Sample, Quantize, and Encode: Timely Estimation Over Noisy Channels

Ahmed Arafa, *Member, IEEE*, Karim Banawan, *Member, IEEE*, Karim G. Seddik, *Senior Member, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*— The effects of *quantization* and *coding* on the estimation quality of Gauss-Markov processes are considered, with a special attention to the Ornstein-Uhlenbeck process. Samples are acquired from the process, quantized, and then encoded for transmission using either *infinite incremental redundancy* (IIR) or *fixed redundancy* (FR) coding schemes. A fixed *processing* time is consumed at the receiver for decoding and sending feedback to the transmitter. Decoded messages are used to construct a minimum mean square error (MMSE) estimate of the process as a function of time. This is shown to be an increasing functional of the *age-of-information* (AoI), defined as the time elapsed since the sampling time pertaining to the latest successfully decoded message. Such functional depends on the quantization bits, codewords lengths and receiver processing time. The goal, for each coding scheme, is to optimize sampling times such that the long-term average MMSE is minimized. This is then characterized in the setting of *general increasing functionals of AoI,* not necessarily corresponding to MMSE, which may be of independent interest in other contexts. We first show that the optimal sampling policy for IIR is such that a new sample is generated only if the AoI exceeds a certain *threshold,* while for FR it is such that a new sample is delivered *just-in-time* as the receiver finishes processing the previous one. *Enhanced* transmissions schemes are then developed in order to exploit the processing times to make new data available at the receiver sooner. For both IIR and FR, it is shown that there exists an optimal number of quantization bits that balances AoI and quantization errors, and hence minimizes the MMSE. It is also shown that for longer receiver processing times, the relatively simpler FR scheme outperforms IIR.**

***Index Terms*— Ornstein-Uhlenbeck process, general age-penalty functional, infinite incremental redundancy, fixed redundancy, receiver processing time.**

Manuscript received November 17, 2020; revised April 23, 2021; accepted June 15, 2021. Date of publication June 25, 2021; date of current version October 18, 2021. This work was supported in part by the U.S. National Science Foundation under Grant CCF-1908308. This article was presented in part at the 2020 International Symposium of Information Theory (ISIT), Los Angeles, CA, June 2020 [1]. The associate editor coordinating the review of this article and approving it for publication was A. Cohen. *(Corresponding author: Ahmed Arafa.)*

Ahmed Arafa is with the Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: aarafa@uncc.edu).

Karim Banawan is with the Department of Electrical Engineering, Alexandria University, Alexandria 21526, Egypt (e-mail: kbanawan@alexu.edu.eg).

Karim G. Seddik is with the Electronics and Communications Engineering Department, The American University in Cairo, New Cairo 11835, Egypt (e-mail: kseddik@aucegypt.edu).

H. Vincent Poor is with the Electrical and Computer Engineering Department, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2021.3092413.

Digital Object Identifier 10.1109/TCOMM.2021.3092413

I. Introduction

RECENT works have drawn connections between remote estimation of a time-varying process and the *age-of-information* (AoI) metric, which assesses the timeliness and freshness of the estimated data. While most works focus on transmitting *analog* samples for the purpose of estimation, this work focuses on using *quantized* and *coded* samples in that regard. We present optimal sampling methods that minimize the long-term average minimum mean square error (MMSE) of a Gauss-Markov, namely Ornstein-Uhlenbeck (OU), process under specific coding schemes, taking into consideration receiver *processing* times consumed in decoding and sending feedback. The OU process is the continuous-time analogue of the first-order autoregressive process [2], [3], and is used to model various physical phenomena, and has relevant applications in control and finance. Our goal in this work is to devise practical sampling and coding schemes for the purpose of real-time tracking of OU processes while taking into consideration the effects of quantization, coding delays, and receiver processing times.

AoI, or merely age, is a time-based metric that measures information freshness by capturing delay from the receiver's perspective; it is defined as the time elapsed since the latest received data at the destination has been generated at its source. Hence, in general, to keep the data fresh, one needs to keep the AoI low. An increasing number of works in the recent literature have used AoI as a latency performance metric in various contexts. These include queuing-theoretic analyses of AoI for single and multiple sources [4]–[12], scheduling and sampling for AoI minimization [13]–[17], status updating under energy harvesting constraints [18]–[23], AoI analysis in multihop networks [24], [25], source coding for AoI minimization [26], and using AoI in other applications such as fresh data pricing [27], cloud computing [28] and federated learning [29], among others, see the recent survey in [30].

There are two main lines of research in the AoI literature that relate to this work. The first is the one pertaining to coding over noisy channels for age minimization, e.g., [31]–[44]. These works can be categorized according to the structure of the code being used to transmit the samples; references [31]–[38] focus on analyzing the usage of (rateless) infinite incremental redundancy (IIR) and fixed redundancy (FR) coding schemes and determined conditions in which both perform relatively well; the works in [39]–[42] analyze the usage of

hybrid ARQ (HARQ) coding schemes for AoI minimization; while those in [43], [44] consider broadcast multi-user settings. In IIR schemes, the transmitter sends its messages using a codewords of some original length, and then adds incremental redundancy (IR) bits one by one when signaled by the receiver until successful decoding is accomplished. This may potentially take a very large number of IR bit transmissions, hence the name IIR.[1] In FR schemes, the transmitter sends its messages using fixed-length codewords, with retransmissions in case of decoding failures, i.e., without adding IR bits. HARQ schemes feature an initial transmission followed by subsequent transmissions (of possibly varying lengths) of IR that are guided by feedback from the receiver to the transmitter, but not necessarily at a granularity of a single bit as in IIR. One main theme in the findings of works [31]–[44] is that optimal codes should strike a balance between using long codewords to minimize channel errors and using short ones to minimize age. Our work in this paper primarily focuses on evaluating the performances of using IIR and FR coding schemes. However, different from all the works in [31]–[38], *we consider the additional presence of fixed non-zero receiver processing times.*

The second line of research related to this work is related to evaluating the role of AoI in remote estimation, e.g., [48]–[58]. The works in [48]–[50] characterize implicit and explicit relationships between mean square error (MSE) and AoI under different estimation contexts; references [51], [52] consider the notion of the value of information (mainly through MSE) and show that optimizing it can be different from optimizing AoI; lossy source coding and distorted updates for AoI minimization is considered in [53]–[55]; reference [56] adds more context to AoI by introducing and analyzing a variant metric termed the age of incorrect information (AoII) to capture error in updates; while the works in [57], [58] consider sampling of Wiener and OU processes for the purpose of remote estimation, and draw connections between MSE and AoI. Our work in this paper also focuses on characterizing the relationship of MSE and AoI, *yet with the additional presence of quantization errors.* It is worth noting that while studying optimal sampling with distortion guarantees is a classical problem, it has been recently approached differently in [59], which characterizes the minimal sampling frequency required to achieve Shannon's rate-distortion function, and concludes that sub-Nyquist sampling can attain the fundamental rate-distortion tradeoff if the energy spectral density of the signal is non-uniform (see [59] and the references therein).

While AoI is a time-based metric that has been originally studied in queuing-theoretic frameworks to assess latency, e.g., [4]–[6], it is relatively easier to analyze for process tracking purposes compared to MSE, since AoI only takes the statistics of the communication channel into consideration, unlike MSE that also takes the statistics of the process itself into account to assess the *quality* of tracking. Under some assumptions, MSE can be shown to have a very dependent behavior on AoI, and hence, minimizing AoI becomes equivalent to minimizing MSE. This is one main idea around which this work revolves, and has been the focus of the works in [57], [58], which are the most closely-related works to ours. References [57], [58] derive optimal sampling methods to minimize the long-term average MMSE for Wiener [57] and OU [58] processes. In both works, the communication channel introduces random delays, before perfect (distortion-free) samples are received. It is shown that if sampling times are independent of the instantaneous values of the process (signal-independent sampling) the MMSE reduces to AoI in case of Wiener [57], and to an increasing functional of AoI (age-penalty) in case of OU [58]. It is then shown that the optimal sampling policy has a threshold structure, in which a new sample is acquired only if the expected AoI in case of Wiener (or age-penalty in case of OU) surpasses a certain value. In addition, signal-dependent optimal sampling policies are also derived [57], [58].

In this work, we consider the transmission of quantized and coded samples of an OU process through a noisy channel. We note that we consider an OU process in our study since, unlike the conventional Wiener process, it has a bounded variance, leading to bounded quantization error as well. Different from [58], not every sample has guaranteed reception, and received samples suffer from quantization noise. The receiver uses the received samples to construct an MMSE estimate for the OU process. Quantization and coding introduce a tradeoff: *few quantization levels and codeword bits would transmit samples faster, yet with high distortion and probability of error.* An optimal choice, therefore, needs to be made, which depends mainly on how fast the OU process varies as well as the channel errors. Different from related works, effects of having (fixed) *non-zero receiver processing times*, mainly due to decoding and sending feedback, are also considered in this work.

We focus on signal-independent sampling, together with an MMSE quantizer, combined with either IIR or FR coding schemes; see Fig. 1. The MMSE of the OU process is first shown to be an increasing functional of AoI, as in [58], parameterized directly by the number of quantization bits ℓ, and indirectly by the number of codeword bits n and the receiver processing time β. We formulate two problems, one for IIR and another for FR, to choose sampling times so that the long-term average MMSE is minimized. Focusing on stationary deterministic policies, we present optimal solutions for both problems in the case of *general increasing age-penalties,* not necessarily corresponding to MMSE, which may be useful in other contexts in which IIR and FR coding schemes are employed. The solution for IIR has a *threshold* structure, as in [16], [58], while that for FR is a *just-in-time* sampling policy that does not require receiver feedback.

We then present what we call *enhanced* IIR and FR schemes, in which we leverage the processing time to our favor through fine-tuning sampling and/or transmission times in such a way that *the receiver never waits for data when necessary.* This allows us to mitigate the negative effects of processing

[1] A clear example of the IIR scheme is the family of fountain (rateless) codes. In a rateless code, the encoder produces limitless (potentially infinite) stream of coded symbols based on the ℓ input symbols. The decoder reconstructs the ℓ bits after receiving *any* n correct symbols. One common rateless code is the systematic Raptor code in [45], which is used in the 3GPP multimedia broadcast multicast services (MBMS), DVB-H IPDC, and DVB-IPTV [46], [47].

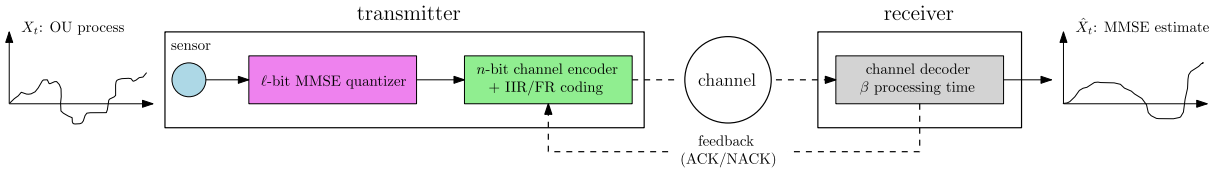


Fig. 1. System model considered for sampling, quantizing and encoding an OU process at the transmitter, and reconstructing it at the receiver.

times to the most extent possible, and produce timely estimates that are able to track the OU process better. We finally discuss how to select ℓ and n, and show that the relatively simpler FR scheme can outperform IIR for relatively large values of β.

The proposed joint optimization of sampling, quantization and coding in this paper takes a step towards achieving the notion of timely real-time tracking of random processes, which can be applied in applications of communications, networks and control. We summarize our main contributions as follows:

- Presenting a thorough analysis of the effects of quantization and coding (with specific focus on IIR and FR) on the estimation error of Gauss-Markov processes (with specific focus on the OU process). Specifically, we show that there is an inherent relationship between the number of quantization levels and codeword lengths used to convey the samples and the OU process statistics (in particular how fast it varies over time).
- Characterizing the optimal (signal-independent) sampling strategy (MSE-minimal) for IIR and FR in this context.
- Introducing, *for the first time in the AoI literature (to the best of our knowledge),* the effects of non-zero processing delays at the receiver (for decoding and sending feedback), based on which we argue that one can enhance the performance of both IIR and FR by carefully tailoring the transmission times to the processing delays.
- Validating our theoretical results by conducting multiple numerical studies and presenting examples that show the effects of the OU process statistics on the optimal quantization levels and coding lengths.

Compared to the conference version [1], this paper adds: (1) novel thorough analyses of the enhanced schemes in Section IV; (2) complete proofs for results and formulas that were omitted in [1]; and (3) multiple numerical studies that showcase the main results and intuitions.

II. System Model

A. Quantization and Coding of the OU Process

We consider a sensor that acquires time-stamped samples from an OU process. Given a value of X_s at time s, the OU process evolves as follows [2], [3]:

$$X_t = X_s e^{-\theta(t-s)} + \frac{\sigma}{\sqrt{2\theta}} e^{-\theta(t-s)} W_{e^{2\theta(t-s)}-1}, \quad t \geq s, \tag{1}$$

where W_t denotes a Wiener process, while $\theta > 0$ and $\sigma > 0$ are fixed parameters. The sensor acquires the ith sample at time S_i and feeds it to an MMSE quantizer that produces an ℓ-bit message ready for encoding. We will use the term *message* to refer to a quantized sample of the OU process. Let $\tilde{X}_{S_i}$ represent the quantized version of the sample X_{S_i}, and let Q_{S_i} denote the corresponding quantization error. Thus,

$$X_{S_i} = \tilde{X}_{S_i} + Q_{S_i}. \tag{2}$$

Each message is encoded and sent over a noisy channel to the receiver. The receiver updates an MMSE estimate of the OU process if decoding is successful. ACKs and NACKs are fed back following each decoding attempt. A fixed receiver processing time β time units is incurred per each decoding attempt, which also includes the time to generate and send feedback. Channel errors are independent and identically distributed (i.i.d.) across time/messages.

Two channel coding schemes are investigated. The first is IIR, in which a message transmission starts with an n-bit codeword, $n \geq \ell$, and then incremental redundancy (IR) bits are added one-by-one if a NACK is received until the message is eventually decoded and an ACK is fed back. The second scheme is FR, in which a message is encoded into fixed n-bit codewords, yet following a NACK the message in transmission is discarded and a *new* sample is acquired and used instead. Following ACKs, the transmitter may idly wait before acquiring a new sample and sending a new message.[2]

B. Communication Channel

Let D_i denote the reception time of the ith *successfully decoded* message. For the IIR scheme, each message is eventually decoded, and therefore

$$D_i = S_i + Y_i \tag{3}$$

for some random variable Y_i that represents the channel delay incurred due to the IR bits added. Let T_b denote the time units consumed per bit transmission. Hence,

$$Y_i = nT_b + \beta + r_i(T_b + \beta), \tag{4}$$

where the random variable $r_i \in \{0, 1, 2, \dots\}$ denotes the number of IR bits used until the ith message is decoded. Note that in the IIR scheme β is added for the original n-bit codeword transmission, and then for each IR transmission until successful decoding. Let

$$\bar{n} \triangleq nT_b + \beta \tag{5}$$

[2]The main reason behind waiting, as will be shown in details in the sequel, is that it leads to sending fresher samples, which can be more rewarding in terms of the *long-term average* MMSE, and *not* the instantaneous MMSE. Note that waiting policies have been generously used in previous works that focus on minimizing average AoI, e.g., [14], [18], [21].

for conciseness. Channel delays Y_i's are i.i.d. $\sim Y$, where

$$\mathbb{P}\left(Y=\bar{n}\right)=p_0, \tag{6}$$

$$\mathbb{P}\left(Y=\bar{n}+k(T_b+\beta)\right)=\prod_{j=0}^{k-1}(1-p_j)p_k, \quad k\geq 1, \tag{7}$$

with p_j denoting the probability that an ACK is received when $r_i=j$. This depends on the channel code being used, and the model of the channel errors, yet it holds that $p_j\leq p_{j+1}$.

For the FR scheme, there can possibly be a number of transmission *attempts* before a message is eventually decoded. Let M_i denote the number of these attempts in between the $(i-1)$th and ith successfully decoded messages, and let $S_{i,j}$ denote the sampling time pertaining to the jth attempt of which, $1\leq j\leq M_i$. Therefore, only the M_ith message is successfully decoded, and the rest are all discarded. Since each message is encoded using fixed n-bit codewords, we have

$$D_i=S_{i,M_i}+\bar{n}, \quad \forall i. \tag{8}$$

Observe that in the FR scheme each successfully-decoded message incurs only *one* β, since each decoding attempt occurs on a message pertaining to a *fresh* sample. According to the notation developed for the IIR channel delays above, M_i's are i.i.d. geometric random variables with parameter p_0.

C. MMSE Estimation and AoI

Based on the above notation so far, the AoI at time t is mathematically defined as follows:

$$\text{AoI}(t)\triangleq t-u_i(t), \quad D_i\leq t<D_{i+1}, \tag{9}$$

where $u_i(t)$ denotes the time stamp of the latest received sample before time t. Thus, for $D_i\leq t<D_{i+1}$, we have $u_i(t)=S_i$ for the IIR scheme, and $u_i(t)=S_{i,M_i}$ for the FR scheme.

Upon successfully decoding a message at time D_i, the receiver constructs an MMSE estimate for the OU process. For the purpose of real-time tracking, do not allow retroactive reconstruction of the process, and restrict our attention to MMSE estimators that only use the latest-received information.[3] For the IIR scheme this is

$$\hat{X}_t=\mathbb{E}\left[X_t\middle|S_i,\tilde{X}_{S_i}\right], \quad D_i\leq t<D_{i+1}. \tag{10}$$

Using (1) and (2), we have

$$\begin{aligned}\hat{X}_t=\mathbb{E}\Bigg[&\tilde{X}_{S_i}e^{-\theta(t-S_i)}+Q_{S_i}e^{-\theta(t-S_i)}\\&+\frac{\sigma}{\sqrt{2\theta}}e^{-\theta(t-S_i)}W_{e^{2\theta(t-S_i)}-1}\Bigg|S_i,\tilde{X}_{S_i}\Bigg] \qquad (11)\\=\tilde{X}_{S_i}&e^{-\theta(t-S_i)}, \quad D_i\leq t<D_{i+1}, \qquad (12)\end{aligned}$$

where the last equality follows by independence of the Wiener noise in $[D_i,t]$ from $(S_i,\tilde{X}_{S_i})$, and that for the MMSE quantizer, the quantization error is zero-mean [60]. The MMSE is now given as follows for $D_i\leq t<D_{i+1}$:

$$\begin{aligned}\texttt{mse}\left(t,S_i\right)&=\mathbb{E}\left[\left(X_t-\hat{X}_t\right)^2\right] \qquad (13)\\&=\mathbb{E}\left[Q_{S_i}^2\right]e^{-2\theta(t-S_i)}+\frac{\sigma^2}{2\theta}\left(1-e^{-2\theta(t-S_i)}\right). \qquad (14)\end{aligned}$$

We see from the above that even if $D_i-S_i=0$, i.e., if the ith sample is transmitted and received instantaneously, the MMSE estimate at $t=D_i$ would still suffer from quantization errors.

In the sequel, we consider $X_0=0$ without loss of generality, and hence, using (1), the variance of X_t is given by $\mathbb{E}\left[X_t^2\right]=\frac{\sigma^2}{2\theta}\left(1-e^{-2\theta t}\right)$, $t>0$. Following a rate-distortion approach (note that X_t is Gaussian), the following relates the number of bits ℓ and the instantaneous mean square quantization error [60][4]:

$$\mathbb{E}\left[Q_t^2\right]=\frac{\sigma^2}{2\theta}\left(1-e^{-2\theta t}\right)2^{-2\ell}, \quad t>0. \tag{15}$$

Using the above in (14) and rearranging, we get that

$$\texttt{mse}(t,S_i)=\frac{\sigma^2}{2\theta}\Big(1-\big(1-2^{-2\ell}\big(1-e^{-2\theta S_i}\big)\big)e^{-2\theta(t-S_i)}\Big), \tag{16}$$

We note that as $\ell\to\infty$, the above expression becomes the same as that derived for the signal-independent sampling scheme analyzed in [58]. However, since we consider practical coding aspects in this work, as $\ell\to\infty$, it holds that $n\to\infty$ as well and no sample will be received.

We focus on dealing with the system in *steady state,* in which both t and S_i are relatively large. In this case, the mean square quantization error in (15) becomes independent of time, and only dependent upon the steady state variance of the OU process $\sigma^2/2\theta$.[5] Hence, in steady state, the MMSE becomes

$$\begin{aligned}\texttt{mse}\left(t,S_i\right)&=\frac{\sigma^2}{2\theta}\left(1-\left(1-2^{-2\ell}\right)e^{-2\theta(t-S_i)}\right) \qquad (17)\\&\triangleq h_\ell\left(t-S_i\right), \quad D_i\leq t<D_{i+1}, \qquad (18)\end{aligned}$$

which is an increasing functional of the AoI $t-S_i$ in (9). One can see from the MMSE expression above that there exists a tension between the number quantization levels and AoI. In particular, as ℓ increases, the quantization noise decreases and the samples transmitted become more precise. However, this necessiates using a larger codeword length n, which in turn increases the age-penalty. Hence, a tradeoff exists between sending slow but precise samples and fast but less accurate ones. We discuss how to optimally characterize this tradeoff in Section V.

For the FR scheme, the analysis follows similarly, after adding one more random variable denoting the number of

[3]Note that the OU process is no longer Markov after quantization. The implication of this is that the MMSE estimator in (10) is potentially suboptimal since it focuses only on the latest received sample. It is, however, simple enough in practice, and admits the analytical solutions derived in the paper. Deriving an optimal MMSE estimator, or showing that considering only the latest received quantized sample performs well enough, e.g., close to optimal, is to be pursued in future work.

[4]There are other works in the literature that consider different kinds of (practical) quantizers and study their effects on filtering stationary Gaussian processes, see, e.g., the uniform quantizer treatment in [61], which may lead to different quantization errors statistics. Our setting naturally focuses on MMSE quantizers, which are relevant to the purpose of MMSE estimation.

[5]Equivalently, one can initiate the OU process by $X_0\sim\mathcal{N}\left(0,\frac{\sigma^2}{2\theta}\right)$, whence $\mathbb{E}\left[X_t^2\right]=\frac{\sigma^2}{2\theta}$, $\forall t$.

transmissions, M_i. Specifically, it holds that

$$\hat{X}_t = \tilde{X}_{S_{i,M_i}} e^{-\theta\left(t-S_{i,M_i}\right)}, \tag{19}$$

$$\mathtt{mse}\left(t, S_{i,M_i}\right) = h_\ell\left(t - S_{i,M_i}\right), \quad D_i \le t < D_{i+1}. \tag{20}$$

We see from (18) and (20) that there are two main contributing factors to the MMSE. The first is due to quantization, represented by the factor $\left(1-2^{-2\ell}\right)$, and the second is due to the channel delay, added mainly because of coding and errors, represented by the AoI $t-S$.

III. Optimal Sampling Policies: General Age-Penalty

The main goal is to choose the sampling times, for given ℓ, n and β, such that the long-term average MMSE is minimized. In this section, we formulate two problems to achieve such goal: one for IIR and another for FR, and present their solutions in the upcoming section. Later on in Section V, we discuss how to choose the best ℓ and n, as well as compare the performances of IIR and FR in general.

For both coding schemes, let us denote by an *epoch* the time elapsed in between two successfully received messages. Thus, the ith epoch starts at D_{i-1} and ends at D_i, with $D_0 \equiv 0$.

Remark 1: *Our analysis does not depend on the specific structure of the MMSE functional $h_\ell(\cdot)$; it extends to any differentiable increasing age-penalty functional $g(\cdot)$. Therefore, in what follows, we formulate our problems and present their solutions for the case of minimizing a long-term average age-penalty, making the results applicable in other contexts.*

A. The IIR Scheme

For the IIR scheme, the problem is formulated as

$$\min_{\{S_i\}} \quad \limsup_{l\to\infty} \frac{\sum_{i=0}^{l} \mathbb{E}\left[\int_{D_i}^{D_{i+1}} g\left(t-S_i\right) dt\right]}{\sum_{i=0}^{l} \mathbb{E}\left[D_{i+1}-D_i\right]}, \tag{21}$$

where the numerator represents the total age-penalty (the MMSE in case of the OU process estimation) across all epochs, and the denominator represents the total time.

Let us define W_i as the waiting time at the beginning of the ith epoch before acquiring the ith sample. That is, $S_i = D_{i-1} + W_i$. Therefore, one can equivalently solve for the waiting times W_i's instead of sampling times S_i's. We focus on a class of *stationary deterministic* policies in which

$$W_i = f\left(g\left(D_{i-1}-S_{i-1}\right)\right), \quad \forall i. \tag{22}$$

That is, *the waiting time in the ith epoch is a deterministic function of its starting age-penalty value.* Such focus is motivated by the fact that channel errors are i.i.d. and by its optimality in similar frameworks, e.g., [14], [19], [21]. Defining $w \triangleq f \circ g$ and noting that $D_{i-1} - S_{i-1} = Y_{i-1}$ we have

$$W_i = w\left(Y_{i-1}\right), \tag{23}$$

which induces a stationary distribution of D_i's and the age-penalty across all epochs. Due to stationarity, we can now drop the epoch's index i, and (re)define notations used in a typical epoch. It starts at time $\overline{D}$ with AoI $\overline{Y}$, and with the latest sample acquired at time $\overline{S}$, such that $\overline{D} = \overline{S} + \overline{Y}$. Then, a waiting time of $w\left(\overline{Y}\right)$ follows, after which a new sample is acquired, quantized, and transmitted, taking Y time units to reach the receiver at time $D = \overline{D} + w\left(\overline{Y}\right) + Y$, which is the epoch's end time. Therefore, problem (21) now reduces to a minimization over a single epoch as follows:

$$\min_{w(\cdot)\ge 0} \quad \frac{\mathbb{E}\left[\int_{\overline{D}}^{\overline{D}+w\left(\overline{Y}\right)+Y} g\left(t-\overline{S}\right) dt\right]}{\mathbb{E}\left[w\left(\overline{Y}\right)+Y\right]}. \tag{24}$$

Given the realization of $\overline{Y}$ at time $\overline{D}$, the transmitter decides on the waiting time $w\left(\overline{Y}\right)$ that minimizes the long-term average age-penalty demonstrated in the fractional program above.[6]

We follow Dinkelbach's approach to transform (24) into the following auxiliary problem for fixed $\lambda \ge 0$ [62]:

$$p^{IIR}(\lambda) \triangleq \min_{w(\cdot)\ge 0} \quad \mathbb{E}\left[\int_{\overline{D}}^{\overline{D}+w\left(\overline{Y}\right)+Y} g\left(t-\overline{S}\right) dt\right] - \lambda \mathbb{E}\left[w\left(\overline{Y}\right)+Y\right]. \tag{25}$$

The optimal solution of (24) is then given by λ^*_{IIR} that solves $p^{IIR}(\lambda^*_{IIR}) = 0$, which can be found via bisection, since $p^{IIR}(\lambda)$ is decreasing [62]. The following theorem characterizes the solution of problem (25). The proof is in Appendix A.

Theorem 1: *The optimal solution of problem (25) is given by*

$$w^*(\bar{y}) = \left[G^{-1}_{\bar{y}}(\lambda)\right]^+, \tag{26}$$

where $[\cdot]^+ \triangleq \max(\cdot, 0)$, $\bar{y}$ is the realization of the starting AoI $\overline{Y}$, and $G_{\bar{y}}(x) \triangleq \mathbb{E}\left[g\left(\bar{y}+x+Y\right)\right]$.

We note that the theorem can be shown using the result reported in [16, Theorem 1]. Our proof approach, however, is different, and is reported here for completeness. Such approach is also used to show parts of Theorem 2 below.

The optimal waiting policy for IIR has a *threshold* structure: a new sample is acquired only when the expected age-penalty by the end of the epoch is at least λ. Note that the optimal λ^*_{IIR} corresponds to the optimal long-term average age-penalty.

B. The FR Scheme

For the FR scheme, the formulated problem can be derived similarly, *with the addition of possible waiting times in between retransmissions.*[7] Specifically, let $W_{i,j}$ represent the waiting time before the jth transmission attempt in the ith epoch. A stationary deterministic policy[8] here is such that $W_{i,j}$ is a determinisitc function $w(\cdot)$ of the instantaneous

[6]We now see explicitly how waiting can be beneficial. Since waiting increases *both* the numerator and denominator of the objective function of problem (24), its optimal value can be non-zero.

[7]This is only amenable for FR since waiting leads to acquiring a fresher sample, and possibly reduced age-penalties. For IIR, waiting after a NACK is clearly suboptimal since it elongates the channel delay for the *same* sample.

[8]We note that [48] shows the optimality of stationary policies in a time-slotted system in which samples are conveyed through an erasure channel. This resembles our FR model yet with no quantization or coding.

age-penalty. This makes the waiting time before the first transmission attempt given by

$$W_{i,1} = f\left(g\left(D_{i-1} - S_{i-1,M_{i-1}}\right)\right) = w\left(\bar{n}\right) \equiv w_1, \quad (27)$$

where $D_{i-1} - S_{i-1,M_{i-1}} = \bar{n}$ represents the starting AoI of the ith (and every) epoch, following M_{i-1} transmission attempts in the previous one. The waiting time before the second attempt, if needed, will then be given by

$$W_{i,2} = w\left(\bar{n} + w_1 + \bar{n}\right) \equiv w_2, \quad (28)$$

since the AoI before the second attempt is given by the starting AoI of the epoch in addition to the time needed to finish the first transmission attempt. In general, the waiting time before the jth attempt in the epoch is given by

$$W_{i,j} = w\left(\sum_{l=1}^{j-1} w_l + j\bar{n}\right) \equiv w_j, \quad (29)$$

and so on. Therefore, under the FR scheme, a stationary deterministic policy reduces to a countable sequence $\{w_j\}$.

Proceeding with the same notations for a given epoch as in the IIR scheme, let us define M as the number of transmission attempts in the epoch, $\bar{M}$ as those in the previous epoch, and $\overline{S}_{\bar{M}}$ as the sampling time of the successful (last) transmission attempt in the previous epoch. The problem now becomes

$$\min_{\{w_j \geq 0\}} \frac{\mathbb{E}\left[\int_{\overline{D}}^{\overline{D}+\sum_{j=1}^{M} w_j + M\bar{n}} g\left(t - \overline{S}_{\bar{M}}\right) dt\right]}{\mathbb{E}\left[\sum_{j=1}^{M} w_j + M\bar{n}\right]}. \quad (30)$$

We follow a similar approach here as in the IIR scheme and consider the following auxiliary problem:

$$p^{FR}(\lambda) \triangleq \min_{\{w_j \geq 0\}} \mathbb{E}\left[\int_{\overline{D}}^{\overline{D}+\sum_{j=1}^{M} w_j + M\bar{n}} g\left(t - \overline{S}_{\bar{M}}\right) dt\right] - \lambda \mathbb{E}\left[\sum_{j=1}^{M} w_j + M\bar{n}\right]. \quad (31)$$

The optimal solution of problem (30) is now given by λ^*_{FR} that solves $p^{FR}\left(\lambda^*_{FR}\right) = 0$, which we will actually provide in *closed-form* this time. The optimal waiting policy structure is provided in the next theorem. The proof is in Appendix B.

Theorem 2: *The optimal solution of problem (31) is given by*

$$w_1^* = \left[G^{-1}(\lambda)\right]^+, \quad (32)$$

$$w_j^* = 0, \ j \geq 2, \quad (33)$$

where $G(x) \triangleq \mathbb{E}\left[g\left(\bar{n} + x + M\bar{n}\right)\right]$. *In addition, the optimal solution of problem (30),* λ^*_{FR}, *is such that* $w_1^* = \left[G^{-1}\left(\lambda^*_{FR}\right)\right]^+ = 0$.

A closed-form expression for λ^*_{FR} can now be found via substituting $w_j = 0, \ \forall j$ in (30).

Theorem 2 shows that *zero-wait* policies are optimal for FR, which is quite intuitive. First, waiting is not optimal in between retransmissions, even though it would lead to acquiring fresher samples, since the AoI is already relatively high following failures. Second, since the epoch always starts with the same AoI, $\bar{n}$, one can optimize the long-term average age-penalty to make waiting not optimal at the beginning of the epoch as well. We note, however, that such results do *not* follow from [14, Theorem 5], since there can be multiple transmissions in the same epoch. We also note that while zero-wait policies have been invoked in other works involving FR coding schemes, e.g., [34], [37], Theorem 2 provides a proof of their *optimality* for general increasing age-penalties. Finally, we note that the results of Theorem 2 are related to those reported in Propositions 3 and 6 in [48]. However, our proof of the optimality of the threshold policy is based on a quite different Lagrangian approach that works for continuous-time systems (different from the time-slotted system considered in [48]).

IV. Enhanced Transmission Schemes

So far the analysis assumed that, naturally, the transmitter must wait for feedback before taking new decisions, e.g., sending IR bits in case of the IIR scheme or acquiring a new sample in case of the FR scheme. In this section, we show that such waiting for receiver processing is unnecessary. We basically take advantage of the processing time β to send extra pieces of information when possible, in order to maintain a smooth information supply *as the receiver decodes and processes previous messages.* We show that with proper timing, this can lead to better results for both the IIR and FR schemes, and hence the name *enhanced* schemes. One assumption here is that the receiver has a (possibly-infinite) queue to store unprocessed data.

A. Enhanced IIR Scheme

The enhanced IIR scheme works as follows. The transmitter sends the original n-bit codeword, consuming nT_b time units, after which the receiver starts decoding. Then, instead of waiting for the β time units processing time, the transmitter goes ahead with transmitting IR bits continuously. This way, if the original n-bit codeword is not successfully decoded, the receiver would have some IR bits awaiting in its queue ready for processing, which saves time. The continuous stream of IR bits transmission stops whenever an ACK is fed back. We note that if the ACK is received in the middle of a bit transmission, this transmission is cut off and stopped immediately.

The next lemma shows that the enhanced IIR scheme described above experiences (almost surely) smaller channel delay for each message transmission. The proof is in Appendix C.

Lemma 1: *For a given value of* r_i, *the enhanced IIR scheme saves the following amount of time in channel delay during the* i*th epoch*:

$$r_i \min\{\beta, T_b\} + (r_i - \kappa_i)\beta \cdot \mathbb{1}_{\beta \geq T_b}, \quad (34)$$

where κ_i *is the smallest integer in* $\{0, 1, \ldots, r_i\}$ *such that* $\lfloor \kappa_i \beta / T_b \rfloor \geq r_i$, *with* $\lfloor x \rfloor$ *denoting the largest integer smaller than or equal to* x, *and* $\mathbb{1}_A = 1$ *if event* A *is true and* 0 *otherwise.*

Lemma 1 shows that the enhanced IIR scheme would achieve smaller long-term average age-penalty relative to the

original IIR scheme discussed previously, owing to (34). The intuition behind this is that once a new sample is generated, its AoI counter starts to increase, and hence the faster it reaches the destination the better. This is different from idle waiting, however, since the waiting occurs *before* the sample is generated.

Let $\tilde{Y}_i$ denote the channel delay experienced by the ith message using the enhanced IIR scheme. Such $\tilde{Y}_i$'s are i.i.d. $\tilde{Y}$. Using the same notation used to describe the distribution of (the original channel delay) Y in (6) and (7), the enhanced IIR channel delay $\tilde{Y}$ has the following distribution according to Lemma 1:

$$\mathbb{P}\left(\tilde{Y}=\bar{n}\right)=p_0, \tag{35}$$

$$\mathbb{P}\left(\tilde{Y}=\bar{n}+kT_b\right)=\prod_{j=0}^{k-1}(1-p_j)p_k, \quad k\geq 1, \tag{36}$$

for $\beta < T_b$, and

$$\mathbb{P}\left(\tilde{Y}=\bar{n}\right)=p_0, \tag{37}$$

$$\mathbb{P}\left(\tilde{Y}=\bar{n}+k\beta\right)=\prod_{j=0}^{k-1}\left(1-p_{\left\lfloor\frac{(k-1)\beta}{T_b}\right\rfloor}\right)p_{\left\lfloor\frac{k\beta}{T_b}\right\rfloor}, \quad k\geq 1, \tag{38}$$

for $\beta \geq T_b$. One would then apply the results of Theorem 1 to find the optimal waiting policy in accordance to the enhanced IIR channel delay distribution $\tilde{Y}$ specified above.

B. Enhanced FR Scheme

For FR, since zero-waiting is optimal by Theorem 2, it could be rewarding therefore, age-wise, to send a new message right away after the previous one is *delivered*, i.e., after nT_b time units instead of $\bar{n}$. However, this may not be optimal if β is relatively large, since it would lead to accumulating *stale* messages at the receiver's end as they wait for decoding to finish.

Let δ denote the waiting time following a message *delivery* at which a new message is transmitted. In the original FR scheme, by Theorem 2, we had $\delta = \beta$. In general though, $\delta \in [0, \beta]$ and should be optimized. The next lemma provides a solution to the optimal δ^*. The proof is in Appendix D.

Lemma 2: *In the FR scheme, it is optimal to send a new message after the previous one's delivery by* $\delta^* = [\beta - nT_b]^+$ *time units.*

Lemma 2 shows that *just-in-time* updating is optimal. For $\beta \leq nT_b$, a new sample is acquired and transmitted just-in-time as the previous message is delivered. While for $\beta > nT_b$, a new sample is acquired and transmitted such that it is delivered just-in-time as the receiver finishes decoding the previous message. This way, delivered samples are always fresh, the receiver is never idle, and feedback is unnecessary.

V. Performance Evaluations and Comparisons

In this section, we discuss how the IIR and FR schemes perform relative to each other under variant system parameters and channel conditions. We do so in the original context of OU process estimation, i.e., when $g(\cdot) \equiv h_\ell(\cdot)$. We note that since the FR scheme has an optimal waiting time of 0, according to Theorem 2, it becomes equivalent to a *uniform* sampling scheme with fixed sampling frequency that depends on ℓ, n, and β. In particular, the enhanced FR scheme generates a new sample every $nT_b + [\beta - nT_b]^+ = \max\{nT_b, \beta\}$ time units. The optimal choice of ℓ and n, therefore, implicitly provides the optimal (uniform) sampling frequency. Due to the wide use of uniform sampling schemes in practice, the FR scheme serves as an implicit uniform sampling benchmark in our evaluations.

Applying Theorem 1 and Lemma 1's result, the optimal waiting policy for enhanced IIR is

$$w^*(\bar{y})=\left[\frac{1}{2\theta}\log\left(\frac{\frac{\sigma^2}{2\theta}\left(1-2^{-2\ell}\right)\mathbb{E}\left[e^{-2\theta\tilde{Y}}\right]}{\frac{\sigma^2}{2\theta}-\lambda^*_{IIR}}\right)-\bar{y}\right]^+, \tag{39}$$

where $\tilde{Y}$ is as defined following Lemma 1.[9] In addition, observing that $\frac{\sigma^2}{2\theta}2^{-2\ell} \leq h_\ell\left(t-\overline{S}\right) \leq \frac{\sigma^2}{2\theta}$ holds true $\forall t \geq \overline{S}$, one can directly see that $\lambda^*_{IIR} \in \left[2^{-2\ell}\frac{\sigma^2}{2\theta}, \frac{\sigma^2}{2\theta}\right]$, facilitating the bisection search. Applying Theorem 2 and Lemma 2's results, the optimal long-term average MMSE for enhanced FR is given by

$$\frac{\sigma^2}{2\theta}\left(1-\frac{\left(1-2^{-2\ell}\right)e^{-2\theta\bar{n}}p_0}{2\theta K_{n,\beta}}\frac{1-e^{-2\theta K_{n,\beta}}}{1-(1-p_0)e^{-2\theta K_{n,\beta}}}\right), \tag{40}$$

where $K_{n,\beta} \triangleq \max\{\beta, nT_b\}$. Derivation details for (39) and (40) are in Appendix E.

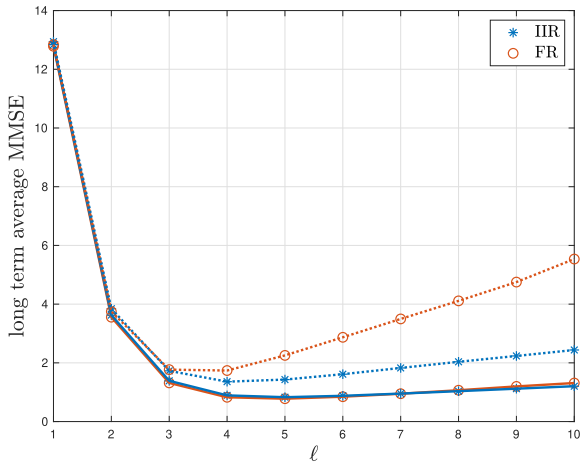
We consider a binary symmetric channel (BSC) with crossover probability $\epsilon \in \left(0, \frac{1}{2}\right)$, and use maximum distance separable (MDS) codes for transmission. This allows us to write $p_j = \sum_{l=0}^{\lfloor\frac{n+j-\ell}{2}\rfloor}\binom{n+j}{l}\epsilon^l(1-\epsilon)^{n+j-l}$. We set $\sigma^2 = 1$, and $T_b = 0.05$ time units. We refer to enhanced IIR and FR without using the word enhanced throughout this section for convenience.

A. Optimal (ℓ, n): Effect of Memory Factor θ

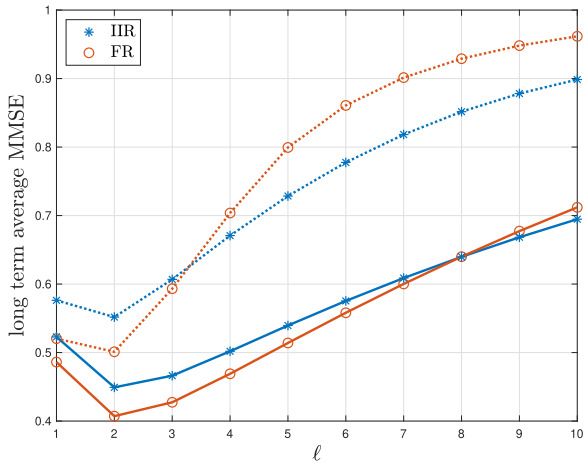
For fixed $\beta = 0.15$, we vary ℓ and numerically choose the best n for IIR and FR. We plot the long-term average MMSE for both IIR and FR versus ℓ in Fig. 2. We do so for $\theta = 0.01$ in Fig. 2a (slowly-varying OU process) and $\theta = 0.5$ in Fig. 2b (fast-varying OU process). For each value of ℓ, the optimal n is evaluated. For both values of θ, we repeat the analysis for $\epsilon = 0.1$ (in solid lines) and $\epsilon = 0.4$ (in dotted lines).

In all of the cases considered, the optimal $n^* = \ell^* + 2$. While the optimal ℓ^* itself depends on whether the OU processes is slowly ($\theta = 0.01$) or fast ($\theta = 0.5$) varying. Specifically, we notice that ℓ^* decreases with θ. This is intuitive since for slowly-varying processes, one can tolerate larger waiting times to get high quality estimates, and vice versa. It is also shown in the figure that IIR performs better than FR for slowly-varying processes, and vice versa for fast-varying ones.

[9] With a slight abuse of notation here, $\bar{y}$ now represents the realization of $\tilde{Y}$ that ended the previous epoch.



(a) $\theta = 0.01$



(b) $\theta = 0.5$

Fig. 2. Performance comparison of IIR and FR vs. ℓ for $\beta = 0.15$, with $\theta = 0.01$ in Fig. 2a (slowly-varying OU process) and $\theta = 0.5$ in Fig. 2b (fast-varying OU process). Solid lines: $\epsilon = 0.1$, and dotted lines: $\epsilon = 0.4$. For $\theta = 0.01$, the optimal (ℓ, n) pair for both schemes is given by $(5, 7)$ for $\epsilon = 0.1$ and by $(4, 6)$ for $\epsilon = 0.4$. While for $\theta = 0.5$, the optimal (ℓ, n) pair for both schemes is given by $(2, 4)$ for both values of ϵ.

This observation settles a goal that this paper is seeking regarding whether one should send fast low-quality samples or slow high-quality ones for the purpose of remote estimation and tracking; it depends on the memory the time-varying process possesses, abstracted by the variable θ in this case. We also note that the relationship $n^* = \ell^* + 2$ does not always hold, neither it is the case that the optimal (ℓ^*, n^*) pairs are the same for IIR and FR; it all depends on the parameters used in the numerical evaluations. If, for instance, we set $\theta = 0.01$, $\epsilon = 0.4$ and $\beta = 1$, we find that the optimal (ℓ^*, n^*) pairs are given by $(4, 10)$ for IIR, and $(4, 18)$ for FR. This can be attributed to the fact that one is estimating a slowly-varying process, over a channel that introduces errors with relatively high rate, with an estimator that incurs a relatively large processing delay ($\beta = 20T_b$).

B. IIR Vs. FR: Effect of Processing Time β

In Fig. 3, we fix $\theta = 0.25$ and plot the long-term average MMSE achieved by IIR and FR versus β. We do so for $\epsilon = 0.1$ (in solid lines) and $\epsilon = 0.4$ (in dotted lines). We observe that IIR performs better than FR for relatively lower values of β, and then the performance switches after some β_{sw} processing time value, marked in black squares. We note that the reason why the curves for $\epsilon = 0.4$ are not very smooth is mainly attributed to the $\lfloor \cdot \rfloor$ (floor) function used in the enhanced schemes' channel delay calculations.

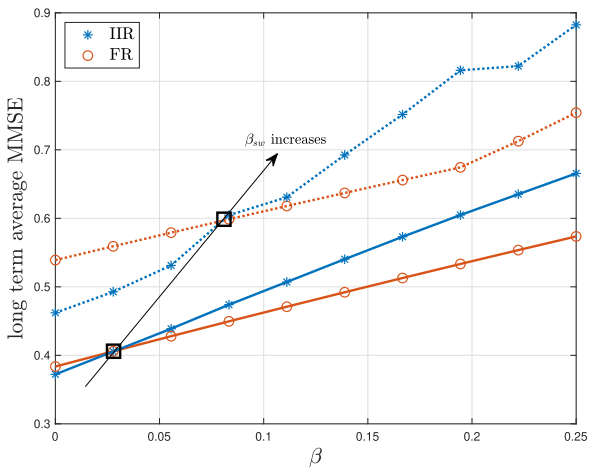


Fig. 3. Performance comparison of IIR and FR vs. β, with $\theta = 0.25$. Solid lines: $\epsilon = 0.1$, and dotted lines: $\epsilon = 0.4$. The processing time value after which FR beats IIR, β_{sw}, is marked in black squares, and is increasing with ϵ.

We notice that the value of β_{sw} increases with ϵ, i.e., when the channel becomes worse. However, the gain due to switching from IIR to FR also increases and becomes more rewarding in this case too. As evident from Figs. 2 and 3, there is no coding scheme that dominantly outperforms the other; it all depends on the system parameters comprising the process, the channel and the processing time.

C. Enhanced Vs. Non-Enhanced Schemes

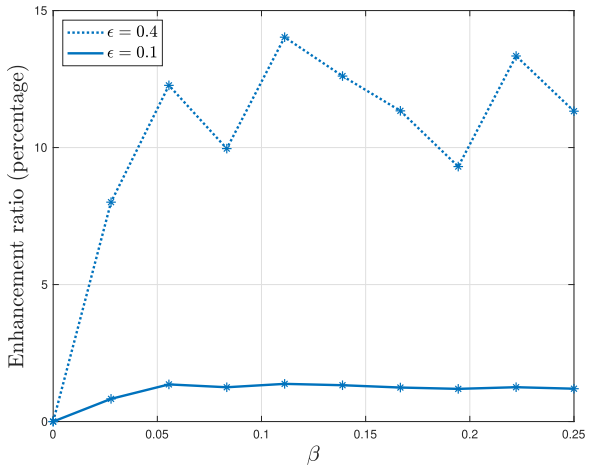
We turn our attention to evaluating the gain achieved (i.e., the loss in MMSE) due to employing the enhanced schemes. Specifically, for fixed $\theta = 0.25$, let us denote by $\widetilde{\mathtt{mmse}}(\beta)$ and $\mathtt{mmse}(\beta)$ the long-term average MMSE achieved by the enhanced and the non-enhanced schemes, respectively. We define the enhancement ratio as

$$1 - \frac{\widetilde{\mathtt{mmse}}(\beta)}{\mathtt{mmse}(\beta)}, \tag{41}$$

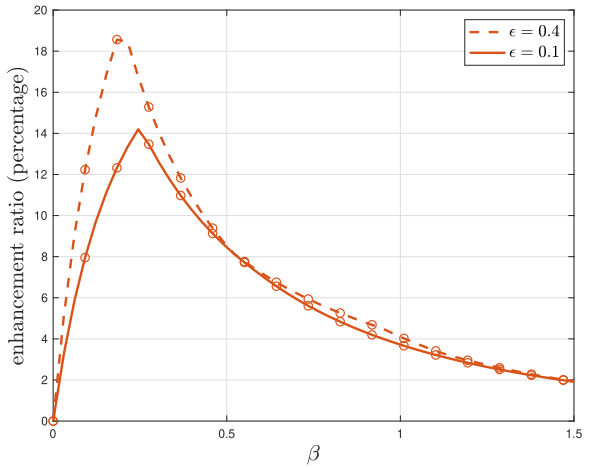
and so the higher this ratio is, the larger the gain due to enhancement. In Fig. 4, we plot the enhancement ratio (in percentage) for both IIR and FR versus β.

For the IIR case in Fig. 4a, we observe that: (1) the enhancement ratio relatively increases with β (again, the non-smoothness effect is mainly due to using the floor function in calculations), because as β increases, one can *fit more data* as the receiver decodes previous ones; and (2) the gain is more apparent for worse channel conditions, which is due to the ability of enhanced IIR to make more data available for reprocessing at the receiver's end following decoding errors, compared to non-enhanced IIR.

Fig. 4b deals with FR, and exhibits some behavioral differences when compared to IIR. In particular, the enhancement



(a) IIR

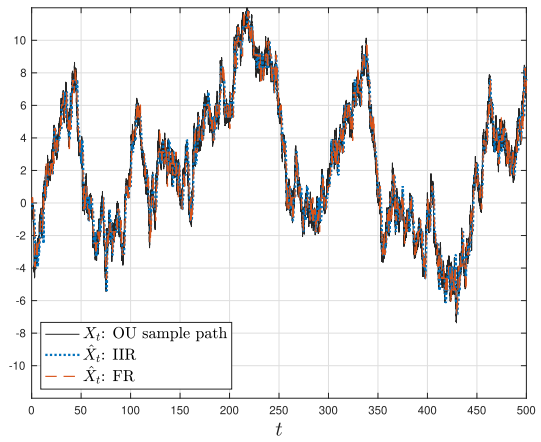


(b) FR

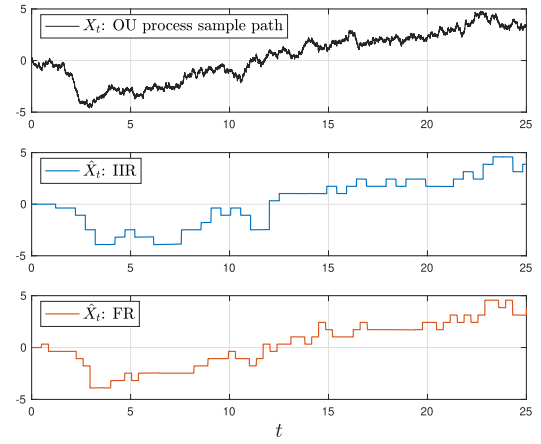
Fig. 4. Evaluating the *gain* due to enhancement, with $\theta = 0.25$. The enhancement ratio is defined as the ratio between the long-term average MMSE of the enhanced scheme to that of the non-enhanced scheme, subtracted from unity.

ratio first increases then decreases with β. The reason for such behavior is that for the enhanced FR scheme there is *only one* extra codeword that can be transmitted as the receiver finishes processing, *regardless* of the value of β. Specifically, according to Lemma 2, the optimal inter-sampling (and transmission) delay for the enhanced FR scheme is given by $\beta - nT_b$ (for $\beta > nT_b$). While for the non-enhanced FR scheme, the inter-sampling delay is given by β. Hence, as β becomes much larger than nT_b, the two inter-sampling delays become equivalent, and the performances of both schemes (enhanced and non-enhanced) become similar. Therefore, for the FR scheme, intermediate values of β (relative to nT_b) provide the highest gain from enhancement. As in the IIR scheme, the enhancement gain is more apparent in worse channel conditions.

In summary, this numerical calculation shows that the enhancement effect is relatively more noticeable for FR ($\approx$18% gain) than it is for IIR ($\approx$14% gain), and that it would better serve both schemes in relatively worse channel conditions.



(a) Full view



(b) Zoomed view

Fig. 5. Tracking an OU sample path by generating an MMSE estimate using IIR and FR. We fix $\beta = 0.15$, $\theta = 0.01$ and $\epsilon = 0.1$, and use (the optimal) $\ell = 5$ and $n = 7$.

D. Timely Real-Time Tracking

We finally apply the techniques developed in this paper to an example sample path of the OU process. In this particular example we fix $\beta = 0.15$, $\theta = 0.01$ and $\epsilon = 0.1$. We first generate an OU process sample path over $t = 500$ time units ($10^4 \times T_b$). Then, we pass it through an MMSE quantizer[10] with $\ell = 5$ (which is the optimal ℓ^* in this case using Fig. 2a). After that, we use either IIR or FR with $n = 7$ (again, this is the optimal n^* in this case) to send the quantized samples through a BSC(0.1). We apply the optimal waiting policies in accordance to the channel delay realizations and receiver processing time.

The results are shown in Fig. 5. The full view in Fig. 5a shows that both IIR and FR are able to allow the receiver to produce MMSE estimates that closely-track the original OU sample path. While the zoomed view in Fig. 5b shows the specifics of how the MMSE estimates look like. Empirically, the MSE for this sample path is ≈ 0.87 for IIR and ≈ 0.74 for

[10]We train a quantizer using 1000 different OU processes sample paths, each over $t \in [0, 500]$, using Lloyd's algorithm to build this [60]. Each sample path realization produces a particular code when Lloyd's algorithm converges. We then average over all the produced codes and use the averaged code to generate the results of this subsection.

FR, which are close to the theoretical values of the long-term average MMSE evaluated in Fig. 2a. This shows the ability of our techniques to achieve *timely tracking* of the process.

VI. Conclusions and Extensions

A study of the effects of sampling, quantization and coding over noisy channels on MMSE estimates of an OU process has been presented. Focusing on MMSE quantizers, together with IIR and FR codes, a joint optimization problem of when to take new samples, how many quantization and codeword bits to use, has been formulated and solved. A fixed non-zero processing time has been considered at the receiver, modeling mainly decoding and feedback transmission times. It is shown how finely tuning the sampling and transmission times could make us of the processing time to send new data in order to save time in case decoding fails. Through numerical evaluations, it is shown that IIR performs relatively better than FR with small processing times, and vice versa, and so neither coding scheme dominates. It is also shown that the techniques developed in this paper can achieve timely tracking of the original process at the receiver's end.

In this work, the focus has been on signal-independent sampling policies. As an extension, one could develop techniques that work for *signal-dependent* sampling policies instead, in which the state of the OU process is observable to the sampler. While this is expected to produce better results, this comes with the challenge of *jointly* designing an MMSE quantizer *and* deriving an MMSE estimate at the receiver in this case. More generally though, there has been a separation-based quantization and coding methodology followed in this work, with focusing on two relatively-simple coding strategies. One could investigate the benefits of jointly optimizing the quantizer and the transmission code being used to convey the samples to the receiver with the smallest MMSE, which can be done for either signal-independent or signal-dependent sampling policies. Some structural properties of the tracked process may also guide the joint design in this case, as in, e.g., the sparse signal framework of [63]. Finally, one can also extend the notion of fixed processing times to more practical models that take into consideration the code rate being used, together with noise in the feedback channel. As a more direct extension focusing on this point, one may consider *random* processing times, which calls for the investigation of whether it is useful to generate a new sample while an old one is still being processed if the processing time becomes relatively large.

Appendix

A. Proof of Theorem 1

We introduce the following Lagrangian [64][11]:

$$\mathcal{L} = \mathbb{E}\left[\int_{\overline{D}}^{\overline{D}+w\left(\overline{Y}\right)+Y} g\left(t-\overline{S}\right)dt\right] - \lambda\mathbb{E}\left[w\left(\overline{Y}\right)+Y\right] - \sum_{\bar{y}} w(\bar{y})\eta(\bar{y}), \quad (42)$$

where $\eta(\bar{y})$ is a Lagrange multiplier. Using Leibniz rule, we take the functional derivative with respect to $w(\bar{y})$ and equate to 0 to get

$$\mathbb{E}\left[g\left(\bar{y}+w^*(\bar{y})+Y\right)\right] = \lambda + \frac{\eta(\bar{y})}{\mathbb{P}\left(\overline{Y}=\bar{y}\right)}. \quad (43)$$

Since g is increasing, the left hand side above is therefore an increasing function of $w^*(\bar{y})$, which we denote $G_{\bar{y}}(\cdot)$ in the theorem statement. Now, if $\lambda \leq G_{\bar{y}}(0)$, then we must have $\eta(\bar{y}) > 0$, and hence $w^*(\bar{y}) = 0$ by complementary slackness [64]. Conversely, if $\lambda > G_{\bar{y}}(0)$, then we must have $w^*(\bar{y}) > 0$, and hence $\eta(\bar{y}) = 0$ also by complementary slackness. In the latter case, $w^*(\bar{y}) = G_{\bar{y}}^{-1}(\lambda)$. Finally, observe that $\lambda \leq G_{\bar{y}}(0) \iff G_{\bar{y}}^{-1}(\lambda) \leq 0$. This concludes the proof.

B. Proof of Theorem 2

We first simplify the terms of the objective function of (31). Using iterated expectations, it can be shown that

$$\mathbb{E}\left[\sum_{j=1}^{M} w_j + M\bar{n}\right] = \sum_{j=1}^{\infty} w_j(1-p_0)^{j-1} + \frac{\bar{n}}{p_0}. \quad (44)$$

Now let us define

$$\zeta_m\left(\boldsymbol{w}_1^m\right) \triangleq \int_{\overline{D}}^{\overline{D}+\sum_{j=1}^{m} w_j + m\bar{n}} g\left(t-\overline{S}_{\bar{M}}\right)dt \quad (45)$$

and, leveraging iterated expectations on the first term of (31), introduce the following Lagrangian[12]:

$$\mathcal{L} = \sum_{m=1}^{\infty} \zeta_m\left(\boldsymbol{w}_1^m\right)(1-p_0)^{m-1}p_0 - \lambda\sum_{j=1}^{\infty} w_j(1-p_0)^{j-1} - \lambda\frac{\bar{n}}{p_0} - \sum_{j=1}^{\infty} w_j\eta_j, \quad (46)$$

where η_j's are Lagrange multipliers. Now observe that, using Leibniz rule, it holds for $j \leq m$ that

$$\frac{\partial\zeta_m\left(\boldsymbol{w}_1^m\right)}{\partial w_j} = g\left(\bar{n}+\sum_{j=1}^{m} w_j + m\bar{n}\right). \quad (47)$$

Taking derivative of the Lagrangian with respect to w_j and equating to 0, we use the above to get

$$\sum_{m=j}^{\infty} g\left(\bar{n}+\sum_{j=1}^{m} w_j + m\bar{n}\right)(1-p_0)^{m-j}p_0 = \lambda + \frac{\eta_j}{(1-p_0)^{j-1}}. \quad (48)$$

Next, let us substitute $j=k$ and $j=k+1$ above, $k \geq 1$, subtract them from each other, and rearrange to get

$$g\left(\bar{n}+\sum_{j=1}^{k} w_j + k\bar{n}\right) = \lambda + \frac{\eta_k - \eta_{k+1}}{(1-p_0)^{k-1}p_0}. \quad (49)$$

Since $g(\cdot)$ is increasing, and λ is fixed, $\left\{\frac{\eta_k-\eta_{k+1}}{(1-p_0)^{k-1}p_0}\right\}$ is increasing. From there, one can conclude that $\eta_j > 0$,

[11] Using the monotonicity of $g(\cdot)$, it can be shown that problem (25) is convex.

[12] Again, as mentioned above, it can be shown that problem (31) is convex using monotonicity of $g(\cdot)$.

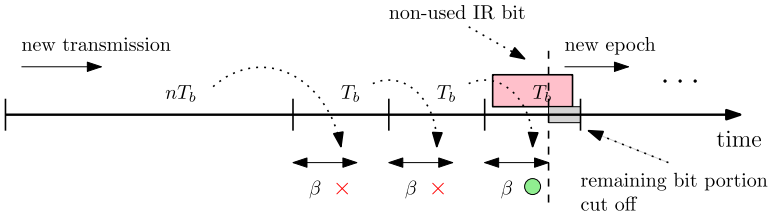


Fig. 6. Example sample path during the ith epoch using the enhanced IIR scheme when $\beta \leq T_b$. In this example $r_i = 2$, and so the third IR bit is non-used and its remaining portion is cut off to start a new epoch. Red crosses denote failed decoding attempts and the green circle denotes success.

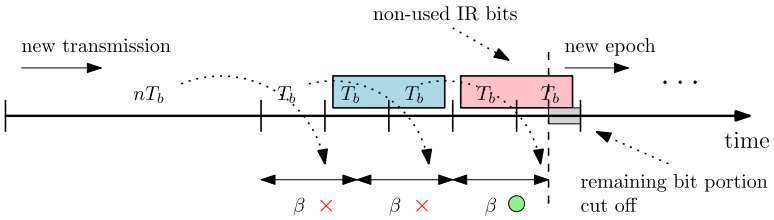


Fig. 7. Example sample path during the ith epoch using the enhanced IIR scheme when $\beta > T_b$. In this example $r_i = 2$ and $\beta = 1.5T_b$, and so the final two IR bits are non-used and the remaining bit portion is cut off to start a new epoch. Red crosses denote failed decoding attempts and the green circle denotes success.

$j \geq 2$ must hold. Hence, by complementary slackness, $w_j^* = 0$, $j \geq 2$ [64]. Using (48) for $j = 1$, the optimal w_1^* now solves

$$G\left(w_1^*\right) = \lambda + \eta_1, \tag{50}$$

where $G(\cdot)$ is as defined in the theorem statement. Observe that $G(\cdot)$ is increasing and therefore the above has a unique solution. Proceeding similarly as in the proof of Theorem 1, if $\lambda \leq G(0)$, then we must have $\eta_1 > 0$, and hence $w_1^* = 0$ by complementary slackness; conversely, if $\lambda > G(0)$, then we must have $w_1^* > 0$, and hence $\eta_1 = 0$ by complementary slackness as well [64]. In the latter case, $w_1^* = G^{-1}(\lambda)$. Finally, observe that $\lambda \leq G(0) \iff G^{-1}(\lambda) \leq 0$. This concludes the proof of the first part of the theorem.

To show the second part, all we need to prove now is that $G^{-1}\left(\lambda_{FR}^*\right) \leq 0$, or equivalently that $\lambda_{FR}^* \leq G(0)$. Toward that end, observe that $p_{FR}(\lambda)$ is decreasing, and therefore if $p_{FR}\left(G(0)\right) \leq 0$ then the premise follows. Now for $\lambda = G(0)$ we know from the first part of the proof that $w_1^* = 0$. Thus,

$$\begin{aligned} & p_{FR}\left(G(0)\right) \\ &= \sum_{m=1}^{\infty} \zeta_m\left(0\right)\left(1-p_0\right)^{m-1} p_0 - G(0)\frac{\bar{n}}{p_0} && (51) \\ &= \mathbb{E}\left[\int_{\overline{D}}^{\overline{D}+M\bar{n}} g\left(t-\overline{S}_{\bar{M}}\right) dt\right] - G(0)\mathbb{E}\left[M\right]\bar{n} && (52) \\ &= \mathbb{E}\left[\int_0^{M\bar{n}} g\left(\bar{n}+t\right) dt\right] - \mathbb{E}\left[\int_0^{M\bar{n}} G(0) dt\right] && (53) \\ &= \mathbb{E}\left[\int_0^{M\bar{n}} \mathbb{E}\left[g\left(\bar{n}+t\right) - g\left(\bar{n}+M\bar{n}\right)\right] dt\right], && (54) \end{aligned}$$

where (53) follows by change of variables and (54) follows by definition of $G(\cdot)$. Finally, observe that by monotonicity of $g(\cdot)$, (54) is non-positive. This concludes the proof.

C. Proof of Lemma 1

Let us consider the ith epoch. We prove the lemma by computing the channel delay experienced by the enhanced scheme for some realization of r_i. The proof can be better-conveyed graphically through Figs. 6 and 7 below. We will consider two cases as follows.

1) $\beta \leq T_b$: In this case, the first feedback following the initial nT_b time units is received while the first IR bit is still being transmitted. If it is an ACK, then the transmitter stops and cuts off the current IR bit transmission and ends the epoch with a channel delay of $nT_b + \beta$. Otherwise, if it is a NACK, then the receiver will begin re-processing with a codeword of length $n+1$ after exactly $T_b - \beta$ time units from the time the feedback is received. Simultaneously, the transmitter will send the second IR bit. The process is repeated till an ACK is received.

In general, an ACK will be received after r_i IR bits, and the (r_i+1)th bit will be cut off (this bit will be a non-used IR bit). This ends the epoch with a channel delay of exactly

$$nT_b + r_i\beta + r_i(T_b - \beta) + \beta = \bar{n} + r_i T_b, \tag{55}$$

which saves $r_i\beta$ time units compared to the original IIR scheme that waits for feedback before sending IR bits. An example sample path is shown in Fig. 6.

2) $\beta > T_b$: Different from the $\beta \leq T_b$ case, the transmitter can now possibly fit more than one IR bit while the receiver is processing previously-received bits. Specifically, a total of $\lfloor \beta/T_b \rfloor$ IR bits would be received by the end of the first decoding attempt, a total of $\lfloor 2\beta/T_b \rfloor$ IR bits would be received by the end of the second decoding attempt, and so on.

Now let κ_i be as defined in the lemma. This way, the required IR bits for successful decoding will be available after exactly $\kappa_i\beta$ time units following the initial nT_b time units, and an ACK will be fed back β time units afterwards. By the time an ACK is received, there would be already

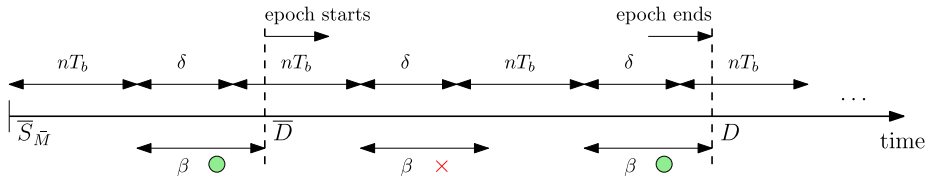


Fig. 8. Example sample path during an epoch using the enhanced FR scheme when $\beta \leq nT_b$. In this example $M = 2$, and so it takes two transmissions to succeed. The red cross denotes a failed decoding attempt and green circles denote success.

some extra IR bits sent to the receiver that were not needed in decoding (these will be non-used IR bits). In addition, there could be an extra bit portion that needs to be cut off belonging to an IR bit that is being transmitted while the ACK is received; this occurs if $(\kappa_i+1)\beta > \lfloor(\kappa_i+1)\beta/T_b\rfloor T_b$. This ends the epoch with a channel delay of exactly

$$nT_b + \kappa_i\beta + \beta = \bar{n} + \kappa_i\beta \tag{56}$$

which saves $r_iT_b + (r_i - \kappa_i)\beta$ time units. An example sample path is shown in Fig. 7.

D. Proof of Lemma 2

Let L denote the epoch length, and let Q denote the cumulative age-penalty in the epoch given by

$$Q = \int_{\overline{D}}^{\overline{D}+L} g\left(t - \overline{S}_{\bar{M}}\right) dt. \tag{57}$$

Recalling the definition of δ, our goal is to characterize $\mathbb{E}[L]$ and $\mathbb{E}[Q]$ in terms of δ and solve the following optimization problem to find δ^*:

$$\min_{0\leq\delta\leq\beta} \quad \frac{\mathbb{E}[Q]}{\mathbb{E}[L]}. \tag{58}$$

Similar to the proof of Lemma 1 in Appendix C, our proof methodology is made clearer through Figs. 8 and 9, and we will consider two cases as follows.

1) $\beta \leq nT_b$: In this case, we need to show $\delta^* = 0$. Right before the epoch starts, there would be $\lfloor(\beta-\delta)/T_b\rfloor$ bits (belonging to a new message) already available. The first decoding attempt in the epoch, therefore, occurs after $nT_b - \beta + \delta$ time units from the epoch's start time. If this decoding attempt is successful, an ACK will be fed back after β time units. Otherwise, a new message will be transmitted through the same manner again, see Fig. 8. From the figure, one can see that the epoch length is given by

$$L = ((nT_b - \beta + \delta) + \beta)\, M \tag{59}$$

$$= (nT_b + \delta)\, M, \tag{60}$$

and therefore

$$\mathbb{E}[L] = \frac{nT_b + \delta}{p_0}, \tag{61}$$

$$\mathbb{E}[Q] = \sum_{m=1}^{\infty}\left(\int_{\overline{D}}^{\overline{D}+(nT_b+\delta)m} g\left(t - \overline{S}_{\bar{M}}\right) dt\right)(1-p_0)^{m-1}p_0. \tag{62}$$

Next, we follow Dinkelbach's approach [62] to solve problem (58) and introduce the auxiliary problem

$$q(\lambda) \triangleq \min_{0\leq\delta\leq\beta} \quad \mathbb{E}[Q] - \lambda\mathbb{E}[L] \tag{63}$$

for some $\lambda \geq 0$. We introduce the following Lagrangian for such problem [64]:

$$\mathcal{L} = \mathbb{E}[Q] - \lambda\mathbb{E}[L] - \eta\delta + \omega(\delta - \beta), \tag{64}$$

where η and ω are Lagrange multipliers. Now using (61) and (62), we take the derivative with respect to δ to get

$$\frac{d\mathcal{L}}{d\delta} = \sum_{m=1}^{\infty} mg\left(\overline{D} + (nT_b+\delta)\, m - \overline{S}_{\bar{M}}\right)(1-p_0)^{m-1}p_0 - \frac{\lambda}{p_0} - \eta + \omega \tag{65}$$

$$= \sum_{m=1}^{\infty} mg\left(\bar{n} + (nT_b+\delta)\, m\right)(1-p_0)^{m-1}p_0 - \frac{\lambda}{p_0} - \eta + \omega \tag{66}$$

$$\triangleq H(\delta) - \frac{\lambda}{p_0} - \eta + \omega. \tag{67}$$

Therefore, the optimal δ^* solves

$$H\left(\delta^*\right) = \frac{\lambda}{p_0} + \eta - \omega. \tag{68}$$

Note that $H(\delta)$ is increasing in δ by monotonicity of $g(\cdot)$. Hence, if $\lambda < p_0H(0)$ then we must have $\eta > 0$, which implies by complementary slackness that $\delta^* = 0$.

We now proceed similarly as in the second part of the proof of Theorem 2 in Appendix B. Specifically, since the optimal λ^* satisfies $q(\lambda^*) = 0$ and $q(\lambda)$ is decreasing [62], it suffices to show that $q\left(p_0H(0)\right) < 0$. Towards that end, we have

$$q\left(p_0H(0)\right) = \sum_{m=1}^{\infty}\left(\int_{\overline{D}}^{\overline{D}+nT_bm} g\left(t - \overline{S}_{\bar{M}}\right) dt\right)(1-p_0)^{m-1}p_0 - p_0H(0)\frac{nT_b}{p_0} \tag{69}$$

$$< \sum_{m=1}^{\infty} nT_bmg\left(\overline{D} + nT_bm - \overline{S}_{\bar{M}}\right)(1-p_0)^{m-1}p_0 - H(0)nT_b \tag{70}$$

$$= 0, \tag{71}$$

where the inequality follows by monotonicity of $g(\cdot)$, and the last equality follows by definition of $H(\cdot)$.

2) $\beta > nT_b$: In this case, we need to show $\delta^* = \beta - nT_b$. We first argue that δ^* cannot be smaller than $\beta - nT_b$. To see this, observe that if $\delta^* < \beta - nT_b$, then there would be a codeword waiting in the receiver's queue for $\beta - nT_b - \delta^*$ time units after being completely received before it gets processed. One can strictly decrease the age-penalty in

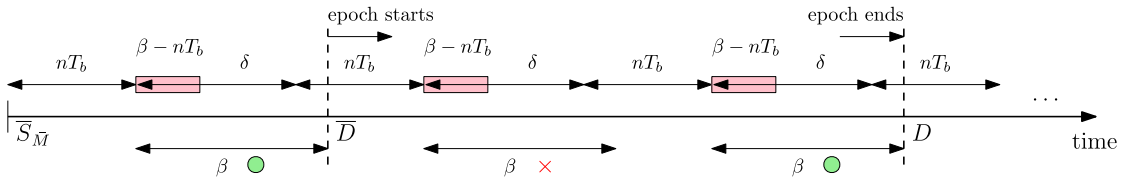


Fig. 9. Example sample path during an epoch using the enhanced FR scheme when $\beta > nT_b$. In this example $M = 2$, and so it takes two transmissions to succeed. Light-red boxes represent the lower bound on δ (idle times). The red cross denotes a failed decoding attempt and green circles denote success.

this case by acquiring *fresher* sample instead of the current one via pushing the sampling time exactly $\beta - nT_b - \delta^*$ time units forward and avoid the unnecessary idle waiting at the receiver. Thus, our goal now is to solve problem (58) over the new bound $\delta \in [\beta - nT_b, \beta]$.

As in the previous case, and now that $\delta \geq \beta - nT_b$, there would also be $\lfloor (\beta - \delta)/T_b \rfloor$ bits available from a new message right before the epoch starts, and the first decoding attempt in the epoch would occur after $nT_b - \beta + \delta$ time units from the epoch's start time. This repeats until an ACK is fed back, see Fig. 9.

This gives rise to the exact same $\mathbb{E}[L]$ and $\mathbb{E}[Q]$ expressions in (61) and (62), respectively. One can thus follow the same analysis for the $\beta \leq nT_b$ case to solve the optimization problem and reach the conclusion that δ^* should be equal to its lower bound, $\beta - nT_b$ in this case.

E. Deriving Equations (39) and (40)

We derive the optimal waiting policy in (39) by solving $G_{\bar{y}}\left(w^*(\bar{y})\right) = \lambda^*_{IIR}$ with $G_{\bar{y}}(\cdot)$ as defined in Theorem 1, with $g(\cdot) \equiv h_\ell(\cdot)$, after replacing the random variable Y with $\tilde{Y}$. That is,

$$\begin{aligned} G_{\bar{y}}\left(w^*(\bar{y})\right) &= \mathbb{E}\left[h_\ell\left(\bar{y} + w^*(\bar{y}) + \tilde{Y}\right)\right] \\ &= \frac{\sigma^2}{2\theta}\left(1 - \left(1 - 2^{-2\ell}\right) e^{-2\theta\left(\bar{y} + w^*(\bar{y})\right)} \mathbb{E}\left[e^{-2\theta\tilde{Y}}\right]\right) \\ &= \lambda^*_{IIR}, \end{aligned} \tag{72}$$

whence (39) directly follows by solving for $w^*(\bar{y})$ above and taking the non-negative part.

Next, we derive the long-term average MMSE expression in (40) through basically evaluating the optimal $\mathbb{E}[L]$ and $\mathbb{E}[Q]$ in (61) and (62), respectively, with $g(\cdot) \equiv h_\ell(\cdot)$, after substituting $\delta^* = [\beta - nT_b]^+$. First, we have

$$\begin{aligned} \mathbb{E}[L] &= \frac{nT_b + [\beta - nT_b]^+}{p_0} \\ &= \frac{K_{n,\beta}}{p_0}. \end{aligned} \tag{73}$$

Next, we have

$$\begin{aligned} &\mathbb{E}[Q] \\ &= \sum_{m=1}^{\infty} \left(\int_{\overline{D}}^{\overline{D} + \left(nT_b + [\beta - nT_b]^+\right)m} h_\ell\left(t - \overline{S}_{\bar{M}}\right) dt \right) \\ &\quad \times (1 - p_0)^{m-1} p_0 \\ &= \sum_{m=1}^{\infty} \left(\int_{\overline{D}}^{\overline{D} + K_{n,\beta} m} \frac{\sigma^2}{2\theta}\left(1 - \left(1 - 2^{-2\ell}\right) e^{-2\theta\left(t - \overline{S}_{\bar{M}}\right)}\right) dt \right) \\ &\quad \times (1 - p_0)^{m-1} p_0 \\ &= \frac{\sigma^2}{2\theta}\left(\frac{K_{n,\beta}}{p_0} - \frac{\left(1 - 2^{-2\ell}\right) e^{-2\theta\bar{n}}}{2\theta} \right. \\ &\quad \left. \times \left(1 - \frac{p_0 e^{-2\theta K_{n,\beta}}}{1 - (1 - p_0) e^{-2\theta K_{n,\beta}}}\right)\right) \\ &= \frac{\sigma^2}{2\theta}\left(\frac{K_{n,\beta}}{p_0} - \frac{\left(1 - 2^{-2\ell}\right) e^{-2\theta\bar{n}}}{2\theta} \frac{1 - e^{-2\theta K_{n,\beta}}}{1 - (1 - p_0) e^{-2\theta K_{n,\beta}}} \right). \end{aligned} \tag{74}$$

Equation (40) now directly follows via dividing $\mathbb{E}[Q]$ above by $\mathbb{E}[L]$.

REFERENCES

[1] A. Arafa, K. Banawan, K. G. Seddik, and H. Vincent Poor, "Timely estimation using coded quantized samples," in *Proc. ISIT*, Jun. 2020, pp. 1812–1817.

[2] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the Brownian motion," *Phys. Rev.*, vol. 36, pp. 823–841, Sep. 1930.

[3] J. L. Doob, "The Brownian movement and stochastic equations," *Ann. Math.*, vol. 43, no. 2, pp. 351–369, 1942.

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

[5] C. Kam, S. Kompella, and A. Ephremides, "Age of information under random updates," in *Proc. IEEE ISIT*, Jul. 2013, pp. 66–70.

[6] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1897–1910, Apr. 2016.

[7] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age and value of information: Non-linear age case," in *Proc. ISIT*, Jun. 2017, pp. 326–330.

[8] R. D. Yates and S. K. Kaul, "The age of information: Real-time status updating by multiple sources," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1807–1827, Mar. 2019.

[9] R. Talak and E. Modiano, "Age-delay tradeoffs in single server systems," in *Proc. ISIT*, Jul. 2019, pp. 340–344.

[10] Y. Inoue, H. Masuyama, T. Takine, and T. Tanaka, "A general formula for the stationary distribution of the age of information and its application to single-server queues," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8305–8324, Dec. 2019.

[11] A. Soysal and S. Ulukus, "Age of information in G/G/1/1 systems: Age expressions, bounds, special cases, and optimization," 2019, *arXiv:1905.13743*. [Online]. Available: https://arxiv.org/abs/1905.13743

[12] P. Zou, O. Ozel, and S. Subramaniam, "Waiting before serving: A companion to packet management in status update systems," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3864–3877, Jun. 2020.

[13] Y.-P. Hsu, E. Modiano, and L. Duan, "Age of information: Design and analysis of optimal scheduling algorithms," in *Proc. ISIT*, Jun. 2017, pp. 561–565.

[14] 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, Nov. 2017.

[15] B. Zhou and W. Saad, "Optimal sampling and updating for minimizing age of information in the Internet of Things," in *Proc. GLOBECOM*, Dec. 2018, pp. 1–6.

[16] Y. Sun and B. Cyr, "Sampling for data freshness optimization: Non-linear age functions," *J. Commun. Netw.*, vol. 21, no. 3, pp. 204–219, Jun. 2019.

[17] H. Tang, J. Wang, L. Song, and J. Song, "Minimizing age of information with power constraints: Multi-user opportunistic scheduling in multi-state time-varying channels," 2020, *arXiv:1912.05947*. [Online]. Available: https://arxiv.org/abs/1912.05947

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

[19] X. Wu, J. Yang, and J. Wu, "Optimal status update for age of information minimization with an energy harvesting source," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 193–204, Mar. 2018.

[20] A. Baknina, O. Ozel, J. Yang, S. Ulukus, and A. Yener, "Sending information through status updates," in *Proc. IEEE ISIT*, Jun. 2018, pp. 2271–2275.

[21] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Age-minimal transmission for energy harvesting sensors with finite batteries: Online policies," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 534–556, Jan. 2020.

[22] B. T. Bacinoglu, Y. Sun, E. Uysal, and V. Mutlu, "Optimal status updating with a finite-battery energy harvesting source," *J. Commun. Netw.*, vol. 21, no. 3, pp. 280–294, Jun. 2019.

[23] S. Leng and A. Yener, "Age of information minimization for an energy harvesting cognitive radio," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 2, pp. 427–439, Jun. 2019.

[24] B. Buyukates, A. Soysal, and S. Ulukus, "Age of information in multihop multicast networks," *J. Commun. Netw.*, vol. 21, no. 3, pp. 256–267, Jun. 2019.

[25] A. M. Bedewy, Y. Sun, and N. B. Shroff, "The age of information in multihop networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 3, pp. 1248–1257, Jun. 2019.

[26] P. Mayekar, P. Parag, and H. Tyagi, "Optimal source codes for timely updates," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3714–3741, Jun. 2020.

[27] M. Zhang, A. Arafa, J. Huang, and H. V. Poor, "Pricing fresh data," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1211–1225, May 2021.

[28] A. Arafa, R. D. Yates, and H. V. Poor, "Timely cloud computing: Preemption and waiting," in *Proc. Allerton*, Sep. 2019, pp. 528–535.

[29] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Age-based scheduling policy for federated learning in mobile edge networks," 2019, *arXiv:1910.14648*. [Online]. Available: https://arxiv.org/abs/1910.14648

[30] R. D. Yates, Y. Sun, D. R. Brown, III, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," 2020, *arXiv:2007.08564*. [Online]. Available: https://arxiv.org/abs/2007.08564

[31] E. Najm, R. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," in *Proc. ISIT*, Jun. 2017, pp. 131–135.

[32] H. Sac, T. Bacinoglu, E. Uysal-Biyikoglu, and G. Durisi, "Age-optimal channel coding blocklength for an M/G/1 queue with HARQ," in *Proc. SPAWC*, Jun. 2018, pp. 1–5.

[33] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal-Biyikoglu, "Delay and peak-age violation probability in short-packet transmissions," in *Proc. ISIT*, Jun. 2018, pp. 2471–2475.

[34] R. D. Yates, E. Najm, E. Soljanin, and J. Zhong, "Timely updates over an erasure channel," in *Proc. ISIT*, Jun. 2017, pp. 316–320.

[35] A. Baknina and S. Ulukus, "Coded status updates in an energy harvesting erasure channel," in *Proc. CISS*, Mar. 2018, pp. 1–6.

[36] S. Feng and J. Yang, "Age-optimal transmission of rateless codes in an erasure channel," in *Proc. ICC*, May 2019, pp. 1–6.

[37] E. Najm, E. Telatar, and R. Nasser, "Optimal age over erasure channels," 2019, *arXiv:1901.01573*. [Online]. Available: https://arxiv.org/abs/1901.01573

[38] A. Javani, M. Zorgui, and Z. Wang, "On the age of information in erasure channels with feedback," 2019, *arXiv:1911.05840*. [Online]. Available: https://arxiv.org/abs/1911.05840

[39] P. Parag, A. Taghavi, and J.-F. Chamberland, "On real-time status updates over symbol erasure channels," in *Proc. WCNC*, Mar. 2017, pp. 1–6.

[40] E. T. Ceran, D. Gunduz, and A. Gyorgy, "Average age of information with hybrid arq under a resource constraint," *IEEE Trans. Commun.*, vol. 18, no. 3, pp. 1900–1913, Mar. 2019.

[41] A. Arafa, K. Banawan, K. G. Seddik, and H. V. Poor, "On timely channel coding with hybrid ARQ," in *Proc. GLOBECOM*, Dec. 2019, pp. 1–6.

[42] K. Huang, W. Liu, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Real-time remote estimation with hybrid ARQ in wireless networked control," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3490–3504, May 2020.

[43] X. Chen and S. S. Bidokhti, "Benefits of coding on age of information in broadcast networks," in *Proc. ITW*, Aug. 2019, pp. 1–5.

[44] S. Feng and J. Yang, "Adaptive coding for information freshness in a two-user broadcast erasure channel," in *Proc. GLOBECOM*, Dec. 2019, pp. 1–6.

[45] M. M. Watson Luby, A. Shokrollahi, and T. Stockhammer, *Raptor Forward Error Correction Scheme for Object Delivery*, document RFC 5053, Oct. 2007.

[46] *Multimedia Broadcast/Multicast Service (MBMS); Protocols and Codes, Version 11.2.0*, document 3GPP TS 26.346, Feb. 2010.

[47] *Digital Video Broadcasting (DVB); IP Datacast Over DVB-H: Content Delivery Protocols, Version 1.3.1*, document ETSI, TS 102 472, Jun. 2009.

[48] M. Klugel, M. H. Mamduhi, S. Hirche, and W. Kellerer, "AoI-penalty minimization for networked control systems with packet loss," in *Proc. INFOCOM*, Apr. 2019, pp. 189–196.

[49] A. Mitra, J. A. Richards, S. Bagchi, and S. Sundaram, "Finite-time distributed state estimation over time-varying graphs: Exploiting the age-of-information," in *Proc. ACC*, Jul. 2019.

[50] J. Chakravorty and A. Mahajan, "Remote estimation over a packet-drop channel with Markovian state," *IEEE Trans. Autom. Control*, vol. 65, no. 5, pp. 2016–2031, May 2020.

[51] 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 *Proc. ICCPS*, Apr. 2019, pp. 109–117.

[52] S. Roth, A. Arafa, H. V. Poor, and A. Sezgin, "Remote short blocklength process monitoring: Trade-off between resolution and data freshness," in *Proc. ICC*, Jun. 2020, pp. 1–6.

[53] D. Ramirez, E. Erkip, and H. Vincent Poor, "Age of information with finite horizon and partial updates," 2019, *arXiv:1910.00963*. [Online]. Available: http://arxiv.org/abs/1910.00963

[54] M. Bastopcu and S. Ulukus, "Age of information for updates with distortion: Constant and age-dependent distortion constraints," 2020, *arXiv:1912.13493*. [Online]. Available: https://arxiv.org/abs/1912.13493

[55] M. Bastopcu and S. Ulukus, "Partial updates: Losing information for freshness," 2020, *arXiv:2001.11014*. [Online]. Available: https://arxiv.org/abs/2001.11014

[56] A. Maatouk, S. Kriouile, M. Assaad, and A. Ephremides, "The age of incorrect information: A new performance metric for status updates," *IEEE/ACM Trans. Netw.*, vol. 28, no. 5, pp. 2215–2228, Oct. 2020.

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

[58] T. Z. Ornee and Y. Sun, "Sampling and remote estimation for the Ornstein–Uhlenbeck process through queues: Age of information and beyond," 2021, *arXiv:1902.03552*. [Online]. Available: https://arxiv.org/abs/1902.03552

[59] A. Kipnis, Y. C. Eldar, and A. J. Goldsmith, "Analog-to-digital compression: A new paradigm for converting signals to bits," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 16–39, May 2018.

[60] T. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.

[61] H. V. Poor, "Quantization effects in filtering of stationary Gaussian processes," in *Proc. IEEE CDC*, Dec. 1984, pp. 1430–1435.

[62] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[63] A. Cohen, N. Shlezinger, Y. C. Eldar, and M. Medard, "Serial quantization for representing sparse signals," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2019, pp. 987–994.

[64] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Ahmed Arafa (Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from Alexandria University, Egypt, in 2010, the M.Sc. degree in wireless technologies from the Wireless Intelligent Networks Center (WINC), Nile University, Egypt, in 2012, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 2016 and 2017, respectively. From 2017 to 2019, he has been a Postdoctoral Research Associate with the Electrical Engineering Department, Princeton University. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of North Carolina at Charlotte.

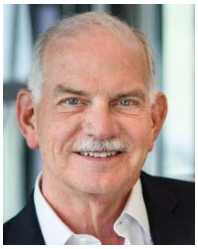
Dr. Arafa's research interests are in communication theory, information theory, machine learning, and signal processing, with recent focus on timely information processing and transfer (age-of-information), energy harvesting communications, information-theoretic security and privacy, and federated learning. He was a recipient of the Distinguished Dissertation Award from the Department of Electrical and Computer Engineering, University of Maryland, in 2017, for his Ph.D. thesis work on optimal energy management policies in energy harvesting communication networks with system costs.

Karim Banawan (Member, IEEE) received the B.Sc. and M.Sc. degrees (Hons.) in electrical engineering from Alexandria University, Alexandria, Egypt, in 2008 and 2012, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 2017 and 2018, respectively, with his Ph.D. thesis on private information retrieval and security in networks.

In 2019, he joined as an Assistant Professor with the Department of Electrical Engineering, Alexandria University. His research interests include information theory, wireless communications, physical layer security, and private information retrieval. He was a recipient of the Distinguished Dissertation Fellowship from the Department of Electrical and Computer Engineering, University of Maryland, for his Ph.D. thesis work.

Karim G. Seddik (Senior Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from Alexandria University, Alexandria, Egypt, in 2001 and 2004, respectively, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2008.

He is currently a Professor with the Electronics and Communications Engineering Department, American University in Cairo (AUC), and the Associate Dean for Graduate Studies and Research, School of Sciences and Engineering (SSE), AUC. Before joining AUC, he was an Assistant Professor with Alexandria University. His research interests include applications of machine learning in communication networks, intelligent reflecting surfaces, age of information, cognitive radio communications, and layered channel coding. He has served on the technical program committees of numerous IEEE conferences in the areas of wireless networks and mobile computing. He was a recipient of the State Encouragement Award in 2016 and the State Medal of Excellence in 2017. In 2002, he was a recipient of the Certificate of Honor from the Egyptian President for being ranked first among all departments in the College of Engineering, Alexandria University. He received the Graduate School Fellowship from the University of Maryland in 2004 and 2005 and the Future Faculty Program Fellowship from the University of Maryland in 2007. He also coauthored a conference articles that received the Best Conference Paper Award from the IEEE Communication Society Technical Committee on Green Communications and Computing in 2019.



H. Vincent Poor (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science (EECS) from Princeton University in 1977.

From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean of the Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests include information theory, machine learning and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the forthcoming book *Machine Learning and Wireless Communications* (Cambridge University Press, 2021). Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences and a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the 2017 IEEE Alexander Graham Bell Medal.