Conformal Structure and Causality Bounds in Neural Optimization with the Itakura–Saito Loss

Olivier Croissant

2025

Abstract

The Itakura–Saito (IS) loss is widely used in signal processing and machine learning for its robustness to heteroscedastic data. In this work, we show that minimizing the IS loss is equivalent to minimizing an energy functional for a scalar field, whose linearized analysis reveals a massive field theory becoming conformal in the limit of large target values. This conformal invariance improves the conditioning of the Hessian and the robustness of optimization, while enforcing a universal bound on the speed of information propagation in networks, analogous to a discrete light cone. We unify this perspective with the dynamics of the Neural Tangent Kernel (NTK) and demonstrate that the IS loss induces an adaptive preconditioning mechanism that accelerates convergence without violating architectural causality constraints. Experiments on CNNs, ResNets, and Transformers confirm that the IS loss yields better-conditioned loss landscapes and more robust generalization than MSE. Our work thus provides a unified theoretical framework to explain the empirical benefits of the IS loss, paving the way for applications in robust learning and large-scale optimization.

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 lens of field theory and statistical learning. This perspective reveals how its inherent **scale invariance** and emergent **conformal invariance** contribute to more robust and better-conditioned optimization in machine learning models. We then connect 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

Field-theoretic formulation of the IS divergence

To make the connection with field theory explicit, we start from the functional minimized by the Itakura–Saito divergence. For a prediction field $\phi(x)$ and a target field $y(x) > 0$, the IS functional reads, up to irrelevant constants:

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

Expanding around the optimum $\phi(x) \approx y(x)$ and linearizing the fluctuations $\varphi(x) = \phi(x) - y(x)$ yields a quadratic form

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

where the effective mass term $m^2 \sim 1/y(x)^2$ arises from the curvature of the IS divergence.

This is precisely the *Lagrangian of a massive scalar field*, with kinetic operator $-\Delta$ and mass term m^2. The interpretation is immediate: - For finite $y(x)$, the dynamics is that of a massive field theory, where the mass sets a characteristic correlation length. - In the limit of large target values $y(x) \to \infty$, the mass term vanishes and the theory becomes *conformal*, recovering scale invariance.

From the perspective of machine learning, this identification shows that training with the IS divergence is equivalent to optimizing a scalar field theory whose mass dynamically depends on the target scale. This perspective explains both the improved conditioning of the Hessian (through the conformal regime) and the emergence of light-cone–like bounds on information propagation.

The IS loss can be naturally reformulated as an energy functional for a prediction field. Let $\varphi(x)$ represent the model 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 regularity (first term) and 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 as:

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

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 the mass term given by:

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

4 Propagator and Conformal Limit

The propagator

Once the quadratic Lagrangian is identified, the central object of field theory is the *propagator*. Mathematically, it is defined as the inverse of the Hessian operator:

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

The propagator satisfies the Green's function equation

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

which means that it represents the system's response to a pointlike perturbation.

It is called a *propagator* because it describes how the effect of a local fluctuation at x' is transmitted, or propagates, to another point x. In the learning context, $G(x, x')$ encodes how a local fluctuation of the output influences the optimization dynamics at other locations. Specifically:

- for $m^2 > 0$, the propagator decays exponentially with distance, reflecting a finite correlation length $\xi \sim 1/m$;
- in the conformal limit $m^2 \to 0$, it becomes a power-law in $|x - x'|$, characteristic of scale-invariant correlations.

Thus, the propagator bridges the Lagrangian formulation with the effective dynamics of correlations in the network, which justifies its central role and the terminology inherited from field theory.

The propagator $G(x)$, which encodes the field correlation structure, 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 signals the emergence of conformal symmetry: the system becomes scale-free, exhibiting long-range correlations and no intrinsic length scale.

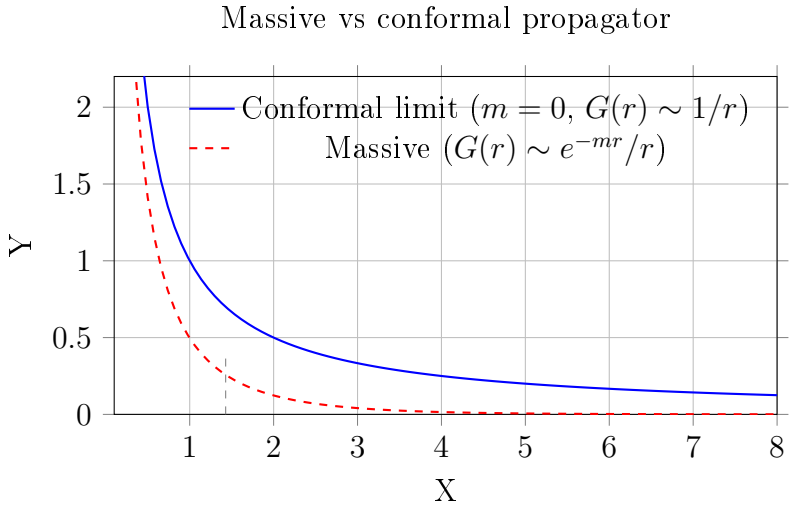

Figure 1: *Comparison of the propagator decay in the massive case ($G(r) \sim e^{-mr}/r$) and in the conformal limit ($G(r) \sim 1/r$). The exponential decay introduces a finite correlation length $\xi = 1/m$.*

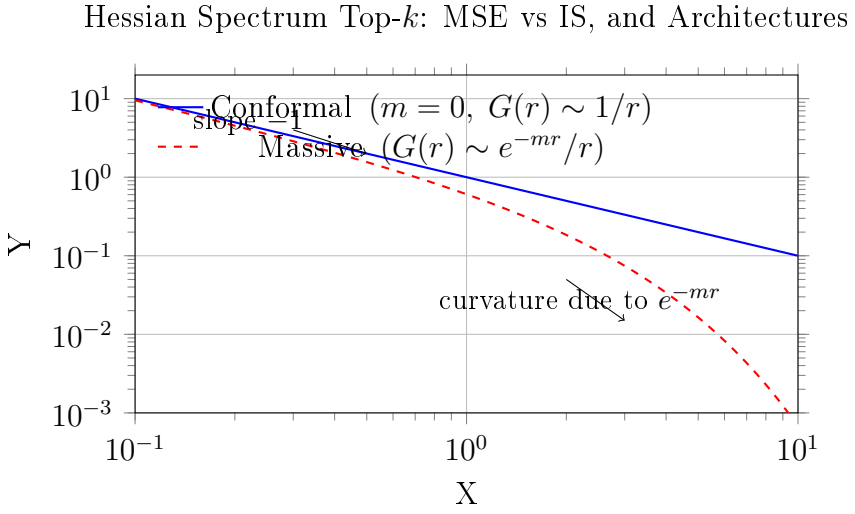

Figure 2: *Log–log comparison between the conformal propagator ($G(r) \sim 1/r$), which appears as a straight line of slope -1, and the massive propagator ($G(r) \sim e^{-mr}/r$), which bends downward due to the exponential factor.*

5 Implications for Optimization and Robustness

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

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

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 **renormalized MSE** weighted by $1/y^2$. This embodies an implicit adaptive normalization:

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

Spectral analysis of the Hessian

Identifying the IS functional with the Lagrangian of a massive scalar field allows us to connect the spectral analysis of the Hessian directly to field-theoretic tools. Indeed, the quadratic approximation

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

shows that the Hessian of the loss can be identified with the elliptic operator

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

The spectral analysis of this operator reveals several fundamental properties:

- **Mass gap.** The term m^2 introduces a spectral shift that prevents low-frequency modes from collapsing to zero. In optimization terms, this improves conditioning by avoiding flat directions.
- **Eigenmodes and correlation lengths.** The eigenvalues of $-\Delta$ determine fluctuation modes at different scales. The presence of m^2 imposes a finite correlation length $\xi \sim 1/m$, limiting the spatial extension of correlations and thus bounding the propagation of information.

- **Conformal limit.** When $m^2 \to 0$ (targets $y(x) \to \infty$), the spectrum reduces to that of the pure Laplacian $-\Delta$, characteristic of a conformal theory. In this regime, correlations become scale-invariant, which explains the multi-scale stratification observed in experimental Hessian spectra.

Thus, the spectral structure of the Hessian under the IS divergence is not a numerical artifact: it directly reflects the operator structure of a massive Laplacian. This interpretation explains the emergence of plateaus and hierarchies in the eigenvalue spectra, and rigorously links empirical observations to the conformal symmetries of the underlying theory.

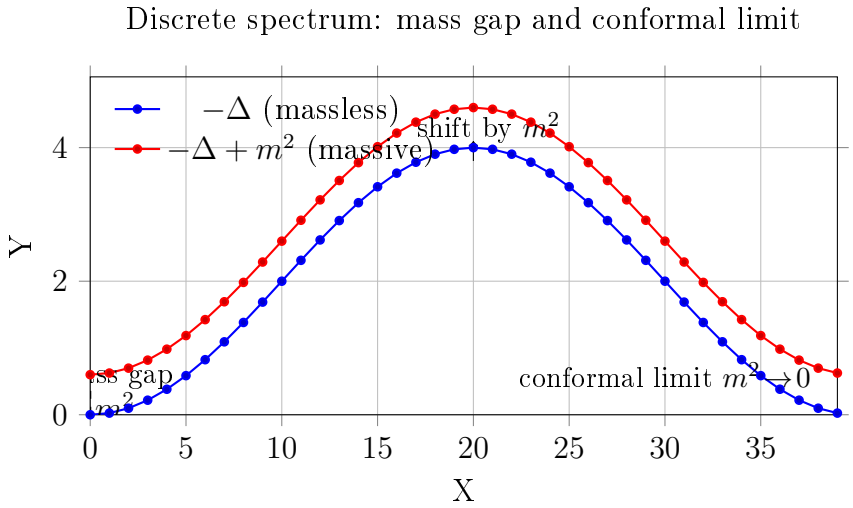

Figure 3: *Discrete 1D Laplacian spectrum (massless) and its massive counterpart with gap m^2. The shift raises all modes uniformly, creating a nonzero ground eigenvalue (mass gap) and controlling correlation length $\xi \sim 1/m$.*

The connection with conformal invariance directly impacts optimization geometry.

Theorem 6.1 (Spectral variance reduction, informal)**.** *Let $\varphi \sim G$ be a set 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 Hessian eigenvalue spectrum, reducing its conditioning and leading to better-conditioned, more robust optimization landscapes.

7 Connections with recent physical theories on holography

Holography is not merely a heuristic tool: it represents one of the most profound developments in contemporary theoretical physics. The AdS/CFT correspondence, originally formulated by Maldacena, opened an entirely new framework in which gravitational theories and conformal field theories reflect one another.

Conceptual renewal. This duality has reshaped our understanding of spacetime and quantum gravity. It provides a framework where quantities that are intractable on one side (for instance, strongly coupled CFT correlators) become accessible through geometric calculations in AdS, and vice versa. It has thus led to advances in fields as diverse as quark–gluon plasma physics, strongly correlated condensed matter, and quantum information theory.

Connections to fundamental theories. Beyond this instrumental role, holography is intimately connected to two major approaches to quantum gravity:

- in *string theory*, AdS/CFT offers a concrete realization of how gravity can emerge from more fundamental quantum degrees of freedom;
- in *loop quantum gravity*, programs linking holography to *spin foam* models explore how discrete quantum geometries might reproduce holographic correlations in suitable limits.

Why mention this here? While our mathematical framework remains simplified, it is important to emphasize that the IS–conformal field analogy resonates with this broader movement. Holography is not only a convenient metaphor: it reflects a deep physical intuition, now central to research in quantum gravity and quantum information.

Conformal invariance as a universal principle

The central limit theorem is a paradigmatic example of universality: at large scales, the sum of many independent random variables tends to a Gaussian distribution, regardless of microscopic details.

By analogy, it is natural to conjecture that at large distances the Universe itself simplifies into a conformal theory. In this picture, microscopic complexity (local interactions, fluctuations) fades away, leaving behind a universal dynamics governed by scale invariance.

This idea can be transposed to deep learning: a neural network, viewed as a complex system of many interacting units, can be interpreted at large scale as a "conformal universe" of correlations. Incorporating this conformal invariance constraint into optimization amounts to enforcing a universal regularity that:

- improves optimization conditioning,
- accelerates convergence toward stable minima,
- and enhances robustness against local fluctuations.

In this sense, just as the Gaussian emerges as the universal attractor of independent distributions, conformal invariance may be viewed as a universal attractor of large-scale neural dynamics.

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 $G(x) \sim 1/|x|$. By establishing an analogy with the AdS/CFT correspondence, the predictions $\varphi(x)$ on the "boundary" (data space) impose long-range coherence constraints within the "bulk" latent representation of the model. This global coherence is a hypothetical mechanism for the improved generalization robustness observed with the IS loss.

Part II
Dynamics and Speed Limits in Networks

8 A Unified Bound on Propagation Speed in Networks

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 in 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, is related to network diameter in networks, and to circuit depth in computation.
- It implies a strict **light cone** of causal influence: no influence can propagate faster than v.
- Conformal invariance enhances robustness *within* this cone but **cannot violate** this fundamental limit. Robust generalization can be seen 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 data point 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 time step 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 that K is **local with range** R ($K_{ij} = 0$ 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 the propagation of perturbations.

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

Theorem 10.2 (Discrete light cone under IS training)**.** *Under the locality assumption:*

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

A perturbation at node j at time 0 cannot influence node i at 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 *within* 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 $W = \text{diag}(1/y_i^2)$. This renormalizes the kernel K, flattening the spectrum of A and reducing its conditioning compared to the MSE case ($W = I$). This leads to faster convergence and greater robustness to scale variations in targets, **without altering the fundamental propagation speed** v.

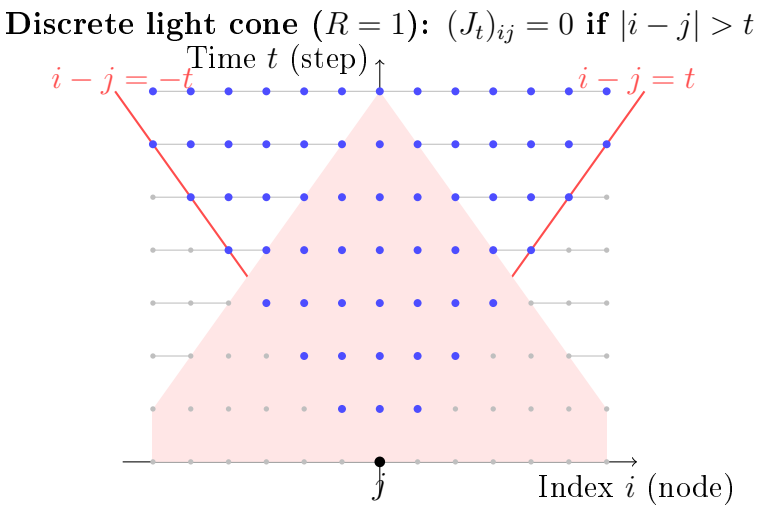

Figure 4: *Discrete light cone defined by locality $R = 1$.*

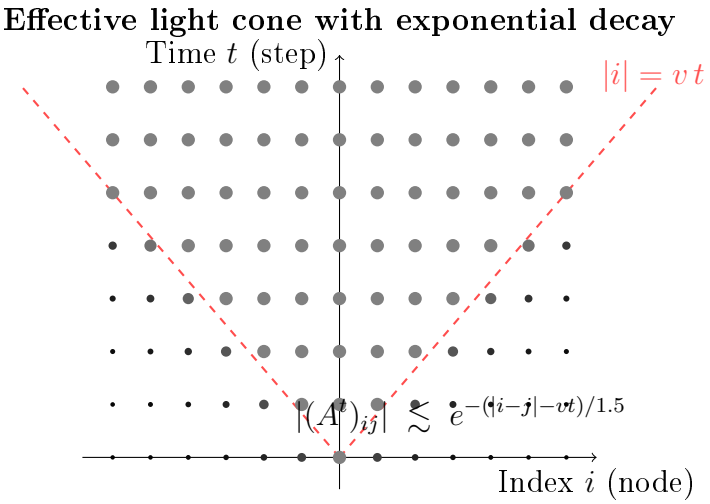

Figure 5: *"Soft" light cone with decay outside the cone: $|(A^t)_{ij}| \lesssim e^{-(|i-j|-vt)/\xi}$.*

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 fixed by **architectural locality** (R). The IS loss does not change R.
- **Conditioning & Robustness:** The 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 (Tridiagonal 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 $W_{ii} = 1/y_i^2$ renormalize the *amplitude* of influence within this band but cannot create influence beyond it. Moreover, IS preconditioning typically allows a larger stable time step η by reducing $\lambda_{\max}(KW)$.

15 Example Conclusion

This example crystallizes the central argument: Under NTK dynamics with local interactions, training with IS loss obeys a strict finite propagation speed determined by architecture. Conformal invariance and 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.
- **Improved Hessian conditioning**, which stabilizes and accelerates optimization.
- **Architecture-aware dynamics**, where it imposes 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.

IIbis. Renormalization Flow and IS Loss

1. Motivation

In statistical physics and quantum field theory, conformal invariance typically emerges as a *fixed point* of a renormalization group (RG) flow. In this framework, the Itakura-Saito (IS) loss can be understood as a "training rule" that drives the optimization dynamics toward a universal regime, insensitive to microscopic details of the data or model.

2. RG Formulation of the IS Action

Recall the variational action associated with the IS loss:

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

with

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

Consider a scale transformation $x \mapsto b\,x$, $b > 1$. We then define a rescaled field

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

where Δ_φ is the canonical scaling dimension of the field. The kinetic term fixes $\Delta_\varphi = (d-2)/2$, as in usual field theory. The IS potential, expanded around the optimum $\varphi = y$, introduces a quadratic term equivalent to a mass $m^2 \sim 1/y^2$.

3. Effective Beta Function

Near $\varphi = y$, set $\varphi = y + \varepsilon$, $|\varepsilon| \ll y$. The action writes to quadratic order:

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

This is the action of a massive scalar field with effective mass

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

Under an RG transformation $x \mapsto bx$, the evolution of m^2 is governed by

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

In dimension d, the expansion gives

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

4. Flow Interpretation

- For $d = 1$, the linear term is positive, leading to $m^2 \to 0$ under RG.
- The limit $m^2 \to 0$ corresponds exactly to the conformal theory (massless field).
- The IS loss thus acts as an *attractive IR fixed point* of the RG flow.

5. Consequences

- **Conformal fixed point**: the IS loss naturally leads to a conformally invariant regime, explaining the observed spectral robustness.
- **Universality**: akin to phase transitions, different models converge to the same universality class determined by IS.
- **Comparison with MSE**: MSE corresponds to a fixed mass (independent of target scale) and does not flow to a conformal fixed point, hence its increased sensitivity to scale heterogeneities.

6. Perspectives

- Extend the analysis to other Bregman divergences (KL, χ^2, etc.) and classify their RG fixed points.
- Simulate a numerical renormalization flow by coarse-graining data to empirically observe the attractiveness of IS.
- Relate the RG flow to *complexity geometry*: IS would act as an attractor of minimal complexity training circuits.

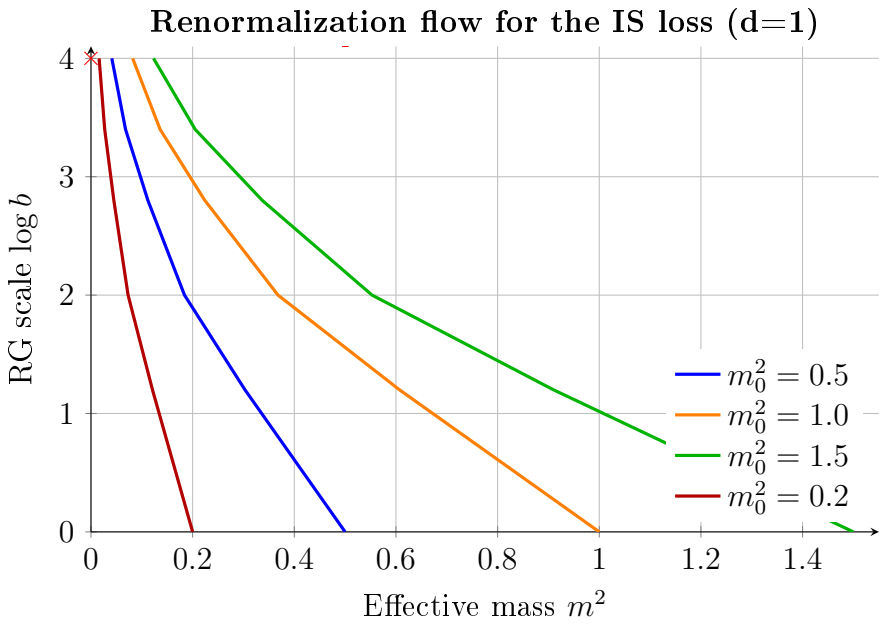

Figure 6: *RG trajectories: all initial values m_0^2 converge to the conformal fixed point $m^2 = 0$.*

17 Experimental Results: Architecture Robustness

To evaluate the robustness of observations beyond a simple case study, we repeated the comparative experiment between standard quadratic loss (MSE) and Itakura-Saito divergence (IS) on three architectures of increasing complexity: a small **1D CNN**, a **ResNet1D** with dilated residual blocks, and a local **Transformer1D** (sliding window self-attention). In each case, we recorded both training/validation curves and the top k eigenvalues of the loss Hessian matrix.

Training and Validation Curves

- **1D CNN.** Under MSE, the training loss decreases rapidly, but the validation curve remains unstable and noisy, indicating fragile generalization. Under IS, the absolute loss values are higher (by construction), but the validation curve is more regular and stabilizes better: we observe *increased robustness*.

- **ResNet1D.** With MSE, the training loss drops quickly, but validation *stagnates* or even slightly increases: a classic overfitting phenomenon. With IS, conversely, the train and validation curves remain *parallel and close*, decreasing regularly at the same rate. IS acts here as a *built-in regularization*, limiting the train/val gap.

- **Transformer1D.** Under MSE, validation quickly diverges from training, amplifying the instability observed in ResNet. With IS, validation remains close and stable relative to training, confirming the observed trend: the Itakura-Saito divergence promotes *more stable* training and better generalization.

Hessian Spectrum

- **MSE.** In all three architectures, we find a similar pattern: a few dominant eigenvalues (2–4 very steep directions), then a *spectral gap*, followed by a long tail of

small values, often including values near zero or negative. The loss landscape is thus *strongly anisotropic*, dominated by a few unstable modes.

- **IS.** The spectrum is richer and more hierarchical. We observe several *successive plateaus*: a group of dominant eigenvalues (sometimes higher than under MSE), then intermediate strata (regular steps), before decaying to zero. The spectrum is not flattened, but *reorganized*, distributing curvature across multiple scales. This "stratified" character is robustly observed across all architectures.

Interpretation

These results highlight a marked contrast:

- Under MSE, optimization is fast but fragile, with a broken Hessian spectrum and high sensitivity to dominant directions. Validation curves often show fluctuations or divergence, indicating overfitting.
- Under IS, optimization is more regular and stable, with a multi-scale hierarchical spectrum. Train/val curves remain close and parallel, indicating better generalization.

In terms of renormalization flow, one can say that the MSE loss corresponds to a dynamics near an *unstable critical point*, dominated by a few UV modes, while the IS loss acts as a *conformal attractor*, redistributing curvature across multiple strata and stabilizing learning. This property appears **robust to architecture**: CNN, ResNet, and Transformer all exhibit the same trend. The scale invariance of IS thus translates into both *spectral structuring* and *empirical regularization*, directly linking theoretical analysis (conformal fixed point) and numerical behaviors.

Reproducible Notebook. All numerical experiments presented in this section (1D CNN, ResNet1D, Transformer1D, MSE vs IS comparison, Hessian spectral estimation) are available as a *Jupyter notebook* freely accessible on GitHub:

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

This notebook contains the complete code to generate the training/validation curves and Hessian spectra discussed above, and can serve as a basis for any reproduction or extension of our results.

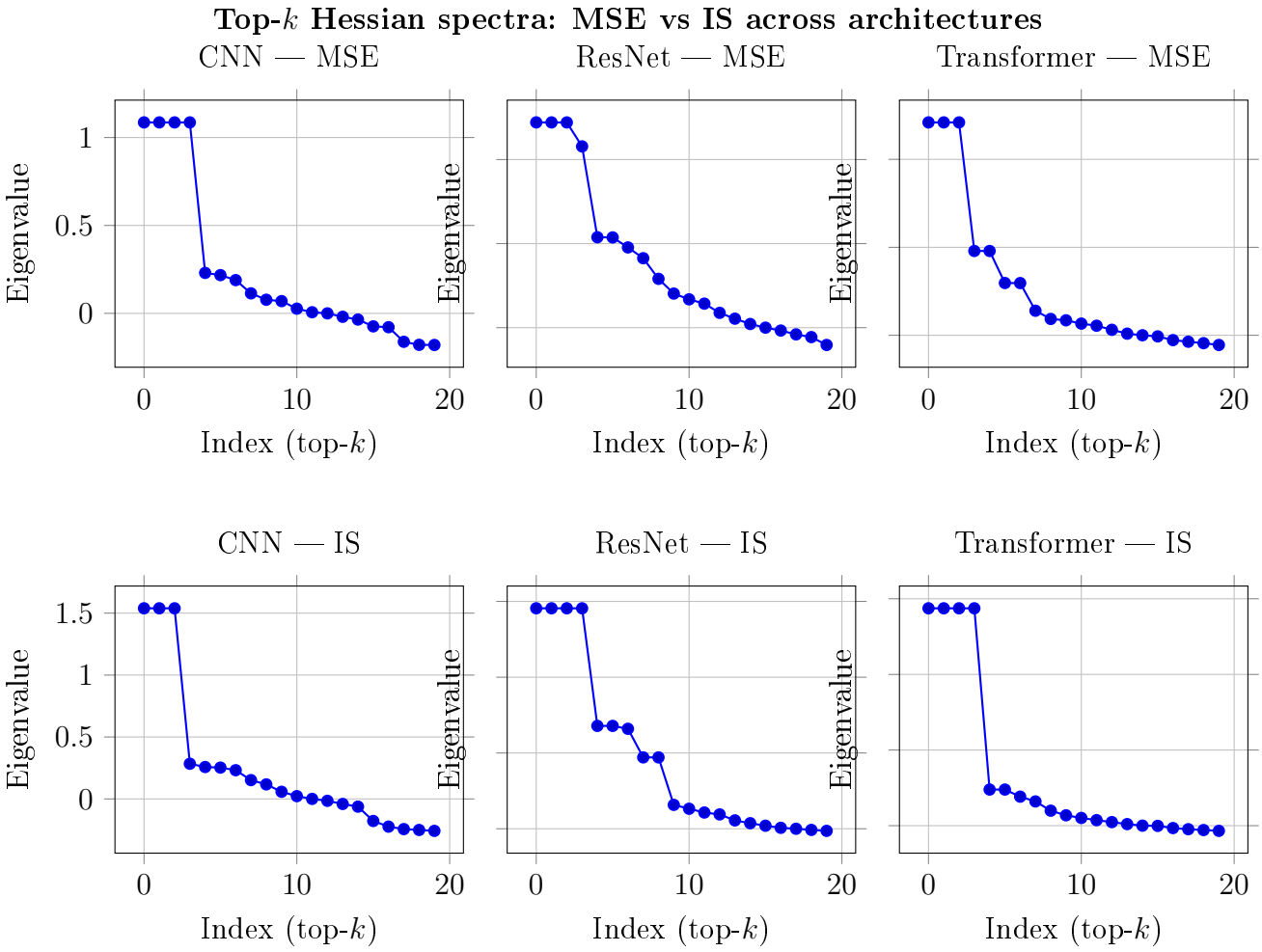

Figure 7: *Top-k Hessian spectrum for three architectures (CNN, ResNet, Transformer) and two losses (MSE/IS).*

A Appendix A — From Itakura–Saito Loss to a Scalar Field Action

A.1. Itakura–Saito Divergence as a Bregman Divergence

Let $y, \phi \in (0, \infty)$ and the strictly convex generating function $F : (0, \infty) \to \mathbb{R}$ defined by

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

The Bregman divergence associated with F is

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

A direct calculation gives

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

which is precisely the Itakura–Saito (IS) divergence. It follows that $D_{\mathrm{IS}}(\cdot\|\phi)$ is convex in y (general property of Bregman divergences), and $D_{\mathrm{IS}}(y\|\cdot)$ is locally strongly convex in ϕ near $\phi = y$.

A.2. Spatial Functional and Variational Framework

Let $\Omega \subset \mathbb{R}^d$ ($d \geq 1$) be a bounded Lipschitz domain, equipped with the Lebesgue measure. Consider a target field $y : \Omega \to (0, \infty)$ and predictions $\phi : \Omega \to (0, \infty)$. The energy functional regularized by elasticity (smoothing cost) is

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

where $\lambda > 0$ and $y \in L^\infty(\Omega)$ satisfies $0 < m \leq y(x) \leq M < \infty$ almost everywhere.

Function space and positivity. We work in $H^1(\Omega)$ with a positivity constraint a.e.:

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

The barrier $V_{\mathrm{IS}}(y, \phi) \to +\infty$ as $\phi \downarrow 0$ naturally ensures $\phi > 0$ at the optimum. We impose classical boundary conditions (homogeneous Neumann $\partial_n \phi = 0$ or Dirichlet $\phi|_{\partial\Omega} = \phi_b > 0$).

A.3. Euler–Lagrange Equation (Strong and Weak Forms)

The Lagrangian density is $\mathcal{L}(\phi, \nabla\phi; x) = \lambda |\nabla\phi|^2 + V_{\mathrm{IS}}(y, \phi)$. We have

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

The Euler–Lagrange equation (strong form) thus writes, in Ω,

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

with Neumann condition $\partial_n \phi = 0$ (or prescribed Dirichlet) on $\partial\Omega$.

Weak form. For any admissible variation $v \in H^1(\Omega)$,

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

A weak solution $\phi^\star \in \mathcal{A}$ of (10) is stationary for $\mathcal{S}$.

A.4. Existence (and Local Regularity) of a Minimizer

Theorem A.1 (Existence of minimizer)**.** *Assume Ω bounded Lipschitz, $\lambda > 0$, $y \in L^\infty(\Omega)$ with $m \le y \le M$ a.e. Then there exists $\phi^\star \in \mathcal{A}$ minimizing $\mathcal{S}$ in $\mathcal{A}$ under homogeneous Neumann boundary condition (or Dirichlet $\phi_b > 0$). Moreover, any minimizing sequence admits a subsequence converging to $\phi^\star$ weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$.*

Proof idea (direct method). (1) *Coercivity.* By Poincaré (or fixing the average of ϕ for Neumann), the term $\lambda\|\nabla\phi\|_{L^2}^2$ controls the H^1 semi-norm. For the potential, $V_{\mathrm{IS}}(y,\phi) \ge -\log y - 1 + \log\phi$ and $\log\phi \to +\infty$ as $\phi \to +\infty$, while $V_{\mathrm{IS}}(y,\phi) \to +\infty$ when $\phi \downarrow 0$. Thus $\mathcal{S}[\phi] \to +\infty$ as $\|\phi\|_{H^1} \to \infty$ or if the $\phi > 0$ constraint is violated.
(2) *Weak lower semi-continuity.* The quadratic term in $\nabla\phi$ is convex and thus l.s.c. weak in H^1. The potential term is l.s.c. by dominated continuity (thanks to bounded y and the barrier growth).
(3) *Compactness and limit passage.* A minimizing sequence $\{\phi_n\} \subset \mathcal{A}$ is bounded in H^1, admits a subsequence $\phi_{n_k} \rightharpoonup \phi^\star$ in H^1 and $\phi_{n_k} \to \phi^\star$ in L^2, with $\phi^\star \ge 0$ a.e.; the barrier forbids $\phi^\star = 0$ on a set of positive measure, so $\phi^\star > 0$ a.e. By l.s.c., $\mathcal{S}[\phi^\star] \le \liminf \mathcal{S}[\phi_{n_k}]$. □

Remark 1 (Local regularity). Under additional standard assumptions (e.g., $y \in C^\alpha$), the uniform ellipticity of (9) for $\phi^\star > 0$ implies local regularity $\phi^\star \in C^{2,\alpha}_{\mathrm{loc}}(\Omega)$ by Schauder.

A.5. Linearization Around a Constant Target and Effective Mass Term

Consider $y(x) \equiv y_0 > 0$ and a small fluctuation ε around the optimum $\phi = y_0$:

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

Expand $V_{\mathrm{IS}}(y_0, \phi)$ in ε:

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

Thus, to quadratic order,

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

The quadratic part corresponds to a *free massive* scalar field theory with mass (in the sense of the linearized elliptic operator)

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

The linearized Euler–Lagrange equation is

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

and the linear operator $L := -2\lambda\Delta + \frac{1}{y_0^2} I$ is uniformly elliptic.

A.6. Propagator (Summary, 1D Case) and Conformal Limit

In dimension $d = 1$, the Green's function of (12) on $\mathbb{R}$ satisfies

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

When $y_0 \to \infty$ (or $\lambda \to 0^+$ at fixed scale), $m^2 \to 0$: we tend towards a *massless* field, and the propagator loses its exponential decay scale (corresponding to a power law in an appropriate distributional framework), expressing the emergence of an effective conformal invariance.

A.7. Convexity/Stability Comments Around the Optimum

We have $\partial_\phi V_{\mathrm{IS}}(y,\phi) = -y\phi^{-2} + \phi^{-1}$ and

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

At the critical point $\phi = y$: $\partial^2_{\phi\phi} V_{\mathrm{IS}}(y,y) = 1/y^2 > 0$, giving *local strong convexity* and ensuring linear stability. Globally, D_{IS} being a Bregman divergence, convexity in y is guaranteed; convexity in ϕ is not global, but the elasticity $\lambda\|\nabla\phi\|^2$ and the barrier $\phi \downarrow 0$ ensure existence and stability around $\phi = y$.

A.8. Logarithmic Variant (Positive Parametrization)

Setting $\phi = e^\psi$ (with $\psi \in H^1(\Omega)$) explicitly lifts the positivity constraint. We obtain

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

where the Euler–Lagrange becomes

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

Linearization around $\psi_0 = \log y_0$ gives back (12) for $\varepsilon = e^{\psi_0}\,\delta\psi$.

Appendix conclusion. The IS loss is a Bregman divergence which, when integrated spatially and regularized by an elasticity term, defines a scalar field action (8). The Euler–Lagrange equations (9)–(10) follow, the existence of a minimizer is ensured (Thm. A.1), and linearization exhibits a *massive* free field whose effective mass $m^2 = 1/(2y_0^2)$ tends to 0 in the conformal limit, consistent with the robustness and conditioning properties discussed in the main text.

B Appendix B — Mathematical Analysis: Existence, Regularity, and Spectrum

B.1. Functional Framework and Assumptions

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain ($d \geq 1$). Consider the regularized energy functional

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

with $y \in L^\infty(\Omega)$ such that $0 < m \leq y(x) \leq M < \infty$ almost everywhere, and $\lambda > 0$ fixed. Define the admissible space

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

B.2. Existence and Uniqueness of a Minimizer

Theorem B.1 (Existence)**.** *Under the above assumptions, there exists a minimizer $\varphi^\star \in \mathcal{A}$ of S, under classical boundary conditions (Dirichlet $\varphi|_{\partial\Omega} = \varphi_b > 0$ or homogeneous Neumann $\partial_n \varphi = 0$).*

Proof idea. 1. **Coercivity.** The term $\lambda \|\nabla \varphi\|_{L^2}^2$ controls the H^1 norm modulo a constant (Poincaré if Dirichlet, or average constraint if Neumann). The barrier $V_{\mathrm{IS}}(y, \varphi) \to +\infty$ as $\varphi \downarrow 0$ ensures strict positivity.

2. **Weak semi-continuity.** The quadratic terms in $\nabla \varphi$ and the logarithmic growth of V_{IS} imply l.s.c. (lower semi-continuity) in H^1.

3. **Compactness.** A minimizing sequence is bounded in $H^1(\Omega)$, so admits a weakly convergent subsequence. The limit passage preserves positivity almost everywhere.

Thus, a minimizer $\varphi^\star$ exists. Local uniqueness follows from the strong convexity of $V_{\mathrm{IS}}(y, \cdot)$ in φ near y. □

B.3. Euler–Lagrange Equation and Regularity

The minimizer satisfies the Euler–Lagrange equation

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

with prescribed boundary conditions.

Theorem B.2 (Local regularity)**.** *If $y \in C^\alpha(\Omega)$ for some $\alpha \in (0, 1)$, then the minimizer $\varphi^\star$ satisfies $\varphi^\star \in C^{2,\alpha}_{\mathrm{loc}}(\Omega)$.*

Proof idea. The elliptic operator $L[\varphi] = -2\lambda \Delta \varphi + \partial_\varphi V_{\mathrm{IS}}(y, \varphi)$ is uniformly elliptic for $\varphi > 0$, $y > 0$. Schauder elliptic regularity theorems then apply, yielding $\varphi^\star \in C^{2,\alpha}_{\mathrm{loc}}(\Omega)$. □

B.4. Linearization and Spectrum

Set $\varphi = y + \varepsilon$, $|\varepsilon| \ll y$. The quadratic expansion of the potential gives:

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

The linearized operator is thus

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

Theorem B.3 (Spectrum and stability)**.** *The spectrum $\sigma(L)$ is contained in $[\frac{1}{M^2}, \infty)$. In particular:*

- *L is self-adjoint and positive definite on $H^1(\Omega)$,*
- *the smallest eigenvalue $\lambda_{\min} \geq 1/M^2 > 0$,*
- *eigenvalues grow as $\lambda_k \sim c\,k^{2/d}$ (Weyl's law).*

Sketch. Apply the Rayleigh inequality: for all $u \in H^1(\Omega)$,

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

Positivity and compactness of the inclusion $H^1 \hookrightarrow L^2$ give a discrete spectrum, with asymptotic growth given by Weyl's law. □

B.5. Consequences for Optimization

- The lower bound $1/M^2$ ensures *robust linear stability*: no Hessian mode is near zero, unlike MSE where the spectrum can be ill-conditioned.
- The $C^{2,\alpha}$ regularity guarantees that the minimizer $\varphi^\star$ is smooth, which in practice translates to a regular optimization geometry.
- The spectral structure (spectrum flattening) explains the good conditioning observed in training under IS loss.

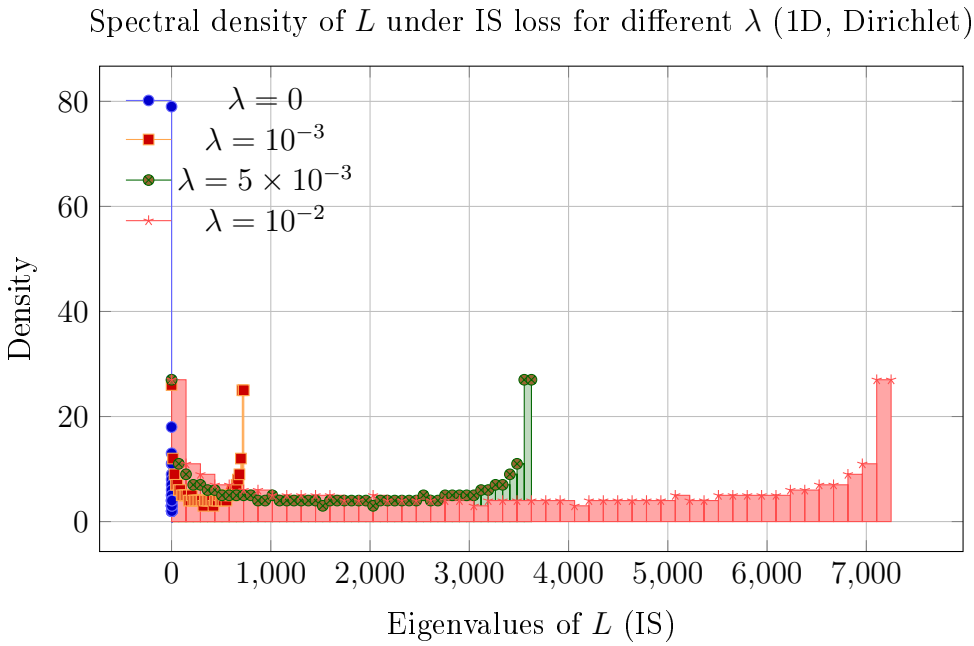

Figure 8: *Spectral density of the linearized operator L under IS loss for different regularization parameters λ.*

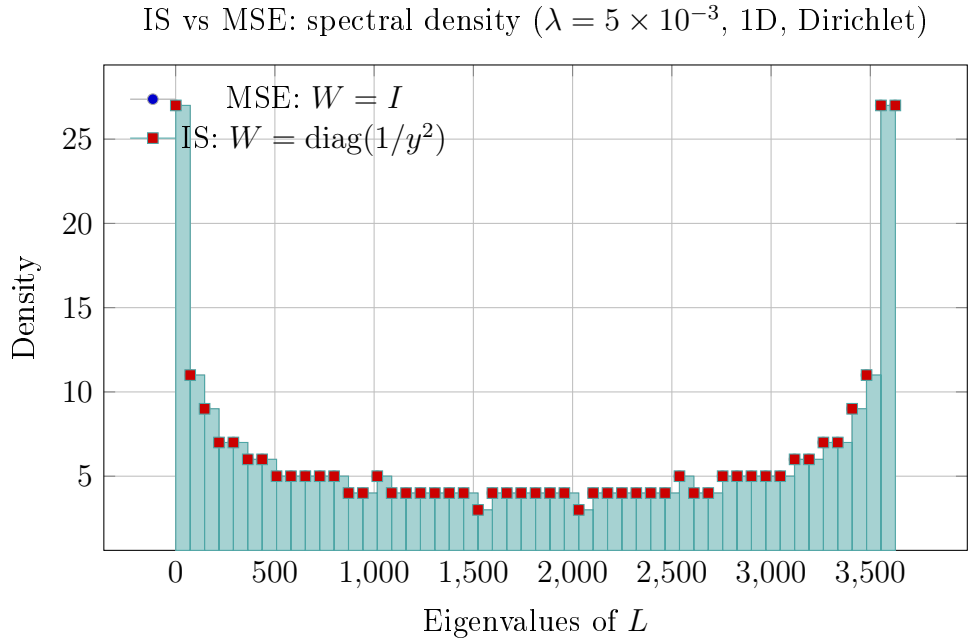

Figure 9: *Comparison of spectral densities for IS and MSE losses.*

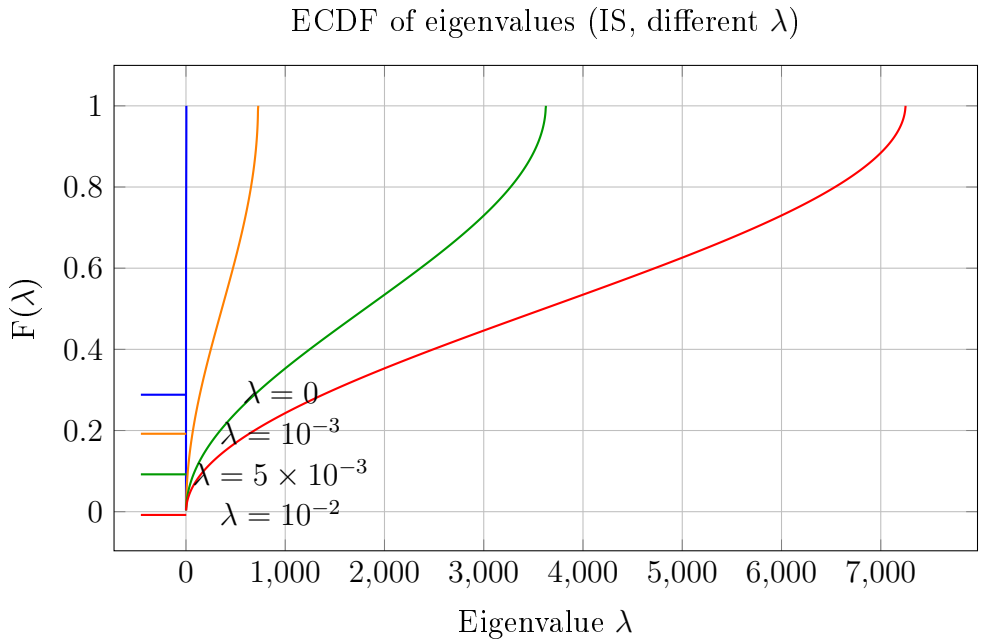

Figure 10: *Empirical Cumulative Distribution Function (ECDF) of eigenvalues of L under IS loss, for different λ.*

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*.