Probabilistic Predictability of Stochastic Dynamical Systems

Tao Xu [a], Yushan Li [a], Jianping He [a]

[a] *Department of Automation, Shanghai Jiao Tong University, Key Laboratory of System Control and Information Processing, Ministry of Education of China, China*

Abstract

To assess the quality of a probabilistic prediction for stochastic dynamical systems (SDSs), scoring rules assign a numerical score based on the predictive distribution and the measured state. In this paper, we propose an ϵ-logarithm score that generalizes the celebrated logarithm score by considering a neighborhood with radius ϵ. We characterize the probabilistic predictability of an SDS by optimizing the expected score over the space of probability measures. We show how the probabilistic predictability is quantitatively determined by the neighborhood radius, the differential entropies of process noises, and the system dimension. Given any predictor, we provide approximations for the expected score with an error of scale $\mathcal{O}(\epsilon)$. In addition to the expected score, we also analyze the asymptotic behaviors of the score on individual trajectories. Specifically, we prove that the score on a trajectory can converge to the expected score when the process noises are independent and identically distributed. Moreover, the convergence speed against the trajectory length T is of scale $\mathcal{O}(T^{-\frac{1}{2}})$ in the sense of probability. Finally, numerical examples are given to elaborate the results.

Key words: Predictability, Probabilistic Prediction, Stochastic Dynamical System.

1 Introduction

1.1 Background

Noises are inevitable in dynamical systems, resulting in prediction uncertainties for future state trajectories. A probabilistic predictor predicts the target by a distribution rather than a single point, which can inherently quantify the prediction uncertainties. Therefore, probabilistic prediction for stochastic dynamical systems (SDSs) has attracted a surge of recent attention [2].

To measure the quality of a probabilistic prediction, scoring rules assign a numerical score based on the predictive distribution and the realized outcome [3, 4]. A scoring rule is called proper if the expected score is maximized when the predictive distribution equals the ground truth, which motivates the predictor to be unbiased in predicting the true distribution. One of the most celebrated proper scoring rules is the logarithm score, which assigns the score as the logarithm value of the predictive probability density function (pdf) or probability mass function (pmf) at the realized outcome.

1.2 Motivations

Due to the potential nonlinear SDS dynamics and non-Gaussian system noises, the target pdfs are typically multimodal in a variety of scenarios. For state estimation tasks with multimodal observation likelihoods, particle filters can provide multimodal estimations [5]; for edge detection tasks in computer vision, the regression-by-classification approach can model multimodal distributions under aleatoric uncertainties [6]. By taking into account the neighborhood with tunable radius ϵ, we propose an ϵ-logarithm score in this paper. It degenerates into the traditional logarithm score when ϵ equals 0.

A popular line of research is to design algorithms to probabilistically predict the state trajectories of SDSs, aiming for feasibility guarantee [7], better robustness [8], higher accuracy [9], etc. Given a scoring rule, the probabilistic predictability of an SDS can be naturally characterized by the optimal expected score. It may greatly boost the efficiency of designing predictors if we have a deeper understanding of the predictability of an SDS, e.g., what system features directly affect predictability and which one possesses the largest weight.

Email addresses: Zerken@sjtu.edu.cn (Tao Xu), yushan_li@sjtu.edu.cn (Yushan Li), jphe@sjtu.edu.cn (Jianping He).

[1] This work was supported by the National Natural Science Foundation of China under Grant 62373247. Preliminary results are published in the 61-th IEEE Conference on Decision and Control [1]. *(Corresponding author: Jianping He.)*

Although the expected score is theoretically appealing in characterizing the system's predictability, practically evaluating its value requires a sufficient amount of repeated samples for averaging. However, the samples generated from a typical SDS prediction scenario are usually temporal (a trajectory of states) rather than spatial (repeated samplings for the state at a fixed time step). While the average of spatial score samplings converges to the expectation as ensured by the law of large numbers, there is no simple guarantee for the average of temporal score samplings. Given any single trajectory generated from an SDS, under what condition can the temporal averaged score converge? Will it converge to the expected score? How fast the convergence can be?

1.3 Contributions

The main contributions are summarized as follows.

- (Metric) We propose an ϵ-logarithm score that generalizes the celebrated logarithm score by considering a neighborhood with radius ϵ. When ϵ equals 0, the proposed score will degenerate to the logarithm score. Given any predictor, we approximate the expected ϵ-logarithm score with the error of scale $\mathcal{O}(\epsilon)$.
- (Optimality) We characterize the probabilistic predictability of an SDS by the optimal expected ϵ-logarithm score, regardless of specific prediction algorithms. It quantitatively strengthens our understanding of how a system's predictability is jointly determined by the neighborhood radius, the differential entropies of process noises and the state dimension.
- (Convergence) We analyze the asymptotic convergence behaviors of the proposed score on any single trajectory generated from an SDS. It is proved that the score can converge to the expected score when the process noises are independent and identically distributed. Furthermore, the convergence speed against the trajectory length T is guaranteed to be of scale $\mathcal{O}(T^{-\frac{1}{2}})$ in the sense of probability.

The remainder of this paper is organized as follows. Section II reviews the related works. Sec. III formulates the problems of interest. Sec. IV characterize the system's predictability by evaluating the optimal expected score. Sec. V introduces a partition-based method to approximate expected scores and analyzes the asymptotic convergence behavior. Simulations are shown in Sec. VI.

2 Related Works

A lot of insightful works contribute to the predictability analysis of deterministic dynamical systems. Lorenz considered prediction performance as the growing rate of initial state uncertainty, then defined predictability as the asymptotic exponential growing rate of initial prediction error [10]. Motivated by this idea, some famous indexes such as Lyapunov exponent and Kolmogorov-Sinai entropy were proposed to characterize the predictability of dynamical systems, see a review of these indexes in [11]. These early predictability analyses have found wide applications in the climatology fields [12]. However, these works do not take noises or state measurements into consideration, and thus can not be directly applied to characterize the predictability of SDSs.

Research on the predictability analysis of discrete-state SDSs mainly bifurcates into two directions. A body of research treats the predictability of SDS from an information-theoretic perspective without first evaluating the prediction performance, thus a lot of entropy-based predictability metrics were proposed. The entropy of stochastic process is defined as the joint entropy in [13], based on which optimal prediction performance analysis and unpredictable system designs were presented in [14]. Another line of research steers the complicated evaluation of prediction performance by approximation techniques. In [15], an upper bound of the accurate prediction probability is derived based on standard Fano's inequality. This bound is applied to the study of large-scale urban vehicular mobility [16].

Research on the predictability analysis of continuous-state SDS is relatively less than the discrete ones. In the field of climate forecasting, the predictability of an SDS is defined as the distance between a predicted distribution and climatological distribution based on entropy, relative entropy and mutual information [17–19]. In the field of state estimation, some concern the predictability as the effect of model mismatch on the steady solution of the Kalman filter [20], some study the predictability by evaluating the worst-case mean square error prediction performance of the Kalman filter [21].

However, existing works on predictability analysis of SDSs mainly serve for point predictions, and it remains open and challenging to analyze the predictability under a probabilistic prediction framework.

3 Preliminaries and Problem Formulation

3.1 Preliminaries

Notations We denote random variables in bold fonts to distinguish them from constant variables. Given a random variable $\mathbf{x}$, we denote its pdf (or pmf) as $p_{\mathbf{x}}(\cdot)$. We also denote a sequence $\{(\cdot)_k\}_{k=1}^T$ by $(\cdot)_{1:T}$, and $(\cdot)_{1:0}$ is defined as an empty set. For a vector v, we use $v_{(i)}$ to deonte the ith element of v. Given a set A, the indicator function $\mathbb{I}_A(x)$ equals 1 if $x \in A$, otherwise equals 0. For two complex-valued functions on $\mathbb{R}^d$, f and g, their convolution is a complex-valued function defined by $(f * g)(x) = \int_{\mathbb{R}^d} f(y)g(x-y)\mathrm{d}y$. The deconvolution of f by g, denoted by $f *\!/\!* g$, is a complex-valued function such that $(f *\!/\!* g) * g = f$. The value of $0\log(0)$ is set to 0.

Entropy The Shannon entropy of a discrete random variable $\mathbf{x}$ with alphabet $\mathcal{X}$ and pmf $p_{\mathbf{x}}$ is $\mathrm{H_s}(p_{\mathbf{x}}) := -\sum_{x\in\mathcal{X}} p_{\mathbf{x}}(x)\log p_{\mathbf{x}}(x)$. The differential entropy of a continuous random variable $\mathbf{x}$ with support $\mathcal{X}$ and pdf $p_{\mathbf{x}}$ is $\mathrm{H_d}(p_{\mathbf{x}}) := -\int_{x\in\mathcal{X}} p_{\mathbf{x}}(x)\log p_{\mathbf{x}}(x)\mathrm{d}x$. The KL-divergence measures how a distribution $p_{\mathbf{x}}$ is different from another distribution $\hat{p}_{\mathbf{x}}$, $D_{\mathcal{KL}}(p_{\mathbf{x}}||\hat{p}_{\mathbf{x}}) := \mathbb{E}_{x\sim p_{\mathbf{x}}}\log(\frac{p_{\mathbf{x}}(x)}{\hat{p}_{\mathbf{x}}(x)})$. The cross entropy of a distribution $p_{\mathbf{x}}$ relateive to another distribution $\hat{p}_{\mathbf{x}}$ is $\mathrm{H}(p_{\mathbf{x}}||\hat{p}_{\mathbf{x}}) := -\mathbb{E}_{x\sim p_{\mathbf{x}}}\log(\hat{p}_{\mathbf{x}}(x))$.

Probabilistic prediction The problem of probabilistic prediction can be generally formulated as follows. Suppose a random variable $\mathbf{x}$ takes value on $\mathcal{X}$ with distribution $p_{\mathbf{x}}$, a probabilistic predictor predicts it by a distribution $\hat{p}_{\mathbf{x}} \in \mathcal{P}_{\mathcal{X}}$, where $\mathcal{P}_{\mathcal{X}}$ is a family of distributions over $\mathcal{X}$. When the value of $\mathbf{x}$ is materialized as x, a scoring rule,

$$S(\hat{p}_{\mathbf{x}}, x) : \mathcal{P}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}, \tag{1}$$

assigns a numerical score $S(\hat{p}_{\mathbf{x}}, x)$ to measure the quality of the predictive distribution $\hat{p}_{\mathbf{x}}$ on the realized value x. The expected scoring rule of S, usually sharing the same operator $S : \mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$ but different operands, is

$$S(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) := \mathbb{E}_{x\sim p_{\mathbf{x}}} S(\hat{p}_{\mathbf{x}}, x). \tag{2}$$

A scoring rule S is proper with respect to the prediction space $\mathcal{P}_{\mathcal{X}}$ if $S(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) \le S(p_{\mathbf{x}}, p_{\mathbf{x}})$ holds for all $\hat{p}_{\mathbf{x}}, p_{\mathbf{x}} \in \mathcal{P}_{\mathcal{X}}$. It is strictly proper if the equality holds only when $\hat{p}_{\mathbf{x}} = p_{\mathbf{x}}$. For example, the logarithm score, $\mathcal{L}(\hat{p}_{\mathbf{x}}, x) := \log \hat{p}_{\mathbf{x}}(x)$, is most celebrated for being essentially the only local proper scoring rule up to equivalence [22]; the linear score, $\mathrm{LinS}(\hat{p}_{\mathbf{x}}, x) := \hat{p}_{\mathbf{x}}(x)$, is not a proper scoring rule, despite its intuitive appeal in both theory and practice [3].

3.2 System and Predictor Model

Consider a discrete-time stochastic dynamical system, denoted by Φ,

$$\Phi : \mathbf{x}_{k+1} = f(\mathbf{x}_k) + \mathbf{w}_k, \tag{3}$$

where $\mathbf{x}_k \in \mathbb{R}^{d_x}$ is the system state and the dynamics $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ is continuous. Let $\mathcal{P}$ be the family of Lebesgue integrable pdfs over $\mathbb{R}^{d_x}$, the process noises $\{\mathbf{w}_k\}_{k=1}^{\infty}$ with pdf $p_{\mathbf{w}_k} \in \mathcal{P}$ are not necessarily required to be independent and identically distributed (i.i.d).

A probabilistic predictor keeps observing the states of Φ and predicting the conditional pdfs of future states. Specifically, given previous observations $x_{1:k}$ at time step k, a probabilistic prediction is denoted as

$$x_{1:k} \to \hat{p}_{\mathbf{x}_{k+1}|\mathbf{x}_{1:k}}(\cdot \mid x_{1:k}) \in \mathcal{P}. \tag{4}$$

After the value of $\mathbf{x}_{k+1}$ is realized as x_{k+1} from the true pdf $p_{\mathbf{x}_{k+1}|\mathbf{x}_{1:k}}(\cdot \mid x_{1:k}) \in \mathcal{P}$, the one-step prediction performance is measured by a scoring rule S as $S(\hat{p}_{\mathbf{x}_{k+1}|\mathbf{x}_{1:k}}(\cdot \mid x_{1:k}), x_{k+1})$.

3.3 Problems of Interest

The logarithm score assesses the prediction performance by the predictive pdf at the exact observed value without considering its neighborhood. Taking a neighborhood with radius ϵ into account, we generalize the logarithm score into an ϵ-logarithm score as follows.

Definition 1 (ϵ-logarithm score) *Given a neighborhood radius $\epsilon \ge 0$, a random variable $\mathbf{x}$ to be predicted, the ϵ-logarithm score evaluates the quality of a predictive distribution $\hat{p}_{\mathbf{x}} \in \mathcal{P}$ on a realized outcome x by*

$$\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, x) := \begin{cases} \log \hat{p}_{\mathbf{x}}(x) & \epsilon = 0, \\ \log \int_{\|s-x\|_\infty \le \epsilon} \hat{p}_{\mathbf{x}}(s)\mathrm{d}s & \epsilon > 0, \end{cases} \tag{5}$$

and the expected ϵ-logarithm score is

$$\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) := \mathbb{E}_{x\sim p_{\mathbf{x}}} \mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, x). \tag{6}$$

While the ϵ-logarithm score $\mathcal{L}_\epsilon(p_{\mathbf{x}}, x)$ scores a one-step prediction, we can naturally extend this definition to the trajectory prediction of SDSs.

Definition 2 (ϵ-logarithm score for SDSs) *Given $\epsilon \ge 0$, a state trajectory $x_{1:T}$ generated from an SDS Φ, the ϵ-logarithm score for a probabilistic predictor $\hat{p}$ on this trajectory is the average of one-step scores, i.e.,*

$$\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T}) := \frac{1}{T}\sum_{k=1}^{T} \mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1}), x_k\right), \tag{7}$$

where $\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} \in \mathcal{P}$ for $k = 1, \ldots, T$. The expected ϵ-logarithm score is denoted as

$$\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}) := \mathbb{E}_{x_{1:T}} \bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T}). \tag{8}$$

Problem 1 *Evaluate the expected ϵ-logarithm score and characterize the probabilistic predictability of an SDS by optimizing the expected ϵ-logarithm score, i.e.,*

$$\max_{\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} \in \mathcal{P}\ \forall k\in[1,T]} \bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}). \tag{9}$$

Though the expected ϵ-logarithm score is theoretically appealing, practically evaluating it requires many samples for averaging. However, the data under a typical SDS prediction scenario is usually a state trajectory rather than repeated samplings for trajectories.

Problem 2 *Given any state trajectory $x_{1:\infty}$ generated from Φ, does the ϵ-logarithm score $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T})$ converge as T approaches infinity? If it does converge, what is the converged value and how fast the convergence is?*

4 Probabilistic Predictability of SDSs

Let $\mathbf{u}_\epsilon$ be a random variable obeying a uniform distribution over the ϵ-ball $\{x \in \mathbb{R}^{d_x} \mid \|x\|_\infty \leq \epsilon\}$, and we denote its pdf as $p_{\mathbf{u}_\epsilon}$. Convolving each pdf in $\mathcal{P}$ by $p_{\mathbf{u}_\epsilon}$, we denote the convolution range set as $\mathcal{R}_\epsilon = \{\tilde{p} \mid \tilde{p} = \hat{p} * p_{\mathbf{u}_\epsilon}, \hat{p} \in \mathcal{P}\}$. Then, we obtain the probabilistic predictability of SDSs by solving (9).

Theorem 1 (Probabilistic predictability) *Given a neighborhood radius $\epsilon \geq 0$ and a trajectory length $T \geq 1$, i) the optimal expected ϵ-logarithm score for predicting the trajectory $\mathbf{x}_{1:T}$ of an SDS Φ is*

$$\begin{aligned}&\max_{\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} \in \mathcal{P}\ \forall k \in [1,T]} \bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}) \\ =&\mathbb{I}_{\epsilon>0} \cdot d_x \log(2\epsilon) - \frac{1}{T}\mathrm{H}(p_{\mathbf{x}_{1:T}} || \tilde{p}^\star_{\mathbf{x}_{1:T}}),\end{aligned} \tag{10}$$

where $\tilde{p}^\star_{\mathbf{x}_{1:T}}$ satisfies the following equations $\forall\, k \in [1,T]$,

$$\tilde{p}^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} = \arg\min_{\tilde{p} \in \mathcal{R}_\epsilon} \mathrm{H}(p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} || \tilde{p}). \tag{11}$$

ii) The optimal predictor $p^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$ attaining the predictability is the deconvolution of $\tilde{p}^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$ by $p_{\mathbf{u}_\epsilon}$, i.e.,

$$p^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} = \tilde{p}^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} *\!/\!* p_{\mathbf{u}_\epsilon}. \tag{12}$$

PROOF. When $\epsilon = 0$, the scoring rule degenerates into the salient logarithm score, whose expectation attains its maximum $\frac{1}{T}\mathrm{H}_\mathrm{d}(p_{\mathbf{x}_{1:T}})$ under the optimal predictor $p^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} = p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$. Notice that $\mathcal{R}_0$ contains all pdfs, (11) adimts that $\tilde{p}^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} = p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$. Moreover, since $p_{\mathbf{u}_0}$ is the Dirac function, the deconvolution (12) has correctly revealed the fact that the optimal predictive pdf should be the ground truth pdf. Finally, since $0\log(0)$ is set to 0, (10) immediately holds. When $\epsilon > 0$, the expected ϵ-logarithm score at any step $k \in [1,T]$ is

$$\begin{aligned}&\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1}), p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1})\right) \\ \overset{(i)}{=}&\int_{x\in\mathbb{R}^{d_x}} q(x) \log\left(\int_{\|s-x\|_\infty \leq \epsilon} \hat{q}(s)\mathrm{d}s\right)\mathrm{d}x \\ =&d_x\log(2\epsilon) + \int_{x\in\mathbb{R}^{d_x}} q(x)\log\left(\frac{\int_{\|s-x\|_\infty\leq\epsilon}\hat{q}(s)\mathrm{d}s}{(2\epsilon)^{d_x}}\right)\mathrm{d}x \\ \overset{(ii)}{=}&d_x\log(2\epsilon) + \int_{x\in\mathbb{R}^{d_x}} q(x)\log\left[\hat{q} * p_{\mathbf{u}_\epsilon}(x)\right]\mathrm{d}x \\ \overset{(iii)}{\leq}&d_x\log(2\epsilon) - \mathrm{H}(q||\tilde{q}^\star),\end{aligned}$$

where (i) follows by denoting $p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1})$ as $q(\cdot)$ and $\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1})$ as $\hat{q}(\cdot)$; (ii) follows from the definition of convolution and (iii) holds by denoting $\tilde{p}^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1})$ as $\tilde{q}^\star$ and the definition in (11). Therefore, $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}})$ can be maximized as follows

$$\begin{aligned}&\mathbb{E}_{x_{1:T}}\frac{1}{T}\sum_{k=1}^{T}\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1}), x_k\right) \\ =&\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}_{x_{1:k-1}}\mathcal{L}_\epsilon\big(\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot\,|\,x_{1:k-1}), p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot\,|\,x_{1:k-1})\big) \\ \overset{(i)}{\leq}&\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}_{x_{1:k-1}}\Big[d_x\log(2\epsilon) - \\ &\mathrm{H}\left(p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1})||\tilde{p}^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot \mid x_{1:k-1})\right)\Big] \\ \overset{(ii)}{=}&d_x\log(2\epsilon) - \frac{1}{T}\mathrm{H}(p_{\mathbf{x}_{1:T}}||\tilde{p}^\star_{\mathbf{x}_{1:T}}),\end{aligned}$$

where (i) follows from the previous one-step inequality, and the equality holds when (11) is satisfied; (ii) holds according to the chain rules for joint entropy and relative entropy [23, pp. 22-24]. Given $\tilde{p}^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} \in \mathcal{R}_\epsilon$, there is $p^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} \in \mathcal{P}$ such that (12) holds, which is the optimal predictor that attains the probalistic predictability.

Remark 1 *One can get an upper bound for* (10)*:*

$$\begin{aligned}&d_x\log(2\epsilon) - \frac{1}{T}\mathrm{H}(p_{\mathbf{x}_{1:T}}||\tilde{p}^\star_{\mathbf{x}_{1:T}}) \\ \leq&d_x\log(2\epsilon) - \frac{1}{T}\mathrm{H}_\mathrm{d}(p_{\mathbf{x}_{1:T}}),\end{aligned} \tag{13}$$

where the equality holds if and only if $p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} \in \mathcal{R}_\epsilon \forall k \in [1,T]$, then $\tilde{p}^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} = p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$ and the optimal predictor $p^\star_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$ is the deconvolution $p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} *\!/\!* p_{\mathbf{u}_\epsilon}$. As for a general $\epsilon \in \mathbb{R}_+$, the density deconvolution may not exist. Following the famous Bochner's theorem [24, p. 291], we immediately have a necessary and sufficient condition that guarantees the existence of density convolution: let the characteristic functions of $p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$ and $p_{\mathbf{u}_\epsilon}$ be $\phi_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}$ and $\phi_{\mathbf{u}_\epsilon}$ respectively, then $p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} *\!/\!* p_{\mathbf{u}_\epsilon}$ is a pdf if and only if $\phi_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}/\phi_{\mathbf{u}_\epsilon}$ is a positive-definite function.

5 Approximation and Convergence Analysis

In this section, we first provide approximations to the expected ϵ-logarithm score with an error of scale $\mathcal{O}(\epsilon)$. Next, while $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}})$ considers the expected performance over all possible trajectories, we also analyze the asymptotic behaviors of the $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T})$ on any single trajectory. In particular, we characterize the convergence speed for the SDSs with i.i.d process noises.

Typically, there is no explicit expression for $\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}})$ under general pdfs. To handle this problem, we first utilize a partition-based method to formally evaluate $\mathcal{L}_\epsilon$.

Definition 3 (Unifrom grid partition) *Given $l \in \mathbb{R}_+$, a uniform grid partition Σ^l is a set that divides $\mathbb{R}^{d_x}$ into disjoint cubes with edge lengths equaling l. Specifically, we denote*

$$\Sigma^\ell := \left\{A_v^\ell\right\}_{v\in\mathbb{Z}^{d_x}}, \quad \Theta_{\Sigma^\ell}(x) := \sum_{v\in\mathbb{Z}^{d_x}} v \cdot \mathbb{I}_{A_v^\ell}(x),$$

where $A_v^\ell = \{a \in \mathbb{R}^{d_x} \mid a_{(i)} \in \left(v_{(i)}\ell, (v_{(i)}+1)\ell\right]\}$ is a cube in $\mathbb{R}^{d_x}$ with size ℓ and index v. $\forall x \in \mathbb{R}^{d_x}$, $\Theta_{\Sigma^\ell}(x)$ is the cube index of x under partition Σ^ℓ.

Given a uniform grid partition Σ^ℓ, a pdf $\hat{p}_{\mathbf{x}}$ can be approximated by a pmf $\hat{p}_{\mathbf{x}}^{\Sigma^\ell}$, where

$$\hat{p}_{\mathbf{x}}^{\Sigma^\ell}(v) := \int_{A_v^\ell} \hat{p}_{\mathbf{x}}(s)\mathrm{d}s, \forall v \in \mathbb{Z}^{d_x}. \tag{14}$$

Then we define the Σ^ℓ-logarithm score as $\mathcal{L}_{\Sigma^\ell}(\hat{p}_{\mathbf{x}}, x) := \log \hat{p}_{\mathbf{x}}^{\Sigma^\ell}(\Theta_{\Sigma^\ell}(x))$, and the expected Σ^ℓ-logarithm score is defined as $\mathcal{L}_{\Sigma^\ell}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) := \mathbb{E}_x \mathcal{L}_{\Sigma^\ell}(\hat{p}_{\mathbf{x}}, x)$.

$\mathcal{L}_{\Sigma^\ell}$ can be evaluated as the negative cross entropy.

Proposition 1 (Evaluation of $\mathcal{L}_{\Sigma^\ell}$) *Given a random variable $\mathbf{x}$, a partition Σ^ℓ on $\mathbb{R}^{d_x}$ and a probabilistic predictor $\hat{p}_{\mathbf{x}}$, the expected Σ^ℓ-logarithm score is*

$$\mathcal{L}_{\Sigma^\ell}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) = -\mathrm{H}(p_{\mathbf{x}}^{\Sigma^\ell} || \hat{p}_{\mathbf{x}}^{\Sigma^\ell}). \tag{15}$$

To evaluate $\mathcal{L}_\epsilon$, we first develop the following lemma to address an inequality relationship between $\mathcal{L}_\epsilon$ and $\mathcal{L}_{\Sigma^\ell}$.

Lemma 1 *Given a neighborhood radius $\epsilon \geq 0$, a pdf $p_{\mathbf{x}} \in \mathcal{P}$ and a predictor $\hat{p}_{\mathbf{x}} \in \mathcal{P}$, $\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}})$ is bounded by*

$$\mathcal{L}_{\Sigma^\epsilon}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) \leq \mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) \leq \max_{\ell>0} \mathcal{L}_{\Sigma^\ell}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}). \tag{16}$$

PROOF. Please see Appendix A.

This lemma provides a coarse way to bound the ϵ-logarithm score by Σ^ℓ-logarithm score. It helps to guarantee the existence of a special partition $\Sigma^\star$ that transforms the evaluation of ϵ-logarithm score to evaluating $\Sigma^\star$-logarithm score.

Theorem 2 (Evaluation of $\mathcal{L}_\epsilon$) *Given $\epsilon \geq 0$, a pdf $p_{\mathbf{x}} \in \mathcal{P}$ and a predictor $\hat{p}_{\mathbf{x}} \in \mathcal{P}$, there exists a uniform grid partition $\Sigma^\star$ such that*

$$\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) = -\mathrm{H}(p_{\mathbf{x}}^{\Sigma^\star} || \hat{p}_{\mathbf{x}}^{\Sigma^\star}). \tag{17}$$

PROOF. Please see Appendix B.

Exploiting the partition-based formal evaluation, we can approximate $\mathcal{L}_\epsilon$ when pdfs are continuous. Let $\mathcal{P}_c$ denote the space of continuous pdfs over $\mathbb{R}^{d_x}$, we have the following lemma.

Lemma 2 (Approximation of $\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}})$) *Given $\epsilon \geq 0$, a pdf $p_{\mathbf{x}} \in \mathcal{P}_c$ and a probabilistic predictor $\hat{p}_{\mathbf{x}} \in \mathcal{P}_c$, the expected ϵ-logarithm score $\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}})$ is approximated as*

$$|\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) - \mathbb{I}_{\epsilon>0} \cdot d_x \log(2\epsilon) + \mathrm{H}(p_{\mathbf{x}} || \hat{p}_{\mathbf{x}})| = \mathcal{O}(\epsilon).$$

PROOF. Please see Appendix C.

Next, we can extend this one-step approximation result to the SDS trajectory.

Theorem 3 (Approximation of $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}})$) *Given $\epsilon \geq 0$, a trajectory $\mathbf{x}_{1:T}$ with continuous pdf $p_{\mathbf{x}_{1:T}}$ and a predictor $\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}} \in \mathcal{P}_c$ at each step k, the expected ϵ-logarithm score $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}})$ is approximated as*

$$\left|\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}) - \mathbb{I}_{\epsilon>0} \cdot d_x \log(2\epsilon) + \frac{1}{T}\mathrm{H}(p_{\mathbf{x}_{1:T}} || \hat{p}_{\mathbf{x}_{1:T}})\right| = O(\epsilon).$$

PROOF. Please see Appendix D.

Particularly when $\hat{p}_{\mathbf{x}_{1:T}} = p_{\mathbf{x}_{1:T}}$, the convergence of $\bar{\mathcal{L}}_\epsilon$ is mainly determined by the convergence of $\frac{1}{T}\mathrm{H_d}(\mathbf{x}_{1:T})$, which is the entropy rate of the stochastic process $\{\mathbf{x}_k\}_{k=1}^\infty$. However, the entropy rate is not guaranteed to always exist when the process noises are not i.i.d [23, p. 74]. Therefore, to analyze the convergence of the ϵ-logarithm score for a given state trajectory, one should focus on the SDSs with i.i.d process noises. Additionally, when the system dynamics are known to the predictor, the predictive pdf for state $\mathbf{x}_k$ can be reduced to a predictive pdf $\hat{p}_{\mathbf{w}} \in \mathcal{P}$ for the noise $\mathbf{w}$, i.e.,

$$\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(x_k \mid x_{1:k-1}) = \hat{p}_{\mathbf{w}}(x_k - f(x_{k-1})). \tag{18}$$

Theorem 4 *Given $\epsilon \geq 0$, a state trajectory $x_{1:T}$ of an SDS subjected to i.i.d process noises with pdf $p_{\mathbf{w}} \in \mathcal{P}$ and a predictor $\hat{p}_{\mathbf{x}_{1:T}}$ satisfying* (18),
i) the expected ϵ-logarithm score is $\mathcal{L}_\epsilon(\hat{p}_{\mathbf{w}}, p_{\mathbf{w}})$, i.e.,

$$\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}) = \mathcal{L}_\epsilon(\hat{p}_{\mathbf{w}}, p_{\mathbf{w}}).$$

ii) As T approaches infinity, the ϵ-logarithm score on any single trajectory converges to the expectation, i.e.,

$$\lim_{T\to\infty} \bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T}) \overset{P}{=} \bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}}).$$

iii) Moreover, if $\mathbb{E}_w \mathcal{L}_\epsilon(\hat{p}_{\mathbf{w}}, w)^2 < \infty$, the convergence speed is $\mathcal{O}_p(\frac{1}{\sqrt{T}})$, i.e., $\forall \delta > 0$, there is

$$\Pr\left\{|\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T}) - \bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}})| \geq \delta\right\} = \mathcal{O}(\frac{1}{\sqrt{T}}).$$

PROOF. Please see Appendix E.

Remark 2 *In practice, there is no need to use the sample average of a large number of trajectories to approximate $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, p_{\mathbf{x}_{1:T}})$ when the SDS has i.i.d process noises. Specifically, as guaranteed by Theorem 4, calculating the ϵ-logarithm score on any single trajectory $x_{1:T}$ will quickly converge to the expectation with the speed $\mathcal{O}_p(\frac{1}{\sqrt{T}})$.*

6 Simulation

6.1 Simulation Setup

We study a linear SDS subjected to i.i.d noises that are uniformly distributed over the cube $[-1, 1]^2$,

$$\Phi: \mathbf{x}_{k+1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \mathbf{x}_k + \mathbf{w}_k, \quad \mathbf{w}_k \overset{\text{i.i.d}}{\sim} U([-1, 1]^2).$$

Then, we randomly generate trajectories $\{x_{1:T}^{(n)}\}_{n=1}^{10^5}$ of Φ with length $T = 30$ starting from a random initial state. The predictor $\hat{p}_{\mathbf{x}_{1:T}}$ satisfies the condition (18) in Theorem 4, and we let $\hat{p}_{\mathbf{w}}$ be a standard two-dimensional Gaussian distribution. We choose the neighborhood radius ϵ as $0.25, 0.5, 0.75$ respectively. Next, we numerically compute the score $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}}, x_{1:T}^{(n)})$ for each trajectory. In Fig. 1, we use the scores' mean to evaluate the expected score and calculate 0.025 and 0.975 quantiles to form a 95% confidence interval. According to Theorem 1, the predictability of Φ can be explicitly calculated as $2\log(2\epsilon) - \log 4$, plotted as red dotted line. According to Theorem 3, the expected score can be approximated by $2\log(2\epsilon) - \log(2\pi) - \frac{1}{3}$, plotted as blue dotted line. Finally, we randomly choose three trajectories for each ϵ and plotted their score curves.

6.2 Results and Analysis

As Fig. 1 shows, the distances between the blue solid lines (the real expected score) and the blue dotted lines (the approximated expected score) are indeed of scale $\mathcal{O}(\epsilon)$. Even for the extreme setting in Fig. 1(c), where $\epsilon = 0.75$ is close to the radius 1 of the noises' support $[-1, 1]^2$, the approximation error is still around 0.1. Nevertheless, a practically reasonable choice for the error tolerance ϵ should always be much smaller than the support of noises, otherwise the performances of different predictors would be indistinguishable. Therefore, our approximation in Theorem 3 is effective.

Next, Fig. 1 shows that all individual trajectories approach fast to the blue dotted lines in less than 20 time steps. This quick convergence is ensured by Theorem 4 in the sense of probability. Benefiting from the quick convergence property of the score on individual trajectories, evaluating the expected score is easy to implement on one trajectory without the need for repeated samplings of different trajectories.

7 Conclusion

In this paper, we have proposed an ϵ-logarithm score to assess the performance of probabilistic predictions in stochastic dynamical systems. We have evaluated the probabilistic predictability of an SDS by optimizing the expected score over the space of probability measures. It has allowed us to quantitatively analyze how the predictability of the system depends on the neighborhood radius, differential entropies of process noises, and system dimension. Additionally, we have provided approximations to the expected score for general non-optimal predictors. We have also analyzed the asymptotic convergence behavior of our score on any individual trajectory. It is proved that the score converges to the expected score when the process noises are independent and identically distributed, with a convergence speed of scale $\mathcal{O}_p(T^{-\frac{1}{2}})$ for the trajectory length T.

References

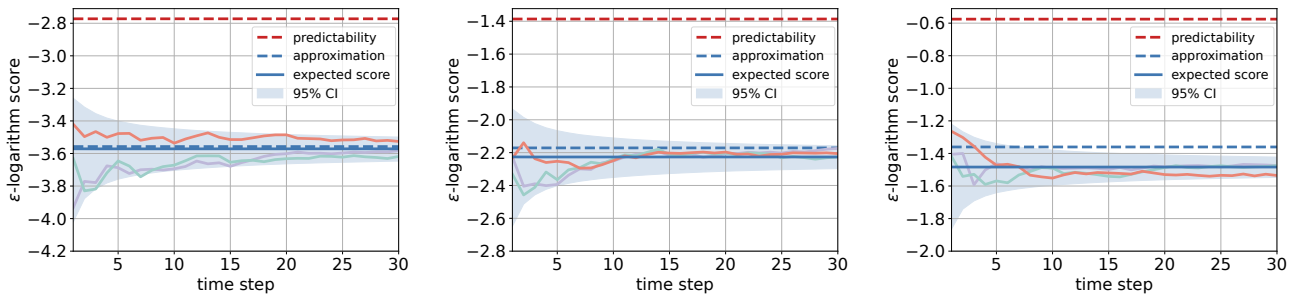
[1] T. Xu and J. He, "Predictability of stochastic dynamical system: A probabilistic perspective," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 5466–5471, Dec. 2022.

[2] D. Landgraf, A. Völz, F. Berkel, K. Schmidt, T. Specker, and K. Graichen, "Probabilistic prediction methods for nonlinear systems with application to stochastic model predictive control," *Annual Reviews in Control*, vol. 56, p. 100905, Jan. 2023.

[3] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, pp. 359–378, Mar. 2007.

[4] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 125–151, 2014.

[5] N. Vaswani, "Particle filtering for large-dimensional state spaces with multimodal observation likelihoods," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4583–4597, 2008.



(a) ϵ=0.25 (b) ϵ=0.50 (c) ϵ=0.75

Fig. 1. ϵ-logarithm scores $\bar{\mathcal{L}}_\epsilon(p_{\mathbf{x}_{1:T}}, x_{1:T}^{(n)})$ v.s. the time step T: given ϵ and trajectories $\{x_{1:T}^{(n)}\}_{n=1}^{10^5}$ of Φ, i) score curves on three individual trajectories are randomly chosen for presentation; ii) the expected scores (blue solid line) and 95% confidence intervals (blue transparent area) at each step are calculated from the scores on 100, 000 trajectories; iii) the predictability (red dotted line) is evaluated by Theorem 1 and the expected score's approximation (blue dotted line) is evaluated by Theorem 3.

[6] Z. Xiong, A. Jonnarth, A. Eldesokey, J. Johnander, B. Wandt, and P.-E. Forssén, "Hinge-wasserstein: Estimating multimodal aleatoric uncertainty in regression tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3471–3480, 2024.

[7] B. Kouvaritakis, M. Cannon, S. V. Raković, and Q. Cheng, "Explicit use of probabilistic distributions in linear predictive control," *Automatica*, vol. 46, pp. 1719–1724, Oct. 2010.

[8] T. Sauder, S. Marelli, and A. J. Sørensen, "Probabilistic robust design of control systems for high-fidelity cyber–physical testing," *Automatica*, vol. 101, pp. 111–119, Mar. 2019.

[9] C. E. Roelofse and C. E. van Daalen, "An accurate and efficient approach to probabilistic conflict prediction," *Automatica*, vol. 153, p. 111021, July 2023.

[10] E. N. Lorenz, "Predictability: A problem partly solved," in *Proc. Seminar on Predictability*, vol. 1, 1996.

[11] G. Boffetta, M. Cencini, M. Falcioni, and A. Vulpiani, "Predictability: A way to characterize complexity," *Physics Reports*, vol. 356, pp. 367–474, Jan. 2002.

[12] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge university press, 2003.

[13] F. Biondi, A. Legay, B. F. Nielsen, and A. Wąsowski, "Maximizing entropy over markov processes," *Journal of Logical and Algebraic Methods in Programming*, vol. 83, pp. 384–399, Sept. 2014.

[14] Y. Savas, M. Hibbard, B. Wu, T. Tanaka, and U. Topcu, "Entropy maximization for partially observable markov decision processes," *IEEE Transactions on Automatic Control*, vol. 67, pp. 6948–6955, Dec. 2022.

[15] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[16] Y. Li, D. Jin, P. Hui, Z. Wang, and S. Chen, "Limits of predictability for large-scale urban vehicular mobility," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, pp. 2671–2682, Dec. 2014.

[17] T. DelSole, "Predictability and information theory. part i: Measures of predictability," *Journal of the Atmospheric Sciences*, vol. 61, pp. 2425–2440, Oct. 2004.

[18] T. DelSole, "Predictability and information theory. part ii: Imperfect forecasts," *Journal of the Atmospheric Sciences*, vol. 62, pp. 3368–3381, Sept. 2005.

[19] T. DelSole and M. K. Tippett, "Predictability: Recent insights from information theory," *Reviews of Geophysics*, vol. 45, no. 4, 2007.

[20] C. Byrnes, A. Lindquist, and T. McGregor, "Predictability and unpredictability in kalman filtering," *IEEE Transactions on Automatic Control*, vol. 36, pp. 563–579, May 1991.

[21] S. Yasini and K. Pelckmans, "Worst-case prediction performance analysis of the kalman filter," *IEEE Transactions on Automatic Control*, vol. 63, pp. 1768–1775, June 2018.

[22] M. Parry, A. P. Dawid, and S. Lauritzen, "Proper local scoring rules," *The Annals of Statistics*, vol. 40, Feb. 2012.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*. Hoboken, N.J: Wiley-Interscience, 2nd edition ed., July 2006.

[24] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis: Volume II: Structure and Analysis for Compact Groups Analysis on Locally Compact Abelian Groups*, vol. 152. Springer, 2013.

[25] W. Rudin, *Principles of Mathematical Analysis*, vol. 3. McGraw-hill New York, 1976.

A Proof of Lemma 1

For the convenience of notation, we omit ∞ in the notation of $\|\cdot\|_\infty$ from now on.

Guaranteed by the law of large numbers, given samples $\{x^r\}_{r=1}^\infty$ where $x^r \overset{i.i.d}{\sim} p_{\mathbf{x}}$, there is

$$\begin{aligned}\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) &= \int_{\mathbb{R}^{d_x}} p_{\mathbf{x}}(x) \log\left(\int_{\|\hat{x}-x\|\le\epsilon} \hat{p}_{\mathbf{x}}(\hat{x})\mathrm{d}\hat{x}\right)\mathrm{d}x \\ &= \lim_{R\to\infty} \frac{1}{R}\sum_{r=1}^{R} \log\left(\int_{\|\hat{x}-x^r\|\le\epsilon} \hat{p}_{\mathbf{x}}(\hat{x})\mathrm{d}\hat{x}\right).\end{aligned}$$

Viewing the integration above as the expectation for an indicator random variable $\mathbf{1}_{\|\hat{\mathbf{x}}-x^r\|\le\epsilon}$ (the indicator takes value 1 when $\|\hat{\mathbf{x}}-x^r\|\le\epsilon$ and takes value 0 otherwise),

it can be expressed as the asymptotic mean based on samples $\{\hat{x}_k^r\}_{k=1}^{\infty}$ where $\hat{x}_k^r \overset{i.i.d}{\sim} \hat{p}_{\mathbf{x}}$, i.e.,

$$\int_{\|\hat{x}-x^r\|\leq\epsilon} \hat{p}_{\mathbf{x}}(\hat{x})\mathrm{d}\hat{x} = \mathbb{E}_{\hat{\mathbf{x}}}\mathbf{1}_{\|\hat{\mathbf{x}}-x^r\|\leq\epsilon} = \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\mathcal{K}_{\epsilon}^{(k,r)},$$

where $\mathcal{K}_{\epsilon}^{(k,r)} = \mathbf{1}_{\|\hat{x}_k^r-x^r\|\leq\epsilon}$. It follows that

$$\mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) = \lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^{R}\log\left(\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\mathcal{K}_{\epsilon}^{(k,r)}\right).$$

Similarly, the expected Σ^{ℓ}-logarithm score is

$$\mathcal{L}_{\Sigma^{\ell}}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) = \lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^{R}\log\left(\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\mathcal{K}_{\Sigma^{\ell}}^{(k,r)}\right),$$

where $\mathcal{K}_{\Sigma^{\ell}}^{(k,r)} = \mathbf{1}_{\Theta_{\Sigma^{\ell}}(\hat{x}_k^r)=\Theta_{\Sigma^{\ell}}(x^r)}$.

For the left inequality, there is

$$\mathcal{K}_{\Sigma^{\epsilon}}^{(k,r)} = 1 \overset{(i)}{\Leftrightarrow} \Theta_{\Sigma^{\epsilon}}(\hat{x}_k^r) = \Theta_{\Sigma^{\epsilon}}(x^r) \overset{(ii)}{\Rightarrow} \|\hat{x}_k^r - x^r\| \leq \epsilon \overset{(iii)}{\Leftrightarrow} \mathcal{K}_{\epsilon}^{(k,r)} = 1,$$

where (i) and (iii) follows from the definitions of $\mathcal{K}_{\Sigma^{\epsilon}}^{(k,r)}$ and $\mathcal{K}_{\epsilon}^{(k,r)}$ respectively; (ii) holds because $\Theta_{\Sigma^{\epsilon}}(\hat{x}_k^r) = \Theta_{\Sigma^{\epsilon}}(x^r)$ indicates the existence of an ϵ-sized cube $A \in \Sigma^{\epsilon}$ such that $\hat{x}_k^r, x^r \in A$, thus $\|\hat{x}_k^r - x^r\| \leq \epsilon$. Therefore, given any $R, K \in \mathbb{R}_+$ and any possible samples $\left\{x^r, \{\hat{x}_k^r\}_{k=1}^{K}\right\}_{r=1}^{R}$, there is $\mathcal{K}_{\Sigma^{\epsilon}}^{(k,r)} \leq \mathcal{K}_{\epsilon}^{(k,r)}$. Thus

$$\frac{1}{R}\sum_{r=1}^{R}\log\left(\frac{1}{K}\sum_{k=1}^{K}\mathcal{K}_{\Sigma^{\epsilon}}^{(k,r)}\right) \leq \frac{1}{R}\sum_{r=1}^{R}\log\left(\frac{1}{K}\sum_{k=1}^{K}\mathcal{K}_{\epsilon}^{(k,r)}\right).$$

Letting $R, K \to \infty$, we have $\mathcal{L}_{\Sigma^{\epsilon}}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) \leq \mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}})$. Considering the extreme case where $\Sigma = \{\mathbb{R}^{d_x}\}$, given any $R, K \in \mathbb{Z}_+$ and samples $\left\{x^r, \{\hat{x}_k^r\}_{k=1}^{K}\right\}_{r=1}^{R}$, there is $\mathcal{K}_{\Sigma}^{(k,r)} \geq \mathcal{K}_{\epsilon}^{(k,r)}$. Thus

$$\mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) \leq \max_{\ell>0}\mathcal{L}_{\Sigma^{\ell}}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}).$$

B Proof of Theorem 2

Given pdfs $\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}$ over $\mathbb{R}^{d_x}$, define function $\mathcal{F}: \mathbb{R}_+ \to \mathbb{R}$ as $\mathcal{F}(\ell) := \mathcal{L}_{\Sigma^{\ell}}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}})$. We first prove that $\mathcal{F}$ is continuous, i.e., for any $l > 0$, we need to prove that given any $\xi > 0$, there exists $\delta > 0$ s.t. $\forall \ell \in (l-\delta, l+\delta)$, $|\mathcal{F}(\ell) - \mathcal{F}(l)| < \xi$. Since

$$\mathcal{F}(\ell) = \lim_{N\to\infty}\sum_{\|v\|\leq N} p_{\mathbf{x}}^{\Sigma^{\ell}}(v)\log\left(\hat{p}_{\mathbf{x}}^{\Sigma^{\ell}}(v)\right),$$

then $\forall \ell > 0$, $\exists N_{\xi}(\ell) \in \mathbb{N}$ such that

$$\left|\mathcal{F}(\ell) - \sum_{\|v\|\leq N_{\xi}(\ell)} p_{\mathbf{x}}^{\Sigma^{\ell}}(v)\log\left(\hat{p}_{\mathbf{x}}^{\Sigma^{\ell}}(v)\right)\right| < \frac{\xi}{4}.$$

Let $\bar{N} = \sup_{\kappa\in[0.9,1.1]} N_{\xi}(\kappa l)$, $\forall v$ with $\|v\| \leq \bar{N}$, one gets

$$\left|p_{\mathbf{x}}^{\Sigma^{\ell}}(v) - p_{\mathbf{x}}^{\Sigma^{l}}(v)\right| \leq \left|\int_{A_v^{\ell}\Delta A_v^{l}} p_{\mathbf{x}}(x)\mathrm{d}x\right|,$$

where Δ denotes the symmetric difference between two measurable sets. Let $\mu(\cdot)$ denote the Lebesgue measure, there is $\lim_{\ell\to l}\mu(A_v^{\ell}\Delta A_v^{l}) = 0$. Thus

$$\lim_{\ell\to l}\left|p_{\mathbf{x}}^{\Sigma^{\ell}}(v) - p_{\mathbf{x}}^{\Sigma^{l}}(v)\right| = 0,$$

which means $p_{\mathbf{x}}^{\Sigma^{\ell}}(v)$ is continuous at l. Similarly, one has $\hat{p}_{\mathbf{x}}^{\Sigma^{\ell}}(v)$ is continuous at l. Then, there exists $\delta > 0$ such that when $|\ell - l| \leq \delta$ one has

$$\left|\sum_{\|v\|\leq\bar{N}}\left\{p_{\mathbf{x}}^{\Sigma^{\ell}}(v)\log\left(\hat{p}_{\mathbf{x}}^{\Sigma^{\ell}}(v)\right) - p_{\mathbf{x}}^{\Sigma^{l}}(v)\log\left(\hat{p}_{\mathbf{x}}^{\Sigma^{l}}(v)\right)\right\}\right| \leq \frac{\xi}{2}.$$

Hence, for any ℓ such that $|\ell - l| \leq \min\{\delta, 0.1l\}$, there is

$$\begin{aligned}&\left|\mathcal{F}(\ell) - \sum_{\|v\|\leq\bar{N}} p_{\mathbf{x}}^{\Sigma^{\ell}}(v)\log\left(\hat{p}_{\mathbf{x}}^{\Sigma^{\ell}}(v)\right) + \sum_{\|v\|\leq\bar{N}} p_{\mathbf{x}}^{\Sigma^{\ell}}(v)\log\left(\hat{p}_{\mathbf{x}}^{\Sigma^{\ell}}(v)\right)\right.\\ &\left. - \sum_{\|v\|\leq\bar{N}} p_{\mathbf{x}}^{\Sigma^{l}}(v)\log\left(\hat{p}_{\mathbf{x}}^{\Sigma^{l}}(v)\right) + \sum_{\|v\|\leq\bar{N}} p_{\mathbf{x}}^{\Sigma^{l}}(v)\log\left(\hat{p}_{\mathbf{x}}^{\Sigma^{l}}(v)\right) - \mathcal{F}(l)\right|\\ &\leq \frac{\xi}{4} + \frac{\xi}{2} + \frac{\xi}{4} = \xi,\end{aligned}$$

and the continuity of $\mathcal{F}$ is proved. Next, as indicated by Lemma 1, $\lim_{\ell\to\infty}\mathcal{F}(\ell) = 0 > \mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}})$. Since $\mathcal{F}$ is continuous, there exists $l_1 > \epsilon$ such that $\mathcal{F}(l_1) > \mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}})$. Applying the intermediate value theorem [25] to $\mathcal{F}$ at interval $[\epsilon, l_1]$, there exists $l^{\star} \in [\epsilon, l_1]$ such that $\mathcal{L}_{\epsilon}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}}) = \mathcal{L}_{\Sigma^{l^{\star}}}(\hat{p}_{\mathbf{x}}, p_{\mathbf{x}})$. Letting $\Sigma^{\star} = \Sigma^{l^{\star}}$, the proof is completed following from the evaluation in Proposition 1.

C Proof of Lemma 2

When $\epsilon = 0$, the result trivially follows from the definition of the expected logarithm score. When $\epsilon > 0$, define an error functional $\delta_{\epsilon}^{N}: \mathcal{P}_c \to \mathbb{R}$ by

$$\delta_{\epsilon}^{N}(h) := \max_{\|x_1-x_2\|\leq\epsilon, \max\{\|x_1\|,\|x_2\|\}\leq N}\left|\log\frac{h(x_1)}{h(x_2)}\right|,$$

where $N \in \mathbb{R}_+$, $x_1, x_2 \in \mathbb{R}^{d_x}$. $\delta_\epsilon^N(h)$ is monotonically increasing regarding N or ϵ because the feasible region for (x_1, x_2) is enlarged when N or ϵ increases.

According to Theorem 2, $\exists \Sigma^\star$ s.t. $\mathcal{L}_\epsilon(\hat{p}_\mathbf{x}, p_\mathbf{x}) = \mathcal{L}_{\Sigma^\star}(\hat{p}_\mathbf{x}, p_\mathbf{x})$. Suppose $\Sigma^\star = \Sigma^{\kappa\epsilon} = \{A_v\}_{v\in\mathbb{Z}^{d_x}}$. Since $p_\mathbf{x}(\cdot)$ is continuous and any cube A_v is bounded, according to the intermediate value theorem, $\forall v \in \mathbb{Z}^{d_x}, \exists a_v \in A_v$ such that $p_\mathbf{x}(a_v)|A_v| = p_\mathbf{x}^{\Sigma^\star}(v)$, where $|A_v|$ denotes the Lebesgue volume of A_v. Then $\mathcal{L}_\epsilon(\hat{p}_\mathbf{x}, p_\mathbf{x})$ equals

$$\begin{aligned}
&\sum_{v\in\mathbb{Z}^{d_x}} p_\mathbf{x}^{\Sigma^\star}(v)\log\left[\hat{p}_\mathbf{x}(a_v)\cdot|A_v|\right]\\
=&\sum_{v\in\mathbb{Z}^{d_x}} p_\mathbf{x}^{\Sigma^\star}(v)\log\left[|A_v|\right] + \sum_{v\in\mathbb{Z}^{d_x}} p_\mathbf{x}(a_v)|A_v|\log\left[\hat{p}_\mathbf{x}(a_v)\right]\\
=&d_x\log(\kappa\epsilon) + \sum_{v\in\mathbb{Z}^{d_x}} p_\mathbf{x}(a_v)|A_v|\log\left[\hat{p}_\mathbf{x}(a_v)\right].
\end{aligned}$$

Notice that $\forall\tau \in (0, \epsilon/2)$, $\exists N_1 > 0$ such that

$$\left|\sum_{\|v\|>N_1/(\kappa\epsilon)} p_\mathbf{x}(a_v)|A_v|\log\left[\hat{p}_\mathbf{x}(a_v)\right]\right| < \tau.$$

Next, notice that $\mathrm{H}(p_\mathbf{x}||\hat{p}_\mathbf{x})$ is defined by an integration over $\mathbb{R}^{d_x}$, $\forall\tau \in (0, \epsilon/2)$, $\exists N_2 > 0$ such that

$$\left|\sum_{\|v\|>N_2/(\kappa\epsilon)} \int_{A_v} p_\mathbf{x}(x)\log\hat{p}_\mathbf{x}(x)\mathrm{d}x\right| < \tau.$$

Let $N_3 = \max\{N_1, N_2\}$, there is

$$\begin{aligned}
&\left|\sum_{v\in\mathbb{Z}^{d_x}} p_\mathbf{x}(a_v)|A_v|\log\left[\hat{p}_\mathbf{x}(a_v)\right] + \mathrm{H}(p_\mathbf{x}||\hat{p}_\mathbf{x})\right|\\
=&\left|\sum_{v\in\mathbb{Z}^{d_x}} p_\mathbf{x}(a_v)|A_v|\log\left[\hat{p}_\mathbf{x}(a_v)\right] - \int_{A_v} p_\mathbf{x}(x)\log\left[\hat{p}_\mathbf{x}(x)\right]\mathrm{d}x\right|\\
\le&\left|\sum_{\|v\|>N_3/(\kappa\epsilon)} p_\mathbf{x}(a_v)|A_v|\log\left[\hat{p}_\mathbf{x}(a_v)\right]\right| +\\
&\left|\sum_{\|v\|\le N_3/(\kappa\epsilon)} \int_{A_v} p_\mathbf{x}(x)\log\left(\frac{\hat{p}_\mathbf{x}(x)}{\hat{p}_\mathbf{x}(a_v)}\right)\mathrm{d}x\right| +\\
&\left|\sum_{\|v\|>N_3/(\kappa\epsilon)} \int_{A_v} p_\mathbf{x}(x)\log\hat{p}_\mathbf{x}(x)\mathrm{d}x\right|\\
<&\tau + \sum_{\|v\|\le N_3/(\kappa\epsilon)} \left|\int_{A_v} p_\mathbf{x}(x)\mathrm{d}x\right| \delta_{\kappa\epsilon}^{N_3}(\hat{p}_\mathbf{x}) + \tau\\
\le&\delta_{\kappa\epsilon}^{N_3}(\hat{p}_\mathbf{x}) + 2\tau
\end{aligned}$$

It follows that

$$|\mathcal{L}_\epsilon(\hat{p}_\mathbf{x}, p_\mathbf{x}) + \mathrm{H}(p_\mathbf{x}||\hat{p}_\mathbf{x}) - d_x\log(\kappa\epsilon)| < \delta_{\kappa\epsilon}^{N_3}(\hat{p}_\mathbf{x}) + 2\tau. \tag{C.1}$$

(C.1) is quite close to our objective except for the $d_x\log(\kappa\epsilon)$ term. Immediately,

$$\begin{aligned}
&|\mathcal{L}_\epsilon(\hat{p}_\mathbf{x}, p_\mathbf{x}) + \mathrm{H}(p_\mathbf{x}||\hat{p}_\mathbf{x}) - d_x\log(2\epsilon)|\\
\le&|\mathcal{L}_\epsilon(\hat{p}_\mathbf{x}, p_\mathbf{x}) + \mathrm{H}(p_\mathbf{x}||\hat{p}_\mathbf{x}) - d_x\log(\kappa\epsilon)| + \left|d_x\log(\frac{2}{\kappa})\right|\\
<&\delta_{\kappa\epsilon}^{N_3}(\hat{p}_\mathbf{x}) + 2\tau + \left|d_x\log(\frac{2}{\kappa})\right|
\end{aligned} \tag{C.2}$$

The next goal is to find an upper bound for $\left|d_x\log(\frac{2}{\kappa})\right|$. Guaranteed by the law of large numbers, given a sequence of samples $\{x^r\}_{r=1}^\infty$ with $x^r \overset{i.i.d}{\sim} p_\mathbf{x}$, there is

$$\begin{aligned}
\mathcal{L}_\epsilon(\hat{p}_\mathbf{x}, p_\mathbf{x}) =& \lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^R \log\left(\int_{\|x^r-s\|\le\epsilon}\hat{p}_\mathbf{x}(s)\mathrm{d}s\right)\\
=&d_x\log(2\epsilon) + \lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^R\log\hat{p}_\mathbf{x}(x_1^r),
\end{aligned} \tag{C.3}$$

where $\|x_1^r - x^r\| \le \epsilon$ s.t. $(2\epsilon)^{d_x}\hat{p}_\mathbf{x}(x_1^r) = \int_{\|x^r-s\|}\hat{p}_\mathbf{x}(s)\mathrm{d}s$. Similarly, there is

$$\mathcal{L}_{\Sigma^\star}(\hat{p}_\mathbf{x}, p_\mathbf{x}) = d_x\log(\kappa\epsilon) + \lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^R\log\hat{p}_\mathbf{x}(x_2^r), \tag{C.4}$$

where $x_2^r \in A_{\Theta_{\Sigma^\star}(x^r)}$ s.t. $(\kappa\epsilon)^{d_x}\hat{p}_\mathbf{x}(x_2^r) = \hat{p}_\mathbf{x}^{\Sigma^\star}(\Theta_{\Sigma^\star}(x^r))$. Since (C.3) and (C.4) are equal, one has

$$d_x\log(\frac{2}{\kappa}) = \lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^R\log\hat{p}_\mathbf{x}(x_2^r) - \lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^R\log\hat{p}_\mathbf{x}(x_1^r).$$

Notice that $\forall\tau \in (0, \epsilon/2)$, $\exists N_4 > 0$ s.t.

$$\begin{aligned}
\tau >& \left|\int_{\|x\|>N_4} p_\mathbf{x}(x)\log\left(\frac{\int_{\|x-s\|\le\epsilon}\hat{p}_\mathbf{x}(s)\mathrm{d}s}{(2\epsilon)^{d_x}}\right)\mathrm{d}x\right|\\
=&\left|\lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^R\mathbb{I}_{\|x^r\|>N_4}\cdot\log\hat{p}_\mathbf{x}(x_1^r)\right|,\\
\tau >& \left|\sum_{\|v\|>N_4/(\kappa\epsilon)} p_\mathbf{x}^{\Sigma^\star}(v)\log\left(\frac{\hat{p}_\mathbf{x}^{\Sigma^\star}(v)}{(\kappa\epsilon)^{d_x}}\right)\mathrm{d}x\right|\\
=&\left|\lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^R\mathbb{I}_{\|x^r\|>N_4}\cdot\log\hat{p}_\mathbf{x}(x_2^r)\right|.
\end{aligned}$$

Therefore, $\left|d_x \log(\frac{2}{\kappa})\right|$ equals

$$\begin{aligned}
&\left|\lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^{R}\log\hat{p}_{\mathbf{x}}(x_2^r)-\lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^{R}\log\hat{p}_{\mathbf{x}}(x_1^r)\right| \\
\leq&\left|\lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^{R}\mathbb{I}_{\|x^r\|>N_4}\cdot\log\hat{p}_{\mathbf{x}}(x_1^r)\right|+ \\
&\left|\lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^{R}\mathbb{I}_{\|x^r\|>N_4}\cdot\log\hat{p}_{\mathbf{x}}(x_2^r)\right|+ \\
&\lim_{R\to\infty}\frac{1}{R}\sum_{r=1}^{R}\mathbb{I}_{\|x^r\|\leq N_4}\cdot|\log\hat{p}_{\mathbf{x}}(x_1^r)-\log\hat{p}_{\mathbf{x}}(x_2^r)| \\
<&\delta^{N_4}_{(\kappa+1)\epsilon}(\hat{p}_{\mathbf{x}})+\epsilon.
\end{aligned} \tag{C.5}$$

Let $\bar{N}=\max\{N_3,N_4\}$, we have

$$|\mathcal{L}_\epsilon(\hat{p}_{\mathbf{x}},p_{\mathbf{x}})+\mathrm{H}(p_{\mathbf{x}}||\hat{p}_{\mathbf{x}})-d_x\log(2\epsilon)|<2\left(\delta^{\bar{N}}_{(\kappa+1)\epsilon}(\hat{p}_{\mathbf{x}})+\epsilon\right).$$

Consider another functional operator $\rho:\mathcal{P}_c\to\mathbb{R}$ as the maximum value of the solution set of an inequality, i.e.,

$$\rho(h):=\max_{z\in\mathbb{R}_+}\left\{z:\left|d_x\log(\frac{2}{z})\right|\leq\delta^{\bar{N}}_{(z+1)\epsilon}(h)+\epsilon\right\}.$$

(C.5) ensures that κ belongs to the solution set. Notice that $\delta^{\bar{N}}_{z\epsilon}(\hat{p}_{\mathbf{x}})$ is upper bounded regarding z while $\left|d_x\log(\frac{2}{z})\right|$ is not, one has the solution set is upper bounded and $\rho(\hat{p}_{\mathbf{x}})$ is well-defined. It follows that

$$\left|d_x\log(\frac{2}{\kappa})\right|\leq\delta^{\bar{N}}_{\kappa\epsilon}(\hat{p}_{\mathbf{x}})+\epsilon\leq\delta^{\bar{N}}_{\rho(\hat{p}_{\mathbf{x}})\epsilon}(\hat{p}_{\mathbf{x}})+\epsilon.$$

Since $\log\hat{p}_{\mathbf{x}}(\cdot)$ is uniformly continuous over the bounded set $\{x\mid\|x\|\leq\bar{N}\}$, we have $\delta^{\bar{N}}_{\rho(\hat{p}_{\mathbf{x}})\epsilon}(\hat{p}_{\mathbf{x}})=\mathcal{O}(\epsilon)$, and the proof is completed.

D Proof of Theorem 3

When $\epsilon=0$, the result trivially follows from the definition of the expected logarithm score. When $\epsilon>0$, let $q_{x_{1:k-1}}(\cdot)=p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot\mid x_{1:k-1})$ and $\hat{q}_{x_{1:k-1}}(\cdot)=\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot\mid x_{1:k-1})$, it follows follows from the the chain rules for joint entropy and relative entropy [23, pp. 22-24] that

$$\begin{aligned}
&\left|\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}},p_{\mathbf{x}_{1:T}})-\left\{d_x\log(2\epsilon)-\frac{1}{T}\mathrm{H}(p_{\mathbf{x}_{1:T}}||\hat{p}_{\mathbf{x}_{1:T}})\right\}\right| \\
=&\left|\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}_{x_{1:k-1}}\mathcal{L}_\epsilon\left(\hat{q}_{x_{1:k-1}},q_{x_{1:k-1}}\right)\right. \\
&\left.-\left\{d_x\log(2\epsilon)-\frac{1}{T}\sum_{k=1}^{T}\mathrm{H}(p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}||\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}})\right\}\right|,
\end{aligned}$$

which can be upper bounded by

$$\begin{aligned}
&\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}_{x_{1:k-1}}\left|\mathcal{L}_\epsilon\left(q_{x_{1:k-1}},q_{x_{1:k-1}}\right)-d_x\log(2\epsilon)+\right. \\
&\left.\mathrm{H}\left(p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot|x_{1:k-1})||\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot|x_{1:k-1})\right)\right|\overset{(i)}{=}\mathcal{O}(\epsilon),
\end{aligned}$$

where (i) holds because Lemma 2 ensures each term is $\mathcal{O}(\epsilon)$, the average of finite sum is also of the scale $\mathcal{O}(\epsilon)$.

E Proof of Theorem 4

According to the assumption that the process noises are i.i.d, $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}},x_{1:T})$ equals

$$\begin{aligned}
&\frac{1}{T}\sum_{k=1}^{T}\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}(\cdot\mid x_{1:k-1}),x_k\right) \\
\overset{(i)}{=}&\frac{1}{T}\sum_{k=1}^{T}\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{x}_k|\mathbf{x}_{k-1}}(\cdot\mid x_{k-1}),x_k\right) \\
\overset{(ii)}{=}&\frac{1}{T}\sum_{k=1}^{T}\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{w}}(\cdot),x_k-f(x_{k-1})\right)\overset{(iii)}{=}\frac{1}{T}\sum_{k=1}^{T}\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{w}},w_k\right),
\end{aligned}$$

where (i) holds because when the process noises $\{\mathbf{w}_k\}_{k=1}^T$ are independent, $\{\mathbf{x}_k\}_{k=1}^T$ is a Markov process, thus $p_{\mathbf{x}_k|\mathbf{x}_{1:k-1}}=p_{\mathbf{x}_k|\mathbf{x}_{k-1}}$; (ii) and (iii) follows from the assumption on the predictors where the prediction is reduced to predicting the noises $\mathbf{w}_k\overset{i.i.d}{\sim}p_{\mathbf{w}}$ by $\hat{p}_{\mathbf{w}}$. Then we can derive $\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}},p_{\mathbf{x}_{1:T}})$ as

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left\{\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{w}},w_k\right)\right\}=\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{w}},p_{\mathbf{w}}\right).$$

Ensured by the strong law of large numbers, one has

$$\begin{aligned}
\lim_{T\to\infty}\bar{\mathcal{L}}(\hat{p}_{\mathbf{x}_{1:T}},x_{1:T})&=\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{w}},w_k\right) \\
&\xrightarrow{a.s.}\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{w}},p_{\mathbf{w}}\right).
\end{aligned}$$

Finally, if $\mathbb{E}_{w\sim p_{\mathbf{w}}}\mathcal{L}_\epsilon(\hat{p}_{\mathbf{w}},w)^2<\infty$, it follows that

$$\begin{aligned}
&\mathrm{Var}\{\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}},x_{1:T})\}=\mathrm{Var}\left\{\frac{1}{T}\sum_{k=1}^{T}\mathcal{L}_\epsilon\left(\hat{p}_{\mathbf{w}},w_k\right)\right\} \\
=&\frac{1}{T}\left\{\mathbb{E}_{w\sim p_{\mathbf{w}}}\mathcal{L}_\epsilon(\hat{p}_{\mathbf{w}},w)^2-\mathcal{L}_\epsilon(\hat{p}_{\mathbf{w}},p_{\mathbf{w}})^2\right\}.
\end{aligned}$$

Ensured by the Chebyshev inequality, the converging speed is $\mathcal{O}_p(\frac{1}{\sqrt{T}})$, i.e., $\forall\delta>0$, there is

$$\Pr\left\{|\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}},x_{1:T})-\bar{\mathcal{L}}_\epsilon(\hat{p}_{\mathbf{x}_{1:T}},p_{\mathbf{x}_{1:T}})|\geq\delta\right\}=\mathcal{O}(\frac{1}{\sqrt{T}}).$$