent sur l'investigation des espaces latents au travers de nouvelles approches en mathématiques pures. En effet, comprendre la structure et la dynamique de ces espaces reste essentiel pour améliorer l'interprétabilité des LLM. Le programme de cette période se concentrera ainsi sur des pistes prometteuses pour avancer dans la compréhension de ces modèles complexes, tout en posant des jalons théoriques pour la suite des travaux.

Les réflexions et propositions qui suivent s'inscrivent dans cette démarche et visent à contribuer à cet objectif central.

Partie I : Approche avec la géométrie différentielle

La géométrie différentielle fournit des outils mathématiques puissants pour analyser des structures complexes, en particulier lorsque ces structures sont représentées dans des espaces à plusieurs dimensions. Dans le cadre de l'intelligence artificielle et de l'apprentissage automatique, les espaces latents, qui représentent des caractéristiques abstraites de données, sont souvent non linéaires et possèdent des propriétés géométriques difficiles à appréhender. En appliquant la géométrie différentielle à ces espaces latents, on peut mieux comprendre la manière dont les modèles apprennent et structurent les informations. Cette approche permet d'explorer les relations entre les points de données, la courbure de l'espace latent et les transformations locales. Finalement, l'utilisation de ces outils constitue une tentative pour percer la nature des représentations internes dans ces espaces, offrant ainsi de nouvelles perspectives pour améliorer les modèles et leur interprétation.

1. Espace latent :

Soit $\mathcal{L}$ un espace latent, représenté comme un espace vectoriel de dimension d, c'est-à-dire $\mathcal{L} \subseteq \mathbb{R}^d$.

Chaque entité textuelle t_i (comme un mot, une phrase ou un document) peut être représentée par un vecteur $z_i \in \mathcal{L}$.

Ainsi, une entité textuelle t_i est associée à un point z_i dans l'espace latent $\mathcal{L}$, où $z_i = f(t_i)$, avec f une fonction d'encodage qui projette une entité textuelle dans cet espace latent.

2. Variété différentielle :

L'espace latent $\mathcal{L}$ peut être vu comme une variété différentielle $\mathcal{M}$, c'est-à-dire un espace qui localement ressemble à un espace euclidien mais qui globalement peut avoir une structure plus complexe.

Formellement, une variété $\mathcal{M}$ est un ensemble muni d'une structure différentielle, permettant la dérivation. Chaque point $z_i \in \mathcal{M}$ est une entité textuelle dans cet espace. En général, nous considérons que la variété $\mathcal{M}$ est plongée dans un espace $\mathbb{R}^d$, donc $\mathcal{M} \subset \mathbb{R}^d$.

3. Transformations et apprentissage :

Lors de l'apprentissage des modèles de langage, des transformations $\mathrm{T} : \mathcal{M} \to \mathcal{M}$ sont appliquées aux points dans l'espace latent. Ces transformations peuvent être vues comme des déplacements sur la variété.

Mathématiquement, une transformation T est une fonction différentiable qui modifie les coordonnées d'une entité textuelle dans l'espace latent :

$$z_i' = \mathrm{T}(z_i)$$

où T est un difféomorphisme (transformation différentiable inversible) de la variété $\mathcal{M}$, assurant que la nouvelle position z_i' reste sur la variété.

4. Distance et géométrie :

La géométrie de la variété $\mathcal{M}$ peut être mesurée en utilisant une métrique g, qui définit la distance entre deux points z_i et z_j sur la variété.

La distance géodésique entre z_i et z_j , notée $d_{\mathcal{M}}(z_i , z_j)$, représente la longueur du chemin le plus court sur la variété reliant ces deux points.

$$d_{\mathcal{M}}(z_i , z_j) = \boldsymbol{inf}_{\gamma} \int_0^1 \left(\sum_{k=1}^{d} \left(\frac{d\gamma^k(t)}{dt} \right)^2 \right)^{1/2} dt$$

où $\gamma : [0,1] \to \mathcal{M}$ est une courbe différentiable reliant z_i à z_j sur la variété,
avec $\gamma(0) = z_i$, $\gamma(1) = z_j$ et g la métrique euclidienne.

5. Note mathématique :

- $\mathcal{L} \subseteq \mathbb{R}^d$: espace latent vectoriel.
- $\mathcal{M} \subset \mathbb{R}^d$: variété différentielle représentant cet espace latent.
- $f(t_i) = z_i$: fonction d'encodage plaçant les entités textuelles dans cet espace.
- $\mathrm{T} : \mathcal{M} \to \mathcal{M}$: transformation appliquée lors de l'apprentissage.
- g : métrique définissant la géométrie de la variété $\mathcal{M}$.

6. Métrique riemannienne et produit scalaire :

Sur une variété différentielle $\mathcal{M}$, on peut définir une métrique riemannienne g, qui associe à chaque point $z \in \mathcal{M}$ un produit scalaire dans l'espace tangent $T_z\mathcal{M}$ à la variété en ce point.

Ce produit scalaire est défini localement dans l'espace tangent $T_z\mathcal{M}$ par une forme bilinéaire :

$$g_z(v, w): T_z\mathcal{M} \times T_z\mathcal{M} \to \mathbb{R}$$

pour deux vecteurs tangents v, w $\in T_z\mathcal{M}$.

Dans le cas particulier où $\mathcal{M}$ est plongée dans un espace euclidien $\mathbb{R}^d$, la métrique riemannienne g_z peut être induite par le produit scalaire euclidien dans $\mathbb{R}^d$. Autrement dit, si v, w $\in T_z\mathcal{M}$, alors le produit scalaire induit est simplement le produit scalaire euclidien :

$$g_z(v, w) = v.w = \sum_{i=1}^{d} v_i.w_i$$

7. Définition de l'angle :

L'angle θ entre deux vecteurs tangents v, w $\in T_z\mathcal{M}$ en un point $z \in \mathcal{M}$ peut être défini à partir de ce produit scalaire.

L'angle est donné par la formule classique du produit scalaire :

$$\cos(\theta) = \frac{g_z(v, w)}{\|v\|.\|w\|}$$

où $\|v\| = \sqrt{g_z(v, v)}$ est la norme de v et $\|w\| = \sqrt{g_z(w, w)}$ est la norme de w, toutes deux calculées à partir de la métrique riemannienne g_z .

Ainsi, l'angle θ entre v et w est donné par :

$$\theta = arccos\left(\frac{g_z(v, w)}{\|v\|.\|w\|}\right)$$

8. Utilisation du produit scalaire pour l'étude de la variété :

Le produit scalaire et la notion d'angle permettent d'étudier plusieurs aspects géométriques de la variété :

- **Alignement des vecteurs tangents** : si deux vecteurs tangents v et w ont un angle proche de 0, cela signifie qu'ils pointent dans des directions similaires sur la variété. Inversement, si l'angle est proche de $\frac{\pi}{2}$, cela indique que les directions sont presque orthogonales.
- **Détection des courbures locales** : la comparaison des angles entre vecteurs tangents dans différentes régions de la variété permet d'étudier la courbure locale. Par exemple, sur une variété plate, les angles entre vecteurs parallèles restent constants, tandis que sur une variété courbée, ces angles peuvent varier.
- **Variation des transformations** : dans le cadre de l'apprentissage des modèles, l'effet des transformations $\mathrm{T} : \mathcal{M} \rightarrow \mathcal{M}$ peut être mesuré en observant comment ces transformations modifient les angles entre les vecteurs tangents. Une transformation qui conserve les angles est appelée isométrique, tandis qu'une transformation qui les modifie pourrait signifier une modification significative de la géométrie locale.

9. Note mathématique :

En résumé, nous avons :

- Un espace latent $\mathcal{L} \subseteq \mathbb{R}^d$, qui est une variété différentielle $\mathcal{M}$.
- Chaque point $z \in \mathcal{M}$ possède un espace tangent $\mathrm{T}_z\mathcal{M}$, où une métrique g_z définit un produit scalaire.
- L'angle θ entre deux vecteurs tangents v, w $\in \mathrm{T}_z\mathcal{M}$ est donné par :

$$\theta = arccos\left(\frac{g_z(v,w)}{\|v\|.\|w\|}\right)$$

Où $g_z($v, w$)$ est le produit scalaire dans l'espace tangent $\mathrm{T}_z\mathcal{M}$ défini par la métrique riemannienne.

Cette formulation permet d'exploiter à la fois la notion de distance géodésique et d'angle pour étudier les transformations sur la variété et la géométrie locale de l'espace latent lors de l'apprentissage.

10. Courbure de Ricci :

Nous décidons d'introduire plusieurs concepts de la géométrie riemannienne, notamment la courbure de Ricci. Cette courbure nous permet de caractériser la manière dont la géométrie locale de la variété se comporte et de comprendre comment les transformations dans l'espace latent affectent la répartition des points.

La courbure de Ricci est une contraction du tenseur de courbure de Riemann. Elle mesure la manière dont les volumes de petites régions géodésiques de la variété se comportent sous l'effet de la courbure. Intuitivement, la courbure de Ricci capture la tendance d'une région locale de la variété à "concentrer" ou "disperser" les points.

Formellement, si $\mathcal{M}$ est une variété riemannienne avec une métrique g, alors la courbure de Ricci Ric est un tenseur de type (0,2), qui à chaque point $z \in \mathcal{M}$ associe une forme bilinéaire sur l'espace tangent $\mathrm{T}_z\mathcal{M}$.

Le tenseur de Ricci Ric peut être vu comme la trace du tenseur de Riemann R sur deux de ses indices :

$$\mathrm{Ric}(\mathrm{u}, \mathrm{w}) = \sum_{i=1}^{dim(\mathrm{M})} R(v, e_i, w, e_i)$$

où $\{e_i\}$ est une base orthonormée locale de l'espace tangent $\mathrm{T}_z\mathcal{M}$, et R est le tenseur de courbure de Riemann, qui mesure la courbure de la variété en fonction du déplacement des vecteurs tangentiels le long des géodésiques.

11. Interprétation de la courbure de Ricci dans l'espace latent :

La courbure de Ricci dans un espace latent peut être interprétée comme une mesure de la tendance des points à se concentrer ou à se disperser dans cet espace :

- **Courbure positive** $\mathrm{Ric} > 0$: une courbure de Ricci positive indique que les géodésiques locales convergent, ce qui signifie que les points dans cette région de la variété ont tendance à se concentrer. Cela pourrait refléter une structure locale où les entités textuelles (mots, phrases, documents) deviennent plus similaires ou plus proches dans l'espace latent
- **Courbure négative** $\mathrm{Ric} < 0$: une courbure de Ricci négative signifie que les géodésiques locales divergent, indiquant une dispersion des points. Dans un contexte de modèle de langage, cela pourrait indiquer que les transformations entraînent un étalement des points dans l'espace latent, créant des différences plus marquées entre les entités textuelles.

12. Exemple d'analyse de la courbure en fonction des transformations :

Lors de l'apprentissage d'un modèle de langage, une transformation $\mathrm{T} : \mathcal{M} \to \mathcal{M}$ agit sur les points z_i de la variété $\mathcal{M}$, ce qui peut changer la géométrie locale, y compris la courbure de Ricci.

Soit $z \in \mathcal{M}$ un point donné, et soit $\{v_1, v_2, \ldots, v_n\}$ un ensemble de vecteurs tangents dans l'espace tangent $\mathrm{T}_z\mathcal{M}$. La courbure de Ricci $\mathrm{Ric}(v, v)$, où v est un vecteur tangent, décrit la façon dont le volume de la variété dans une petite boule autour de z évolue sous l'effet de la courbure.

12.1 Dispersion des points - Courbure négative :

Si $\mathrm{Ric}(v, v) < 0$ dans une région de l'espace latent, cela signifie que les points textuels tendent à se disperser sous l'effet des transformations T, les géodésiques divergeant dans cette région. Cela se traduit mathématiquement par une dilatation du volume dans cette région, c'est-à-dire que le volume d'une petite boule géodésique autour d'un point augmente plus rapidement que dans un espace plat.

Formellement, pour une petite région géodésique de rayon r centrée sur z le volume est donné par :

$$Vol\big(B_r(z)\big) \approx cr^n \left(1 + \frac{1}{6}\mathrm{Ric}(z)r^2\right)$$

où c est une constante et n est la dimension de la variété. Si $\mathrm{Ric}(z) < 0$, le volume augmente plus rapidement, indiquant une dispersion.

Cette dispersion pourrait refléter une plus grande variété dans les représentations latentes, ce qui peut pousser le LLM à fournir des réponses plus spécifiques et concises. En effet, chaque représentation textuelle est éloignée, ce qui pourrait suggérer un contenu distinct et non ambigu.

Conjectures :

- Les questions ou prompts traitant de sujets bien distincts génèrent des représentations fortement séparées, poussant le modèle à donner des réponses plus nettes et directes.
- Les transformations de l'espace latent favorisent une disjonction entre concepts similaires, ce qui améliore la précision et réduit l'ambiguïté.

12.2 Concentration des points - Courbure positive :

En revanche, si $\mathrm{Ric}(v, v) > 0$, les géodésiques dans cette région de l'espace latent convergent, et le volume local autour de z tend à se contracter. Les points textuels dans l'espace latent se rapprochent les uns des autres, indiquant une plus grande similarité ou une concentration de l'information.

Cela peut conduire le LLM à produire des réponses plus générales et laconiques. Les représentations similaires sont rapprochées, ce qui pourrait rendre difficile pour le modèle de différencier des concepts proches, le poussant à une réponse plus vague ou synthétique.

Conjectures :

- Des sujets connexes ou ambigus génèrent des représentations similaires, menant à des réponses générales ou moins spécifiques.
- Les transformations compriment les représentations latentes, réduisant la capacité du modèle à distinguer des nuances subtiles, ce qui entraîne des réponses plus courtes et plus synthétiques.

13. Note mathématique :

L'analyse de la courbure via la courbure de Ricci permet d'étudier la manière dont les transformations dans l'espace latent affectent la géométrie locale :

- **Courbure de Ricci positive** $\mathrm{Ric} > 0$: indique une concentration des points dans l'espace latent, ce qui peut signifier que les entités textuelles deviennent plus proches les unes des autres, reflétant une certaine homogénéité locale dans l'espace latent.
- **Courbure de Ricci négative** $\mathrm{Ric} < 0$: indique une dispersion des points, traduisant une hétérogénéité ou un étalement dans l'espace latent, où les entités textuelles tendent à se séparer.

En utilisant la courbure de Ricci, on peut donc comprendre comment les transformations au sein du modèle de langage affectent la structure géométrique de l'espace latent, et comment cela se traduit par une concentration ou une dispersion des points.

Partie II : Approche avec la géométrie algébrique

La géométrie algébrique, qui étudie les solutions d'équations polynomiales et leurs structures géométriques, offre des outils précieux pour analyser des espaces complexes. Dans le contexte des espaces latents utilisés en intelligence artificielle, ces représentations internes sont souvent décrites par des relations non linéaires que la géométrie algébrique permet de formaliser et de mieux comprendre. En utilisant ces méthodes, on peut explorer les structures cachées et les symétries présentes dans ces espaces, révélant ainsi des aspects essentiels du fonctionnement des modèles. Cette approche constitue une tentative pour approfondir la compréhension des espaces latents et des représentations qu'ils encapsulent.

Nous devons d'abord comprendre quelques concepts clés de la topologie algébrique, notamment la théorie des nœuds et comment cela peut s'appliquer à l'analyse des représentations internes dans un espace latent.

1. Théorie des nœuds :

La théorie des nœuds est une branche de la topologie qui étudie les propriétés des courbes fermées (ou lacets) dans l'espace tridimensionnel $\mathbb{R}^3$. Un nœud est formellement une immersion continue d'un cercle S^1 dans $\mathbb{R}^3$, c'est-à-dire une fonction injective et continue $K : S^1 \to \mathbb{R}^3$, où S^1 est le cercle unité.

Deux nœuds sont dits équivalents s'ils peuvent être transformés l'un en l'autre par une isotopie, c'est-à-dire une déformation continue de $\mathbb{R}^3$ qui évite toute auto-intersection du nœud. La théorie des nœuds se concentre sur l'étude des invariants topologiques qui restent constants lors de ces transformations.

2. Représentations des modèles de langage et enchevêtrement :

Dans le contexte des modèles de langage, les représentations internes des entités textuelles (comme des mots, phrases ou documents) peuvent être considérées comme des trajectoires complexes dans un espace latent $\mathcal{L} \subseteq \mathbb{R}^d$, qui est généralement de très grande dimension d.

Ces trajectoires peuvent être vues comme des courbes enchevêtrées dans cet espace. En particulier, lorsque les transformations appliquées dans le cadre de l'apprentissage des modèles modifient ces trajectoires, elles peuvent introduire des structures complexes qui rappellent les nœuds en topologie. L'idée est donc d'analyser ces enchevêtrements en utilisant des outils de la théorie des nœuds pour comprendre la structure sous-jacente de l'espace latent et ses invariants topologiques.

3. Invariants topologiques des représentations internes :

Un invariant topologique est une propriété d'un espace topologique qui reste constante sous des déformations continues (comme une isotopie). Dans la théorie des nœuds, les invariants les plus couramment utilisés sont :

- **Le polynôme de Jones** : un invariant qui associe à chaque nœud un polynôme qui reste constant sous isotopie.
- **Le nombre de croisement minimal** : le nombre minimum de points de croisement dans une projection d'un nœud sur un plan.
- **La classe d'homologie** : une manière de classer les nœuds en fonction de leurs propriétés topologiques dans des espaces de dimensions supérieures.

Dans le contexte d'un modèle de langage, ces invariants peuvent être adaptés pour mesurer la complexité des trajectoires latentes des représentations internes.

3.1. Invariant topologique pour l'espace latent :

Considérons un ensemble de trajectoires dans l'espace latent $\mathcal{L}$, qui représentent l'évolution des représentations des entités textuelles au fur et à mesure qu'elles passent par les différentes couches du modèle.

Soit $\gamma_i : [0,1] \to \mathcal{L} \subset \mathbb{R}^d$ une courbe paramétrée représentant la trajectoire de l'entité i dans cet espace latent. Ces courbes peuvent se croiser et former des enchevêtrements complexes.

L'objectif est de définir des invariants topologiques qui permettent de caractériser ces enchevêtrements. Pour cela, nous pouvons :

- Associer un polynôme de Jones $J(\gamma_i)$ à chaque trajectoire γ_i , en projetant cette trajectoire dans un espace à trois dimensions approprié.
- Utiliser le nombre de croisement minimal pour quantifier la complexité des enchevêtrements entre les trajectoires γ_i .
- Analyser les classes d'homologie de ces trajectoires dans l'espace latent pour capturer les propriétés topologiques invariantes.

4. Applications à la Simplification des Modèles :

Ces invariants topologiques peuvent être utilisés pour simplifier les représentations internes dans les modèles de langage tout en conservant leur fonction primaire.

4.1 Isotopie dans l'espace latent :

Soit une transformation $T : \mathcal{L} \to \mathcal{L}$, représentant une modification des représentations internes au cours de l'apprentissage du modèle. Si cette transformation correspond à une isotopie, alors les invariants topologiques des trajectoires restent invariants.

Cela signifie que même si les représentations internes sont modifiées par T, leur structure topologique sous-jacente (en termes de nœuds ou d'enchevêtrement) est conservée. En identifiant de tels invariants, on pourrait proposer des méthodes de simplification des modèles qui modifient les trajectoires de manière isotopique, permettant ainsi de réduire la complexité sans perdre la fonction principale.

4.2 Transformation isotopique et simplification :

Une transformation isotopique T pourrait être définie mathématiquement comme une famille continue de transformations T_i: $\mathcal{L} \to \mathcal{L}$ pour $i \in [0,1]$, où T_0 est l'identité et T_1 est la transformation finale appliquée au modèle. Cette continuité assure que les invariants topologiques ne changent pas et que la fonction du modèle est préservée.

5. Note mathématique :

- **Théorie des nœuds** : Modélise les trajectoires des représentations internes des mots/phrases comme des courbes fermées dans l'espace latent $\mathcal{L} \subset \mathbb{R}^d$.
- **Invariant topologique** : Utilise des outils comme le polynôme de Jones et le nombre de croisement minimal pour caractériser la complexité des enchevêtrements dans cet espace.
- **Isotopie** : Les transformations isotopiques de l'espace latent permettent de simplifier les représentations internes tout en conservant leur fonction, en préservant les invariants topologiques.

Ainsi, l'analyse topologique des données dans l'espace latent via la théorie des nœuds peut fournir une manière élégante de comprendre et de simplifier la dynamique des représentations internes dans les modèles de langage.

Partie III : Autre approche avec la géométrie algébrique

Cette approche combine à la fois l'homologie persistante et les complexes de Čech.

1. Homologie persistante :

L'homologie persistante est une méthode qui étudie la manière dont des structures topologiques (comme des composants connexes, des trous ou des cavités) apparaissent et disparaissent à différentes échelles dans un ensemble de données. Dans le cas des données textuelles représentées par des vecteurs dans un espace latent, cela permet de découvrir des motifs et des relations persistants entre des groupes de données (comme des mots, des phrases ou des documents).

1.1. Construction d'un filtrage :

Pour capturer cette structure multi-échelle, nous introduisons un filtrage. Soit X un ensemble de points de données (représentant par exemple des vecteurs latents) dans un espace $\mathcal{L} \subseteq \mathbb{R}^{d}$. L'idée est de construire une suite de complexes simpliciaux $\mathrm{K}(\mathrm{r})$, paramétrée par un paramètre d'échelle r, à partir de cet ensemble de points.

- On commence par définir une boule de rayon r centrée en chaque point $\mathrm{x} \in \mathrm{X}$, c'est-à-dire $\mathrm{B}(\mathrm{x}, \mathrm{r})$.
- À chaque valeur de r, on connecte deux points $\mathrm{x}, \mathrm{y} \in \mathrm{X}$ par une arête si leurs boules $\mathrm{B}(\mathrm{x}, \mathrm{r})$. et $\mathrm{B}(\mathrm{y}, \mathrm{r})$. s'intersectent, et ainsi de suite pour des simplexes de dimension supérieure.

  Cela forme une famille croissante de complexes simpliciaux $\mathrm{K}(\mathrm{r})$, où r est le rayon qui détermine le degré de connexion des points.

1.2. Persistance des homologies :

Pour chaque dimension k, l'homologie $\mathrm{H}_{\mathrm{k}}(\mathrm{K}(\mathrm{r}))$ des complexes simpliciaux $\mathrm{K}(\mathrm{r})$ capture les caractéristiques topologiques de dimension k, comme :

- H_0 : composantes connexes (groupes de points connectés).
- H_1 : trous (cycles qui ne peuvent pas être contractés en un point).
- H_2 : cavités (volumes enfermés par des cycles en dimension 2).

L'idée est d'analyser la naissance et la mort de ces caractéristiques au fur et à mesure que r augmente. Les caractéristiques qui persistent sur une large gamme de r sont considérées comme des motifs significatifs, tandis que celles qui apparaissent et disparaissent rapidement sont considérées comme du bruit.

Matériellement, on résume cette information sous la forme d'un diagramme de persistance, qui représente les caractéristiques topologiques par des points ($r_{naissance}$, r_{mort}) dans le plan.

1.3. Expression de l'homologie persistante :

Si nous notons $H_k(K(r))$ pour la k-ème homologie à une échelle r, alors l'homologie persistante est une suite d'inclusions :

$$H_k(K(r_1)) \rightarrow H_k(K(r_2)) \rightarrow \cdots \rightarrow H_k(K(r_n))$$

pour $r_1 < r_2 < \cdots < r_n$, qui capture l'apparition (naissance) et la disparition (mort) des classes d'homologie. Les diagrammes de persistance associés à ces suites visualisent les structures topologiques persistantes.

2. Complexes simpliciaux et complexes de Čech :

Un complexe simplicial est une structure combinatoire composée de simplexes (comme des points, des arêtes, des triangles, et des tétraèdres) qui sont utilisés pour approximer la forme de données complexes.

Le complexe de Čech est un type particulier de complexe simplicial, construit à partir de boules de rayon r centrées sur les points de données. Il est défini de manière à capturer les interactions topologiques entre les points en fonction de la taille des boules.

2.1. Simplexes :

Un simplexe est une généralisation d'un triangle dans n'importe quelle dimension :

- Un 0-simplexe est un point.
- Un 1-simplexe est une arête connectant deux points.
- Un 2-simplexe est un triangle formé par trois points connectés par des arêtes.
- Un k-simplexe est la généralisation à k points.

Un complexe simplicial K est un ensemble de simplexes qui satisfont les deux propriétés suivantes :

- **Propriété 1 :**

Si un simplexe appartient K, alors tous ses sous-simplexes appartiennent aussi à K.

(Preuve)
Par définition, un complexe simplicial est un ensemble de simplexes. Si un simplexe σ appartient à K, alors σ est un élément de cet ensemble. Un sous-simplexe de σ est par définition un sous-ensemble des sommets de σ. Puisque K contient tous les simplexes dont σ est composé, il contient nécessairement tous les sous-ensembles de sommets de σ, c'est-à-dire tous les sous-simplexes de σ.

- **Propriété 2 :**

L'intersection de deux simplexes est soit vide, soit un simplexe de dimension inférieure.

(Preuve)
Soient σ et τ deux simplexes de K. Leur intersection σ ∩ τ est l'ensemble des sommets communs à σ et τ.

- **Cas 1 :** Si σ ∩ τ est vide, la propriété est vérifiée.
- **Cas 2 :** Si σ ∩ τ n'est pas vide, alors cet ensemble de sommets communs définit un nouveau simplexe, disons ρ. Ce simplexe ρ est un sous-simplexe à la fois de σ et de τ. Par la propriété 1, comme σ et τ appartiennent à K, ρ appartient aussi à K. De plus, ρ ne peut pas avoir une dimension supérieure à σ ou τ, car il est défini par un sous-ensemble des sommets de σ ou τ.

Dans tous les cas, l'intersection de deux simplexes d'un complexe simplicial est soit vide, soit un simplexe de dimension inférieure appartenant au complexe.

Remarque : Ces deux propriétés assurent la cohérence structurelle d'un complexe simplicial. Elles garantissent que les simplexes d'un complexe s'emboîtent de manière bien définie, sans créer d'incohérences topologiques.

2.2. Complexe de Čech :

Le complexe de Čech $\mathcal{C}(\mathrm{r})$ est un type particulier de complexe simplicial défini à partir d'un ensemble de points $\mathrm{X} \subseteq \mathbb{R}^{d}$. Pour un rayon r, on considère une boule $\mathrm{B}(\mathrm{x}, \mathrm{r})$

autour de chaque point $x \in X$. Un simplexe est formé entre un ensemble de points $\{x_0, x_1, \ldots, x_k\}$ si et seulement si l'intersection des boules $B(x_0, r)$, $B(x_1, r)$,..., $B(x_k, r)$.

Formellement :

$$C(r) = \left\{ \{x_0, x_1, \ldots, x_k\} \subset X \mid \bigcap_{i=0}^{k} B(x_k, r) = \emptyset \right\}$$

2.3. Interprétation en termes de modèles de langage :

En analysant les complexes de Čech formés à partir des vecteurs représentant des mots dans un espace latent, on peut découvrir des relations topologiques entre ces vecteurs que les méthodes traditionnelles ne détectent pas. Par exemple, si un groupe de vecteurs forme un simplexe de dimension k, cela indique qu'ils partagent une interaction sémantique importante, au-delà de simples relations de proximité mesurées par la distance euclidienne.

3. Application de l'homologie persistante aux espaces latents :

L'idée centrale de l'homologie persistante est de détecter des motifs structurels multi-échelles dans les espaces latents, qui ne sont pas évidents à des échelles fixes (par exemple, en utilisant uniquement des distances euclidiennes).

3.1 Étude des clusters et structures persistantes dans l'espace latent :

- **Problème :** Les vecteurs latents issus d'un modèle de langage (par exemple, les vecteurs de mots ou de phrases) forment un nuage de points dans un espace de haute dimension $\mathbb{R}^d$. Ces points sont censés capturer des relations sémantiques, mais il peut être difficile de distinguer les clusters significatifs, les sous-groupes ou les anomalies à différentes échelles.
- **Approche par l'homologie persistante :** En appliquant un filtrage à ces vecteurs latents et en construisant une suite de complexes simpliciaux (comme les complexes de Čech), on peut analyser comment des clusters ou trous topologiques apparaissent et persistent à travers différentes échelles r.
    - **Naissance et mort des structures** : Si un certain nombre de clusters apparaissent dans l'espace latent à une petite échelle et persistent à mesure que r augmente, cela révèle des relations stables entre certains mots ou phrases, même lorsque la granularité change. Ces relations stables peuvent correspondre à des concepts sémantiques cohérents dans l'espace latent.
    - **Signification des caractéristiques topologiques persistantes** : Par exemple, si un groupe de mots forme un cycle H_1 dans un diagramme

de persistance, cela peut indiquer une relation sémantique circulaire ou cyclique entre ces mots, suggérant des synonymes ou des associations contextuelles. Si cette structure persiste à travers plusieurs échelles, elle pourrait révéler une organisation sémantique robuste dans l'espace latent.

3.2 Identifier les zones de complexité ou de bruit :

L'homologie persistante peut également être utilisée pour distinguer les structures importantes du bruit dans l'espace latent.

- **Exemple :** Si des caractéristiques topologiques apparaissent et disparaissent rapidement (c'est-à-dire des cycles ou des clusters de courte durée), elles peuvent être considérées comme du bruit ou des artefacts des représentations. En éliminant ces artefacts, il devient possible de simplifier l'espace latent en conservant uniquement les structures stables et significatives. Cela pourrait aider à identifier les sous-espaces du modèle où les représentations sont plus cohérentes.
- **Simplification des espaces latents** : L'élimination des structures topologiques de courte durée peut réduire la dimensionnalité ou simplifier les représentations latentes sans perdre d'informations importantes, améliorant ainsi l'efficacité computationnelle tout en préservant la performance.

4. Application des complexes de Čech aux espaces latents :

Les complexes de Čech permettent d'étudier les interactions topologiques entre les vecteurs latents en analysant les relations entre eux au fur et à mesure que l'échelle r varie. Cela peut révéler des groupements sémantiques complexes que les méthodes traditionnelles comme la distance euclidienne ou la cosinus-similarité ne détectent pas.

4.1 Détection de structures sémantiques complexes :

Les complexes de Čech permettent de capturer des interactions complexes entre les points de l'espace latent. Par exemple, dans un espace latent où les mots "*chien*", "*chat*", et "*animal*" ont des représentations proches, un simplexe de dimension 2 pourrait être formé entre ces trois points. Cela reflète une relation triangulaire entre les concepts, capturant une interaction sémantique sous-jacente.

- **Problème :** Les méthodes classiques de regroupement (clustering) des vecteurs latents pourraient rater ces interactions sémantiques complexes, se contentant de regrouper les vecteurs en fonction de la proximité euclidienne.

- **Solution par les complexes de Čech :** En utilisant les complexes simpliciaux pour construire ces interactions, on peut révéler des structures plus riches. Un simplexe de dimension k dans un complexe de Čech révèle une interaction entre $k + 1$ mots ou phrases dans l'espace latent, mettant en évidence des groupes sémantiques ou des relations d'association plus profondes.

4.2 Analyse des relations sémantiques cachées :

En analysant les complexes de Čech, on peut découvrir des groupements implicites ou des relations cachées dans l'espace latent.

- **Exemple :** Si un modèle de langage capture des relations sémantiques indirectes (comme des associations entre des synonymes ou des relations contextuelles), ces relations peuvent se manifester sous forme de simplexes de haute dimension dans les complexes de Čech. Par exemple, les mots "*docteur*", "*médical*" et "*soin*" pourraient ne pas être proches selon une simple métrique euclidienne, mais former un simplexe car leurs boules d'interaction dans l'espace latent se chevauchent.
- **Application :** En étudiant ces simplexes de différentes dimensions, on peut mieux comprendre comment le modèle encode des interactions sémantiques plus riches, et identifier des groupes de mots ou de phrases qui partagent une signification commune mais ne sont pas directement liés dans l'espace euclidien.

5. Amélioration des modèles basée sur l'analyse topologique :

En utilisant l'homologie persistante et les complexes de Čech, il est possible d'améliorer la compréhension et la performance des modèles de langage de plusieurs manières :

5.1 Amélioration de l'organisation sémantique :

En analysant les structures topologiques persistantes dans l'espace latent, on peut ajuster les représentations pour mieux capturer des relations sémantiques importantes à différentes échelles. Cela pourrait se faire en ajustant l'objectif de l'entraînement du modèle pour qu'il maximise la cohérence des groupes sémantiques détectés par l'homologie persistante.

- **Action possible :** Ajuster les transformations dans l'espace latent pour favoriser la stabilité des clusters ou des cycles sémantiques importants, en modifiant la fonction de perte du modèle pour qu'elle prenne en compte la persistance des caractéristiques topologiques.

5.2 Simplification des modèles en éliminant le bruit topologique :

Les artefacts ou les structures de courte durée détectés par l'homologie persistante peuvent être éliminés pour simplifier le modèle et ses représentations sans sacrifier la performance.

- **Action possible :** Entraîner un modèle pour qu'il conserve uniquement les motifs topologiques persistants (par exemple, en réduisant la dimensionnalité ou en affinant les représentations latentes), ce qui pourrait réduire la taille du modèle ou améliorer son efficacité computationnelle.

5.3 Découverte de nouvelles relations sémantiques :

L'analyse des complexes simpliciaux comme les complexes de Čech permet de découvrir des groupes sémantiques cachés ou des interactions complexes qui ne sont pas visibles avec des méthodes traditionnelles. Ces découvertes pourraient être utilisées pour enrichir les bases de données sémantiques (thésaurus, bases de synonymes, etc.) ou améliorer les algorithmes de génération de texte en capturant des relations plus complexes entre les mots.

6. Note mathématique :

- **Homologie persistante** : Capture les structures topologiques persistantes dans les données textuelles, à travers des filtrages successifs. Les caractéristiques persistantes à travers plusieurs échelles sont visualisées par des diagrammes de persistance.

$$\mathrm{H_k(K(r_1)) \rightarrow H_k(K(r_2)) \rightarrow \cdots \rightarrow H_k(K(r_n))}$$

  Les diagrammes de persistance résument la naissance et la mort des caractéristiques topologiques.

- **Complexes de Čech** : Sont utilisés pour construire des complexes simpliciaux à partir de points de données, révélant des interactions topologiques que les méthodes traditionnelles pourraient manquer. Ces complexes se basent sur des intersections de boules centrées sur les points de données.

$$C(\mathrm{r}) = \left\{ \{x_0, x_1, \ldots, x_k\} \subset \mathrm{X} \mid \bigcap_{i=0}^{k} B(x_k, \mathrm{r}) = \emptyset \right\}$$

Ces outils permettent de capturer et de comprendre la structure topologique multi-échelle des représentations internes des modèles de langage, révélant des motifs sémantiques cachés et des relations entre les entités textuelles.