Optimal Source Codes for Timely Updates

Prathamesh Mayekar, *Student Member, IEEE*, Parimal Parag, *Member, IEEE*,
and Himanshu Tyagi, *Senior Member, IEEE*

***Abstract*— A transmitter observing a sequence of independent and identically distributed random variables seeks to keep a receiver updated about its latest observations. The receiver need not be apprised about each symbol seen by the transmitter, but needs to output a symbol at each time instant t. If at time t the receiver outputs the symbol seen by the transmitter at time $U(t) \leq t$, the age of information at the receiver at time t is $t - U(t)$. We study the design of lossless source codes that enable transmission with minimum average age at the receiver. We show that the asymptotic minimum average age can be attained up to a constant gap by the Shannon codes for a tilted version of the original pmf generating the symbols, which can be computed easily by solving an optimization problem. Furthermore, we exhibit an example with alphabet $\mathcal{X}$ where Shannon codes for the original pmf incur an asymptotic average age of a factor $O(\sqrt{\log |\mathcal{X}|})$ more than that achieved by our codes. Underlying our prescription for optimal codes is a new variational formula for integer moments of random variables, which may be of independent interest. Also, we discuss possible extensions of our formulation to randomized schemes and to the erasure channel, and include a treatment of the related problem of source coding for minimum average queuing delay.**

***Index Terms*— Timely updates, source codes, Gibbs variational formula, age of information.**

I. Introduction

TIMELINESS is emerging as an important requirement for communication in cyber-physical systems (CPS). Broadly, it refers to the requirement of having the latest information from the transmitter available at the receiver in a timely fashion. It is important to distinguish the requirement of timeliness from that of low delay transmission: The latter places a constraint on the delay in transmission of each message, while timeliness is concerned about how recent is the current information at the receiver. In particular, the instantaneous staleness at the receiver is low if a message is received with low delay. However, the instantaneous staleness increases linearly at the receiver until a subsequent message is decoded successfully. A heuristically appealing metric that can capture the notion of timeliness of information in a variety of applications, termed its *age*, was first used in [12] for a setting involving queuing and link layer delays and was analyzed systematically for a queuing model in the pioneering work [13]; see [1], [3], [10], [14], [21], [23] for a sampling of subsequent developments in problems related to minimum age scheduling. In this paper, we initiate a systematic study of the design of source codes with the goal of minimizing the age of the information at the receiver.

As a motivating application, consider remote sensor data monitoring where at each instant the sensor observes real-valued, time-series measurements. For concreteness, the reader may consider voltage and current data recording using intelligent electronic devices in a power distribution network. The sensor communicates to a center over a network to enable fault detection and fault analysis. On the one hand, the communication protocol and buffer constraints at the sensor limits the rate at which the sensor can send data packets to the center. On the other hand, it is not very important for the center to get all the packets from the sensor. Rather the center wants timely updates about the sensor observations. In fact, when operating with cheap hardware with limited front-end buffers, it is common to have observation values in the buffer overwritten as new recordings are made even before the previous one waiting in the buffer has been picked-up for processing. Our work focuses on data compression for such applications where there is no direct cost of skipping packets and the interest is only in timely updates.

Specifically, we consider the problem of source coding where a transmitter receives symbols generated from a known distribution and seeks to communicate them to a receiver in a timely fashion.[1] To that end, it encodes a symbol x to $e(x)$ using a variable length prefix-free code e. The coded sequence is then transmitted over a noiseless communication channel that sends one bit per unit time. We restrict our treatment to a simple class of deterministic[2] update schemes, termed *memoryless update schemes*, where the transmitter does not have have a buffer to store the symbols it has seen previously and simply sends the next observed symbol once the channel is free.

Manuscript received October 12, 2018; revised October 2, 2019; accepted March 1, 2020. Date of publication March 25, 2020; date of current version May 20, 2020. This work was supported in part by the indigenous 5G Testbed project of the Department of Telecommunications, Ministry of Communications, India; in part by the Science and Engineering Research Board (SERB) under Grant DSTO-1677; and in part by a research grant from the Robert Bosch Center for Cyber-Physical Systems at the Indian Institute of Science, Bangalore. This article was presented at the IEEE International Symposium on Information Theory. *(Corresponding author Himanshu Tyagi.)*

The authors are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India (e-mail: prathamesh@iisc.ac.in; parimal@iisc.ac.in; htyagi@iisc.ac.in).

Communicated by V. Vaishampayan, Associate Editor for Source Coding.

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

Digital Object Identifier 10.1109/TIT.2020.2983151

[1] This assumption of known distribution is realized in practice by building a model for sensor data offline, before initiating the live monitoring process.

[2] Our analysis of average age extends to randomized schemes as well; see Section VI.

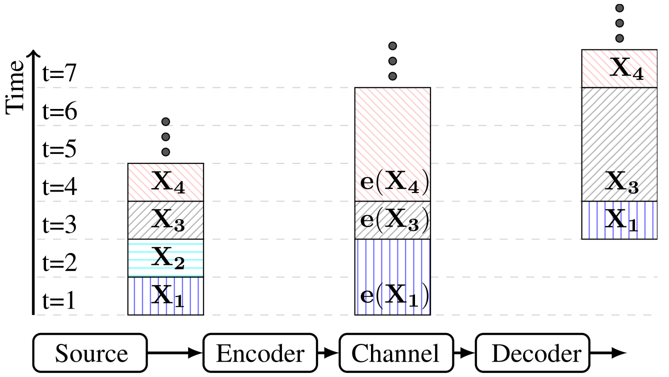


Fig. 1. Illustration of a memoryless update scheme for the first 4 packets in the source-queue.

Specifically, denoting the source alphabet by $\mathcal{X}$, the transmitter observes a symbol $X_t \in \mathcal{X}$ at each discrete time t. At time $t = 1$, the transmitter communicates the symbol $X_1 = x_1$ by encoding it to codeword $e(x_1)$ of length $\ell(x_1)$ bits. This transmission requires $\ell(x_1)$ channel uses and is received perfectly at the decoder at time $1 + \ell(x_1)$. Since the channel remains busy sending $e(x_1)$ for time instants 1 to $\ell(x_1)$, the transmitter cannot send any new symbols during this period. At time $t' = 1 + \ell(x_1)$, the transmitter observes the symbol $X_{t'} = x_{t'}$. Under a memoryless update scheme, the transmitter cannot store the symbols seen during the time interval $\{2, \ldots, \ell(x)\}$ and communicates codeword $e(x_{t'})$ over the next $\ell(x_{t'})$ channel uses, starting from the time instant $t' = 1 + \ell(x_1)$. The communication process continues repeatedly in this fashion.

We emphasize that under memoryless schemes, the source symbols generated and observed by the transmitter while the channel is busy sending a previous symbol are simply skipped. This skipping is only allowed when the channel is busy, and not at the will of the encoder when the channel is free (see Section VI for discussion on randomized schemes that allow the transmitted to skip symbols even when channel is free). Furthermore, the encoder need not indicate to the decoder that a symbol has been skipped using a special symbol – the decoder can ascertain this from the received communication since the channel is noiseless and compression is done using prefix-free codes.

On the receiver side, at each instance t the decoder outputs a time $U(t)$ and the symbol $X_{U(t)}$ seen by the transmitter at time $U(t)$. Thus, the *age of information* at the receiver at time t is given by $A(t) = t - U(t)$. We note that age of information measures timeliness at the receiver. When the transmitter skips source symbols, $U(t)$ remains unchanged at the receiver and the age $A(t)$ increases. Therefore, the age metric implicitly penalizes for skipping symbols.

We illustrate the setup in Figure 1. In this example, the symbol X_1 generated at time $t = 1$ is encoded to a two-bit codeword $e(X_1)$ and received at the decoder at time $t = 3$ after two channel uses. At time $t = 2$, the transmitter skips symbol X_2 since the channel was busy sending X_1 when it arrived. Further, the decoder retains $U(t) = 0$ since it has not received any symbol. At time $t = 3$, the decoder receives the codeword $e(X_1)$, updates $U(3) = 1$, and outputs the corresponding symbol X_1. Thus, the age of information at the receiver at time $t = 3$ is $A(3) = 2$. Since the channel becomes available at time $t = 3$, the transmitter encodes the symbol X_3 and transmits the one-bit codeword $e(X_3)$, which is received after a single channel-use at time $t = 4$. At time $t = 4$, the decoder outputs time $U(4) = 3$ with outputs the corresponding symbol X_3, and the age of information at the receiver is $A(4) = 1$. Once again, the channel becomes available at time $t = 4$ for the transmitter. It encodes the current symbol X_4 into the codeword $e(X_4)$ of length 3 bits and sends $e(X_4)$ over the channel; $e(X_4)$ is received at time $t = 7$. The decoder retains the output $U(t) = 3$ and $X_{U(t)} = X_3$ for times $t \in \{4, 5, 6\}$. At time $t = 7$, the decoder outputs time $U(7) = 4$ and the corresponding symbol X_4; the age of information at the receiver is $A(7) = 3$.

Our goal in this paper is to design prefix-free codes for which the average age of the memoryless scheme above is minimized; namely codes e that minimize

$$\bar{A}(e) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} A(t).$$

This formulation is apt for the timely update problem where the transmitter need not send each update and strives only to reduce the average age of the information at the receiver.

Using a simple extension of the renewal reward theorem, we derive a closed form formula for the asymptotic average age attained by a prefix-free code. Interestingly, this formula is a rational function of the first and the second moment of the random codeword length. Our main technical contribution in this paper is a variational formula for the second moment of random variables that enables an algorithm for finding the code that attains the minimum asymptotic average age up to a constant gap. The variational formula is of independent interest and may be useful in other settings where such cost functions arise; we point-out one such setting in Section VI. In fact, our prescribed prefix-free code is a Shannon code[3] for a tilted version of the original pmf. See (10) below for the description of the tilted version; it can be computed by solving an optimization problem entailing entropy maximization.

The formula for average age that we derive yields an $O(\log|\mathcal{X}|)$ upper bound on the minimum average age, attained by a fixed length code. We show that the same upper bound of $O(\log|\mathcal{X}|)$ holds for the average age of a Shannon code for the original distribution as well. However, we exhibit an example where Shannon codes for the original distribution have $\Omega(\log|\mathcal{X}|)$ age, while our aforementioned proposed code yields an average age of $O(\sqrt{\log|\mathcal{X}|})$.

The problem of designing update codes with low average age is related to real-time source coding ($cf.$ [16]) where we seek to transmit a stream of data under strict delay bounds. A related formulation has emerged in the control over communication network literature ($cf.$ [22]) where an observation is quantized and sent to an estimator/controller to enable control.

[3] A Shannon code for P is a prefix-free code that assigns lengths $\ell_S(x) = \lceil -\log P(x) \rceil$ to a symbol x ($cf.$ [5]).

Here, too, the requirement is that of communication under bounded delay.

An alternative formulation for minimum age source coding is considered in the recent work [25]. Unlike our formulation, skipping of symbols is prohibited in [25]. Instead, the authors consider fixed-to-variable length block codes and require that each coded symbol be transmitted over a constant rate, noiseless bit-pipe. In this setting, an exact expression for average age is not available, and the authors take recourse to an approximation for average age. This approximate average age is then optimized numerically over a set of prefix-free codes using the algorithm in [15]. The authors further reduce the computational complexity of this algorithm by using the algorithm in [2].

A recent paper [24] extends this problem to include random arrival times of source symbols and applies the algorithm from [15] for optimizing the cost function. Note that the cost function optimized in [15] is similar to the approximate average age of [24], [25], but with one crucial difference: While the former is monotonic in both first and second moments of random lengths, the latter is not. In absence of this monotonicity, the optimality of the solution produced by algorithm in [15] is not guaranteed for the cost functions in [24], [25]. In a related work [26], the same authors point-out that the average age can be further reduced by allowing the encoder to dynamically control the block-length of the fixed-to-variable length codes.

In contrast to [25], which is perhaps closest to our work, we derive an exact expression for average age and rigorously establish the structural properties of the optimal solution to the relaxed problem. In fact, our proposed minimum average age problem differs from all these prior formulations since we need not send the entire stream and are allowed to skip some symbols. In our applications of interest, such as that of real-time sensor data monitoring outlined earlier, the allowed communication rates are much lower than the rate at which data is generated. Thus, there is no hope of transmitting all the data at bounded delay, as mandated by the formulations available hitherto. Nonetheless, our setting is related closely to that in [25] and provides a complementary formulation for age optimal source coding. We note that our focus is on settings where the alphabet size of the streaming symbols is large. In such settings, the average age for any memoryless update scheme would be much larger than a small constant. Therefore, it suffices to establish optimality up to small additive constants.

In addition to our basic formulation, we present a few extensions of our formulations and other use cases for our proposed variational formula. Specifically, while we restrict to deterministic schemes for the most part, our analysis can be extended easily to analyze randomized schemes where the encoder can choose to skip an available transmission slot randomly. This idea of skipping transmission slots arises also in the recent work [21], albeit in a slightly different context. We exhibit an example where a particular randomized scheme outperforms every deterministic scheme. However, our analysis is limited and does not completely clarify the role of randomization; for instance, it remains unclear for which distributions can randomized schemes strictly outperform deterministic ones.

In another direction, we consider the case where the transmission channel is not error-free, but can erase each bit with a known probability. Furthermore, an ACK-NACK feedback indicating the success of transmission is available. Note that for the standard transmission problem, the simple repeat-until-succeed scheme is optimal in this setting. Our analysis can be used to design the optimal source code when we restrict our channel coding to this simple scheme. However, the optimality of the ensuing source-channel coding scheme remains unclear.

Finally, we study the related problem of source coding for ensuring minimum queuing delays. This problem, introduced in [11], is closely related to the minimum age formulation of this paper. Interestingly, our recipe for designing update codes with minimum average age can be extended to this setting as well. However, here, too, our results are somewhat unsatisfactory: Our approach only provides a solution to the real-relaxation of the underlying integer-valued optimization problem and naive rounding-off is far from optimal. Nonetheless, we have included these extensions in the current paper since they indicate the rich potential for our proposed techniques and provide new formulations for future research.

The next section contains a formal description of our setting and a formula for asymptotic average age of a code. Our main technical tool is presented in Section III, and we apply it to the minimum average age code design problem in Section IV. Numerical evaluations of our proposed scheme for the family of Zipf distributions is presented in Section V. Section VI contains a discussion on extensions to randomized schemes and erasure channel, along with a treatment of source codes for minimum average waiting time. We provide all the proofs in the final section.

Notation and Some Preliminaries: Random variables are denoted by capital letters X, Y etc., their realizations by small letters x, y etc., and their range sets by $\mathcal{X}, \mathcal{Y}$. The cardinality of the set $\mathcal{X}$ is denoted by $|\mathcal{X}|$. The set of all finite length binary sequences is denoted by $\{0,1\}^*$.

The logarithm to the base 2 is denoted by $\log a$ and the logarithm to the base e is denoted by $\ln a$. All the information theoretic measures considered in this paper – such as Entropy, Rényi divergence, and Kullback-Leibler divergence – are defined with logarithm to the base 2.

Next, we recall the notions of Shannon lengths and Shannon codes, which will be used throughout. A source code is called *prefix-free* if no codeword is a prefix of another.

Definition I.1 (Shannon Lengths and Shannon Codes for P): For a pmf P on an alphabet $\mathcal{X}$, the real-values $\ell(x) = -\log P(x), x \in \mathcal{X}$, are called the *Shannon lengths* for the pmf P. A prefix-free source code for P with codeword lengths $\ell(x) = \lceil -\log P(x) \rceil$, $\forall x \in \mathcal{X}$, is called a *Shannon code*[4] for the pmf P.

[4] There can be different codes with codeword lengths required in our definition of a Shannon code. We simply refer to all of them as a Shannon code, since any of these can serve our purpose in this paper.

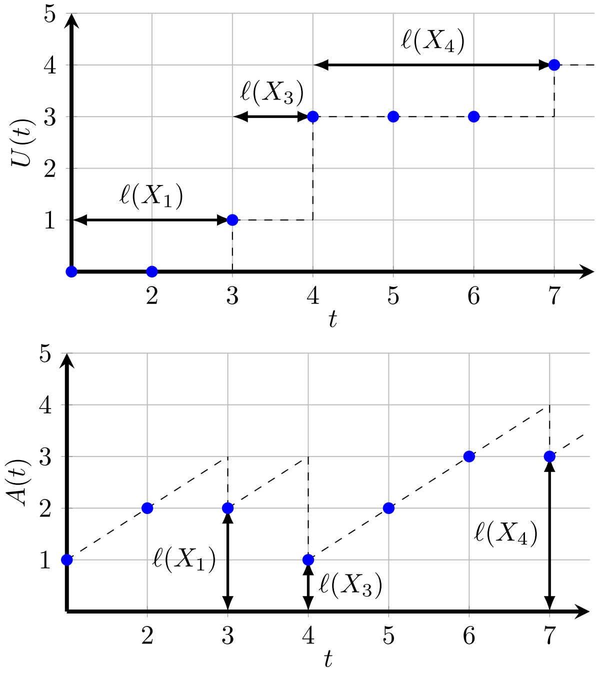


Fig. 2. A sample path of $U(t)$, $A(t)$ corresponding to Figure 1 starting with $U(1) = 0$.

II. Average Age for Memoryless Update Schemes

Consider a discrete-time system in which at every time instant t, a transmitter observes a symbol X_t generated from a finite alphabet $\mathcal{X}$ with pmf P. We assume that the sequence $\{X_t\}_{t=1}^{\infty}$ is independent and identically distributed (iid). The transmitter has a noiseless communication channel at its disposal over which it can transmit one bit per unit time. A *memoryless update scheme* consists of a prefix-free code, represented by its encoder $e : \mathcal{X} \to \{0,1\}^*$, and a decoder which at each time instant t declares a time index $U(t) \le t$ and an estimate $\hat{X}_{U(t)}$ for the symbol $X_{U(t)}$ that was observed by the encoder at time $U(t)$. We focus on error-free schemes and require $\hat{X}_{U(t)}$ to equal $X_{U(t)}$ with probability 1.

In a memoryless update scheme, once the encoder starts communicating a symbol x, encoded as $e(x)$, it only picks up the next symbol once all the bits in $e(x)$ have been transmitted successfully to the receiver. The time index $U(t)$ is updated to a new value only upon receiving all the encoded bits for the current symbol. That is, if the transmission of a symbol is completed at time $t-1$, the encoder will start transmitting $e(X_t)$ in the next instant. Moreover, if the final bit of $e(X_t)$ is received at time t', $U(t')$ is updated to t. A typical sample path for $U(t)$ is given in Figure 2. The age $A(t)$ of the symbol available at the receiver at time t is given by

$$A(t) = t - U(t).$$

A more general treatment can allow errors in estimates of $X_{U(t)}$ as well as encoders with memory, but we limit ourselves to the simple error-free and memoryless setting in this paper.

We are interested in designing prefix-free codes e that minimize the average age for the memoryless update scheme described above.

Definition II.1: The *average age* for a prefix-free code e, denoted $\bar{A}(e)$, is given by

$$\bar{A}(e) = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(t - U(t)).$$

We remark that $\bar{A}(e)$ can be viewed as the average area under the curve of $A(t)$ (w.r.t. t). Note that $\bar{A}(e)$ is random variable, nevertheless we will prove that this random variable is a constant almost surely. For any symbol $x \in \mathcal{X}$, we denote the length of the codeword $e(x)$ by $\ell(x)$. Let $X \in \mathcal{X}$ be a random symbol with pmf P over the alphabet $\mathcal{X}$, then the length of the random codeword $e(X)$ is denoted by

$$L = \ell(X).$$

The result below uses a simple extension of the classical renewal reward theorem ($cf.$ [19]) to provide a closed form expression for $\bar{A}(e)$ in terms of the first and the second moments of L.

Theorem II.2: Consider a random variable X with pmf P on $\mathcal{X}$. For a prefix-free code e, the average age $\bar{A}(e)$ is given by

$$\bar{A}(e) = \mathbb{E}\left[L\right] + \frac{\mathbb{E}\left[L^2\right]}{2\mathbb{E}\left[L\right]} - \frac{1}{2} \quad a.s. \quad . \tag{1}$$

The proof is deferred to Section VII-A.

Denoting by $\bar{A}^*$ the minimum average age over all prefix-free codes e, as a corollary of the characterization above, we can obtain the following bounds for $\bar{A}^*$.

Corollary II.3: For any pmf P over $\mathcal{X}$, the optimal average age $\bar{A}^*$ is bounded as

$$\frac{3}{2}H(P) - \frac{1}{2} \le \bar{A}^* \le \frac{3}{2}\log|\mathcal{X}| + 1.$$

The proof of lower bound simply uses Jensen's inequality $\mathbb{E}\left[L^2\right] \ge \mathbb{E}\left[L\right]^2$ and the fact that $\mathbb{E}\left[L\right] \ge H(P)$ for a prefix free code; the upper bound is obtained by using codewords of constant length $\lceil \log|\mathcal{X}| \rceil$.

Note that the lengths $\ell(x)$ are required to be nonnegative integers. However, for any set of real-valued lengths $\ell(x) \ge 0$, we can obtain integer-valued lengths by using the rounded-off values $\lceil \ell(x) \rceil$. Unlike the average length cost, the average age cost function identified in (1) is not an increasing function of the lengths. Nevertheless, by (1), the average age $\bar{A}(e)$ achieved when we use the rounded-off values can be bounded as follows: Denoting $\bar{L} := \lceil \ell(X) \rceil$, we have

$$\begin{aligned}\mathbb{E}\left[\bar{L}\right] + \frac{\mathbb{E}\left[\bar{L}^2\right]}{2\mathbb{E}\left[\bar{L}\right]} - \frac{1}{2} &\le \mathbb{E}\left[L+1\right] + \frac{\mathbb{E}\left[(L+1)^2\right]}{2\mathbb{E}\left[L\right]} - \frac{1}{2}\\ &\le \mathbb{E}\left[L\right] + \frac{\mathbb{E}\left[L^2\right]}{2\mathbb{E}\left[L\right]} + \frac{2\mathbb{E}\left[L\right]}{2\mathbb{E}\left[L\right]}\\ &\quad + \frac{1}{2\mathbb{E}\left[L\right]} + \frac{1}{2}\\ &\le \mathbb{E}\left[L\right] + \frac{\mathbb{E}\left[L^2\right]}{2\mathbb{E}\left[L\right]} + 2. \end{aligned} \tag{2}$$

Accordingly, in our treatment below we shall ignore the integer constraints and allow nonnegative real-valued length assignments.

Returning now to the bound of Corollary II.3, the upper and lower bounds differ only by a constant 1.5 when P is uniform. In view of the foregoing discussion, Shannon codes for a uniform distribution attain the minimum average age up to a constant gap. The next result gives an upper bound on average age for Shannon codes for an arbitrary P on $\mathcal{X}$.

Lemma II.4: Given a pmf P on $\mathcal{X}$, a Shannon code e for P has average age $\overline{A}(e)$ at most $O(\log|\mathcal{X}|)$.

Proof: Let $\ell(X)$ denote the lengths of Shannon code corresponding to P (see Definition I.1). We establish the claim using the standard bound $H(P') \leq \log|\mathcal{X}|$ for an appropriately chosen pmf P' on $\mathcal{X}$. Specifically, for the tilting of P given by $P'(x) \propto \ell(x)P(x)$, we get

$$\begin{aligned}\log|\mathcal{X}| &\geq \sum_{x\in\mathcal{X}} \frac{P(x)\ell(x)}{\mathbb{E}\left[\ell(X)\right]} \log\frac{\mathbb{E}\left[\ell(X)\right]}{P(x)\ell(x)} \\ &= \sum_{x\in\mathcal{X}} \frac{P(x)\ell(x)(-\log P(x))}{\mathbb{E}\left[\ell(X)\right]} \\ &\qquad - \sum_{x\in\mathcal{X}} \frac{P(x)\ell(x)}{\mathbb{E}\left[\ell(X)\right]} \log\frac{\ell(x)}{\mathbb{E}\left[\ell(X)\right]} \\ &\geq \sum_{x\in\mathcal{X}} \frac{P(x)\ell(x)(-\log P(x))}{\mathbb{E}\left[\ell(X)\right]} \\ &\qquad - \sum_{x\in\mathcal{X}:\ell(x)\geq\mathbb{E}[\ell(X)]} \frac{P(x)\ell(x)}{\mathbb{E}\left[\ell(X)\right]} \log\frac{\ell(x)}{\mathbb{E}\left[\ell(X)\right]}.\end{aligned}$$

Using $-\log P(x) \geq \ell(x) - 1$ and $\ln x \leq \frac{x^2-1}{2x}$ for $x \geq 1$, we obtain

$$\begin{aligned}\log|\mathcal{X}| &\geq \frac{\mathbb{E}\left[\ell^2(X)\right]}{\mathbb{E}\left[\ell(X)\right]} - 1 \\ &\qquad - \frac{1}{2\ln 2}\cdot \sum_{x\in\mathcal{X}:\ell(x)\geq\mathbb{E}[\ell(X)]} P(x)\left(\frac{\ell^2(x)}{\mathbb{E}\left[\ell(X)\right]^2} - 1\right) \\ &\geq \frac{\mathbb{E}\left[\ell^2(X)\right]}{\mathbb{E}\left[\ell(X)\right]} - 1 \\ &\qquad - \frac{1}{2\ln 2}\cdot \sum_{x\in\mathcal{X}:\ell(x)\geq\mathbb{E}[\ell(X)]} P(x)\cdot\frac{\ell^2(x)}{\mathbb{E}\left[\ell(X)\right]^2} \\ &\geq \frac{\mathbb{E}\left[\ell^2(X)\right]}{\mathbb{E}\left[\ell(X)\right]} - 1 - \frac{1}{2\ln 2}\cdot\sum_{x\in\mathcal{X}} \frac{P(x)\ell^2(x)}{\mathbb{E}\left[\ell(X)\right]^2} \\ &\geq \frac{\mathbb{E}\left[\ell^2(X)\right]}{\mathbb{E}\left[\ell(X)\right]} - 1 - \frac{1}{2\ln 2}\cdot\sum_{x\in\mathcal{X}} \frac{P(x)\ell^2(x)}{\mathbb{E}\left[\ell(X)\right]} \\ &\geq \left(1 - \frac{1}{2\ln 2}\right)\cdot\frac{\mathbb{E}\left[\ell^2(X)\right]}{\mathbb{E}\left[\ell(X)\right]} - 1,\end{aligned}$$

where the second-last inequality follows from the fact that $\mathbb{E}\left[\ell^2(X)\right] \geq \mathbb{E}\left[\ell(X)\right]$, which in turn follows from the fact that $\ell(X) \geq 1$. The proof is completed by rearranging the terms. □

It is of interest to examine if, in general, a Shannon code for P itself has average age close to $\bar{A}^*$, as was the case for the uniform distribution. In fact, it is not the case. Below we exhibit a pmf P where the average age of a Shannon code for P is $\Omega(\log|\mathcal{X}|)$, namely the previous bound is tight, and yet a Shannon code for another distribution (when evaluated for P) has an average age of only $O(\sqrt{\log|\mathcal{X}|})$.

Example II.5: Consider $\mathcal{X} = \{0, \ldots, 2^n\}$ and a pmf P on $\mathcal{X}$ given by

$$P(x) = \begin{cases} 1 - \frac{1}{n}, & x = 0 \\ \frac{1}{n2^n}, & x \in \{1, \ldots, 2^n\}. \end{cases}$$

Using (1), the average age $\bar{A}(e_P)$ for a Shannon code for P can be seen to satisfy $\bar{A}(e_P) \approx (n + 2\log n)/2$. On the other hand, if we instead use a Shannon code for the pmf Q given by

$$Q(x) = \begin{cases} \frac{1}{2^{\sqrt{n}}}, & x = 0 \\ \frac{1-2^{-\sqrt{n}}}{2^n}, & x \in \{1, \ldots, 2^n\}, \end{cases}$$

we get $\mathbb{E}\left[L\right] \approx \sqrt{n}$ and $EL^2 \approx 2n$, whereby $\bar{A}(e_Q) \approx 2\sqrt{n}$, just $O(\sqrt{\log|\mathcal{X}|})$. □

Thus, one needs to look beyond the standard Shannon codes for P to find codes with minimum average age. Interestingly, we show that Shannon codes for a tilted version of P attain the optimal asymptotic average age (up to the constant loss of at most 2.5 bits incurred by rounding-off lengths to integers). In particular, for the example above, our proposed optimal codes will have an average age of only $O(\sqrt{\log|\mathcal{X}|})$ in comparison to $\Omega(\log|\mathcal{X}|)$ of Shannon codes for P.

A key technical tool in design of our codes is a variational formula that will allow us to linearize the cost function in (1), thereby rendering Shannon codes for a tilted distribution optimal. We present this in the next section.

III. A Variational Formula for p-Norm

The expression for average age identified in Theorem II.2 involves the second moment of the random codeword length L. This is in contrast to the traditional variable length source coding problem where the goal is to minimize the average codeword length $\mathbb{E}\left[L\right]$. For this standard cost, Shannon codes which assign a codeword of length $\lceil -\log P(x)\rceil$ to the symbol x come within 1-bit of the optimal cost (see, for instance, [5]). A variant of this standard problem was studied in [4], where the goal was to minimize the log-moment generating function $\log\mathbb{E}\left[\exp(\lambda L)\right]$. A different approach for solving this problem is given in [9] where the *Gibbs variational principle* is used to linearize the nonlinear cost function $\log\mathbb{E}\left[\exp(\lambda L)\right]$. The next result provides the necessary variational formula to extend the aforementioned approach to another nonlinear function, namely $\|L\|_p := (\mathbb{E}\left[L^p\right])^{\frac{1}{p}}$ for $p > 1$.

We believe that our result is of independent interest, and present it in a general form that applies to general distributions (and not just the discrete random variables considered in this paper). To state the general result, we recall a basic notation from probability theory. For two probability measures P and Q on the same probability space such that Q is absolutely continuous with respect to P, denoted $Q \ll P$, denote by $\frac{dQ}{dP}$ the Radon-Nikodym derivative of Q with respect to P. Note that $\frac{dQ}{dP}$, too, is a random variable measurable with respect to the underlying sigma-algebra. A reader not familiar with these notions can see a standard textbook on probability theory for definitions. For the discrete case, $Q \ll P$ corresponds to the condition[5] $\texttt{supp}(Q) \subset \texttt{supp}(P)$ and $\frac{dQ}{dP}$ equals the ratio of the pmfs of the distributions Q and P.

Note that expectations are always taken with respect to the reference measure. In particular, the expectations without any subscript in Theorem III.1 below and its proof denote the expectation with respect to P, which is the reference measure in this case. The expectation in Remark 1 denotes the expectation with respect to R.

Theorem III.1: For a real-valued random variable X with distribution P and $p \geq 1$ such that $\|X\|_p < \infty$, we have

$$\|X\|_p = \max_{Q \ll P} \mathbb{E}\left[\left(\frac{dQ}{dP}\right)^{\frac{1}{p'}} |X|\right],$$

where $p' = p/(p-1)$ is the Hölder conjugate of p.

Proof: For $Q \ll P$ and $0 < \alpha \neq 1$, let $D_\alpha(P, Q)$ denote the Rényi divergence of order α between distributions Q and P (see [18]), defined by

$$D_\alpha(P, Q) := \frac{1}{\alpha - 1} \log \mathbb{E}\left[\left(\frac{dQ}{dP}\right)^{\alpha}\right].$$

It is well-known that $D_\alpha(P, Q) \geq 0$ with equality if and only if $P = Q$. Consider the probability measure $P_p \ll P$ defined by

$$\frac{dP_p}{dP} := \frac{1}{\|X\|_p^p} \cdot |X|^p.$$

Then, for $\alpha = 1/p'$,

$$\begin{aligned} 0 \leq D_\alpha(P_p, Q) &= \frac{1}{\alpha - 1} \log \mathbb{E}\left[\left(\frac{dQ}{dP}\right)^{\alpha}\left(\frac{dP_p}{dP}\right)^{1-\alpha}\right] \\ &= -p \log \mathbb{E}\left[\left(\frac{dQ}{dP}\right)^{\alpha} |X|\right] + p \log \|X\|_p, \end{aligned}$$

where the previous equality holds since $p(1 - \alpha) = 1$. Thus, for every $Q \ll P$,

$$\mathbb{E}\left[\left(\frac{dQ}{dP}\right)^{\alpha} |X|\right] \leq \|X\|_p,$$

with equality if and only if $P_p = Q$. □

Remark 1: The given definition of Rényi divergence restricts Theorem III.1 to the case $P(X = 0) = 0$. To remove this restriction, the following general definition of Rényi divergence with respect to a common measure can be used: For all $Q, P \ll R$, define

$$D_\alpha(P, Q) := \frac{1}{\alpha - 1} \log \mathbb{E}\left[\left(\frac{dQ}{dR}\right)^{\alpha}\left(\frac{dP}{dR}\right)^{1-\alpha}\right].$$

The proof then follows by using the positivity of $D_\alpha(P_p, Q)$, then by proceeding in the same manner as the previous proof.

Returning to the problem at hand, we apply the variational formula above to the L_2 norm of a discrete random variable. We highlight this special case separately below.

Corollary III.2: For a nonnegative discrete random variables X with a pmf P such that $\|X\|_2 < \infty$, we have

$$\|X\|_2 = \max_{\texttt{supp}(Q) \subset \texttt{supp}(P)} \sum_{x \in \mathcal{X}} \sqrt{Q(x)P(x)}\, x,$$

where $\texttt{supp}(P)$ denotes the support-set of the distribution P.

IV. Prefix-Free Codes With Minimum Average Age

We now present a recipe for designing prefix-free codes with minimum average age. By Theorem II.2, we seek prefix-free codes that minimize the cost

$$\mathbb{E}[L] + \frac{\mathbb{E}[L^2]}{2\mathbb{E}[L]}, \tag{3}$$

where $L = \ell(X)$ for X with pmf P. Recall that a prefix-free code with lengths $\{\ell(x) \in \mathbb{N}, x \in \mathcal{X}\}$ exists if and only if lengths satisfy Kraft's inequality (*cf.* [5]), *i.e.*, if and only if

$$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1. \tag{4}$$

Following the discussion leading to (2), we relax the integral constraints for $\ell(x)$ and search over all real-valued $\ell(x) \geq 0$ satisfying (4). Specifically, we solve the relaxed optimization problem

$$\min_{\ell \in \Lambda} \mathbb{E}[L] + \frac{\mathbb{E}[L^2]}{2\mathbb{E}[L]}, \tag{5}$$

where

$$\Lambda = \left\{\ell \in \mathbb{R}^{|\mathcal{X}|} : \sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1,\ \ell(x) \geq 0\ \forall x \in \mathcal{X}\right\}.$$

As noticed in (2), this can incur a loss of only a constant. A key challenge in minimizing (3) is that it is nonlinear. We linearize this cost as follows:

1) Note first the identity below, which is obtained by maximizing the expression on the right-side:

$$\mathbb{E}[L] + \frac{\mathbb{E}[L^2]}{2\mathbb{E}[L]} = \max_{z \geq 0} \left(1 - \frac{z^2}{2}\right) \mathbb{E}[L] + z\|L\|_2. \tag{6}$$

2) Then, Corollary III.2 yields

$$\|L\|_2 = \max_{Q \ll P} \sum_{x \in \mathcal{X}} \sqrt{Q(x)P(x)}\, \ell(x),$$

which further leads to

$$\begin{aligned} &\mathbb{E}[L] + \frac{\mathbb{E}[L^2]}{2\mathbb{E}[L]} \\ &= \max_{z \geq 0} \left(1 - \frac{z^2}{2}\right) \mathbb{E}[L] + z \max_{Q \ll P} \sum_{x \in \mathcal{X}} \sqrt{Q(x)P(x)}\, \ell(x) \\ &= \max_{z \geq 0} \max_{Q \ll P} \sum_{x \in \mathcal{X}} g_{z,Q,P}(x) \ell(x), \end{aligned}$$

where

$$g_{z,Q,P}(x) := \left(1 - \frac{z^2}{2}\right) P(x) + z\sqrt{Q(x)P(x)}. \tag{7}$$

As remarked earlier, as the source distribution P is discrete, the constraint $Q \ll P$ simplifies to $\texttt{supp}(Q) \subset \texttt{supp}(P)$. Thus, our goal is to identify the minimizer ℓ^* that achieves

$$\Delta^*(P) = \min_{\ell \in \Lambda} \max_{z \geq 0} \max_{Q \ll P} \sum_{x \in \mathcal{X}} g_{z,Q,P}(x) \ell(x). \tag{8}$$

[5] $\texttt{supp}(P)$ denotes the support of distribution P over an alphabet $\mathcal{X}$, *i.e.*, $\texttt{supp}(P) := \{x \in \mathcal{X} : P(x) > 0\}$.

The result below captures our main observation and facilitates the computation of optimal lengths attaining the minmax cost $\Delta^*(P)$.

Theorem IV.1 (*Structure of optimal codes):* The optimal minmax cost $\Delta^*(P)$ in (8) satisfies

$$\begin{aligned}\Delta^*(P) &= \max_{z\geq 0}\max_{Q\ll P}\min_{\ell\in\Lambda}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x)\\ &= \max_{\substack{z\geq 0, Q\ll P,\\ (z,Q)\in\mathcal{G}}}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\log\frac{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}{g_{z,Q,P}(x)},\end{aligned} \tag{9}$$

where

$$\mathcal{G} := \{z\geq 0, Q\in\mathbb{R}^{|\mathcal{X}|} : g_{z,Q,P}(x)\geq 0 \quad \forall x\in\mathcal{X}\}.$$

Furthermore, if (z^*, Q^*) is the maximizer of the right-side of (9), then the minmax cost (8) is achieved uniquely by the Shannon lengths[6] for the pmf P^* on $\mathcal{X}$ given by

$$P^*(x) = \frac{g_{z^*,Q^*,P}(x)}{\sum_{x'\in\mathcal{X}} g_{z^*,Q^*,P}(x')}. \tag{10}$$

Thus, our prescription for design of source codes is simple: Use a Shannon code for P^* instead of P. To compute P^*, we need to solve the optimization problem in (9). Note that is unclear a priori that the minimum average age for the problem in (5) would correspond to Shannon lengths for some pmf since our cost function is not monotonic in expected length, whereby the optimal solution may not satisfy Kraft's inequality with equality. Nonetheless, we show that the Shannon lengths $-\log P^*(x)$ are optimal for the relaxed problem given by (5).

We note that our formal result above only provides a structural result for the optimal solution. But we believe that this structural result leads to a recipe to design practical algorithms for finding the optimal solution; we describe this recipe below. Specifically, note that the resulting optimization problem for finding P^* is one of entropy maximization for which several heuristic recipes are available. Furthermore, we note the following structural simplification for the optimal solution which shows that if $P(x) = P(y)$, then $P^*(x) = P^*(y)$ must hold as well; the proof is relegated to the Appendix. Thus, the dimension of the optimization problem (9) can be reduced from $|\mathcal{X}| + 1$ to $M_P + 1$, where M_P denotes the number of distinct elements in the probability multiset $\{P(x) : x\in\mathcal{X}\}$. Let $A_1\cdots A_{M_P}$ denote the partition of $\mathcal{X}$ such that

$$P(x) = P(y) \quad \forall x, y\in A_i, \quad \forall i\in[M_P].$$

Lemma IV.2: Suppose that Q^* is an optimal Q for (9). Then, Q^* must satisfy

$$Q^*(x) = Q^*(y) \quad \forall x, y\in A_i, \quad \forall i\in[M_P]. \tag{11}$$

In proving Lemma IV.2, we use the fact that the cost function in (9) is concave in Q for each fixed z and is concave in z for each fixed Q (see Lemma .2). However, it may not be jointly concave in (z, Q). Nevertheless, we apply standard numerical packages to optimize it in the next section to quantify the performance of our proposed codes and compare it with Shannon codes for the original distribution P.

V. Numerical Results for Zipf Distribution

We program all our optimization problems in *AMPL* [7] and solve it using *SNOPT* [8] and *CONOPT* [6] solvers. Specifically, for the pmfs P we consider in this section, we solve the optimization problem given by (9) to find the corresponding optimal (z^*, Q^*). In order to check if we have indeed found the optimal (z^*, Q^*), we once again use Theorem IV.1. In particular, it follows from Theorem IV.1 that the necessary and sufficient condition for a particular (z, Q) to be the optimal solution is that the value of the maximization problem (9) at (z, Q) equals

$$\mathbb{E}[-\log P'(X)] + \frac{\mathbb{E}\left[(\log P'(X))^2\right]}{2\mathbb{E}[-\log P'(X)]},$$

where

$$P'(X) = \frac{g_{z,Q,P}(x)}{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')};$$

in all our numerical evaluations, the solution found by the solver satisfies this condition, which establishes its optimality.

We now illustrate our recipe for construction of prefix-free codes that yield minimum average age for memoryless update schemes when P is a Zipf distribution. Specifically, we illustrate our qualitative results using the $\texttt{Zipf}(s, N)$ distribution with alphabet $\mathcal{X} = \{1, \cdots, N\}$ and given by

$$P(i) = \frac{i^{-s}}{\sum_{j=1}^{N} j^{-s}}, \quad 1\leq i\leq N.$$

Heuristically, the average age formula (1) suggests that the differences between the performances of a code under average codeword length cost and the average age cost will be the most for "peaky distribution," namely for distributions with heavy elements. The parameter s of the Zipf distribution allows us to vary from a uniform distribution to a "peaky distribution," making this family apt for our numerical study. Indeed, our numerical results confirm that our proposed scheme outperforms a Shannon code for P when the parameter s is high; see Figure 3. When we round-off real lengths to integers, the gains are subsided but still exist. Further, when the parameter s is close to 0, Shannon codes for P are close to optimal. With increase in s, the gain of our proposed schemes over Shannon codes starts becoming more prominent. As an aside, Figure 3 also provides an illustration of the non-monotonic nature of the average age function with respect to code lengths.

The distribution P^* we use to construct our codes seems to be a flattened version of the original Zipf distribution; we illustrate the two distributions for $\texttt{Zipf}(1, 8)$ in Figure 4. As we see in Figure 4, P^* and P are very close in this case. Indeed, we illustrate in Figure 5 that the average length $\mathbb{E}[L]$ when Shannon lengths $-\log P(x)$ are used and when $-\log P^*(x)$ are used are very close[7]. In Figure 5, we note the dependence of average age on the entropy of the underlying

[6] Recall that Shannon lengths for the pmf P on $\mathcal{X}$ are given by $\ell(x) = -\log P(x)$, $x\in\mathcal{X}$, and are not necessarily integers.

[7] The difference of these two average lengths (averaged w.r.t. P) is given by the Kullback-Leibler divergence $D(P\|P^*)$; see [5].

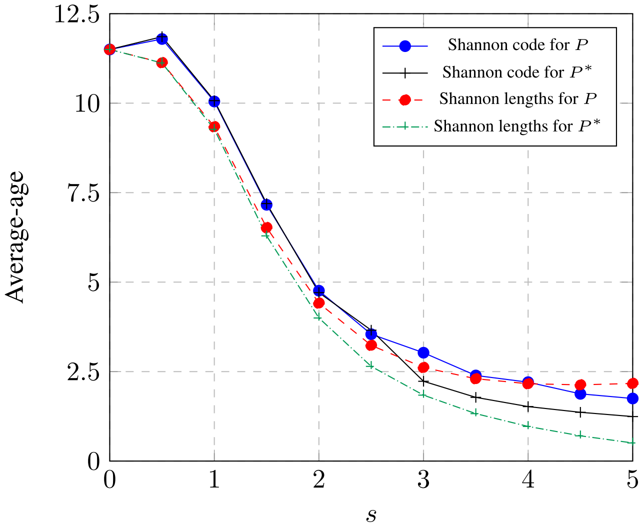


Fig. 3. Comparison of proposed codes and Shannon codes for $\texttt{Zipf}(s, 256)$ with varying s. The average age is computed using real-valued lengths as well as lengths rounded-off to integer values.

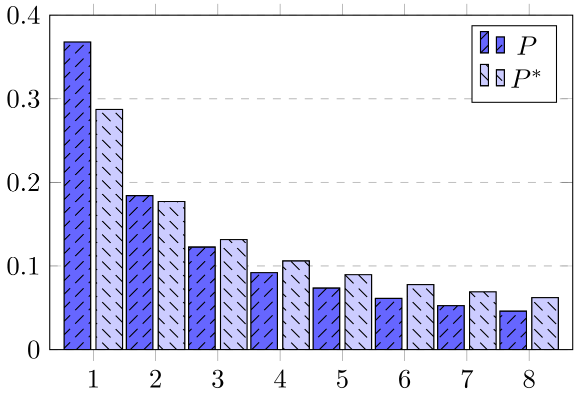


Fig. 4. The pmf for P^* and P for $\texttt{Zipf}(1, 8)$.

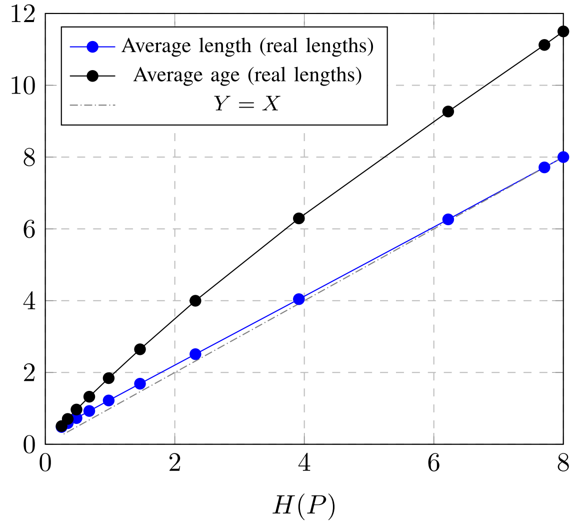


Fig. 5. Average age and average length for our update codes as a function of $H(P)$ for $\texttt{Zipf}(s, 256)$ with s varying from 0 to 5 at step sizes of 0.5.

distribution P. As expected, average age increases as $H(P)$ increases.

Thus, while Example II.5 illustrated high gains of the proposed code over Shannon codes for P, for the specific case of Zipf distributions the gains may not be large. Characterizing this gain for any given distribution is a direction for future research.

VI. Extensions

A. Randomization for Timely Updates

We have restricted our treatment to deterministic memoryless update schemes. A natural extension to randomized memoryless schemes would entail allowing the encoder to make a randomized decision to skip transmission of a symbol even when the channel is free (we can allocate a special symbol $\emptyset$ to signify no transmission to the receiver). Specifically, assume that we transmit the symbol $\emptyset$ using a codeword of length $\ell(\emptyset)$ when we choose not to transmit the observed symbol $x \in \mathcal{X}$. Denoting by $\theta(x)$ the probability with which the encoder will transmit the symbol x, the average age $\bar{A}(e, \theta)$ for the randomized scheme is given by

$$\bar{A}(e, \theta) = \frac{\mathbb{E}\left[L(\theta)\right]}{\mathbb{E}\left[\theta(X)\right]} + \frac{\mathbb{E}\left[L(\theta)^2\right]}{2\mathbb{E}\left[L(\theta)\right]} - \frac{1}{2}, \tag{12}$$

where the random variable $L(\theta)$ is defined as follows:

$$L(\theta) := \begin{cases} \ell(x), & w.p \quad P(x)\theta(x) \\ \ell(\emptyset), & w.p \quad 1 - \mathbb{E}\left[\theta(X)\right]. \end{cases} \tag{13}$$

Note that the expression in (12) is a slight generalization of Theorem II.2 and is derived in Section VII-A.

Example VI.1: Consider $\mathcal{X} = \{1, \ldots, 64\}$ and the following pmf;

$$P(x) = \begin{cases} 1/4, & x \in \{1, \ldots, 3\}, \\ 1/244, & x \in \{4, \ldots, 64\}. \end{cases}$$

Since $H(P) = 3.483$, Corollary II.3 yields that the average age of the deterministic memoryless update scheme is bounded below by 4.724. Next, consider a randomized update scheme with $\theta(x) = 1$ for $x \in \{1, 2, 3\}$ and 0 otherwise. For this choice, the effective pmf P_θ is uniformly distributed over the symbols $\{1, 2, 3\} \cup \{\phi\}$. Thus, the optimal length assignment for this case assigns $\ell(x) = 2$ to all the symbols and the average age equals 3.17, which is less than the lower bound of 4.724 for the deterministic scheme.

The idea of skipping available transmission opportunities, i.e., not transmitting even when the channel is free, to minimize average age appears in the recent work [21] as well, albeit in a slightly different setting. Heuristically, the randomization scheme above operates as we expect – it ignores the rare symbols which will require longer codeword lengths. In practice, however, these rare symbols might be the ones we are interested in. But keep in mind that our prescribed solution only promises to minimize the average age and does not pay heed to any other consideration. Furthermore, for a given randomization vector θ, we can establish a result similar to Theorem IV.1. This will lead to the design of almost optimal source codes for a given randomization vector θ. However, the joint optimality over the class of randomized schemes and source coding schemes is still unclear.

In a more comprehensive treatment, one can study the design of update codes with other constraints imposed. We foresee the use of Corollary III.2 in these more general settings as well. In another direction, we can consider the extension of our results to the case when the transmission channel is an erasure channel with probability of erasure ε. If we assume the availability of perfect feedback, a natural model for the link or higher layer in a network, and restrict to simple repetition schemes where the transmitter keeps on transmitting the coded symbol until it is received, our formula for average age extends with (roughly) an additional multiplicative factor of $1/(1-\varepsilon)$. Formally the average age over an erasure channel with ε probability of erasure; a source code e, along with a randomization vector θ and a repetition channel-coding scheme yields the following average age

$$\bar{A}_{\varepsilon}(e,\theta) = \frac{1}{1-\varepsilon}\cdot \bar{A}(e,\theta) + \frac{\varepsilon}{2(1-\varepsilon)}.$$

However, the optimality of repetition scheme is unclear, and the general problem constitutes a new formulation in joint-source channel coding which is of interest for future research.

B. Source Coding for Minimum Queuing Delay

Next, we point out a use case for Corollary III.2 in a minimum queuing delay problem introduced in [11]. The setting is closely related to our minimum average age update formulation with two differences: First, the arrival process of source symbols is a Poisson process of rate λ; and second, the encoder is not allowed to skip source symbols. Instead, each symbol is encoded and scheduled for transmission in a first-come-first-serve (FCFS) queue. Our goal is to design a source code that minimizes the average queuing delay encountered by the source sequence. Formally, the symbols $\{X_n\}_{n=1}^{\infty}$ are generated iid from a finite alphabet $\mathcal{X}$, using a common pmf P. Every incoming symbol x is encoded as $e(x)$ using a prefix-free code specified by the encoder mapping $e : \mathcal{X} \to \{0,1\}^*$, and the bit string $e(x)$ is placed in a queue. The queue schedules bits for transmission using a FCFS policy. Each bit in the queue is transmitted over a noiseless communication channel. Denote by A_n the time of successful arrival of the nth symbol. Also, denote by D_n the time instant of successful reception of the nth symbol X_n. That is, D_n is the instant at which the last bit of $e(X_n)$ is received[8]. The delay for the nth symbol is given by $D_n - A_n$; see Figure 6 for an illustration.

Thus, if $\ell(x)$ is the length of the encoded symbol $e(x)$ in bits, then the number of channel uses to transmit this symbol is $\ell(x)$, whereby the service time of the n^{th} arriving symbol is given by $S_n = \ell(X_n)$. Since $\{X_n\}_{n=1}^{\infty}$ is iid and the encoder mapping e is fixed, the sequence $(S_n)_{n\in\mathbb{N}}$, too, is iid with common mean $\mathbb{E}[L]$. Therefore, the resulting queue is an M/G/1 queuing system with Poisson arrivals of rate λ and

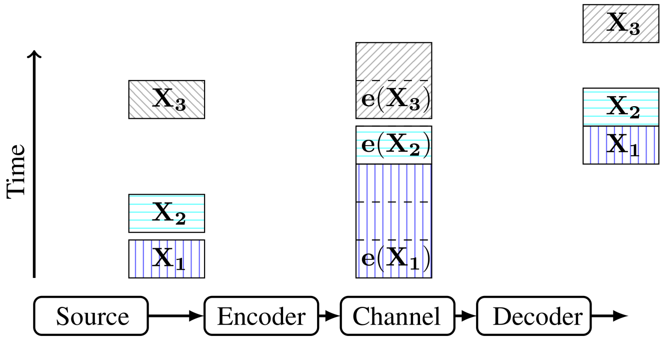


Fig. 6. Figure describes a typical sample-path for transmission of encoded symbols over a FCFS queuing system. Symbol X_1 arrives at some time instant 1, it is encoded and transmitted over the channel. Recall that unlike the slotted setup of Figure 1, the setup here is that of continuous time with Poisson arrivals. It is decoded at time instant 4. Symbol X_2 arrives in between time instants 2 and 3, and is placed in the queue, as the channel is busy transmitting X_1. As soon as the channel becomes free at time instant 4, an encoded version of X_2 is transmitted over it. Symbol X_3 arrives when the channel is free and is transmitted immediately.

iid service times $(S_n)_{n\in\mathbb{N}}$. Note that this queue will be stable only if $\lambda\mathbb{E}[S_n] = \lambda\mathbb{E}[L] < 1$.

We are interested in designing prefix-free codes e that minimize the average waiting time defined as follows:

Definition VI.2: The *average waiting time* $D(e)$ of a source code e is given by

$$D(e) := \limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[D_n - A_n],$$

where the expectation is over source symbol realizations $\{X_n\}_{n=1}^{\infty}$ and arrival instants $\{A_n\}_{n\in\mathbb{N}}$.

We seek prefix-free codes e with the least possible average waiting time $D(e)$. In fact, a closed-form expression for $D(e)$ was obtained in [11]. For clarity of exposition, we denote the load for the queuing system above for a fixed λ by $\rho(L) := \lambda\mathbb{E}[L]$. Since $\rho(L) < 1$ for the queue to be stable, the average codeword length $\mathbb{E}[L]$ must be strictly less than a threshold $L_{\mathtt{th}} := \frac{\mathbb{E}[L]}{\rho(L)} = \frac{1}{\lambda}$ for the queue to be stable.

Theorem VI.3 ([11]): Consider a random variable X with pmf P and a source code e which assigns a bit sequence of length $\ell(x)$ to $x \in \mathcal{X}$. Let L denote the random variable $\ell(X)$. Then, the average waiting time $D(e)$ for e is given by

$$D(e) = \begin{cases} \frac{\mathbb{E}\left[L^2\right]}{2(L_{\mathtt{th}}-\mathbb{E}[L])} + \mathbb{E}[L], & \mathbb{E}[L] < L_{\mathtt{th}}, \\ \infty, & \mathbb{E}[L] \geq L_{\mathtt{th}}. \end{cases} \tag{14}$$

Thus, the problem of designing source codes with minimum average waiting time reduces to that of designing a prefix-free code that minimizes the cost in (14). This problem was first considered in [11]. In fact, it was noted in [11, Chapter 1, Section 3] that codes which minimize the first moment are robust for (14). We will justify this empirical observation in Corollary VI.5. However, optimal codes can differ from Shannon codes for P. Indeed, an algorithm for finding the optimal length assignments $\ell(x)$, $x \in \mathcal{X}$, for a prefix-free code that minimizes $\bar{D}(e)$ was presented in [15] and the optimal code can be seen to outperform Shannon codes for P. While this algorithm has complexity that is polynomial in

[8] Note both A_n and D_n may not be integer valued, unlike the age setup.

the alphabet size, it is computationally expensive for large alphabet sizes – the case of interest for our problem.

Interestingly, the cost function in (14) resembles closely the expression we obtained for asymptotic average age and our recipe used to design minimum average age codes can be applied to design minimum average delay codes as well. The underlying optimization problem can be solved numerically rather quickly, much faster than the optimization in [15]. However, as before, our procedure can only handle the real-relaxation of the underlying optimization problem, and unlike the previous case, naive rounding-off to integer lengths yields a sub-optimal solution when $(1 - \rho(L))$ is small. Nonetheless, the minimum average waiting time computed using our recipe serves as an easily computable lower bound for the optimal $D(e)$. In fact, we observe in our numerical simulations that the resulting lower bound is rather close to the optimal cost obtained using [15].

Now, we describe the modification of our recipe to design codes with $\mathbb{E}\left[L\right] < L_{\mathtt{th}}$ that minimize the cost

$$\|L\|_1 + \frac{\|L\|_2^2}{2(L_{\mathtt{th}} - \|L\|_1)}, \tag{15}$$

where $L = \ell(X)$ for X with pmf P. As before, we first obtain a variational form of (15) which entails a linear function of lengths. Specifically, we have the following steps.

1) First, we obtain a polynomial form from the rational function:

$$\frac{\|L\|_2^2}{2(L_{\mathtt{th}} - \|L\|_1)} = \max_{z \geq 0} z\,\|L\|_2 - \frac{z^2}{2}(L_{\mathtt{th}} - \|L\|_1).$$

2) Then, Corollary III.2 yields that the cost in (15) equals

$$\max_{z\geq 0}\max_{Q\ll P}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) - \frac{z^2}{2}L_{\mathtt{th}}$$

where the $g_{z,Q,P}(x)$ is defined as

$$g_{z,Q,P}(x) := \left(1 + \frac{z^2}{2}\right)P(x) + z\sqrt{Q(x)P(x)}.$$

Thus, our goal reduces to identifying the minimizer $\ell^* \in \Lambda$ that achieves

$$\Delta^*(P) = \min_{\substack{\ell\in\Lambda,\\ \mathbb{E}[L]<L_{\mathtt{th}}}}\max_{z\geq 0}\max_{Q\ll P}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) - \frac{z^2}{2}L_{\mathtt{th}}. \tag{16}$$

The result below is the counterpart of Theorem IV.1 for minimum delay source codes and is proved in Section VII-C.

Theorem VI.4: Under the condition

$$H(X) + \log(1 + 1/\sqrt{2}) < L_{\mathtt{th}}, \tag{17}$$

the optimal minmax cost $\Delta^*(P)$ in (16) satisfies

$$\begin{aligned}\Delta^*(P) &= \max_{z\geq 0}\max_{Q\ll P}\min_{\substack{\ell\in\Lambda,\\ \mathbb{E}[L]<L_{\mathtt{th}}}}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) - \frac{z^2}{2}L_{\mathtt{th}}\\ &= \max_{z\geq 0}\max_{Q\ll P}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\log\frac{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}{g_{z,Q,P}(x)}\\ &\qquad - \frac{z^2}{2}L_{\mathtt{th}}.\end{aligned} \tag{18}$$

Furthermore, if (z^*, Q^*) is the maximizer of the right-side of (18), then the minmax cost (16) is achieved uniquely by Shannon lengths for pmf P^* on $\mathcal{X}$ given by

$$P^*(x) = \frac{g_{z^{\prime}Q^*,P}(x)}{\sum_{x'\in\mathcal{X}} g_{z^*,Q^*,P}(x')}.$$

We remark that (14) implies that $H(X) < L_{\mathtt{th}}$ is essential for the existence of a prefix free source coding scheme with finite average delay. Thus, the condition $H(X) + \log$ $(1 + 1/\sqrt{2}) < L_{\mathtt{th}}$ is a mild one.

Thus, as before, the optimal codeword lengths for the relaxed problem (allowing real-valued lengths) correspond, once again, to Shannon lengths for a titled distribution P^*. As remarked earlier, the performance of the optimal source code is known to be not too far from the Shannon code for P. This observation can be justified by the following simple corollary of Theorem VI.4.

Corollary VI.5: The KL-Divergence between P, P^* is bounded as

$$D(P||P^*) \leq \log\left(1 + \frac{1}{\sqrt{2}}\right).$$

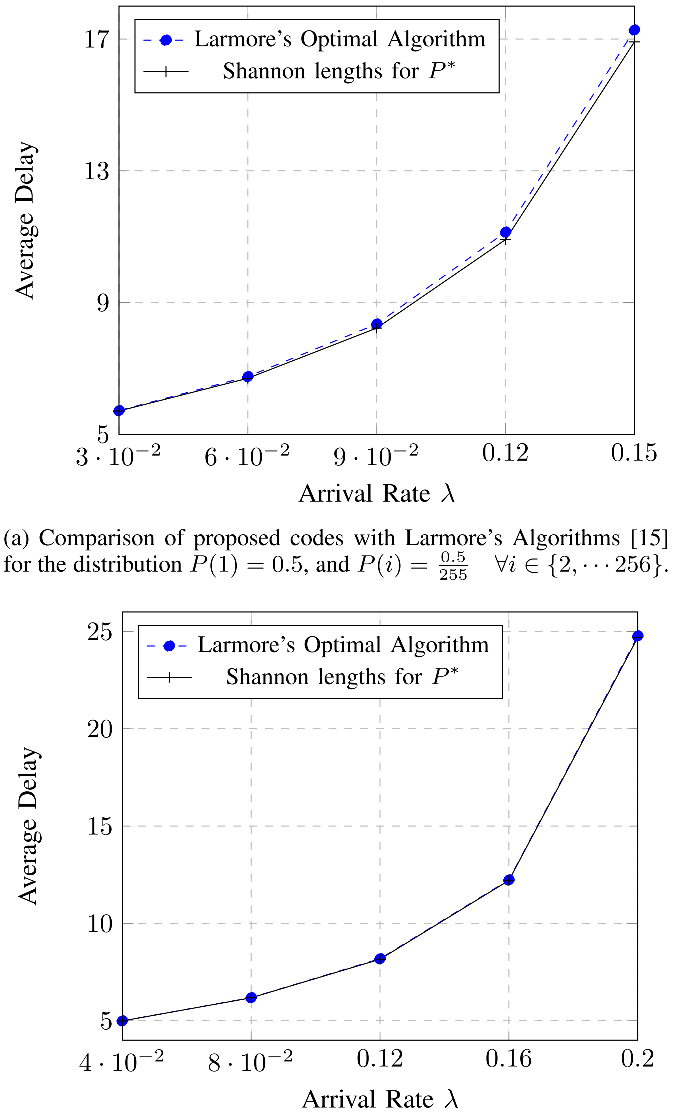
Proof: The proof follows from (37), which is in turn derived in the proof of theorem VI.4 in section VII-C. □

Thus, the average length for Shannon codes and our codes do not differ by more than $\log(1 + 1/\sqrt{2})$ $(cf.$ [5]). Indeed, we note in Figures 7a, 7b via numerical simulations that the optimal cost in (18) is very close to the performance of optimal codes designed using [15]. This suggests that possibly there is an appropriate rounding-off procedure for real-valued lengths that can yield integer lengths with close to optimal performance; devising such a rounding-off procedure is an interesting research direction for the future. We close this section by noting that analogous versions of Lemma IV.2 and Lemma .2 in the Appendix can be obtained for optimization problem (18).

VII. Proofs

A. Proof of Theorem II.2

We establish the expression for average age given in (12) for the more general class of randomized schemes; Theorem II.2 will follow upon setting $\theta(x) = 1$, for all $x \in \mathcal{X}$. Recall that the symbol $\emptyset$ is available only in the extended model in Section VI, and not in the original model discussed in rest of the paper. Note that the formula for average age given in Theorem II.2 is similar in form to the expressions for average age derived in other settings; see [13] for an example.



(a) Comparison of proposed codes with Larmore's Algorithms [15] for the distribution $P(1) = 0.5$, and $P(i) = \frac{0.5}{255} \quad \forall i \in \{2, \cdots 256\}$.

(b) Comparison of proposed codes with Larmore's Algorithms [15] for the distribution $P(1) = 0.6$, and $P(i) = \frac{0.4}{255} \quad \forall i \in \{2, \cdots 256\}$.

Fig. 7. Comparison of proposed codes with Larmore's Algorithms.

We will first set up some notation. Let $S_0 := 0$ and

$$S_k := \inf\{t > S_{k-1} : U(t) > U(t-1)\},\ k \in \mathbb{N}.$$

Namely, S_k is the time at which the decoder updates its estimate for the symbol for the kth time. Recall that $U(t)$ is incremented only on successful reception at the receiver and is strictly increasing in t. For brevity, we introduce the notation $Y_k := S_k - S_{k-1}$ for the time between the $(k-1)$th and the kth information update at the decoder. Further, denote by $Z_k := S_k - U(S_k)$ the age at time S_k, which is simply the time taken for the successful reception of the symbol[9] $x \in \mathcal{X}$ transmitted at time $U(S_k)$. Also, denote by R_k the sum of instantaneous age between S_{k-1} and S_k (the kth reward), namely

$$R_k := \sum_{t=S_{k-1}+1}^{S_k} (t - U(t)).$$

Heuristically, our proof can be understood as follows. We note that the asymptotic average age is roughly

$$\frac{\sum_{k=1}^{\infty} R_k}{\lim_{k\to\infty} S_k}.$$

It is easy to see that $\{Y_k\}_{k=1}^{\infty}$ is an iid sequence. Thus, if $\{R_k\}_{k=1}^{\infty}$, too, was an iid sequence, we would obtain the asymptotic average age to be $\mathbb{E}[R_1]/\mathbb{E}[Y_1]$ by the standard Renewal Reward Theorem [19]. Unfortunately, this is not the case. But it turns out that the dependence in sequence $\{R_k\}$ is only between consecutive terms. Therefore, we can obtain the same conclusion as above by dividing the sum $\sum_{k=1}^{\infty} R_k$ into the sum of odd terms and even terms, each of which is in turn a sum of iid random variables.

We will now proceed to prove that dependence in R_k is between consecutive terms. Since $U(t)$ remains $U(S_{k-1})$ for all $t < S_k$, we get for $k \geq 1$ that

$$\begin{aligned} R_k = &\frac{(S_k - S_{k-1} - 1)(S_k - S_{k-1})}{2} \\ &+ (S_k - S_{k-1} - 1) \cdot (S_{k-1} - U(S_{k-1})) \\ &+ S_k - U(S_k) \\ = &\frac{1}{2}Y_k^2 + Y_k\left(Z_{k-1} - \frac{1}{2}\right) + Z_k - Z_{k-1}, \end{aligned} \tag{19}$$

with Z_0 set to 0.

Note that since the source sequence $\{X_n\}$ is iid and the randomization θ is stationary, the sequences Y_k and Z_k are iid, too. Therefore, the $(R_{2n})_{n\in\mathbb{N}}$ and $(R_{2n+1})_{n\in\mathbb{N}}$ are both[10] iid sequences with $\mathbb{E}[R_{2n}] = \mathbb{E}[R_{2n+1}] = \mathbb{E}[R_2]$ for all n.

Using this observation, we can obtain the following expression for the average age:

$$\bar{A}(e, \theta) = \frac{\mathbb{E}[R_2]}{\mathbb{E}[Y_1]}. \tag{20}$$

Before we prove (20), which is the main ingredient of our proof, we evaluate the expression on the right-side.

For $\mathbb{E}[Y_1]$, note that Y_1 gets incremented by $\ell(\emptyset)$ each time $\emptyset$ is sent, and gets incremented finally by $\ell(x)$ once a symbol $x \in \mathcal{X}$ is sent. Thus, Y_1 takes the value $\ell(x) + r\ell(\emptyset)$ with probability $(1 - \mathbb{E}[\theta(X)])^r \theta(x) P(x)$. Denoting $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, we get

$$\begin{aligned} \mathbb{E}[Y_1] &= \sum_{x\in\mathcal{X}} \sum_{r\in\mathbb{N}_0} (\ell(x) + r\ell(\phi))P(x)\theta(x)(1 - \mathbb{E}[\theta(X)])^r \\ &= \sum_{x\in\mathcal{X}} \sum_{r\in\mathbb{N}_0} \ell(x)P(x)\theta(x)(1 - \mathbb{E}[\theta(X)])^r \\ &\quad + \sum_{x\in\mathcal{X}} \sum_{r\in\mathbb{N}_0} r\ell(\phi)P(x)\theta(x)(1 - \mathbb{E}[\theta(X)])^r \\ &= \frac{\sum_{x\in\mathcal{X}} \ell(x)P(x)\theta(x)}{\mathbb{E}[\theta(X)]} + \frac{\ell(\phi)(1 - \mathbb{E}[\theta(X)])}{\mathbb{E}[\theta(X)]} \\ &= \frac{\mathbb{E}[L(\theta)]}{\mathbb{E}[\theta(X)]}. \end{aligned}$$

[9] This must be a symbol in $\mathcal{X}$ and not $\emptyset$ by the definition of S_k.

[10] The initial term R_1 has a different distribution since $Z_0 = 0$.

For $\mathbb{E}[R_2]$, it follows from (19) that

$$\mathbb{E}[R_2] = \frac{1}{2}\mathbb{E}\left[Y_2^2\right] + \mathbb{E}[Y_2 Z_1] - \frac{1}{2}\mathbb{E}[Y_2],$$

since $\mathbb{E}[Z_2] = \mathbb{E}[Z_1]$. Also, note that Z_1 only depends on the symbol $x \in \mathcal{X}$ received at time S_1 which in turn can depend only on the symbols X_n for $n \leq S_1 - 1$. On the other hand, $Y_2 = S_2 - S_1$ depends on symbols X_n for $n \geq S_1$ and the outputs of the independent coin tosses corresponding to randomization θ. Therefore, Z_1 is independent of Y_2, whereby

$$\mathbb{E}[R_2] = \frac{1}{2}\mathbb{E}\left[Y_2^2\right] + \mathbb{E}[Y_2]\left(\mathbb{E}[Z_1] - \frac{1}{2}\right).$$

Next, note that Z_1 takes the value $\ell(x)$, $x \in \mathcal{X}$, when the symbol received at S_1 is x. This latter event happens with probability

$$\sum_{r=0}^{\infty}(1 - \mathbb{E}[\theta(X)])^r \theta(x)P(x) = \frac{\theta(x)P(x)}{\mathbb{E}[\theta(X)]},$$

and so, by the definition of $L(\theta)$ in (13),

$$\begin{aligned}\mathbb{E}[Z_1] &= \frac{\sum_x \ell(x)\theta(x)P(x)}{\mathbb{E}[\theta(X)]}\\ &= \frac{\mathbb{E}[L(\theta)]}{\mathbb{E}[\theta(X)]} - \frac{\ell(\emptyset)(1 - \mathbb{E}[\theta(X)])}{\mathbb{E}[\theta(X)]}.\end{aligned}$$

Then by denoting $p_\emptyset = 1 - \mathbb{E}[\theta(X)]$, the second moment $\mathbb{E}\left[Y_1^2\right]$ can be computed by observing the following recursion:

$$\begin{aligned}&\mathbb{E}\left[Y_1^2\right]\\ &= \sum_{x\in\mathcal{X}}\sum_{r\in\mathbb{N}_0}(\ell(x) + r\ell(\emptyset))^2 P(x)\theta(x)p_\emptyset^r\\ &= \sum_{x\in\mathcal{X}}\ell(x)^2 P(x)\theta(x)\\ &\qquad + p_\emptyset\sum_{x\in\mathcal{X}}\sum_{r\in\mathbb{N}}(\ell(x) + r\ell(\emptyset))^2 P(x)\theta(x)p_\emptyset^{r-1}\\ &= \sum_{x\in\mathcal{X}}\ell(x)^2 P(x)\theta(x)\\ &\qquad + p_\emptyset\sum_{x\in\mathcal{X}}\sum_{r\in\mathbb{N}}\big(\ell(x) + (r-1)\ell(\emptyset)\big)^2 P(x)\theta(x)p_\emptyset^{r-1}\\ &\qquad + 2\ell(\emptyset)p_\emptyset\sum_{x\in\mathcal{X}}\sum_{r\in\mathbb{N}}\big(\ell(x) + (r-1)\ell(\emptyset)\big)P(x)\theta(x)p_\emptyset^{r-1}\\ &\qquad + p_\emptyset\sum_{x\in\mathcal{X}}\sum_{r\in\mathbb{N}}\ell(\emptyset)^2 P(x)\theta(x)p_\emptyset^{r-1}\\ &= \sum_{x\in\mathcal{X}}\ell(x)^2 P(x)\theta(x)\\ &\qquad + p_\emptyset\mathbb{E}\left[Y_1^2\right] + 2\ell(\emptyset)(1 - \mathbb{E}[\theta(X)])\mathbb{E}[Y_1] + \ell(\emptyset)^2 p_\emptyset,\end{aligned}$$

which upon rearrangement yields

$$\mathbb{E}\left[Y_1^2\right] = \frac{\mathbb{E}\left[L(\theta)^2\right]}{\mathbb{E}[\theta(X)]} + 2\mathbb{E}[Y_1]\cdot\frac{\ell(\emptyset)p_\emptyset}{\mathbb{E}[\theta(X)]}.$$

Upon combining the relations derived above, we get

$$\frac{\mathbb{E}[R_2]}{\mathbb{E}[Y_1]} = \frac{\mathbb{E}\left[L(\theta)^2\right]}{2\mathbb{E}[L(\theta)]} + \frac{\mathbb{E}[L(\theta)]}{\mathbb{E}[\theta(X)]} - \frac{1}{2},$$

which with (20) completes the proof.

It remains to establish (20). The proof is a simple extension of the renewal reward theorem to our sequence of rewards R_n in which adjacent terms may be dependent. We include it here for completeness. Note that $(Y_n)_{n\in\mathbb{N}}$ is a sequence of non-negative iid random variables with mean $\mathbb{E}[Y_1]$, and $S_n = \sum_{k=1}^{n} Y_k$ for all $n \in \mathbb{N}$. The sequence $\{S_n\}$ serves as a sequence of renewal times and R_n denotes the reward accumulated in the nth renewal interval (though not in the standard iid sense). Define $N(t)$ to be the number of receptions up to time $t > 0$, $i.e.$,

$$N(t) = \sup\{n : S_n \leq t\},$$

and $R(t)$ to be the cumulative reward accumulated till time t, $i.e.$,

$$R(t) = \sum_{k=1}^{N(t)} R_k.$$

With this notation, we have

$$\begin{aligned}\frac{R(t)}{t} &= \frac{\sum_{k=1}^{N(t)} R_k}{t} &(21)\\ &= \frac{\sum_{k=1}^{N(t)} R_k}{N(t)}.\frac{N(t)}{t}. &(22)\end{aligned}$$

Note that

$$\begin{aligned}\frac{\sum_{k=1}^{\lfloor\frac{N(t)}{2}\rfloor}\sum_{i\in\{0,1\}} R_{2k+i}}{N(t)} &\leq \frac{\sum_{k=2}^{N(t)} R_k}{N(t)}\\ &\leq \frac{\sum_{k=1}^{\lceil\frac{N(t)}{2}\rceil}\sum_{i\in\{0,1\}} R_{2k+i}}{N(t)}.\end{aligned}$$

We now analyze the two bounds in the previous set of inequalities. Since $\mathbb{E}[Y_1]$ is finite, we get (see [19] for a proof)

$$\lim_{t\to\infty}\frac{N(t)}{t} \to \frac{1}{\mathbb{E}[Y_1]} \quad a.s., \tag{23}$$

which also shows that $N(t) \to \infty \quad a.s.$ as $t \to \infty$. Therefore, for $i \in \{0, 1\}$,

$$\frac{\sum_{k=1}^{\lceil\frac{N(t)}{2}\rceil} R_{2k+i}}{N(t)} = \frac{\sum_{k=1}^{\lceil\frac{N(t)}{2}\rceil} R_{2k+i}}{\lceil\frac{N(t)}{2}\rceil}\cdot\frac{\lceil\frac{N(t)}{2}\rceil}{N(t)}.$$

Since $(R_{2k+i})_{k\in\mathbb{N}}$ is iid and $N(t) \to \infty \quad a.s.$ as $t \to \infty$, strong law of large numbers yields

$$\lim_{t\to\infty}\frac{\sum_{k=1}^{\lceil\frac{N(t)}{2}\rceil} R_{2k+i}}{\lceil\frac{N(t)}{2}\rceil} = \mathbb{E}[R_2] \quad a.s. \quad \forall i \in \{0, 1\},$$

which further gives

$$\lim_{t\to\infty}\frac{\sum_{k=1}^{\lceil\frac{N(t)}{2}\rceil}\sum_{i\in\{0,1\}} R_{2k+i}}{N(t)} = \mathbb{E}[R_2] \quad a.s..$$

Similarly,

$$\lim_{t\to\infty}\frac{\sum_{k=1}^{\lfloor\frac{N(t)}{2}\rfloor}\sum_{i\in\{0,1\}} R_{2k+i}}{N(t)} = \mathbb{E}[R_2] \quad a.s..$$

Combining the observations above, we get

$$\lim_{t\to\infty} \frac{\sum_{k=1}^{N(t)} R_k}{N(t)} = \mathbb{E}\left[R_2\right] \quad a.s.,$$

which together with (22) and (23) yields (20).

B. Proof of Theorem IV.1

Our proof is based on noticing that the minmax cost $\Delta^*(P)$ in (8) involves weighted average length with weights $g_{z,Q,P}(x)$. In fact, we will see below that there is no loss in restricting to nonnegative weights, whereby our cost has a form of average length with respect to a distribution that depends on (z, Q). The broad idea of the proof is to establish that a optimal code corresponding to the *least favorable* choice of (z, Q) is minmax optimal. However, the proof is technical since our cost function may not satisfy the assumptions in a standard saddle-point theorem.

A simpler form of the minmax cost $\Delta^*(P)$ from (6) is given by

$$\Delta^*(P) = \min_{\ell\in\Lambda} \max_{z\geq 0} f(\ell, z), \tag{24}$$

where

$$f(\ell, z) := -z^2 \frac{\mathbb{E}\left[L\right]}{2} + z\sqrt{\mathbb{E}\left[L^2\right]} + \mathbb{E}\left[L\right]. \tag{25}$$

We seek to apply the following version of Sion's minmax theorem to the function f.

Theorem VII.1 (Sion's Minmax Theorem [20]): Let $\mathcal{X}$ be convex space and $\mathcal{Y}$ be a convex, compact space. Let h be a function on $\mathcal{X}\times\mathcal{Y}$ which is convex on $\mathcal{X}$ for every fixed y in $\mathcal{Y}$ and concave on $\mathcal{Y}$ for every fixed x in $\mathcal{X}$. Then,

$$\inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} h(x, y) = \sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} h(x, y).$$

Indeed, the following lemma shows that our function f satisfies the convexity requirements of Sion's minmax theorem.

Lemma VII.2: $f(\ell, z)$ is convex in ℓ for every fixed $z \geq 0$ and concave in z for a fixed $\ell \in \Lambda$.

Proof: To show that $f(\ell, z)$ is a convex function of ℓ for every fixed $z \geq 0$, it suffices to show that $\sqrt{\mathbb{E}\left[L^2\right]}$ is convex in $L = \ell(X)$. To that end, let $L_1 = \ell_1(X)$ and $L_2 = \ell_2(X)$, for some ℓ_1 and ℓ_2 in λ. For all $\lambda \in [0, 1]$,

$$\sqrt{\mathbb{E}\left[(\lambda L_1 + (1-\lambda)L_2)^2\right]} \leq \lambda\sqrt{\mathbb{E}\left[L_1^2\right]} + (1-\lambda)\sqrt{\mathbb{E}\left[L_2^2\right]},$$

where the inequality is by Minkowski inequality for $\|L\|_2$.

The concavity in z can be seen easily by noticing that $\frac{\partial^2 f(\ell,z)}{\partial z^2} \leq 0$ for all ℓ in λ. □

However, our underlying domains of optimization are not compact. Our proof below circumvents this difficulty by showing that we may replace one of the domains by a compact set. For ease of reading, we divide the proof into 3 steps; we begin by summarize the flow at a high-level. The first step is to show that this minmax cost remains unchanged when the domain of z is restricted to a bounded interval $[0, K]$ for a sufficiently large K. This will allow us to interchange $\min_{\ell\in\Lambda}$ and $\max_{z\in[0,K]}$ in the second step by using Theorem VII.1 to obtain

$$\Delta^*(P) = \max_{z\in[0,K]} \min_{\ell\in\Lambda} f(\ell, z). \tag{26}$$

Furthermore, we then use Corollary III.2 to linearize the cost. But this brings in the maximization over an additional parameter Q, which we again interchange with the minimum over ℓ using Sion's minmax theorem (Theorem VII.1). Note that the required convexity of the cost function is easy to see; we note it in the following lemma.

Lemma VII.3: For every fixed $z \geq 0$, $\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x)$ is convex in ℓ for a fixed $Q \ll P$ and concave in Q for a fixed $\ell \in \Lambda$.

Proof: For every fixed $z \geq 0$, the cost function $\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x)$ is linear, and thereby convex, in ℓ for a fixed Q. For concavity in Q, note that for a fixed $\ell \in \Lambda$, the function $\sqrt{Q(x)}$ is a concave function of $Q(x)$, for all x in $\mathcal{X}$. □

Thus, we obtain

$$\Delta^*(P) = \max_{z\in[0,K], Q\ll P} \min_{\ell\in\Lambda} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x).$$

In the final step, we will establish that the optimal code for linear cost with weights corresponding to the least favorable (z, Q) is minmax optimal. We now present each step in detail.

Step 1: We begin by noting that there is no loss in restricting to codes with[11] $\mathbb{E}\left[L\right] \leq \log|\mathcal{X}|$. Indeed, note that for $\mathbb{E}\left[L\right] > \log|\mathcal{X}|$ the average age is bounded as

$$\mathbb{E}\left[L\right] + \frac{\mathbb{E}\left[L^2\right]}{2\mathbb{E}\left[L\right]} \geq \frac{3}{2}\mathbb{E}\left[L\right] > \frac{3}{2}\log|\mathcal{X}|, \tag{27}$$

where we have used Jensen's inequality. On the other hand, a fixed-length code with $\ell(x) = \log|\mathcal{X}|$ attains

$$\mathbb{E}\left[L\right] + \frac{\mathbb{E}\left[L^2\right]}{2\mathbb{E}\left[L\right]} = \frac{3}{2}\log|\mathcal{X}|, \tag{28}$$

which gives the desired form

$$\begin{aligned}\Delta^*(P) &= \min_{\ell\in\Lambda, \mathbb{E}[L]\leq\log\mathcal{X}} \mathbb{E}\left[L\right] + \frac{\mathbb{E}\left[L^2\right]}{2\mathbb{E}\left[L\right]} \\ &= \min_{\ell\in\Lambda, \mathbb{E}[L]\leq\log\mathcal{X}} \max_{z\in\mathbb{R}} f(\ell, z),\end{aligned} \tag{29}$$

where $f(\ell, z)$ is defined in (25). Also, for a fixed ℓ in Λ the function $f(\ell, z)$ attains its maximum at $z^*(\ell)$ given by

$$z^*(\ell) := \frac{\sqrt{\mathbb{E}\left[L^2\right]}}{\mathbb{E}\left[L\right]}.$$

[11] For simplicity, we assume that $\log\mathcal{X}$ is an integer.

For $\mathbb{E}[L] \leq \log|\mathcal{X}|$, the maximizer $z^*(\ell)$ is bounded as[12]

$$\begin{aligned} z^*(\ell) &\leq \frac{\sqrt{\mathbb{E}[L^2]}}{H(X)} \\ &= \frac{\sqrt{\sum_x P(x)\ell(x)^2}}{H(X)} \\ &\leq \frac{\mathbb{E}[L]}{H(X)}\sqrt{\max_{x\in\mathcal{X}} \frac{1}{P(x)}} \\ &\leq \frac{\log|\mathcal{X}|}{H(X)}\sqrt{\frac{1}{\min_{x\in\mathcal{X}} P(x)}}, \end{aligned}$$

where the first inequality uses $\mathbb{E}[L] \geq H(X)$, which holds for every prefix-free code, and the second holds since $\|a\|_2 \leq \|a\|_1$ for any sequence $a = (a_1, \ldots, a_n)$. Denoting

$$K := \frac{\log|\mathcal{X}|}{H(X)}\sqrt{\frac{1}{\min_{x\in\mathcal{X}} P(x)}},$$

(29) yields

$$\Delta^*(P) = \min_{\ell\in\Lambda, \mathbb{E}[L]\leq \log|\mathcal{X}|} \max_{z\in[0,K]} f(\ell, z).$$

Next, we show that the minmax cost above remains unchanged when we drop the constraint $\mathbb{E}[L] \leq \log|\mathcal{X}|$ in the outer minimum, which will complete the first step of the proof and establish (26). Indeed, since by (28) the minimum over $\ell \in \Lambda$ such that $\mathbb{E}[L] \leq \log|\mathcal{X}|$ is at most $(3/2)\log|\mathcal{X}|$, it suffices to show that

$$\min_{\ell\in\Lambda, \mathbb{E}[L]>\log|\mathcal{X}|} \max_{z\in[0,K]} f(\ell, z) > \frac{3}{2}\log|\mathcal{X}|. \tag{30}$$

We divide the proof of this fact into two cases. First consider the case when ℓ in Λ is such that $\mathbb{E}[L] > \log|\mathcal{X}|$ and $K \geq z^*(\ell)$. Then, $\max_{z\in[0,K]} f(\ell, z)$ equals $\max_{z\geq 0} f(\ell, z)$, which is bounded below by $(3/2)\log|\mathcal{X}|$ using (27) and the definition of $f(\ell, z)$. For the second case when $\mathbb{E}[L] > \log|\mathcal{X}|$ and $K < z^*(\ell)$, we have

$$\begin{aligned} \max_{z\in[0,K]} f(\ell, z) &= -K^2\frac{\mathbb{E}[L]}{2} + K\sqrt{\mathbb{E}[L^2]} + \mathbb{E}[L] \\ &> K^2\frac{\mathbb{E}[L]}{2} + \mathbb{E}[L] \\ &> \frac{3}{2}\cdot\mathbb{E}[L] \\ &> \frac{3}{2}\cdot\log|\mathcal{X}|, \end{aligned}$$

where the first inequality uses $K < z^*(\ell) = \sqrt{\mathbb{E}[L^2]}/\mathbb{E}[L]$ and the second holds since $K \geq 1$ from its definition. Therefore, we have established (30), and so we have

$$\Delta^*(P) = \min_{\ell\in\Lambda, \mathbb{E}[L]\leq\log|\mathcal{X}|} \max_{z\in[0,K]} f(\ell, z) = \min_{\ell\in\Lambda} \max_{z\in[0,K]} f(\ell, z).$$

Step 2: By lemma VII.2, $f(\ell, z)$ is convex in ℓ for every fixed $z \geq 0$ and concave in z for a fixed $\ell \in \Lambda$, z takes values in a convex compact set $[0, K]$, and the set $\{\ell : \ell \in \Lambda\}$ is convex, we get from Sion's minmax theorem (Theorem VII.1) that

$$\Delta^*(P) = \min_{\ell\in\Lambda} \max_{z\in[0,K]} f(\ell, z) = \max_{z\in[0,K]} \min_{\ell\in\Lambda} f(\ell, z).$$

Using Corollary III.2, we have

$$\|L\|_2 = \max_{Q\ll P} \sum_{x\in\mathcal{X}} Q(x)^{\frac{1}{2}} P(x)^{\frac{1}{2}} \ell(x),$$

which by the definition of f in (25) further yields

$$f(\ell, z) = \max_{Q\ll P} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x), \tag{31}$$

where

$$g_{z,Q,P}(x) = \left(1 - \frac{z^2}{2}\right)P(x) + z\sqrt{Q(x)P(x)}.$$

We have obtained

$$\Delta^*(P) = \max_{z\in[0,K]} \min_{\ell\in\Lambda} \max_{Q\ll P} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x). \tag{32}$$

From Lemma VII.3, $\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x)$ is convex in ℓ, for all $Q \ll P$, and concave in Q, for a fixed $\ell \in \Lambda$. Furthermore, since the set $\{Q : Q \ll P\}$ is convex compact for a pmf P on finite alphabet, using Sion's minmax theorem (Theorem VII.1) once again, we get

$$\Delta^*(P) = \max_{z\in[0,K]} \max_{Q\ll P} \min_{\ell\in\Lambda} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x), \tag{33}$$

which completes our second step.

Step 3: By (33), we get

$$\Delta^*(P) \leq \max_{z\geq 0} \max_{Q\ll P} \min_{\ell\in\Lambda} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x).$$

On the other hand, by (24) and (31) we have

$$\begin{aligned} \Delta^*(P) &= \min_{\ell\in\Lambda} \max_{z\geq 0} \max_{Q\ll P} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) \\ &\geq \max_{z\geq 0} \max_{Q\ll P} \min_{\ell\in\Lambda} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x), \end{aligned}$$

whereby

$$\begin{aligned} \Delta^*(P) &= \min_{\ell\in\Lambda} \max_{z\geq 0} \max_{Q\ll P} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) \\ &= \max_{z\geq 0} \max_{Q\ll P} \min_{\ell\in\Lambda} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x), \end{aligned} \tag{34}$$

which proves the first part of theorem IV.1.

Next, we claim that in the maxmin formula above, the maximum is attained by a (z, Q) for which $g_{z,Q,P}(x)$ is non-negative for every x. Indeed, if for some z, Q there exists an x' in $\mathcal{X}$ such that $g_{z,Q,P}(x')$ is negative, then the cost $\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x)$ is minimized by any ℓ such that $\ell(x') = \infty$ and the minimum value is $-\infty$. Such z, Q clearly can't be the optimizer of the maxmin problem, since for $z = 0$, we have $g_{z,Q,P} \geq 0$, which in turn leads to $\min_{\ell\in\Lambda} \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) \geq 0$.

[12] We assume without loss of generality that $P(x) > 0$ for every $x \in \mathcal{X}$.

Finally, consider (z,Q) such that $g_{z,Q,P}(x) \geq 0$ for all $x \in \mathcal{X}$. For such a (z,Q), we seek to identify the minimized ℓ below:

$$\begin{aligned}&\min_{\ell\in\Lambda}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x)\\ &= \sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')\min_{\ell\in\Lambda}\sum_{x\in\mathcal{X}}\frac{g_{z,Q,P}(x)}{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}\ell(x).\end{aligned} \tag{35}$$

Thus, our optimization problem reduces to the standard problem of designing minimum average length prefix-free codes for the pmf

$$P_{z,Q}(x) = \frac{g_{z,Q,P}(x)}{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}.$$

By Shannon's source coding theorem for variable length codes, the minimum is achieved by

$$\ell^*_{z,Q}(x) := \log\frac{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}{g_{z,Q,P}(x)}.$$

Furthermore, $\ell^*_{z,Q}$ is the unique minimizer in Λ.

Consider now a maximizer (z^*,Q^*) of the maxmin problem in (34), and let $\ell^o = \ell^*_{z^*,Q^*}$. Then, by Lemma .1 in the appendix, $(\ell^o,(z^*,Q^*))$ is a saddle-point for the minmax problem in (34). Moreover, ℓ^o is the unique minmax optimal solution.

C. Proof of Theorem VI.4

Denoting

$$f(\ell,z) = -z^2\frac{(L_{\mathtt{th}}-\mathbb{E}[L])}{2} + z\sqrt{\mathbb{E}[L^2]} + \mathbb{E}[L], \tag{36}$$

the optimal cost $\Delta^*(P)$ can be written as

$$\begin{aligned}\Delta^*(P) &= \inf_{\ell\in\Lambda,\mathbb{E}[L]<L_{\mathtt{th}}}\frac{\mathbb{E}[L^2]}{2(L_{\mathtt{th}}-\mathbb{E}[L])} + \mathbb{E}[L]\\ &= \min_{\ell\in\Lambda,\mathbb{E}[L]<L_{\mathtt{th}}}\max_{z\geq 0} f(\ell,z).\end{aligned}$$

This form is similar to the one we had in Theorem IV.1. But the proof there does not extend to the case at hand. Specifically, note that for each ℓ, $f(\ell,z)$ attains its maximum value for $z^*(\ell) = \frac{\sqrt{\mathbb{E}[L^2]}}{(L_{\mathtt{th}}-\mathbb{E}[L])}$ which, unlike the quantity that we obtained in the proof of Theorem IV.1, is unbounded over the set of $\ell \in \Lambda$ such that $\mathbb{E}[L] \leq L_{\mathtt{th}}$. However, under the additional assumption $H(X) + \log(1+1/\sqrt{2}) < L_{\mathtt{th}}$, we can provide a simpler alternative proof. We rely on the following lemma.

Lemma VII.4: Consider a function $h : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ such that the set $\mathcal{X}$ is compact convex, the set $\mathcal{Y}$ is convex, $h(x,y)$ is a convex function of x for every fixed y and a concave function of y for every fixed x. Suppose additionally that there exist a convex subset $\mathcal{X}_0$ of $\mathcal{X}$ and a compact convex subset $\mathcal{Y}_0$ of $\mathcal{Y}$ such that

1) for every for every $x \in \mathcal{X}_0$, an optimizer $y^*(x) \in \arg\max_{y\in\mathcal{Y}} h(x,y)$ belongs to $\mathcal{Y}_0$; and
2) for every $y \in \mathcal{Y}_0$, an optimizer $x^*(y) \in \arg\min_{x\in\mathcal{X}} h(x,y)$ belongs to $\mathcal{X}_0$.

Then,

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} h(x,y) = \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} h(x,y).$$

Proof: Note that since for x in $\mathcal{X}_0$, the y that maximizes $h(x,y)$ over $\mathcal{Y}$ is in $\mathcal{Y}_0$, we get

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} h(x,y) \leq \min_{x\in\mathcal{X}_0}\max_{y\in\mathcal{Y}} h(x,y) = \min_{x\in\mathcal{X}_0}\max_{y\in\mathcal{Y}_0} h(x,y).$$

Further, by Sion's minmax theorem (Theorem VII.1), the right-side equals $\max_{y\in\mathcal{Y}_0}\min_{x\in\mathcal{X}_0} h(x,y)$. But by our second assumption, the restriction $x \in \mathcal{X}_0$ can be dropped, and we have

$$\max_{y\in\mathcal{Y}_0}\min_{x\in\mathcal{X}_0} h(x,y) = \max_{y\in\mathcal{Y}_0}\min_{x\in\mathcal{X}} h(x,y) \leq \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} h(x,y).$$

Thus, we have shown $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} h(x,y) \leq \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} h(x,y)$, which completes the proof since the inequality in the other direction holds as well. □

For our minmax cost, we will verify that both the conditions of the lemma above hold under the assumption $H(X)+\log(1+1/\sqrt{2}) < L_{\mathtt{th}}$. Indeed, first note that for any fixed $\ell \in \Lambda$ with $\mathbb{E}[L] \leq H(X) + \log(1+1/\sqrt{2})$, the maximizer z of $f(\ell,z)$ given by $\sqrt{\mathbb{E}[L^2]}/(L_{\mathtt{th}}-\mathbb{E}[L])$ satisfies

$$\begin{aligned}&\frac{\sqrt{\mathbb{E}[L^2]}}{L_{\mathtt{th}}-\mathbb{E}[L]}\\ &\leq \sqrt{\frac{1}{\min_x P(x)}}\cdot\frac{\mathbb{E}[L]}{L_{\mathtt{th}}-\mathbb{E}[L]}\\ &\leq \sqrt{\frac{1}{\min_x P(x)}}\cdot\frac{H(X)+\log(1+1/\sqrt{2})}{L_{\mathtt{th}}-H(X)-\log(1+1/\sqrt{2})}.\end{aligned}$$

Denote the right-side above by K and $L'_{\mathtt{th}} = H(X)+\log(1+1/\sqrt{2})$. Therefore, with the set $\{\ell\in\Lambda, \mathbb{E}[L] \leq L'_{\mathtt{th}}\}$ in the role of $\mathcal{X}_0$ in Lemma VII.4, the set $[0,K]$ can play the role of $\mathcal{Y}_0$.

To apply Lemma VII.4, we require two conditions to hold: first, that $f(l,z)$ is a convex function of ℓ for every fixed z and a concave function of z for every fixed ℓ, second, that for every $z \in [0,K]$, the minimizing ℓ satisfies $\mathbb{E}[L] \leq L'_{\mathtt{th}}$.The first easily follows from (36). The proof of this fact is exactly the same as Lemma VII.2. However, while the second condition can be shown to be true, the proof of this fact is almost the same as the proof of our theorem. For simplicity of presentation, we instead present an alternative proof of the theorem that uses a slight extension of the lemma above. Note that from our foregoing discussion and following the proof of the lemma, we already have obtained

$$\Delta^*(P) \leq \max_{z\in[0,K]}\min_{\ell\in\Lambda,\mathbb{E}[L]\leq L'_{\mathtt{th}}} f(\ell,z).$$

By using Corollary III.2 and using Sion's minmax theorem once again, we get

$$\begin{aligned}&\Delta^*(P)\\ &\leq \max_{z\in[0,K]}\max_{Q\ll P}\min_{\ell\in\Lambda,\mathbb{E}[L]\leq L'_{\mathtt{th}}}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) - \frac{z^2}{2}L_{\mathtt{th}},\end{aligned}$$

where

$$g_{z,Q,P}(x) := \left(1+\frac{z^2}{2}\right)P(x) + z\sqrt{Q(x)P(x)}.$$

In the preceding argument, we can use Sion's minmax theorem as the following two conditions hold. First, for every fixed

$z \geq 0$, the function $\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) - \frac{z^2}{2}L_{\tt th}$ is concave in Q for a fixed $\ell \in \Lambda$ and convex in ℓ for a fixed $Q \ll P$. Second, the sets $\{Q : Q \ll P\}$ and $\{\ell \in \Lambda : \mathbb{E}[L] \leq L'_{\tt th}\}$ are compact and convex. Proof of the first is exactly the same as that of VII.3. Second is true as we have restricted to a finite alphabet $\mathcal{X}$. Thus, we can proceed as in the proof of the lemma, but we need to show now that for every $z \in [0, K]$ and $Q \ll P$, the optimal $\ell^*(z, Q)$ satisfies $\mathbb{E}[L^*] \leq L'_{\tt th}$. Indeed, consider the following optimization problem for a fixed z, Q:

$$\begin{aligned}\min_{\ell\in\Lambda}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) \\ &= \left(\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')\right)\min_{\ell\in\Lambda}\sum_{x\in\mathcal{X}}\frac{g_{z,Q,P}(x)}{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}\ell(x).\end{aligned}$$

Since $\frac{g_{z,Q,P}(x)}{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}$ are nonnegative and add to 1, in the optimization problem above, we are minimizing the expected prefix free lengths for a finite alphabet for a particular distribution. Thus, by Shannon's Source Coding Theorem, the optimal $\ell^*_{z,Q}$ is given by

$$\ell^*_{z,Q}(x) := \log\frac{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}{g_{z,Q,P}(x)};$$

in fact, this optimizer is unique. But then for every x in $\mathcal{X}$,

$$\begin{aligned}&\ell^*_{z,Q}(x) \\ &= \log\frac{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}{g_{z,Q,P(x)}} \\ &= \log\frac{\sum_{x'\in\mathcal{X}}\left(1+\frac{z^2}{2}\right)P(x') + \sum_{x\in\mathcal{X}} z\sqrt{Q(x')P(x')}}{\left(1+\frac{z^2}{2}\right)P(x) + z\sqrt{Q(x)P(x)}} \\ &\leq \log\frac{1}{P(x)} \\ &\quad + \log\left(\frac{\left(1+\frac{z^2}{2}\right)}{\left(1+\frac{z^2}{2}\right)+z\sqrt{\frac{Q(x)}{P(x)}}} + \frac{z}{\left(1+\frac{z^2}{2}\right)+z\sqrt{\frac{Q(x)}{P(x)}}}\right) \\ &\leq \log\frac{1}{P(x)} + \log\left(\frac{\left(1+\frac{z^2}{2}\right)}{\left(1+\frac{z^2}{2}\right)} + \frac{z}{\left(1+\frac{z^2}{2}\right)}\right) \\ &\leq \log\frac{1}{P(x)} + \log\left(1+\frac{1}{\sqrt{2}}\right),\end{aligned}$$

where the first inequality is by the Cauchy-Schwarz inequality, the second inequality follows upon noting that $\frac{Q(x)}{P(x)}$ is nonnegative, and the last inequality follows from the fact that $z^2/2+1 \geq \sqrt{2}z$ (which holds with equality at $z = \sqrt{2}$). Thus as a consequence of this inequality the expected code length of such a code is upper bounded as follows,

$$\mathbb{E}\left[L^*_{z,Q}\right] \leq H(x) + \log\left(1+\frac{1}{\sqrt{2}}\right), \tag{37}$$

which in the manner of Lemma VII.4 gives

$$\Delta^*(P) = \max_{z\geq 0}\max_{Q\ll P}\min_{\ell\in\Lambda,\mathbb{E}[L]\leq L_{\tt th}}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x) - \frac{z^2}{2}L_{\tt th}.$$

Finally, it remains to establish that $\ell^*_{z^*,Q^*}$ is the unique minmax optimal solution. This can be shown in exactly the same manner as it was shown for Theorem IV.1 in the previous section; we skip the details. □

Appendix

A Saddle-Point Lemma

The following simple result is needed to establish the minmax optimality of our scheme. The first part of the result claims that any pair of minmax optimal x and maxmin optimal y forms a saddle point, a well-known fact. The second part claims that if the minimizer for the maxmin optimal y is unique, then it must also be minmax optimal and thereby constitute a saddle-point with y.

Lemma .1: Consider the minmax problem $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} h(x, y)$, and assume that

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} h(x, y) = \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} h(x, y).$$

Then, for every pair (x^*, y^*) such that $x^* \in \arg\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} h(x, y)$ and $y^* \in \arg\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} h(x, y)$ constitutes a saddle-point. Furthermore, if the minimizer $x^o(y^*)$ of $\min_{x\in\mathcal{X}} h(x, y^*)$ is unique, then $x^* = x^o(y^*)$ is the unique minmax optimal solution.

Proof: Since minmax and maxmin costs are assumed to be equal, by the definition of x^* and y^*, we have

$$\begin{aligned}h(x, y^*) &\geq \max_{y'\in\mathcal{Y}}\min_{x'\in\mathcal{X}} h(x', y') \\ &= \min_{x'\in\mathcal{X}}\max_{y'\in\mathcal{Y}} h(x', y') \geq h(x^*, y),\end{aligned} \tag{38}$$

for all x in $\mathcal{X}$ and y in $\mathcal{Y}$. Upon substituting x^* for x and y^* for y, we get that x^* is a minimizer of $h(x, y^*)$ and y^* a maximizer of $h(x^*, y)$. Therefore, (x^*, y^*) forms a saddle-point and $h(x^*, y^*) = \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} h(x, y)$.

Turning now to the second part, suppose that x', too, is minmax optimal. Then, using (38) with $x = x'$ and $y = y^*$, we get that x' must be a minimizer of $h(x, y^*)$ as well. But since this minimizer is unique, x' must coincide with x^o. □

Proof of Lemma IV.2

Denoting

$$c_P(z, Q) := \sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\log\frac{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}{g_{z,Q,P}(x)},$$

we begin by observing the concavity of $c_P(z, Q)$. Recall the notations $\mathcal{G} = \{z \geq 0, Q \in \mathbb{R}^{|\mathcal{X}|} : g_{z,Q,P}(x) \geq 0 \quad \forall x \in \mathcal{X}\}$ and $g_{z,Q,P}(x) = (1 - z^2/2)P(x) + z\sqrt{Q(x)P(x)}$.

Lemma .2: The function $c_P(z, Q)$ is concave in Q for each fixed z and is concave in z for each fixed Q, over the set $\mathcal{G}$.

Proof: For the first part, (35) yields that for every $(z, Q) \in \mathcal{G}$,

$$\begin{aligned}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\log\frac{\sum_{x'\in\mathcal{X}} g_{z,Q,P}(x')}{g_{z,Q,P}(x)} \\ = \min_{\ell\in\Lambda}\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x).\end{aligned}$$

Also, for every fixed z, the function $g_{z,Q,P}(x)$ is concave in Q, and thereby $\sum_{x\in\mathcal{X}} g_{z,Q,P}(x)\ell(x)$, is concave in Q.

Thus, since the minimum of concave functions is concave, $c_P(z,Q)$ is concave in Q for a fixed z. Similarly, we can show concavity in z for a fixed Q since $g_{z,Q,P}(x)$ is concave in z, too, for every fixed Q. □

We now complete the proof of Lemma IV.2. We will show that for any (z,Q) which is feasible for optimization problem (9), we can find a feasible (z,Q') with Q' satisfying (11), and

$$c_P(z,Q) \leq c_P(z,Q').$$

Indeed, consider $Q'(x) := Q(A_i)/|A_i|$ for all $x \in \mathcal{X}$. The remainder of the proof is divided into two parts, the first proving the feasibility of Q' and the second proving $c_P(z,Q) \leq c_P(z,Q')$.

a) Feasibility of (z,Q'): From the feasibility of (z,Q), for all symbols x in A_i and for all i in $[M_P]$, $g_{z,Q,P}(x) \geq 0$, whereby

$$\begin{aligned}\sum_{x\in A_i} g_{z,Q,P}(x) &= \sum_{x\in A_i}\left(1-\frac{z^2}{2}\right)P(x) \\ &\qquad + z\sum_{x\in A_i}\sqrt{Q(x)P(x)} \\ &= \left(1-\frac{z^2}{2}\right)P(A_i) + z\sum_{x\in A_i}\sqrt{Q(x)P(x)} \\ &\geq \left(1-\frac{z^2}{2}\right)P(A_i) + z\sqrt{Q'(A_i)P(A_i)} \\ &= |A_i| g_{z,Q',P}(x) \\ &\geq 0,\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the positivity of z, and the assumption that $P(x) = P(A_i)/|A_i|$ for every x in A_i, and the final identity uses definition of Q'. This proves the feasibility of (z,Q') for the optimization problem (9).

b) Proof of optimality: Denoting by $\Pi(A_1)$ the set of all permutations of the elements of A_1, let Q^π be the distribution given by

$$Q^\pi(x) = \begin{cases} Q(\pi(x)), & \forall x \in A_1 \\ Q(x), & \text{otherwise.}\end{cases}$$

Then, the distribution $\overline{Q} = (1/|\Pi(A_1)|)\cdot\sum_{\pi\in\Pi(A_1)} Q^\pi$ satisfies

$$\overline{Q}(x) = \begin{cases} \frac{1}{|A_1|}\cdot Q(A_1), & \forall x \in A_1 \\ Q(x), & \text{otherwise.}\end{cases}$$

Since by Lemma .2 $c_P(z,Q)$ is concave in Q for every fixed z, we get

$$c_P(z,\overline{Q}) \geq \frac{1}{|\Pi(A_1)|}\cdot\sum_{\pi\in\Pi(A_1)} c_P(z,Q^\pi).$$

Furthermore, note that $g_{z,Q^\pi,P}(x) = g_{z,Q,P}(\pi(x))$ since $P(x) = P(A_1)/|A_1|$ for every x in A_1, and thereby $c_P(z,Q^\pi) = c_P(z,Q)$ for every $\pi \in \Pi(A_1)$. Therefore, combining the observations above, we obtain $c_P(z,\overline{Q}) \geq c_P(z,Q)$.

Repeating this argument by iteratively using permutations of A_i for $i \geq 2$, we obtain the required inequality

$$c_P(z,Q') \geq c_P(z,Q).$$

□

Acknowledgment

The authors thank R. Sundaresan for pointing out connection to Campbell's work and to the reference [9] which led to Theorem III.1, the main technical contribution of this work. Also, they thank R. Yates for bringing to their notice references [25], [26], and illuminating discussions on connection between their formulation and that in [25].

References

[1] B. T. Bacinoglu and E. Uysal-Biyikoglu, "Scheduling status updates to minimize age of information with an energy harvesting sensor," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1122–1126.

[2] M. B. Baer, "Source coding for quasiarithmetic penalties," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4380–4393, Oct. 2006.

[3] S. Bhambay, S. Poojary, and P. Parag, "Differential encoding for real-time status updates," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[4] L. L. Campbell, "A coding theorem and Rényi's entropy," *Inf. Control*, vol. 8, no. 4, pp. 423–429, 1965.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[6] A. S. Drud, "CONOPT—A large-scale GRG code," *ORSA J. Comput.*, vol. 6, no. 2, pp. 207–216, May 1994.

[7] R. Fourer, D. M. Gay, and B. Kernighan, *AMPL*, vol. 117. Danvers, MA, USA: Boyd & Fraser, 1993.

[8] P. E. Gill, W. Murray, and M. A. Saunders, "SNOPT: An SQP algorithm for large-scale constrained optimization," *SIAM J. Optim.*, vol. 12, no. 4, pp. 979–1006, Jan. 2002.

[9] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 70–78, Jan. 2011.

[10] Q. He, D. Yuan, and A. Ephremides, "Optimal link scheduling for age minimization in wireless systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5381–5394, Jul. 2018.

[11] P. A. Humblet, "Source coding for communication concentrators," Ph.D. dissertation, MIT, Cambridge, MA, USA, 1978.

[12] S. Kaul, R. Yates, and M. Gruteser, "On piggybacking in vehicular networks," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2011, pp. 1–5.

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

[14] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Found. Trends Netw.*, vol. 12, no. 3, pp. 162–259, 2017.

[15] L. L. Larmore, "Minimum delay codes," *SIAM J. Comput.*, vol. 18, no. 1, pp. 82–94, Feb. 1989.

[16] A. Mahajan and D. Teneketzis, "Optimal design of sequential real-time communication systems," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5317–5338, Nov. 2009.

[17] P. Mayekar, P. Parag, and H. Tyagi, "Optimal lossless source codes for timely updates," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1246–1250.

[18] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, vol. 1. Berkeley, CA, USA: Univ. of California Press, 1961, pp. 547–561.

[19] S. M. Ross, *Stochastic Processes*. New York, NY, USA: Wiley, 1996.

[20] M. Sion, "On general minimax theorems," *Pacific J. Math.*, vol. 8, no. 1, pp. 171–176, Mar. 1958.

[21] 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.

[22] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Trans. Autom. Control*, vol. 49, no. 7, pp. 1056–1068, Jul. 2004.

[23] R. D. Yates, “Lazy is timely: Status updates by an energy harvesting source,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 3008–3012.

[24] J. Zhong, R. D. Yates, and E. Soljanin, “Timely lossless source coding for randomly arriving symbols,” 2018, *arXiv:1810.01533*. [Online]. Available: http://arxiv.org/abs/1810.01533

[25] J. Zhong and R. D. Yates, “Timeliness in lossless block coding,” in *Proc. Data Compress. Conf. (DCC)*, Mar. 2016, pp. 339–348.

[26] J. Zhong, R. D. Yates, and E. Soljanin, “Backlog-adaptive compression: Age of information,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 566–570.

Prathamesh Mayekar (Student Member, IEEE) received the B.E. degree in electronics and telecommunication engineering from Mumbai University, India, in 2013, and the M.Tech. degree in industrial engineering and operation research from the Indian Institute of Technology Bombay, India, in 2015. He is currently pursuing the Ph.D. degree with the Department of Electrical Communication Engineering, Indian Institute of Science, India. Broadly, his research interests lie at the intersection of information theory and optimization. He was a recipient of the 2018 Jack Keil Wolf ISIT Student Paper Award and a Wipro Ph.D. fellowship.

Parimal Parag (Member, IEEE) received the B.Tech. and M.Tech. degrees from IIT Madras in 2004 and the Ph.D. degree from the Texas A&M University in 2011, all in electrical engineering. He was a Senior System Engineer (R&D) with Assia Inc., Redwood City, from 2011 to 2014. He is currently an Assistant Professor with the ECE Department, Indian Institute of Science. He was a coauthor of the 2018 IEEE International Symposium on Information Theory student best paper. His research interests lie in the design and analysis of large-scale distributed systems. He was a recipient of the 2017 Early Career Award from the Science and Engineering Research Board.

Himanshu Tyagi (Senior Member, IEEE) received the B.Tech. degree in electrical engineering and the M.Tech. degree in communication and information technology from the Indian Institute of Technology, Delhi, India, in 2007, and the Ph.D. degree from The University of Maryland, College Park, in 2013. From 2013 to 2014, he was a Post-Doctoral Researcher with the Information Theory and Applications (ITA) Center, University of California, San Diego. Since January 2015, he has been a faculty member with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore. His research interests broadly lie in information theory and its application in cryptography, statistics, machine learning, and computer science. Also, he is interested in communication and automation for city-scale systems.