Scale Invariance and Itakura-Saito Loss: A Field-Theoretic Interpretation, a Unified Bound, and a Worked Example

Olivier Croissant

2025

Abstract

This work explores the geometric and statistical properties of the Itakura-Saito (IS) loss through a field-theoretic lens. We demonstrate that minimizing the IS loss is equivalent to minimizing an energy functional for a prediction field. A linearized analysis reveals this functional corresponds to a massive scalar field theory, which becomes conformally invariant in the limit of large target values. This invariance is shown to enhance robustness and optimization by flattening the Hessian spectrum. Furthermore, we derive a fundamental, architecture-dependent bound on the speed of information propagation in networks trained with IS loss, proving that while conformal invariance improves conditioning, it cannot violate this causal speed limit. The theory is unified with Neural Tangent Kernel (NTK) dynamics and illustrated with a concrete example.

Contents

1 Introduction

This work synthesizes and extends the geometric insights from Appendices D and E of "Risk-Averse Reinforcement Learning with Itakura-Saito Loss" by Udovichenko et al. (2025). We present a self-contained account that interprets the Itakura-Saito (IS) loss through the dual lenses of field theory and statistical learning. This perspective reveals how its inherent **scale invariance** and emergent **conformal invariance** contribute to more robust and well-conditioned optimization in machine learning models. We further bridge this to a fundamental limit on learning dynamics by deriving a unified bound on information propagation speed in networks, demonstrating that architectural causality constraints persist even under optimal conditioning.

Part I
Field-Theoretic Foundation

2 From Loss to Field Theory

The IS loss can be naturally recast as an energy functional for a prediction field. Let $\varphi(x)$ represent the model's predictions and $y(x)$ the target function. We define the action:

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

where λ is a regularization parameter. This variational formulation frames the minimization of the IS loss as the search for a field φ that balances smoothness (first term) against local fidelity to the target (second group of terms).

3 Linearized Regime and Mass Term

To analyze the behavior, we expand around a constant target $y(x) = y_0$. Setting $\varphi(x) = y_0 + \varepsilon(x)$ with $\varepsilon \ll y_0$, the IS potential expands to:

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

The linearized action then becomes:

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

This is equivalent to the action for a free massive scalar field, with mass term given by:

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

4 Propagator and Conformal Limit

The propagator $G(x)$, which encodes the correlation structure of the field, satisfies:

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

In one dimension, the solution is:

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

Notably, in the conformal limit where $y_0 \to \infty$ (or equivalently $\lambda \to 0$), the mass vanishes $m^2 \to 0$, and the propagator reduces to a power law:

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

This signifies the emergence of conformal symmetry: the system becomes scale-free, exhibiting long-range correlations and no intrinsic length scale.

5 Implications for Optimization and Robustness

The IS loss offers favorable optimization properties compared to Mean Squared Error (MSE):

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

Under a local expansion $f = y + \varepsilon$, the IS loss behaves as:

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

which is a **rescaled MSE** weighted by $1/y^2$. This embodies an implicit, adaptive normalization:

- **Large target values** $\Rightarrow$ **weaker penalty** on absolute error (prioritizing relative error).
- **Small target values** $\Rightarrow$ **stronger penalty** on absolute error.

This automatic scaling acts as a form of **natural gradient descent**, improving training stability, especially for heteroscedastic data.

6 Spectral Conditioning Theorem

The connection to conformal invariance directly impacts optimization geometry.

Theorem 6.1 (Spectral Variance Reduction, informal)**.** *Let $\varphi \sim G$ be an ensemble of models, and let $G_{conf} \subset G$ be the subset where the IS-induced action is conformally invariant. For a spectral measure $\lambda(\varphi)$ of the Hessian, we have:*

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

Conformal invariance flattens the eigenspectrum of the Hessian, reducing its condition number and leading to better-conditioned, more robust optimization landscapes.

7 Holographic Analogy

In the conformal limit ($y_0 \to \infty$), the action reduces to that of a massless scalar field:

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

a 1D Conformal Field Theory (CFT) with correlation function $G(x) \sim 1/|x|$. Drawing an analogy with the AdS/CFT correspondence, the predictions $\varphi(x)$ on the "boundary" (data space) enforce long-range consistency constraints within the latent "bulk" representation space of the model. This global consistency is a hypothesized mechanism for the improved generalization robustness observed with IS loss.

Part II
Dynamics and Speed Limits in Networks

8 A Unified Bound on Network Propagation Speed

We now consider a network (a graph, lattice, or computational circuit) with nodes V, edges E, a metric distance $d(x, y)$, and local update rules. Each node $x \in V$ has a state

$\varphi_x(t)$ evolving in time. Let $\mathcal{O}_x$ be an observable at node x.

Theorem 8.1 (Unified Network Propagation Bound)**.** *If the network satisfies:*

(i) ***Locality:*** *Updates depend only on nodes within a bounded neighborhood.*

(ii) ***Finite Interaction Strength:*** *Updates are Lipschitz-bounded by a constant g.*

(iii) ***Well-defined Metric:*** *A distance function $d(x, y)$ exists.*

Then, there exists a finite propagation speed $v > 0$ such that for any two observables $\mathcal{O}_x(t), \mathcal{O}_y(0)$, their correlation is bounded:

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

where C, ξ are system-specific constants.

8.1 Interpretations and Consequences

- This is a universal **speed limit** (v) for information propagation.
- It reduces to the Lieb-Robinson bound in quantum systems, relates to network diameter in networking, and to circuit depth in computation.
- It implies a strict **light-cone** of causal influence: no influence can propagate faster than v.
- Conformal invariance improves robustness *within* this cone but **cannot violate** this fundamental limit. Robust generalization may be viewed as a consequence of such generalized causality constraints.

9 NTK Dynamics with IS Loss

Consider training data $\{(x_i, y_i)\}_{i=1}^n$ and a model f_θ. The IS loss per datum is:

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

Its gradient and Hessian near convergence ($f_i \approx y_i$) are:

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

Under the NTK linearization regime with step size η, the prediction vector f_t evolves as:

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

where K is the fixed Neural Tangent Kernel and $W = \text{diag}(1/y_1^2, \ldots, 1/y_n^2)$. Defining the error $e_t = f_t - y$, the dynamics simplify:

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

Locality Assumption: We assume K is **range-R local** ($K_{ij} = 0$ if $d(i, j) > R$), which holds for CNNs, GNNs, and other localized architectures.

10 A Discrete Light-Cone for Prediction Influence

The Jacobian $J_t = \partial f_t / \partial f_0 = (I - \eta A)^t$ governs how perturbations propagate.

Lemma 10.1 (Bandedness under Locality)**.** *If K is range-R local and W is diagonal, then $A^t = (KW)^t$ is range-tR local.*

Theorem 10.2 (Discrete Light-Cone under IS Training)**.** *Under the locality assumption:*

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

A perturbation at node j at time 0 cannot influence node i by time t if they are separated by a distance greater than tR.

This establishes an **exact light-cone** with propagation speed $v = R$ nodes per step. This speed is an **architectural property**, independent of the loss function or optimization parameters.

11 Spectral Stability and Convergence Rate

While the *speed* v is fixed by architecture, the *rate* of error decay *inside* the light-cone depends on the spectrum of $A = KW$.

Lemma 11.1 (Linear Stability)**.** *The dynamics are stable if $0 < \eta < 2/\lambda_{\max}(A)$. The convergence rate is then governed by the smallest positive eigenvalue $\lambda^+_{\min}(A)$.*

The IS loss induces the preconditioning matrix $W = \mathrm{diag}(1/y_i^2)$. This rescales the kernel K, flattening the spectrum of A and reducing its condition number compared to the MSE case ($W = I$). This leads to faster convergence and greater robustness to varying target scales, **without altering the fundamental propagation speed** v.

12 Exponential Tails beyond Strict Locality

If the kernel K is not strictly banded but has exponential decay ($|K_{ij}| \leq C_0 e^{-d(i,j)/\xi_0}$), then the light-cone becomes "fuzzy". One can derive a Lieb-Robinson-type bound:

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

showing exponential suppression of influence outside the effective light-cone defined by speed v.

Part III
Synthesis and Example

13 Synthesis: Causality, Conditioning, and Conformal Invariance

- **Causality (Speed):** The maximum propagation speed v is set by **architectural locality** (R). IS loss does not change R.

- **Conditioning & Robustness:** IS loss induces adaptive preconditioning (W) via scale invariance. This **improves spectral conditioning**, accelerates convergence *within* the light-cone, and reduces sensitivity to heteroscedastic data.
- **Conformal Invariance:** In the continuum limit, scale invariance promotes conformal symmetry, which further flattens the Hessian spectrum. However, the **causal speed limit v remains a fundamental constraint**.

14 A Concrete 1D Example (Tri-diagonal Kernel)

Consider a 1D chain of data points. Let the NTK K be a tridiagonal matrix (nearest-neighbor interactions, $R = 1$):

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

By Theorem 2, the Jacobian J_t is *exactly zero* outside a band of width $2t+1$: $(J_t)_{ij} = 0$ if $|i - j| > t$. A perturbation propagates at most one node per step ($v = 1$). The IS weights $W_{ii} = 1/y_i^2$ rescale the *magnitude* of influence within this band but cannot create influence beyond it. Furthermore, the IS preconditioning typically allows for a larger stable step size η by reducing $\lambda_{\max}(KW)$.

15 Conclusion of the Example

This example crystallizes the core argument: Under NTK dynamics with local interactions, IS loss training obeys a strict finite propagation speed determined by architecture. Conformal invariance and the associated preconditioning enhance robustness and convergence efficiency *within* the causal horizon but cannot surpass the speed limit imposed by locality.

16 General Conclusion

The Itakura-Saito loss provides a powerful alternative to standard losses like MSE due to its foundational properties:

- **Scale Invariance**, which penalizes relative errors.
- **Emergent Conformal Invariance** in a key limit, leading to robust, scale-free learning.
- **Enhanced Hessian Conditioning**, which stabilizes and accelerates optimization.
- **Architecture-aware Dynamics**, where it enforces a strict causal speed limit on learning.

This field-theoretic interpretation provides a unified and principled framework for understanding the robustness and efficiency gains observed when using the IS loss in machine learning.

References

[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*.