

# Invariance d'échelle et perte d'Itakura-Saito : Une interprétation field-théorique, une borne unifiée et un exemple détaillé

Olivier Croissant

2025

## Résumé

Ce travail explore les propriétés géométriques et statistiques de la perte d'Itakura-Saito (IS) à travers le prisme de la théorie des champs. Nous démontrons que minimiser la perte IS équivaut à minimiser un fonctionnel d'énergie pour un champ de prédiction. Une analyse linéarisée révèle que ce fonctionnel correspond à une théorie de champ scalaire massif, qui devient invariant conforme dans la limite des grandes valeurs cibles. Cette invariance améliore la robustesse et l'optimisation en aplatissant le spectre du Hessien. De plus, nous dérivons une borne fondamentale, dépendante de l'architecture, sur la vitesse de propagation de l'information dans les réseaux entraînés avec la perte IS, prouvant que bien que l'invariance conforme améliore le conditionnement, elle ne peut violer cette limite de vitesse causale. La théorie est unifiée avec la dynamique du Noyau Tangent Neural (NTK) et illustrée par un exemple concret.

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>I</b>	<b>Fondation Field-Théorique</b>	<b>2</b>
<b>2</b>	<b>De la perte à la théorie des champs</b>	<b>2</b>
<b>3</b>	<b>Régime linéarisé et terme de masse</b>	<b>3</b>
<b>4</b>	<b>Propagateur et limite conforme</b>	<b>3</b>
<b>5</b>	<b>Implications pour l'optimisation et la robustesse</b>	<b>4</b>
<b>6</b>	<b>Théorème de conditionnement spectral</b>	<b>4</b>
<b>7</b>	<b>Analogie holographique</b>	<b>4</b>
<b>II</b>	<b>Dynamique et limites de vitesse dans les réseaux</b>	<b>5</b>

8	Une borne unifiée sur la vitesse de propagation dans les réseaux	5
8.1	Interprétations et conséquences . . . . .	5
9	Dynamique NTK avec perte IS	5
10	Un cône de lumière discret pour l'influence des prédictions	6
11	Stabilité spectrale et taux de convergence	6
12	Queues exponentielles au-delà de la localité stricte	7
III	Synthèse et exemple	7
13	Synthèse : Causalité, conditionnement et invariance conforme	7
14	Un exemple concret 1D (Noyau tridiagonal)	7
15	Conclusion de l'exemple	8
16	Conclusion générale	8
	Références	8

## 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 Field-Théorique

## 2 De la perte à la théorie des champs

La perte IS peut être naturellement reformulée comme un fonctionnel 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  $\lambda$  est un paramètre de régularisation. Cette formulation variationnelle cadre la minimisation de la perte IS comme la recherche d'un champ  $\varphi$  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 :

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

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  $G(x)$ , 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  $y_0 \rightarrow \infty$  (ou équivalamment  $\lambda \rightarrow 0$ ), la masse s'annule  $m^2 \rightarrow 0$ , 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.

## 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) :

$$L_{\text{MSE}}(\theta) = \frac{1}{2} \|f(\theta) - y\|^2,$$

$$L_{\text{IS}}(\theta) = \sum_i \left( \frac{y_i}{f_i(\theta)} - \log \frac{y_i}{f_i(\theta)} - 1 \right).$$

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 adaptative implicite :

- **Grandes valeurs cibles**  $\Rightarrow$  **pénalité plus faible** sur l'erreur absolue (priorisant l'erreur relative).
- **Petites valeurs cibles**  $\Rightarrow$  **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

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  $G_{\text{conf}} \subset G$  le sous-ensemble où l'action induite par IS est invariante conforme. Pour une mesure spectrale  $\lambda(\varphi)$  du Hessien, nous avons :*

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

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

Dans la limite conforme ( $y_0 \rightarrow \infty$ ), 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  $C, \xi$  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  $f_\theta$ . 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  $\eta$ , 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  $W = \text{diag}(1/y_1^2, \dots, 1/y_n^2)$ . 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 :*

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

*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  $K$ , aplatisant 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**  $v$ .

## 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,  $R = 1$ ) :

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

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  $\eta$  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.

## 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*.