A Systematic Approach to Incremental Redundancy With Application to Erasure Channels

Anoosheh Heidarzadeh, *Member, IEEE*, Jean-Francois Chamberland, *Senior Member, IEEE*, Richard D. Wesel, *Senior Member, IEEE*, and Parimal Parag, *Member, IEEE*

***Abstract*— This paper focuses on the design and evaluation of pragmatic schemes for delay-sensitive communication. Specifically, this contribution studies the operation of data links that employ incremental redundancy as a means to shield information bits from the degradation associated with unreliable channels. While this inquiry puts forth a general methodology, exposition centers around erasure channels because they are well suited for analysis. Nevertheless, the goal is to identify both structural properties and design guidelines that are broadly applicable. Conceptually, this paper leverages a methodology, termed sequential differential optimization, aimed at identifying near-optimal block sizes for hybrid ARQ. This technique is applied to erasure channels and it is extended to scenarios where throughput is maximized subject to a constraint on the feedback rate. The analysis shows that the impact of the coding strategy adopted and the propensity of the channel to erase symbols naturally decouple when maximizing throughput. Ultimately, block size selection is informed by approximate distributions on the probability of decoding success at every stage of the incremental transmission process. This novel perspective, which rigorously bridges hybrid automatic repeat request and coding, offers a computationally efficient framework to select code rates and blocklengths for incremental redundancy. These findings are supported through numerical results.**

***Index Terms*— Feedback communication, optimization methods, variable length codes, error correction coding, channel coding, automatic repeat request, maximum likelihood decoding.**

I. Introduction

AS THE reach of the Internet stretches beyond traditional applications to integrate sensing, actuation, and cyber-physical systems, there is a need to better understand delay-sensitive communication over unreliable channels. The rising popularity of interactive communications, live gaming over mobile devices, and augmented reality contributes to a growing interest in low-latency connections. These circumstances have been a key motivating factor underlying several recent inquiries pertaining to information transfers under stringent delay constraints. Such contributions include the divergence framework for short blocklengths [1], [2], the interplay between coding and queueing [3], and ongoing work on the age of information [4], [5].

Hybrid automatic repeat request (ARQ) has been identified as a central approach to deliver information in a timely manner over unreliable channels [6]. It can be designed to adapt gracefully to channel degradations associated with fading and interference, and it has found wide application in theory and practice [7], [8]. Conceptually, hybrid ARQ is a means to leverage limited feedback between a source and its destination to ensure the timely delivery of information, especially in short blocklength regimes. Researchers have developed techniques to analyze the benefits of communication systems with hybrid ARQ [9], [10]. Yet, until recently, brute force searches, simulation studies, and ad hoc schemes remained the primary means of parameter selection in terms of blocklengths and code rate for such systems [11]. This situation changed when Vakilinia *et al.* [12], [13] introduced a novel approach for parameter selection. Their proposed methodology captures the effects of the physical channel on code performance by defining an approximate empirical distribution on the probability that a rate compatible code decodes successfully at each of its available rates. Based on the ensuing distribution, the authors then put forth a numerically efficient, sequential differential optimization (SDO) algorithm that yields best operational parameters for hybrid ARQ.

In [14], SDO is applied to erasure channels where the objective is to maximize throughput subject to a limit on the number of hybrid ARQ sub-blocks. Extending this recent contribution, the present article offers a novel geometric interpretation for the SDO technique, and it introduces a novel framework for constraining the feedback rate as opposed to the maximum number of hybrid ARQ sub-blocks. To illustrate the potential of the proposed technique, we employ the augmented framework on an erasure channel and demonstrate the value of our approach by characterizing overall performance for a class of random linear codes. System performance is measured in terms of channel throughput and feedback overhead, while maintaining the probability of decoding failure below a

Manuscript received July 19, 2018; revised November 7, 2018; accepted December 13, 2018. Date of publication December 24, 2018; date of current version April 16, 2019. This material is based on work supported by the National Science Foundation under Grants No. CCF-1619085, CCF-1618272, CNS-1642983, and CCF-1718658, and by the Defence Research and Development Organization under the Grant No. DRDO-0654. This paper was presented in part at the IEEE International Symposium on Information Theory 2018. The associate editor coordinating the review of this paper and approving it for publication was Q. Huang. *(Corresponding author: Anoosheh Heidarzadeh.)*

A. Heidarzadeh and J.-F. Chamberland are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: anoosheh@tamu.edu; chmbrlnd@tamu.edu).

R. D. Wesel is with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: wesel@ucla.edu).

P. Parag is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru 560012, India (e-mail: parimal@iisc.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2018.2889254

prescribed threshold. This contribution is significant in that it provides a widely applicable algorithmic blueprint for parameter selection in hybrid ARQ with rate compatible codes, a popular combination in the literature [15]. In addition, we offer a new visual interpretation for this optimization problem and its solution.

It should be noted that the proposed framework can also be applied to scenarios with different codes and other types of channels, some of which have been studied in the context of hybrid ARQ. This includes coding schemes such as LDPC codes [16], rate-compatible LDPC codes [17], [18], polar codes [19]–[21], and rate-compatible polar codes [22]. Likewise, alternate channel models have received attention, including AWGN and Rayleigh fading [17]–[22]. For combinations of such codes and channels, the probability of decoding success at any given time can be evaluated numerically, and the proposed SDO-based framework can potentially be utilized to optimize the parameters of a hybrid ARQ scheme. Still, our choice of random linear codes and the erasure channels in this article is primarily motivated by ease of exposition. In particular, not only is the theoretical analysis of our system tractable, it also provides a pragmatic proxy for more sophisticated codes and channels.

For erasure channels, the analysis reveals a clear separation between the effects of the unreliable channel and the attributes of the underlying code in selecting block sizes. In this context, a systematic approach that links decoding success to the number of observed symbols is derived based on moment matching. This proposed technique builds on the asymptotic behavior of random linear codes, and their connection to well-known constants in number theory, namely, the Erdös-Borwein constant (OEIS: A065442) and the digital search tree constant (OEIS: A065443). Altogether, the performance of a system with incremental redundancy hinges on three main components: the coding scheme employed, the behavior of the channel, and the quantization effects associated with hybrid ARQ blocks. Using the tools developed herein, it is possible to revisit many scenarios where the performance of traditional systems is compared to that of hybrid ARQ, albeit using optimal design parameters.

II. System Model and Renewal Structure

The scenario we wish to explore is a classical point-to-point communication system where a source seeks to transmit information to a destination over an unreliable, memoryless channel. Information bits are protected from the effects of channel variations through forward error correction. The focus is on practical schemes with finite block lengths [23], [24]. Specifically, suitable performance is realized using incremental redundancy in the form of hybrid ARQ. The system architecture assumes that the destination is capable of supplying acknowledgement bits (ACK/NACK) to the source in a faithful, timely manner. While feedback is present, it is pertinent to mention that feedback rate can be tuned via a cost structure in the upcoming analysis. The design goal is to maximize throughput subject to constraints on the probability of decoding failure, the maximum number of feedback messages and, possibly, the average feedback rate.

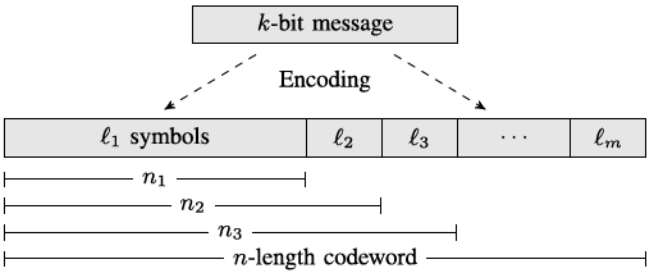

Fig. 1. This diagram shows how a k-bit message is encoded into a codeword with n symbols. The codeword is then partitioned into sub-blocks. These blocks are sent sequentially to the destination, as dictated by hybrid ARQ.

Conceptually, this article extends the sequential differential optimization (SDO) methodology [13] to account for feedback rate. As mentioned above, this technique provides an algorithmic platform to select sizes for sub-blocks in incremental redundancy in an efficient manner. In particular, SDO offers a straightforward iterative procedure to identify admissible assignments for optimally solving this resource allocation problem, a solution to which would otherwise demand a high-dimensional search. The contribution of this article is threefold. We show how an extended version of SDO can be employed to control average feedback rate. In a novel application of the SDO framework, we demonstrate that this approach is naturally suited to erasure channels. Thirdly, we introduce a geometric interpretation for SDO that offers new insight about the design task at hand. Before discussing these results in detail, we must review modeling assumptions, notation, and other preliminaries.

A. Forward Error Correction and Hybrid ARQ

The source wishes to convey a k-bit message to the destination. This message is encoded into a codeword of length n for eventual transmission over the unreliable channel. Coded symbols are sent in waves using hybrid ARQ. That is, the codeword is partitioned into m blocks of symbols, each of size $\ell_i \geq 0$. The total length of the codeword being fixed, we necessarily have $\sum_{i=1}^{m} \ell_i = n$. For notational convenience, we introduce the partial sums $n_j = \sum_{i=1}^{j} \ell_i$. A graphical illustration of these quantities appears in Fig. 1.

The source initiates the transmission process by sending the first n_1 coded symbols. After completing this initial phase, the destination attempts to decode the original message, treating unaccounted symbols as erasures. If decoding succeeds, the receiver acknowledges reception of the message (ACK), and the source proceeds to the next message. Otherwise, the receiver notifies the source of its failed attempt (NACK), and thereby requests transmission of an additional sub-block of ℓ_2 symbols. Once received, these extra symbols, which can be regarded as incremental redundancy, improve the probability of decoding success at the destination. At every intermediate stage, a similar process takes place with a supplemental block of symbols being sent, followed by a decoding attempt, and a feedback notification (ACK/NACK). If decoding fails at the last step, then the n received symbols are discarded and the process begins anew. Implicit in this scheme is the capability

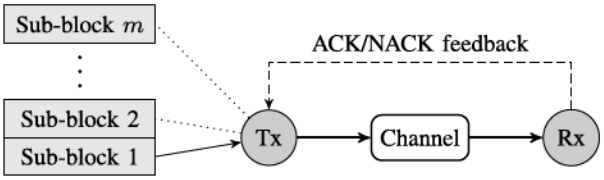

Fig. 2. Under hybrid ARQ, a communication round begins with the transmission of a sub-block. The destination tries to decode based on the received information. If unsuccessful, an additional sub-block is requested; otherwise, the source moves on to the next message. The hybrid ARQ round continues until the original message is recovered at the destination or all available sub-blocks have been exhausted.

by the receiver to accurately assess the outcome of a decoding attempt and, potentially, a resilience to the rare occurrence of an undetected decoding failure. The elapsed time between the onset of the k-bit transmission process and its eventual conclusion, either through an early ACK or once maximum sub-block m has been passed on (whichever comes first), is referred to as one round of the hybrid ARQ process. The parameters of this standard hybrid ARQ scheme are summarized in Fig. 2.

B. Performance Analysis for Memoryless Channels

The performance criteria involve system throughput, probability of decoding failure, and feedback rate. These measures are determined by the nature of the underlying communication channel and the properties of the forward error correction scheme put in place to protect information bits. For memoryless channels, a fundamental attribute that ties hybrid ARQ to these quantities is the probability of decoding success.

Suppose that the hybrid ARQ scheme employs a length assignment $\mathbf{n} = (n_1, \ldots, n_m) \in \mathbb{N}^m$, where m is the index of the last possible sub-block. Furthermore, let $P_{\mathrm{ACK}}(n_i)$ designate the probability that the destination decodes the original message successfully using at most i sub-blocks. We write $P_{\mathrm{NACK}}(n_i) = 1 - P_{\mathrm{ACK}}(n_i)$ to denote the probability that the destination requests an additional set of symbols after i sub-blocks have already been transmitted. The number of sub-blocks sent within one instance of the hybrid ARQ process is random; its value may depend on the channel and code realizations. As such, we introduce random variable S and denote the probability mass function (PMF),

$$\Pr(S = i) = \begin{cases} P_{\mathrm{ACK}}(n_i) - P_{\mathrm{ACK}}(n_{i-1}) & i = 1, \ldots, m-1 \\ 1 - P_{\mathrm{ACK}}(n_{m-1}) & i = m. \end{cases}$$

In general, the behavior of S adequately captures the number of sub-blocks used within one round of the hybrid ARQ process and, for sensible coding strategies over erasure channels, their distributions match exactly. Since S is a non-negative discrete random variable, we can compute its mean as

$$\mathrm{E}[S] = \sum_{i=1}^{m} \Pr(S \geq i) = m - \sum_{i=1}^{m-1} P_{\mathrm{ACK}}(n_i). \tag{1}$$

Another pertinent expectation for the problem at hand is the expected block length,

$$\begin{aligned} \mathrm{E}[n_S] &= \sum_{t=1}^{\infty} \Pr(n_S \geq t) = \sum_{t=0}^{\infty} \Pr(n_S > t) \\ &= n_1 + \sum_{i=1}^{m-1} (n_{i+1} - n_i) \Pr(n_S > n_i) \\ &= n_1 + \sum_{i=1}^{m-1} (n_{i+1} - n_i) \left(1 - \Pr(n_S \leq n_i)\right) \\ &= n_m - \sum_{i=1}^{m-1} (n_{i+1} - n_i) P_{\mathrm{ACK}}(n_i). \end{aligned} \tag{2}$$

Having established these expressions, we turn to renewal theory to calculate average throughput and feedback rate for this point-to-point communication system.

C. Renewal Structure

Owing to the structure of a memoryless channel, the inter-completion times for hybrid ARQ rounds are independent and identically distributed. From this perspective, the number of hybrid ARQ rounds as a function of time forms a renewal process [25]. A similar statement applies to the number of feedback bits sent within a round, one bit from the destination to the source per sub-block, as the hybrid ARQ scheme progresses.

Formally, consider a transmission setting where the completion of an hybrid ARQ round immediately leads to the beginning of the next round. This corresponds to an infinite backlog at the source, the standard setting to examine maximum throughput. Let S_r be the number of sub-blocks used in the rth hybrid ARQ round. We emphasize that $\{n_{S_r}, r = 1, 2, \ldots\}$ can then be interpreted as the time between the completion of the $(r-1)$th hybrid ARQ round and that of the rth round. Following common renewal notation, we let $R_0 = 0$ and

$$R_r = \sum_{q=1}^{r} n_{S_q} \quad r \geq 1.$$

Accordingly, R_r becomes the completion time of the rth round. Since the number of finished rounds by time t amounts to the largest value of r for which the rth round is completed before or at time t, we can write

$$R(t) = \sup\{r : R_r \leq t\}.$$

In words, $R(t)$ denotes the number of completed hybrid ARQ rounds at time t. Furthermore, given that n_S is a non-negative random variable with finite support, we immediately get

$$\lim_{t \to \infty} R(t) = \infty \quad \text{almost surely.}$$

Expressing the renewal function as $\mathrm{E}[R(t)]$, we can apply the elementary renewal theorem [25, Th. 3.3.4], which yields

$$\lim_{t \to \infty} \frac{\mathrm{E}[R(t)]}{t} = \frac{1}{\mathrm{E}[n_S]}.$$

Average throughput and feedback rate can be analyzed based on this renewal structure. In these latter two cases, the renewal reward framework applies.

For throughput, the reward structure is k information bits when the message is decoded successfully at the destination; and no information bits otherwise. The completed work at time t can be expressed as

$$W(t) = \sum_{r=1}^{R(t)} W_r,$$

where W_r is the number of information bits successfully received at the destination during round r. Then, we have

$$\mathrm{E}[W_r] = \mathrm{E}[W] = kP_{\mathrm{ACK}}(n_m).$$

The renewal theorem for reward processes [25, Th. 3.6.1] delivers the desired expression for throughput,

$$\lim_{t\to\infty} \frac{\mathrm{E}[W(t)]}{t} = \frac{\mathrm{E}[R]}{\mathrm{E}[n_S]} = \frac{kP_{\mathrm{ACK}}(n_m)}{\mathrm{E}[n_S]}. \tag{3}$$

We emphasize that the necessary conditions for the theorem, $\mathrm{E}[R] < \infty$ and $\mathrm{E}[n_S] < \infty$, are readily satisfied in view of the fact that these random variables have finite support.

Regarding feedback bits as cost, one can also compute the average feedback rate using the renewal theorem for reward processes. In this case, the cost is captured by S_r, the number of feedback bits employed in round r. The number of feedback bits accumulated by time t is

$$S(t) = \sum_{r=1}^{R(t)} S_r,$$

and the feedback rate is therefore given by

$$\lim_{t\to\infty} \frac{\mathrm{E}[S(t)]}{t} = \frac{\mathrm{E}[S]}{\mathrm{E}[n_S]}. \tag{4}$$

As before, necessary conditions $\mathrm{E}[S] < \infty$ and $\mathrm{E}[n_S] < \infty$ for the renewal reward theorem are immediate because S and n_S have finite support.

D. Consolidated Optimization Framework

The expressions derived in Section II-C are accurate for any specific length assignment $\mathbf{n} = (n_1, \ldots, n_m)$. Yet, in comparing potential assignments, it is crucial to develop a unified framework. This is accomplished by choosing a smooth approximation for the cumulative distribution function (CDF) for the initial point at which a message becomes decodable. Mathematically, the extended SDO methodology derived herein relies on the availability of a strictly increasing, differentiable function $F(\cdot)$ such that

$$P_{\mathrm{ACK}}(t) \approx F(t) \tag{5}$$

for every vector assignment $\mathbf{n}$ and integer $t \geq 0$. Fortunately, as we will see shortly, finding such an approximation is straightforward for the operational scenarios we wish to study. As a side note, we stress that the same type of approximations that underlie the vast body of work on dispersion [26] can be leveraged in the current context as well. Moreover, the rapid concentration of empirical measures for memoryless channels too points at the existence of accurate approximations for most practical scenarios.

III. Sequential Differential Optimization

At this stage, we are in a position to formally state the class of optimization problems we wish to study and, subsequently, extend sequential differential optimization (SDO) as a platform to obtain appropriate solutions. Throughout this section, we embrace approximation (5) as a proxy for code performance. In particular, $F(\cdot)$ denotes the CDF of a continuous probability distribution; and $F(t)$ captures the probability of the receiver being able to successfully decode the original message after at most t symbols have been transmitted. As is customary, we use $f(\cdot)$ to denote the PDF associated with $F(\cdot)$; that is,

$$f(t) = \left.\frac{dF(x)}{dx}\right|_{x=t}.$$

To prevent confusion and because it appears in several expressions, we retain the use of $P_{\mathrm{ACK}}(n_m)$ at the maximum length of a codeword, $n = n_m$.

A. Throughput Optimization

As mentioned above, our initial design goal is to select $\mathbf{n}$ as to maximize average throughput, while maintaining the probability of decoding failure for any given round below a prescribed threshold δ.

Problem 1: Find an optimal block assignment vector $\mathbf{n} = (n_1, \ldots, n_m)$ for the following optimization problem,

$$\begin{aligned} &\underset{n_1,\ldots,n_m}{\text{maximize}} \ \frac{kP_{\mathrm{ACK}}(n_m)}{\mathrm{E}[n_S]} \\ &\text{subject to } P_{\mathrm{NACK}}(n_m) \leq \delta. \end{aligned}$$

Pragmatically, the solution to Problem 1 must assume an integer form, $\mathbf{n} = (n_1, \ldots, n_m) \in \mathbb{N}^m$. Yet, integer programs are known to be challenging and, consequently, we first consider the relaxed version of the problem where $\mathbf{n} \in \mathbb{R}_+^m$. We can employ the method of Lagrange multipliers to identify candidate local maxima corresponding to this constrained optimization problem. We note that the throughput and the probability of decoding failure have continuous partial derivatives for the relaxed version of Problem 1. Moving forward, we introduce multiplier λ_δ into the formulation, and we examine the Lagrangian expression defined by

$$J(\mathbf{n}, \lambda_\delta) = \frac{kP_{\mathrm{ACK}}(n_m)}{\mathrm{E}[n_S]} - \lambda_\delta \left(P_{\mathrm{NACK}}(n_m) - \delta\right).$$

The Karush-Kuhn-Tucker (KKT) conditions associated with $J(\mathbf{n}, \lambda_\delta)$ are given by equation $\nabla J(\mathbf{n}, \lambda_\delta) = \mathbf{0}$. We note that $P_{\mathrm{NACK}}(n_m)$ is completely determined by n_m. This considerably simplifies the form of these necessary conditions. Taking the partial derivative of $J(\mathbf{n}, \lambda_\delta)$ with respect to n_1 and setting it equal to zero, we get

$$n_2 = n_1 + \frac{F(n_1)}{f(n_1)}.$$

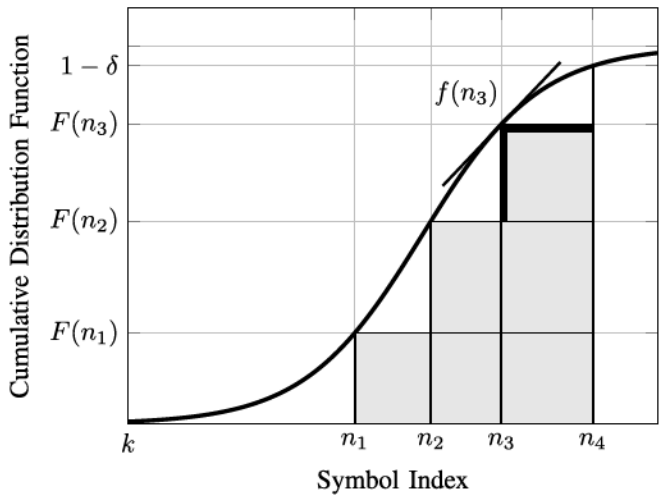

Fig. 3. The task of selecting vector $\mathbf{n} = (n_1, \ldots, n_4)$ is mathematically equivalent to finding the best 3-level Lebesgue integral of the CDF $F(\cdot)$ over the range $[0, n_4]$. The KKT conditions require the two black bands to have a same area.

In evaluating the derivative, we make use of expression (2) and approximation (5). Performing similar actions for n_i, $i \in \{2, \ldots, m-1\}$, we obtain the iterative form

$$n_{i+1} = n_i + \frac{F(n_i) - F(n_{i-1})}{f(n_i)}.$$

Taking the partial derivative with respect to n_m and setting it equal to zero yields the value

$$\lambda_\delta = \frac{kP_{\text{ACK}}(n_m)}{\left(\text{E}[n_S]\right)^2}\left(\frac{1-F(n_{m-1})}{f(n_m)}\right) - \frac{k}{\text{E}[n_S]}.$$

Differentiating $J(\mathbf{n}, \lambda_\delta)$ with respect to λ_δ and equating it to zero gives $P_{\text{NACK}}(n_m) = \delta$ or, equivalently, the condition

$$n_m = F^{-1}(1-\delta). \tag{6}$$

Adopting the convention $n_0 = -\infty$, the above necessary conditions produce the recursive formula

$$n_{i+1} = n_i + \frac{F(n_i) - F(n_{i-1})}{f(n_i)} \quad i = 1, \ldots, m-1. \tag{7}$$

Since $F(\cdot)$ is chosen to be a distribution with $f(\cdot) > 0$ over the range of interest, the values generated by (7) form a strictly increasing sequence $n_1 < n_2 < \cdots < n_m$ for any admissible n_1. Hence, using this approach, the multi-dimensional optimization introduced in Problem 1 reduces to a one-dimensional optimization challenge. The ensuing task becomes finding a value of n_1 for which (6) and (7) are solved concurrently. This can readily be accomplished by performing a one-dimensional exhaustive search over $n_1 \in [k, F^{-1}(1-\delta))$.

Interestingly, the task of selecting vector $\mathbf{n} = (n_1, \ldots, n_m)$ is mathematically equivalent to finding the best $(m-1)$-level Lebesgue integral approximation to the CDF $F(\cdot)$ over the interval $[0, F^{-1}(1-\delta)]$; this is illustrated in Fig. 3 for $m = 4$. This alternate interpretation stems from rewriting the expected block length of (2) using $F(\cdot)$,

$$\text{E}[n_S] = n_m - \sum_{i=1}^{m-1}(n_{i+1} - n_i)F(n_i).$$

The subtracted sum above corresponds to the gray area in Fig. 3. Thus, maximizing throughput becomes equivalent to minimizing $\text{E}[n_S]$ or, alternatively, maximizing $\sum_{i=1}^{m-1}(n_{i+1} - n_i)F(n_i)$. As mentioned earlier, parameter n_m is given implicitly by the constraint $P_{\text{NACK}}(n_m) = \delta$. The KKT conditions in (7) can be construed as two (infinitesimal) rectangular regions having a same area,

$$(F(n_i) - F(n_{i-1}))\,\epsilon = (n_{i+1} - n_i)f(n_i)\epsilon.$$

In other words, the ratio of $F(n_i) - F(n_{i-1})$ over $n_{i+1} - n_i$ should be equal to the derivative $f(n_i)$. This is depicted by the black bands in Fig. 3.

B. Throughput Optimization With Constrained Feedback

The optimization described in Problem 1 sets a hard limit on the number of increments. Yet, the aforementioned formulation does not take into account the feedback rate induced by the ACK/NACK structure. An alternate and more encompassing viewpoint is to maximize average throughput while constraining both the number of increments and the feedback rate. Conceptually, the extended framework offers a means to trade off realized throughput against the implicit cost of feedback on the reverse link. It can also be regarded as a way to identify a larger candidate set for assignment vector $\mathbf{n}$, which then translates into a refined selection of optimal operating points in terms of throughput and feedback rate. This is detailed below.

Problem 2: Find an optimal increment assignment vector $\mathbf{n} = (n_1, \ldots, n_m)$ for the following constrained optimization problem,

$$\begin{aligned} &\underset{n_1,\ldots,n_m}{\text{maximize}} \ \frac{kP_{\text{ACK}}(n_m)}{\text{E}[n_S]} \\ &\text{subject to } P_{\text{NACK}}(n_m) \le \delta \\ &\text{and} \quad \frac{\text{E}[S]}{\text{E}[n_S]} \le \rho. \end{aligned}$$

Paralleling our earlier approach, we again turn to a Lagrangian formulation. The augmented objective function, which takes into account feedback rate, changes into

$$J(\mathbf{n}, \lambda_\delta, \lambda_\rho) = \frac{kP_{\text{ACK}}(n_m)}{\text{E}[n_S]} - \lambda_\delta\left(P_{\text{NACK}}(n_m) - \delta\right) - \lambda_\rho\left(\frac{\text{E}[S]}{\text{E}[n_S]} - \rho\right).$$

In deriving the corresponding KKT conditions, we will make use of the following convenient expression for $\text{E}[S]$,

$$\text{E}[S] = m - \sum_{i=1}^{m-1} F(n_i).$$

The first set of conditions associated with $\nabla J(\mathbf{n}, \lambda_\delta, \lambda_\rho) = \mathbf{0}$ can be written as

$$n_{i+1} = n_i + \frac{F(n_i) - F(n_{i-1})}{f(n_i)} - \frac{\lambda_\rho \text{E}[n_S]}{kP_{\text{ACK}}(n_m) - \lambda_\rho \text{E}[S]} \tag{8}$$

where $i = 1, \ldots, m-1$ and $n_0 = -\infty$. We note that the last term in (8) implicitly depends on $\mathbf{n}$ through $\text{E}[n_S]$

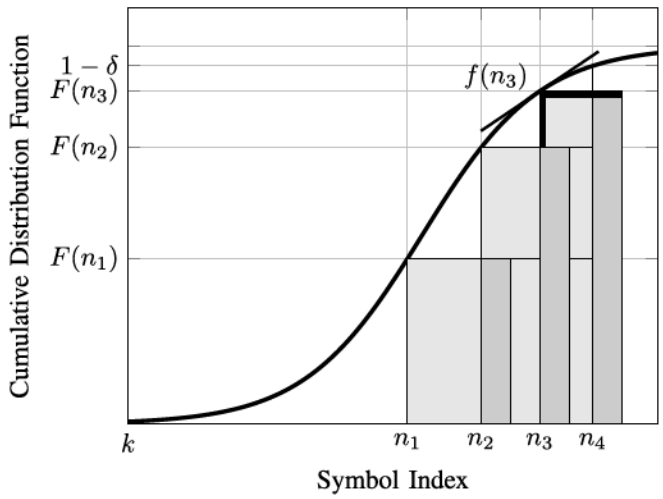

Fig. 4. When the optimization objective accounts for feedback, the task remains selecting $\mathbf{n} = (n_1, \ldots, n_4)$ as to maximize the shaded area. However, in this case, the shape of the rectangle is not only determined by the derivative $f(\cdot)$; a strip of width γ is added to every rectangle as to limit feedback. This simultaneously reduces throughput and feedback rate.

and $\mathrm{E}[S]$. Furthermore, it assumes the same value for sub-blocks $i \in \{1, \ldots, m-1\}$. The necessary conditions for an optimal solution can then be simplified by introducing auxiliary variable

$$\gamma = \frac{\lambda_\rho \mathrm{E}[n_S]}{kP_{\mathrm{ACK}}(n_m) - \lambda_\rho \mathrm{E}[S]}. \tag{9}$$

We emphasize that $\gamma > 0$ in the Lagrangian formulation whenever the rewards associated with throughput exceed the cost of feedback, i.e.,

$$\frac{kP_{\mathrm{ACK}}(n_m)}{\mathrm{E}[n_S]} - \lambda_\rho \frac{\mathrm{E}[S]}{\mathrm{E}[n_S]} > 0.$$

Under this expanded notation, (8) becomes

$$n_{i+1} = n_i + \frac{F(n_i) - F(n_{i-1})}{f(n_i)} - \gamma \quad i = 1, \ldots, m-1. \tag{10}$$

The partial derivative of $J(\mathrm{n}, \lambda_\delta, \lambda_\rho)$ with respect to n_m gives

$$\lambda_\delta = \frac{kP_{\mathrm{ACK}}(n_m) - \lambda_\rho \mathrm{E}[S]}{(\mathrm{E}[n_S])^2} \left(\frac{1 - F(n_{m-1})}{f(n_m)} \right) - \frac{k}{\mathrm{E}[n_S]}.$$

Taking the derivative of the objective function with respect to λ_δ yields condition $n_m = F^{-1}(1-\delta)$, as before. The feedback rate constraint in Problem 2 is recovered by differentiating with respect to λ_ρ.

For Problem 2, the Lagrangian analysis showcases that the multi-dimensional optimization can be solved by performing an exhaustive search over a two-dimensional set. The search takes place over all admissible $n_1 \in [k, n_m]$ and $\gamma \geq 0$. The geometric interpretation of this optimization task is similar to that of Problem 1 in that the aim is to maximize the shaded area. In this latter formulation, optimal rectangular shapes are again governed by $f(\cdot)$, but they are altered by a constant width γ as illustrated in Fig. 4. Intuitively, the role of the γ-bands is to limit feedback rate.

IV. SDO Applied to Erasure Channels

In this section, we illustrate the value of the extended SDO methodology with constrained feedback by applying it to binary erasure channels [27], [28]. These channels are memoryless and, as such, erasures form sequences of independent and identically distributed random variables. When an erasure occurs, the corresponding symbol is lost; otherwise, the channel input is received unaltered at the destination. Throughout, we represent the probability of an erasure by ϵ. For a fixed erasure probability, the number of observed (non-erased) symbols available to the receiver after t symbols are transmitted is a random variable, which we denote by R_t. This random variable is characterized by a binomial distribution,

$$P_{R_t}(r) = \binom{t}{r} \epsilon^{t-r} (1-\epsilon)^r \quad r = 0, \ldots, t \tag{11}$$

where r designates the number of unerased symbols. Note that we adopt the convention $0^0 = 1$ and hence, when $\epsilon = 0$, we have $P_{R_t}(t) = 1$ and $P_{R_t}(r) = 0$ for all $r \neq t$.

To shield information bits from channel erasures, redundancy is added to the original message using random linear coding. The encoding of a message involves a sequence of steps. First, a random parity-check matrix of size $(n-k) \times n$ is generated, with individual entries selected uniformly over a binary alphabet, independently from one another. The nullspace of the realized matrix produces a codebook. A message is then mapped to a codeword using an arbitrary choice function known to both the source and the destination [29]. To recover the original message, the destination employs maximum-likelihood decoding. This coding strategy is known to perform well, and it serves as an analytically tractable proxy for more pragmatic codes [3], [28]. One of the attractive aspects of random linear coding lies in the flexibility it affords in terms of selecting block length and code rate. This enables a unified analysis of overall performance as a function of design parameters. Furthermore, the statistical symmetry in this random linear coding scheme produces a probability of decoding success that depends solely on the number of erased symbols, rather than their precise locations. These attributes make random linear codes ideally suited for an explicative case study of SDO. To apply the SDO methodology in the context of binary erasure channels with random linear coding, we need to obtain expressions for $P_{\mathrm{ACK}}(\cdot)$ and its smooth approximation $F(\cdot)$. This is best accomplished by treating the properties of random linear coding and the effects of channel erasures separately.

A. Asymptotic Analysis of Random Linear Codes

For the random linear coding scheme at hand, we use $P_{\mathrm{s}}(k, n, r)$ to represent the probability of decoding success as a function of the number of unerased symbols r available at the destination.

Lemma 1: The probability of decoding success for the random linear coding scheme described above is

$$P_{\mathrm{s}}(k, n, r) = \begin{cases} 0, & r < k \\ \prod_{\ell=0}^{n-r-1} \left(1 - 2^{\ell-(n-k)}\right), & k \leq r \leq n \\ 1, & r > n. \end{cases} \tag{12}$$

Proof: See Appendix, Section A. ∎

Although the number of sent symbols and, consequently, the number of symbols available at the destination cannot exceed the blocklength, we find it useful to extend $P_{\mathrm{s}}(k,n,r)$ in (12) to cases where $r > n$. The purpose of this slight abuse of notation will become manifest shortly when we compare systems with alternate coding schemes.

B. Asymptotic Behavior Over Reliable Channels

We initiate our analysis by focusing on the special case of a lossless channel, with $\epsilon = 0$. We examine an elementary version of the problem where symbols are obtained in a sequential manner, and a decoding attempt takes place after every new symbol arrives (not only upon the completion of sub-blocks). For system parameters k and n, let M_n be a random variable that denotes the number of symbols needed for the message to become decodable, following chronological ordering. Under these circumstances, we have $k \leq M_n \leq n$ and $\Pr(M_n \leq r) = P_{\mathrm{s}}(k,n,r)$. We wish to analyze the asymptotic behavior of the mean and variance of M_n as n grows unbounded. To achieve this objective, we leverage two known constants. We denote the Erdös-Borwein constant (OEIS: A065442) by

$$c_0 = \sum_{i=1}^{\infty} \frac{1}{2^i - 1} = 1.6066951524\ldots$$

and the digital search tree constant (OEIS: A065443) by

$$c_1 = \sum_{i=1}^{\infty} \frac{1}{(2^i - 1)^2} = 1.1373387363\ldots$$

The following infinite sums of products, presented in the form of a lemma, are key components in our impending derivations.

Lemma 2: For infinite product $a_i = 2^{-i} \prod_{j=i+1}^{\infty} \left(1 - 2^{-j}\right)$, it holds that

$$\sum_{i=0}^{\infty} a_i = 1 \quad \sum_{i=0}^{\infty} i a_i = c_0$$
$$\sum_{i=0}^{\infty} i^2\, a_i = c_0^2 + c_0 + c_1 = 5.3255032015\ldots$$

Proof: See Appendix, Section B. ∎

Let $P_{M_n}(\cdot)$ represent the PMF associated with M_n; that is, $P_{M_n}(r) = P_{\mathrm{s}}(k,n,r) - P_{\mathrm{s}}(k,n,r-1)$. This function can be rewritten as

$$P_{M_n}(r) = 2^{k-r} P_{\mathrm{s}}(k,n,r) = 2^{k-r} \prod_{\ell=0}^{n-r-1} \left(1 - 2^{\ell-(n-k)}\right)$$

for $k \leq r \leq n$. Moreover, $P_{M_n}(r) = 0$ for $r < k$ or $r > n$. The normalization axiom applied to this problem ensures that $\sum_{r=k}^{n} P_{M_n}(r) = 1$. Thus, we can compute the mean of M_n as

$$\mathrm{E}[M_n] = \sum_{r=k}^{n} r P_{M_n}(r) = \sum_{i=0}^{n-k} (k+i) 2^{-i} \prod_{j=i+1}^{n-k} \left(1 - 2^{-j}\right).$$

Similarly, the second moment of M_n is equal to

$$\mathrm{E}\left[M_n^2\right] = \sum_{r=k}^{n} r^2\ P_{M_n}(r) = \sum_{i=0}^{n-k} (k+i)^2 2^{-i} \prod_{j=i+1}^{n-k} \left(1 - 2^{-j}\right)$$

and its variance can be evaluated based on the first two moments. Passing to the limit, as n goes to infinity, we get the following result.

Theorem 1: For k fixed, the limiting mean and variance of M_n are given by

$$\lim_{n\to\infty} \mathrm{E}[M_n] = k + c_0 \tag{13}$$
$$\lim_{n\to\infty} \mathrm{Var}[M_n] = c_0 + c_1. \tag{14}$$

Proof: As n becomes large, we get the expressions

$$\lim_{n\to\infty} \mathrm{E}[M_n] = \sum_{i=0}^{\infty} (k+i) 2^{-i} \prod_{j=i+1}^{\infty} \left(1 - 2^{-j}\right)$$
$$\lim_{n\to\infty} \mathrm{E}\left[M_n^2\right] = \sum_{i=0}^{\infty} (k^2 + 2ki + i^2) 2^{-i} \prod_{j=i+1}^{\infty} \left(1 - 2^{-j}\right).$$

Then, by Lemma 2, we get $\lim_{n\to\infty} \mathrm{E}[M_n] = k + c_0$. Likewise, $\lim_{n\to\infty} \mathrm{E}\left[M_n^2\right] = k^2 + 2\ kc_0 + (c_0^2 + c_0 + c_1)$. Since the variance of M_n can be derived as $\mathrm{Var}[M_n] = \mathrm{E}\left[M_n^2\right] - \left(\mathrm{E}[M_n]\right)^2$, we readily obtain (14), as desired. ∎

C. Asymptotic Behavior Over Unreliable Channels

At this stage, we are ready to address the more elaborate problem where symbols are transmitted over an unreliable channel. That is, individual symbols are erased with probability $\epsilon > 0$. For k, n, and ϵ fixed, we represent the length of a communication round by N_n. Note that $k \leq N_n \leq n$. We can partition rounds into two categories: (i) the receiver is able to decode before all the symbols are transmitted, and N_n corresponds to the first instant at which the message can be successfully recovered; (ii) all the symbols are exhausted during the transmission phase, and $N_n = n$ irrespective of the outcome of the decoding process. Mirroring the steps above, we inspect the asymptotic behavior of the mean and variance of N_n as n increases to infinity.

Define E_r as the number of symbols lost prior to observing the rth unerased symbols at the destination. We write $P_{E_r}(\cdot)$ to refer to the PMF of E_r, and we emphasize that this random variable possesses a negative binomial distribution with parameters r and ϵ. In other words, we have

$$P_{E_r}(e) = \binom{r+e-1}{e} \epsilon^e (1-\epsilon)^r \quad e \geq 0.$$

Then, we get $\Pr(N_n = t) = \sum_{r=k}^{t} P_{E_r}(t-r) P_{M_n}(r)$ for $k \leq t < n$ and, necessarily,

$$\Pr(N_n = n) = 1 - \sum_{t=k}^{n-1} \Pr(N_n = t)$$
$$= \sum_{t=n}^{\infty} \sum_{r=k}^{t} P_{E_r}(t-r) P_{M_n}(r).$$

Consequently, we can write

$$\begin{aligned}\mathrm{E}[N_n] &= \sum_{t=k}^{n} t \Pr(N_n = t) \\ &= \sum_{t=k}^{\infty}\sum_{r=k}^{t} \min(t,n) P_{E_r}(t-r) P_{M_n}(r) \\ &= \sum_{r=k}^{n}\sum_{e=0}^{\infty} \min(r+e,n) P_{E_r}(e) P_{M_n}(r).\end{aligned}$$

The second moment can be expressed as

$$\mathrm{E}\left[N_n^2\right] = \sum_{r=k}^{n}\sum_{e=0}^{\infty} \min((r+e)^2, n^2) P_{E_r}(e) P_{M_n}(r).$$

Collecting these results and evaluating limit expressions, we arrive at the following theorem.

Theorem 2: Given parameters k and ϵ,

$$\mu(k,\epsilon) = \lim_{n\to\infty} \mathrm{E}[N_n] = \frac{k+c_0}{1-\epsilon} \tag{15}$$

$$\sigma^2(k,\epsilon) = \lim_{n\to\infty} \mathrm{Var}[N_n] = \frac{(k+c_0)\epsilon + c_0 + c_1}{(1-\epsilon)^2}. \tag{16}$$

Proof: We initiate this argument by establishing bounds on $\mathrm{E}[N_n]$. Observing that $\min(r+e,n) \le r+e$, we get

$$\mathrm{E}[N_n] \le \sum_{r=k}^{n}\sum_{e=0}^{\infty} (r+e) P_{E_r}(e) P_{M_n}(r) \quad \forall n \ge k.$$

For a memoryless erasure channel, E_r possesses a negative binomial distribution with parameters r and ϵ. Thus,

$$\begin{aligned}\sum_{e=0}^{\infty} (r+e) P_{E_r}(e) &= r\sum_{e=0}^{\infty} P_{E_r}(e) + \sum_{e=0}^{\infty} e P_{E_r}(e) \\ &= r + \mathrm{E}[E_r] = \frac{r}{1-\epsilon} \quad \forall r \ge 0.\end{aligned}$$

Substituting this expression into the double summation above, we get

$$\mathrm{E}[N_n] \le \frac{1}{1-\epsilon}\sum_{r=k}^{n} r P_{M_n}(r) = \frac{\mathrm{E}[M_n]}{1-\epsilon} \quad \forall n \ge k. \tag{17}$$

We turn to establishing a lower bound for $\mathrm{E}[N_n]$. Restricting the number of non-negative summands, we get

$$\mathrm{E}[N_n] \ge \sum_{r=k}^{n}\sum_{e=0}^{n-r} (r+e) P_{E_r}(e) P_{M_n}(r).$$

Given any k and r, it is easy to show that $P_{\mathrm{s}}(k,n,r)$ is monotone decreasing in n. Further, $P_{M_n}(r) = 2^{k-r} P_{\mathrm{s}}(k,n,r)$ for all $k \le r \le n$. Then, we gather that $P_{M_n}(r)$ is monotone decreasing in n for any r such that $k \le r \le n$. This implies that $P_{M_n}(r) \ge \lim_{n\to\infty} P_{M_n}(r)$ for all n and all $k \le r \le n$. We note that

$$\lim_{n\to\infty} P_{M_n}(r) = 2^{k-r} \prod_{j=r-k+1}^{\infty} (1-2^{-j}).$$

Therefore, for any n, we can write

$$\begin{aligned}\mathrm{E}[N_n] &\ge \sum_{r=k}^{n}\sum_{e=0}^{n-r} (r+e) P_{E_r}(e) 2^{k-r} \prod_{j=r-k+1}^{\infty} (1-2^{-j}) \\ &= \sum_{r=k}^{n} 2^{k-r} \sum_{e=0}^{n-r} (r+e) P_{E_r}(e) \prod_{j=r-k+1}^{\infty} (1-2^{-j}).\end{aligned} \tag{18}$$

Using Theorem 1, we deduce that the RHS of (17) converges to $(k+c_0)/(1-\epsilon)$ as n grows unbounded. Furthermore, in view of Lemma 2, we see that the RHS of (18) converges to

$$\begin{aligned}&\sum_{r=k}^{\infty} 2^{k-r} \sum_{e=0}^{\infty} (r+e) P_{E_r}(e) \prod_{j=r-k+1}^{\infty} (1-2^{-j}) \\ &\quad= \frac{1}{1-\epsilon}\sum_{r=k}^{\infty} r 2^{k-r} \prod_{j=r-k+1}^{\infty} (1-2^{-j}) \\ &\quad= \frac{1}{1-\epsilon}\sum_{i=0}^{\infty} (k+i) 2^{-i} \prod_{j=i+1}^{\infty} (1-2^{-j}) = \frac{k+c_0}{1-\epsilon}.\end{aligned}$$

Combining (17) and (18), the sandwich theorem yields (15).

By adopting an analogous strategy, we can produce upper bound

$$\begin{aligned}\mathrm{E}\left[N_n^2\right] &\le \sum_{r=k}^{n}\sum_{e=0}^{\infty} (r+e)^2 P_{E_r}(e) P_{M_n}(r) \\ &= \frac{\mathrm{E}\left[M_n^2\right] + \epsilon \mathrm{E}[M_n]}{(1-\epsilon)^2},\end{aligned} \tag{19}$$

and corresponding lower bound

$$\begin{aligned}\mathrm{E}\left[N_n^2\right] &\ge \sum_{r=k}^{n}\sum_{e=0}^{n-r} (r+e)^2 P_{E_r}(e) P_{M_n}(r) \\ &\ge \sum_{r=k}^{\infty} 2^{k-r} \sum_{e=0}^{\infty} (r+e)^2 P_{E_r}(e) \prod_{j=r-k+1}^{\infty} (1-2^{-j})\end{aligned} \tag{20}$$

for any n. As n goes to infinity, the RHS of (19) converges to $\left((k+c_0)^2 + (k+c_0)\epsilon + c_0 + c_1\right)/(1-\epsilon)^2$ by Theorem 1. Likewise, by Lemma 2, the RHS of (20) converges to

$$\begin{aligned}&\frac{1}{(1-\epsilon)^2}\sum_{r=k}^{\infty} r(r+\epsilon) 2^{k-r} \prod_{j=r-k+1}^{\infty} (1-2^{-j}) \\ &\quad= \frac{1}{(1-\epsilon)^2}\sum_{i=0}^{\infty} (k+i)^2 2^{-i} \prod_{j=i+1}^{\infty} (1-2^{-j}) \\ &\qquad+ \frac{\epsilon}{(1-\epsilon)^2}\sum_{i=0}^{\infty} (k+i) 2^{-i} \prod_{j=i+1}^{\infty} (1-2^{-j}) \\ &\quad= \frac{(k+c_0)^2 + (k+c_0)\epsilon + c_0 + c_1}{(1-\epsilon)^2}.\end{aligned}$$

Combining (19) and (20), the sandwich theorem offers a tight characterization of the asymptotic second moment of N_n,

$$\lim_{n\to\infty} \mathrm{E}\left[N_n^2\right] = \frac{(k+c_0)^2 + (k+c_0)\epsilon + c_0 + c_1}{(1-\epsilon)^2}.$$

From its first two moments, we can infer the limiting variance of N_n,

$$\lim_{n\to\infty} \mathrm{Var}[N_n] = \lim_{n\to\infty} \mathrm{E}\left[N_n^2\right] - \lim_{n\to\infty} \left(\mathrm{E}[N_n]\right)^2 = ((k+c_0)\epsilon + c_0 + c_1)/(1-\epsilon)^2,$$

as desired. ∎

D. Approximate Distribution via Moment Matching

For the application of the (n, k) random linear coding scheme at hand over an erasure channel (with erasure probability ϵ), the probability that the destination decodes the original message successfully at time t or earlier is given by

$$P_{\mathrm{ACK}}(t) = \begin{cases} 1 - \sum_{r=0}^{t} (1 - P_{\mathrm{s}}(k, n, r)) P_{R_t}(r), & k \le t \le n, \\ 0, & 0 \le t < k, \end{cases} \tag{21}$$

where $P_{R_t}(\cdot)$ and $P_{\mathrm{s}}(k, n, \cdot)$ are given in (11) and (12), respectively.

The extended SDO framework relies on a smooth approximation for $P_{\mathrm{ACK}}(\cdot)$ as indicated in (5). A natural approach to obtaining such a distribution consists of identifying a fitting distribution family, like the collection of Gaussian distributions, and subsequently apply moment matching to get suitable parameters [30]. Since Gaussian distributions are determined by two parameters, it suffices to compute the mean and variance to select a member within the Gaussian family. Hereafter, we adopt the Gaussian distribution for illustrative purposes. This choice can be motivated, partly, through the Central Limit Theorem.

Leveraging results from the previous sections, we let $F(\cdot)$ be a Gaussian distribution with mean $\mu(k, \epsilon)$ and variance $\sigma^2(k, \epsilon)$, as defined in Theorem 2. Mathematically, we take

$$F(t) = 1 - Q\left(\frac{t - \mu(k,\epsilon)}{\sigma(k,\epsilon)}\right)$$

where $\sigma(k, \epsilon) = \sqrt{\sigma^2(k, \epsilon)}$ and $Q(\cdot)$ is the complementary CDF of a standard Gaussian random variable,

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-\xi^2/2} d\xi.$$

In view of the geometric interpretation of the SDO algorithm introduced in Section III, we gather that the smaller is the maximum point-wise distance between the functions $F(t)$ and $P_{\mathrm{ACK}}(t)$, i.e., $\sup_{t\in[0,n]} |F(t) - P_{\mathrm{ACK}}(t)|$, the smaller would be the difference in throughput between the optimal solution and the solution derived via the extended SDO algorithm. Luckily, these two functions tend to be very close, as illustrated in Fig. 5.

The approximate CDF $F(\cdot)$ enables the application of the extended SDO algorithm to find near optimal values for sub-block sizes $\mathbf{n}$. For numerical analysis, we adopt parameters $k = 64$ and $n = 127$. We assume that the binary erasure channel features an erasure probability given by $\epsilon = 0.358$. The Shannon capacity of this particular channel is 0.642 bits per channel use [27]. Under the random linear coding scheme

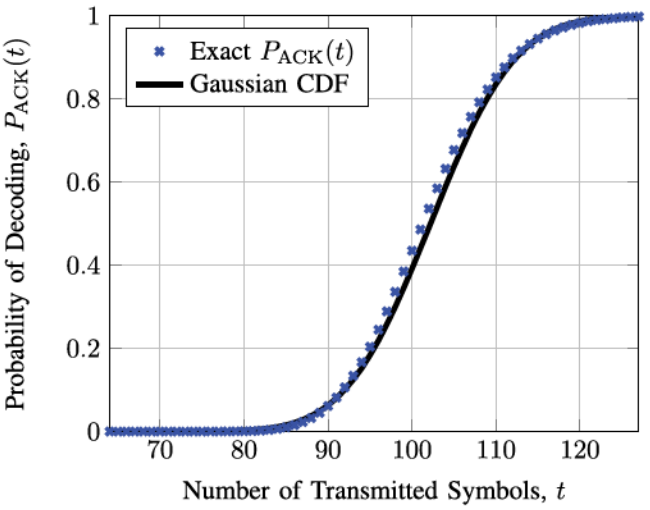

Fig. 5. This graph showcases how the approximate CDF obtained through moment matching is very close to the exact CDF for random linear coding over an erasure channel. In this case, parameters $k = 64$, $n = 127$, and $\epsilon = 0.358$ are chosen. The small gap between the two functions hints at a near-optimal SDO performance.

of Section IV, but with unlimited feedback, the maximum throughput becomes

$$\frac{kP_{\mathrm{ACK}}(n)}{\mathrm{E}[n_S]} = \frac{kP_{\mathrm{ACK}}(n)}{n - \sum_{t=1}^{n-1} P_{\mathrm{ACK}}(t)} = 0.624756\ldots,$$

where $n = 127$, $k = 64$, $\epsilon = 0.358$, and $P_{\mathrm{ACK}}(t)$ is given by (21). We refer to this throughput value as the Random Code Limit in Fig. 6 and Fig. 8. The difference between this throughput value and the Shannon capacity of the erasure channel is attributable to the limitations of the coding scheme and the finite block length. The latter value serves as an optimistic upper bound on the performance of incremental redundancy applied to this particular setting.

E. Performance Analysis and Validation

We use this same setting to present the performance of the extended SDO algorithm applied to the formulation of Problem 2. The constraint on the maximum number of feedback messages takes value in $m \in \{1, 2, 4, 8, 16\}$. In addition, we also consider the unconstrained setting where $m = \infty$. We study performance for an average feedback rate varying between zero and 0.2 bits per channel use. Note that the upper boundary on the feedback rate, 0.2 bits per channel use, effectively reduces to having no constraint on the feedback rate because performance saturates. Every instance of this problem essentially entails an exhaustive search over a two-dimensional set, as discussed in Section III-B. This leads to a very rapid execution, much faster than an exhaustive search over all sub-block sizes for large m. Furthermore, the y-intercepts on the RHS of the plot correspond to the realized throughput values associated with the SDO algorithm applied to the formulation of Problem 1. These numerical findings are shown in Fig. 6.

These numerical findings are conceptually appealing because they support the use of incremental redundancy based on one-bit feedback messages. They also attest to the value

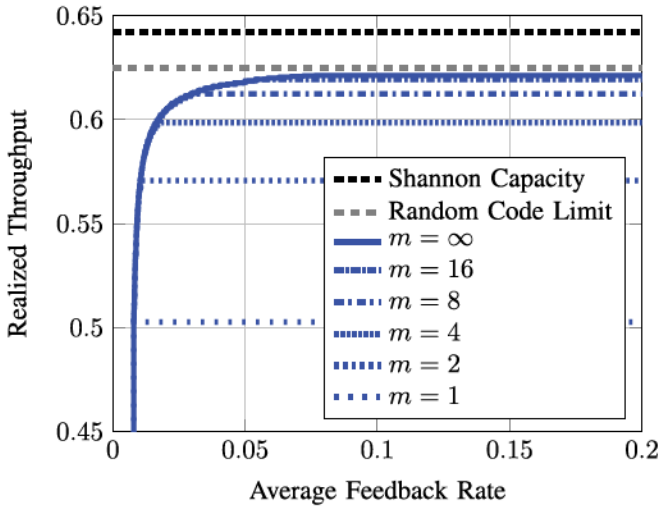

Fig. 6. This graph plots realized throughput as functions of maximum feedback rate for SDO-based system optimization. The curves demonstrate diminishing returns as functions of both, the maximum number of messages, m, and the limit on average feedback rate. Performance rapidly gets close to Shannon, but then saturates due to the limitations of the coding scheme.

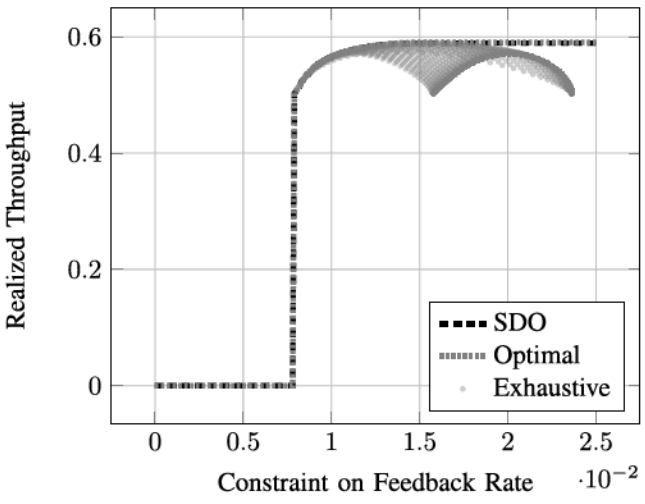

Fig. 7. The performance achieved using SDO is nearly indistinguishable from the throughput associated with the optimal hybrid ARQ schemes obtained via exhaustive searches. The graph also plots the performance point for every admissible $\mathbf{n}$; collectively these points form the umbrella shape. The curves correspond to the case where the maximum number of feedback messages is limited to three.

of the extended SDO algorithm in finding appropriate sizes for sub-blocks. The structural properties of the extended SDO scheme, along with its iterative nature, enable the design and analysis of incremental redundancy in the form of hybrid ARQ in many contexts. Two important issues remain. First, we wish to know how the performance of an SDO-based system compares to that of a hybrid ARQ implementation with optimal parameters. Numerical methods suggest that the realized throughput values of comparable systems are very close, at least for values of m where an exhaustive search is possible. To substantiate this claim, we provide a comparative plot of SDO-based performance against optimal throughput for the case where $m = 3$; as seen from Fig. 7, the realized throughput between optimal parameters and SDO-derived values for any given feedback constraint is essentially indistinguishable. For illustrative purposes, the figure also includes the performance point for every admissible selection of $\mathbf{n}$. The optimal curve is then obtained as the maximum throughput among all the points with average feedback rate below the prescribed limit.

The second persisting issue is related to the design decision to employ one-bit feedback messages. Arguably, enhanced performance could potentially be obtained by using larger feedback messages, albeit less often. This question raises technical issues. While it is straightforward to assign a meaning to one-bit (ACK/NACK) feedback, mapping a multi-bit feedback message to a particular set of actions is more involved. To circumvent this difficulty and showcase the suitability of one-bit feedback in the context of hybrid ARQ, we adopt the following approach. We compare the performance of the original SDO-based, one-bit feedback implementation to that of a system where the maximum number of feedback messages remains the same, but the size of individual messages is unbounded. In the latter case, we assume that the decoder feeds back the total number of unerased symbols received at the destination thus far, thereby enabling the source to

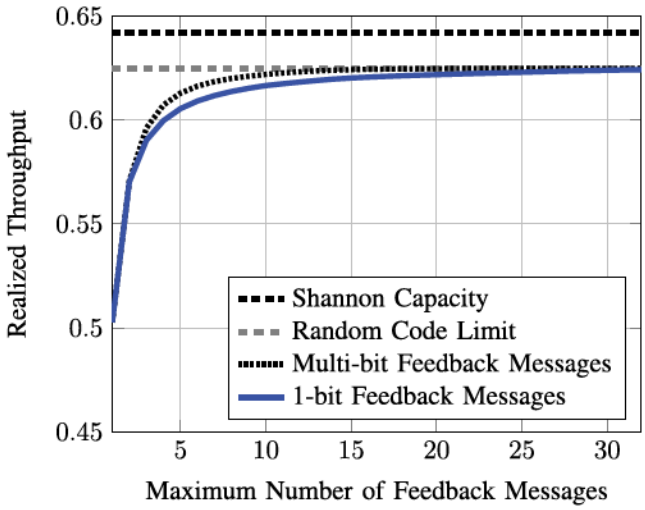

Fig. 8. This figure offers supportive evidence to the fact that one-bit (ACK/NACK) feedback is a suitable paradigm for incremental redundancy over erasure channels with limited feedback. The one-bit and multi-bit implementations above are subject to the same restriction on the number of feedback messages, yet the messages are unlimited in size in the latter system, whereas they are constrained to one bit in the former one.

select the most suitable block size for the system under current conditions. Mathematically, solving the problem for unbounded message size warrants the application of a finite-horizon dynamic program [31] whereby the system selects the optimal size for the next sub-block after every feedback message is received. The state of this dynamic program contains the decision time with respect to the onset of the round, the number of encoded bits received thus far, and the number of feedback messages used in the past. Based on this information, the system determines the size of the next increment and the source transmits the corresponding symbols over the erasure channel. Fig. 8 contrasts the performance of the SDO-based, one-bit feedback system to that

of the implementation with unlimited packet sizes. Despite the drastic information asymmetry between the two schemes, their overall performance is very close. This can be explained, partly, through the fact that for the purpose of inference and decision making, the first few bits of a message are often the most informative. In any case, the one-bit implementation is evidently a reasonable pragmatic approach.

V. Discussion

This article casts SDO as a classic optimization problem and extends this methodology to include constraints on feedback rate. It also offers a novel geometric interpretation that showcases how decisions regarding the sizes of hybrid ARQ sub-blocks are related to Lebesgue approximations of the area under the CDF of the first decoding success. The power of the extended SDO algorithm is exemplified by applying this approach to hybrid ARQ over erasure channels. Due to their structure, erasure channels are especially well suited to SDO and the resulting realized throughput is essentially indistinguishable from optimal performance. While throughput increases with the maximum number of sub-blocks as anticipated, numerical results suggest that only a small number of feedback messages suffice to achieve a performance close to the maximum throughput obtained with a potentially unbounded number of feedback messages. This an encouraging conclusion for pragmatic systems, as it favors simplicity over overly complex implementations.

There are several possible avenues of future research. While this contribution offers an in-depth treatment of SDO over classical channels, it may be possible to extend the technique to fading channels. In the latter context, both the size of sub-blocks and the role of side information warrant further attention. In particular, the renewal problem structure will have to be revisited. In addition, SDO may offer a principled approach to assessing the potential benefits of incremental redundancy in the context of age of information and delay-sensitive communications with queues. Some preliminary steps have been taken along these lines in the literature, yet these topics are still not fully developed. Finally, the methodology may apply to uncoordinated multiple access systems where the access point is given the opportunity to broadcast one-bit feedback to active devices. Despite being collectively very promising, these candidate directions lie outside the scope of this article and are therefore relegated to future inquiries.

Appendix

A. Proof of Lemma 1

Let H be a random matrix of size $(n-k)\times n$, where each entry is selected independently and uniformly from $\{0,1\}$. Consider an (n,k) linear code with parity-check matrix H. For any codeword $\mathbf{c}$, we have $H\mathbf{c}^T=0$. The destination can decode the message from any r received symbols $c_{i_1},\ldots,c_{i_r}$ provided that the $n-r$ columns of H with indices $[n]\setminus\{i_1,\ldots,i_r\}$ are linearly independent. That is, $P_{\mathrm{s}}(k,n,r)$ is equal to the probability that the $n-r$ randomly generated binary column vectors of length $n-k$ are linearly independent. This event has probability

$$P_{\mathrm{s}}(k,n,r)=\prod_{l=1}^{n-r}\frac{\left(2^{n-k}-2^{l-1}\right)}{2^{n-k}}=\prod_{l=0}^{n-r-1}\left(1-2^{l-(n-k)}\right)$$

for $k\le r\le n$, and $P_{\mathrm{s}}(k,n,r)=0$ for $r<k$.

B. Proof of Lemma 2

First, we recall notation $a_i=2^{-i}\prod_{j=i+1}^{\infty}\left(1-2^{-j}\right)$ from the statement of Lemma 2. We start this proof with the simplest of the three sums, namely $\sum_{i=0}^{\infty}a_i=1$. To this end, we define

$$b_l=\sum_{i=0}^{\infty}2^{-i}\prod_{j=i+1}^{i+l}\left(1-2^{-j}\right)\quad\forall l\in\mathbb{N}_0.$$

We observe that $b_0=2$ and $\lim_{l\to\infty}b_l=\sum_{i=0}^{\infty}a_i$. Leveraging the fact that $2^{-i}=2^{-(i-1)}-2^{-i}$ for any $i\in\mathbb{Z}$ and $\prod_{j=0}^{l-1}(1-2^{-j})=0$, it can be shown that $b_l-b_{l-1}=2^{-(l+1)}(b_l-2b_{l-1})$.

Thus, for any $l\in\mathbb{N}_0$, we have $b_l\left(1-2^{-(l+1)}\right)=b_0(1-2^{-1})$. Taking the limit as l grows unbounded, we get $\lim_{l\to\infty}b_l=b_0/2=1$. Thus, $\sum_{i=0}^{\infty}a_i=\lim_{l\to\infty}b_l=1$, as desired.

Next, we consider the equation $\sum_{i=0}^{\infty}ia_i=c_0$. Using Euler's pentagonal number theorem (see, e.g., [32, p. 20]), it can be shown that

$$\prod_{i=0}^{\infty}\frac{1}{1-2^{-i}x}=\sum_{i=0}^{\infty}x^i\prod_{j=1}^{i}\frac{1}{1-2^{-j}}.\tag{22}$$

Differentiating with respect to x on both sides, we get

$$\sum_{i=0}^{\infty}\frac{2^{-i}}{1-2^{-i}x}\times\prod_{j=0}^{\infty}\frac{1}{1-2^{-j}x}=\sum_{i=1}^{\infty}ix^{i-1}\prod_{j=1}^{i}\frac{1}{1-2^{-j}}.$$

Setting $x=1/2$, we obtain

$$\sum_{i=0}^{\infty}\frac{2^{-i}}{1-2^{-i-1}}\times\prod_{j=0}^{\infty}\frac{1}{1-2^{-j-1}}=\sum_{i=1}^{\infty}i2^{-i+1}\prod_{j=1}^{i}\frac{1}{1-2^{-j}}.$$

By a simple change of variables and rearranging the terms, we arrive at

$$c_0=\sum_{i=1}^{\infty}\frac{2^{-i}}{1-2^{-i}}=\sum_{i=1}^{\infty}i2^{-i}\prod_{j=i+1}^{\infty}\left(1-2^{-j}\right)=\sum_{i=0}^{\infty}ia_i.$$

The procedure followed to get the third expression is similar in nature. First, we take derivatives with respect to x twice on both sides of identity (22). Then, we evaluate these expressions at $x=1/2$. After rearranging terms, this yields

$$\left(\sum_{i=1}^{\infty}\left(2^i-1\right)^{-1}\right)^2+\left(\sum_{i=1}^{\infty}\left(2^i-1\right)^{-2}\right)=\sum_{i=0}^{\infty}i(i-1)a_i.$$

Noticing that these terms are related to constants introduced earlier, with $c_0=\sum_{i=1}^{\infty}\left(2^i-1\right)^{-1}$ and $c_1=\sum_{i=1}^{\infty}\left(2^i-1\right)^{-2}$, we readily obtain the desired expression.

REFERENCES

[1] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[2] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Dispersion of the Gilbert-Elliott channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1829–1848, Apr. 2011.

[3] S. Kumar, J.-F. Chamberland, and H. D. Pfister, "First-passage time and large-deviation analysis for erasure channels with memory," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5547–5565, Sep. 2013.

[4] 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.

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

[6] D. J. Love, R. W. Heath, Jr., V. K. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[7] A. Chockalingam, M. Zorzi, L. B. Milstein, and P. Venkataram, "Performance of a wireless access protocol on correlated Rayleigh-fading channels with capture," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 644–655, May 1998.

[8] C. Shen, T. Liu, and M. P. Fitz, "On the average rate performance of hybrid-ARQ in quasi-static fading channels," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3339–3352, Nov. 2009.

[9] L. B. Le, E. Hossain, and M. Zorzi, "Queueing analysis for GBN and SR ARQ protocols under dynamic radio link adaptation with non-zero feedback delay," *IEEE Trans. Wireless Commun.*, vol. 6, no. 9, pp. 3418–3428, Sep. 2007.

[10] F. Hamidi-Sepehr, J.-F. Chamberland, and H. D. Pfister, "On the performance of block codes over finite-state channels in the rare-transition regime," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 3974–3990, Nov. 2015.

[11] A. R. Williamson, T.-Y. Chen, and R. D. Wesel, "Variable-length convolutional coding for short blocklengths with decision feedback," *IEEE Trans. Commun.*, vol. 63, no. 7, pp. 2389–2403, Jul. 2015.

[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 Inf. Theory Workshop (ITW)*, Hobart, TAS, Australia, Nov. 2014, pp. 167–171.

[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.*, vol. 64, no. 6, pp. 2245–2257, Jun. 2016.

[14] A. Heidarzadeh, J.-F. Chamberland, R. D. Wesel, and P. Parag, "A systematic approach to incremental redundancy over erasure channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 1176–1180.

[15] N. Ahmed, M. A. Khojastepour, A. Sabharwal, and B. Aazhang, "Outage minimization with limited feedback for the fading relay channel," *IEEE Trans. Commun.*, vol. 54, no. 4, pp. 659–669, Apr. 2006.

[16] N. Varnica, E. Soljanin, and P. Whiting, "LDPC code ensembles for incremental redundancy hybrid ARQ," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Sep. 2005, pp. 995–999.

[17] J. Kim, W. Hur, A. Ramamoorthy, and S. W. McLaughlin, "Design of rate-compatible irregular LDPC codes for incremental redundancy hybrid ARQ systems," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2006, pp. 1139–1143.

[18] D. Kim, "Hybrid-ARQ with rate compatible LDPC codes," in *Convergence and Hybrid Information Technology*, G. Lee, D. Howard, J. J. Kang, and D. Ślęzak, Eds. Berlin, Germany: Springer, 2012, pp. 25–32.

[19] K. Chen, K. Niu, and J. Lin, "A hybrid ARQ scheme based on polar codes," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1996–1999, Oct. 2013.

[20] H. Saber and I. Marsland, "An incremental redundancy Hybrid ARQ scheme via puncturing and extending of polar codes," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 3964–3973, Nov. 2015.

[21] L. Ma, J. Xiong, Y. Wei, and M. Jiang. (Aug. 2017). "An incremental redundancy HARQ scheme for polar code." [Online]. Available: https://arxiv.org/abs/1708.09679

[22] M. El-Khamy, H. P. Lin, J. Lee, H. Mahdavifar, and I. Kang, "HARQ rate-compatible polar codes for wireless channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[23] K. Ausavapattanakun and A. Nosratinia, "Analysis of selective-repeat ARQ via matrix signal-flow graphs," *IEEE Trans. Commun.*, vol. 55, no. 1, pp. 198–204, Jan. 2007.

[24] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529–532, Oct. 2014.

[25] S. M. Ross, *Stochastic Processes*, 2nd ed. Hoboken, NJ, USA: Wiley, 1995.

[26] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Dispersion of Gaussian channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Seoul, South Korea, Jun./Jul. 2009, pp. 2204–2208.

[27] R. G. Gallager, *Information Theory and Reliable Communication*. Hoboken, NJ, USA: Wiley, 1968.

[28] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[29] J. Munkres, *Topology*, 2nd ed. London, U.K.: Pearson, 2000.

[30] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA, USA: Duxbury Thomson Learning, 2001.

[31] R. Bellman, *Dynamic Programming*. New York, NY, USA: Dover, 2003.

[32] D. E. Knuth, *The Art of Computer Programming: Sorting and Searching*, vol. 3, 2nd ed. Redwood City, CA, USA: Addison-Wesley, 1998.

Anoosheh Heidarzadeh received the Ph.D. degree in electrical and computer engineering from Carleton University, Ottawa, ON, Canada, in 2012. He was a Post-Doctoral Scholar with the California Institute of Technology, Pasadena, CA, USA, from 2013 to 2014. He is currently a Visiting Assistant Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. His research interests include network coding and network information theory, private information retrieval, group testing, compressed sensing, distributed computing, and game theory.

Jean-Francois Chamberland (S'98–M'04–SM'09) received the Ph.D. degree from the University of Illinois at Urbana–Champaign. He is currently a Professor with the Department of Electrical and Computer Engineering, Texas A&M University. His research interests are in the areas of computing, information, and inference. He has been a recipient of the IEEE Young Author Best Paper Award from the IEEE Signal Processing Society and the Faculty Early Career Development (CAREER) Award from the National Science Foundation.

Richard D. Wesel received the B.S. and M.S. degrees in electrical engineering from MIT. He is currently a Professor with Electrical Engineering Department, UCLA. After receiving the Ph.D. degree in electrical engineering from Stanford, he joined UCLA in 1996. He is currently the Associate Dean of academic and student affairs with the Henry Samueli School of Engineering and Applied Science, UCLA. His research interests are in the areas of communication theory with particular interest in low-density parity-check coding, short-blocklength communication with feedback, and coding for storage. He has received the National Science Foundation CAREER Award, the Okawa Foundation Award for research in information theory and telecommunications, and the Excellence in Teaching Award from the Henry Samueli School of Engineering and Applied Science.

Parimal Parag (S'04–M'11) received the M.Tech. and B.Tech. degrees in electrical engineering from IIT Madras in 2004 and the Ph.D. degree in electrical engineering from Texas A&M University in 2011. He is currently an Assistant Professor with the Department of Electrical and Communication Engineering, Indian Institute of Science. Prior to that, he was a Senior System Engineer in research and development with Assia, Inc., Redwood City, from 2011 to 2014.