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

Olivier Croissant (2025)

Contents

1 Introduction

This section summarizes and reformulates the results of Appendix D and E of "*Risk-Averse Reinforcement Learning with Itakura–Saito Loss*" by Udovichenko et al. (2025). It provides a self-

contained account of the geometric, field-theoretic, and statistical interpretations of the Itakura–Saito (IS) loss, highlighting how scale and conformal invariance enhance robustness and optimization in machine learning.

2 From Loss to Field Theory

The IS loss can be recast as an energy functional. Let $\varphi(x)$ be a field representing predictions, and $y(x)$ the target. We define the action

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

where λ is a regularization parameter. This variational formulation highlights that minimizing IS loss corresponds to minimizing an energy functional with both local and global consistency terms.

3 Linearized Regime and Mass Term

Consider expansion around a constant target $y(x) = y_0$. Let $\varphi(x) = y_0 + \varepsilon(x)$ with $\varepsilon \ll y_0$. The IS potential expands as

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

The linearized action reads

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

This is equivalent to a Gaussian field theory with mass

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

4 Propagator and Conformal Limit

The propagator $G(x)$ satisfies

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

The solution in 1D is

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

In the conformal limit $y_0 \to \infty$ (or $\lambda \to 0$), this reduces to

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

showing absence of intrinsic scale, power-law correlations, and emergent conformal symmetry.

5 Implications for Optimization

The IS loss has favorable conditioning compared to MSE:

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

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

Locally, if $f = y + \varepsilon$,

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

which is equivalent to a rescaled MSE, with weighting $\propto 1/y^2$. Thus:

- Large $y \Rightarrow$ weaker penalty,
- Small $y \Rightarrow$ stronger penalty.

This implicit normalization acts like *natural gradient descent*, improving stability.

6 Spectral Conditioning and Robustness

Theorem (informal). Let $\varphi \sim G$ be a Gaussian ensemble of models, and let $G_{\text{conf}} \subset G$ be the subset where the IS-induced action is conformally invariant. Define $\lambda(\varphi)$ as a spectral measure of the Hessian. Then

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

Hence conformal invariance reduces spectral variance, flattening the Hessian spectrum, and leading to better-conditioned optimization.

7 Holographic Analogy

In the limit $y_0 \to \infty$, the IS-induced action reduces to

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

a massless scalar field theory (a 1D CFT). This exhibits long-range correlations $G(x) \sim 1/|x|$. The analogy with AdS/CFT suggests that the IS-trained boundary predictions enforce global consistency in the latent "bulk" representation space, contributing to robustness.

8 Generalized Speed Limit in Networks

Consider a network (graph, lattice, or computational circuit) represented as a set of nodes V and edges E, with a metric distance $d(x, y)$ between nodes and local interaction rules. Each node $x \in V$ has a state $\varphi_x(t)$ evolving in time t according to local update dynamics. Let $\mathcal{O}_x$ denote an observable localized at node x.

We are interested in the maximal speed at which influence, information, or correlations can propagate across the network.

Theorem 1 (Unified Network Propagation Bound)**.** *Suppose the network satisfies:*

(i) ***Locality:*** *updates of φ_x depend only on nodes within a bounded neighborhood of x;*

(ii) ***Finite interaction strength:*** *each local update is Lipschitz-bounded with constant g (energy, bandwidth, or Lipschitz factor);*

(iii) ***Well-defined metric:*** *the graph admits a distance $d(x, y)$.*

Then there exists a finite velocity $v > 0$, called the network propagation speed*, such that for any two observables $\mathcal{O}_x(t), \mathcal{O}_y(0)$ localized at nodes $x, y \in V$, one has*

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

where C, ξ are constants depending on the system.

Interpretations

- In **quantum spin systems**, this reduces to the Lieb–Robinson bound, with v the Lieb–Robinson velocity.
- In **communication networks**, v corresponds to maximum edge capacity and $d(x, y)$ to path length (network diameter).
- In **distributed computing / consensus**, v^{-1} is proportional to the spectral gap of the Laplacian (algebraic connectivity).
- In **neural or Boolean circuits**, v corresponds to *layers per unit time*, so $d(x, y)/v$ gives the minimal circuit depth for influence.
- In **machine learning with IS loss**, conformal invariance enforces scale-free conditioning, but the Hessian spectrum still constrains v, preventing arbitrarily fast convergence.

Consequences

This theorem implies:

1. No network, physical or computational, can propagate information instantaneously; there is always a finite "light-cone" of influence.
2. Conformal/scale invariance improves robustness by smoothing propagation and flattening spectra, but does not remove the fundamental bound v.
3. Robust generalization in learning systems may thus be interpreted as emerging from a *generalized causality constraint.*

9 Setup: NTK Dynamics with IS Loss

Let $\{(x_i, y_i)\}_{i=1}^n$ be training samples and let $f_\theta(x) \in \mathbb{R}$ be a scalar model output. Denote the vector of predictions $f_\theta := (f_\theta(x_1), \ldots, f_\theta(x_n))^\top \in \mathbb{R}^n$ and targets $y \in \mathbb{R}^n$.

For a single datum (y_i, f_i), the Itakura–Saito loss is

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

One computes

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

Near $f \approx y$, the loss is locally quadratic with diagonal curvature $W := \mathrm{diag}(1/y_1^2, \ldots, 1/y_n^2)$ and gradient $\nabla_f L \approx W(f-y)$.

Assume training by (full-batch) gradient descent with step size $\eta > 0$ and *lazy training*/NTK linearization:

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

where $K \in \mathbb{R}^{n\times n}$ is the (empirical) Neural Tangent Kernel (NTK) at initialization (held fixed in the linearization). Define the *prediction error* $e_t := f_t - y$. Then the error dynamics is linear:

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

Locality assumption. Let the training points be nodes of a graph (V, E) with graph distance $d(i, j)$. We assume K is *range-R local*:

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

This covers common cases: convolutional architectures on grids (finite receptive field), graph neural networks with R-hop message passing, and localized kernels. Note that W is diagonal and hence preserves sparsity patterns when multiplied.

10 A Discrete Light-Cone for Prediction Influence

We study how a perturbation at sample j at step 0 can influence sample i after t steps. Let $J_t := \partial f_t / \partial f_0$ be the Jacobian of the prediction map under (14). With $e_t = (I - \eta A)^t e_0$ and $f_t = y + e_t$, we have $J_t = (I - \eta A)^t$.

Lemma 1 (Bandedness under locality)**.** *If K satisfies* (15) *with range R and W is diagonal, then for all integers $t \geq 0$*

$$A^t = (KW)^t \text{ is range-}tR \text{ local:} \quad (A^t)_{ij} = 0 \quad \text{whenever } d(i,j) > tR.$$

Proof. By (15), $A = KW$ has the same sparsity as K (since W is diagonal). Products of range-R local matrices convolve ranges additively: if B and C are range-R_B and range-R_C local, then $(BC)_{ij} = \sum_k B_{ik} C_{kj}$ can be nonzero only if there exists k with $d(i,k) \leq R_B$ and $d(k,j) \leq R_C$, hence $d(i,j) \leq R_B + R_C$ by triangle inequality. Iterating gives that A^t is range-tR local. □

Theorem 2 (Discrete light-cone under IS training)**.** *Under* (15)*, for any $t \geq 0$,*

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

Equivalently, a unit perturbation at node j at step 0 cannot affect $f_t(i)$ if $d(i,j) > tR$.

Proof. Expand $(I - \eta A)^t = \sum_{k=0}^{t} \binom{t}{k} (-\eta)^k A^k$. By Lemma 1, A^k is range-kR local. If $d(i,j) > tR$, then for all $k \leq t$ we have $d(i,j) > kR$, so $(A^k)_{ij} = 0$, hence the sum has zero (i,j) entry. □

Theorem 2 gives an *exact* light-cone with discrete speed $v = R$ nodes per step. Thus, the maximum propagation speed is an architectural/locality property; it is *independent* of curvature or step size (subject to stability below).

11 Spectral Stability, Convergence Rate, and Hessian Conditioning

While v is set by locality (R), the *rate* at which errors decay inside the cone depends on the spectrum of $A = KW$.

Lemma 2 (Linear stability)**.** *If $\rho(A)$ denotes the spectral radius of A, then gradient descent* (14) *is linearly stable if*

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

Under this condition, $\|e_t\| \;\le\; \kappa\,(1 - \eta\lambda^{+}_{\min})^t\,\|e_0\|$ in any A-compatible norm, where $\lambda^{+}_{\min}$ is the smallest positive eigenvalue of A and κ depends on the eigenbasis conditioning.

Role of IS curvature. Near $f \approx y$, the Hessian w.r.t. predictions is $H_f \approx W = \mathrm{diag}(1/y_i^2)$. Thus $A = KW$ rescales rows/columns of K by $1/y_i^2$, acting as an *implicit preconditioner*. If the target magnitudes $\{|y_i|\}$ vary widely, W *flattens* the effective spectrum of A relative to MSE ($W = I$), reducing the condition number and improving convergence rates—*without* changing the locality range R (and hence without increasing v).

12 Exponential Tails beyond Strict Locality

In many architectures, K is not strictly banded but has fast off-diagonal decay:

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

Since W is diagonal and bounded (assume $0 < w_{\min} \le W_{ii} \le w_{\max} < \infty$), standard submultiplicative estimates yield constants $C, \xi > 0$ such that

$$\left|\left(A^t\right)_{ij}\right| \;\le\; C\, e^{-\frac{d(i,j)-vt}{\xi}}, \quad v := \frac{R}{\Delta t} \text{ for an effective step-time } \Delta t = 1,$$

i.e. a Lieb–Robinson type bound: exponential suppression outside a linear light-cone. This recovers the qualitative form used in quantum spin systems.

13 Putting It Together

- **Speed bound (causality).** The maximum propagation speed v is set by architectural locality (receptive field/message-passing radius R) and is *independent* of the Hessian spectrum. IS loss does not change R.
- **Conditioning & robustness.** IS loss induces $W = \mathrm{diag}(1/y^2)$, normalizing errors by target scale. This *improves spectral conditioning* of $A = KW$ and hence accelerates convergence *inside* the light-cone; it also reduces sensitivity to heteroscedastic targets.
- **Conformal invariance.** Because IS depends only on the ratio y/f, it is scale-invariant: $(y, f) \mapsto (ay, af)$ leaves the loss unchanged. In the continuum/regularized limit, this connects to the conformal picture (Appendix D/E): conditioning improves via spectrum flattening, yet the causal speed limit v persists.

14 Concrete 1D Example (Tri-diagonal Kernel)

Let the samples lie on a 1D chain with distance $d(i,j) = |i-j|$. Suppose

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

i.e. $R = 1$ (nearest-neighbor). Then by Lemma 1, $(KW)^t$ is $(2t{+}1)$-banded, so

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

Thus a perturbation at index j can influence at most indices i with $|i-j| \leq t$ after t steps: $v = 1$ node/step. If $|y_i|$ varies, IS sets $W_{ii} = 1/y_i^2$, which rescales influence magnitudes but cannot create entries outside the $(2t{+}1)$-band. Stability requires $\eta < 2/\lambda_{\max}(KW)$; the IS normalization typically reduces $\lambda_{\max}$ relative to MSE, widening the stable step-size range.

15 Conclusion of the Example

Under NTK linearization and locality, neural prediction dynamics trained with IS loss obey a strict finite propagation speed:

$$\textbf{Speed:} \quad v = R \text{ (nodes/step)} \quad \text{(architectural, not spectral)}.$$

IS/conformal invariance improves *conditioning* (Hessian spectrum), enhancing robustness and convergence *within* the cone, but cannot exceed the causal speed set by locality.

16 Conclusion

The IS loss:

- Encodes **scale invariance** by penalizing relative errors,
- Exhibits **conformal invariance** in the continuum/high-target limit,
- Statistically improves **Hessian conditioning**,
- Provides **robust optimization and generalization**.

This field-theoretic and spectral interpretation gives a principled explanation of why IS-based training is more stable and robust than standard MSE.