IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) | 979-8-3503-8447-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/INFOCOMWKSHPS61880.2024.10620843

On the Monotonicity of Information Aging

MD Kamran Chowdhury Shisher and Yin Sun Department of ECE, Auburn University, AL, USA

Abstract—In this paper, we analyze the monotonicity of information aging in a remote estimation system, where historical observations of a Gaussian autoregressive AR(p) process are used to predict its future values. We consider two widely used loss functions in estimation: (i) logarithmic loss function for maximum likelihood estimation and (ii) quadratic loss function for MMSE estimation. The estimation error of the AR(p) process is written as a generalized conditional entropy which has closedform expressions. By using a new information-theoretic tool called *ϵ-Markov chain*, we can evaluate the divergence of the AR(p) process from being a Markov chain. When the divergence ϵ is large, the estimation error of the AR(p) process can be far from a non-decreasing function of the Age of Information (AoI). Conversely, for small divergence ϵ, the estimation error is close to a non-decreasing AoI function. Each observation is a short sequence taken from the AR(p) process. As the observation sequence length increases, the parameter ϵ progressively reduces to zero, and hence the estimation error becomes a non-decreasing AoI function. These results underscore a connection between the monotonicity of information aging and the divergence of from being a Markov chain.

This work was supported in part by NSF grant CNS-2239677 and ARO grant W911NF-21-1-0244.

979-8-3503-8447-5/24/$31.00 ©2024 IEEE

I. Introduction

Timely updates of sensor measurements are crucial for real-time state estimation and decision-making in networked controlled and cyber-physical systems, such as UAV navigation, real-time surveillance, factory automation, and weather monitoring systems. To evaluate the timeliness of sensor measurements received from a remote sensor, the concept of *Age of Information* (AoI) was introduced in [1], [2]. Let $U(t)$ be the generation time of the freshest sensor measurement delivered to the receiver by time t. The AoI $\Delta(t)$, as a function of time t, is defined as

$$\Delta(t) := t - U(t), \tag{1}$$

which is the time difference between the current time t and the generation time $U(t)$ of the most recently delivered sensor data. A smaller AoI indicates the presence of recently generated sensor data at the receiver. There has been a significant research efforts on analyzing and optimizing AoI on communication networks [2]–[29].

In this paper, we investigate a remote estimation system where a time-varying target is estimated based on observations collected from a sensor. Due to communication delays and transmission errors, the observations delivered at the receiver may not be fresh. Previous studies assumed that system performance degrades monotonically as observations become stale [7], [9], [13], [15], [23]. This assumption was justified for Markov sources [9]: However, recent machine learning experimental studies [24], [25], [27] showed that this monotonic assumption does not always hold. In certain scenarios, it was found that stale data with AoI > 0 can even achieve a smaller inference error than fresh data with AoI $= 0$, which is counter-intuitive. Information-theoretic tools was developed in [24], [25], [27] to interpret such non-monotonic information aging phenomena in machine learning experiments. To further understand in what scenarios information aging could be non-monotonic, we use a model-based approach to analyze information aging in this paper. Specifically, we derive closedform expressions for the remote estimation error of Gaussian autoregressive AR(p) processes and study how the monotonicity of information aging is affected by the parameters of the AR(p) process. The contributions of this paper are summarized as follows:

- We analyze the impact of fresh observations on the remote estimation of a p-th order Gaussian autoregressive process (AR(p)). The AR(p) process is widely used in modeling channel state information [30], economic forecasting [31], biomedical signals [32], and control systems [14], [33], [34]. Our study is more general than the earlier model-based studies in AoI literature [14], [33], [34], which are centered on AR(1) processes.
- The estimation error of the AR(p) process is formulated as a generalized conditional entropy (refer to Lemma 1). Closed-form expressions are provided for computing the estimation error (see Propositions 1-2). These expressions are provided for two commonly used loss functions in machine learning and remote estimation: (i) quadratic loss and (ii) logarithmic loss.
- By using a new information-theoretic tool called *ϵMarkov chain* [24]–[26], we evaluate the divergence of the AR(p) process from being Markovian. We then characterize the monotonicity of the estimation error with respect to AoI using the parameter ϵ (refer to Lemma 2). Specifically, if ϵ is close to zero, the target

Authorized licensed use limited to: Auburn University. Downloaded on April 25,2025 at 18:34:18 UTC from IEEE Xplore. Restrictions apply.

process is close to being Markov and the estimation error becomes a non-decreasing function of AoI; otherwise, if ϵ deviates significantly from zero, the target process is far from being Markov and the estimation error can exhibit highly non-monotonic behavior in AoI.

- A closed-form expression is provided to compute ϵ from AR(p) process (see Proposition 3(a)). Additionally, we characterize the parameter ϵ as a function of the observation time-sequence length. As the observation timesequence length of the AR(p) process increases to p, ϵ reduces to zero, and hence the estimation error becomes a non-decreasing function of AoI (See Proposition 3(b)). • Numerical results verify our theoretical findings (see Fig. 2 and Table I).

II. System Model

Consider the remote estimation system composed of a sensor, a transmitter, and an estimator, as illustrated in Fig. 1. The goal of the system is to estimate a time-varying target $Y_t \in \mathrm{R}$. We consider that the target Y_t evolves as

$$Y_t = X_t + N_t, \tag{2}$$

where $X_t \in \mathrm{R}$ follows a discrete-time p-th order autoregressive (AR(p)) linear time-invariant system:

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \ldots + a_p X_{t-p} + W_t, \tag{3}$$

$N_t \in \mathrm{R}$ and $W_t \in \mathrm{R}$ are i.i.d. Gaussian noises over time with zero mean, and $a_k \in \mathrm{R}$ for all $k = 1,2,\ldots,p$. Let $\sigma^2_{Y_t}$ and $\sigma^2_{X_t}$ be the variances of Y_t and X_t, respectively.

At every time slot t, the sensor observes X_t and feeds the observation to the transmitter. The transmitter progressively sends the sensory data to the estimator through a communication channel. Due to communication delays and channel errors, the delivered sensor observations may not be fresh. The most recently received sensor observation at the estimator is

$X_{\text{ence}_{t-\Delta(t)}}(\Delta(t)t)$that was generated at time$\in \mathrm{Z}_+$ between the generation time$t-\Delta(t)$. The time differ-$t-\Delta(t)$ and the

current time t is the AoI defined in (1). The estimator takes a consecutive sequence of sensor observations (also called

Rfeature sequence)$_l$ $(X_{t-\Delta(t)},X_{+t-}$as inputs and generates

an$_{\Delta(t)-1,\ldots,X_{t-\Delta(t)-l+1}}) \in$ and the AoI $\Delta(t) \in \mathrm{Z}$

output $a = \phi(\mathbf{X}^l_{t-\Delta(t)}, \Delta(t)) \in \mathcal{A}$, where $\mathbf{X}^l_{t-\Delta(t)} = [X_{t-\Delta(t)},X_{t-\Delta(t)-1},\ldots,X_{t-\Delta(t)-l+1}]$ is the feature sequence vector and the estimator is represented by the function $\phi : \mathrm{R}^l \times \mathrm{Z}^+ \to \mathrm{A}7$. The performance of the estimator is measured by a loss function $L : \mathrm{R} \times \mathrm{A}\,7 \to \mathrm{R}$, where $L(y,a)$ is the incurred loss if the output $a \in \mathrm{A}$ is used for estimation when $Y_t = y$. The loss function L is determined by the *goal* of the remote estimation system.

We assume that the age process $\{\Delta(t),t = 0,1,2,\ldots\}$ is signal-agnostic and the signal process $\{(Y_t,X_t),t = 0,1,\ldots\}$ is stationary. Under these assumptions, if $\Delta(t) = \delta$, then the minimum estimation error at time slot t can be expressed as a function of AoI δ and feature sequence length l [25], [27], given by

$$\mathrm{err}_{\text{estimation}}(\delta,l) := \min_{\phi\in\Phi} \mathbb{E}_{Y,\mathbf{X}^l\sim P_{Y_t,\mathbf{X}^l_{t-\delta}}} \left[L\left(Y, \phi\left(\mathbf{X}^l, \delta\right)\right)\right], \tag{4}$$

where the set of functions Φ consists of all functions that map from $\mathrm{R}^l \times \mathrm{Z}^+$ to A and $P_{Y_t,\mathbf{X}_{lt-\delta}}$ is the joint distribution of the target Y_t and the feature $\mathbf{X}^l_{t-\delta}$.

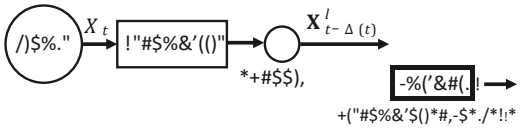


Fig. 1: A remote estimation system.

III. Monotonicity of Estimation Error with AoI: An ϵ-Markov Chain Approach

In this section, we employ an information-theoretic analysis and an ϵ-Markov chain model introduced in [24], [25], [27] to interpret how the estimation error $\mathrm{err}_{\text{estimation}}(\delta,l)$ varies with the AoI δ and the feature length l.

A. Information-theoretic Metrics for Estimation Error

We begin with the definitions of L-entropy and Lconditional entropy. The L-entropy of a random variable Y is defined as [25], [27], [35], [36]

$$H_L(Y) := \min_{a\in\mathrm{A}} \mathrm{E}_{Y\sim P_Y}[L(Y,a)]. \tag{5}$$

The L-conditional entropy of Y given X is defined as [25], [27], [35], [36]

$$H_L(Y|X) := \mathrm{E}_{x\sim P_X}[H_L(Y|X = x)], \tag{6}$$

where $H_L(Y|X = x)$ is given by

$$H_L(Y|X = x) = \min_{a\in\mathrm{A}} \mathrm{E}_{Y\sim P_{Y|X=x}}[L(Y,a)]. \tag{7}$$

Lemma 1. *Estimation error* $\mathrm{err}_{\text{estimation}}(\delta,l)$ *is equal to* L*conditional entropy of* Y_t *given* $\mathbf{X}^l_{t-\delta}$*, i.e.,*

Authorized licensed use limited to: Auburn University. Downloaded on April 25,2025 at 18:34:18 UTC from IEEE Xplore. Restrictions apply.

$$\text{err}_{\text{estimation}}(\delta, l) = H_L(Y_t|\mathbf{X}^l_{t-\delta}). \quad (8)$$

Proof. See Appendix A. □

Lemma 1 implies that we can directly use the Lconditional entropy $H_L(Y_t|\mathbf{X}^l_{t-\delta})$ to analyze the estimation error $\text{err}_{\text{estimation}}(\delta,l)$ of a remote estimation system. By directly using the properties of L-information theoretic metrics [25], [27], the estimation error can be analyzed conveniently. The information-theoretic metrics in the prior studies [8]–[11], [37] cannot be directly used to evaluate system performance.

B. Evaluating L-conditional Entropy

We will evaluate the L-conditional entropy associated with the loss function L. The loss function L is determined based on the objective of a remote estimation system. For example, in minimum mean-squared estimation (MMSE), the loss function is $L_2(y,\hat{y}) = (y-\hat{y})^2$, where the output $a = \hat{y}$ is an estimate of the target $Y_t = y$. In maximum likelihood estimation of the target distribution, the action $a = Q_{Y_t}$ is a distribution of Y_t and the loss function $L_{\log}(y,Q_{Y_t}) = -\log Q_{Y_t}(y)$ is the negative log-likelihood of the target value $Y_t = y$.

1) Logarithmic Loss (log loss): For log loss $L_{\log}(y,Q_{Y_t}) = -\log Q_{Y_t}(y)$, the L-entropy is the differential entropy [38], [39], defined as

$$H_{\log}(Y_t) = -\int_{y\in\mathbb{R}} p_{Y_t}(y)\, \log\, p_{Y_t}(y)\, \mathrm{d}\, y, \quad (9)$$

where p_{Y_t} is the density function of the distribution P_{Y_t} of Y_t. Because Y_t is a Gaussian random variable with zero mean, one can obtain [39]

$$H_{\log}(Y_t) = \frac{1}{2}\log\left(2\pi e\ \mathbb{E}[Y_t^2]\right). \quad (10)$$

The L-entropy for a discrete random variable associated with log loss is the well known Shannon entropy [25], [27], [36]. The Shannon entropy is always non-negative. However, the differential entropy can be negative, positive, and zero [39].

Proposition 1. *The L-conditional entropy* $H_{\log}(Y_t|\mathbf{X}^l_{t-\delta})$ *is given by*

$$H_{\log}(Y_t|\mathbf{X}^l_{t-\delta}) = \frac{1}{2}\log\left(\frac{\det(\mathbf{R}_{[Y_t,\mathbf{X}^l_{t-\delta}]})}{\det(\mathbf{R}_{\mathbf{X}^l_t})}\right) + \frac{1}{2}\log 2\pi e, \quad (11)$$

where $\det(\mathbf{A})$ *denotes the determinant of a square matrix* $\mathbf{A}$*,*

$$\mathbf{R}_{\mathbf{X}^l_t} = \mathbb{E}[(\mathbf{X}^l_t)^T\mathbf{X}^l_t] \quad (12)$$

is an $l \times l$ *dimensional auto-correlation matrix of a random vector* $\mathbf{X}^l_t$*, and*

$$\mathbf{R}^{[Y_t,\mathbf{X}^l_{t-\delta}]} = \mathbb{E}\left[[Y_t,\mathbf{X}^l_{t-\delta}]^T[Y_t,\mathbf{X}^l_{t-\delta}]\right] \quad (13)$$

is an $(l+1)\times(l+1)$ *dimensional auto-correlation matrix of a random vector* $[Y_t,\mathbf{X}^l_{t-\delta}] = [Y_t, X_{t-\delta},\ldots,X_{t-\delta-l+1}]$.

Proof. See Appendix B. □

In the special case of feature length $l = 1$, from (11), it can be shown that

$$H_{\log}(Y_t|X_{t-\delta}) = \frac{1}{2}\left(\log\left(\mathbb{E}[Y_t^2] - \frac{\mathbb{E}[X_tX_{t-\delta}]^2}{\mathbb{E}[X_t^2]}\right) + \log 2\pi e\right). \quad (14)$$

2) Quadratic Loss: For quadratic loss function $L_2(y,\hat{y}) = (y-\hat{y})^2$, the L-entropy of Y_t is the variance of Y_t, given by

$$H_2(Y_t) = \sigma^2_{Y_t}. \quad (15)$$

Because $\mathsf{E}[Y_t] = 0$, we have

$$H_2(Y_t) = \mathsf{E}[Y_t^2]. \quad (16)$$

Proposition 2. *The L-conditional entropy* $H_2(Y_t|\mathbf{X}^l_{t-\delta})$ *is given by*

$$H_2(Y_t|\mathbf{X}^l_{t-\delta}) = \mathbb{E}[(Y_t - \mathbb{E}[Y_t|\mathbf{X}^l_{t-\delta}])^2] = \mathbb{E}[Y_t^2] - \mathbb{E}[X_t\mathbf{X}^l_{t-\delta}](\mathbf{R}_{\mathbf{X}^l_t})^{-1}\mathbb{E}[X_t\mathbf{X}^l_{t-\delta}]^T, \quad (17)$$

where $\mathbb{E}[X_t\mathbf{X}^l_{t-\delta}] = \left[\mathbb{E}[X_tX_{t-\delta}],\ldots,\mathbb{E}[X_tX_{t-\delta-l+1}]\right]$ *is a* $1\times l$ *dimensional vector and* $\mathbf{R}_{\mathbf{X}^l_t}$ *is an* $l\times l$ *dimensional auto-correlation matrix of* $\mathbf{X}^l_t$ *defined in* (12)*.*

Due to space limitation, the proof of Proposition 2 is relegated to our technical report [40].

In the special case of feature length $l = 1$, from (17), it can be shown that

$$H_2(Y_t|X_{t-\delta}) = \mathbb{E}[Y_t^2] - \frac{\mathbb{E}[X_t\mathbb{X}^l_{t-\delta}]^2}{\mathbb{E}[X_t^2]}. \quad (18)$$

By utilizing Propositions 1-2, one can evaluate the Lconditional entropy of a data sequence that is generated using a Gaussian AR(p) system. The L-conditional entropy for an AR(4) model is depicted in Fig. 2. The model parameters of the AR(4) model is presented in Section V. Fig. 2 reveals that the L-conditional entropy can be a non-monotonic function of AoI δ.

C. L-conditional Entropy vs. AoI

If $Y_t \leftrightarrow \mathbf{X}^l_{t-\mu} \leftrightarrow \mathbf{X}^l_{t-\mu-\nu}$ is a Markov chain for all $\mu,\nu \geq 0$, by the data processing inequality [35, Lemma 12.1], $H_L(Y_t|\mathbf{X}^l_{t-\delta})$ is a non-decreasing function of δ. Nevertheless, the results in Fig. 2 show that the L-conditional entropy is not always a non-decreasing function

Authorized licensed use limited to: Auburn University. Downloaded on April 25,2025 at 18:34:18 UTC from IEEE Xplore. Restrictions apply.

of δ. This is because $Y_t \leftrightarrow \mathbf{X}_{t-\mu}^l \leftrightarrow \mathbf{X}_{t-\mu-\nu}^l$ is not a Markov chain for all $\mu,\nu \geq 0$, particularly when $l < p$ and $p \geq 2$ in AR(p) process. A relaxation of the data processing inequality is needed to analyze how $H_L(Y_t|\mathbf{X}_{t-\delta}^l)$ varies with δ for both Markovian and non-Markovian time-series data. To that end, an ϵ-Markov chain model is proposed in [25], [27].

Definition 1 (ϵ-Markov Chain). *Given $\epsilon \geq 0$, a sequence of three random variables Y,X, and Z is said to be an ϵ-*Markov chain*, denoted as $Y \leftrightarrow^{\epsilon} X \leftrightarrow^{\epsilon} Z$, if*

$$I_{\log}(Y;Z|X) = D_{\log}\left(P_{Y,X,Z}||P_{Y|X}P_{Z|X}P_X\right) \leq \epsilon^2, \quad (19)$$

where $I_{\log}(Y;Z|X)$ is Shannon conditional mutual information, $D_{\log}\left(P_{Y,X,Z}||P_{Y|X}P_{Z|X}P_X\right)$ is KL-divergence between two distributions $P_{Y,X,Z}$ and $P_{Y|X}P_{Z|X}P_X$, KL-divergence $D_{\log}(P_Y||Q_Y)$ between two distributions P_Y and Q_Y is defined as

$$D_{\log}(P_Y||Q_Y) := \int_{y\in\mathcal{Y}} p(y)\ \log\left(\frac{p(y)}{q(y)}\right) \mathrm{d}y, \quad (20)$$

p and q are the probability densities of P_Y and Q_Y, respectively.

Notice that the KL-divergence in (19) can be also equivalently expressed as

$$\begin{aligned} &D_{\log}(P_{Y,X,Z}||P_{Y|X}P_{Z|X}P_X) \\ &= \mathrm{E}_X[D_{\log}(P_{Y,Z|X}||P_{Y|X}P_{Z|X})] \\ &= \mathrm{E}_{X,Z}[D_{\log}(P_{Y|X,Z}||P_{Y|X})], \end{aligned} \quad (21)$$

Lemma 2. *The following assertions are true:*

(a) If $Y_t \leftrightarrow^{\epsilon} \mathbf{X}^l_{t-\mu} \leftrightarrow^{\epsilon} \mathbf{X}^l_{t-\mu-\nu}$ is an ϵ-Markov chain for every $\mu,\nu \geq 0$, then the L-conditional entropy is given by

$$H_L(Y_t|\mathbf{X}_{t-\delta}^l) = g_1(\delta) + O(\epsilon^2) \quad (22)$$

where $g_1(\delta)$ is a non-decreasing function of δ, given by

$$g_1(\delta) = H_L(Y_t|\mathbf{X}_t^l) + \sum_{k=0}^{\delta-1} I_L(Y_t;\mathbf{X}_{t-k}^l|\mathbf{X}_{t-k-1}^l), \quad (23)$$

the L-conditional mutual information $I_L(Y;X|Z)$ between Y and X given Z is

$$I_L(Y;X|Z) = H_L(Y|Z) - H_L(Y|X,Z). \quad (24)$$

(b) Given $\delta \geq 0$, $H_L(Y_t|\mathbf{X}_{t-\delta}^l)$ is a non-increasing function of feature length l, i.e., for all $1 \leq l_1 \leq l_2$,

$$H_L(Y_t|\mathbf{X}_{t-\delta}^{l_1}) \geq H_L(Y_t|\mathbf{X}_{t-\delta}^{l_2}). \quad (25)$$

Lemma 2 was introduced in our earlier work [25], [26] for discrete random variables. To ensure completeness of the paper, we restate Lemma 2 for continuous random variables.

According to Lemma 2(a), the monotonicity of $H_L(Y_t|\mathbf{X}_{t-\delta}^l)$ in δ is characterized by the parameter $\epsilon \geq 0$ in the ϵ-Markov chain model. If ϵ is close to zero, then $Y_t \overset{\epsilon}{\leftrightarrow} \mathbf{X}_{t-\mu}^l \overset{\epsilon}{\leftrightarrow} \mathbf{X}_{t-\mu-\nu}^l$ is close to a Markov chain, and $H_L(Y_t|\mathbf{X}_{t-\delta}^l)$ is non-decreasing in AoI δ. If ϵ is large, then $Y_t \overset{\epsilon}{\leftrightarrow} \mathbf{X}_{t-\mu}^l \overset{\epsilon}{\leftrightarrow} \mathbf{X}_{t-\mu-\nu}^l$ is far from a Markov chain, and $H_L(Y_t|\mathbf{X}_{t-\delta}^l)$ could be non-monotonic in AoI δ.

Lemma 2(b) states that $H_L(Y_t|\mathbf{X}^l_{t-\delta})$ decreases with increasing feature length l. A longer feature sequence adds more information that results in better estimation. Nevertheless, increasing the feature length also increases data size, necessitating more communication resources. For example, a longer feature sequence may require a longer transmission time and may end up being stale when delivered, thus resulting in worse inference performance. Recently, a study [26] has investigated a learning and communications co-design problem that jointly optimizes the timeliness and length of feature sequences.

IV. Characterizing the Parameter ϵ of An ϵ-Markov Chain

In this section, we show how to evaluate the value of the parameter ϵ from an AR(p) process. We also analyzed the impact of feature length l on the parameter ϵ.

The parameter ϵ in $Y_t \leftrightarrow^{\epsilon} \mathbf{X}^l_{t-\mu} \leftrightarrow^{\epsilon} \mathbf{X}^l_{t-\mu-\nu}$ depends on μ,ν, and l. We denote $\epsilon_{\mu,\nu}(l)$ as the minimum value of ϵ for which $Y_t \overset{\epsilon}{\leftrightarrow} \mathbf{X}_{t-\mu}^l \overset{\epsilon}{\leftrightarrow} \mathbf{X}_{t-\mu-\nu}^l$ is an ϵ-Markov chain. By using Definition 1, we have

$$\epsilon_{\mu,\nu}(l) = \sqrt{I_{\log}(Y_t;\mathbf{X}_{t-\mu-\nu}^l|\mathbf{X}_{t-\mu}^l)}. \quad (26)$$

We also denote $\epsilon(l)$ as the minimum value of ϵ for which $Y_t \leftrightarrow^{\epsilon} \mathbf{X}^l_{t-\mu} \leftrightarrow^{\epsilon} \mathbf{X}^l_{t-\mu-\nu}$ is an ϵ-Markov chain for all $\mu,\nu \geq 0$. Then, we can write

$$\epsilon(l) = \max_{\mu,\nu\geq 0} \epsilon_{\mu,\nu}(l). \quad (27)$$

Proposition 3. *The following assertions are true for the Gaussian AR(p) model defined in* (2)-(3)*.*

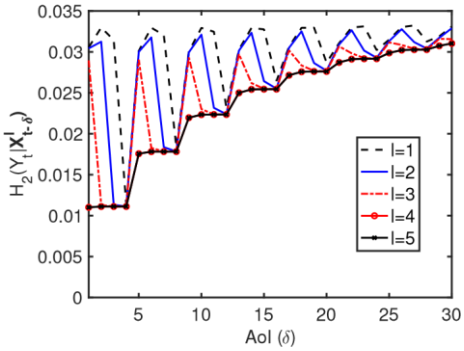
(a) The minimum value of ϵ for which the data sequence $(Y_t, \mathbf{X}_{t-\mu}^l, \mathbf{X}_{t-\mu-\nu}^l)$ satisfies an ϵ-Markov chain property, i.e., $Y_t \overset{\epsilon}{\leftrightarrow} \mathbf{X}_{t-\mu}^l \overset{\epsilon}{\leftrightarrow} \mathbf{X}_{t-\mu-\nu}^l$, for all $\mu,\nu \geq 0$ is given by

$$\epsilon(l) = \max_{\mu,\nu\geq 0} \epsilon_{\mu,\nu}(l), \quad (28)$$

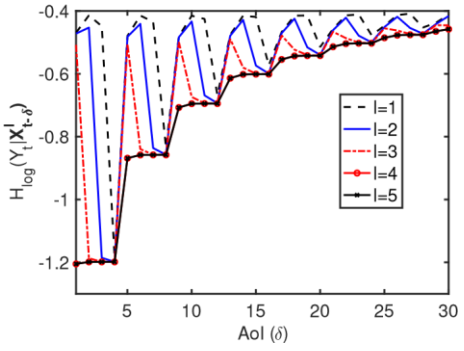
where $\epsilon_{\mu,\nu}(l)$ is determined by

$$\epsilon_{\mu,\nu}(l) = \sqrt{\frac{1}{2}\log\left(\frac{\det(\mathbf{R}_{[\mathbf{X}_{t-\mu-\nu}^l,\mathbf{X}_{t-\mu}^l]})\det(\mathbf{R}_{[Y_t,\mathbf{X}_{t-\mu}^l]})}{\det(\mathbf{R}_{\mathbf{X}_{t-\mu}^l})\det(\mathbf{R}_{[Y_t,\mathbf{X}_{t-\mu-\nu}^l,\mathbf{X}_{t-\mu}^l]})}\right)}, \quad (29)$$

Authorized licensed use limited to: Auburn University. Downloaded on April 25,2025 at 18:34:18 UTC from IEEE Xplore. Restrictions apply.



(a) $H_2(Y_t|\mathbf{X}^l_{t-\delta})$ vs. AoI (δ)



(b) $H_{\log}(Y_t|\mathbf{X}^l_{t-\delta})$ vs. AoI (δ)

Fig. 2: L-conditional entropy vs. AoI with (a) quadratic loss function and (b) log loss function (base 2). The L-conditional entropy is not always a monotonic function of AoI. An AR(4) model as discussed in Section V is considered for this simulation.

$\det(\mathbf{A})$ *denotes the determinant of a square matrix* $\mathbf{A}$*, and* $\mathbf{R_X} = \mathrm{E}[\mathbf{X}^T\mathbf{X}]$ *is the auto-correlation matrix of a random vector* $\mathbf{X}$.

(b) If $l \geq p$*, then*

$$\epsilon(l) = 0. \tag{30}$$

Due to space limitation, the proof of Proposition 3 is relegated to our technical report [40].

In Proposition 3(a), we present a closed-form expression for computing the parameter $\epsilon(l)$. Utilizing Proposition 3(a), one can derive $\epsilon(l)$ from the auto-correlation function of a data sequence generated from the AR(p) model. Proposition 3(b) implies that if the feature length l is greater than or equal to the order p of the AR(p) model, then $\epsilon(l)$ equals 0. By integrating Proposition 3(b) with Lemma 2, we can conclude that if the feature length l is greater than or equal to the order p, the L-conditional entropy becomes a non-decreasing function of AoI. However, transmitting longer features demands more communication resources [26].

V. Numerical Results

In this section, we utilize Proposition 1-3 to compute the estimation error and the parameter $\epsilon(l)$ for the following AR(4) process:

$$X_t = 0.1X_{t-1} + 0.8X_{t-p} + W_t, \tag{31}$$

$$Y_t = X_t + N_t, \tag{32}$$

where $W_t \in \mathrm{R}$ and $N_t \in \mathrm{R}$ are i.i.d. Gaussian noises over time with zero mean and variances 0.01 and 0.001, respectively. The goal is to estimate Y_t using a feature sequence $\mathbf{X}^l_{t-\delta} = [X_{t-\delta}, X_{t-\delta-1}, \ldots, X_{t-\delta-l+1}]$.

A. Evaluating L-conditional Entropy Using Propositions 1-2

We compute the L-conditional entropy of Y_t given $\mathbf{X}^l_{t-\delta}$ for two different loss functions: (a) quadratic loss and (b) log loss, using (17) and (11), respectively. In Fig. 2, we illustrate the L-conditional entropy $H_2(Y_t|\mathbf{X}^l_{t-\delta})$ associated with quadratic loss and the L-conditional entropy $H_{\log}(Y_t|\mathbf{X}^l_{t-\delta})$ associated with log loss (base 2). Both $H_2(Y_t|\mathbf{X}^l_{t-\delta})$ and $H_{\log}(Y_t|\mathbf{X}^l_{t-\delta})$ exhibit similar behavior with respect to AoI δ and feature length l, but they are measured in different scale and are used in different applications.

B. Evaluating $\epsilon(l)$ Using Proposition 3

We determine $\epsilon(l)$ through the following steps: Firstly, we calculate $\epsilon_{\mu,\nu}(l)$ using (29) given μ,ν, and l. Subsequently, we compute $\epsilon(l)$ by maximizing $\epsilon_{\mu,\nu}(l)$ over all $\mu,\nu \geq 0$. However, this needs to compute $\epsilon_{\mu,\nu}(l)$ for an infinite number of μ and ν, which is not possible. We find that when μ or ν exceed a large value, $\epsilon_{\mu,\nu}(l)$ becomes either 0 or close to 0 for all l. Therefore, we can choose an upper bound denoted as M and compute $\epsilon(l)$ by maximizing $\epsilon_{\mu,\nu}(l)$ over all $0 \leq \mu,\nu \leq M$. In our simulation, we set $M = 50$. The outcomes of $\epsilon(l)$ for feature length $l = 1,2,3,4,5$ are presented in Table I.

Feature length l	1	2	3	4	5
$\epsilon(l)$	1.55	1.49	1.39	0	0

TABLE I: $\epsilon(l)$ for $l = 1,2,3,4,5$.

C. Analysis of the Numerical Results

Fig. 2 and Table I illustrate that as the feature length l increases, the parameter $\epsilon(l)$ tends to zero, and the Lconditional entropy becomes a monotonic function of AoI δ. Specifically, when the feature length l reaches the order p of the AR(p) process, the parameter $\epsilon(l)$ equals zero and hence, the L-conditional entropy becomes a monotonic function of AoI. Moreover, as the feature length l increases, the Lconditional entropy reduces. However, beyond the order p, further increases in feature length do not result in the reduction of the L-conditional entropy. It is evident from

Authorized licensed use limited to: Auburn University. Downloaded on April 25,2025 at 18:34:18 UTC from IEEE Xplore. Restrictions apply.

Fig. 2 that the L-conditional entropy for $l = 4$ and $l = 5$ remains the same for the AR(4) model.

VI. Conclusion

This paper investigates the impact of information freshness on the remote estimation of AR(p) processes. Employing a new ϵ-Markov chain model, we demonstrate that the estimation error does not always degrade monotonically as the observations become stale. We provide closed-form expressions for computing both the estimation error and the parameter ϵ for AR(p) processes. Both theoretical analyses and numerical results illustrate that, with an increasing feature length, ϵ converges to zero and the estimation error converges to a nondecreasing function of AoI.

Appendix A Proof of Lemma 1

By using (5)-(7), we can obtain from (4) that

$$\begin{aligned}
&\mathrm{err}_{\mathrm{estimation}}(\delta,l)\\
&= \min_{\phi\in\Phi} \mathbb{E}_{Y,\mathbf{X}^l\sim P_{Y_t,\mathbf{X}^l_{t-\delta}}}\left[L\left(Y,\phi\left(\mathbf{X}^l,\delta\right)\right)\right]\\
&= \mathbb{E}_{\mathbf{x}^l\sim P_{\mathbf{X}^l_{t-\delta}}}\left[\min_{\phi(x^l,\delta)\in\mathcal{A}} \mathbb{E}_{Y\sim P_{Y_t|\mathbf{X}^l_{t-\delta}=\mathbf{x}^l}}\left[L\left(Y,\phi\left(\mathbf{x}^l,\delta\right)\right)\right]\right]\\
&= \mathbb{E}_{\mathbf{x}^l\sim P_{\mathbf{X}^l_{t-\delta}}}\left[\min_{a\in\mathcal{A}} \mathbb{E}_{Y\sim P_{Y_t|\mathbf{X}^l_{t-\delta}=\mathbf{x}^l}}\left[L(Y,a)\right]\right]\\
&= \mathbb{E}_{\mathbf{x}^l\sim P_{\mathbf{X}^l_{t-\delta}}}\left[H_L(Y_t|\mathbf{X}^l_{t-\delta}=\mathbf{x}^l)\right]\\
&= H_L(Y_t|\mathbf{X}^l_{t-\delta}),
\end{aligned} \tag{33}$$

where the second equality holds because Φ contains all functions that map from $\mathsf{R}^l \times \mathsf{Z}^+$ to A.

Appendix B Proof of Proposition 1

We begin with the definitions of *L-divergence* and *L-mutual information*. The L-divergence $D_L(P_Y||Q_Y)$ of P_Y from Q_Y can be expressed as [25], [36], [41]

$$\begin{aligned}
&D_L(P_Y||Q_Y)\\
&\quad = \mathsf{E}_{Y\sim P_Y}[L(Y,a_{P_Y})] - \mathsf{E}_{Y\sim P_Y}[L(Y,a_{Q_Y})],
\end{aligned} \tag{34}$$

where a_{P_Y} is the optimal solution to

$$\min_{a\in\mathsf{A}} \mathsf{E}_{Y\sim P_Y}[L(Y,a)]. \tag{35}$$

The *L-mutual information* $I_L(Y;X)$ is defined as [25], [36], [41]

$$\begin{aligned}
I_L(Y;X) &= \mathbb{E}_{X\sim P_X}\left[D_L\left(P_{Y|X}||P_Y\right)\right]\\
&= H_L(Y) - H_L(Y|X) \geq 0,
\end{aligned} \tag{36}$$

which measures the performance gain in estimating Y by observing X. The L-conditional mutual information $I_L(Y;X|Z)$ is given by

$$\begin{aligned}
I_L(Y;X|Z) &= \mathbb{E}_{X,Z\sim P_{X,Z}}\left[D_L\left(P_{Y|X,Z}||P_{Y|Z}\right)\right]\\
&= H_L(Y|Z) - H_L(Y|X,Z) \geq 0.
\end{aligned} \tag{37}$$

Using (36), the L-conditional entropy $H_{\log}(Y_t|\mathbf{X}^l_{t-\delta})$ associated with log loss can be expressed as

$$H_{\log}(Y_t|\mathbf{X}^l_{t-\delta}) = H_{\log}(Y_t) - I_{\log}(Y_t;\mathbf{X}^l_{t-\delta}). \tag{38}$$

For jointly Gaussian random vectors $\mathbf{Y} \in \mathsf{R}^m$ and $\mathbf{X} \in \mathsf{R}^n$, we can obtain [39]

$$I_{\log}(\mathbf{Y};\mathbf{X}) = \frac{1}{2}\log\frac{\det(\Sigma_{\mathbf{X}})\det(\Sigma_{\mathbf{Y}})}{\det(\Sigma_{[\mathbf{X},\mathbf{Y}]})}, \tag{39}$$

where $\Sigma_{\mathbf{X}} := \mathsf{E}[(\mathbf{X} - \mathsf{E}[\mathbf{X}])]^T\mathsf{E}[(\mathbf{X} - \mathsf{E}[\mathbf{X}])]$ denotes the covariance matrix of the row vector $\mathbf{X}$. If $\mathsf{E}[\mathbf{X}] = 0$, then $\Sigma_{\mathbf{X}} = \mathbf{R}_{\mathbf{X}}$.

By $\mathbb{E}[Y_t] = 0,\ \mathbb{E}[\mathbf{X}^l_{t-\delta}] = 0$ using, (10), (38), and (39), we obtain (11).

References

[1] X. Song and J. W.-S. Liu, "Performance of multiversion concurrency control algorithms in maintaining temporal consistency," in *IEEE Fourteenth Annual International Computer Software and Applications Conference*, 1990, pp. 132–133.

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

[3] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Select. Areas in Commun.*, vol. 39, no. 5, pp. 1183–1210, 2021.

[4] N. Pappas, M. A. Abd-Elmagid, B. Zhou, W. Saad, and H. S. Dhillon, *Age of Information: Foundations and Applications*. Cambridge University Press, 2022.

[5] 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 Trans. Inf. Theory*, vol. 63, no. 11, pp. 7492–7508, 2017.

[6] R. D. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *IEEE ISIT*, 2015, pp. 3008–3012.

[7] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraints," in *IEEE INFOCOM*, 2018, pp. 1844–1852.

[8] Y. Sun and B. Cyr, "Information aging through queues: A mutual information perspective," in *Proc. IEEE SPAWC Workshop*, 2018.

[9] ——, "Sampling for data freshness optimization: Non-linear age functions," *J. Commun. Netw.*, vol. 21, no. 3, pp. 204–219, 2019.

[10] G. Chen, S. C. Liew, and Y. Shao, "Uncertainty-of-information scheduling: A restless multiarmed bandit framework," *IEEE Trans. Inf. Theory*, vol. 68, no. 9, pp. 6151–6173, 2022.

[11] Z. Wang, M.-A. Badiu, and J. P. Coon, "A framework for characterizing the value of information in hidden markov models," *IEEE Trans. Inf. Theory*, vol. 68, no. 8, pp. 5203–5216, 2022.

[12] T. Z. Ornee and Y. Sun, "Sampling and remote estimation for the Ornstein-Uhlenbeck process through queues: Age of information and beyond," *IEEE/ACM Trans. Netw.*, vol. 29, no. 5, pp. 1962–1975, 2021.

[13] V. Tripathi and E. Modiano, "A Whittle index approach to minimizing functions of age of information," in *IEEE Allerton*, 2019, pp. 1160–1167.

[14] M. Klugel, M. H. Mamduhi, S. Hirche, and W. Kellerer, "AoI-penalty¨ minimization for networked control systems with packet loss," in *IEEE INFOCOM Age of Information Workshop*, 2019, pp. 189–196.

[15] A. M. Bedewy, Y. Sun, S. Kompella, and N. B. Shroff, "Optimal sampling and scheduling for timely status updates in multi-source networks," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 4019–4034, 2021.

[16] J. Sun, Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Closed-form Whittle's index-enabled random access for timely status update," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1538–1551, 2019.

[17] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2637–2650, 2018.

Authorized licensed use limited to: Auburn University. Downloaded on April 25,2025 at 18:34:18 UTC from IEEE Xplore. Restrictions apply.

[18] T. Z. Ornee and Y. Sun, "A Whittle index policy for the remote estimation of multiple continuous Gauss-Markov processes over parallel channels," *ACM MobiHoc*, 2023.

[19] J. Pan, Y. Sun, and N. B. Shroff, "Sampling for remote estimation of the Wiener process over an unreliable channel," *ACM Sigmetrics*, 2023.

[20] Y. Sun and S. Kompella, "Age-optimal multi-flow status updating with errors: A sample-path approach," *J. Commun. Netw.*, vol. 25, no. 5, pp. 570–584, 2023.

[21] Y. Sun, I. Kadota, R. Talak, and E. Modiano, *Age of information: A new metric for information freshness*. Springer Nature, 2020.

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

[23] V. Tripathi, L. Ballotta, L. Carlone, and E. Modiano, "Computation and communication co-design for real-time monitoring and control in multiagent systems," in *IEEE WiOpt*, 2021, pp. 1–8.

[24] M. K. C. Shisher, H. Qin, L. Yang, F. Yan, and Y. Sun, "The age of correlated features in supervised learning based forecasting," in *IEEE INFOCOM Age of Information Workshop*, 2021.

[25] M. K. C. Shisher and Y. Sun, "How does data freshness affect real-time supervised learning?" *ACM MobiHoc*, 2022.

[26] M. K. C. Shisher, B. Ji, I.-H. Hou, and Y. Sun, "Learning and communications co-design for remote inference systems: Feature length selection and transmission scheduling," *IEEE J. Sel. Areas Inf. Theory*, vol. 4, pp. 524–538, 2023.

[27] M. K. C. Shisher, Y. Sun, and I.-H. Hou, "Timely communications for remote inference," *submitted*, 2023.

[28] T. Z. Ornee, M. K. C. Shisher, C. Kam, and Y. Sun, "Contextaware status updating: Wireless scheduling for maximizing situational awareness in safety-critical systems," in *IEEE MILCOM*, 2023, pp. 194– 200.

[29] C. Ari, M. K. C. Shisher, E. Uysal, and Y. Sun, "Goal-oriented communications for remote inference with two-way delay," *arXiv preprint arXiv:2311.11143*, 2023.

[30] W. C. Jakes and D. C. Cox, *Microwave mobile communications*. WileyIEEE press, 1994.

[31] J. H. Stock and M. W. Watson, "Vector autoregressions," *Journal of Economic perspectives*, vol. 15, no. 4, pp. 101–115, 2001.

[32] A. Isaksson, A. Wennberg, and L. H. Zetterberg, "Computer analysis of eeg signals with parametric models," *Proceedings of the IEEE*, vol. 69, no. 4, pp. 451–461, 1981.

[33] J. P. Champati, M. H. Mamduhi, K. H. Johansson, and J. Gross, "Performance characterization using aoi in a single-loop networked control system," in *IEEE INFOCOM Age of Information Workshop*, 2019, pp. 197–203.

[34] O. Ayan, M. Vilgelm, M. Klugel, S. Hirche, and W. Kellerer, "Age-of-¨ information vs. value-of-information scheduling for cellular networked control systems," in *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 2019, pp. 109–117.

[35] A. P. Dawid, "Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design," *Technical Report 139*, 1998.

[36] F. Farnia and D. Tse, "A minimax approach to supervised learning," *NIPS*, vol. 29, pp. 4240–4248, 2016.

[37] T. Soleymani, S. Hirche, and J. S. Baras, "Optimal self-driven sampling for estimation based on value of information," in *IEEE WODES*, 2016, pp. 183–188.

[38] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 1999.

[39] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes for MIT (6.441), UIUC (ECE 563), Yale (STAT 664)*, no. 20122017, 2014.

[40] M. K. C. Shisher and Y. Sun, "On the monotonicity of information aging," *Technical Report*, 2024, arXiv:2403.03380.

[41] P. D. Grunwald 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.

Authorized licensed use limited to: Auburn University. Downloaded on April 25,2025 at 18:34:18 UTC from IEEE Xplore. Restrictions apply.