Real-Time Variable-to-Fixed Lossless Source Coding of Randomly Arriving Symbols

Uri Abend and Anatoly Khina

Abstract—We address the recently suggested problem of causal lossless coding of a randomly arriving source samples. We construct variable-to-fixed coding schemes and show that they outperform the previously considered fixed-to-variable schemes when traffic is high both in terms of delay and Age of Information by appealing to tools from queueing theory. We supplement our theoretical bounds with numerical simulations.

Index Terms—Age of Information, delay, lossless source coding, queueing theory.

I. INTRODUCTION

Real-time communications has become a topic of growing interest in recent years due to rising demand in applications such as vehicular communications [1], telemedicine [2] and satellite control [3]. Under this regime, the objective is to convey messages in a timely fashion, in contrast to classic communications where relatively long delays may be tolerated.

Two figures of merit have been suggested to measure the performance of real-time communication systems: *Age of Information* (AoI) [4]–[6] and *delay* [4], [7]. The AoI of a system quantifies the freshness of the data at the receiver by assigning an "age" to the most recent update of the system and measuring the time difference between updates. Delay, on the other hand, quantifies the timeliness of a system by measuring the elapsed end-to-end time difference from a symbol arrival to the encoder to its decoding at the decoder. These two quantities are similar in nature and indeed in some schemes [8], [9], they can be simultaneously minimized.

In this work, we concentrate on the recently suggested setting of lossless transmission of causal source samples with random arrival times [9], [10]. This setting is simple to formulate and can apply to many different settings, e.g., a multi-user scenario where data is transmitted through a single channel.

So far, the majority of works dealing with real-time lossless source coding, have focused on one-to-variable encoding, following the work of Larmore [7]. Additionally, for deterministic (periodic) arrivals, some work has been done on the more general case of block-to-variable coding [8]. These coding schemes, while effective in low-traffic scenarios, exhibit large delays as the traffic increases, and the waiting time becomes
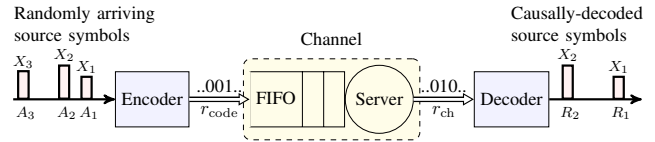
Fig. 1. Block diagram of the system model.

dominant. However, in the case of deterministic arrivals, a variable-to-variable scheme in which the block size is set on the fly according to the number of symbols waiting to be coded, can be used to facilitate lower latency as was suggested in [11].

To alleviate this problem, we propose in this work a variable-to-fixed coding scheme. The motivation for variable-to-fixed codes stems from two main reasons: First, a work in queueing theory by Rogozin [12] states that, under some restrictions, the waiting time of a queue is minimized by a deterministic service time. While this does not prove that variable-to-fixed codes are optimal, it suggests that they might perform well when the waiting time is the dominant component of the AoI and the delay, i.e., in high-traffic scenarios. Secondly, fixed-length codes synergize well with error correcting codes when real-time transmission is carried over noisy channels.

The rest of the paper is organized as follows. Sec. II presents the system model and objectives. Sec. III provides necessary background from queueing theory. In Sec. IV, we provide an analysis of the delay for variable-to-fixed codes, with Sec. IV-A dedicated to an analysis of the waiting time. Sec. V contains some simulation results. We conclude with a summary of the work and suggestions for future work in Sec. VI.

II. PROBLEM SETUP

We formulate here the real-time source coding setup that will be treated in this work, depicted in Fig. 1. Operation of the system is set to begin at time 0.

Source: At each (positive) time step, the source generates a new sample with known probability $q \in (0, 1)$. The generation process samples across time are assumed independent and identically distributed (i.i.d.). We define the n-th arrival time A_n as the generation time of the n-th symbol. We further define the n-th time difference between arrivals by

$$D_n = A_n - A_{n-1},$$

with A_0 set to 0. Clearly, $\{D_n\}_{n=1}^{\infty}$ are i.i.d. according to a geometric distribution with parameter q.

The sample values, $\{X_n\}_{n=1}^{\infty}$, are i.i.d. and are drawn from a finite alphabet $\mathcal{X}$ according to a known probability mass function (p.m.f.) p_X.

Encoder: The memoryless encoder accumulates $B_i \in \mathbb{N}$ samples and maps them into $L_i \in \mathbb{N}$ bits using a prefix-free one-to-one mapping (i being the block index):

$$\mathscr{E} : \mathcal{I} \subseteq \mathcal{X}^* \to \{0,1\}^*,$$

such that $\mathcal{X}^* \subseteq \times_{i=1}^{\infty} \mathcal{I}$, i.e., any stream of source symbols is mapped to a stream of code bits. The encoder then sends the L_i-length codeword over the channel. Note that both B_i and L_i are finite with probability 1 and that $\{B_i\}_{i=1}^{\infty}$ and $\{L_i\}_{i=1}^{\infty}$ have i.i.d. samples. The average rate of the memoryless encoder (average number of coded bits per symbol) is given by

$$r_{\text{code}} = \frac{\mathbb{E}[L]}{\mathbb{E}[B]}, \tag{1}$$

where B and L are general samples of their respective processes.

It is customary to divide the class of lossless encoders into three families:

- *Variable-to-fixed.* This is the encoding scheme used in this work. The output of this encoder is of fixed length $L_i \equiv \ell$. The best known representative of this family is the Tunstall code [13], featuring a memoryless encoder which minimizes $\mathbb{E}[B]$ given ℓ; this code will be discussed and used in later sections.
- *Fixed-to-variable.* This family of codes was considered previously for the problem at hand in [8], [9], [11]. The input length of this encoder is of fixed length $B_i \equiv b$. The best known representative of this family is the Huffman code [14, Ch. 5.6], featuring a memoryless encoder which minimizes $\mathbb{E}[L]$ given b.
- *Variable-to-variable.* This is the most general family in which both the input and the output lengths are not fixed and depend on the encoded sequence. The most famous code in this family is the arithmetic code [14, Ch 13.3]. Note that an encoder employing this code does not fall under the definitions given here.

Channel: The channel is modeled by a first-in, first-out (FIFO) queue, which receives codewords from the encoder and outputs them to the decoder at a fixed known rate $r_{\text{ch}} \in \mathbb{R}$. From a queueing perspective, we can view the codewords as clients entering the queue. This will be analyzed in Sec. IV.

Remark 1. In general, the rate of emission from the FIFO may be any positive real number. This is the case for asynchronous FIFOs which have different input and output clock domains.

Decoder: Once L_i bits are received from the FIFO, they are mapped back to B_i symbols, applying the inverse of $\mathscr{E}$

$$\mathscr{E}^{-1} : \{0,1\}^* \to \mathcal{X}^*.$$

Because the channel is noiseless and the encoder is prefix-free and lossless, we are guaranteed perfect reconstruction.

The time of decoding of the n-th symbol is denoted R_n. Note that the decoding time is shared between B_i symbols. We shall denote the shared decoding time by R_i.

Remark 2. Throughout this work we assume that the decoder is privy to the FIFO's status, e.g., via a side-channel or a special symbol reserved for this scenario, as is common in FIFO architecture.

Delay: The delay of the n-th symbol, $\Delta_n = R_n - A_n$, is defined as the difference between its decoding time and arrival time. The mean delay of the system can then be defined by averaging over all symbol delays

$$\bar{\Delta} = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \Delta_k. \tag{2}$$

AoI: The (peak) AoI of block i, $\Gamma_i = R_i - A_{i-1}$, is defined as the time difference between said decoding time of block i and the arrival of the most recent symbol in block $i-1$. [1] The mean peak AoI is then given by

$$\bar{\Gamma} = \lim_{i \to \infty} \frac{1}{i} \sum_{k=1}^{i} \Gamma_k. \tag{3}$$

We refer the reader to [9] for a graphical illustration of AoI.

Objectives: The objective of this encoding scheme is to minimize either the delay or the AoI. Note that for the case of fixed-to-variable encoding, the two quantities are simultaneously minimized.

III. BACKGROUND: RESULTS FROM QUEUEING THEORY

We now introduce some notations and known results from queueing theory to be used later in our analysis. We define the inter-arrival time T_i as the time passing between the arrival of client $(i-1)$ and client i, and the service time S_i—as the amount of time client i spends inside the queue.

Remark 3. The definitions given here hold for any distributions of T_i and S_i with finite means [15]. For example, T_i may receive only integer values and S_i may receive values that are multiples of $1/r_{\text{ch}}$ as described in Sec. II.

We are interested in results for the GI/GI/1 queue [16], i.e., the setting of a single server, and independent i.i.d. processes $\{T_i\}$ and $\{S_i\}$ with known (possibly different) distributions. General samples of theses processes are denoted by T and S.

We now present some known results for the GI/GI/1 queue waiting time, W_i, which is defined as the amount of time that client i waits in order to enter the queue.

To that end, first define U_i as the difference between the service and arrival times of client i

$$U_i \triangleq S_i - T_i. \tag{4}$$

Note that U_i is also an i.i.d. process; we denote a general sample from the process as U. The waiting time of the i-th client, W_i, is given by

$$W_i = \max(0, W_{i-1} + U_{i-1}), \tag{5}$$

---

[1]Some works, e.g., [8], [11], study the average AoI. Both peak and average AoI exhibit similar behavior and we therefore treat only the former in the interest of space.

with W_1 set to 0. A queue is said to be *stable* if the waiting time series $\{W_i\}_{i=1}^{\infty}$ is bounded with probability 1. It was shown in [16] that a queue is stable iff[2]

$$\mathbb{E}\left[U\right] < 0. \tag{6}$$

Moreover, if the queue is stable, then the waiting time series tends to a random variable, W, that is bounded with probability 1. Finding the mean value of W requires solving integral equations that generally have no analytic solutions [16]. To overcome this, several upper bounds for $\mathbb{E}\left[W\right]$ have been derived in the literature. We now present two of them, which will be used throughout the paper. The first bound, which will be refererd to as the *low-moment bound*, is given by [16]

$$\mathbb{E}\left[W\right] \leq \frac{\text{Var}\left(U\right)}{-\mathbb{E}\left[U\right]}. \tag{7}$$

To derive the second bound, we start by defining the moment generating function (m.g.f.) of U:

$$\phi_U(\theta) = \mathbb{E}\left[e^{\theta U}\right]. \tag{8}$$

Then, the *m.g.f. bound* is given by

$$\mathbb{E}\left[W\right] \leq \frac{1}{\nu}, \tag{9}$$

where $\nu \triangleq \sup\left\{\theta > 0 | \phi_U(\theta) < 1\right\}$.

IV. DELAY AND AOI OF VARIABLE-TO-FIXED CODES

As was stated in Sec. II, a memoryless variable-to-fixed binary encoder is a function that encodes a block of symbols of varying length B which depends on the symbol values, to a binary codeword of fixed length ℓ. The coding rate (1) specializes, therefore, to

$$r_{\text{code}} = \frac{\ell}{\mathbb{E}\left[B\right]}. \tag{10}$$

From a queuing perspective, we can think of the noiseless channel as a queue with the encoder inserting clients with a fixed service time

$$S_i \equiv \frac{\ell}{r_{\text{ch}}}, \tag{11}$$

and an inter-arrival time of

$$T_i = \sum_{n=1}^{B_i} D_n. \tag{12}$$

Note that $\{D_n\}_{n=1}^{\infty}, \{B_i\}_{i=1}^{\infty}$ are independent processes. This will be used later in our analysis.

By substituting (4), (10), (11), and (12) in (6) we arrive at the following stability condition: The queue is stable iff

$$r_{\text{code}} < \frac{r_{\text{ch}}}{q}, \tag{13}$$

where $\mathbb{E}\left[T\right] = \mathbb{E}\left[B\right]/q$ by Wald's identity [17].

Remark 4. This result is similar to the one-to-variable scheme ($b = 1$ and random L_i) of [9], where the code rate (1), which

specializes to $r_{\text{code}} = \mathbb{E}\left[L\right]$ for that scheme, was to be kept smaller than r_{ch}/q to maintain stability.

Furthermore, because the encoder is lossless, the code rate satisfies $H(X) \leq r_{\text{code}}$, where $H(X)$ denotes the source entropy [14, Ch. 5.3]. Consequently, r_{ch}/q of a stabilizable system must be bounded from below by the source entropy

$$H(X) < \frac{r_{\text{ch}}}{q}, \tag{14}$$

which leads to the following lemma.

Lemma 1. *The AoI and delay are bounded iff* (14) *holds.*

Proof. The necessity of (14) follows from (13) and the aforementioned bound $H(X) \leq r_{\text{code}}$. To prove the sufficiency, consider a sequence of Tunstall codes with increasing block-length ℓ. The rate of this sequence is known to converge to the entropy of the source [13]:

$$\lim_{\ell \to \infty} r_{\text{code}} = H(X).$$

Therefore, (13) is satisfied for a Tunstall code with a sufficiently large $\ell \in \mathbb{N}$. $\qquad\square$

We now return to the system objectives described in Sec. II. To that end, we describe the delay and the AoI, each, as the sum of three known quantities, which are defined next.

Tarry time: The tarry time of the ℓ-th symbol, J_ℓ, is defined as the amount of time it waits to be coded. To inject bits into the channel, the encoder has to accumulate B symbols. Consequently, $J_B = 0$, whereas for other symbols:

$$J_\ell = \sum_{n=\ell+1}^{B} D_n, \qquad \ell \in \{1, \dots, B-1\}. \tag{15}$$

The mean tarry time, $\bar{J}$, is equal to

$$\bar{J} \triangleq \limsup_{k \to \infty} \frac{\sum_{n=1}^{k} J_n}{k} \tag{16a}$$

$$= \lim_{m \to \infty} \frac{\frac{1}{m}\sum_{i=1}^{m}\sum_{\ell=1}^{B_i} J_\ell}{\frac{1}{m}\sum_{i=1}^{m} B_i} \tag{16b}$$

$$= \frac{\mathbb{E}\left[\sum_{\ell=1}^{B} J_\ell\right]}{\mathbb{E}\left[B\right]} \tag{16c}$$

$$= \frac{\mathbb{E}\left[\sum_{n=2}^{B}(n-1)D_n\right]}{\mathbb{E}\left[B\right]} \tag{16d}$$

$$= \frac{\mathbb{E}\left[B^2\right] - \mathbb{E}\left[B\right]}{2q\mathbb{E}\left[B\right]}, \tag{16e}$$

where (16b) is due to the encoder construction (recall Sec. II), (16c) follows from ergodicity by recalling that $\{B_i\}_{i=1}^{\infty}$ and $\{D_n\}_{n=1}^{\infty}$ are i.i.d. and independent of each other, (16d) follows from (15) and exchange of order of summation, and (16e) follows from Wald's identity and the sum of an arithmetic series.

Inter-arrival time: The inter-arrival time is given by (12).
Service time: The service time is fixed and is given by (11).

---

[2] [16] showed that the queue is stable if (6) holds. The other direction is trivial by noting that $W_i \geq \sum_{k=0}^{i-1} U_k$ by (5).

Waiting time: The waiting time is defined as the time that passes between a codeword entering the FIFO and the beginning of its service as described in Sec. III. Bounds on the mean waiting time are provided Sec. IV-A.

The mean delay (2) of the system can now be expressed as the sum of the means of three delay elements

$$\bar{\Delta} = \bar{J} + \mathbb{E}[S] + \mathbb{E}[W]. \tag{17}$$

The mean peak AoI (3) can be expressed similarly as

$$\bar{\Gamma} = \mathbb{E}[T] + \mathbb{E}[S] + \mathbb{E}[W].$$

A. Analysis of the Mean Waiting Time

As was stated in Sec. III, obtaining an analytic expression for the waiting time is difficult, in general. Instead, we evaluate the two upper bounds of Sec. III for the suggested scheme.

To derive low-moment bound (7), we first derive an expression for U by substituting (11) and (12) in (4):

$$U = \sum_{n=1}^{B} D_n - \frac{\ell}{r_{\mathrm{ch}}}.$$

Thus, the low-moment bound (7) specializes to

$$\mathbb{E}[W] \leq \frac{\mathrm{Var}(B) + (1-q)\mathbb{E}[B]}{q(\mathbb{E}[B] - q\frac{\ell}{r_{\mathrm{ch}}})},$$

with the variance and mean of T given by Wald's identity.

Next, we derive the m.g.f. bound (9). To that end, we start by deriving the m.g.f. of U (8):

$$\begin{aligned} \phi_U(\theta) &= \mathbb{E}\left[e^{\theta\left(\frac{\ell}{r_{\mathrm{ch}}} - \sum_{n=1}^{B} D_n\right)}\right] \\ &= e^{\frac{\ell}{r_{\mathrm{ch}}}\theta}\mathbb{E}\left[\left(\frac{qe^{-\theta}}{1-(1-q)e^{-\theta}}\right)^B\right], \end{aligned} \tag{18}$$

where the second equality follows from the law of total expectation, the fact that $\{D_n\}_{n=1}^{\infty}$ are i.i.d. according to a geometric distribution with parameter q and are independent of B, and by substituting the m.g.f. of D_n—$\phi_D(\theta) \triangleq \mathbb{E}[e^{\theta D_n}] = \frac{qe^{\theta}}{1-(1-q)e^{\theta}}$. Because $\phi_U(\theta)$ is a continuous function, the bound can be found by solving the equation

$$\phi_U(\theta) = 1,$$

which, by (18), can be rewritten as

$$\sum_{b=1}^{b_{\max}}\left(\frac{qe^{-\theta}}{1-(1-q)e^{-\theta}}\right)^b P(B=b) = e^{-\frac{\ell}{r_{\mathrm{ch}}}\theta},$$

or, equivalently, as

$$\left(1-(1-q)e^{-\theta}\right)^{b_{\max}}e^{-\theta\frac{\ell}{r_{\mathrm{ch}}}}$$
$$- \sum_{b=1}^{b_{\max}} P(B=b)\left(1-(1-q)e^{-\theta}\right)^{b_{\max}-b}qe^{-\theta b} = 0,$$

with $b_{\max}$ being the maximal value that B can attain with a non-zero probability. The solutions for this equation can be found using an appropriate root finding algorithm, e.g.,

the Newton–Raphson algorithm [18].[3] Once the solutions are found, the bound is derived by taking the minimum of the real solutions, and substituting it in (9).[4]

V. SIMULATION STUDY

In this section, we simulate variable-to-fixed Tunstall codes and compare their performance to those of the optimized fixed-to-variable codes of [8].

To that end, we generated several simulations of 2×10^6 i.i.d. Bernoulli distributed source samples of $\{X_n | n = 1, \ldots, 2 \times 10^6\}$, each simulation with a different parameter $p \triangleq \mathrm{Pr}(X_i = 1) \in (0, 1/2]$, corresponding to a different source entropy. Note that an increase in the entropy corresponds to an increase in traffic since a higher (average) rate is required to describe a source sample. The probability of arrival in all simulations was set to $q = 1/2$.

To encode the source, we used an off-the-shelf Tunstall code for each parameter p, which minimizes the code rate r_{code} (10) for a given ℓ; note that the queue is stabilizable with some variable-to-fixed code with a given ℓ iff it is stabilizable with a Tunstall code with this ℓ for the given distribution. The output length was set to $\ell = 4$, which is the minimal length for the chosen region of p that satisfies (13), and therefore minimizes the service time.

In order to test the performance of the waiting time bounds, we compare between the empiric results and the analytic expression of the delay given in (17), with the waiting time bounded as described in Sec. IV-A.

We compare our results to a fixed-to-variable scheme adapted from a work by Zhong et al. [8]. To that end, we derive a bound for the waiting time using (7)

$$\mathbb{E}[W] \leq \frac{\mathrm{Var}(L)/r_{\mathrm{ch}}^2 + (1-q)b/q^2}{b/q - \mathbb{E}[L]/r_{\mathrm{ch}}}.$$

The mean delay is then bounded by

$$\bar{\Delta} \leq \frac{\mathrm{Var}(L)/r_{\mathrm{ch}}^2 + (1-q)b/q^2}{b/q - \mathbb{E}[L]/r_{\mathrm{ch}}} + \frac{\mathbb{E}[L]}{r_{\mathrm{ch}}} + \frac{b-1}{2q}.$$

The mean peak AoI is bounded similarly by

$$\bar{\Gamma} \leq \frac{\mathrm{Var}(L)/r_{\mathrm{ch}}^2 + (1-q)b/q^2}{b/q - \mathbb{E}[L]/r_{\mathrm{ch}}} + \frac{\mathbb{E}[L]}{r_{\mathrm{ch}}} + \frac{b}{q}.$$

Note that a fixed-to-variable code that minimizes each of the two bounds will necessarily minimize the other as they differ only by a constant. Such a fixed-to-variable code was found using Larmore's convex hull algorithm [7]. The block size was set to $b = 4$, which is the minimal block size that maintains stability for the chosen region of p (as per [8]).

The mean delay $\bar{\Delta}$ (2) and mean peak AoI $\bar{\Gamma}$ (3) versus $H(X)$ of each scheme are depicted in Fig. 2, along with the analytical upper bounds of Sec. IV.

---

[3]If $\frac{\ell}{r_{\mathrm{ch}}}$ is integer, the equation can be solved by substituting $z = e^{-\theta}$ and solving the resulting polynomial equation.

[4]The bound is guaranteed to exist since $\phi_U(0) = 1$. In the case that this is the only real solution, the bound trivializes to ∞.
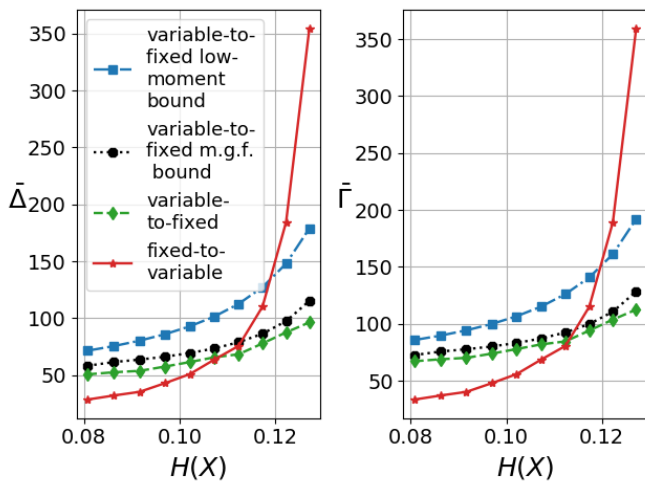
Fig. 2. Mean delay and mean peak AoI as a function of source entropy for random arrivals. $r_{\text{ch}} = 1/6.5$. fixed-to-variable block size $b = 4$. variable-to-fixed output length $\ell = 4$.

As we can see, the variable-to-fixed Tunstall code outperforms the fixed-to-variable coding scheme in the high-traffic region, while the fixed-to-variable scheme offers better results in the low-traffic region. Furthermore, we observe that the low-moment bound, while easier to obtain than the m.g.f. bound, is less tight, especially in the high-traffic region.

VI. DISCUSSION AND FUTURE WORK

We put forward variable-to-fixed coding for real-time source coding as an alternative to the hitherto used fixed-to-variable coding. We analyze our proposed method and demonstrate that it outperforms its fixed-to-variable counterparts when traffic is high both in terms of delay and AoI. We suggest the following topics for future work which are currently under investigation:

- We have used off-the-shelf Tunstall codes and demonstrated that even such codes outperform optimized fixed-to-variable codes when traffic is high (large $H(X)$). However, although these codes are optimal in terms of stabilizability (cf. Huffman codes in fixed-to-variable coding) they are not necessarily optimal in terms of minimum mean delay (cf. optimized codes of Zhong et al. [11]). Designing minimum mean delay codes is an interesting research avenue.

- Extension of the problem setup to allow several symbols arriving at the same time instant.

- As suggested by Sec. V, variable-to-fixed codes outperform fixed-to-variable codes for some parameters, and vice versa. Designing variable-to-variable coding schemes that outperform both of these classes of codes holds promise. This approach was explored in [11] where an adaptive choice of the block size was investigated for the simpler setting of periodic (deterministic) arrivals.

- Throughout this work, an implicit indicator signal was assumed available at the decoder that states whether the FIFO queue is empty or not. Studying this problem in the absence of such an indicator signal, in which case the event of not enough (or none at all) source symbols to encode needs to be taken into account (and encoded). This has been done for fixed-to-variable coding in [9].

- Formulation of the m.g.f. waiting-time bound for fixed-to-variable codes to use in lieu of the low-moment bound for the design of optimal fixed-to-variable codes.

VII. ACKNOWLEDGMENTS

REFERENCES

[1] P. Papadimitratos, A. De La Fortelle, K. Evenssen, R. Brignolo, and S. Cosenza, "Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation," *IEEE communications magazine*, vol. 47, no. 11, pp. 84–95, 2009.

[2] J. Bodner, H. Wykypiel, A. Greiner, W. Kirchmayr, M. C. Freund, R. Margreiter, and T. Schmid, "Early experience with robot-assisted surgery for mediastinal masses," *The Annals of Thoracic Surgery*, vol. 78, no. 1, pp. 259–265, 2004.

[3] M. Lovera, E. De Marchi, and S. Bittanti, "Periodic attitude control techniques for small satellites with magnetic actuators," *IEEE Transactions on Control Systems Technology*, vol. 10, no. 1, pp. 90–95, 2002.

[4] P. Mayekar, P. Parag, and H. Tyagi, "Optimal source codes for timely updates," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3714–3731, 2020.

[5] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE Int. Conf. on Comp. Comm. (INFOCOM)*, 2012, pp. 2731–2735.

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

[7] L. L. Larmore, "Minimum delay codes," *SIAM Journal on Computing*, vol. 18, no. 1, pp. 82–94, 1989.

[8] J. Zhong and R. D. Yates, "Timeliness in lossless block coding," in *Data Compression Conference (DCC)*. IEEE, 2016, pp. 339–348.

[9] J. Zhong, R. D. Yates, and E. Soljanin, "Timely lossless source coding for randomly arriving symbols," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2018.

[10] A. Lalitha, A. Khina T. Javidi, and V. Kostina, "Real-time binary posterior matching," in *Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*, Paris, France, Jul. 2019, pp. 2239–2243.

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

[12] B. Rogozin, "Some extremal problems in the theory of mass service," *Theory of Prob. & Its Apps.*, vol. 11, no. 1, pp. 144–151, 1966.

[13] B. P. Tunstall, "Synthesis of Noiseless Compression Codes," Ph.D. dissertation, School of Electrical Engineering, Georgia Institute of Technology, Oct. 1967.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory, Second Edition*. New York: Wiley, 2006.

[15] D. V. Lindley, "The theory of queues with a single server," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, no. 2. Cambridge University Press, 1952, pp. 277–289.

[16] J. Kingman, "Some inequalities for the queue GI/G/1," *Biometrika*, vol. 49, no. 3/4, pp. 315–324, 1962.

[17] A. Wald, "On cumulative sums of random variables," *The Annals of Mathematical Statistics*, vol. 15, no. 3, pp. 283–296, 1944.

[18] A. Gil, J. Segura, and N. M. Temme, *Numerical methods for special functions*. SIAM, 2007.