Invariance d'échelle et perte d'Itakura–Saito : Une interprétation théorie des champs, une borne unifiée et un exemple détaillé

Olivier Croissant (2025)

Contents

1 Introduction

Cette section résume et reformule les résultats des annexes D et E de « *Risk-Averse Reinforcement Learning with Itakura–Saito Loss* » par Udovichenko et al. (2025). Elle propose un exposé

autonome des interprétations géométrique, théorie des champs et statistique de la perte d'Itakura–Saito (IS), mettant en lumière comment l'invariance d'échelle et conforme améliore la robustesse et l'optimisation en apprentissage automatique.

2 De la perte à la théorie des champs

La perte IS peut être reformulée comme une fonctionnelle d'énergie. Soit $\varphi(x)$ un champ représentant les prédictions, et $y(x)$ la cible. Nous définissons l'action

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

où λ est un paramètre de régularisation. Cette formulation variationnelle souligne que minimiser la perte IS correspond à minimiser une fonctionnelle d'énergie avec des termes de cohérence locaux et globaux.

3 Régime linéarisé et terme de masse

Considérons un développement autour d'une cible constante $y(x) = y_0$. Soit $\varphi(x) = y_0 + \varepsilon(x)$ avec $\varepsilon \ll y_0$. Le potentiel IS se développe comme

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

L'action linéarisée s'écrit

$$S[\varepsilon] \approx \int dx \left[\lambda \left(\frac{d\varepsilon}{dx} \right)^2 + \frac{1}{2y_0^2} \varepsilon(x)^2 \right]. \tag{3}$$

Ceci est équivalent à une théorie des champs gaussienne avec une masse

$$m^2 = \frac{1}{2y_0^2}. \tag{4}$$

4 Propagateur et limite conforme

Le propagateur $G(x)$ satisfait

$$\left(-\lambda \frac{d^2}{dx^2} + \frac{1}{2y_0^2} \right) G(x) = \delta(x). \tag{5}$$

La solution en 1D est

$$G(x) = \frac{y_0}{\sqrt{2\lambda}} \exp\left(-\frac{|x|}{\sqrt{2\lambda} y_0} \right). \tag{6}$$

Dans la limite conforme $y_0 \to \infty$ (ou $\lambda \to 0$), ceci se réduit à

$$G(x) \sim \frac{1}{|x|}, \tag{7}$$

montrant l'absence d'échelle intrinsèque, des corrélations en loi de puissance, et une symétrie conforme émergente.

5 Implications pour l'optimisation

La perte IS a un conditionnement plus favorable que l'erreur quadratique moyenne (MSE) :

$$L_{\text{MSE}}(\theta) = \tfrac{1}{2}\|f(\theta) - y\|^2, \tag{8}$$

$$L_{\text{IS}}(\theta) = \sum_i \left(\frac{y_i}{f_i(\theta)} - \log \frac{y_i}{f_i(\theta)} - 1 \right). \tag{9}$$

Localement, si $f = y + \varepsilon$,

$$L_{\text{IS}}(f) \approx \frac{1}{2y^2}\varepsilon^2, \tag{10}$$

ce qui est équivalent à une MSE renormalisée, avec une pondération $\propto 1/y^2$. Ainsi :

- Grand $y \Rightarrow$ pénalité plus faible,
- Petit $y \Rightarrow$ pénalité plus forte.

Cette normalisation implicite agit comme une *descente de gradient naturelle*, améliorant la stabilité.

6 Conditionnement spectral et robustesse

Théorème (informel). Soit $\varphi \sim G$ un ensemble gaussien de modèles, et soit $G_{\text{conf}} \subset G$ le sous-ensemble où l'action induite par IS est conformément invariante. Définissons $\lambda(\varphi)$ comme une mesure spectrale du Hessien. Alors

$$\mathbb{E}_{\varphi \in G_{\text{conf}}}[\text{Var}(\lambda(\varphi))] < \mathbb{E}_{\varphi \in G}[\text{Var}(\lambda(\varphi))]\,. \tag{11}$$

Ainsi, l'invariance conforme réduit la variance spectrale, aplatissant le spectre du Hessien, et conduisant à une optimisation mieux conditionnée.

7 Analogie holographique

Dans la limite $y_0 \to \infty$, l'action induite par IS se réduit à

$$S[\varepsilon] = \int dx\, \lambda \left(\frac{d\varepsilon}{dx} \right)^2, \tag{12}$$

une théorie des champs scalaires sans masse (une CFT 1D). Celle-ci exhibe des corrélations à longue portée $G(x) \sim 1/|x|$. L'analogie avec AdS/CFT suggère que les prédictions au bord entraînées par IS imposent une cohérence globale dans l'espace de représentation latent « bulk », contribuant à la robustesse.

8 Borne de propagation généralisée dans les réseaux

Considérons un réseau (graphe, réseau ou circuit computationnel) représenté comme un ensemble de nœuds V et d'arêtes E, avec une distance métrique $d(x, y)$ entre les nœuds et des règles d'interaction locales. Chaque nœud $x \in V$ a un état $\varphi_x(t)$ évoluant dans le temps t selon une dynamique de mise à jour locale. Soit $\mathcal{O}_x$ une observable localisée au nœud x.

Nous nous intéressons à la vitesse maximale à laquelle l'influence, l'information ou les corrélations peuvent se propager à travers le réseau.

Théorème 1 (Borne de propagation unifiée pour les réseaux)**.** *Supposons que le réseau satisfait :*

(i) ***Localité :*** *les mises à jour de φ_x dépendent seulement des nœuds dans un voisinage borné de x ;*

(ii) ***Intensité d'interaction finie :*** *chaque mise à jour locale est bornée Lipschitz avec une constante g (énergie, bande passante, ou facteur de Lipschitz) ;*

(iii) ***Métrique bien définie :*** *le graphe admet une distance $d(x, y)$.*

Alors il existe une vitesse finie $v > 0$, appelée vitesse de propagation du réseau*, telle que pour deux observables quelconques $\mathcal{O}_x(t), \mathcal{O}_y(0)$ localisées aux nœuds $x, y \in V$, on ait*

$$\left|\langle \mathcal{O}_x(t), \mathcal{O}_y(0)\rangle\right| \;\leq\; C \exp\!\left(-\frac{d(x,y) - vt}{\xi}\right), \tag{13}$$

où C, ξ sont des constantes dépendant du système.

Interprétations

- Dans les **systèmes de spins quantiques**, ceci se réduit à la borne de Lieb–Robinson, avec v la vitesse de Lieb–Robinson.
- Dans les **réseaux de communication**, v correspond à la capacité maximale des arêtes et $d(x, y)$ à la longueur de chemin (diamètre du réseau).
- En **informatique distribuée / consensus**, v^{-1} est proportionnel à la gap spectral du Laplacien (connectivité algébrique).
- Dans les **circuits neuronaux ou booléens**, v correspond à *couches par unité de temps*, donc $d(x, y)/v$ donne la profondeur minimale de circuit pour l'influence.
- En **apprentissage automatique avec perte IS**, l'invariance conforme impose un conditionnement sans échelle, mais le spectre du Hessien contraint toujours v, empêchant une convergence arbitrairement rapide.

Conséquences

Ce théorème implique :

1. Aucun réseau, physique ou computationnel, ne peut propager l'information instantanément ; il y a toujours un « cône de lumière » fini d'influence.
2. L'invariance conforme/d'échelle améliore la robustesse en lissant la propagation et en aplatissant les spectres, mais ne supprime pas la borne fondamentale v.
3. La généralisation robuste dans les systèmes d'apprentissage peut ainsi être interprétée comme émergeant d'une *contrainte de causalité généralisée.*

9 Cadre : Dynamique NTK avec perte IS

Soit $\{(x_i, y_i)\}_{i=1}^n$ des échantillons d'entraînement et soit $f_\theta(x) \in \mathbb{R}$ la sortie scalaire d'un modèle. Désignons le vecteur des prédictions $f_\theta := (f_\theta(x_1), \dots, f_\theta(x_n))^\top \in \mathbb{R}^n$ et des cibles $y \in \mathbb{R}^n$.

Pour une donnée unique (y_i, f_i), la perte d'Itakura–Saito est

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

On calcule

$$\frac{\partial \ell_{\mathrm{IS}}}{\partial f_i} \;=\; \frac{f_i - y_i}{f_i^2}, \qquad \frac{\partial^2 \ell_{\mathrm{IS}}}{\partial f_i^2}\Big|_{f_i = y_i} \;=\; \frac{1}{y_i^2}.$$

Près de $f \approx y$, la perte est localement quadratique avec une courbure diagonale $W := \mathrm{diag}(1/y_1^2, \dots, 1/y_n^2)$ et un gradient $\nabla_f L \approx W(f - y)$.

Supposons un entraînement par descente de gradient (lot complet) avec un pas $\eta > 0$ et une *approche paresseuse*/linéarisation NTK :

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

où $K \in \mathbb{R}^{n \times n}$ est le noyau tangent neuronal (NTK) empirique à l'initialisation (maintenu fixe dans la linéarisation). Définissons l'*erreur de prédiction* $e_t := f_t - y$. Alors la dynamique de l'erreur est linéaire :

$$e_{t+1} \;=\; \Big(I - \eta A\Big)\, e_t, \qquad A \;:=\; KW. \tag{14}$$

Hypothèse de localité. Soient les points d'entraînement les nœuds d'un graphe (V, E) avec une distance de graphe $d(i, j)$. Nous supposons que K est *local de portée* R :

$$K_{ij} = 0 \quad \text{si } d(i, j) > R. \tag{15}$$

Ceci couvre les cas courants : architectures convolutionnelles sur grilles (champ réceptif fini), réseaux de neurones sur graphes avec passage de messages à R sauts, et noyaux localisés. Notez que W est diagonale et préserve donc les motifs de parcimonie lorsqu'elle est multipliée.

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

Nous étudions comment une perturbation à l'échantillon j à l'étape 0 peut influencer l'échantillon i après t étapes. Soit $J_t := \partial f_t / \partial f_0$ la Jacobienne de l'application de prédiction sous (14). Avec $e_t = (I - \eta A)^t e_0$ et $f_t = y + e_t$, nous avons $J_t = (I - \eta A)^t$.

Lemme 1 (Bande sous localité)**.** *Si K satisfait* (15) *avec une portée R et W est diagonale, alors pour tous les entiers $t \geq 0$*

$$A^t = (KW)^t \text{ est local de portée-}tR : \quad (A^t)_{ij} = 0 \quad \text{si } d(i, j) > tR.$$

Proof. Par (15), $A = KW$ a la même parcimonie que K (puisque W est diagonale). Les produits de matrices locales de portée R convoluent les portées additivement : si B et C sont locaux de portée R_B et R_C, alors $(BC)_{ij} = \sum_k B_{ik} C_{kj}$ peut être non nul seulement s'il existe k avec $d(i, k) \leq R_B$ et $d(k, j) \leq R_C$, donc $d(i, j) \leq R_B + R_C$ par l'inégalité triangulaire. Itérer donne que A^t est local de portée tR. □

Théorème 2 (Cône de lumière discret sous entraînement IS)**.** *Sous* (15)*, pour tout* $t \geq 0$*,*

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

Équivalemment, une perturbation unitaire au nœud j *à l'étape* 0 *ne peut affecter* $f_t(i)$ *si* $d(i,j) > tR$*.*

Proof. Développons $(I-\eta A)^t = \sum_{k=0}^{t} \binom{t}{k}(-\eta)^k A^k$. Par le Lemme 1, A^k est local de portée kR. Si $d(i,j) > tR$, alors pour tout $k \leq t$ nous avons $d(i,j) > kR$, donc $(A^k)_{ij} = 0$, et ainsi la somme a une entrée (i,j) nulle. □

Le Théorème 2 donne un *cône de lumière exact* avec une vitesse discrète $v = R$ nœuds par étape. Ainsi, la vitesse de propagation maximale est une propriété architecturale/localité ; elle est *indépendante* de la courbure ou de la taille du pas (sous réserve de stabilité ci-dessous).

11 Stabilité spectrale, taux de convergence et conditionnement du Hessien

Alors que v est fixée par la localité (R), la *vitesse* à laquelle les erreurs décroissent à l'intérieur du cône dépend du spectre de $A = KW$.

Lemme 2 (Stabilité linéaire)**.** *Si* $\rho(A)$ *désigne le rayon spectral de* A*, alors la descente de gradient* (14) *est linéairement stable si*

$$0 \;<\; \eta \;<\; \frac{2}{\lambda_{\max}(A)}.$$

Sous cette condition, $\|e_t\| \leq \kappa\,(1-\eta\lambda^{+}_{\min})^t\,\|e_0\|$ *dans toute norme compatible avec* A*, où* $\lambda^{+}_{\min}$ *est la plus petite valeur propre positive de* A *et* κ *dépend du conditionnement de la base propre.*

Rôle de la courbure IS. Près de $f \approx y$, le Hessien par rapport aux prédictions est $H_f \approx W = \mathrm{diag}(1/y_i^2)$. Ainsi $A = KW$ renormalise les lignes/colonnes de K par $1/y_i^2$, agissant comme un *préconditionneur implicite.* Si les magnitudes cibles $\{|y_i|\}$ varient largement, W *aplatit* le spectre effectif de A relativement à MSE $(W = I)$, réduisant le conditionnement et améliorant les taux de convergence—*sans* changer la portée de localité R (et donc sans augmenter v).

12 Queue exponentielle au-delà de la localité stricte

Dans de nombreuses architectures, K n'est pas strictement bandée mais a une décroissance rapide hors diagonale :

$$|K_{ij}| \;\leq\; C_0\, e^{-d(i,j)/\xi_0}.$$

Puisque W est diagonale et bornée (supposons $0 < w_{\min} \leq W_{ii} \leq w_{\max} < \infty$), des estimations standard sous-multiplicatives donnent des constantes $C, \xi > 0$ telles que

$$\big|\big(A^t\big)_{ij}\big| \;\leq\; C\, e^{-\frac{d(i,j)-vt}{\xi}}, \quad v := \frac{R}{\Delta t} \text{ pour un temps de pas effectif } \Delta t = 1,$$

c'est-à-dire une borne de type Lieb–Robinson : suppression exponentielle à l'extérieur d'un cône de lumière linéaire. Ceci retrouve la forme qualitative utilisée dans les systèmes de spins quantiques.

13 Synthèse

- **Borne de vitesse (causalité).** La vitesse de propagation maximale v est fixée par la localité architecturale (champ réceptif/rayon de message-passing R) et est *indépendante* du spectre du Hessien. La perte IS ne change pas R.
- **Conditionnement & robustesse.** La perte IS induit $W = \mathrm{diag}(1/y^2)$, normalisant les erreurs par l'échelle cible. Ceci *améliore le conditionnement spectral* de $A = KW$ et donc accélère la convergence *à l'intérieur* du cône de lumière ; cela réduit aussi la sensibilité aux cibles hétéroscédastiques.
- **Invariance conforme.** Parce que IS dépend seulement du rapport y/f, elle est invariante d'échelle : $(y, f) \mapsto (ay, af)$ laisse la perte inchangée. Dans la limite continue/régularisée, ceci se connecte à l'image conforme (Annexe D/E) : le conditionnement s'améliore via l'aplatissement spectral, mais la limite de vitesse causale v persiste.

14 Exemple concret 1D (Noyau tridiagonal)

Soient les échantillons sur une chaîne 1D avec une distance $d(i, j) = |i - j|$. Supposons

$$K = \begin{bmatrix} \ddots & \ddots & & & \\ \ddots & \alpha & \beta & & \\ & \beta & \alpha & \beta & \\ & & \beta & \alpha & \ddots \\ & & & \ddots & \ddots \end{bmatrix}, \quad \beta \neq 0,$$

c'est-à-dire $R = 1$ (voisins les plus proches). Alors par le Lemme 1, $(KW)^t$ est de bande $(2t{+}1)$, donc

$$\big(J_t\big)_{ij} = 0 \quad \text{si } |i - j| > t.$$

Ainsi, une perturbation à l'indice j peut influencer au plus les indices i avec $|i - j| \leq t$ après t étapes : $v = 1$ nœud/étape. Si $|y_i|$ varie, IS définit $W_{ii} = 1/y_i^2$, qui renormalise les magnitudes d'influence mais ne peut pas créer des entrées en dehors de la bande $(2t{+}1)$. La stabilité requiert $\eta < 2/\lambda_{\max}(KW)$; la normalisation IS réduit typiquement $\lambda_{\max}$ relativement à MSE, élargissant la plage de pas stable.

15 Conclusion de l'exemple

Sous linéarisation NTK et localité, la dynamique de prédiction neuronale entraînée avec perte IS obéit à une vitesse de propagation finie stricte :

$$\textbf{Vitesse :} \quad v = R \text{ (nœuds/étape)} \quad \text{(architectural, non spectral).}$$

L'invariance IS/conforme améliore le *conditionnement* (spectre du Hessien), renforçant la robustesse et la convergence *à l'intérieur* du cône, mais ne peut excéder la vitesse causale fixée par la localité.

16 Conclusion

La perte IS :

- Encode l'**invariance d'échelle** en pénalisant les erreurs relatives,
- Exhibe une **invariance conforme** dans la limite continuum/haute cible,
- Améliore statistiquement le **conditionnement du Hessien**,
- Fournit une **optimisation et généralisation robustes**.

Cette interprétation théorie des champs et spectrale donne une explication principielle pourquoi l'entraînement basé sur IS est plus stable et robuste que la MSE standard.