Predictability of Stochastic Dynamical System: A Probabilistic Perspective

Tao Xu and Jianping He

***Abstract*—In this paper, to characterize the predictability of discrete-time continuous-state stochastic dynamical system (SDS) and use predictability to explain the probability of making accurate predictions, we have proposed a new performance metric, i.e., predictability exponent. This metric quantifies the asymptotic exponential decaying rate of the probability that prediction errors never exceed a fixed ϵ. The novelties of this predictability metric lie in that it is proposed from an asymptotic probabilistic perspective, which i) minimizes the exponential decaying rate of prediction probability, ii) holds an approximated expression deeply related with differential entropy within an error of $O(\epsilon)$, and iii) possesses a fast converging speed from finite decaying rate of $O(e^{-K})$ in the sense of probability. Numerical examples are given to illustrate the efficiency of the obtained results.**

I. Introduction

A Stochastic Dynamical System is a dynamical system subjected to the effects of stochastic noise. Predictability of stochastic dynamical system aims to characterize to what extend the future states of an SDS can be accurately predicted. With rapidly blooming of various modern prediction techniques, such as deep learning and reinforcement learning, it brings emerging theoretical challenges to the predictability analysis of an SDS. On the one hand, different prediction methods have different performance limits, black-box structures of deep neuron networks make prediction performance analysis difficult. On the other hand, large scale stochastic dynamical system suffers from uncertainty and volatility, therefore prediction and validation tasks cost a lot in practice.

Entropy, as an information-theoretic quantity, is intuitively used to measure the predictability of outcomes in a stochastic event. A large amount of researches (e.g., [1]–[9]) have been devoted to characterizing predictability by entropy. [1]–[3] consider predictability as system's sensitivity to initial error and have found great applications in meteorology. [4]–[7] either directly use entropy or define entropy-based quantities to measure the predictability of various Markov decision processes (MDP), and use these entropy-based indexes to give performance analysis on specific prediction tasks. These analyses view the prediction performance from the perspective of information theory rather than from traditional methods, such as root-mean-square-error (RMSE), patter correlation (PC) and probability, etc. The authors in [8], [9] characterize the uncertainty of system trajectories by entropy, then use Fano's inequality to derive a probabilistic upper bound to estimate predictability. Due to the probabilistic description, these works circumvent the indirect information-theoretic explanation caused by entropy. However, most of these exciting conclusions about predictability are based on stochastic systems with finite states, such as finite states MDP, finite targets of motion, etc. Therefore, this characterization cannot describe stochastic dynamical systems with not necessarily finite states. To evaluate the predictive performance for SDS with infinite states, [10] specified some proper scoring rules to evaluate the asymptotic performance, in a sense that any sequential prediction cannot do better than the oracle based on the knowledge the data-generating process in advance.

Recently, [11], [12] have studied the one-step accurate prediction probability and designed unpredictable strategies accordingly, and their min-max strategies are robust when noises are not necessarily bounded. However, a specific definition on predictability considering prediction accuracy is still an open and challenging issue. In this article, we extend the characterization on predictability of stochastic dynamical systems from the finite state case to the the infinite state case. Moreover, our metric is defined directly from prediction performance, i.e., the asymptotic exponential decaying rate for the probability that prediction errors never exceed a fixed ϵ.

This new metric is partially motivated by the following reasons. First, entropy cannot be directly used in the case where stochastic systems have infinite states. A simple example is that Shannon entropy of a uniform distribution on N states will explode to infinity as N approaches to infinity, not to mention the case that state is of uncountable scale. Second, even in the finite case, entropy is indirect in quantifying prediction performance and gives an information-theoretic explanation on uncertainty. Therefore, we need a predictability metric that can directly quantify the prediction performance, such as the probability of making accurate predictions. Third, a trajectory of states generated by a stochastic dynamical system, instead of completely irregular, actually possess some asymptotic statistical rules. These rules usually do not rely on the realization of state trajectory, but rather, the stochastic system itself. Hence asymptotic characterization of stochastic dynamical system is reasonable.

The main contributions are summarized as follows.

- We propose a new metric to characterize the predictability of stochastic dynamical system with infinite states and

This work was supported in part by the NSF of China under Grant 61973218. The authors are with the Dept. of Automation, Shanghai Jiao Tong University, the Key Laboratory of System Control and Information Processing, Ministry of Education of China, and Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai, 200240, China. E-mail address: {zerken, jphe}@sjtu.edu.cn.

use predictability to explain the probability of making accurate predictions.

- We prove an approximated expression of the proposed metric, which is deeply related with differential entropy. Moreover, it shows that the approximation error is bounded of scale $O(\epsilon)$.
- It is found that the finite-steps decaying rate of accurate prediction probability converges to the asymptotic decaying rate with a quick speed of scale $O(e^{-K})$, and this convergence is ensured in the sense of probability.

The remainder of this paper is organized as follows. Section II gives basic preliminaries and describes the problem of interest. The discretization method and evaluation of predictability are presented in Sec. III. Simulation results are shown in Sec. IV, followed by the concluding remarks and further research issues in Sec. V.

Through out this paper, we always use capital letter to denote random variable while lower-case letter to denote numerical variable. For a sequence $\{a_i\}_{i=1}^n$, we use $a_{[s:t]}$ to denote its sub-sequence $\{a_s, a_{s+1}, \cdots, a_t\}$. Specifically, we use $x_{[1:K]}$ to denote a possible path of Φ's states.

II. Problem Formulation and Preliminaries

A. Problem Formulation

Consider a discrete stochastic dynamical system, denoted by Φ, as follows

$$\Phi : \begin{cases} X_{k+1} & = & f(X_k, u_k) + W_k \quad W_k \sim q \\ Y_k & = & g(X_k, u_k) \end{cases} \tag{1}$$

where $X_k \in \mathbb{R}^d$ is state vector, $W_k \in \mathbb{R}^d$ is noise, u_k is input, Y_k is output, q is a d dimensional probability density function of W_k, and f and g are two given functions.

Assumption 1. *Using historical data $Y_{[1:k]}$ can deduce the value of x_k, i.e., $x_{[1:k]}$ can be obtained when $Y_{[1:k]}$ are given.*

Suppose that a predictor will make a rolling prediction on system Φ. Rolling prediction from time step 1 to time step K means iterative one-step prediction at time step t, where $1 \leq t \leq K$. However, even though Φ is known to this predictor, error is unavoidable. This unavoidable error is an inherent property of a stochastic dynamical system, which is also an upper bound of rolling prediction performance. This concept is widely referred to as predictability, but badly in short of quantitative research.

To characterize the upper bound of rolling prediction performance, we use probability as prediction performance.

Definition 1 (rolling prediction probability)**.** *Suppose Φ has generated a series of states $x_{[1:K]}$ during the time interval $[1, K]$, then the probability that rolling prediction error at each step doesn't exceed ϵ is denoted as $\mathcal{P}_\epsilon(x_{[1:K]})$:*

$$\mathcal{P}_\epsilon(x_{[1:K]}) \triangleq \prod_{k=1}^{K} \mathbb{P}(\|X_{k|k-1} - x_k\|_\infty \leq \epsilon) \tag{2}$$

where $X_{k|k-1}$ is a prediction of X_k at step $k-1$.

Remark 1. *Although rolling prediction probability is self-evident and powerful to characterize the ability to make accurate rolling prediction on Φ, it fails to characterize system predictability because of its dependence on the value of $x_{[1:K]}$.*

Notice that $\mathcal{P}_\epsilon(x_{[1:K]})$ will converge to 0 when K approaches infinity regardless of $x_{[1:K]}$, a natural guess is whether the converging speed still doesn't depend on $x_{[1:K]}$. This hypothesis will be confirmed to be right in the following sections. Therefore, we characterize the system predictability by evaluating the asymptotic exponential decaying rate of $\mathcal{P}_\epsilon(x_{[1:K]})$, and give the following definition.

Definition 2 (predictability exponent)**.** *Given accuracy ϵ, the predictability exponent of Φ is denoted as $\mathcal{I}_\epsilon(\Phi)$:*

$$\mathcal{I}_\epsilon(\Phi) \triangleq \liminf_{K\to\infty} -\frac{1}{K} \ln \mathcal{P}_\epsilon(x_{[1:K]}) \tag{3}$$

Actually, $\mathcal{I}_\epsilon(\Phi)$ is actually the limit of the exponential decaying rate of rolling prediction probability, it not only has an elegant expression related with entropy, but also has a fast converging speed of $O(e^{-K})$. To fully taste this impressive interconnection between dynamical system and information theory, some preliminaries on entropy is necessary. To bridge the discrete world and continuous world and successfully evaluate $\mathcal{I}_\epsilon(\Phi)$, some preliminaries on discrete approximation theory is indispensable.

B. Preliminaries on Entropy

In this part, we define Shannon entropy, differential entropy and KL-divegence in reference to the classical textbook [13].

The entropy of a probability vector $\nu = [\nu_1, \cdots, \nu_n]^T$ is

$$H(\nu) \triangleq -\sum_{i=1}^{n} \nu_i \log \nu_i$$

The KL-divergence of a probability vector $\nu = [\nu_1, \cdots, \nu_n]^T$ with respect to another probability vector $\mu = [\mu_1, \cdots, \mu_n]^T$ is given by

$$D_{\mathcal{KL}}(\nu \mid \mu) \triangleq \sum_{i=1}^{n} \nu_i \log \frac{\nu_i}{\mu_i}$$

An important property of KL-divergence is that $D_{\mathcal{KL}}(\nu \mid \mu) \geq 0$ and the equality hold if and only if $\nu = \mu$.

The differential entropy $H(X)$ or $H(f)$ of a continuous random variable X with density $f(x)$ is defined as

$$H(X) = -\int_S f(x) \log f(x) dx$$

where S is the support set of the random variable.

C. Preliminaries on Discrete Approximation

A finite partition of a space Ω is defined as

$$\Sigma \triangleq \left\{ A_1, \cdots, A_{|\Sigma|} \mid \Omega = \bigcup_{i=1}^{|\Sigma|} A_i, A_i \cap A_j = \emptyset \; \forall i \neq j \right\}.$$

where we use $|\Sigma|$ to denote the number of elements in Σ. There are many methods to partition Ω, for example, uniform grid partition of a square into $k \times k$ equal-sized squares is the most familiar one.

Generally we treat Σ as a set composed of small regions, sometimes it's more convenient to treat it as a label function, which assigns each point in Σ a label. This is because partition Σ naturally induces a map $\Theta : \Omega \mapsto \{1, 2, \cdots, |\Sigma|\}$, where

$$\Theta(x) = \sum_{i=1}^{|\Sigma|} \mathbb{I}_{A_i}(x) i \tag{4}$$

Since each point x in Ω uniquely belongs to a region of Σ, and suppose this region is A_i, then the label function Θ assigns it value i.

A discretization of Ω naturally induces a step-wise function q_Σ to approximate q:

$$q_\Sigma(x) = \sum_{i=1}^{|\Sigma|} \mathbb{I}_{A_i}(x) \int_{A_i} q(u)du \tag{5}$$

where q_Σ assigns each point x in Ω with the value of $\int_{A_i} q(u)du$ if $\Theta(x) = i$. The finer the partition is, the better the approximation will be.

III. Main Results

First, we apply discretization methods to rolling prediction probability, which leads to a discrete version of rolling prediction probability. Second, we evaluate discrete rolling prediction probability and use it to approximate the continuous one. Third, an asymptotic analysis on rolling prediction probability is made to derive expression of predictability exponent. Finally, a large deviation result is given on predictability exponent to show that exponential decaying rate of rolling prediction probability converges to predictability exponent with the speed of $O(e^{-K})$.

A. Discretization

Given a partition Σ of $\Re^d$ and a label function $\Theta(x)$ induced from it, we can propose a discrete version of rolling prediction.

Definition 3 (discrete rolling prediction)**.** *During time* 1 *to time* K*,* Φ *has generated a trajectory of states* $x_{[1:K]}$*, where* $x_i \sim X_i \; \forall 1 \le i \le K$*. Then,*

$$\mathcal{P}_\Sigma(w_{[1:K]}) \triangleq \prod_{k=1}^{K} \mathbb{P}[\Theta(W_k) = \Theta(w_k)] \tag{6}$$

The only difference between discrete rolling prediction and continuous version is the accuracy requirement: the continuous version requires the distance between predicted state and real state less than a given tolerance ϵ, while the discrete version requires predicted state and real state to be located in the same region defined by partition Σ. However, this small difference has not only transformed a continuous problem to a discrete one, but also made the evaluation of rolling prediction probability tractable.

From the perspective of empirical distribution, the probability of discrete rolling prediction can be evaluated neatly as the combination of Shannon entropy and KL-divergence.

Theorem 1 (evaluation of $\mathcal{P}_\Sigma(w_{[1:K]})$)**.** *Consider the discrete rolling prediction, we have*

$$\mathcal{P}_\Sigma(w_{[1:K]}) = \exp\{-K[H(q_{\Sigma,K}) + D_{KL}(q_{\Sigma,K}||q_\Sigma)]\} \tag{7}$$

where $q_{\Sigma,K}$ *is the empirical distribution of* $w_{[1:K]}$*:*

$$q_{\Sigma,K}(u) = \frac{1}{K} \sum_{j=1}^{|\Sigma|} \mathbb{I}_{A_j}(u) \sum_{i=1}^{K} \mathbb{I}_{A_j}(w_i)$$

Proof. Suppose the discrete distribution on $\{A_i\}_{i=1}^{|\Sigma|}$ induced by the step-wise function q_Σ is:

$$p(i) = q_\Sigma(x_0) \text{ where } x_0 \in A_i$$

And the discrete distribution on $\{A_i\}_{i=1}^{|\Sigma|}$ induced by the step-wise function $q_{\Sigma,K}$ is:

$$p_K(i) = q_{\Sigma,K}(x_0) \text{ where } x_0 \in A_i$$

According to the definition of discrete rolling prediction probability, there is

$$\begin{aligned}
\mathcal{P}_\Sigma(w_{[1:K]}) &= \prod_{k=1}^{K} \mathbb{P}[\Theta(W_k) = \Theta(w_k)] \\
&= \prod_{k=1}^{K} q_\Sigma(w_k) \\
&= \prod_{n=1}^{|\Sigma|} p(n)^{N_n} \quad \text{where } N_n = \sum_{k=1}^{K} \mathbb{I}_{A_n}(w_k)
\end{aligned}$$

The point to be emphasized is that, the last equality is doing a classification of series $\{q_\Sigma(w_k)\}_{k=1}^K$ based their values, and gather the same values together. For example, after enumeration of this series we found that N_n values are exactly $p(n)$, so their product can be gathered simply as $p(n)^{N_n}$.

Another observation is, the distribution induced by step-wise function $q_{\Sigma,K}$ has the property that

$$\begin{aligned}
p_K(n) &= \frac{1}{K} \sum_{j=1}^{|\Sigma|} \mathbb{I}_{A_j}(u) \sum_{i=1}^{K} \mathbb{I}_{A_j}(w_i) \quad \text{where } u \in A_n \\
&= \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}_{A_n}(w_k) \\
&= \frac{N_n}{K}
\end{aligned}$$

Therefore, we continue to find that

$$\begin{aligned}\mathcal{P}_\Sigma(w_{[1:K]}) &= \prod_{n=1}^{|\Sigma|} p(n)^{N_n} \\ &= \exp\{K \sum_{n=1}^{|\Sigma|} p_K(n)\ln(p(n))\} \\ &= \exp\{-K[-\sum_{n=1}^{|\Sigma|} p_K(n)\ln(p_K(n)) \\ &\quad + \sum_{n=1}^{|\Sigma|} p_K(n)\ln(\frac{p_K(n))}{p(n)}]\} \\ &= \exp\{-K[H(q_{\Sigma,K}) + D_{KL}(q_{\Sigma,K}||q_\Sigma)]\}\end{aligned}$$

the proof is completed. □

The introduction of empirical distribution $q_{\Sigma,K}$ gives $\mathcal{P}_\Sigma(w_{[1:K]})$ a beautiful expression related with information theory. This theorem is quite instructive in two perspectives: on the one hand, discrete rolling prediction is decided by the uncertainty of $q_{\Sigma,K}$ and the distance between $q_{\Sigma,K}$ and q_Σ. While q_Σ is fixed, only $q_{\Sigma,K}$ will change as K approaches infinity. On the other hand, discrete rolling prediction probability has an exponential decaying trend against time, and the decaying rate at each time step is explicitly expressed.

The from of discrete rolling prediction probability drives us to find the relationship between it and continuous version. We have discovered an inequality relationship between $\mathcal{P}_\Sigma(w_{[1:K]})$ and $\mathcal{P}_\epsilon(w_{[1:K]})$:

Lemma 1 (sandwich lemma)**.** *$\mathcal{P}_\epsilon(w_{[1:K]})$ is bounded by $\mathcal{P}_\Sigma(w_{[1:K]})$ in the following sense*

$$\max_{\{\Sigma|\operatorname{diam}(\Sigma)\le\epsilon\}} \mathcal{P}_\Sigma(w_{[1:K]}) \le \mathcal{P}_\epsilon(w_{[1:K]}) \le \max_{\Sigma} \mathcal{P}_\Sigma(w_{[1:K]}) \tag{8}$$

where $\operatorname{diam}(\Sigma) \triangleq \max_{A\in\Sigma}\max_{x,y\in A} \|x-y\|_\infty$.

Proof. The proof is omitted due to the limit of space. See [14] for details. □

Motivated by Lemma 1, we confirm the existence of a special partition to evaluate $\mathcal{P}_\epsilon(w_{[1:K]})$ in the following theorem.

Theorem 2 (existence)**.** *For $\forall\ \epsilon > 0$, there exist a partition $\Sigma^\star(K)$, such that*

$$\mathcal{P}_\epsilon(w_{[1:K]}) = \mathcal{P}_{\Sigma^\star(K)}(w_{[1:K]}) \tag{9}$$

Proof. Suppose space Y is composed of all partitions of $\mathbb{R}^d$, to make Y a metric space , we use the partition distance metric $D(\cdot,\cdot)$ which has been defined in [15] [16]:

$$D(P,Q) = \min\left\{\mathcal{M}(A^c) : \emptyset \subset A \subseteq \mathbb{R}^d, P^A = Q^A\right\}.$$

where P^A is a partition of set A induced by P, which means if $P = \bigcup_{i=1}^m B_i$ then $P^A = \bigcup_{i=1}^m (B_i \bigcap A)$.

Intuitively, partition distance $D(P,Q)$, between P and Q is the minimum measure of set that must be deleted from $\mathbb{R}^d$, so that the two induced partitions (P and Q restricted to the remaining elements) are identical. It's trivial to verify that $D(\cdot,\cdot)$ satisfies all the three requirement of a distance metric.

Consider this functional operator:

$$\begin{aligned}\mathcal{F} :& Y \to \Re \\ & \Sigma \mapsto \mathcal{P}_\Sigma(x_{[1:K]})\end{aligned}$$

According to theorem 1, there is

$$\mathcal{F}(\Sigma) = \exp\{-N[H(q_{\Sigma,K}) + D_{KL}(q_{\Sigma,K}||q_\Sigma)]\}$$

where $q_{\Sigma,K}$ is the empirical distribution of $w_{[1:K]}$:

$$q_{\Sigma,K}(u) = \frac{1}{K}\sum_{j=1}^{|\Sigma|} \mathbb{I}_{A_j}(u) \sum_{i=1}^{K} \mathbb{I}_{A_j}(w_i)$$

and q_K is the discretized step-wise function of q:

$$q_\Sigma(x) = \sum_{i=1}^{n} \mathbb{I}_{A_i}(x) \int_{A_i} q(u)du$$

Next, we prove that $\mathcal{F}$ is continuous at any point in Y. Continuity means that, for any $\Sigma = \bigcup_{i=1}^{|\Sigma|} A_i \in Y$ and any converging partition series $\{\Sigma^n = \bigcup_{i=1}^{|\Sigma^n|} A_i^n\}_{n=1}^\infty$, if $\lim_{n\to\infty} D(\Sigma_n', \Sigma) = 0$, we have

$$\lim_{n\to\infty} \mathcal{F}(\Sigma^n) = \mathcal{F}(\Sigma)$$

According to the converging definition of partition series, $\forall\gamma > 0$, $\exists N \in \mathbb{N}$ such that $\forall n > N$ there is $D(\Sigma^n, \Sigma) < \gamma$. Consider $\forall x \in \mathbb{R}^d$, and suppose $x \in A_t$. When $n > N$ we have $A_t^n \in \Sigma^n$ such that $\mathcal{M}(A_t \Delta A_t^n) \le \gamma$, here $\mathcal{M}(\cdot)$ means the Lebesgue measure.

Firstly, we prove the continuity of q_Σ:

$$\begin{aligned}&|q_\Sigma(x) - q_{\Sigma^n}(x)| \\ =&|\sum_{i=1}^{n} \mathbb{I}_{A_i}(x)\int_{A_i} q(u)du - \sum_{i=1}^{n} \mathbb{I}_{A_i^n}(x)\int_{A_i^n} q(u)du| \\ =&|\int_{A_t} q(u)du - \int_{A_t^n} q(u)du| \\ \le& \int_{A_t\Delta A_t^n} q(u)du\end{aligned}$$

Since $\lim_{n\to\infty} \mathcal{M}(A_t \Delta A_t^n) = 0$, then we have

$$\lim_{n\to\infty} |q_\Sigma(x) - q_{\Sigma^n}(x)| = 0$$

Secondly, we prove the continuity of $q_{\Sigma,K}$:

$$\begin{aligned}&|q_{\Sigma,K}(x) - q_{\Sigma^n,K}(x)| \\ =&|\frac{1}{K}\sum_{j=1}^{|\Sigma|} \mathbb{I}_{A_j}(x)\sum_{i=1}^{K} \mathbb{I}_{A_j}(w_i) - \frac{1}{K}\sum_{j=1}^{|\Sigma^n|} \mathbb{I}_{A_j}(x)\sum_{i=1}^{K} \mathbb{I}_{A_j}(w_i)| \\ =&\frac{1}{K}|\sum_{i=1}^{K}(\mathbb{I}_{A_t}(w_i) - \mathbb{I}_{A_t^n}(w_i))|\end{aligned}$$

Since $\lim_{n\to\infty} \mathcal{M}(A_t \Delta A_t^n) = 0$, we have

$$\lim_{n\to\infty} \sum_{i=1}^{K} (\mathbb{I}_{A_t}(w_i) - \mathbb{I}_{A_t^n}(w_i)) = 0$$

Finally, the above two results lead to the continuity of $\mathcal{F}(\Sigma)$ because $H(\cdot)$ and $D_{\mathcal{KL}(\cdot)}$ are continuous functions. The bounded inequality

$$\max_{\{\Sigma|\,\mathrm{diam}(\Sigma)\le\epsilon\}} \mathcal{P}_\Sigma(w_{[1:K]}) \le \mathcal{P}_\epsilon(w_{[1:K]}) \le \max_{\Sigma} \mathcal{P}_\Sigma(w_{[1:K]})$$

suggests the existence of $\Sigma_a, \Sigma_b \in Y$ such that

$$\mathcal{P}_{\Sigma_a}(w_{[1:K]}) \le \mathcal{P}_\Sigma(w_{[1:K]}) \le \mathcal{P}_{\Sigma_b}(w_{[1:K]})$$

A general intermediate value theorem [17] admits the existence of $\Sigma^\star(K) \in Y$ such that $\mathcal{P}_{\Sigma^\star(K)}(w_{[1:K]}) = \mathcal{P}_\epsilon(w_{[1:K]})$, and the proof is completed. □

The significance of knowing the existence of $\Sigma^\star(K)$ is that the problem of continuous rolling prediction probability has been completely transformed to the evaluation of discrete version. However, this transformation is only a formal method, and existence theorem itself doesn't ensure a practical algorithm to design $\Sigma^\star(K)$.

B. Evaluation

When K is approaching infinity, the law of large number ensures that empirical distribution q_K shall have no difference with $q_{\Sigma^\star}$, therefore the rolling prediction probability will only depend on $q_{\Sigma^\star}$, and predictability exponent can be derived by the entropy of $q_{\Sigma^\star}$.

Theorem 3 (predictability exponent)**.** *Let $\Sigma^\star(K)$ be the special partition ensured by theorem 2, then*

$$\mathcal{I}_\epsilon(\Phi) = \lim_{K\to\infty} H(q_{\Sigma^\star(K)}) \tag{10}$$

Let $L = \max_x -\frac{\ln(q_{\Sigma^\star(K)}(x))}{q_{\Sigma^\star(K)}(x)}$, the converging speed toward $\mathcal{I}_\epsilon(\Phi;\Phi)$ is $O(e^{-K})$ in the sense of probability:

$$\mathbb{P}(|-\frac{1}{K}\ln \mathcal{P}_{\Sigma^\star(K)}(w_{[1:K]}) - \mathcal{I}_\epsilon(\Phi)| \ge t) \le 2e^{-2\frac{Kt^2}{L^2}} \tag{11}$$

Proof. Suppose the discrete distribution on $\{A_i\}_{i=1}^{|\Sigma^\star(K)|}$ induced by the step-wise function q_Σ is:

$$p(i) = q_{\Sigma^\star(K)}(x_0) \text{ where } x_0 \in A_i$$

And the discrete distribution on $\{A_i\}_{i=1}^{|\Sigma^\star(K)|}$ induced by the step-wise function $q_{\Sigma^\star,K}$ is:

$$p_K(i) = q_{\Sigma^\star(K),K}(x_0) \text{ where } x_0 \in A_i$$

Then

$$\begin{aligned}
\mathcal{I}_\epsilon(\Phi) &= \limsup_{K\to\infty} -\frac{1}{K}\ln \mathcal{P}_\epsilon(x_{[1:K]}) \\
&= \limsup_{K\to\infty} -\frac{1}{K}\ln \mathcal{P}_{\Sigma^\star(K)}(x_{[1:K]}) \\
&= \limsup_{K\to\infty} -\sum_{n=1}^{|\Sigma^\star(K)|} p_K(n)\ln(p(n))
\end{aligned}$$

According to strong law of large number, we have $\lim_{K\to\infty} p_K(n) \overset{\text{a.s.}}{=} p(n)$, therefore

$$\begin{aligned}
\mathcal{I}_\epsilon(\Phi) &= \limsup_{K\to\infty} -\sum_{n=1}^{|\Sigma^\star(K)|} p_K(n)\ln(p(n)) \\
&= \lim_{K\to\infty} -\sum_{n=1}^{|\Sigma^\star(K)|} p_K(n)\ln(p(n)) \\
&= -\sum_{n=1}^{|\Sigma^\star(K)|} p(n)\ln(p(n)) \\
&= H(q_{\Sigma^\star(K)})
\end{aligned}$$

Now consider

$$\begin{aligned}
&\mathbb{P}(|-\frac{1}{K}\ln \mathcal{P}_{\Sigma^\star(K)}(w_{[1:K]}) - \mathcal{I}_\epsilon(\Phi)| \ge t) \\
=&\mathbb{P}(|H(q_{\Sigma^\star(K),K}) + D_{KL}(q_{\Sigma^\star(K),K}||q_\Sigma) - H(q_{\Sigma^\star(K)})| \ge t) \\
=&\mathbb{P}(|\sum_{n=1}^{|\Sigma^\star(K)|} p(n)(p(n)-p_K(n))\frac{\ln(p(n))}{p(n)}| \ge t) \\
\le&\mathbb{P}(\sum_{n=1}^{|\Sigma^\star(K)|} p(n)|p(n)-p_K(n)| \ge \frac{t}{L})
\end{aligned}$$

According to Hoeffding inequality, $\forall n \in [1, |\Sigma^\star(K)|]$ we have

$$\mathbb{P}(|p(n)-p_K(n)| \ge \frac{t}{L}) \le 2e^{-2\frac{Kt^2}{L^2}}$$

Together with the fact that $\sum_{n=1}^{|\Sigma^\star(K)|} p(n)|p(n)-p_K(n)| \le \max_n\{|p(n)-p_K(n)|\}$, it follows that

$$\mathbb{P}(\sum_{n=1}^{|\Sigma^\star(K)|} p(n)|p(n)-p_K(n)| \ge \frac{t}{L}) \le 2e^{-2\frac{Kt^2}{L^2}}$$

The proof is completed. □

When $\Sigma^\star$ is given, an explicit expression of $\mathcal{I}_\epsilon(\Phi)$ is derived. However, we usually only know the existence of $\Sigma^\star$ but not the accurate form of it. Thus, an approximation of $\mathcal{I}_\epsilon(\Phi)$ is needed.

Theorem 4 (approximation of $\mathcal{I}_\epsilon(\Phi)$)**.** *For a SDS, Φ, with known structure, $\mathcal{I}_\epsilon(\Phi)$ can be approximated as $H(q) - d\log(2\epsilon)$ with error guaranteed by following inequality*

$$|\mathcal{I}_\epsilon(\Phi) - H(q) + d\log(2\epsilon)| \le \delta_{2\epsilon}(q) \tag{12}$$

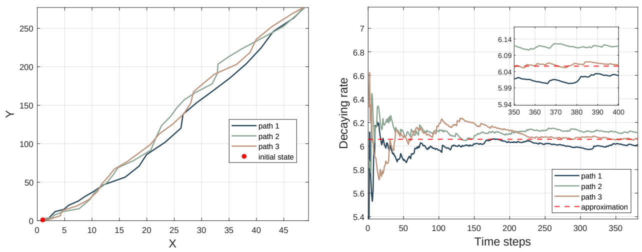
where $H(q)$ is the differential entropy of q, d is the dimension of state, and

$$\delta_\epsilon(q) \triangleq \max_{\|x-y\|_\infty\le\epsilon} |\ln(q(x)) - \ln(q(y))|$$

Proof. The proof is omitted due to the limit of space. See [14] for details.

□

This theorem gives an accurate approximation to $\mathcal{I}_\epsilon(\Phi)$, and the error is controlled by $\delta_{2\epsilon}(q)$. It should be noted that $\delta_\epsilon(q)$



(a) Trajectories (b) Asymptotic decaying rate

Fig. 1. Asymptotic decaying rate simulations for three different trajectories which have the same settings on model parameter and initial state.

reflects to what extend q can vibrate in a local region with diameter less than ϵ. Take Gaussian distribution $\bar{q} = \mathcal{N}(\mu, \sigma^2)$ for example, there is $\delta_\epsilon(\bar{q}) \leq \frac{\epsilon^2}{2\sigma^2}$. Furthermore, for any distribution q with bounded support, the probability distribution must be uniformly continuous, therefore $\delta_\epsilon(q) = O(\epsilon)$. As a result, this approximation leads to an error approximately $O(\epsilon)$, and that's quite satisfying.

IV. Simulation

In this section, we simulate rolling prediction on a randomly generated linear stochastic dynamical system driven by Gaussian noises, then verify that asymptotic prediction probability decaying rate can be approximated by $H(q) - d\log(2\epsilon)$ with error of scale $O(\epsilon)$.

A. Simulation Setup

We consider a linear SDS as follows.

$$\Phi : \begin{cases} X_{k+1} &= AX_k + W_k \quad W_k \sim \mathcal{N}(\mu, \Sigma) \\ Y_k &= X_k \end{cases} \tag{13}$$

where $X_k \in \mathbb{R}^2$. We randomly generate A,μ, Σ, and normalize Σ to have spectral radius 1 without loss of generality. Setting the accuracy tolerance ϵ to be 0.1 and the time step to be 400, we generate several state paths of Φ starting from a random initial state. Then we calculate the decaying rates of $\mathcal{P}_\epsilon(x_{[1:K]})$ for each state path.

B. Results and Analysis

Driven by the stochastic noises, the same initial state can evolve into totally different trajectories even they have the same SDS model. Fig. 1(a) has simulated three possible state trajectories of Φ, and Fig. $1(b)$ presents the relationship between their exponential decaying rate and time steps.

First, the approximation performance of $H(q) - d\log(2\epsilon)$ is verified to be correct. As $1(b)$ shows, all the trajectories of decaying rate converge to the same red dotted line (which is the theoretical asymptotic decaying rate). Moreover, if we zoom in the graph, it is clear that the approximation error is bounded by 0.05, which is of the same scale as ϵ. Therefore, Theorem 4 indeed gives an both elegant and accurate approximation to the predictability exponent $\mathcal{I}_\epsilon(\Phi)$.

Second, the converging rate is quite fast. Fig. $1(b)$ shows that all trajectories approaches red dotted line in less than 50 time steps, this quick convergence is ensured by the concentration inequality in Theorem 3 in the sense of probability.

Third, the fact that different trajectories generated from the same SDS hold the same asymptotic decaying rate has reconfirmed the advantage of predictability exponent. This metric is defined directly from probabilistic prediction performance and does not depend on specific state trajectory generated from a SDS, therefore it views different trajectories in Fig. 1(a) as having the same predictability.

V. Conclusion

In this paper, we studied predictability of a SDS by introducing a new probabilistic performance metric, i.e., the exponential decaying rate of the probability that the prediction error never exceed a given ϵ. The proposed metric is defined directly from the prediction performance rather than intuitive entropy-based characterization, therefore able to directly quantify to what extend a SDS can be accurately predicted. Moreover, an elegant approximation of this metric is available, which is deeply related with the differential entropy of SDS. This impressive fact provides a better understanding of the relationship between predictability and information theory.

The future work includes two interesting directions. One is to extend the existing results into the case where model is unknown to the predictor, while the other one is to apply the predictability metric to the design of unpredictable stochastic dynamical system.

References

[1] R. Kleeman, "Information Theory and Dynamical System Predictability," *Entropy*, vol. 13, no. 3, pp. 612–649, Mar. 2011.

[2] D. Giannakis and A. J. Majda, "Quantifying the Predictive Skill in Long-Range Forecasting. Part I: Coarse-Grained Predictions in a Simple Ocean Model," *Journal of Climate*, vol. 25, no. 6, pp. 1793–1813, Mar. 2012, publisher: American Meteorological Society Section: Journal of Climate.

[3] N. Chen and A. J. Majda, "Conditional Gaussian Systems for Multiscale Nonlinear Stochastic Systems: Prediction, State Estimation and Uncertainty Quantification," *Entropy*, vol. 20, no. 7, p. 509, Jul. 2018, number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[4] 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, no. 5, pp. 384–399, Sep. 2014.

[5] Y. Savas, M. Ornik, M. Cubuktepe, M. O. Karabag, and U. Topcu, "Entropy Maximization for Markov Decision Processes Under Temporal Logic Constraints," *IEEE Transactions on Automatic Control*, vol. 65, no. 4, pp. 1552–1567, Apr. 2020, conference Name: IEEE Transactions on Automatic Control.

[6] M. Hibbard, Y. Savas, B. Wu, T. Tanaka, and U. Topcu, "Unpredictable Planning Under Partial Observability," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, Dec. 2019, pp. 2271–2277, iSSN: 2576-2370.

[7] M. Ornik, "Measuring Target Predictability for Optimal Environment Design," in *2020 59th IEEE Conference on Decision and Control (CDC)*, Dec. 2020, pp. 5023–5028, iSSN: 2576-2370.

[8] 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, publisher: American Association for the Advancement of Science.

[9] 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, no. 6, pp. 2671–2682, Dec. 2014, conference Name: IEEE Transactions on Intelligent Transportation Systems.

[10] J. Ding, J. Zhou, and V. Tarokh, "Asymptotically Optimal Prediction for Time-Varying Data Generating Processes," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3034–3067, May 2019, conference Name: IEEE Transactions on Information Theory.

[11] J. Li, J. He, Y. Li, and X. Guan, "Unpredictable Trajectory Design for Mobile Agents," in *2020 American Control Conference (ACC)*, Jul. 2020, pp. 1471–1476, iSSN: 2378-5861.

[12] J. He, L. Cai, C. Zhao, P. Cheng, and X. Guan, "Privacy-Preserving Average Consensus: Privacy Analysis and Algorithm Design," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 1, pp. 127–138, Mar. 2019, conference Name: IEEE Transactions on Signal and Information Processing over Networks.

[13] M. Thomas and A. T. Joy, *Elements of information theory*. Wiley-Interscience, 2006.

[14] T. Xu, J. He, and Y. Li, "Predictability exponent of stochastic dynamical systems," *arXiv preprint arXiv:2205.06006*, 2022.

[15] D. Gusfield, "Partition-distance: A problem and class of perfect graphs arising in clustering," *Information Processing Letters*, vol. 82, no. 3, pp. 159–164, 2002.

[16] G. Rossi, "Partition distances," *arXiv preprint arXiv:1106.4579*, 2011.

[17] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.