The Age of Correlated Features in Supervised Learning based Forecasting

Md Kamran Chowdhury Shisher*, Heyang Qin†, Lei Yang†, Feng Yan†, and Yin Sun*
*Department of ECE, Auburn University, AL, USA
†Department of CSE, University of Nevada, Reno, NV, USA

***Abstract*—In this paper, we analyze the impact of information freshness on supervised learning based forecasting. In these applications, a neural network is trained to predict a time-varying target (e.g., solar power), based on multiple correlated features (e.g., temperature, humidity, and cloud coverage). The features are collected from different data sources and are subject to heterogeneous and time-varying ages. By using an information-theoretic approach, we prove that the minimum training loss is a function of the ages of the features, where the function is not always monotonic. However, if the empirical distribution of the training data is close to the distribution of a Markov chain, then the training loss is approximately a non-decreasing age function. Both the training loss and testing loss depict similar growth patterns as the age increases. An experiment on solar power prediction is conducted to validate our theory. Our theoretical and experimental results suggest that it is beneficial to (i) combine the training data with different age values into a large training dataset and jointly train the forecasting decisions for these age values, and (ii) feed the age value as a part of the input feature to the neural network.**

I. Introduction

Recently, the proliferation of artificial intelligence and cyber physical systems has engendered a significant growth in machine learning techniques for time-series forecasting applications, such as autonomous driving [1], [2], energy forecasting [3]–[5], and traffic prediction [6]. In these applications, a predictor (e.g., a neural network) is used to infer the status of a time-varying target (e.g., solar power) based on several features (e.g., temperature, humidity, and cloud coverage). Fresh features are desired, because they could potentially lead to a better forecasting performance. For example, recent studies on pedestrian intent prediction [1] and autonomous driving [2] showed that prediction accuracy can be greatly improved if fresher data is used. Similarly, it was found in [4], [5] that the performance of energy forecasting degrades as the observed feature becomes stale. This phenomenon has also been observed in other applications of time-series forecasting, such as traffic control [7] and financial trading [8].

Age of information (AoI), or simply *age*, is a performance metric that measures the freshness of the information that a receiver has about the status of a remote source [9]. Recent research efforts on AoI have been focused on analyzing and optimizing the AoI in, e.g., communication networks [10]–[16], remote estimation [17], [18], and control systems [19], [20]. However, the impact of AoI on supervised learning based forecasting has not been well-understood, despite its significance in a broad range of applications. Recently, an *age of features* concept was studied in [21], where a stream of features are collected progressively from a single data source and, at any time, only the freshest feature is used for prediction. Meanwhile, in many applications, the forecasting decision is made by jointly using multiple features that are collected in real-time from different data sources (e.g., temperature readings from thermometers and wind speed measured from anemometers). These features are of diverse measurement frequency, data formats, and may be received through separated communication channels. Hence, their AoI values are different. This motivated us to analyze the performance of supervised learning algorithms for time-series forecasting, where the features are subject to heterogeneous and time-varying ages. The main contributions of this paper are summarized as follows:

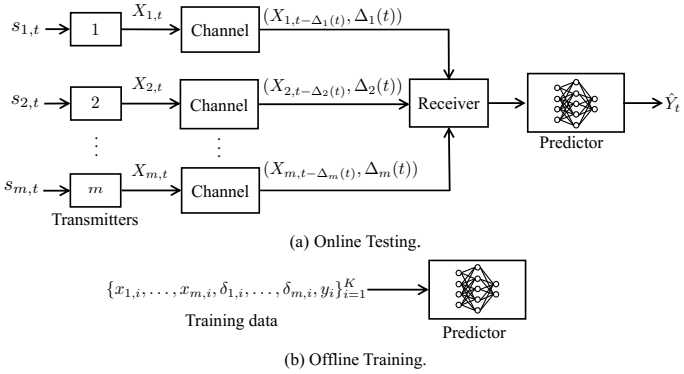

Fig. 1: System Model.

- We present an information theoretic approach to interpret the influence of age on supervised learning based forecasting. Our analysis shows that the minimum training loss is a multi-dimensional function of the age vector of the features, but the function is not necessarily monotonic. This is a key difference from the non-decreasing age metrics considered in earlier work, e.g., [12], [19], [22] and the references therein.
- Moreover, by using a local information geometric analysis,

This work was supported in part by NSF grants CCF-1813050, IIS-1838024, EEC-1801727, CNS-1950485, CCF-1756013, and ONR grant N00014-17-1-2417.

we prove that if the empirical distribution of training data samples can be accurately approximated as a Markov chain, then the minimum training loss is close to a non-decreasing function of the age. The testing loss performance is analyzed in a similar way.
- We compare the performance of several training approaches and find that it is better to (i) combine the training data with different age values into a large training dataset and jointly train the forecasting decisions for these age values, and (ii) add the age value as a part of the feature. This training approach has a lower computational complexity than the separated training approach used in [23], where the forecasting decision for each age value is trained individually. Experimental results on solar power prediction are provided to validate our findings.

II. System Model

Consider the learning-based time-series forecasting system in Fig. 1, which consists of m transmitters and one receiver. The system time is slotted. Each transmitter l takes measurements from a discrete-time signal process $s_{l,t}$. The processes $s_{1,t}, \ldots, s_{m,t}$ contain useful information for inferring the behavior of a target process Y_t. Transmitter l progressively generates a sequence of features $X_{l,t}$ from the process $s_{l,t}$. Each feature $X_{l,t} = f(s_{l,t-\tau}, s_{l,t-1-\tau}, \ldots, s_{l,t-b+1-\tau})$ is a function of a finite-length time sequence from the process $s_{l,t}$, where b is the length of the sequence and τ is the processing time needed for creating the feature. The processes $s_{1,t}, \ldots, s_{m,t}$ may be correlated with each other, and so are the features $X_{1,t}, \ldots, X_{m,t}$. The features are sent from the transmitters to the receiver through one or multiple channels. The receiver feeds the features to a predictor (e.g., a trained neural network), which infers the current target value Y_t.

Due to transmission errors and random transmission time, freshly generated features may not be immediately delivered to the receiver. Let $G_{l,i}$ and $D_{l,i}$ be the creation time and delivered time of the i-th feature of the process $s_{l,t}$, respectively, such that $G_{l,i} \leq G_{l,i+1}$ and $G_{l,i} \leq D_{l,i}$. Then, $U_l(t) = \max\{G_{l,i} : D_{l,i} \leq t\}$ is the creation time of the freshest feature that was generated from process $s_{l,t}$ and has been delivered to the receiver by time t. At time t, the receiver uses the m freshest delivered features $(X_{1,U_1(t)}, \ldots, X_{m,U_m(t)})$, each from a transmitter, to predict the current target Y_t. The age of the features generated from process $s_{l,t}$ is defined as

$$\Delta_l(t) = t - U_l(t) = t - \max\{G_{l,i} : D_{l,i} \leq t\}, \qquad (1)$$

which is the time elapsed since the creation time $U_l(t)$ of the freshest delivered feature $X_{l,U_l(t)}$ up to the current time t. If $\Delta_l(t)$ is small, then there exists a fresh delivered feature that was created recently from process $s_{l,t}$. The evolution of $\Delta_l(t)$ over time is illustrated in Fig. 2.

The predictor is trained by using an Empirical Risk Minimization (ERM) based supervised learning algorithm, such as logistic regression and linear regression. A supervised learning algorithm consists of two phases: *offline training* and *online testing*. In offline training phase, the predictor is trained by minimizing an expected loss function under the empirical distribution of a training dataset. Each entry $(x_{1,i}, \ldots, x_{m,i}, \delta_{1,i}, \ldots, \delta_{m,i}, y_i)$ of the training dataset contains m features $(x_{1,i}, \ldots, x_{m,i})$, the age values of the features $(\delta_{1,i}, \ldots, \delta_{m,i})$, and the target y_i. In Section IV, we will see that it is important to add the age values into training data. In online testing, the trained predictor is used to predict the target in real-time, as explained above.

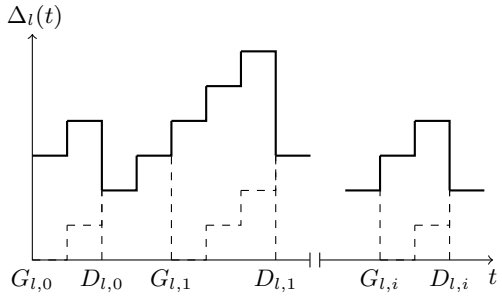

Fig. 2: Evolution of the age $\Delta_l(t)$ in discrete-time.

The goal of this paper is to interpret how the age processes $\Delta^m(t) = (\Delta_1(t), \ldots, \Delta_m(t))$ of the features affect the performance of time-series forecasting.

III. Performance of Supervised Learning From An Information Theoretic Perspective

In this section, we introduce several information theoretic measures that characterize the fundamental limits for the training and testing performance of supervised learning. Based on these information theoretic measures, the influence of information freshness on supervised learning based forecasting will be analyzed subsequently in Section IV.

Let $X^m = (X_1, \ldots, X_m)$ represent a vector of m random features, which takes value $x^m = (x_1, \ldots, x_m)$ from the finite space $\mathcal{X}^m = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$. As the standard approach for supervised learning, ERM is a stochastic decision problem $(\mathcal{X}^m, \mathcal{Y}, \mathcal{A}, L)$, where a decision-maker predicts $Y \in \mathcal{Y}$ by taking an action $a = \psi(X^m) \in \mathcal{A}$ based on features $X^m \in \mathcal{X}^m$. The performance of ERM is measured by a loss function $L : \mathcal{Y} \times \mathcal{A} \mapsto \mathbb{R}$, where $L(y, a)$ is the incurred loss if action a is chosen when $Y = y$. For example, L is a logarithmic function $L_{\log}(y, P_Y) = -\log P_Y(y)$ in logistic regression and a quadratic function $L_2(y, \hat{y}) = (y - \hat{y})^2$ in linear regression. Let $P_{X^m,Y}$ and $P_{\tilde{X}^m,\tilde{Y}}$, respectively, denote the empirical distributions of the training data and testing data, where $\tilde{X}^m$ and $\tilde{Y}$ are random variables with a joint distribution $P_{\tilde{X}^m,\tilde{Y}}$. We restrict our analysis in which the marginals P_{X^m} and $P_{\tilde{X}^m}$ are strictly positive.

A. Minimum Training Loss

The objective of training in ERM-based supervised learning is to solve the following problem:

$$H_L(Y|X^m) = \min_{\psi \in \Psi} \mathbb{E}_{X^m,Y \sim P_{X^m,Y}}[L(Y, \psi(X^m))], \qquad (2)$$

where Ψ is the set of allowed decision functions and $H_L(Y|X^m)$ is the *minimum training loss*. We consider a case

that Ψ contains *all* functions from $\mathcal{X}^m$ to $\mathcal{A}$. Such a choice of Ψ is of particular interest for two reasons: (i) Since Ψ is quite large, $H_L(Y|X^m)$ provides a fundamental lower bound of the training loss for any ERM based learning algorithm. (ii) When Ψ contains all functions from $\mathcal{X}^m$ to $\mathcal{A}$, Problem (2) can be reformulated as

$$\begin{aligned} & H_L(Y|X^m) \\ &= \min_{\substack{\psi(x^m)\in\mathcal{A},\\ \forall\ x^m\in\mathcal{X}^m}} \sum_{x^m\in\mathcal{X}^m} P_{X^m}(x^m)\mathbb{E}_{Y\sim P_{Y|X^m=x^m}}[L(Y,\psi(x^m))] \\ &= \sum_{x^m\in\mathcal{X}^m} P_{X^m}(x^m) \min_{\psi(x^m)\in\mathcal{A}} \mathbb{E}_{Y\sim P_{Y|X^m=x^m}}[L(Y,\psi(x^m))], \end{aligned} \tag{3}$$

where, in the last step, the training problem is decomposed into a sequence of separated optimization problems, each optimizing an action $\psi(x^m)$ for given $x^m \in \mathcal{X}^m$. For the considered Ψ, $H_L(Y|X^m)$ in (2) is termed the generalized conditional entropy of Y given X^m [24]. Similarly, the generalized (unconditional) entropy $H_L(Y)$ is defined as [24]–[26]

$$H_L(Y) = \min_{a\in\mathcal{A}} \mathbb{E}_{Y\sim P_Y}[L(Y,a)]. \tag{4}$$

The optimal solution to (4) is called *Bayes action*, which is denoted as a_{P_Y}. If the Bayes actions are not unique, one can pick any such Bayes action as a_{P_Y} [27]. The generalized conditional entropy for Y given $X^m = x^m$ is [24]

$$\begin{aligned} H_L(Y|X^m = x^m) &= \min_{a\in\mathcal{A}} \mathbb{E}_{Y\sim P_{Y|X^m=x^m}}[L(Y,a)] \\ &= \min_{\psi(x^m)\in\mathcal{A}} \mathbb{E}_{Y\sim P_{Y|X^m=x^m}}[L(Y,\psi(x^m))]. \end{aligned} \tag{5}$$

Substituting (5) into (3), yields the relationship

$$H_L(Y|X^m) = \sum_{x^m\in\mathcal{X}^m} P_{X^m}(x^m)\ H_L(Y|X^m = x^m). \tag{6}$$

We assume that entropy and conditional entropy discussed in this paper are bounded. We also need to define the generalized mutual information, which is given by

$$I_L(Y;X^m) = H_L(Y) - H_L(Y|X^m). \tag{7}$$

Examples of the loss function L and the associated generalized entropy $H_L(Y)$ were discussed in [21], [24]–[26].

B. Testing Loss

The testing loss, also called validation loss, of supervised learning is the expected loss on testing data using the trained predictor. In the sequel, we use the concept of generalized cross entropy to characterize the *testing loss*. The generalized cross entropy between Y and $\tilde{Y}$ is defined as

$$H_L(\tilde{Y};Y) = \mathbb{E}_{Y\sim P_{\tilde{Y}}}\left[L(Y,a_{P_Y})\right], \tag{8}$$

where a_{P_Y} is the Bayes action defined above. Similarly, the generalized conditional cross entropy between $\tilde{Y}$ and Y given $\tilde{X}^m = x^m$ is

$$H_L(\tilde{Y};Y|\tilde{X}^m = x^m) = \mathbb{E}_{Y\sim P_{\tilde{Y}|\tilde{X}^m=x^m}}\left[L(Y,a_{P_{Y|X^m=x^m}})\right], \tag{9}$$

where $a_{P_{Y|X^m=x^m}}$ is the Bayes action of a predictor that was trained by using the empirical conditional distribution $P_{Y|X^m=x^m}$ of the training data and $P_{\tilde{Y}|\tilde{X}^m=x^m}$ is the empirical conditional distribution of the testing data. Using (9), the testing loss of supervised learning can be expressed as

$$H_L(\tilde{Y};Y|\tilde{X}^m) = \sum_{x^m\in\mathcal{X}^m} \tilde{P}_{\tilde{X}^m}(x^m) H_L(\tilde{Y};Y|\tilde{X}^m = x^m), \tag{10}$$

which is also termed the generalized conditional cross entropy between $\tilde{Y}$ and Y given $\tilde{X}^m$.

IV. Impact of Information Freshness on Supervised Learning based Forecasting

In this section, we analyze the training and testing loss performance of supervised learning under different age values. It is shown that the minimum training loss is a function of the age, but the function is not always monotonic. By using a local information geometric approach, we provide a sufficient condition under which the training loss can be closely approximated as a non-decreasing age function. Similar conditions for the monotonicity of the testing loss on age are also discussed.

A. Training Loss under Constant AoI

For ease of understanding, we start by analyzing the minimum training loss under a constant AoI, i.e, $\Delta^m(t) = \delta^m$, for all time t. The more practical case of time-varying AoI will be studied subsequently in Section IV-B.

Markov chain has been a widely-used model in time-series analysis [28]–[30]. Define $X^m_{t-\tau^m} = (X_{1,t-\tau_1},\ldots,X_{m,t-\tau_m})$ for $\tau^m = (\tau_1,\ldots,\tau_m)$, where $X_{l,t-\tau_l}$ is the feature generated from transmitter l at time $t-\tau_l$. If $Y_t \leftrightarrow X^m_{t-\tau^m} \leftrightarrow X^m_{t-\tau^m-\mu^m}$ is a Markov chain for all $\mu^m, \tau^m \geq 0$ (assume the Markov chain is also stationary), then by using the data processing inequality [25], one can show that the minimum training loss $H_L(Y_t|X^m_{t-\delta^m})$ is a non-decreasing function of the age vector δ^m. However, practical time-series signals are rarely Markovian [31]–[34] and, as a result, the minimum training loss $H_L(Y_t|X^m_{t-\delta^m})$ is not always a monotonic function of δ^m.

To develop a unified framework for analyzing the minimum training loss $H_L(Y_t|X^m_{t-\delta^m})$, we consider a relaxation of the Markov chain model, called *ϵ-Markov chain*, which was proposed recently in [21].

Definition 1. ***ϵ-Markov Chain:*** *[21] Assume that the distributions $P_{Y|X}, P_{Z|X}$, and P_X are strictly positive. Given $\epsilon \geq 0$, a sequence of random variables Y, X, and Z is said to be an ϵ-Markov chain, denoted as $Y \overset{\epsilon}{\leftrightarrow} X \overset{\epsilon}{\leftrightarrow} Z$, if*

$$D_{\chi^2}\left(P_{Y,X,Z}||P_{Y|X}P_{Z|X}P_X\right) \leq \epsilon^2, \tag{11}$$

where $D_{\chi^2}(P_Y||Q_Y)$ is Neyman's χ^2-divergence, given by

$$D_{\chi^2}(P_Y||Q_Y) = \sum_{y\in\mathcal{Y}} \frac{(P_Y(y)-Q_Y(y))^2}{Q_Y(y)}. \tag{12}$$

The inequality (11) can be also expressed as

$$I_{\chi^2}(Y;Z|X) \leq \epsilon^2, \tag{13}$$

where $I_{\chi^2}(Y;Z|X)$ is the χ^2-conditional mutual information. If $\epsilon = 0$, then $Y \overset{\epsilon}{\leftrightarrow} X \overset{\epsilon}{\leftrightarrow} Z$ reduces to a Markov chain. Hence, ϵ-Markov chain is more general than Markov chain.

For ϵ-Markov chain, a relaxation of the data processing inequality is provided in the following lemma:

Lemma 1 (ϵ-data processing inequality)**.** *[21] If $Y \overset{\epsilon}{\leftrightarrow} X \overset{\epsilon}{\leftrightarrow} Z$ is an ϵ-Markov chain and $H_L(Y)$ for the loss function L is twice differentiable in P_Y, then we have (as $\epsilon \to 0$)*

$$I_L(Y;Z|X) = O(\epsilon^2), \tag{14}$$

$$H_L(Y|X) \leq H_L(Y|Z) + O(\epsilon^2). \tag{15}$$

By using Lemma 1, the training loss performance of supervised learning based forecasting is characterized in the following theorem. For notational simplicity, the theorem is presented for the case of $m = 2$ features, whereas it can be easily generalized to any positive integer value of m.

Theorem 1. *Let $\{(X_{1,t}, X_{2,t}, Y_t), t \geq 0\}$ be a stationary stochastic process.*

(a) The minimum training loss $H_L(Y_t|X_{1,t-\delta_1}, X_{2,t-\delta_2})$ is a function of δ_1 and δ_2, determined by

$$H_L(Y_t|X_{1,t-\delta_1}, X_{2,t-\delta_2}) = f_1(\delta_1,\delta_2) - f_2(\delta_1,\delta_2), \tag{16}$$

where $f_1(\delta_1,\delta_2)$ and $f_2(\delta_1,\delta_2)$ are two non-decreasing functions of δ_1 and δ_2, given by

$$\begin{aligned} &f_1(\delta_1,\delta_2) \\ =&H_L(Y_t|X_{1,t}, X_{2,t}) \\ &+ \sum_{k=0}^{\delta_1-1} I_L(Y_t; X_{1,t-k}, X_{2,t-\delta_2}|X_{1,t-k-1}, X_{2,t-\delta_2}) \\ &+ \sum_{k=0}^{\delta_2-1} I_L(Y_t; X_{1,t}, X_{2,t-k}|X_{1,t}, X_{2,t-k-1}), \end{aligned} \tag{17}$$

$$\begin{aligned} &f_2(\delta_1,\delta_2) \\ =&\sum_{k=0}^{\delta_1-1} I_L(Y_t; X_{1,t-k-1}, X_{2,t-\delta_2}|X_{1,t-k}, X_{2,t-\delta_2}) \\ &+ \sum_{k=0}^{\delta_2-1} I_L(Y_t; X_{1,t}, X_{2,t-k-1}|X_{1,t}, X_{2,t-k}). \end{aligned} \tag{18}$$

(b) If $H_L(Y)$ is twice differentiable in P_Y, and

$$Y_t \overset{\epsilon}{\leftrightarrow} (X_{1,t-\tau_1}, X_{2,t-\tau_2}) \overset{\epsilon}{\leftrightarrow} (X_{1,t-\tau_1-\mu_1}, X_{2,t-\tau_2-\mu_2})$$

is an ϵ-Markov chain for all $\mu_l, \tau_l \geq 0$, then $f_2(\delta_1,\delta_2) = O(\epsilon^2)$ and hence (as $\epsilon \to 0$)

$$H_L(Y_t|X_{1,t-\delta_1}, X_{2,t-\delta_2}) = f_1(\delta_1,\delta_2) + O(\epsilon^2). \tag{19}$$

Proof. See Appendix A. □

According to Theorem 1, the minimum training loss $H_L(Y_t|X^m_{t-\delta^m})$ is a function of the age vector δ^m. In addition, $H_L(Y_t|X^m_{t-\delta^m})$ is the difference between two non-decreasing functions of δ^m. Furthermore, the monotonicity of $H_L(Y_t|X^m_{t-\delta^m})$ is characterized by the parameter ϵ in the ϵ-Markov chain $Y_t \overset{\epsilon}{\leftrightarrow} X^m_{t-\tau^m} \overset{\epsilon}{\leftrightarrow} X^m_{t-\tau^m-\mu^m}$. If ϵ is small, then the empirical distribution of training data samples can be accurately approximated as a Markov chain. As a result, the term $f_2(\delta^m)$ is close to zero and $H_L(Y_t|X^m_{t-\delta^m})$ tends to be a non-decreasing function of δ^m. On the other hand, for large ϵ, $H_L(Y_t|X^m_{t-\delta^m})$ is unlikely monotonic in δ^m.

As depicted later in Figs. 3-4, the training loss can indeed be non-monotonic on δ^m. This finding suggests that it is beneficial to investigate non-monotonic age penalty functions, which are more general than the non-decreasing age penalty metrics studied in, e.g., [12], [19], [22].

B. Training Loss under Dynamic AoI

In practice, the AoI $\Delta^m(t)$ varies dynamically over time, as shown in Fig. 2. Let P_{Δ^m} denote the empirical distribution of the AoI in the training dataset and Δ^m be a random vector with the distribution P_{Δ^m}. In the case of dynamic AoI, there are two approaches for training: (i) *Separated training*: the Bayes action for the AoI value $\Delta^m = \delta^m$ is trained by only using the training data samples with AoI value δ^m [23]. The minimum training loss of separate training for $\Delta^m = \delta^m$ is $H_L(Y_t|X^m_{t-\delta^m})$, which has been analyzed in Section IV-A. (ii) *Joint training*: the training data samples of all AoI values are combined into a large training dataset and the Bayes actions for different AoI values are trained together. In joint training, the AoI value can either be included as part of the feature, or be excluded from the feature. If the AoI value is included in the training data (i.e., as part of the feature), the minimum training loss of joint training is $H_L(Y_t|X^m_{t-\Delta^m}, \Delta^m)$. If the AoI value is excluded from the training data, the minimum training loss of joint training is $H_L(Y_t|X^m_{t-\Delta^m})$. Because conditioning reduces the generalized entropy [24], [25], we have

$$H_L(Y_t|X^m_{t-\Delta^m}, \Delta^m) \leq H_L(Y_t|X^m_{t-\Delta^m}). \tag{20}$$

Hence, a smaller training loss can be achieved by including the AoI in the feature. Moreover, similar to (6), one can get

$$H_L(Y_t|X^m_{t-\Delta^m}, \Delta^m) = \sum_{\delta^m} P_{\Delta^m}(\delta^m) H_L(Y_t|X^m_{t-\delta^m}). \tag{21}$$

Therefore, the minimum training loss of joint training is simply the expectation of the training loss of separated training. Our experiment results show that joint training can have a much smaller computational complexity than separated training (see the discussions in Section V-D). Therefore, we suggest to use joint training and add AoI into the feature.

The results in Theorem 1 can be directly generalized to the scenario of dynamic AoI. In particular, if the age processes of two experiments, denoted by subscripts c and d, satisfy a

sample-path ordering[1]

$$\Delta_c^m(t) \leq \Delta_d^m(t), \ \forall \ t, \tag{22}$$

then, similar to (19), one can obtain (as $\epsilon \to 0$)

$$H_L(Y_t|X^m_{t-\Delta_c^m}, \Delta_c^m) \leq H_L(Y_t|X^m_{t-\Delta_d^m}, \Delta_d^m) + O(\epsilon^2). \tag{23}$$

Next, we show that (23) can be also proven under a weaker stochastic ordering condition (26).

Definition 2. ***Univariate Stochastic Ordering:*** *[35] A random variable X is said to be stochastically smaller than another random variable Z, denoted as $X \leq_{st} Z$, if*

$$P(X > x) \leq P(Z > x), \quad \forall x \in \mathbb{R}. \tag{24}$$

Definition 3. ***Multivariate Stochastic Ordering:*** *[35] A set $U \subseteq \mathbb{R}^m$ is called upper if $z^m \in U$ whenever $z^m \geq x^m$ and $x^m \in U$. A random vector X^m is said to be stochastically smaller than another random vector Z^m, denoted as $X^m \leq_{st} Z^m$, if*

$$P(X^m \in U) \leq P(Z^m \in U), \quad \textit{for all upper sets } U \subseteq \mathbb{R}^m. \tag{25}$$

Theorem 2. *If $\{(X_t^m, Y_t), t \geq 0\}$ is a stationary random process, $Y_t \overset{\epsilon}{\leftrightarrow} X^m_{t-\tau^m} \overset{\epsilon}{\leftrightarrow} X^m_{t-\tau^m-\mu^m}$ is an ϵ-Markov chain for all $\mu^m, \tau^m \geq 0$, $H_L(Y)$ is twice differentiable in P_Y, and the empirical distributions of the training datasets in two experiments c and d satisfy*

$$\Delta_c^m \leq_{st} \Delta_d^m, \tag{26}$$

then (23) *holds.*

Proof. See Appendix B. □

According to Theorem 2, if Δ_c^m is stochastically smaller than Δ_d^m, then the minimum training loss of joint training in Experiment c is approximately smaller than that in Experiment d, where the approximation error is of the order $O(\epsilon^2)$.

C. Testing Loss under Dynamic AoI

Let $\mathcal{P}^{\mathcal{Y}}$ denote the space of distributions on $\mathcal{Y}$ and relint($\mathcal{P}^{\mathcal{Y}}$) denote the relative interior of $\mathcal{P}^{\mathcal{Y}}$, i.e., the subset of strictly positive distributions.

Definition 4. ***β-neighborhood:*** *[36] Given $\beta \geq 0$, the* β-neighborhood *of a reference distribution $Q_Y \in relint(\mathcal{P}^{\mathcal{Y}})$ is the set of distributions that are in a Neyman's χ^2-divergence ball of radius β^2 centered on Q_Y, i.e.,*

$$\mathcal{N}_\beta^{\mathcal{Y}}(Q_Y) = \left\{P_Y \in \mathcal{P}^{\mathcal{Y}} : D_{\chi^2}(P_Y||Q_Y) \leq \beta^2\right\}. \tag{27}$$

Theorem 3. *Let $\{(X_t^m, Y_t), t \geq 0\}$ be a stationary stochastic process.*

(a) If the empirical distributions of training data and testing data are close to each other such that

$$D_{\chi^2}\left(P_{\tilde{Y}_t, \tilde{X}^m_{t-\tilde{\Delta}^m}, \tilde{\Delta}^m} || P_{Y_t, X^m_{t-\Delta^m}, \Delta^m}\right) \leq \beta^2, \tag{28}$$

then the testing loss is close to the minimum training loss, i.e.,

$$H_L(\tilde{Y}_t; Y_t|\tilde{X}^m_{t-\tilde{\Delta}^m}, \tilde{\Delta}^m) = H_L(Y_t|X^m_{t-\Delta^m}, \Delta^m) + O(\beta), \tag{29}$$

provided that the testing loss is bounded.

(b) In addition, if $H_L(Y)$ is twice differentiable in P_Y, $Y_t \overset{\epsilon}{\leftrightarrow} X^m_{t-\tau^m} \overset{\epsilon}{\leftrightarrow} X^m_{t-\tau^m-\mu^m}$ is an ϵ-Markov chain for all $\mu^m, \tau^m \geq 0$, and the empirical distributions of the testing datasets in two experiments c and d satisfy $\tilde{\Delta}_c^m \leq_{st} \tilde{\Delta}_d^m$, then the corresponding testing loss satisfies

$$H_L(\tilde{Y}_t; Y_t|\tilde{X}^m_{t-\tilde{\Delta}_c^m}, \tilde{\Delta}_c^m) \leq H_L(\tilde{Y}_t; Y_t|\tilde{X}^m_{t-\tilde{\Delta}_d^m}, \tilde{\Delta}_d^m) + O(\max\{\epsilon^2, \beta\}). \tag{30}$$

Proof. See Appendix C. □

As shown in Theorem 3, if the empirical distributions of training data and testing data are close to each other, then the minimum training loss and testing loss have similar growth patterns as the AoI grows.

V. Case Studies

We provide several case studies using a real-world solar power prediction task. A Long Short-Term Memory (LSTM) neural network is used as the prediction model. Four cases are studied: (i) training loss under constant AoI, (ii) sensitivity analysis of input sequence length b under constant AoI, (iii) training loss under dynamic AoI, and (iv) testing loss under dynamic AoI.

A. Experimental Setup

1) Environment: We use Tensorflow 2 [37] on Ubuntu 16.04, and a server with two Intel Xeon E5-2630 v4 CPUs and four GeForce GTX 1080 Ti GPUs to perform the experiments.

2) Dataset: A real-world dataset from the probabilistic solar power forecasting task of the Global Energy Forecasting Competition (GEFCom) 2014 [38] is used for evaluation. The dataset contains 2 years of measurement data for 13 signal processes, including humidity, thermal, wind speed and direction, and other weather metrics, as explained in [38]. Feature $X_{l,t} = (s_{l,t}, \ldots, s_{l,t-b+1})$ of signal l is a time sequence of length b. The task is to predict the solar power at an hourly level. We have used two different AoI values δ_1 and δ_2, where 6 features are of the AoI δ_1 and 7 features are of the AoI δ_2. In jointly training, we use an aggregated dataset with samples of constant AoI $\delta_1 = \delta_2$ ranging from 1 to 48 (hours). The first year of data is used for training and the second year of data is used for testing. During preprocessing, both datasets are normalized such that the training dataset has a zero mean and a unity standard derivation.

3) Prediction Model: A Long Short-Term Memory (LSTM) neural network is used for prediction. It composes of one input layer, one hidden layer with 32 LSTM cells, and one fully-connected (dense) output layer. The object of training is to minimize the expected quadratic loss under the empirical distribution of the training samples. In other words, empirical

[1]We say $x^m \leq z^m$, if $x_l \leq z_l$ for every $l = 1, 2, \ldots, m$.

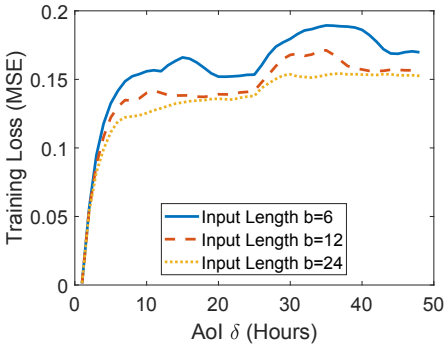

Fig. 3: Training loss vs. AoI $\delta_1 = \delta_2 = \delta$, where the input length is $b = 6$, 12, and 24.

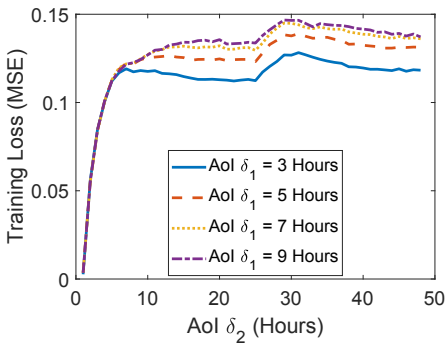

Fig. 4: Training loss vs. AoI δ_2, where the input length is $b = 24$ and δ_1 =3, 5, 7, and 9 hours.

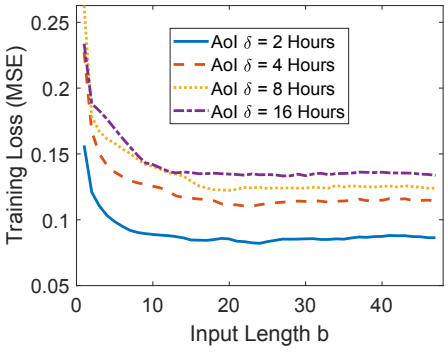

Fig. 5: Training loss vs. input length b, where the AoI is $\delta_1 = \delta_2$ =2, 4, 8, and 16 hours.

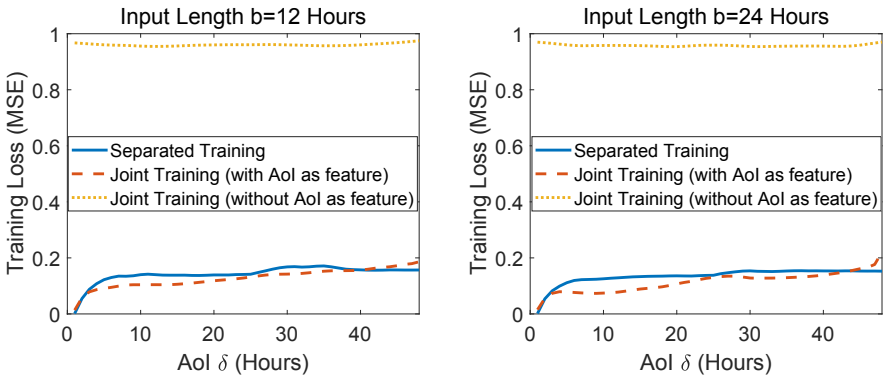

Fig. 6: Training loss under dynamic AoI $\Delta_1(t) = \Delta_2(t)$, where the input length is $b = 12$ and 24. Three training approaches are illustrated: (i) separated training, (ii) joint training with AoI as part of feature, and (iii) joint training without using AoI as part of feature.

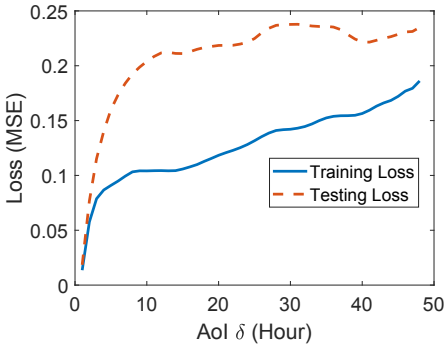

Fig. 7: Testing loss and training loss vs. dynamic AoI $\Delta_1(t) = \Delta_2(t)$, where the input length is $b = 12$.

mean square error (MSE) $\frac{1}{K}\sum_{i=1}^{K}(y_i - \hat{y}_i)^2$ is minimized, where K is the number of training data entries, y_i is the actual label, and $\hat{y}_i$ is the predicted label. All the experimental results are averaged over multiple runs to reduce the randomness that occurs during the training of LSTM. This training setting and the hyper-parameters of Tensorflow 2 training algorithm are consistently used across all evaluations. Note that in theoretical analysis, we have considered that Ψ consists of all possible functions from $\mathcal{X}^m$ to $\mathcal{A}$. In practice, neural networks cannot represent such a large function space and the trained weights of the neural network may not be globally optimal. Hence, the training loss (MSE) in the experimental study is larger than the minimum training loss analyzed in our theory, but their patterns should be similar.

B. Training Loss under Constant AoI

We start with the scenario of separated training, where the AoI is the same across all input features, i.e., $\delta_1 = \delta_2 = \delta$. As shown in Fig. 3, the training loss is not a monotonic function of AoI δ. Moreover, with the increase of input sequence length b, the training loss becomes close to a non-decreasing function. However, with the increase of input sequence length b, the training loss tends to be close to a non-decreasing function. This phenomenon can be interpreted by using Shannon's high-order Markov model for information sources [39]. Specifically, the feature process $X_{l,t}$ can be approximated as a Markov chain model with order b, where the approximation error reduces as the order b grows. Because of this, the training loss is far away from an increasing function of the AoI when b = 6 or 12, and is nearly an increasing function of the AoI when b = 24.

Next, we consider a more general case where δ_1 and δ_2 can have different values. The results are shown in Fig. 4, where both δ_1 and δ_2 affect the training loss in accordance to Theorem 1.

C. Sensitivity Analysis of Input Sequence Length b Under Constant AoI

With the increase of input length sequence b, the training loss is expected to decrease, because conditioning reduces the generalized conditional entropy. To show this, we evaluate the training loss for a wide range of input lengths, as shown in Fig. 5. The observations are two-folded. First, for a given AoI value, the training loss is a non-increasing function of the input length b. Second, even with the increase of input length, a

larger AoI leads to a larger training loss. The observed pattern agrees with our theoretical analysis.

D. Training Loss under Dynamic AoI

In the *separated training* method considered above, one predictor (i.e., an LSTM neural network) is trained for every AoI value. Hence, a number of predictors are needed. Training these predictors may incur a huge computational cost, which would be impractical for prediction tasks with huge datasets. As discussed in Section IV-B, *joint training* is a better approach, where a single predictor is trained by jointly using the input samples of different AoI values. As plotted in Fig. 6, if the AoI is excluded from the feature, joint training has a significant performance degradation compared to separated training as described in (20). However, with AoI as a part of the input feature, joint training has comparable performance as separated training, which agrees with the relationship in (21). For a wide range of AoI values, the performance of joint training is slightly better than separated training because it uses all training data where separated predictor can only be trained on the data with certain AoI. This phenomenon is in alignment with data augmentation [40]. In current training dataset for jointly training, we mainly focus on small AoIs so the joint model may not be as good as separated training for very large AoI values. But such long AoI is rare in real-world applications. Thus, the idea of appending AoI to the input features is a good idea for time-varying AoI.

E. Testing Loss under Dynamic AoI

Training loss and testing loss are compared in Fig. 7 for joint training under dynamic AoI. One can observe that the testing loss is not monotonic in AoI, but it has a growing trend that is similar to the training loss. In addition, the testing loss is larger than the training loss, which is caused by the difference between the empirical distributions of training and testing datasets. Such a difference is quite normal in machine learning, which occurs, for instance, if the datasets for training/testing are not sufficiently large or if there is concept shift [41].

VI. Conclusion

In this paper, we have presented a unified theoretical framework to analyze how the age of correlated features affects the performance in supervised learning based forecasting. It has been shown that the minimum training loss is a function of the age, which is not necessarily monotonic. Conditions have been provided under which the training loss and testing loss are approximately non-decreasing in age. Our investigation suggests that, by (i) jointly training the forecasting actions for different age values and (ii) adding the age value into the input feature, both forecasting accuracy and computational complexity can be greatly improved.

References

[1] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[2] T. Do, M. Duong, Q. Dang, and M. Le, "Real-time self-driving car navigation using deep neural network," in *2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*, 2018, pp. 7–12.

[3] T. Hong and P. Pinson, "Energy forecasting in the big data world," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1387–1388, Jan. 2019.

[4] D. Kaur, S. N. Islam, M. A. Mahmud, and Z. Dong, "Energy forecasting in smart grid systems: A review of the state-of-the-art techniques," 2020, https://arxiv.org/abs/2011.12598.

[5] A. Dobbs, T. Elgindy, B.-M. Hodge, A. Florita, and J. Novacheck, "Short-term solar forecasting performance of popular machine learning algorithms," *National Renewable Energy Laboratory*, 2017.

[6] C. Zhang, J. J. Q. Yu, and Y. Liu, "Spatial-temporal graph attention networks: A deep learning approach for traffic forecasting," *IEEE Access*, vol. 7, pp. 166 246–166 256, 2019.

[7] H. Dia, "An object-oriented neural network approach to short-term traffic forecasting," *European Journal of Operational Research*, vol. 131, no. 2, pp. 253 – 261, 2001.

[8] T. Gao, Y. Chai, and Y. Liu, "Applying long short term momory neural networks for predicting stock closing price," in *8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2017, pp. 575–578.

[9] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *IEEE INFOCOM*, 2012, pp. 2731–2735.

[10] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7492–7508, 2017.

[11] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, 2011, pp. 350–358.

[12] Y. Sun and B. Cyr, "Sampling for data freshness optimization: Non-linear age functions," *Journal of Communications and Networks*, vol. 21, no. 3, pp. 204–219, 2019.

[13] B. Zhou and W. Saad, "Optimal sampling and updating for minimizing age of information in the internet of things," in *IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[14] A. Arafa, K. Banawan, K. G. Seddik, and H. V. Poor, "On timely channel coding with hybrid ARQ," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[15] R. D. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 3008–3012.

[16] X. Wu, J. Yang, and J. Wu, "Optimal status update for age of information minimization with an energy harvesting source," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 193–204, 2018.

[17] Y. Sun, Y. Polyanskiy, and E. Uysal, "Sampling of the Wiener process for remote estimation over a channel with random delay," *IEEE Transactions on Information Theory*, vol. 66, no. 2, pp. 1118–1135, 2020.

[18] T. Z. Ornee and Y. Sun, "Sampling for remote estimation through queues: Age of information and beyond," in *International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2019, pp. 1–8.

[19] M. Klügel, M. H. Mamduhi, S. Hirche, and W. Kellerer, "AoI-penalty minimization for networked control systems with packet loss," in *IEEE Conference on Computer Communications Workshops*, 2019, pp. 189–196.

[20] T. Soleymani, J. S. Baras, and K. H. Johansson, "Stochastic control with stale information–part i: Fully observable systems," in *IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 4178–4182.

[21] M. K. C. Shisher and Y. Sun, "The age of features in time-series forecasting based on supervised learning," in preparation, 2021.

[22] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age and value of information: Non-linear age case," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 326–330.

[23] R. P. Masini, M. C. Medeiros, and E. F. Mendes, "Machine learning advances for time series forecasting," 2020, https://arxiv.org/abs/2012.12802.

[24] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 4240–4248.

[25] A. P. Dawid, "Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design," *Technical Report 139*, 1998, https://www.ucl.ac.uk/statistics/research/reports.

[26] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory," *Annals of Statistics*, vol. 32, no. 4, pp. 1367–1433, 08 2004.

[27] A. P. Dawid and M. Musio, "Theory and applications of proper scoring rules," *Metron*, vol. 72, no. 2, pp. 169–183, 2014.

[28] F. O. Hocaoglu and F. Serttas, "A novel hybrid (Mycielski-Markov) model for hourly solar radiation forecasting," *Renewable Energy*, vol. 108, pp. 635 – 643, 2017.

[29] J. Munkhammar, D. van der Meer, and J. Widén, "Probabilistic forecasting of high-resolution clear-sky index time-series using a Markov-chain mixture distribution model," *Solar Energy*, vol. 184, pp. 688–695, 2019.

[30] C. M. Turner, R. Startz, and C. R. Nelson, "A Markov model of heteroskedasticity, risk, and learning in the stock market," *Journal of Financial Economics*, vol. 25, no. 1, pp. 3–22, 1989.

[31] N. G. van Kampen, "Remarks on non-Markov processes," *Brazilian Journal of Physics*, vol. 28, no. 2, pp. 90–96, 1998.

[32] J. Łuczka, "Non-Markovian stochastic processes: Colored noise," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 15, no. 2, p. 026107, 2005.

[33] M. Deshpande and G. Karypis, "Selective Markov models for predicting web page accesses," *ACM transactions on internet technology (TOIT)*, vol. 4, no. 2, pp. 163–184, 2004.

[34] K. Krishna, D. Jain, S. V. Mehta, and S. Choudhary, "An LSTM based system for prediction of human activities with durations," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–31, 2018.

[35] M. Shaked and J. G. Shanthikumar, *Stochastic orders*. Springer Science & Business Media, 2007.

[36] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," 2019, https://arxiv.org/abs/1911.09105.

[37] M. Abadi, A. Agarwal, P. Barham, *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, https://arxiv.org/abs/1603.04467.

[38] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *International Journal of Forecasting*, vol. 32, no. 3, pp. 896 – 913, 2016.

[39] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[40] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" in *international conference on digital image computing: techniques and applications (DICTA)*, 2016, pp. 1–6.

[41] P. Vorburger and A. Bernstein, "Entropy-based concept shift detection," in *Sixth International Conference on Data Mining (ICDM'06)*, 2006, pp. 1113–1118.

APPENDIX A
PROOF OF THEOREM 1

Part (a): By the definitions of generalized conditional entropy and generalized mutual information, one can obtain

$$\begin{aligned} &H_L(Y|X_1,X_2,Z_1,Z_2)\\ =&H_L(Y|Z_1,Z_2)-I_L(Y;X_1,X_2|Z_1,Z_2)\\ =&H_L(Y|X_1,X_2)-I_L(Y;Z_1,Z_2|X_1,X_2), \end{aligned} \tag{31}$$

which yields

$$\begin{aligned} H_L(Y|Z_1,Z_2)=H_L(Y|X_1,X_2)&+I_L(Y;X_1,X_2|Z_1,Z_2)\\ &-I_L(Y;Z_1,Z_2|X_1,X_2). \end{aligned} \tag{32}$$

Now, if we replace Z_1, Z_2, X_1, X_2, and Y in (32) with $X_{1,t-k-1}, X_{2,t-\delta_2}, X_{1,t-k}, X_{2,t-\delta_2}$, and Y_t, respectively, then (32) becomes

$$\begin{aligned} &H_L(Y_t|X_{1,t-k-1},X_{2,t-\delta_2})\\ =&H_L(Y_t|X_{1,t-k},X_{2,t-\delta_2})\\ &+I_L(Y_t;X_{1,t-k},X_{2,t-\delta_2}|X_{1,t-k-1},X_{2,t-\delta_2})\\ &-I_L(Y_t;X_{1,t-k-1},X_{2,t-\delta_2}|X_{1,t-k},X_{2,t-\delta_2}). \end{aligned} \tag{33}$$

Equation (33) is valid for any value of t and k. Therefore, taking summation of $H_L(Y_t|X_{1,t-k-1},X_{2,t-\delta_2})$ from $k=0$ to δ_1-1, we get

$$\begin{aligned} &H_L(Y_t|X_{1,t-\delta_1},X_{2,t-\delta_2})\\ =&H_L(Y_t|X_{1,t},X_{2,t-\delta_2})\\ &+\sum_{k=0}^{\delta_1-1} I_L(Y_t;X_{1,t-k},X_{2,t-\delta_2}|X_{1,t-k-1},X_{2,t-\delta_2})\\ &-\sum_{k=0}^{\delta_1-1} I_L(Y_t;X_{1,t-k-1},X_{2,t-\delta_2}|X_{1,t-k},X_{2,t-\delta_2}). \end{aligned} \tag{34}$$

Similarly, we can establish that

$$\begin{aligned} &H_L(Y_t|X_{1,t},X_{2,t-\delta_2})\\ =&H_L(Y_t|X_{1,t},X_{2,t})\\ &+\sum_{k=0}^{\delta_2-1} I_L(Y_t;X_{1,t},X_{2,t-k}|X_{1,t},X_{2,t-k-1})\\ &-\sum_{k=0}^{\delta_2-1} I_L(Y_t;X_{1,t},X_{2,t-k-1}|X_{1,t},X_{2,t-k}). \end{aligned} \tag{35}$$

Combining (34) and (35), we get (16).

Because, mutual information is a non-negative term, one can observe from (17) that for a fixed δ_2, the functions $f_1(\delta_1,\delta_2)$ is a non-decreasing function of δ_1. Moreover, $f_1(\delta_1,\delta_2)$ can be written in another form as

$$\begin{aligned} &f_1(\delta_1,\delta_2)\\ =&H_L(Y_t|X_{1,t},X_{2,t})\\ &+\sum_{k=0}^{\delta_2-1} I_L(Y_t;X_{1,t-\delta_1},X_{2,t-k}|X_{1,t-\delta_1},X_{2,t-k-1})\\ &+\sum_{k=0}^{\delta_1-1} I_L(Y_t;X_{1,t-k},X_{2,t}|X_{1,t-k-1},X_{2,t}). \end{aligned} \tag{36}$$

From the above equation, it is also observed that for a fixed δ_1, the functions $f_1(\delta_1,\delta_2)$ is a non-decreasing function of δ_2. Therefore, (17) and (36) imply that $f_1(\delta_1,\delta_2)$ is a non-decreasing function of δ_1 and δ_2. Similarly, from (18), we can deduce that $f_2(\delta_1,\delta_2)$ is a non-decreasing function of δ_1 and δ_2.

Part (b): Because $H_L(Y)$ is twice differentiable in P_Y and $Y_t \overset{\epsilon}{\leftrightarrow} (X_{1,t-\tau_1},X_{2,t-\tau_2}) \overset{\epsilon}{\leftrightarrow} (X_{1,t-\tau_1-\mu_1},X_{2,t-\tau_2-\mu_2})$ is an ϵ-Markov chain for all $\mu_l,\tau_l \geq 0$, by using Lemma 1,

$f_2(\delta_1, \delta_2)$ satisfies

$$f_2(\delta_1, \delta_2) = \sum_{k=0}^{\delta_1 - 1} O(\epsilon^2) + \sum_{k=0}^{\delta_2 - 1} O(\epsilon^2) = O(\epsilon^2). \quad (37)$$

This concludes the proof.

Appendix B
Proof of Theorem 2

By using Theorem 1 and substituting Equation (19) into (21), we obtain

$$H_L(Y_t|X_{t-\Delta^m}^m, \Delta^m) = \mathbb{E}_{\Delta^m \sim P_{\Delta^m}}[f_1(\Delta^m)] + O(\epsilon^2). \quad (38)$$

The expectations in the above equation exist as entropy and conditional entropy are considered bounded and $H_L(Y_t|X_{t-\Delta^m}^m, \Delta^m) \leq H_L(Y_t) < \infty$. Moreover, $f_1(\delta^m)$ is a non-decreasing function of δ^m. Therefore, as $\Delta_c^m \leq_{st} \Delta_d^m$, we get [35]

$$\mathbb{E}_{\Delta_c^m \sim P_{\Delta_c^m}}[f_1(\Delta_c^m)] \leq \mathbb{E}_{\Delta_d^m \sim P_{\Delta_d^m}}[f_1(\Delta_d^m)]. \quad (39)$$

By using (38) and (39), we obtain

$$H_L(Y_t|X_{t-\Delta_c^m}^m, \Delta_c^m) \leq H_L(Y_t|X_{t-\Delta_d^m}^m, \Delta_d^m) + O(\epsilon^2). \quad (40)$$

This concludes the proof.

Appendix C
Proof of Theorem 3

Part (a): First, we provide the following lemma.

Lemma 2. *If*

$$D_{\chi^2}\left(P_{\tilde{Y},\tilde{X}}||P_{Y,X}\right) \leq \beta^2, \quad (41)$$

then

$$P_{\tilde{X}}(x) = P_X(x) + O(\beta), \ \forall x \in \mathcal{X}, \quad (42)$$

$$P_{\tilde{Y}|\tilde{X}}(y|x) = P_{Y|X}(y|x) + O(\beta), \ \forall y \in \mathcal{Y}, \forall x \in \mathcal{X}. \quad (43)$$

In addition, if $H_L(\tilde{Y}; Y|\tilde{X})$ is bounded, then

$$H_L(\tilde{Y}; Y|\tilde{X}) = H_L(Y|X) + O(\beta). \quad (44)$$

Proof. See Appendix D. □

Using Lemma 2, we prove part (a) of Theorem 3. Because the condition (28) holds and the testing loss is assumed bounded, if we replace X, Y, $\tilde{X}$, and $\tilde{Y}$ in (44) with $(X_{t-\Delta^m}^m, \Delta^m)$, Y_t, $(\tilde{X}_{t-\tilde{\Delta}^m}^m, \tilde{\Delta}^m)$, and $\tilde{Y}_t$, then (44) becomes (29).

Part (b): By using Theorem 2 and (29), we get

$$H_L(\tilde{Y}_t; Y_t|\tilde{X}_{t-\tilde{\Delta}_a^m}^m, \tilde{\Delta}_a^m) \leq H_L(\tilde{Y}_t; Y_t|\tilde{X}_{t-\tilde{\Delta}_b^m}^m, \tilde{\Delta}_b^m) + O(\beta) + O(\epsilon^2), \quad (45)$$

where $O(\beta) + O(\epsilon^2) = O(\max\{\epsilon^2, \beta\})$. This concludes the proof.

Appendix D
Proof of Lemma 2

Because

$$D_{\chi^2}(P_{\tilde{Y},\tilde{X}}||P_{Y,X}) \geq D_{\chi^2}(P_{\tilde{X}}||P_X), \quad (46)$$

from (41), we have

$$D_{\chi^2}(P_{\tilde{X}}||P_X) \leq \beta^2. \quad (47)$$

By using the definition of Neyman's χ^2-divergence, from inequality (47), we can show (42). Now, from inequality (41), we get

$$\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \frac{\left[P_{\tilde{Y}|\tilde{X}}(y|x)P_{\tilde{X}}(x) - P_{Y|X}(y|x)P_X(x)\right]^2}{P_{Y|X}(y|x)P_X(x)} \leq \beta^2. \quad (48)$$

Substituting (42) into the above inequality and after some algebraic operations, we have

$$\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} P_X(x) \frac{\left[P_{\tilde{Y}|\tilde{X}}(y|x) - P_{Y|X}(y|x)\right]^2}{P_{Y|X}(y|x)} \leq O(\beta^2). \quad (49)$$

From (48), it is simple to observe that $P_{Y|X}(y|x) \neq 0$ and $P_X(x) \neq 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Because $P_X(x) \neq 0$, $\forall x \in \mathcal{X}$, we can further obtain

$$\sum_{y \in \mathcal{Y}} \frac{\left[P_{\tilde{Y}|\tilde{X}}(y|x) - P_{Y|X}(y|x)\right]^2}{P_{Y|X}(y|x)} \leq \frac{O(\beta^2)}{P_X(x)}, \forall x \in \mathcal{X}. \quad (50)$$

An equivalent representation of the above inequality is that there exists a function $C : \mathcal{Y} \mapsto \mathbb{R}$ such that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$P_{\tilde{Y}|\tilde{X}}(y|x) - P_{Y|X}(y|x) = \frac{O(\beta)}{\sqrt{P_X(x)}} C(y) \sqrt{P_{Y|X}(y|x)} = O(\beta), \quad (51)$$

where

$$\sum_{y \in \mathcal{Y}} C^2(y) \leq 1. \quad (52)$$

Equation (51) implies (43).

Define a function $g(P_{\tilde{Y}|\tilde{X}=x})$ as

$$g(P_{\tilde{Y}|\tilde{X}=x}) = H_L(\tilde{Y}; Y|\tilde{X} = x) - H_L(Y|X = x) = \mathbb{E}_{\tilde{Y} \sim P_{\tilde{Y}|\tilde{X}=x}}[L(Y, a_{P_{Y|X=x}})] - H_L(Y|X = x)]. \quad (53)$$

In this lemma, we assume that conditional cross entropy $H_L(\tilde{Y}; Y|\tilde{X})$ is bounded. Also, conditional entropy $H_L(Y|X)$ is considered bounded in this paper. Thus, $g(P_{\tilde{Y}|\tilde{X}=x})$ is bounded for all $x \in \mathcal{X}$. By the definition of conditional entropy

and conditional cross entropy, we obtain

$$\begin{aligned} & g(P_{\tilde{Y}|\tilde{X}=x}) \\ =& \sum_{y\in\mathcal{Y}} \left(P_{\tilde{Y}|\tilde{X}}(y|x) - P_{Y|X}(y|x)\right) L(y, a_{P_{Y|X=x}}) \\ =& O(\beta). \end{aligned} \tag{54}$$

The last equality is obtained by substituting the value of $P_{\tilde{Y}|\tilde{X}}(y|x)$ from (43). Substituting (53) into (54), we get

$$H_L(\tilde{Y}; Y|\tilde{X} = x) = H_L(Y|X = x) + O(\beta). \tag{55}$$

By using (42) and (55), yields

$$\begin{aligned} H_L(\tilde{Y}; Y|\tilde{X}) &= \sum_{x\in\mathcal{X}} P_{\tilde{X}}(x) H_L(\tilde{Y}; Y|\tilde{X} = x) \\ &= \sum_{x\in\mathcal{X}} \left(P_X(x) + O(\beta)\right) H_L(\tilde{Y}; Y|\tilde{X} = x) \\ &= \sum_{x\in\mathcal{X}} P_X(x) H_L(\tilde{Y}; Y|\tilde{X} = x) + O(\beta) \\ &= \sum_{x\in\mathcal{X}} P_X(x) \left(H_L(Y|X = x) + O(\beta)\right) + O(\beta) \\ &= \sum_{x\in\mathcal{X}} P_X(x) H_L(Y|X = x) + O(\beta) \\ &= H_L(Y|X) + O(\beta). \end{aligned} \tag{56}$$

This concludes the proof.