Timely Transmissions Using Optimized Variable Length Coding

Ahmed Arafa[1] and Richard D. Wesel[2]

[1]Electrical and Computer Engineering Department, University of North Carolina at Charlotte, NC 28223
[2]Department of Electrical and Computer Engineering, University of California at Los Angeles, CA 90095

***Abstract*—A status updating system is considered in which a variable length code is used to transmit messages to a receiver over a noisy channel. The goal is to optimize the codewords lengths such that successfully-decoded messages are *timely*. That is, such that the *age-of-information* (AoI) at the receiver is minimized. A hybrid ARQ (HARQ) scheme is employed, in which variable-length incremental redundancy (IR) bits are added to the originally-transmitted codeword until decoding is successful. With each decoding attempt, a non-zero processing delay is incurred. The optimal codewords lengths are analytically derived utilizing a *sequential differential optimization* (SDO) framework. The framework is general in that it only requires knowledge of an analytical expression of the positive feedback (ACK) probability as a function of the codeword length.**

I. Introduction

Status updating over noisy communication channels calls for careful coding design such that the delivered status update messages are as timely as possible. Using an age-of-information (AoI) metric to assess timeliness, defined as the time elapsed since the latest successfully-decoded message has been generated, our goal in this paper is to provide an analytical framework to optimize codewords lengths for variable length codes used in delivering timely updates.

Most previous work on systems that seek to optimize codewords for AoI minimization, as in, e.g., [1]–[8], have mainly focused on two distinct approaches, fixed redundancy (FR), in which the message is communicated with a single fixed-length transmission, and infinite incremental redundancy (IIR) schemes in which the transmission length is increased one symbol at a time until decoding is successful. Real systems often use a hybrid ARQ (HARQ) approach, as in, e.g., [9]–[11], in which the message length can be variable-length, but not at a granularity of a single symbol. HARQ systems feature an initial transmission followed by subsequent transmissions (of possibly varying lengths) of incremental redundancy that are guided by feedback from the receiver to the transmitter.

With no delay associated with decoding or requesting incremental redundancy, the pure IIR scheme is expected to provide a better AoI than the HARQ scheme that restricts the number of incremental redundancy transmissions. However, most real systems include a nonzero processing delay β corresponding to the time that it takes to decode the received codeword, transmit a negative acknowledgement (NACK) to the transmitter, and receive a subsequent incremental redundancy transmission. For a large enough β, this overhead significantly increases the AoI of the IIR approach and makes the HARQ approach preferable.

Optimizing the HARQ approach requires determination of the length of the initial transmission and each subsequent transmission of incremental redundancy. Sequential differential optimization (SDO) [12]–[14] identifies a sequence of HARQ transmission lengths that optimizes throughput. For a specified maximum number of feedback transmissions and a maximum probability that the decoder fails to produce a positive acknowledgement (ACK) even when all possible incremental redundancy has been received, SDO finds the transmission lengths that minimize average blocklength. SDO requires a known probability distribution on the probability of ACK at each cumulative blocklength, but works equally well for the variety of distributions that arise from different variable-length codes operating on different channels [14]–[16]. The original formulation of SDO minimizes the average blocklength for a fixed *maximum* number of feedback transmissions. The recent paper [17] re-frames the optimization problem using a Lagrangian approach to provide a closed-form expression for the optimal transmission lengths under a constraint on the *average* number of feedback transmissions.

This paper extends the SDO approach to determine transmission lengths that explicitly optimize AoI. Using AoI as the SDO objective function yields different optimal transmission lengths than using throughput as the objective function as in [17], since the two objectives behave differently, see, e.g., [18].

One can differentiate between the works in [1]–[11] according to 1) whether status updates are exogenous or generated at will, depending on the ability to control transmission times; and 2) whether or not replacements are allowed, depending on the ability to let new updates replace the ones in service. Our work in this paper is categorized as a *generate-at-will HARQ scheme without replacement,* and is different from related works in that a nonzero processing delay β is considered, and that the optimal set of codewords lengths that minimize the long-term average AoI is *analytically* derived.

Our case study for tail-biting convolutional codes shows that optimized HARQ beats optimized IIR and FR without replacement for all values of processing delay β.

This research is supported by National Science Foundation (NSF) grant CCF-1955660. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect views of the NSF.

II. System Model and Problem Formulation

We consider a transmitter-receiver pair communicating over a noisy memoryless channel. The transmitter generates k-bit measurements, at will, from a time-varying process. Measurements are time-stamped and sent to the receiver using ℓ_1-bit codewords, $\ell_1 \geq k$. We use the term *message* to denote a transmitted codeword. The receiver sends an ACK (a NACK) feedback following successful (unsuccessful) decoding attempts. Feedback messages are assumed to be free of errors, which is a mild assumption given the low information rate of the ACKs and NACKs. In addition to the time for message transmission, a fixed β amount of time is consumed per decoding attempt, which includes the roundtrip time for sending feedback and processing it at the transmitter. We term β the *processing delay*. A HARQ scheme is employed, in which IR bits are transmitted to help the receiver re-attempt decoding in case a NACK is fed back. IR lengths are denoted by $\{\ell_2, \ell_3, \ldots, \ell_m\}$, where m is the maximum number of transmission attempts per message. A system model overview is shown in Fig. 1.

Let us denote the cumulative blocklength by

$$N_f \triangleq \sum_{i=1}^{f} \ell_i, \quad 1 \leq f \leq m, \tag{1}$$

and let $P_{ACK}^{(N_f)}$ denote the probability of receiving an ACK while using a blocklength of N_f bits. Clearly, such probability increases with N_f. The value of N_m is chosen to be large-enough that $P_{ACK}^{(N_m)} \approx 1$, which depends on the specific code being used and the channel statistics.[2] We note that N_m is fixed, yet the value of m is *not*; it is to be optimally-determined. Our SDO methodology, however, can be altered to work for fixed N_m *and* m (cf. Section V-B).[3]

Let τ_i denote the ith *service time*: time consumed in transmitting the ith message. We consider a normalized setting in which sending a message using N_f bits consumes N_f time units. The channel is memoryless, and hence τ_i's are independent and identically distributed (i.i.d.) $\sim \tau$, which is approximately given by

$$\tau = \begin{cases} N_1 + \beta, & \text{w.p. } P_{ACK}^{(N_1)} \\ N_f + f\beta, & \text{w.p. } P_{ACK}^{(N_f)} - P_{ACK}^{(N_{f-1})}, \ f \geq 2 \end{cases}. \tag{2}$$

The above serves as a close approximation to τ under the reasonable assumption that receiving an ACK using N_f bits implies receiving an ACK using N_{f+1} bits as well. For instance, for $f = 2$, one can write

$$\begin{aligned} &\mathbb{P}\left(\tau = N_2 + 2\beta\right) = \mathbb{P}\left(\text{NACK at } N_1, \text{ ACK at } N_2\right) \\ &= \mathbb{P}\left(\text{ACK at } N_2\right) - \mathbb{P}\left(\text{ACK at } N_1, \text{ ACK at } N_2\right) \\ &= P_{ACK}^{(N_2)} - P_{ACK}^{(N_1)} + \mathbb{P}\left(\text{ACK at } N_1, \text{ NACK at } N_2\right), \end{aligned} \tag{3}$$

[2]We assume an ACK always corresponds to a successful (correct) decoding event. We ignore events in which an error bypasses the receiver undetected.

[3]Other cases, such as when N_m is variable and m is fixed, or when both are variable, are to be studied in future work.

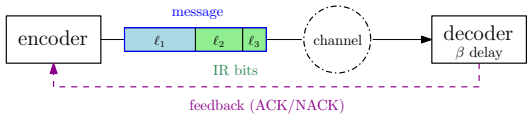

Fig. 1. Overview of the considered HARQ system model. In this example, 3 transmissions are made before successful decoding, thereby requiring $\ell_2 + \ell_3$ IR bits to be transmitted on top of the original ℓ_1 bits. A processing delay of 3β time units is incurred in total (β per decoding attempt).

whence the last term is assumed having probability ≈ 0. Similar arguments can be followed for $f > 2$.

Our goal is to design the blocklengths $\{N_f\}$ such that the long-term average AoI is minimized. The AoI at time t is

$$a(t) \triangleq t - u(t), \tag{4}$$

where $u(t)$ represents the time stamp of the latest successfully-decoded message. To minimize AoI, therefore, the transmitter should not acquire the $(i+1)$th measurement until the ith message is transmitted successfully, i.e., after (at least) τ_i time units starting from the transmission time of the ith message.

Remark 1 *It is important to note that we focus on analyzing a HARQ scheme without replacement. Specifically, it might be better, AoI-wise, to replace the current message in transmission after a certain number of NACKs, and replace it by a new, fresher, one instead. This idea has been studied in, e.g., [11] for a system with fixed $m = 2$. In this paper, we do not focus on systems that allow replacements. Instead, we aim at providing an analytical framework to design the blocklengths $\{N_f\}$ through a novel SDO approach discussed in Section III.*

Let us denote by an epoch the time elapsed in between two successful transmissions. At the beginning of the ith epoch, the transmitter idly waits for W_i time units before acquiring a new sample. Idle waiting can indeed minimize the average AoI as shown in various results of the literature, e.g., [19], [20]. In Fig. 2, we show an example of how the AoI may evolve during the ith epoch. From the figure, one can see that the ith epoch length is given by

$$L_i = W_i + \tau_i, \tag{5}$$

and the corresponding area under the AoI curve is

$$Q_i = \tau_{i-1} L_i + \frac{1}{2} L_i^2. \tag{6}$$

The sequence $\{W_i\}$ denotes a *waiting policy*. Our goal is find the optimal blocklenghts and waiting policy that minimize the long-term average AoI given by

$$\limsup_{j \to \infty} \frac{\sum_{i=1}^{j} \mathbb{E}\left[Q_i\right]}{\sum_{i=1}^{j} \mathbb{E}\left[L_i\right]}. \tag{7}$$

Since τ_i's are i.i.d., one can then conclude using the results in [19] that the optimal waiting policy has a threshold structure, in which

$$W_i = \left[\gamma - \tau_{i-1}\right]^+, \tag{8}$$

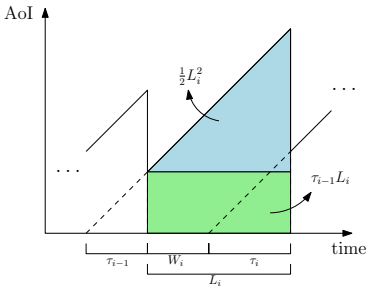

Fig. 2. An example of how the AoI may evolve in the ith epoch.

where $\gamma \geq 0$ is some threshold, and $[\cdot]^+ \triangleq \max(\cdot, 0)$. This induces a *stationary* distribution $L_i \sim L$ and $Q_i \sim Q$ for all epochs, and thereby reduces the focus to a typical epoch through removing the summations in the numerator and denominator of (7). Let us define $\overline{\tau}$ as the starting AoI of such an epoch. This allows us to write

$$\mathbb{E}[L] = \mathbb{E}\left[[\gamma - \overline{\tau}]^+\right] + \mathbb{E}[\tau], \tag{9}$$

$$\mathbb{E}[Q] = \mathbb{E}\left[\overline{\tau}\,[\gamma - \overline{\tau}]^+\right] + (\mathbb{E}[\tau])^2 + \frac{1}{2}\mathbb{E}\left[\left([\gamma - \overline{\tau}]^+ + \tau\right)^2\right]. \tag{10}$$

Our optimization problem is therefore given by

$$\begin{aligned} \min_{\{N_f\},\ \gamma \geq 0} \quad & \frac{\mathbb{E}[Q]}{\mathbb{E}[L]} \\ \text{s.t.} \quad & N_f > N_{f-1},\ N_f \in \mathbb{Z}_{++},\ \forall f \end{aligned} \tag{11}$$

with $N_0 \triangleq k$; $\mathbb{E}[L]$ and $\mathbb{E}[Q]$ given by (9) and (10), respectively; and $\overline{\tau}$ and τ i.i.d. as in (2).

One can possibly follow a decomposition approach to solve problem (11) by fixing the threshold γ and solving for the blocklengths $\{N_f\}$ in terms of γ, and then finding the optimal threshold afterwards. We realize, however, that such approach would not yield a clear analytical solution for the blocklengths, *which is one fundamental goal for this paper.* Thereby, in Section III, we focus on problem (11) in the special case of a *zero-wait policy,* i.e., when $\gamma = 0$, and present a novel SDO framework to find the optimal blocklengths. After that, in Section IV, we discuss how to find the threshold based on the SDO solution (which may be suboptimal). Under a zero-wait policy, the objective function of problem (11) is simplified to

$$\rho_0 \triangleq \mathbb{E}[\tau] + \frac{\mathbb{E}\left[\tau^2\right]}{2\mathbb{E}[\tau]}. \tag{12}$$

III. The SDO Approach

In this section, we solve problem (11) for $\gamma = 0$. The SDO approach basically solves for all the blocklengths sequentially in terms of N_1. A one-dimensional search is then followed to find the optimal N_1^*, and subsequently all the other blocklengths. Such approach, however, will *not* work if we optimize ρ_0 in its current fractional form. The reason, for instance, is that the partial derivative of ρ_0 with respect to N_1 is a function of all the blocklengths, while it should only be a function of N_1 and N_2 so that the optimal N_2 can be completely characterized in terms of N_1.

In fact, as we will show, the SDO approach *will* work if ρ_0 is represented in an equivalent yet non-fractional way. Towards that end, we follow a Dinkelbach-like approach [21], and introduce the following auxiliary problem for fixed $\lambda \geq 0$:

$$\begin{aligned} p(\lambda) \triangleq \min_{\{N_f\}} \quad & (1-\lambda)\mathbb{E}[\tau] + \frac{1}{2}\mathbb{E}\left[\tau^2\right] \\ \text{s.t.} \quad & N_f > N_{f-1},\ N_f \in \mathbb{Z}_{++},\ \forall f. \end{aligned} \tag{13}$$

Let ρ_0^* denote the optimal long-term average AoI in (12). We now have the following result:

Lemma 1 *Let $\{N_f^\lambda\}$ denote the solution of problem (13), and τ_λ be the corresponding service time. It then holds that*

$$\rho_0^* = p(\lambda^*) + \lambda^*, \tag{14}$$

where $\lambda^* \triangleq \arg\min\{p(\lambda) + \lambda :\ p(\lambda) = \mathbb{E}[\tau_\lambda]\}$.

Proof: First, it is direct to see that $p(\lambda) = \mathbb{E}[\tau_\lambda] \iff \lambda = \frac{\mathbb{E}[\tau_\lambda^2]}{2\mathbb{E}[\tau_\lambda]}$, and that at such case ρ_0 would be equal to $p(\lambda) + \lambda$. It therefore follows that ρ_0^* is given by minimizing the expression $p(\lambda) + \lambda$ over all values of λ that satisfy $p(\lambda) = \mathbb{E}[\tau_\lambda]$. Next, one can show that $p(\lambda)$ is decreasing in λ. In particular, there exists some $\lambda_{\max}$ such that $p(\lambda_{\max}) < 0$. This shows that the set $\{\lambda :\ p(\lambda) = \mathbb{E}[\tau_\lambda]\}$ is non-empty and λ^* exists. ■

Lemma 1 shows that one can find the optimal long-term average AoI in (12) by focusing on solving problem (13) at a specific λ^*. The value of λ^* can be found via, e.g., a one-dimensional search over the interval $[0, \lambda_{\max}]$, where $\lambda_{\max}$ is a large-enough value of λ such that $p(\lambda_{\max}) < 0$. We observe that for the case of the convolutional codes studied in Section V, such λ^* is also unique (cf. Fig. 3).

Given this auxiliary result, we now discuss how to use SDO to find the optimal codewords lengths for fixed λ by solving problem (13). First, let us relax the problem by ignoring the integer constraints on the blocklengths and solving for real values of $\{N_f\}$. Imposing the integer constraints back on the acquired solutions can be handled, e.g., via the dithering approach proposed in [17, Section IV-B]. In our work, we follow a rounding approach instead to project the optimal blocklengths onto $\mathbb{Z}_{++}$, yet we do so *simultaneously* after solving for all of them. We observe that such rounding approach has a negligible effect on optimality especially for relatively large blocklengths, as discussed in Section V.

Next we elaborate on the partial derivatives of the first and second moments of τ with respect to the blocklenghts $\{N_f\}$. Using (2), the first moment is given by

$$\begin{aligned} \mathbb{E}[\tau] = {} & (N_1 + \beta)\, P_{ACK}^{(N_1)} \\ & + \sum_{f=2}^{m-1} (N_f + f\beta)\left(P_{ACK}^{(N_f)} - P_{ACK}^{(N_{f-1})}\right) \\ & + (N_m + m\beta)\left(1 - P_{ACK}^{(N_{m-1})}\right), \end{aligned} \tag{15}$$

whose partial derivatives are given by

$$\frac{\partial \mathbb{E}\left[\tau\right]}{\partial N_1} = P_{ACK}^{(N_1)} + \left(N_1 + \beta - \left(N_2 + 2\beta\right)\right) P_{ACK}^{\prime(N_1)}, \tag{16}$$

$$\frac{\partial \mathbb{E}\left[\tau\right]}{\partial N_f} = P_{ACK}^{(N_f)} - P_{ACK}^{(N_{f-1})} + \left(N_f + f\beta - \left(N_{f+1} + (f+1)\beta\right)\right) P_{ACK}^{\prime(N_f)}, \tag{17}$$

for $2 \leq f \leq m-1$, where $P_{ACK}^{\prime(N_f)}$ denotes the derivative $\frac{dP_{ACK}^{(N_f)}}{dN_f}$. Similarly, the second moment is expressed as

$$\mathbb{E}\left[\tau^2\right] = \left(N_1 + \beta\right)^2 P_{ACK}^{(N_1)} + \sum_{f=2}^{m-1} \left(N_f + f\beta\right)^2 \left(P_{ACK}^{(N_f)} - P_{ACK}^{(N_{f-1})}\right) + \left(N_m + m\beta\right)^2 \left(1 - P_{ACK}^{(N_{m-1})}\right), \tag{18}$$

whose partial derivatives are given by

$$\frac{\partial \mathbb{E}\left[\tau^2\right]}{\partial N_1} = 2\left(N_1 + \beta\right) P_{ACK}^{(N_1)} + \left(\left(N_1 + \beta\right)^2 - \left(N_2 + 2\beta\right)^2\right) P_{ACK}^{\prime(N_1)}, \tag{19}$$

$$\frac{\partial \mathbb{E}\left[\tau^2\right]}{\partial N_f} = 2\left(N_f + f\beta\right)\left(P_{ACK}^{(N_f)} - P_{ACK}^{(N_{f-1})}\right) + \left(\left(N_f + f\beta\right)^2 - \left(N_{f+1} + (f+1)\beta\right)^2\right) P_{ACK}^{\prime(N_f)}, \tag{20}$$

for $2 \leq f \leq m-1$.

Now let us take the partial derivative of the objective function of problem (13) with respect to N_1 and equate it to 0. Using the above, after some algebra we get that

$$\left(N_2 + 2\beta\right)^2 + 2(1-\lambda)\left(N_2 + 2\beta\right) - c\left(N_1, \lambda\right) = 0 \tag{21}$$

must hold, where

$$c\left(N_1, \lambda\right) \triangleq 2(1-\lambda)\left(\frac{P_{ACK}^{(N_1)}}{P_{ACK}^{\prime(N_1)}} + \left(N_1 + \beta\right)\right) + 2\left(N_1 + \beta\right)\left(\frac{P_{ACK}^{(N_1)}}{P_{ACK}^{\prime(N_1)}} + \frac{\left(N_1 + \beta\right)}{2}\right). \tag{22}$$

Now let us fix the value of N_1 ($\geq k$). If the discriminant of the quadratic equation in (21), i.e., if

$$\left(1-\lambda\right)^2 + c\left(N_1, \lambda\right) \tag{23}$$

is negative, then there do not exist any real solutions for N_2 that solve (21). This means that the fixed value of N_1 is *not optimal,* and has to change. On the other hand, if the above discriminant is non-negative, then one can get the following two solutions for N_2:

$$N_2 = -\left(1-\lambda\right) \pm \sqrt{\left(1-\lambda\right)^2 + c\left(N_1, \lambda\right)} - 2\beta. \tag{24}$$

Similarly, one can show that taking the partial derivative of the objective function of problem (13) with respect to N_f, $2 \leq f \leq m-1$, and equating it to 0 results in a quadratic equation to solve for N_{f+1} in terms of N_f and N_{f-1}. The two solutions of such equation are given by

$$N_{f+1} = -\left(1-\lambda\right) \pm \sqrt{\left(1-\lambda\right)^2 + c\left(N_f, N_{f-1}, \lambda\right)} - (f+1)\beta, \tag{25}$$

where

$$c(N_f, N_{f-1}, \lambda) \triangleq 2(1-\lambda)\left(\frac{P_{ACK}^{(N_f)} - P_{ACK}^{(N_{f-1})}}{P_{ACK}^{\prime(N_f)}} + \left(N_f + f\beta\right)\right) + 2\left(N_f + f\beta\right)\left(\frac{P_{ACK}^{(N_f)} - P_{ACK}^{(N_{f-1})}}{P_{ACK}^{\prime(N_f)}} + \frac{\left(N_f + f\beta\right)^2}{2}\right), \tag{26}$$

provided that the discriminant below is non-negative:

$$\left(1-\lambda\right)^2 + c\left(N_f, N_{f-1}, \lambda\right). \tag{27}$$

Therefore, using (24) and (25), one can characterize optimal solutions for $\{N_2, N_3, \ldots, N_{m-1}\}$ in terms of N_1. These sequential solutions would eventually stop if N_{f^*+1} surpasses N_m, for some f^*, at which point one may truncate the excess IR bits and set $N_{f^*+1} = N_m$.

Now for the solutions to be meaningful, we need to make sure that the obtained blocklengths are monotonically increasing. In most scenarios, such as in the one discussed in Section V, this would automatically cross-out the smaller solutions in (24) and (25), especially for large values of f.

For $2 \leq f \leq m-1$, in case both solutions obtained for N_f are smaller than N_{f-1}, or in case the discriminant of the quadratic equation to solve for N_{f+1} is negative, then the whole solution sequence leading to such N_f is *rejected.* If it so happens that all solution sequences are rejected, then the fixed value of N_1 is not optimal, and has to change. As noted in Section V, we observe that for large values of β, one needs to initiate SDO with a relatively large value of N_1 to get meaningful (unrejected) solution sequences. Finally, in case two or more solution sequences are obtained, we pick the one that yields a smaller objective function of problem (13).

We now summarize the SDO approach used to characterize the optimal long-term average AoI ρ_0^*. For a given λ, we first fix N_1 and sequentially solve for $\{N_2, N_3, \ldots, N_{m-1}\}$ using equations (24) and (25). We then find the best N_1, which gives $p(\lambda)$. Finally, the optimal λ^* is found as discussed in Lemma 1, which gives $\rho_0^* = p(\lambda^*) + \lambda^*$.

IV. Waiting Policy

We now consider optimizing the waiting policy by going back to problem (11). As discussed towards the end of Section II, jointly optimizing the waiting threshold γ and the blocklenghts $\{N_f\}$ would not directly yield a sequential solution as done in the previous section. We instead follow a potentially-suboptimal approach in which we first find the optimal blocklengths via SDO for a zero-wait policy, *then* we optimize the waiting threshold based on that. Therefore, in this section we assume that we already have a set of blocklengths $\{N_f\}$, with a corresponding service time random variable τ.

Now the task of finding the optimal γ^* can be accomplished by the techniques introduced in [19]. In what follows, we reiterate the procedure of finding γ^* according to our own notation, and approach it slightly differently, for completeness.

To analytically determine the optimal threshold γ^*, one can leverage (the original) Dinkelbach's approach [21] for some fixed $\eta \geq 0$ and define

$$q(\eta) \triangleq \min_{\gamma \geq 0} \mathbb{E}[Q] - \eta \mathbb{E}[L], \tag{28}$$

with $\mathbb{E}[L]$ and $\mathbb{E}[Q]$ given by (9) and (10), respectively. Next, one can show that the following holds:

$$\frac{d\mathbb{E}[Q]}{d\gamma} = (\gamma + \mathbb{E}[\tau])\,\mathbb{P}(\tau \leq \gamma), \quad \frac{d\mathbb{E}[L]}{d\gamma} = \mathbb{P}(\tau \leq \gamma). \tag{29}$$

Therefore, after setting $\frac{d(\mathbb{E}[Q]-\eta\mathbb{E}[L])}{d\gamma} = 0$, the optimal threshold will be given by

$$\gamma^* = \eta^* - \mathbb{E}[\tau], \tag{30}$$

where η^* is the unique solution of $q(\eta^*) = 0$, which can be found via, e.g., a bisection search [21].

We note that $\gamma^* > 0$, and is therefore a meaningful threshold. This can be seen by observing that

$$q(\mathbb{E}[\tau]) = \mathbb{E}\left[\overline{\tau}\,[\gamma - \overline{\tau}]^+\right] + \frac{1}{2}\mathbb{E}\left[\left([\gamma-\overline{\tau}]^+\right)^2\right] + \frac{1}{2}\mathbb{E}[\tau^2], \tag{31}$$

which is strictly positive. Since $q(\eta)$ is decreasing [21], we must have $\eta^* > \mathbb{E}[\tau]$ in order for $q(\eta^*) = 0$ to hold.

V. Case Study: Convolutional Codes

We apply the above analysis to the case of tail-biting convolutional codes over additive white Gaussian noise (AWGN) channels. As shown in [14] for binary inputs with a signal-to-noise ratio (SNR) of 2 dB, the Gaussian distribution closely-approximates the ACK probability as follows:

$$P_{ACK}^{(N_f)} \approx Q\left(\frac{k/N_f - 0.5666}{0.0573}\right), \tag{32}$$

where $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{\frac{-u^2}{2}} du$ is the Q-function. We set the measurement length to $k = 64$ bits and $N_m = 192$ bits. Our results are in the context of the model in (2) and (32).

A. Verifying Lemma 1

We first verify the results of Lemma 1. For a system with $\beta = 10$ time units, we plot both $\mathbb{E}[\tau^\lambda]$ and $p(\lambda)$ versus λ in Fig. 3. We see that $\mathbb{E}[\tau_\lambda]$ is increasing with λ. This makes the set $\{\lambda : p(\lambda) = \mathbb{E}[\tau_\lambda]\}$ basically a *singleton,* which further facilitates evaluating λ^* through a bisection search over $[0, \lambda_{\max}]$. We note that such case holds for all values of β.

Next, we show how the optimal long-term average AoI behaves as a function of N_1. That is, we solve for $\rho_0^*(N_1)$ as opposed to ρ_0^*. We do so via slightly modifying the SDO approach. Specifically, now that N_1 is fixed, we substitute in (12) to get a relatively new metric $\rho_0(N_1)$ to be optimized by choosing $\{N_2, N_3, N_4, \ldots, N_{m-1}\}$. For that, we follow the same SDO approach discussed in Section III, yet after

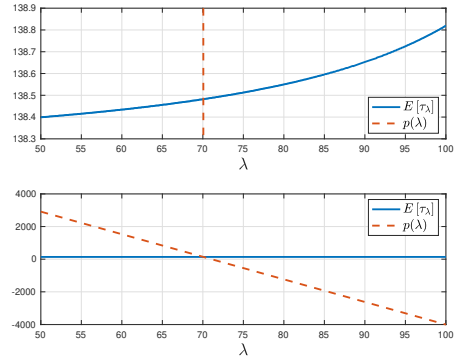

Fig. 3. $p(\lambda)$ and the optimal average service time $\mathbb{E}[\tau_\lambda]$ vs. λ, with $\beta = 10$ time units. Top plot is a zoomed-in version of bottom plot. There exists a unique $\lambda^* \approx 70$ such that $p(\lambda^*) = \mathbb{E}[\tau_{\lambda^*}]$, at which $p(\lambda^*) \approx 138$.

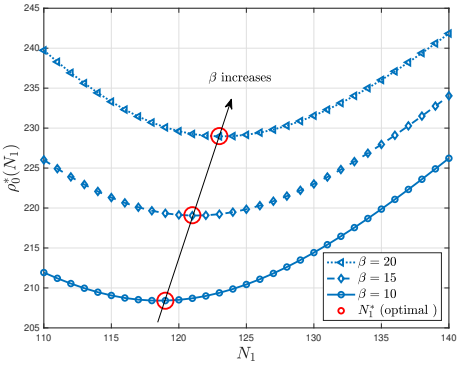

Fig. 4. Optimal long-term average AoI as a function of N_1, with different β's. The optimal N_1^* is denoted by red circles. For $\beta = 10$, the optimal $N_1^* = 119$ bits, with $\rho_0^*(119) \approx 208$ time units.

replacing N_1 with N_2. The result is shown in Fig. 4 for $\beta \in \{10, 15, 20\}$. We see that the optimal N_1^* that minimizes $\rho_0^*(N_1)$ is relatively mid-range and, intuitively, increases with β. Combining the results of Fig. 3 and Fig. 4, we observe that at $\beta = 10$, $p(\lambda^*) + \lambda^* = \rho_0^*(N_1^*)$, as asserted in Lemma 1.

B. A Methodology for fixed m

In Fig. 5, we show how the optimal blocklengths vary with N_1 for $\beta = 10$. We see that as N_1 increases, the set of blocklengths becomes sparser, i.e., fewer number of IR transmissions leads to reaching N_m. This figure, together with Fig. 4 can be used to solve the problem with *fixed number of transmissions per message* m, which may be relevant in some practical systems. For instance, at $N_1^* = 119$ we have $m = 6$ transmissions. If we have a constraint of only $m = 5$, then we would have to use $N_1 \geq 137$ according to Fig. 5. We would then examine Fig. 4 to conclude that $N_1 = 137$ is the optimal choice in this case since it attains the smallest AoI for $\beta = 10$ when compared to higher values of N_1.

C. Comparison to Baseline Schemes: IIR and FR

We compare the proposed HARQ scheme with other baseline schemes. The first is IIR, in which incremental bits are added *one-by-one* until success. This is a special case of HARQ in which $N_{f+1} = N_f + 1$, $\forall f$ (presuming that m

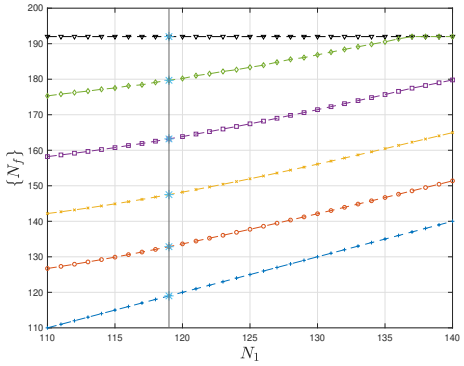

Fig. 5. Optimal IR lengths vs. N_1 using SDO, with $\beta = 10$ time units. The optimal set of blocklengths is at $N_1^* = 119$ and are denoted by $*$.

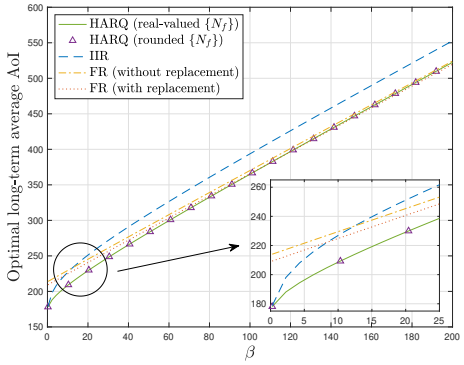

Fig. 6. Proposed HARQ and baselines (IIR, and FR with and without replacement) vs. β. Triangles denote rounded (integer) blocklengths.

can be arbitrarily large). The second baseline scheme is FR, for which we consider two subcases: with and without replacement. FR without replacement is basically using a fixed N_1 to transmit each message, with repetition in case of failures. This makes the service time given by $(N_1+\beta)M$, where M is a geometric random variable with parameter $P_{ACK}^{(N_1)}$. FR with replacement is strictly better than FR without replacement in the sense it uses fresh measurements after failures. This makes the epoch length also given by $(N_1+\beta)M$, yet the service time is *fixed* at $N_1 + \beta$. For IIR and FR without replacement, one can jointly optimize N_1 and the optimal waiting threshold in (30).[4] For FR with replacement, a zero-wait policy is optimal, see [20, Theorem 2], and the long-term average AoI can be shown to be equal to $(N_1 + \beta)\left(1/P_{ACK}^{(N_1)} + 1/2\right)$.

In Fig. 6, we show how the optimal long-term average AoI for the proposed HARQ scheme performs as a function of β, compared to the baselines. We also plot the AoI achieved by HARQ after rounding the blocklengths to their nearest integer values; we see that the performance is almost identical after rounding as noted in Section III. The HARQ scheme outperforms IIR and FR without replacement for all values of β. In addition, it outperforms FR with replacement for $\beta \lesssim 120$, and performs very close to it for $\beta \gtrsim 120$. This latter slight under performance is due to the fact that we do not allow replacements in the current analysis of HARQ.

[4]Different from HARQ, this joint optimization can be optimally solved.

VI. Conclusion

An SDO-based analytical framework has been developed to produce AoI-minimal HARQ transmission lengths. Different from almost all of the AoI-related literature on coding design, a nonzero processing delay is considered in our system, which includes the time to decode a message, send feedback and initiate the transmission of IR bits if needed. The optimized HARQ scheme beats multiple baselines such as IIR and FR.

Future work includes developing an SDO-based framework for HARQ in systems that allow message replacement.

References

[1] E. Najm, R. D. Yates, and E. Soljanin. Status updates through M/G/1/1 queues with HARQ. In *Proc. IEEE ISIT*, June 2017.

[2] H. Sac, B. T. Bacinoglu, E. Uysal-Biyikoglu, and G. Durisi. Age-optimal channel coding blocklength for an M/G/1 queue with HARQ. In *Proc. IEEE SPAWC*, June 2018.

[3] 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. IEEE ISIT*, June 2018.

[4] R. D. Yates, E. Najm, E. Soljanin, and J. Zhong. Timely updates over an erasure channel. In *Proc. IEEE ISIT*, June 2017.

[5] A. Baknina and S. Ulukus. Coded status updates in an energy harvesting erasure channel. In *Proc. CISS*, March 2018.

[6] S. Feng and J. Yang. Age-optimal transmission of rateless codes in an erasure channel. In *Proc. IEEE ICC*, May, 2019.

[7] E. Najm, E. Telatar, and R. Nasser. Optimal age over erasure channels. Available Online: arXiv:1901.01573.

[8] A. Javani, M. Zorgui, and Z. Wang. On the age of information in erasure channels with feedback. Available Online: arXiv:1911.05840.

[9] P. Parag, A. Taghavi, and J.-F. Chamberland. On real-time status updates over symbol erasure channels. In *Proc. IEEE WCNC*, March 2017.

[10] E. T. Ceran, D. Gunduz, and A. Gyorgy. Average age of information with hybrid ARQ under a resource constraint. In *Proc. IEEE WCNC*, April 2018.

[11] A. Arafa, K. Banawan, K. G. Seddik, and H. V. Poor. On timely channel coding with hybrid ARQ. In *Proc. IEEE Globecom*, December 2019.

[12] K. Vakilinia, A. R. Williamson, S. V. S. Ranganathan, D. Divsalar, and R. D. Wesel. Feedback systems using non-binary LDPC codes with a limited number of transmissions. In *Proc. IEEE ITW*, November 2014.

[13] K. Vakilinia, S. V. S. Ranganathan, D. Divsalar, and R. D. Wesel. Optimizing transmission lengths for limited feedback with nonbinary LDPC examples. *IEEE Trans. Commun.*, 64(6):2245–2257, June 2016.

[14] N. Wong, K. Vakilinia, H. Wang, S. V. S. Ranganathan, and R. D. Wesel. Sequential differential optimization of incremental redundancy transmission lengths: An example with tail-biting convolutional codes. In *Proc. ITA*, February 2017.

[15] H. Wang, N. Wong, A. M. Baldauf, C. K. Bachelor, S. V. S. Ranganathan, D. Divsalar, and R. D. Wesel. An information density approach to analyzing and optimizing incremental redundancy with feedback. In *Proc. IEEE ISIT*, June 2017.

[16] A. Heidarzadeh, J.-F. Chamberland, P. Parag, and R. D. Wesel. A systematic approach to incremental redundancy over erasure channel. In *Proc. IEEE ISIT*, June 2018.

[17] R. D. Wesel, N. Wong, A. M. Baldauf, A. Belhouchat, A. Heidarzadeh, and J.-F. Chamberland. Transmission lengths that maximize throughput of variable-length coding & ACK/NACK feedback. In *Proc. IEEE Globecom*, December 2018.

[18] S. K. Kaul, R. D. Yates, and M. Gruteser. Real-time status: How often should one update? In *Proc. IEEE Infocom*, March 2012.

[19] 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*, 63(11):7492–7508, November 2017.

[20] A. Arafa, K. Banawan, K. G. Seddik, and H. V. Poor. Timely estimation using coded quantized samples. In *Proc. IEEE ISIT*, June 2020.

[21] W. Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967.