How Patterns Dictate Learnability in Sequential Data

Mario Morawski[1*] Anaïs Despres[2*] Rémi Rehm[2]

[1]Université Paris Dauphine–PSL
[2]Unaffiliated

November 5, 2025

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

A substantial body of work studies generalization error bounds, often via Bayesian theory and Rademacher complexity. Yet, two critical questions remain:

1. What is the **minimal achievable risk** for a predictor modeling sequential data?
2. Can we distinguish whether poor performance stems from **model limitations** or from **data unpredictability**?

Our work aims to address these two questions by providing an information-theoretic framework to quantify the minimal achievable risk in sequential prediction.

Prior Work

- **ForeCA**: measures the uncertainty of the entropy of the spectral density.
- **EvoRate**: mutual information-based metric quantifying evolving patterns in sequential data.
- **Prospective Learning**: determines under what conditions learning under non-i.i.d. stochastic processes remains feasible.
- **Mutual Information Estimators**: k-NN, MINE, InfoNCE, CLUB, SMILE.
- **Predictive Information & Universal Learning Curve**: generalization of EvoRate & discrete derivative of EvoRate.
- **Lowest Possible Error Rate**: how to bound the gap between empirical and true risk? Previous approaches use Rademacher complexity.

Motivations

Limitations of existing metrics:

1. **EvoRate:** focuses on how the metric evolves with window size, but its absolute values are hard to interpret and it cannot be linked to model performance or risk.
2. **ForeCA:** suffers from high computational cost, making it impractical for deep learning or high-dimensional data. Can only detect cyclic patterns, failing to capture trends or more complex temporal behaviors.

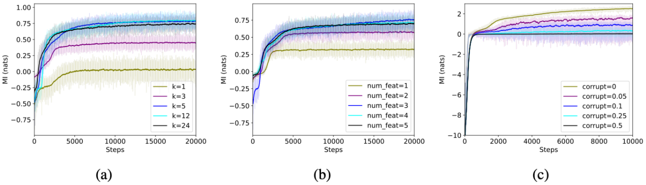


Figure 2: (a) k-order EvoRate estimation. (b) EvoRate estimation on a different number of features. (c) EvoRate estimation of the video prediction tasks with a different corruption rate.

Our Approach

We generalize `EvoRate` by introducing the **predictive information $\mathbf{I}_{\text{pred}}$**.

$$\mathbf{I}_{\text{pred}}(k, k') = \int p(\mathbf{X}_{t-k+1}^{t+k'}) \ln \frac{p(\mathbf{X}_{t-k+1}^{t+k'})}{p(\mathbf{X}_{t-k+1}^{t})p(\mathbf{X}_{t+1}^{t+k'})} \, d\mathbf{X}. \tag{1}$$

We then relate $\mathbf{I}_{\text{pred}}$ to the **universal learning curve** $\Lambda(k) = \ell(k) - \ell_0$ (with $\ell(k)$ the entropy rate of order k) , which measures the reduction in uncertainty about the future when conditioning on k past observations.

$$\mathbf{I}_{\text{pred}}(k+1, k') - \mathbf{I}_{\text{pred}}(k, k') \longrightarrow \Lambda(k) \quad \text{as} \quad k' \to \infty. \tag{2}$$

$\mathbf{I}_{pred}$ quantifies how much information from the past can be used to predict the future.

Asymptotic Behavior of $\Lambda(k)$ for Markov Processes

For a Markov process of order m, dependencies are limited to the past m observations. This is naturally captured by the predictive information $\mathbf{I}_{\text{pred}}$.

Let $\mathbf{X}_t^T$ be a Markov process of order m. For $k' \geq k \geq m$:

$$\text{(i)} \quad \mathbf{I}_{\text{pred}}(k, k') = \mathbb{E}_{\mathbf{X}_{t-m+1}^{t+m}} \left[\ln \frac{P\big(X_{t+1}^{t+m} \mid \mathbf{X}_{t-m+1}^{t}\big)}{P\big(X_{t+1}^{t+m}\big)} \right], \tag{3}$$

$$\text{(ii)} \quad \forall\, k \geq m, \quad \Lambda(k) = 0. \tag{4}$$

- For first-order Markov processes ($m = 1$), $\mathbf{I}_{\text{pred}}(k, k') = \texttt{EvoRate}(1)$ for all $k \geq 1$.
- $\Lambda(k)$ identifies the **true Markov order** by vanishing once $k \geq m$.

Link Between Learning Curve and Minimal Achievable Risk

We connect the predictive information $\mathbf{I}_{\text{pred}}$ to model performance through the k^{th}-order forecasting risk:

$$\mathcal{R}^k(Q) = \mathcal{L}^k_{\text{mle}} = -\mathbb{E}_{P(X_{t+1}, \mathbf{X}^t_{t-k+1})} \ln Q(X_{t+1} \mid \mathbf{X}^t_{t-k+1}). \quad (5)$$

We then show that for any $k \in \mathbb{N}$ and any $Q \in \mathcal{H}_k$,

$$\mathcal{R}^\infty(Q^*) \;\leq\; \mathcal{R}^k(Q) - \Lambda(k).$$

This leads us to present an estimator of this minimal risk:

$$\text{(i)} \quad \hat{\mathcal{R}}^\infty(Q^*) = \min_{1 \leq k \leq M} \{\hat{\mathcal{R}}^k(Q_k) - \Lambda(k)\}, \quad (6)$$

$$\text{(ii)} \quad k^* = \arg \min_{1 \leq k \leq M} \{\hat{\mathcal{R}}^k(Q_k) - \Lambda(k)\}. \quad (7)$$

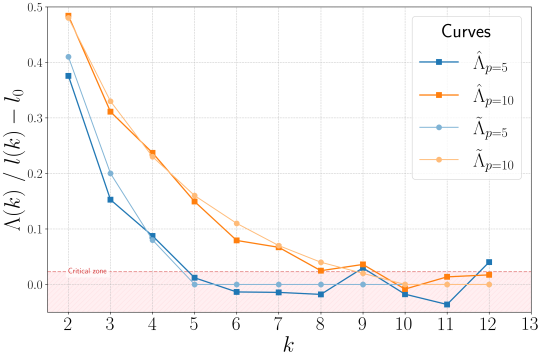
with $\mathcal{R}^\infty(Q^*)$ is the minimal risk achievable by the optimal predictor $Q^* = P(X_{t+1} \mid \mathbf{X}_{\text{past}})$.

Experimental Learning Curve $\Lambda(k)$

We simulate a stationary vector autoregressive process $\{X_t\}_{t=0}^{N-1} \subset \mathbb{R}^3$ of order $p \in \{5, 10\}$:

$$X_t = \frac{\rho}{p} \sum_{j=t-p}^{t-1} X_j + \sqrt{1-\rho^2}\,\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I_3), \tag{8}$$

with initial states $X_0, \ldots, X_{p-1} \sim \mathcal{N}(0, I_3)$ and $\rho \in (0, 1)$ controlling temporal dependence.



Learning curves $\Lambda(k)$ for AR processes with $p = 5$ and $p = 10$.

Estimating $\mathcal{R}^{\infty}(Q^*)$ for Ising Spin Sequences

Setup. We study binary spin sequences $\mathbf{X}_t^T = \{X_u\}_{u=t}^T$ with $X_i \in \{-1, +1\}$, generated as:

$$P(X_i = +1 \mid X_{i-1}, J) = \frac{\exp(JX_{i-1})}{\exp(JX_{i-1}) + \exp(-JX_{i-1})}, \tag{9}$$

where $J \sim \mathcal{N}(0,1)$ is resampled every M steps.

This creates a *blockwise-random Ising process* — piecewise-stationary with block length $M \in \{10^4, 10^5, 10^6, 10^7\}$.

We train **MLP** and **LSTM** models to predict X_{t+1} from k past values ($1 \leq k \leq 19$), using cross-entropy loss and $\dim \Theta = 1$.

M	`EvoRate`(10)	$\hat{\mathcal{R}}^{\infty}_{\text{LSTM}}(Q^*)$	$\hat{\mathcal{R}}^{\infty}_{\text{MLP}}(Q^*)$
10^4	0.28	0.37	0.37
10^5	0.29	0.37	0.37
10^6	0.33	0.36	0.34
10^7	0.48	0.07	0.09

Insights from Ising Spin Sequence Experiments

- **Data complexity decreases with block size M:** For $M = 10^7$, the coupling J is fixed and the process reduces to a first-order Markov chain.
- **`EvoRate` reflects structure:** Higher `EvoRate` indicates stronger underlying predictability, leading to lower prediction loss.
- **Estimator consistency:** $\hat{\mathcal{R}}^{\infty}(Q^*)$ remains consistent across LSTM and MLP, closely aligning with `EvoRate`.
- **Model adequacy:** The ratio $\hat{\mathcal{R}}^{k}(Q)/\hat{\mathcal{R}}^{\infty}(Q^*)$ approaches 1 as M grows, indicating improved prediction performance.
- **Estimator instability at low complexity:** Negative $\Lambda(k)$ for $M = 10^7$ arises from instability in $\hat{\Lambda}(k)$ when $k \gg p$ (true Markov order p). Refining this estimator is needed to avoid misinterpretation.

Conclusion

- Addressed minimal achievable risk in sequential modeling and sources of poor predictive performance.
- Introduced an information-theoretic framework using the learning curve $\Lambda(k)$ to link statistical dependencies and predictive performance.
- Proposed estimator $\hat{\mathcal{R}}^{\infty}(Q^*)$ to diagnose whether performance is limited by model capacity or intrinsic unpredictability.
- Theoretical and empirical validation confirms the framework across parametric and Markovian regimes.