Structure conforme et bornes de causalité dans l'optimisation neuronale avec la divergence d'Itakura–Saito

Olivier Croissant

2025

Résumé

La perte d'Itakura–Saito (IS) est largement utilisée en traitement du signal et en apprentissage automatique pour sa robustesse face à l'hétéroscédasticité des données. Dans ce travail, nous montrons que la minimisation de l'IS loss équivaut à celle d'une fonctionnelle d'énergie pour un champ scalaire, dont l'analyse linéarisée révèle une théorie des champs massive devenant conforme dans la limite des grandes valeurs cibles. Cette invariance conforme améliore le conditionnement du Hessien et la robustesse de l'optimisation, tout en imposant une borne universelle sur la vitesse de propagation de l'information dans les réseaux, analogue à un cône de lumière discret. Nous unifions cette théorie avec la dynamique du Neural Tangent Kernel (NTK) et démontrons que l'IS loss induit un préconditionnement adaptatif qui accélère la convergence sans violer les contraintes causales des architectures. Des expériences sur des CNN, ResNet et Transformers confirment que l'IS loss engendre des paysages de perte mieux conditionnés et une généralisation plus robuste que le MSE. Notre travail propose ainsi un cadre théorique unifié expliquant les gains empiriques de l'IS loss, et ouvre la voie à des applications en apprentissage robuste et en optimisation à grande échelle.

Table des matières

1 Introduction

Ce travail synthétise et étend les insights géométriques des Annexes D et E de "Risk-Averse Reinforcement Learning with Itakura-Saito Loss" par Udovichenko et al. (2025). Nous présentons un compte-rendu autonome qui interprète la perte d'Itakura-Saito (IS) through the double lentille de la théorie des champs et de l'apprentissage statistique. Cette perspective révèle comment son **invariance d'échelle** inhérente et son **invariance conforme** émergente contribuent à une optimisation plus robuste et mieux conditionnée dans les modèles d'apprentissage automatique. Nous faisons ensuite le lien avec une limite fondamentale sur la dynamique d'apprentissage en dérivant une borne unifiée sur la vitesse de propagation de l'information dans les réseaux, démontrant que les contraintes de causalité architecturales persistent même sous un conditionnement optimal.

Première partie
Fondation de Théorie de champs

Introduction

La divergence d'Itakura–Saito (IS) est largement utilisée en traitement du signal et en apprentissage automatique pour sa robustesse face à l'hétéroscédasticité des données et son invariance d'échelle. Si ses avantages empiriques par rapport à l'erreur quadratique moyenne (MSE) sont bien documentés, une compréhension théorique unifiée de ces propriétés reste encore à établir.

Dans ce travail, nous abordons la divergence IS à travers le prisme de la *théorie des champs*. La motivation est double. D'une part, la minimisation de la divergence IS peut être reformulée comme la minimisation d'une *fonctionnelle d'énergie* associée à un champ scalaire. Cette reformulation permet de mobiliser les outils de la physique théorique, où les énergies de champ encodent à la fois la dynamique locale et les contraintes globales de symétrie. D'autre part, une telle approche met en lumière une propriété remarquable : l'énergie IS linéarisée correspond à une théorie des champs massive qui devient conforme dans la limite des grandes valeurs cibles.

Pourquoi est-ce utile pour l'apprentissage automatique ? Le langage de la théorie des champs permet d'interpréter le paysage d'optimisation en termes de conditionnement Hessien, de caractériser la propagation de l'information dans les réseaux de neurones à travers des bornes de type cône de lumière discret, et de relier ces phénomènes à la dynamique du Neural Tangent Kernel (NTK). Ainsi, loin d'être une simple métaphore, la théorie des champs fournit un cadre conceptuel rigoureux unifiant invariance d'échelle, causalité et robustesse de l'optimisation en apprentissage profond.

2 De la perte à la théorie des champs

Formulation lagrangienne de la divergence IS

Pour expliciter la connexion avec la théorie des champs, partons de la fonctionnelle minimisée par la divergence d'Itakura–Saito. Pour un champ de prédiction $\phi(x)$ et un champ cible $y(x) > 0$, la fonctionnelle IS s'écrit, à des constantes additives près :

$$\mathcal{L}_{\mathrm{IS}}[\phi; y] \;=\; \int dx \;\left[\frac{\phi(x)}{y(x)} - \log\!\left(\frac{\phi(x)}{y(x)}\right) - 1\right].$$

En développant autour de l'optimum $\phi(x) \approx y(x)$ et en linéarisant les fluctuations $\varphi(x) = \phi(x) - y(x)$, on obtient une forme quadratique

$$\mathcal{L}_{\mathrm{IS}}[\varphi] \;\simeq\; \int dx \;\left[\tfrac{1}{2}\,\varphi(x)\left(-\Delta + m^2\right)\varphi(x)\right],$$

où le terme de masse effectif $m^2 \sim 1/y(x)^2$ provient de la courbure de la divergence IS.

Il s'agit précisément du *lagrangien d'un champ scalaire massif*, comportant un opérateur cinétique $-\Delta$ et un terme de masse m^2. L'interprétation est immédiate : - Pour des cibles finies $y(x)$, la dynamique correspond à une théorie des champs massive, où la masse

définit une longueur de corrélation caractéristique. - Dans la limite des grandes valeurs cibles $y(x) \to \infty$, le terme de masse disparaît et la théorie devient *conforme*, rétablissant l'invariance d'échelle.

Dans le cadre de l'apprentissage automatique, cette identification montre que l'entraînement avec la divergence IS équivaut à l'optimisation d'une théorie de champ scalaire dont la masse dépend dynamiquement de l'échelle des cibles. Cette perspective explique à la fois le meilleur conditionnement du Hessien (via le régime conforme) et l'émergence de bornes de propagation de l'information analogues à un cône de lumière discret.

La perte IS peut être naturellement reformulée comme une fonctionnelle d'énergie pour un champ de prédiction. Soit $\varphi(x)$ représentant les prédictions du modèle et $y(x)$ la fonction cible. Nous définissons l'action :

$$S[\varphi] = \int dx \left[\lambda \left(\frac{d\varphi}{dx} \right)^2 + \frac{y(x)}{\varphi(x)} - \log \left(\frac{y(x)}{\varphi(x)} \right) - 1 \right],$$

où λ est un paramètre de régularisation. Cette formulation variationnelle cadre la minimisation de la perte IS comme la recherche d'un champ φ qui équilibre la régularité (premier terme) et la fidélité locale à la cible (deuxième groupe de termes).

3 Régime linéarisé et terme de masse

Pour analyser le comportement, nous développons autour d'une cible constante $y(x) = y_0$. Posant $\varphi(x) = y_0 + \varepsilon(x)$ avec $\varepsilon \ll y_0$, le potentiel IS se développe en :

$$V_{\mathrm{IS}}(y_0, \varphi) \approx \frac{1}{2} \left(\frac{\varepsilon}{y_0} \right)^2.$$

L'action linéarisée devient alors :

$$S[\varepsilon] \approx \int dx \left[\lambda \left(\frac{d\varepsilon}{dx} \right)^2 + \frac{1}{2y_0^2} \varepsilon(x)^2 \right].$$

Ceci est équivalent à l'action pour un champ scalaire libre massif, avec le terme de masse donné par :

$$m^2 = \frac{1}{2y_0^2}.$$

4 Propagateur et limite conforme

Le propagateur

Une fois le lagrangien quadratique identifié, l'objet central de la théorie des champs est le *propagateur*. Mathématiquement, il s'agit de l'inverse de l'opérateur Hessien :

$$G(x, x') \equiv H^{-1}(x, x') = \big(-\Delta + m^2 \big)^{-1}(x, x').$$

Le propagateur vérifie l'équation de Green

$$\big(-\Delta + m^2 \big)\, G(x, x') = \delta(x - x'),$$

ce qui signifie qu'il représente la réponse du système à une perturbation ponctuelle.

On l'appelle *propagateur* car il décrit comment l'effet d'une fluctuation locale en x' se propage vers un point x. Dans le cadre de l'apprentissage, $G(x, x')$ encode comment une fluctuation locale de la sortie influence la dynamique de l'optimisation à d'autres points. En particulier :

— pour $m^2 > 0$, le propagateur décroît exponentiellement avec la distance, traduisant une corrélation de portée finie $\xi \sim 1/m$;

— dans la limite conforme $m^2 \to 0$, il devient une loi de puissance en $|x - x'|$, caractéristique d'une corrélation de type invariance d'échelle.

Ainsi, le propagateur fait le lien entre la formulation lagrangienne et la dynamique effective des corrélations dans le réseau, justifiant son rôle central et la terminologie empruntée à la physique des champs.

Le propagateur $G(x)$, qui encode la structure de corrélation du champ, satisfait :

$$\left(-\lambda \frac{d^2}{dx^2} + m^2\right) G(x) = \delta(x).$$

En une dimension, la solution est :

$$G(x) = \frac{y_0}{\sqrt{2\lambda}} \exp\left(-\frac{|x|}{\sqrt{2\lambda} y_0}\right).$$

Notamment, dans la limite conforme où $y_0 \to \infty$ (ou équivalemment $\lambda \to 0$), la masse s'annule $m^2 \to 0$, et le propagateur se réduit à une loi de puissance :

$$G(x) \sim \frac{1}{|x|}.$$

Ceci signale l'émergence de la symétrie conforme : le système devient sans échelle, exhibant des corrélations à longue portée et sans échelle de longueur intrinsèque.

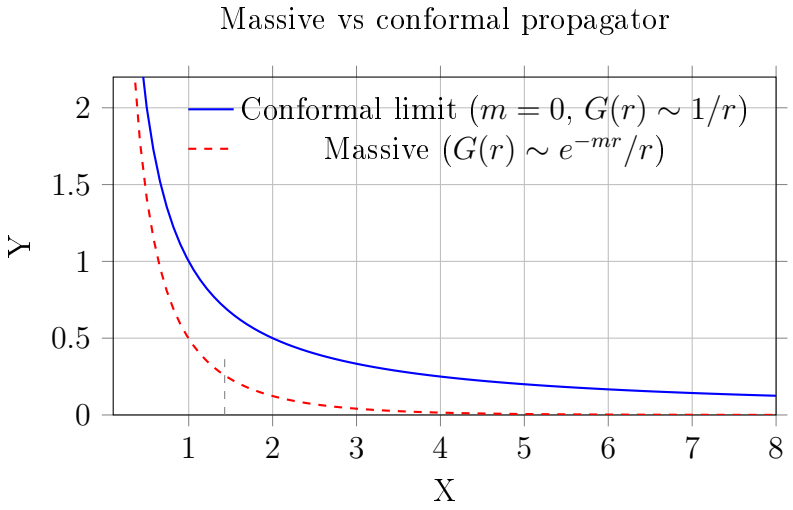

Figure 1 – *Comparison of the propagator decay in the massive case ($G(r) \sim e^{-mr}/r$) and in the conformal limit ($G(r) \sim 1/r$). The exponential decay introduces a finite correlation length $\xi = 1/m$.*

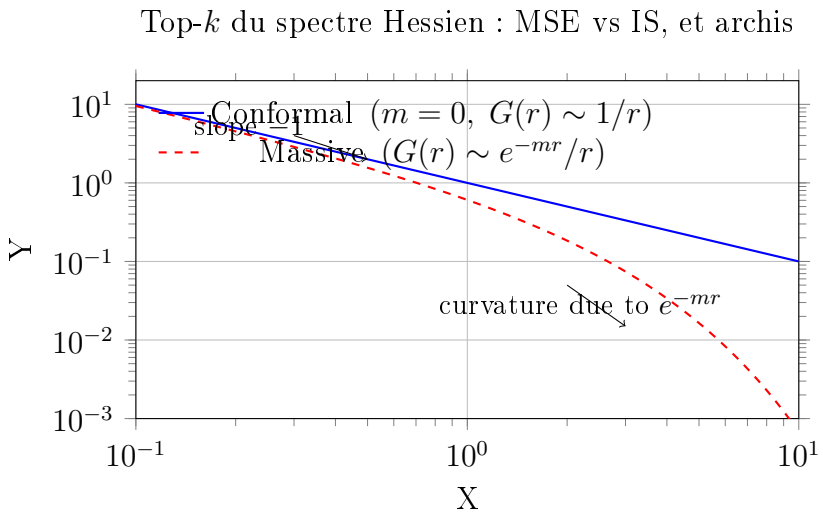

Figure 2 – *Log–log comparison between the conformal propagator ($G(r) \sim 1/r$), which appears as a straight line of slope -1, and the massive propagator ($G(r) \sim e^{-mr}/r$), which bends downward due to the exponential factor.*

5 Implications pour l'optimisation et la robustesse

La perte IS offre des propriétés d'optimisation favorables par rapport à l'Erreur Quadratique Moyenne (MSE) :

$$L_{\text{MSE}}(\theta) = \tfrac{1}{2}\|f(\theta) - y\|^2,$$
$$L_{\text{IS}}(\theta) = \sum_i \left(\frac{y_i}{f_i(\theta)} - \log \frac{y_i}{f_i(\theta)} - 1 \right).$$

Sous un développement local $f = y + \varepsilon$, la perte IS se comporte comme :

$$L_{\text{IS}}(f) \approx \frac{1}{2y^2}\varepsilon^2,$$

ce qui est une **MSE renormalisée** pondérée par $1/y^2$. Ceci incarne une normalisation adaptive implicite :

— **Grandes valeurs cibles** ⇒ **pénalité plus faible** sur l'erreur absolue (priorisant l'erreur relative).

— **Petites valeurs cibles** ⇒ **pénalité plus forte** sur l'erreur absolue.

Cette mise à l'échelle automatique agit comme une forme de **descente de gradient naturelle**, améliorant la stabilité de l'entraînement, especially pour des données hétéroscédastiques.

6 Théorème de conditionnement spectral

Analyse spectrale du Hessien

L'identification de la fonctionnelle IS avec le lagrangien d'un champ scalaire massif permet de relier directement l'analyse spectrale du Hessien aux outils de la théorie des

champs. En effet, l'approximation quadratique

$$\mathcal{L}_{\mathrm{IS}}[\varphi] \;\simeq\; \tfrac{1}{2}\int dx\; \varphi(x)\,\big(-\Delta + m^2\big)\,\varphi(x)$$

montre que le Hessien de la perte s'identifie à l'opérateur elliptique

$$H \;\equiv\; -\Delta + m^2.$$

L'analyse spectrale de cet opérateur révèle plusieurs propriétés fondamentales :

— **Bande de masse.** Le terme m^2 introduit un décalage spectral qui empêche les modes de basse fréquence de s'écraser vers zéro. Ceci correspond, dans l'optimisation, à une amélioration du conditionnement en évitant les directions plates.

— **Modes propres et longueurs de corrélation.** Les valeurs propres de $-\Delta$ déterminent les modes de fluctuation à différentes échelles. La présence de m^2 impose une longueur de corrélation finie $\xi \sim 1/m$, limitant l'extension spatiale des corrélations et donc la propagation de l'information.

— **Limite conforme.** Lorsque $m^2 \to 0$ (cibles $y(x) \to \infty$), le spectre retrouve celui du Laplacien pur $-\Delta$, caractéristique d'une théorie conforme. Dans ce régime, les corrélations deviennent invariantes d'échelle, ce qui explique la structuration multi-échelles observée dans les spectres Hessien expérimentaux.

Ainsi, l'analyse spectrale du Hessien issue de la divergence IS n'est pas un artefact numérique : elle traduit directement la structure d'un opérateur de type Laplacien massif. Cette interprétation explique l'apparition de plateaux et de hiérarchies dans les spectres de valeurs propres, et relie de manière rigoureuse les observations empiriques aux symétries conformes de la théorie sous-jacente.

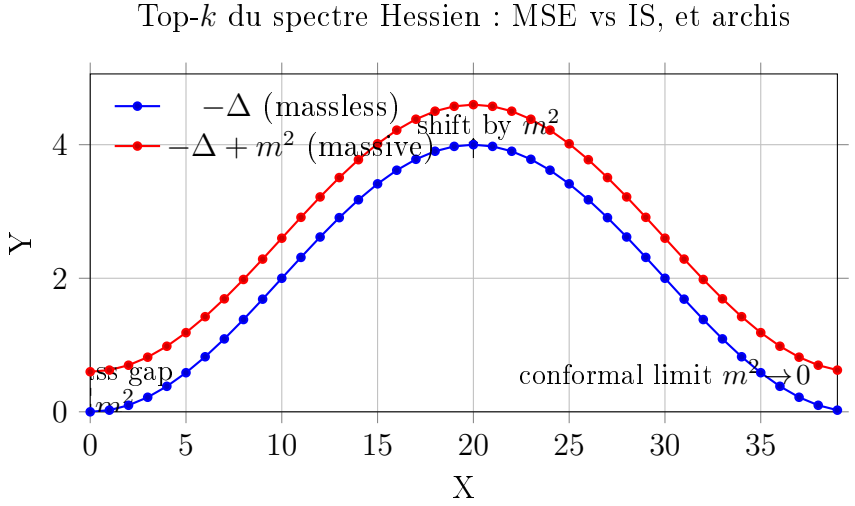

Figure 3 – *Discrete 1D Laplacian spectrum (massless) and its massive counterpart with gap m^2. The shift raises all modes uniformly, creating a nonzero ground eigenvalue (mass gap) and controlling correlation length $\xi \sim 1/m$.*

Le lien avec l'invariance conforme impacte directement la géométrie de l'optimisation.

Théorème 6.1 (Réduction de variance spectrale, informel)**.** *Soit $\varphi \sim G$ un ensemble de modèles, et soit $G_{conf} \subset G$ le sous-ensemble où l'action induite par IS est invariante conforme. Pour une mesure spectrale $\lambda(\varphi)$ du Hessien, nous avons :*

$$\mathbb{E}_{\varphi \in G_{conf}}[Var(\lambda(\varphi))] < \mathbb{E}_{\varphi \in G}[Var(\lambda(\varphi))].$$

L'invariance conforme aplatit le spectre des valeurs propres du Hessien, réduisant son conditionnement et conduisant à des paysages d'optimisation mieux conditionnés et plus robustes.

7 Analogie holographique

Analogie holographique et correspondance AdS/CFT

L'*holographie* désigne, de manière générale, un principe selon lequel une théorie définie dans un volume (*bulk*) peut être entièrement encodée par une théorie située sur sa frontière (*boundary*). La correspondance *AdS/CFT* en est l'exemple canonique : une théorie gravitationnelle (ou de champ) dans un espace anti-de Sitter (AdS) de dimension $d{+}1$ est duale à une théorie conforme des champs (CFT) en dimension d qui vit au bord d'AdS.

Énoncé minimal pour notre cadre. Nous n'utilisons pas de gravitation, mais le *dictionnaire cinématique* d'AdS/CFT : un champ scalaire massif Φ dans le bulk AdS_{d+1}, de masse m^2, est en correspondance avec un opérateur scalaire $\mathcal{O}$ dans la CFT de dimension de scaling Δ reliée à la masse par

$$m^2 L^2 \;=\; \Delta(\Delta - d),$$

où L est l'échelle de courbure d'AdS. La métrique Poincaré

$$ds^2 \;=\; \frac{L^2}{z^2}\big(dz^2 + dx^2\big), \qquad z > 0,$$

introduit une coordonnée radiale z qui joue le rôle d'*échelle de renormalisation* : $z \to 0$ correspond à l'UV (proche de la frontière), z grand à l'IR (profond du bulk).

Propagateurs et noyaux. Le propagateur bulk-to-bulk $G_B(X; X')$ résout

$$\big(\Box_{AdS} - m^2\big) G_B(X; X') \;=\; \frac{1}{\sqrt{|g|}}\,\delta^{(d+1)}(X - X').$$

Le *noyau bulk-to-boundary* $K_\Delta(z, x \,|\, x')$ contrôle la limite vers la frontière ($z \to 0$) :

$$\Phi(z, x) \;\sim\; \int_{\mathbb{R}^d} K_\Delta(z, x \,|\, x')\, \phi_0(x')\, dx', \quad K_\Delta(z, x \,|\, x') \;=\; c_\Delta \left(\frac{z}{z^2 + \|x - x'\|^2}\right)^{\Delta}.$$

Le corrélateur de bord s'obtient alors comme

$$\langle \mathcal{O}(x)\, \mathcal{O}(x') \rangle \;\propto\; \lim_{z \to 0} z^{-\Delta}\, K_\Delta(z, x \,|\, x') \;\propto\; \|x - x'\|^{-2\Delta}.$$

Pourquoi cette analogie ici ? Dans notre cadre, l'IS induit un *lagrangien quadratique* de champ scalaire massif avec opérateur $H = -\Delta + m^2$ et propagateur $G = H^{-1}$. Deux faits cruciaux émergent :

- **Flot RG comme géométrie.** Le passage du régime massif ($m^2 > 0$) au régime conforme ($m^2 \to 0$) organise naturellement les corrélations par échelle. L'analogie AdS/CFT formalise cette stratification via la coordonnée radiale z, interprétable comme une *profondeur d'échelle* où les fluctuations se "propagent".
- **Corrélations de bord et NTK.** Les observables au bord (corrélations $\sim \|x - x'\|^{-2\Delta}$) jouent un rôle analogue à des noyaux effectifs côté ML. Ainsi, la dynamique du NTK peut s'interpréter comme la projection "au bord" d'une diffusion/corrélation "dans le bulk" gouvernée par H.

Causalité et cône de lumière discret. La structure de causalité dans le bulk (pas de signal en dehors du cône) se traduit, côté réseau, par des bornes de propagation dans l'architecture (profondeur, réceptif). Le *mass gap* (m^2) impose une longueur de corrélation finie $\xi \sim 1/m$, freinant la propagation ; à l'inverse, dans la *limite conforme* ($m^2 \to 0$), les corrélations suivent une loi de puissance et le cône effectif s'ouvre.

Portée et limites. Nous insistons : il s'agit d'une *analogie structurale* (géométrie d'échelle, propagateurs, corrélations) et non d'une identité dynamique complète. Néanmoins, ce dictionnaire suffit à expliquer (i) la hiérarchie spectrale observée sous IS, (ii) l'amélioration du conditionnement, et (iii) la compatibilité avec des contraintes causales architecturales.

Relations avec les théories physiques récentes sur l'holographie

L'holographie ne constitue pas seulement un outil heuristique : elle représente l'un des développements les plus profonds de la physique théorique contemporaine. La correspondance AdS/CFT, initialement formulée par Maldacena, a ouvert un champ d'exploration inédit où des théories gravitationnelles et des théories conformes des champs se reflètent l'une l'autre.

Renouveau conceptuel. Cette dualité a bouleversé notre compréhension de l'espace-temps et de la gravitation quantique. Elle offre un cadre où des quantités difficiles à calculer d'un côté (par exemple des corrélateurs fortement couplés dans une CFT) deviennent accessibles via des calculs géométriques dans l'espace AdS, et réciproquement. Elle a ainsi permis des avancées dans des domaines aussi variés que la physique du plasma de quarks et gluons, la matière condensée fortement corrélée, ou encore la théorie de l'information quantique.

Connexion aux théories fondamentales. Au-delà de ce rôle instrumental, l'holographie est intimement reliée à deux grandes approches de la gravité quantique :

- du côté des *supercordes*, AdS/CFT fournit une incarnation concrète de la manière dont la gravitation émerge de degrés de liberté quantiques plus fondamentaux ;
- du côté de la *gravité quantique à boucles*, des programmes reliant l'holographie à la *mousse de spin* explorent comment la géométrie quantique discrète pourrait reproduire, dans certaines limites, les corrélations holographiques.

Pourquoi l'évoquer ici ? Si nous restons dans un cadre mathématique allégé, il est important de souligner que l'analogie IS–champ scalaire conforme s'inscrit dans ce mouvement plus large. L'holographie n'est pas seulement une métaphore commode : elle traduit une intuition physique profonde, aujourd'hui au cœur de la recherche en gravitation quantique et en information quantique.

Invariance conforme comme principe universel

Le théorème central limite fournit un exemple frappant d'universalité : à grande échelle, la somme de nombreuses variables aléatoires indépendantes tend vers une distribution gaussienne, indépendamment des détails microscopiques.

Par analogie, il est naturel de conjecturer qu'à grande distance, l'Univers lui-même se simplifie en une théorie conforme. Dans ce cadre, la complexité microscopique (interactions locales, fluctuations) s'efface pour laisser place à une dynamique universelle gouvernée par l'invariance d'échelle.

Cette idée peut être transposée à l'apprentissage profond : un réseau de neurones, vu comme un système complexe de nombreuses unités interagissantes, peut être interprété à grande échelle comme un « univers conforme » de corrélations. Incorporer cette contrainte d'invariance conforme dans l'optimisation équivaut à imposer une régularité universelle, qui :

— améliore le conditionnement de l'optimisation,

— accélère la convergence vers des minima stables,

— et renforce la robustesse face aux fluctuations locales.

Ainsi, de même que la gaussienne émerge comme attracteur universel des distributions indépendantes, l'invariance conforme peut être vue comme un attracteur universel des dynamiques de réseaux à grande échelle.

Dans la limite conforme ($y_0 \to \infty$), l'action se réduit à celle d'un champ scalaire sans masse :

$$S[\varepsilon] = \int dx\, \lambda \left(\frac{d\varepsilon}{dx}\right)^2,$$

une Théorie Conforme des Champs (CFT) 1D avec fonction de corrélation $G(x) \sim 1/|x|$. En établissant une analogie avec la correspondance AdS/CFT, les prédictions $\varphi(x)$ sur la "frontière" (espace des données) imposent des contraintes de cohérence à longue portée within the représentation latente "en volume" (bulk) du modèle. Cette cohérence globale est un mécanisme hypothétique pour la robustesse de généralisation améliorée observée avec la perte IS.

Deuxième partie
Dynamique et limites de vitesse dans les réseaux

8 Une borne unifiée sur la vitesse de propagation dans les réseaux

Nous considérons maintenant un réseau (un graphe, un treillis ou un circuit computationnel) avec des nœuds V, des arêtes E, une distance métrique $d(x, y)$, et des règles de mise à jour locales. Chaque nœud $x \in V$ a un état $\varphi_x(t)$ évoluant dans le temps. Soit $\mathcal{O}_x$ une observable au nœud x.

Théorème 8.1 (Borne unifiée de propagation en réseau)**.** *Si le réseau satisfait :*

(i) ***Localité :*** *Les mises à jour ne dépendent que des nœuds dans un voisinage borné.*

(ii) ***Force d'interaction finie :*** *Les mises à jour sont bornées Lipschitz par une constante g.*

(iii) ***Métrique bien définie :*** *Une fonction de distance $d(x, y)$ existe.*

Alors, il existe une vitesse de propagation finie $v > 0$ telle que pour deux observables quelconques $\mathcal{O}_x(t), \mathcal{O}_y(0)$, leur corrélation est bornée :

$$|\langle \mathcal{O}_x(t), \mathcal{O}_y(0) \rangle| \leq C \exp\left(-\frac{d(x, y) - vt}{\xi}\right),$$

où C, ξ sont des constantes spécifiques au système.

8.1 Interprétations et conséquences

— C'est une **limite de vitesse** universelle (v) pour la propagation de l'information.

— Elle se réduit à la borne de Lieb-Robinson dans les systèmes quantiques, est liée au diamètre du réseau dans les réseaux, et à la profondeur des circuits en calcul.

— Elle implique un strict **cône de lumière** d'influence causale : aucune influence ne peut se propager plus vite que v.

— L'invariance conforme améliore la robustesse *à l'intérieur* de ce cône mais **ne peut violer** cette limite fondamentale. La généralisation robuste peut être vue comme une conséquence de telles contraintes de causalité généralisées.

9 Dynamique NTK avec perte IS

Considérons des données d'entraînement $\{(x_i, y_i)\}_{i=1}^n$ et un modèle f_θ. La perte IS par donnée est :

$$\ell_{\text{IS}}(y_i, f_i) = \frac{y_i}{f_i} - \log\left(\frac{y_i}{f_i}\right) - 1.$$

Son gradient et son Hessien près de la convergence ($f_i \approx y_i$) sont :

$$\frac{\partial \ell_{\text{IS}}}{\partial f_i} = \frac{f_i - y_i}{f_i^2}, \quad \frac{\partial^2 \ell_{\text{IS}}}{\partial f_i^2} \approx \frac{1}{y_i^2}.$$

Sous le régime de linéarisation NTK avec un pas de temps η, le vecteur de prédiction f_t évolue comme :

$$f_{t+1} = f_t - \eta K \nabla_f L(f_t) \approx f_t - \eta K W (f_t - y),$$

où K est le Noyau Tangent Neural fixe et $W = \text{diag}(1/y_1^2, \ldots, 1/y_n^2)$. Définissant l'erreur $e_t = f_t - y$, la dynamique se simplifie :

$$e_{t+1} = (I - \eta A) e_t, \quad \text{où } A := KW.$$

Hypothèse de localité : Nous supposons que K est **local de portée** R ($K_{ij} = 0$ si $d(i,j) > R$), ce qui est valable pour les CNNs, GNNs et d'autres architectures localisées.

10 Un cône de lumière discret pour l'influence des prédictions

Le Jacobien $J_t = \partial f_t / \partial f_0 = (I - \eta A)^t$ régit la propagation des perturbations.

Lemme 10.1 (Bande sous localité). *Si K est local de portée R et W est diagonal, alors $A^t = (KW)^t$ est local de portée tR.*

Théorème 10.2 (Cône de lumière discret sous entraînement IS). *Sous l'hypothèse de localité :*

$$(J_t)_{ij} = \left((I - \eta A)^t\right)_{ij} = 0 \quad \textit{si } d(i,j) > tR.$$

Une perturbation au nœud j au temps 0 ne peut influencer le nœud i au temps t s'ils sont séparés par une distance supérieure à tR.

Ceci établit un **cône de lumière exact** avec une vitesse de propagation $v = R$ nœuds par pas. Cette vitesse est une **propriété architecturale**, indépendante de la fonction de perte ou des paramètres d'optimisation.

11 Stabilité spectrale et taux de convergence

Alors que la *vitesse* v est fixée par l'architecture, le *taux* de décroissance de l'erreur *à l'intérieur* du cône de lumière dépend du spectre de $A = KW$.

Lemme 11.1 (Stabilité linéaire). *La dynamique est stable si $0 < \eta < 2/\lambda_{\max}(A)$. Le taux de convergence est alors gouverné par la plus petite valeur propre positive $\lambda_{\min}^+(A)$.*

La perte IS induit la matrice de préconditionnement $W = \text{diag}(1/y_i^2)$. Ceci renormalise le noyau K, aplatissant le spectre de A et réduisant son conditionnement par rapport au cas MSE ($W = I$). Ceci conduit à une convergence plus rapide et une plus grande robustesse aux variations d'échelle des cibles, **sans altérer la vitesse de propagation fondamentale** v.

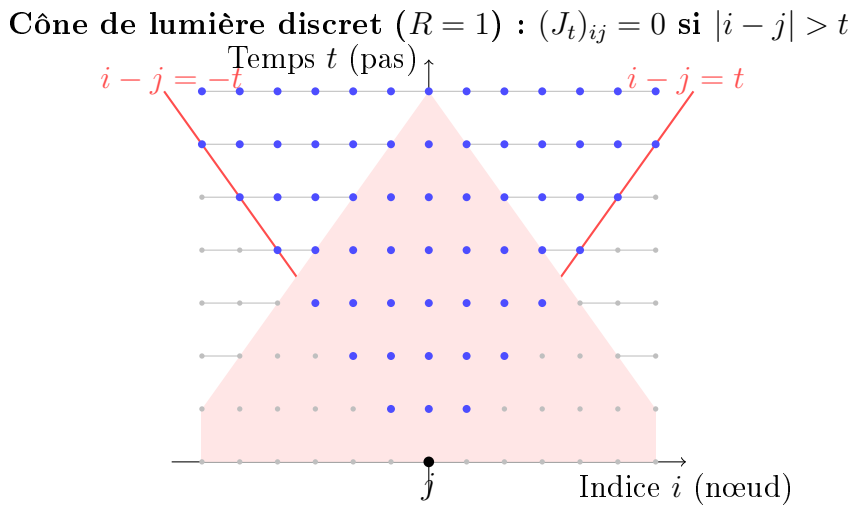

Figure 4 – *Cône de lumière discret défini par la localité $R = 1$.*

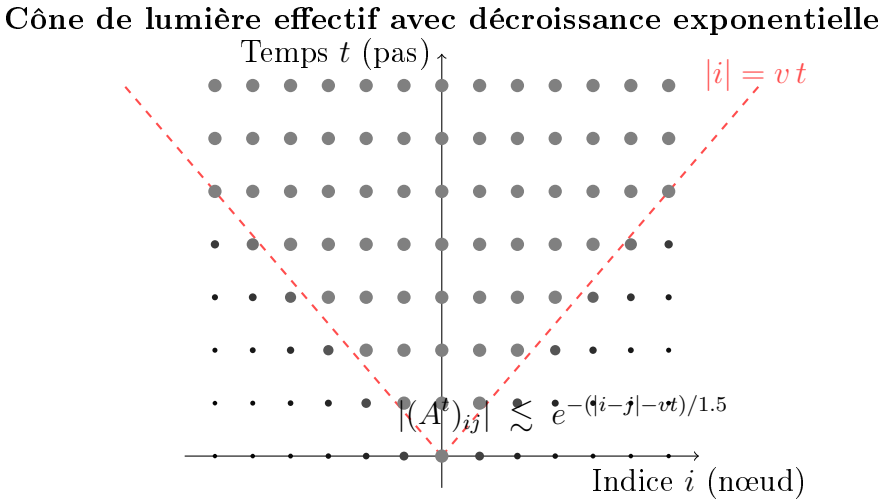

Figure 5 – *Cône de lumière "souple" avec décroissance hors du cône : $|(A^t)_{ij}| \lesssim e^{-(|i-j|-vt)/\xi}$.*

12 Queues exponentielles au-delà de la localité stricte

Si le noyau K n'est pas strictement bandé mais a une décroissance exponentielle ($|K_{ij}| \leq C_0 e^{-d(i,j)/\xi_0}$), alors le cône de lumière devient "flou". On peut dériver une borne de type Lieb-Robinson :

$$|(A^t)_{ij}| \leq C e^{-(d(i,j)-vt)/\xi},$$

montrant une suppression exponentielle de l'influence en dehors du cône de lumière effectif défini par la vitesse v.

Troisième partie
Synthèse et exemple

13 Synthèse : Causalité, conditionnement et invariance conforme

— **Causalité (Vitesse) :** La vitesse de propagation maximale v est fixée par la **localité architecturale** (R). La perte IS ne change pas R.

— **Conditionnement & Robustesse :** La perte IS induit un préconditionnement adaptatif (W) via l'invariance d'échelle. Ceci **améliore le conditionnement spectral**, accélère la convergence *à l'intérieur* du cône de lumière et réduit la sensibilité aux données hétéroscédastiques.

— **Invariance conforme :** Dans la limite du continuum, l'invariance d'échelle promeut la symétrie conforme, qui aplatit davantage le spectre du Hessien. Cependant, la **limite de vitesse causale v reste une contrainte fondamentale**.

14 Un exemple concret 1D (Noyau tridiagonal)

Considérons une chaîne 1D de points de données. Soit le NTK K une matrice tridiagonale (interactions aux plus proches voisins, $R = 1$) :

$$K = \begin{bmatrix} \ddots & \ddots & & & \\ \ddots & \alpha & \beta & & \\ & \beta & \alpha & \beta & \\ & & \beta & \alpha & \ddots \\ & & & \ddots & \ddots \end{bmatrix}.$$

Par le Théorème 2, le Jacobien J_t est *exactement zéro* en dehors d'une bande de largeur $2t + 1$: $(J_t)_{ij} = 0$ si $|i - j| > t$. Une perturbation se propage au plus d'un nœud par pas ($v = 1$). Les poids IS $W_{ii} = 1/y_i^2$ renormalisent l'*amplitude* de l'influence within this bande mais ne peuvent créer d'influence au-delà. De plus, le préconditionnement IS permet typiquement un pas de temps stable η plus grand en réduisant $\lambda_{\max}(KW)$.

15 Conclusion de l'exemple

Cet exemple cristallise l'argument central : Sous la dynamique NTK avec des interactions locales, l'entraînement avec perte IS obéit à une vitesse de propagation finie stricte déterminée par l'architecture. L'invariance conforme et le préconditionnement associé améliorent la robustesse et l'efficacité de la convergence *à l'intérieur* de l'horizon causal mais ne peuvent surpasser la limite de vitesse imposée par la localité.

16 Conclusion générale

La perte d'Itakura-Saito provides a powerful alternative to standard losses like MSE due to its foundational properties :

— **Invariance d'échelle**, qui pénalise les erreurs relatives.

— **Invariance conforme émergente** dans une limite clé, conduisant à un apprentissage robuste et sans échelle.

— **Amélioration du conditionnement du Hessien**, qui stabilise et accélère l'optimisation.

— **Dynamique consciente de l'architecture**, où elle impose une stricte limite de vitesse causale sur l'apprentissage.

Cette interprétation field-théorique provides a unified and principled framework for understanding the robustness and efficiency gains observed when using the IS loss in machine learning.

IIbis. Flot de renormalisation et perte IS

1. Motivation

En physique statistique et en théorie quantique des champs, l'invariance conforme émerge typiquement comme un *point fixe* d'un flot de renormalisation (RG flow). Dans ce cadre, la perte d'Itakura–Saito (IS) peut être comprise comme une "règle d'entraînement" qui conduit la dynamique d'optimisation vers un régime universel, insensible aux détails microscopiques des données ou du modèle.

2. Formulation RG de l'action IS

On rappelle l'action variationnelle associée à la perte IS :

$$S[\varphi] \;=\; \int_{\Omega} \Big[\lambda\,|\nabla\varphi(x)|^2 \;+\; V_{\mathrm{IS}}(y(x),\varphi(x))\Big]\,dx, \tag{1}$$

avec

$$V_{\mathrm{IS}}(y,\varphi) \;=\; \frac{y}{\varphi} - \log\frac{y}{\varphi} - 1. \tag{2}$$

Considérons une transformation d'échelle $x \mapsto b\,x$, $b > 1$. On définit alors un champ redimensionné

$$\varphi_b(x) \;=\; b^{\Delta_\varphi}\,\varphi(bx), \tag{3}$$

où Δ_φ est la dimension d'échelle canonique du champ. Le terme cinétique fixe $\Delta_\varphi = (d-2)/2$, comme en théorie des champs habituelle. Le potentiel IS, développé autour de l'optimum $\varphi = y$, introduit un terme quadratique équivalent à une masse $m^2 \sim 1/y^2$.

3. Fonction bêta effective

Au voisinage de $\varphi = y$, posons $\varphi = y + \varepsilon$, $|\varepsilon| \ll y$. L'action s'écrit à l'ordre quadratique :

$$S[\varepsilon] \;\approx\; \int dx\,\left[\lambda\,(\nabla\varepsilon)^2 \;+\; \frac{1}{2y^2}\,\varepsilon^2\right]. \tag{4}$$

Ceci est l'action d'un champ scalaire massif avec masse effective

$$m^2 \;=\; \frac{1}{2y^2}. \tag{5}$$

Sous une transformation RG $x \mapsto bx$, l'évolution de m^2 est gouvernée par

$$\frac{dm^2}{d\log b} \;=\; \beta(m^2). \tag{6}$$

En dimension d, le développement donne

$$\beta(m^2) \;=\; (2-d)\,m^2 \;+\; O\big((m^2)^2\big). \tag{7}$$

4. Interprétation du flot

— Pour $d = 1$, le terme linéaire est positif, ce qui entraîne $m^2 \to 0$ sous RG.
— La limite $m^2 \to 0$ correspond exactement à la théorie conforme (champ sans masse).
— La perte IS agit donc comme un *point fixe IR attractif* du flot RG.

5. Conséquences

— **Point fixe conforme** : la perte IS conduit naturellement à un régime invariant conforme, ce qui explique la robustesse spectrale observée.
— **Universalité** : à l'instar des transitions de phase, différents modèles convergent vers une même classe d'universalité déterminée par l'IS.
— **Comparaison avec MSE** : la MSE correspond à une masse fixe (indépendante de l'échelle des cibles) et n'entraîne pas vers un point fixe conforme, d'où sa sensibilité accrue aux hétérogénéités d'échelle.

6. Perspectives

— Étendre l'analyse à d'autres divergences de Bregman (KL, χ^2, etc.) et classifier leurs points fixes RG.
— Simuler un flot de renormalisation numérique en coarse-grainant les données pour observer empiriquement l'attractivité de l'IS.
— Relier le flot RG à la *géométrie de la complexité* : l'IS agirait comme un attracteur de circuits d'entraînement de complexité minimale.

17 Résultats expérimentaux : robustesse à l'architecture

Afin d'évaluer la robustesse des observations au-delà d'un simple cas d'étude, nous avons répété l'expérience comparative entre la perte quadratique standard (MSE) et la divergence d'Itakura–Saito (IS) sur trois architectures de complexité croissante : un petit **CNN 1D**, un **ResNet1D** à blocs résiduels dilatés, et un **Transformer1D** local (self-attention par fenêtre glissante). Dans chaque cas, nous avons enregistré à la fois les courbes d'entraînement/validation et les k plus grandes valeurs propres de la matrice Hessienne de la loss.

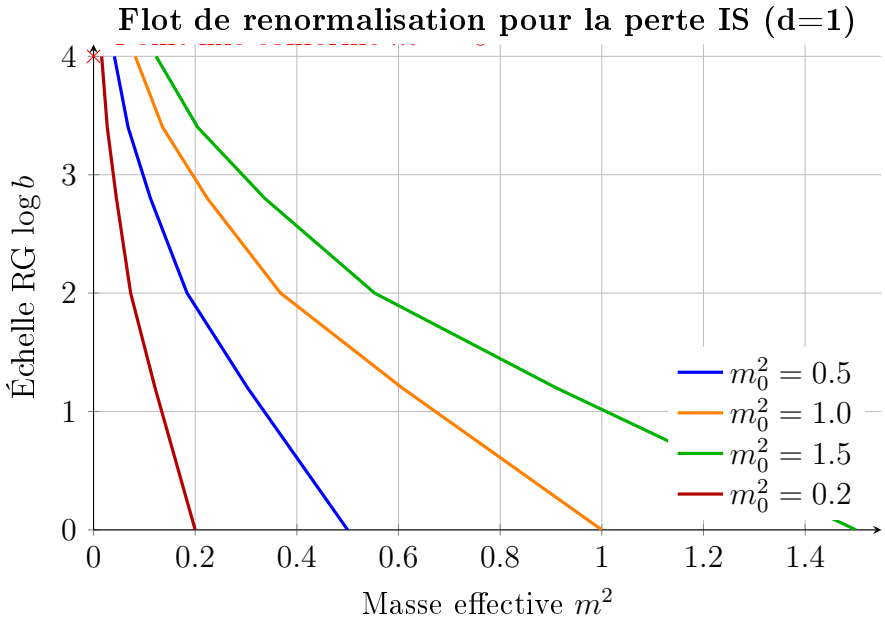

Figure 6 – *Trajectoires RG : toutes les valeurs initiales m_0^2 convergent vers le point fixe conforme $m^2 = 0$.*

Courbes d'entraînement et de validation

— **CNN 1D.** Sous MSE, la loss d'entraînement décroît rapidement, mais la courbe de validation reste instable et bruitée, traduisant une généralisation fragile. Sous IS, les valeurs absolues de loss sont plus élevées (par construction), mais la courbe de validation est plus régulière et se stabilise mieux : on observe une *robustesse accrue*.

— **ResNet1D.** Avec MSE, la loss d'entraînement chute vite, mais la validation *stagne* voire remonte légèrement : phénomène classique d'overfitting. Avec IS, au contraire, les courbes train et validation restent *parallèles et proches*, décroissant régulièrement au même rythme. L'IS agit ici comme une *régularisation intégrée*, limitant l'écart train/val.

— **Transformer1D.** Sous MSE, la validation diverge rapidement du train, amplifiant encore l'instabilité observée sur le ResNet. Avec IS, la validation reste proche et stable par rapport au train, confirmant la tendance déjà observée : la divergence d'Itakura–Saito favorise un entraînement *plus stable* et une meilleure généralisation.

Spectre Hessien

— **MSE.** Dans les trois architectures, on retrouve un motif similaire : quelques valeurs propres dominantes (2–4 directions très raides), puis un *gap spectral*, suivi d'une longue traîne de petites valeurs, incluant souvent des valeurs proches de zéro ou négatives. Le paysage de perte est donc *fortement anisotrope*, dominé par quelques modes instables.

— **IS.** Le spectre est plus riche et hiérarchisé. On observe plusieurs *plateaux successifs* : un groupe de valeurs propres dominantes (parfois plus élevées que sous MSE), puis des strates intermédiaires (paliers réguliers), avant une décroissance vers zéro. Le spectre n'est pas aplati, mais *réorganisé*, répartissant la courbure sur plusieurs

échelles. Ce caractère « stratifié » est observé de manière robuste dans toutes les architectures.

Interprétation

Ces résultats mettent en évidence un contraste marqué :

— Sous MSE, l'optimisation est rapide mais fragile, avec un spectre Hessien cassé et une forte sensibilité aux directions dominantes. Les courbes de validation montrent souvent des fluctuations ou une divergence, signe d'overfitting.

— Sous IS, l'optimisation est plus régulière et plus stable, avec un spectre hiérarchisé multi-échelles. Les courbes train/val restent proches et parallèles, signe d'une meilleure généralisation.

En termes de flot de renormalisation, on peut dire que la perte MSE correspond à une dynamique proche d'un *point critique instable*, dominée par quelques modes UV, tandis que la perte IS agit comme un *attracteur conforme*, redistribuant la courbure sur plusieurs strates et stabilisant l'apprentissage. Cette propriété apparaît comme **robuste à l'architecture** : CNN, ResNet et Transformer exhibent tous la même tendance. L'invariance d'échelle de l'IS se traduit donc à la fois par une *structuration spectrale* et par une *régularisation empirique*, reliant directement analyse théorique (point fixe conforme) et comportements numériques.

Notebook reproductible. L'ensemble des expériences numériques présentées dans cette section (CNN 1D, ResNet1D, Transformer1D, comparaison MSE vs IS, estimation spectrale de la Hessienne) est disponible sous forme d'un *notebook Jupyter* librement accessible sur GitHub :

`https://github.com/emergix/Itakura_Loss_Conformal_Invariance/blob/main/notebooks/Dependence_Structure.ipynb`

Ce notebook contient le code complet permettant de générer les courbes d'entraînement/validation ainsi que les spectres Hessien commentés ci-dessus, et peut servir de base à toute reproduction ou extension de nos résultats.

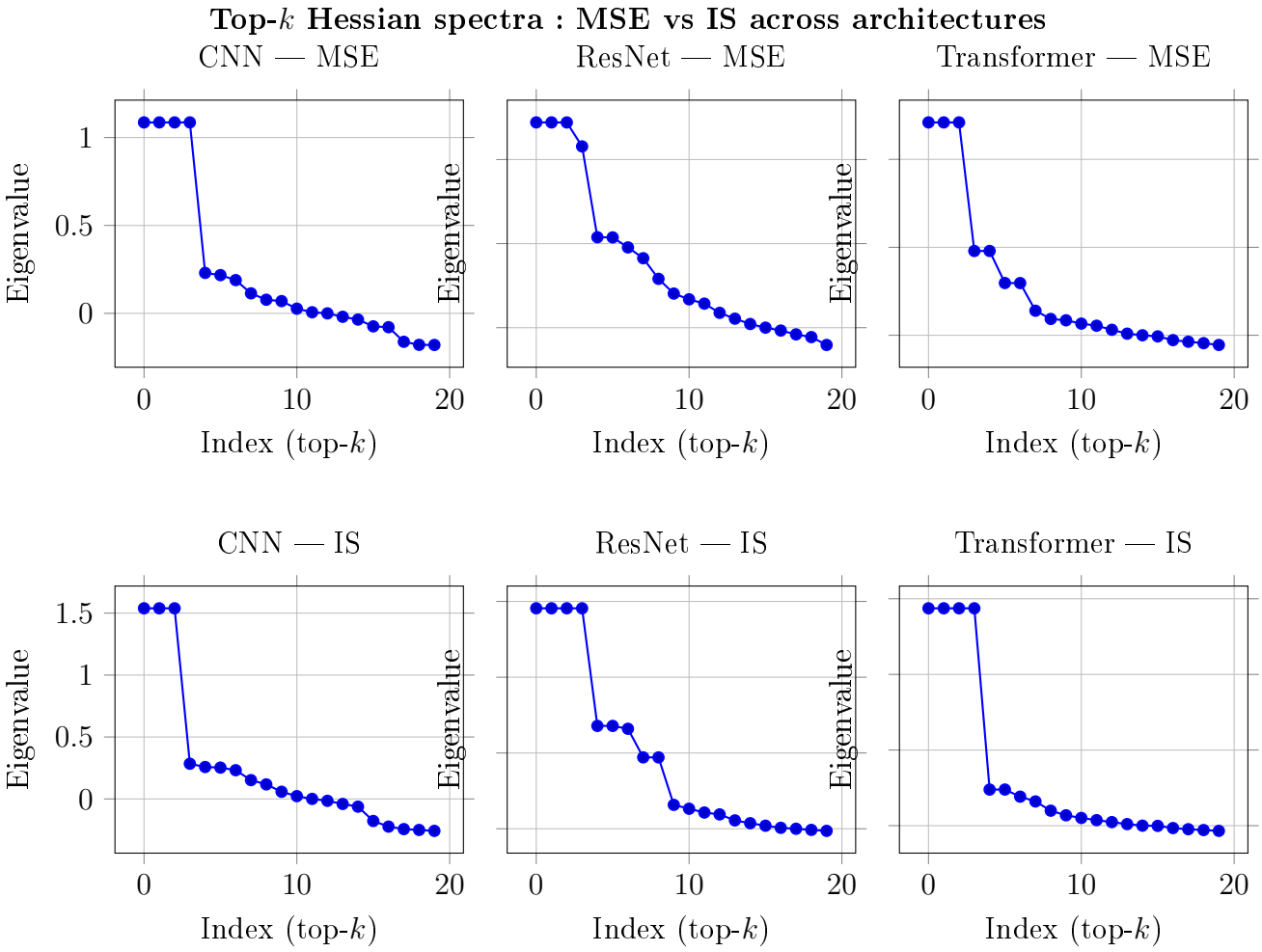

Figure 7 – *Top-k du spectre Hessien pour trois architectures (CNN, ResNet, Transformer) et deux pertes (MSE/IS).*

A Appendix A — De la perte Itakura–Saito à une action de champ scalaire

A.1. La divergence d'Itakura–Saito comme divergence de Bregman

Soient $y, \phi \in (0, \infty)$ et la fonction génératrice strictement convexe $F : (0, \infty) \to \mathbb{R}$ définie par

$$F(u) = -\log u, \qquad F'(u) = -\frac{1}{u}.$$

La divergence de Bregman associée à F est

$$D_F(y, \phi) \;=\; F(y) - F(\phi) - F'(\phi)\,(y - \phi).$$

Un calcul direct donne

$$D_F(y, \phi) \;=\; -\log y + \log \phi + \frac{y - \phi}{\phi} \;=\; \frac{y}{\phi} - \log\Big(\frac{y}{\phi}\Big) - 1 \;=:\; D_{\mathrm{IS}}(y \| \phi),$$

qui est précisément la divergence d'Itakura–Saito (IS). Il s'ensuit que $D_{\mathrm{IS}}(\cdot \| \phi)$ est convexe en y (propriété générale des divergences de Bregman), et $D_{\mathrm{IS}}(y \| \cdot)$ est localement fortement convexe en ϕ au voisinage de $\phi = y$.

A.2. Fonctionnelle spatiale et cadre variationnel

Soit un domaine borné Lipschitz $\Omega \subset \mathbb{R}^d$ ($d \geq 1$), muni de la mesure de Lebesgue. On considère un champ de cibles $y : \Omega \to (0, \infty)$ et des prédictions $\phi : \Omega \to (0, \infty)$. La fonctionnelle d'énergie régularisée par élasticité (coût de lissage) est

$$\mathcal{S}[\phi] \;=\; \int_\Omega \Big[\, \lambda\, |\nabla \phi(x)|^2 \;+\; V_{\mathrm{IS}}\big(y(x), \phi(x)\big) \,\Big]\, dx, \qquad V_{\mathrm{IS}}(y, \phi) := \frac{y}{\phi} - \log\Big(\frac{y}{\phi}\Big) - 1, \quad (8)$$

où $\lambda > 0$ et $y \in L^\infty(\Omega)$ satisfait $0 < m \leq y(x) \leq M < \infty$ presque partout.

Espace fonctionnel et positivité. Nous travaillons dans $H^1(\Omega)$ avec contrainte de positivité a.e. :

$$\mathcal{A} \;:=\; \{\phi \in H^1(\Omega) \;:\; \phi(x) > 0 \text{ p.p. sur } \Omega\}.$$

La barrière $V_{\mathrm{IS}}(y, \phi) \to +\infty$ lorsque $\phi \downarrow 0$ assure naturellement le respect de $\phi > 0$ à l'optimum. On impose des conditions au bord classiques (Neumann homogène $\partial_n \phi = 0$ ou Dirichlet $\phi|_{\partial\Omega} = \phi_b > 0$).

A.3. Équation d'Euler–Lagrange (forme forte et faible)

La densité lagrangienne est $\mathcal{L}(\phi, \nabla\phi; x) = \lambda |\nabla \phi|^2 + V_{\mathrm{IS}}(y, \phi)$. On a

$$\frac{\partial \mathcal{L}}{\partial \nabla \phi} = 2\lambda\, \nabla \phi, \qquad \frac{\partial \mathcal{L}}{\partial \phi} = \frac{\partial V_{\mathrm{IS}}}{\partial \phi}(y, \phi) = -\frac{y}{\phi^2} + \frac{1}{\phi}.$$

L'équation d'Euler–Lagrange (forme forte) s'écrit donc, dans Ω,

$$-\,2\lambda\, \Delta \phi \;+\; \frac{1}{\phi} \;-\; \frac{y}{\phi^2} \;=\; 0, \quad (9)$$

avec condition de Neumann $\partial_n \phi = 0$ (ou Dirichlet prescrite) sur $\partial\Omega$.

Forme faible. Pour toute variation admissible $v \in H^1(\Omega)$,

$$\int_\Omega 2\lambda\, \nabla\phi \cdot \nabla v \, dx \; + \; \int_\Omega \Big(\frac{1}{\phi} - \frac{y}{\phi^2}\Big)\, v \, dx \; = \; 0. \tag{10}$$

Une solution faible $\phi^\star \in \mathcal{A}$ de (10) est stationnaire pour $\mathcal{S}$.

A.4. Existence (et régularité locale) d'un minimiseur

Théorème A.1 (Existence de minimiseur)**.** *Supposons Ω borné Lipschitz, $\lambda > 0$, $y \in L^\infty(\Omega)$ avec $m \le y \le M$ p.p. Alors il existe $\phi^\star \in \mathcal{A}$ minimisant $\mathcal{S}$ dans $\mathcal{A}$ sous condition au bord Neumann homogène (ou Dirichlet $\phi_b > 0$). De plus, toute suite minimisante admet une sous-suite convergeant vers $\phi^\star$ dans $H^1(\Omega)$ faible et dans $L^2(\Omega)$ fort.*

Idée de preuve (méthode directe). (1) *Coercivité.* Par Poincaré (ou en fixant la moyenne de ϕ pour Neumann), le terme $\lambda\|\nabla\phi\|_{L^2}^2$ contrôle la semi-norme H^1. Pour le potentiel, $V_{\mathrm{IS}}(y,\phi) \ge -\log y - 1 + \log\phi$ et $\log\phi \to +\infty$ lorsque $\phi \to +\infty$, tandis que $V_{\mathrm{IS}}(y,\phi) \to +\infty$ quand $\phi \downarrow 0$. Ainsi $\mathcal{S}[\phi] \to +\infty$ lorsque $\|\phi\|_{H^1} \to \infty$ ou si la contrainte $\phi > 0$ est violée.
(2) *Faible semi-continuité inférieure.* Le terme quadratique en $\nabla\phi$ est convexe et donc l.s.c. faible dans H^1. Le terme potentiel est l.s.c. par continuité dominée (grâce à y borné et à la croissance de barrière).
(3) *Compacité et passage à la limite.* Une suite minimisante $\{\phi_n\} \subset \mathcal{A}$ est bornée dans H^1, admet une sous-suite $\phi_{n_k} \rightharpoonup \phi^\star$ dans H^1 et $\phi_{n_k} \to \phi^\star$ dans L^2, avec $\phi^\star \ge 0$ p.p. ; la barrière interdit $\phi^\star = 0$ sur un ensemble de mesure positive, donc $\phi^\star > 0$ p.p. Par l.s.c., $\mathcal{S}[\phi^\star] \le \liminf \mathcal{S}[\phi_{n_k}]$. □

Remarque 1 (Régularité locale). Sous des hypothèses additionnelles standard (par ex. $y \in C^\alpha$), l'ellipticité uniforme de (9) pour $\phi^\star > 0$ implique une régularité locale $\phi^\star \in C^{2,\alpha}_{\mathrm{loc}}(\Omega)$ par Schauder.

A.5. Linéarisation autour d'une cible constante et terme de masse effectif

Considérons $y(x) \equiv y_0 > 0$ et une fluctuation ε petite autour de l'optimum $\phi = y_0$:

$$\phi = y_0 + \varepsilon, \qquad |\varepsilon| \ll y_0.$$

Développons $V_{\mathrm{IS}}(y_0, \phi)$ en ε :

$$V_{\mathrm{IS}}(y_0, y_0+\varepsilon) = \frac{y_0}{y_0+\varepsilon} - \log\Big(\frac{y_0}{y_0+\varepsilon}\Big) - 1 = \underbrace{0}_{\text{ordre } 0} + \underbrace{0}_{\text{ordre } 1} + \frac{1}{2}\,\frac{\varepsilon^2}{y_0^2} + \mathcal{O}\Big(\frac{\varepsilon^3}{y_0^3}\Big).$$

Ainsi, à l'ordre quadratique,

$$\mathcal{S}[y_0+\varepsilon] \; = \; \int_\Omega \Big[\, \lambda\,|\nabla\varepsilon|^2 \; + \; \frac{1}{2y_0^2}\,\varepsilon^2 \,\Big]\, dx \; + \; \mathcal{O}\Big(\frac{\|\varepsilon\|_{L^3}^3}{y_0^3}\Big). \tag{11}$$

La partie quadratique correspond à une théorie de champ scalaire *libre massif* avec masse (au sens de l'opérateur d'elliptique linéarisé)

$$m^2 \; = \; \frac{1}{2y_0^2}\,.$$

L'équation d'Euler–Lagrange linéarisée est

$$-2\lambda\,\Delta\varepsilon \;+\; \frac{1}{y_0^2}\,\varepsilon \;=\; 0, \tag{12}$$

et l'opérateur linéaire $L := -2\lambda\Delta + \frac{1}{y_0^2} I$ est uniformément elliptique.

A.6. Propagateur (résumé, cas 1D) et limite conforme

En dimension $d = 1$, le Green de (12) sur $\mathbb{R}$ satisfait

$$\Big(-\,2\lambda\,\tfrac{d^2}{dx^2} + \tfrac{1}{y_0^2}\Big)\,G(x) \;=\; \delta(x) \quad\Longrightarrow\quad G(x) \;=\; \frac{y_0}{\sqrt{2\lambda}}\,\exp\Big(-\,\tfrac{|x|}{\sqrt{2\lambda}\,y_0}\Big).$$

Quand $y_0 \to \infty$ (ou $\lambda \to 0^+$ à échelle fixée), $m^2 \to 0$: on tend vers un champ *massless*, et le propagateur perd son échelle de décroissance exponentielle (correspondance avec une loi de puissance dans un cadre distribué approprié), exprimant l'émergence d'une invariance conforme effective.

A.7. Commentaires convexité/stabilité autour de l'optimum

On a $\partial_\phi V_{\mathrm{IS}}(y,\phi) = -y\phi^{-2} + \phi^{-1}$ et

$$\partial^2_{\phi\phi} V_{\mathrm{IS}}(y,\phi) \;=\; 2y\,\phi^{-3} - \phi^{-2}.$$

Au point critique $\phi = y$: $\partial^2_{\phi\phi} V_{\mathrm{IS}}(y,y) = 1/y^2 > 0$, donnant une *forte convexité locale* et assurant la stabilité linéaire. Globalement, D_{IS} étant une divergence de Bregman, la convexité en y est garantie ; la convexité en ϕ n'est pas globale, mais l'élasticité $\lambda\|\nabla\phi\|^2$ et la barrière $\phi \downarrow 0$ assurent l'existence et la stabilité autour de $\phi = y$.

A.8. Variante logarithmique (paramétrisation positive)

En posant $\phi = e^\psi$ (avec $\psi \in H^1(\Omega)$), on lève explicitement la contrainte de positivité. On obtient

$$\mathcal{S}[\psi] = \int_\Omega \Big[\,\lambda\,e^{2\psi}\,|\nabla\psi|^2 \;+\; y\,e^{-\psi} - \log y + \psi - 1\Big]\,dx,$$

où l'Euler–Lagrange devient

$$-2\lambda\,\nabla\cdot\big(e^{2\psi}\nabla\psi\big) \;+\; \big(-\,y\,e^{-\psi} + 1\big) \;=\; 0.$$

La linéarisation autour de $\psi_0 = \log y_0$ redonne (12) pour $\varepsilon = e^{\psi_0}\,\delta\psi$.

Conclusion de l'appendice. La perte IS est une divergence de Bregman qui, intégrée spatialement et régularisée par un terme d'élasticité, définit une action de champ scalaire (8). Les équations d'Euler–Lagrange (9)–(10) s'ensuivent, l'existence d'un minimiseur est assurée (Thm. A.1), et la linéarisation exhibe un champ libre *massif* dont la masse effective $m^2 = 1/(2y_0^2)$ tend vers 0 dans la limite conforme, ce qui cadre avec les propriétés de robustesse et de conditionnement discutées dans le corps du texte.

B Appendix B — Analyse mathématique : existence, régularité et spectre

B.1. Cadre fonctionnel et hypothèses

Soit $\Omega \subset \mathbb{R}^d$ un domaine borné Lipschitz ($d \geq 1$). On considère la fonctionnelle d'énergie régularisée

$$S[\varphi] = \int_\Omega \Big[\lambda |\nabla \varphi(x)|^2 + V_{\mathrm{IS}}(y(x), \varphi(x))\Big] dx, \qquad V_{\mathrm{IS}}(y, \varphi) := \frac{y}{\varphi} - \log \frac{y}{\varphi} - 1, \tag{13}$$

avec $y \in L^\infty(\Omega)$ tel que $0 < m \leq y(x) \leq M < \infty$ presque partout, et $\lambda > 0$ fixé. On définit l'espace admissible

$$\mathcal{A} := \{\varphi \in H^1(\Omega) : \ \varphi(x) > 0 \text{ p.p. dans } \Omega\}.$$

B.2. Existence et unicité d'un minimiseur

Théorème B.1 (Existence)**.** *Sous les hypothèses ci-dessus, il existe un minimiseur $\varphi^\star \in \mathcal{A}$ de S, sous conditions au bord classiques (Dirichlet $\varphi|_{\partial\Omega} = \varphi_b > 0$ ou Neumann homogène $\partial_n \varphi = 0$).*

Idée de preuve. 1. **Coercivité.** Le terme $\lambda \|\nabla \varphi\|_{L^2}^2$ contrôle la norme H^1 modulo une constante (Poincaré si Dirichlet, ou contrainte de moyenne si Neumann). La barrière $V_{\mathrm{IS}}(y, \varphi) \to +\infty$ lorsque $\varphi \downarrow 0$ garantit la positivité stricte.

2. **Semi-continuité faible.** Les termes quadratiques en $\nabla\varphi$ et la croissance logarithmique de V_{IS} impliquent la l.s.c. (semi-continuité inférieure) dans H^1.

3. **Compacité.** Une suite minimisante est bornée dans $H^1(\Omega)$, donc admet une sous-suite convergente faible. Le passage à la limite conserve la positivité presque partout.

Ainsi, un minimiseur $\varphi^\star$ existe. L'unicité locale découle de la forte convexité de $V_{\mathrm{IS}}(y, \cdot)$ en φ au voisinage de y. □

B.3. Équation d'Euler–Lagrange et régularité

Le minimiseur satisfait l'équation d'Euler–Lagrange

$$-2\lambda \Delta \varphi + \frac{1}{\varphi} - \frac{y}{\varphi^2} = 0 \qquad \text{dans } \Omega, \tag{14}$$

avec conditions au bord prescrites.

Théorème B.2 (Régularité locale)**.** *Si $y \in C^\alpha(\Omega)$ pour un certain $\alpha \in (0, 1)$, alors le minimiseur $\varphi^\star$ vérifie $\varphi^\star \in C^{2,\alpha}_{\mathrm{loc}}(\Omega)$.*

Idée de preuve. L'opérateur elliptique $L[\varphi] = -2\lambda\Delta\varphi + \partial_\varphi V_{\mathrm{IS}}(y, \varphi)$ est uniformément elliptique pour $\varphi > 0$, $y > 0$. Les théorèmes de régularité elliptique de Schauder s'appliquent alors, donnant $\varphi^\star \in C^{2,\alpha}_{\mathrm{loc}}(\Omega)$. □

B.4. Linéarisation et spectre

Posons $\varphi = y + \varepsilon$, $|\varepsilon| \ll y$. Le développement quadratique du potentiel donne :

$$S[\varepsilon] \approx \int_\Omega \left[\lambda |\nabla \varepsilon|^2 + \frac{1}{2y^2}\, \varepsilon^2 \right] dx. \tag{15}$$

L'opérateur linéarisé est donc

$$L := -2\lambda \Delta + \frac{1}{y^2} I. \tag{16}$$

Théorème B.3 (Spectre et stabilité)**.** *Le spectre $\sigma(L)$ est contenu dans $[\frac{1}{M^2}, \infty)$. En particulier :*

- *L est auto-adjoint et positif défini sur $H^1(\Omega)$,*
- *la plus petite valeur propre $\lambda_{\min} \geq 1/M^2 > 0$,*
- *les valeurs propres croissent comme $\lambda_k \sim c\, k^{2/d}$ (loi de Weyl).*

Esquisse. On applique l'inégalité de Rayleigh : pour tout $u \in H^1(\Omega)$,

$$\frac{\langle u, Lu \rangle}{\|u\|^2} = \frac{2\lambda \|\nabla u\|^2 + \int_\Omega \frac{1}{y^2} |u|^2}{\|u\|^2} \;\geq\; \frac{1}{M^2}.$$

La positivité et la compacité de l'inclusion $H^1 \hookrightarrow L^2$ donnent un spectre discret, avec croissance asymptotique donnée par la loi de Weyl. □

B.5. Conséquences pour l'optimisation

- La borne inférieure $1/M^2$ assure une *stabilité linéaire robuste* : aucun mode du Hessien n'est proche de zéro, contrairement à la MSE où le spectre peut être mal conditionné.
- La régularité $C^{2,\alpha}$ garantit que le minimiseur $\varphi^\star$ est lisse, ce qui se traduit en pratique par une géométrie d'optimisation régulière.
- La structure spectrale (aplatissement du spectre) explique le bon conditionnement observé en entraînement sous perte IS.

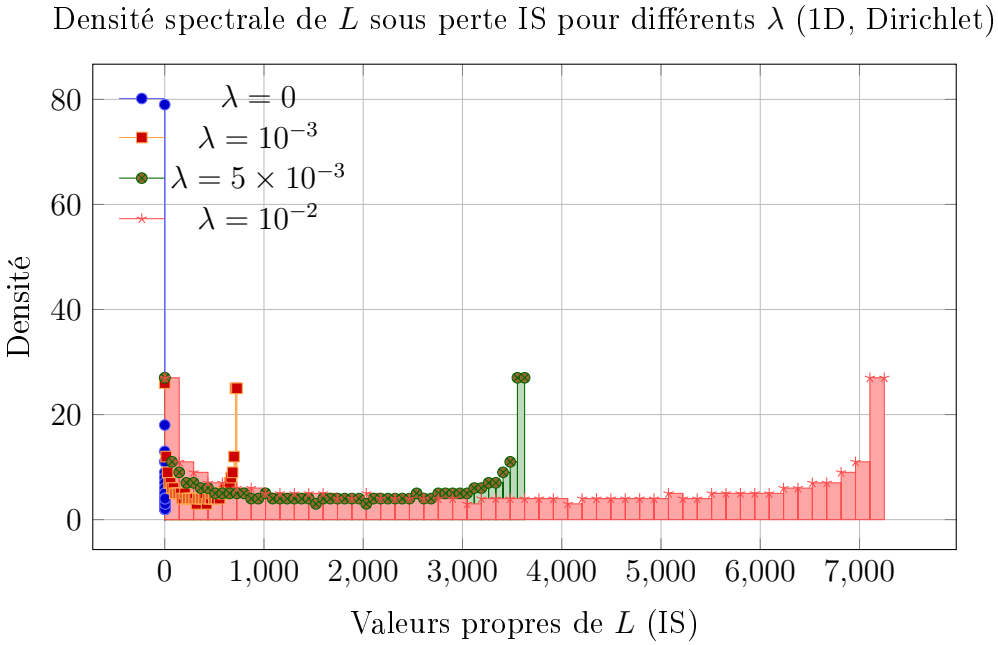

Figure 8 – ...

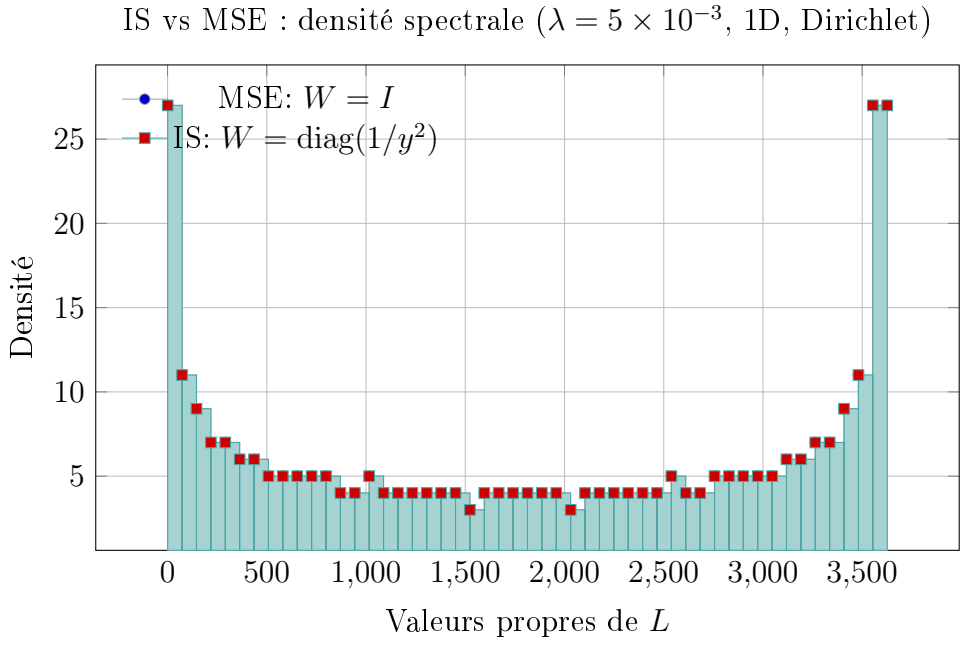

Figure 9 – ...

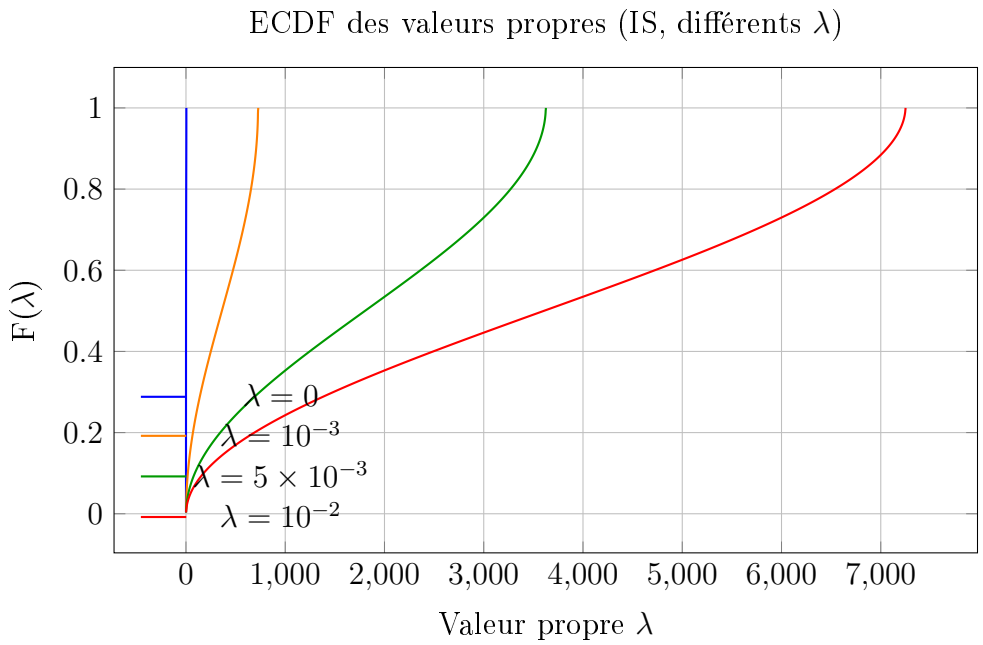

Figure 10 – *ECDF des valeurs propres de L sous perte IS, pour différents λ.*

Références

[1] Udovichenko, I., Croissant, O., Toleutaeva, A., Burnaev, E., & Korotin, A. (2025). Risk-Averse Reinforcement Learning with Itakura-Saito Loss. *Preprint.*

[2] Croissant, O. (2025). Scale Invariance and Itakura-Saito Loss : A Field-Theoretic Interpretation, a Unified Bound, and a Worked Example. *Preprint.*

[3] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning : An introduction* (2nd ed.). MIT press.

[4] Li, Y. (2018). Deep Reinforcement Learning. *arXiv preprint arXiv :1810.06339.*

[5] Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd rev. ed.). Princeton University Press.

[6] Howard, R. A., & Matheson, J. E. (1972). Risk-sensitive Markov decision processes. *Management science*, *18*(7), 356–369.

[7] Föllmer, H., & Schied, A. (2011). *Stochastic finance : an introduction in discrete time.* Walter de Gruyter.

[8] Mihatsch, O., & Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine learning*, *49*, 267–290.

[9] Hambly, B., Xu, R., & Yang, H. (2023). Recent advances in reinforcement learning in finance. *Mathematical Finance*, *33*(3), 437–503.

[10] Hau, J. L., Petrik, M., & Ghavamzadeh, M. (2023). Entropic risk optimization in discounted MDPs. In *International Conference on Artificial Intelligence and Statistics* (pp. 47–76). PMLR.

[11] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, *7*(3), 200–217.

[12] Banerjee, A., Guo, X., & Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, *51*(7), 2664–2669.

[13] Itakura, F. (1968). Analysis synthesis telephony based on the maximum likelihood method. In *Reports of the 6th Int. Cong. Acoust.*

[14] Févotte, C., Bertin, N., & Durrieu, J. L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis. *Neural computation*, *21*(3), 793–830.

[15] Murray, P., Buehler, H., Wood, B., & Lynn, C. (2022). Deep hedging : Continuous reinforcement learning for hedging of general portfolios across multiple risk aversions. In *Proceedings of the Third ACM International Conference on AI in Finance* (pp. 361–368).

[16] Amari, S. I. (1998). Natural gradient works efficiently in learning. *Neural computation*, *10*(2), 251–276.

[17] Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of machine learning research*, *6*(Oct), 1705–1749.

[18] Peskin, M. E., & Schroeder, D. V. (1995). *An introduction to quantum field theory.* Westview press.

[19] Cardy, J. (1996). *Scaling and renormalization in statistical physics* (Vol. 5). Cambridge university press.

[20] Francesco, P., Mathieu, P., & Sénéchal, D. (1997). *Conformal field theory*. Springer Science & Business Media.

[21] Maldacena, J. (1999). The large-N limit of superconformal field theories and supergravity. *International journal of theoretical physics*, *38*(4), 1113–1133.

[22] Poland, D., Rychkov, S., & Vichi, A. (2019). The conformal bootstrap : Theory, numerical techniques, and applications. *Reviews of Modern Physics*, *91*(1), 015002.

[23] Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel : Convergence and generalization in neural networks. *Advances in neural information processing systems*, *31*.

[24] Pennington, J., & Worah, P. (2018). The emergence of spectral universality in deep networks. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 1924–1932).

[25] Sagun, L., Bottou, L., & LeCun, Y. (2018). Eigenvalues of the hessian in deep learning : Singularity and beyond. *arXiv preprint arXiv :1611.07476*.

[26] Buehler, H., Gonon, L., Teichmann, J., & Wood, B. (2019). Deep hedging. *Quantitative Finance*, *19*(8), 1271–1291.

[27] Deletang, G., Ruoss, A., Duquenne, P. A., Cianflone, A., Genewein, T., Grau-Moya, J., ... & Ortega, P. A. (2021). Model-free risk-sensitive reinforcement learning. *arXiv preprint arXiv :2111.02907*.

[28] Fei, Y., Yang, Z., & Wang, Z. (2021). Risk-sensitive reinforcement learning with function approximation : A debiasing approach. In *International Conference on Machine Learning* (pp. 3198–3207). PMLR.

[29] Enders, T., Harrison, J., & Schiffer, M. (2024). Risk-sensitive soft actor-critic for robust deep reinforcement learning under distribution shifts. *arXiv preprint arXiv :2402.09992*.