Low-Latency Preamble-Free Transmission of Short Messages via Quickest Change Detection

Giles Bischoff and Chih-Chun Wang
Elmore Family School of ECE, Purdue University; Email: {gbischo,chihw}@purdue.edu

***Abstract*—The latency and control overhead of sending the preamble in synchronous communications can be excessive when transmitting short sensing/control messages. To reduce these overheads, this work proposes a preamble-free solution based on the framework of *quickest change detection*. Specific contributions include a joint decoding/demodulation scheme that is provably asymptotically optimal, and a more practical CuSum-like implementation. Numerical results show that the proposed scheme reduces the latency by 47%–79% when compared to the preamble-based solutions. The scheme is also inherently robust and automatically adapts to any unknown underlying SNRs.**

I. Introduction

Low-latency transmission has been a key objective for 5G communications and beyond [1]–[3]. In general, a modern communication scheme first transmits the preamble, followed by the payload, where the former is used for synchronization and acquisition of channel parameters. Despite its throughput advantages over asynchronous solutions, for short sensing/control messages, say 8–32 bits, the latency overhead of sending a preamble, typically ranging between 139 to 839 symbols [4], [5], may be severe. It is with such settings in mind that this work proposes a preamble-free transceiver based on the framework of *quickest change detection* (QCD).

Our main idea is to recognize that this deficiency of synchronous schemes is due to the fact that the preamble is designed exclusively for synchronization and channel state acquisition, while the payload, with the help of error correcting codes, is used exclusively for combating noise of the channel. Such a rigid separation diminishes the inference power for either mission. In contrast, we assign each of the 2^q messages to an unending sequence of symbols, where q is the number of bits being transmitted and is assumed to be small $8 \le q \le 16$. The receiver then listens to the transmitted sequence indefinitely until it can decode the message with high confidence. Broadly speaking, the transmitted symbol sequence is now used simultaneously for synchronization (resolving any ambiguity caused by delay) and for conveyance of the messages (resolving any ambiguity among the 2^q messages).

The contributions of our results are summarized as follows:

1) We have formulated a low-latency preamble-free communication problem under the QCD framework.

2) We have designed a joint decoding/demodulation scheme, and proved its asymptotic optimality in terms of the tradeoff among *delay*, *error probability*, and *false reception* frequency. To further lower the complexity, a simplified CuSum implementation has also been devised.

3) Numerical evaluation is used to compare our scheme versus the preamble-based solutions. Specifically, our scheme lowers the latency by 47%–79% under a standard Additive White Gaussian Noise Channel (AWGNC) model.

4) Our scheme is robust and automatically adapts to the underlying channel condition. Namely, when operated in a noisier (resp. cleaner) environment, our receiver automatically lengthens (resp. shortens) the decoding delay by accumulating more (resp. less) observations before making a decision. This is in sharp contrast to the preamble-based solutions, for which the delay, i.e., the length of the preamble, is pre-determined.[1]

Remark: The application of QCD to communications, and in particular, networking, is widespread [6]–[8]. Further comparison to existing QCD solutions will be discussed in Sec. II-B.

II. Problem Formulation

We consider the problem of transmitting a short q-bit message $8 \le q \le 16$ over an AWGNC. For any $m \in [1, 2^q]$, our scheme assigns an *unending* complex-valued "symbol" sequence $\mathbf{s}^m \triangleq \{s_n^m \in \mathbb{C} : n \ge 0\}$ to the m-th message. And we assume $\mathbf{s}^m$ is periodic with period N, where N is chosen by the system designer.[2] We assume $\mathbf{s}^m$ is of unit power:

$$\sum_{n=0}^{N-1} |s_n^m|^2 = N, \quad \forall m \in [1, 2^q]. \tag{1}$$

The "codebook" $\{\mathbf{s}^m : 1 \le m \le 2^q\}$ is known *a priori* to both the transmitter (Tx) and the receiver (Rx).

For any specific message m, the Tx sends the following "sample" sequence $\mathbf{x}^{m,t_0} \triangleq \{x_t^{m,t_0} : t \text{ is an integer}\}$:

$$x_t^{m,t_0} = \sqrt{P} \cdot \sum_{n=0}^{\infty} s_n^m \cdot \operatorname{sinc}\left(\frac{t - t_0 - n \cdot F}{F}\right), \tag{2}$$

where P is the transmission power, F is the number of samples per symbol,[3] and $t_0 \ge 1$ (unit: samples) is the starting time of the transmission. Our scheme uses $\operatorname{sinc}(\cdot)$-based interpolation in (2) and can be easily applied to other pulse-shaping filters, e.g., root-raised cosine.

[1]The optimal length of a preamble can be decided if SNR is known to the transmitter. However, before sending a preamble, it is difficult to know the channel condition, especially when one is attempting communication for the very first time. It is essentially a chicken-and-egg problem.

[2]For future work, it may be advantageous to also consider aperiodic $\mathbf{s}^m$. However, as discussed in Sec. III, the periodicity enables an elegant Fourier-series-based design and the corresponding low-complexity CuSum algorithms.

[3]As a theoretical exploration, we assume a sufficiently large F so that the discrete model in (2) is reasonably close to the continuous model in real life.

The Rx observes the signal $\mathbf{Y} \triangleq \{Y_t \in \mathbb{C} : t \geq 1\}$:

$$Y_t = e^{j\theta} x_{t-\delta_t}^{m,t_0} + W_t, \tag{3}$$

where δ_t is the propagation delay, W_t is a complex Gaussian variable with independent real and imaginary components, both of which being $\mathcal{N}(0, \frac{\sigma^2}{2})$, and θ is the phase shift. If we use N_0 to denote the *one-sided noise spectral density*, we then set

$$\sigma^2 = \frac{N_0 F}{2}. \tag{4}$$

Note that the propagation delay δ_t in (3) has the same time-shifting effect as the (unknown) starting time t_0 in (2). As a result, we simply set $\delta_t = 0$ for the rest of the discussion.

To make our setting a fully discrete one, we assume θ belongs to a discrete set $\Theta_d \triangleq \left\{\frac{2\pi h}{H} : 1 \leq h \leq H\right\}$ for some sufficiently large but fixed H.

Remark: As a first-order approximation model, we assume no frequency drift (no Doppler), and the phase shift θ is constant throughout the transmission. A future work is to allow θ changes over time t, e.g., $\theta(t)$ being a drifted random walk.

Since Rx does not know m, t_0 and θ, the goal is to design a *stopping time* T and a decision $\hat{M}$ at time T that minimize the conditional decoding *delay* for the worst (m, t_0, θ) value:

$$D(T) \triangleq \sup_{m,t_0,\theta} \mathbb{E}^{m,t_0,\theta} \left\{ T - t_0 + 1 \middle| \, T \geq t_0 \right\}, \tag{5}$$

minimize the conditional *error probability* for the worst (m, t_0, θ) value:

$$p_e(T, \hat{M}) \triangleq \sup_{m,t_0,\theta} \mathbb{P}^{m,t_0,\theta} \left(\hat{M} \neq m \middle| \, T \geq t_0 \right), \tag{6}$$

and maximize the average run length to *false reception*:

$$T_{\mathsf{FR}}(T) \triangleq \lim_{t_0 \to \infty} \mathbb{E}^{m,t_0,\theta}\{T\} = \mathbb{E}^{m,\infty,\theta}\{T\}. \tag{7}$$

Note that because the starting time is now $t_0 = \infty$, the expectation in (7) is indepependent of the m value.

More rigorously, the goal is to design $\{\mathbf{s}^m : m\}$, T, and $\hat{M}$ that optimize the tradeoff among D, p_e and T_{FR} by fixing two of the above three quantities (usually (6) and (7)), and then optimizing the third (usually (5)). Herein, we use the superscript (m, t_0, θ) to denote the probability law under message m, starting time t_0, and phase shift θ.

A. A Transient-Free Approximation Model

Since the symbol sequence $\mathbf{s}^m$ has period N (unit: symbols), one might expect that the post-t_0 portion of the sample sequence $\mathbf{x}^{m,t_0}$ is periodic with period $N \cdot F$ (unit: samples). Unfortunately, it is not the case since the $\mathrm{sinc}(\cdot)$ function in (2) has unbounded support and the transient behavior from pre-t_0 to post-t_0 will propagate indefinitely within $\{x_t^{m,t_0} : t\}$, which complicates our analysis. To circumvent this challenge, we first simplify the transmission model in (2) by defining

$$x_{\mathsf{ss},i}^{m,t_0} \triangleq \lim_{k \to \infty} x_{NF \cdot k + i}^{m,t_0} \quad \text{for all } i \in [0, NF) \tag{8}$$

as the steady-state (ss) waveform when $t \to \infty$. Note that $x_{\mathsf{ss},i}^{m,t_0}$ now has period NF (unit: samples). We then define

$$\tilde{x}_t^{m,t_0} \triangleq \begin{cases} 0 & \text{if } t < \lfloor t_0 \rfloor_{NF} \\ x_{\mathsf{ss},i}^{m,t_0} & \text{if } t \geq \lfloor t_0 \rfloor_{NF} \text{ and } t = NF \cdot k + i \end{cases} \tag{9}$$

where $\lfloor t_0 \rfloor_{NF} \triangleq \lfloor \frac{t_0}{NF} \rfloor \cdot NF$ is the *floor function* over the multiples of NF.

Intuitively speaking, *the new sample sequence* $\tilde{\mathbf{x}}^{m,t_0} \triangleq \{\tilde{x}_t^{m,t_0} : t\}$ *bypasses any of the transient behavior from pre-*t_0 *to post-*t_0 *and directly appends the all-zero samples (as a proxy of any pre-*t_0 *waveform) with the steady-state waveform* $x_{\mathsf{ss},i}^{m,t_0}$ *(as a proxy of any post-*t_0 *waveform).* One can verify that for $t \ll t_0$ or $t \gg t_0$, we have x_t^{m,t_0} in (2) and $\tilde{x}_t^{m,t_0}$ in (9) are nearly identical. The biggest difference between the two occurs during the interval $t \in [\lfloor t_0 \rfloor_{NF}, \lfloor t_0 \rfloor_{NF} + NF)$ when the transient behavior is at its peak.

We use $\tilde{\mathbb{E}}^{m,t_0,\theta}$ and $\tilde{\mathbb{P}}^{m,t_0,\theta}$ to denote the distribution of $\mathbf{Y} = \{Y_t : t\}$ when we replace $x_{t-\delta_t}^{m,t_0}$ in (3) by the new $\tilde{x}_{t-\delta_t}^{m,t_0}$ in (9). Similarly, we use $\tilde{D}$, $\tilde{p}_e$, and $\tilde{T}_{\mathsf{FR}}$ to denote the counterparts of D, p_e and T_{FR} by replacing the distributions $\mathbb{E}$ and $\mathbb{P}$ in (5)–(7) with the new $\tilde{\mathbb{E}}$ and $\tilde{\mathbb{P}}$.

As will be seen in Sec. III-B, our performance discussion is in the asymptotic regime. Since $\mathbf{x}^{m,t_0}$ in (2) and $\tilde{\mathbf{x}}^{m,t_0}$ (9) are almost identical except for the transient behavior around time t_0, we expect the performance and analysis using (9) and the corresponding probability law $\tilde{\mathbb{P}}^{m,t_0,\theta}$ to be quite relevant for the more realistic model in (2) and the corresponding $\mathbb{P}^{m,t_0,\theta}$.

B. Existing Work on QCD

There are many variations of QCD analyses, including Bayesian vs non-Bayesian [8], different definitions of Average Detection Delay (ADD) [8], [9], iterative CuSum algorithms [9]–[11], and stochastically dependent observations [8]. In this context, our model (5)–(7) takes a non-Bayesian approach with the delay objective (5) closely related to the *Pollak criterion*. Our achievability and converse results are related to the asymptotic optimality in [10] and [12], respectively. Our work extends these QCD results to the preamble-free transmission problem by generalizing the multiple post-change hypotheses settings [6], [9], [10], [13] to incorporate composite sub-hypotheses like unknown timing t_0 and phase shift θ.

III. Main Results

We first assume that the codebook $\{\mathbf{s}^m : m \in [1, 2^q]\}$ is fixed, and discuss how to design asymptotically optimal $(T, \hat{M})$. We then describe how to optimize $\{\mathbf{s}^m : m \in [1, 2^q]\}$.

A. Asymptotically Optimal Algorithm

Given any σ, we use $f_\sigma(y|x)$ to denote the pdf of a complex Gaussian variable Y with mean $x \in \mathbb{C}$ and independent real and imaginary components with per-component variance $\frac{\sigma^2}{2}$. We then define

$$f_{\mathsf{ss},t}^{m,t_0,\theta}(y) \triangleq \begin{cases} f_\sigma(y|0) & \text{if } t < t_0 \\ f_\sigma(y|e^{j\theta} x_{\mathsf{ss},i}^{m,t_0}) & \text{if } t \geq t_0 \text{ and } t = kNF + i. \end{cases} \tag{10}$$

Namely, $f_{\mathsf{ss},t}^{m,t_0,\theta}(y)$ assumes that an all-zero waveform is transmitted before t_0, and the steady-state waveform $x_{\mathsf{ss},i}^{m,t_0}$ is transmitted right after t_0 and it goes through the phase shift $e^{j\theta}$ in (3). Also see similar discussion[4] around (9).

We then define the log-likelihood ratio (LLR)

$$L^{m,t_0,\theta}(t) \triangleq \ln\left(\frac{f_{\mathsf{ss},t}^{m,t_0,\theta}(Y_t)}{f_\sigma(Y_t|0)}\right),\ \forall m,t_0,\theta. \tag{11}$$

Subsequently, for any arbitrarily given m, we define

$$g^m(t) \triangleq \begin{cases} \max\limits_{1\le t_0\le t+1,\ \theta\in\Theta_d} \sum\limits_{\tau=1}^{t} L^{m,t_0,\theta}(\tau) & \text{if } m\in[1,2^q] \\ 0 & \text{if } m=0. \end{cases} \tag{12}$$

Herein we use $m=0$ to denote the scenario of $t_0=\infty$ (no message is ever transmitted), and the corresponding LLR is thus 0. For any $c>0$, we then define the *stopping time*

$$T^m(c) \triangleq \inf\left\{t\in\mathbb{N}: g^m(t) - \left(\max_{m'\in[0,2^q]\setminus m} g^{m'}(t)\right)\ge c\right\}. \tag{13}$$

Finally, we define the overall stopping time and the message decision pair $(T,\hat{M})$ by

$$T(c) \triangleq \min_{m\in[1,2^q]} T^m(c), \tag{14}$$

$$\hat{M}(c) \triangleq \text{argmin}_{m\in[1,2^q]}\, T^m(c). \tag{15}$$

The description of our scheme $(T(c),\hat{M}(c))$ is complete.

B. Performance Analysis of The Proposed Scheme

To analyze the performance of our scheme $(T(c),\hat{M}(c))$, we assume that there exists a finite constant $B<\infty$ such that

$$\tilde{\mathbb{P}}^{m,t_0,\theta}\left(\left|L^{m,t_0,\theta}(t)\right|\le B\right)=1,\ \forall m,t_0,\theta. \tag{16}$$

Namely, the LLR $L^{m,t_0,\theta}(t)$ is globally bounded within $[-B,B]$. Note that this assumption is technically not true since AWGNCs have unbounded support. However, with the tail probability decaying at rate $e^{-\frac{x^2}{2}}$, any impact of having extremely large $L^{m,t_0,\theta}(t)$ is likely to be negligible. Also, in practice, the received sample Y_t in (3) is often quantized to be within a bounded value, which automatically ensures (16).

For any given $\mathbf{s}^m$, we define $\mathbf{b}^m \triangleq (b_0^m, b_1^m, \cdots, b_{N-1}^m)$ as the unique N-dimensional vector satisfying

$$s_n^m = \sum_{i=0}^{N-1} b_i^m e^{j\frac{2\pi n\cdot i}{N}}. \tag{17}$$

That is, $\mathbf{s}^m$ and $\mathbf{b}^m$ are related through *discrete Fourier transforms*. For the purposes of the below analysis, we allow the sequence $\mathbf{b}^m$ to be tailbiting so that $b_{-1}^m=b_{N-1}^m$, $b_{-2}^m=b_{N-2}^m$ and so on. Given the codebook $\{\mathbf{s}^m: m\in[1,2^q]\}$, we solve the corresponding $\{\mathbf{b}^m: m\in[1,2^q]\}$ in (17) and compute

$$\rho_{m_1,m_2} \triangleq \max_{r\in[0,1]} \left|\sum_{i=-\lceil\frac{N}{2}\rceil+1}^{\lfloor\frac{N}{2}\rfloor} b_i^{m_1}\,(b_i^{m_2})^*\, e^{-j2\pi\cdot i\cdot r}\right| \tag{18}$$

$$\rho \triangleq \max_{1\le m_1\ne m_2\le 2^q} \rho_{m_1,m_2} \tag{19}$$

$$\psi \triangleq \min(1-\rho, 0.5) \tag{20}$$

where the maximization of the real-valued r is over the continuous interval $[0,1]$, and $(\cdot)^*$ is the complex conjugate.

We now characterize the performance of $(T(c),\hat{M}(c))$ under the probability law $\tilde{\mathbb{P}}^{m,t_0,\theta}$ and $\tilde{\mathbb{E}}^{m,t_0,\theta}$ in Sec. II-A.

Proposition 1: Under the assumptions of sufficiently large F, sufficiently large $H=|\Theta_d|$, and the global boundedness constraint (16), our scheme $(T(c),\hat{M}(c))$ satisfies

$$\tilde{D}(T(c)) \le c\cdot\frac{\sigma^2}{P\cdot\psi}\cdot(1+o(1)) \quad \text{(unit: samples)} \tag{21}$$

$$\tilde{p}_e(T(c),\hat{M}(c)) \le c\cdot e^{-c}\cdot 2^q\cdot NFH\cdot e^{NFB} \cdot\left(\frac{\sigma^2}{P\cdot\psi}+1\right)\cdot(1+o(1)) \tag{22}$$

$$\tilde{T}_{\mathsf{FR}}(T(c)) \ge e^c\frac{1}{2^qH}\cdot(1+o(1)) \quad \text{(unit: samples)} \tag{23}$$

where $o(1)\to 0$ when $c\to\infty$.

The proof of Proposition 1 is omitted due to space constraints. Intuitively, the performance of a QCD scheme is decided by the Kullback-Leibler (KL) divergence between competing hypotheses. However, because our hypotheses consist of not only the message m to decode but also the unknown starting time t_0 and phase shift θ, one must use the KL divergence to also resolve the ambiguity caused by (t_0,θ).

More technically speaking, the real part of the summation $\sum_i b_i^{m_1}(b_i^{m_2})^*$ within the definition of ρ_{m_1,m_2} in (18) represents the inner product between two competing symbol sequences $\mathbf{s}^{m_1}$ and $\mathbf{s}^{m_2}$. The additional $e^{j2\pi\cdot i\cdot r}$ term in (18) corresponds to the possibility that the unknown time shift t_0 could further rotate the Fourier-domain coefficients $b_i^{m_1}$ by $e^{j2\pi\cdot i\cdot r}$. Since one must resolve the ambiguity caused by t_0, we take $\max_r$ in (18) to find the closest pair of $\mathbf{s}^{m_1}$ and $\mathbf{s}^{m_2}$ under unknown time shift t_0. Finally, the amplitude operator in (18) takes into account that the unknown phase θ could rotate the entire summation, i.e., *the maximum value of the real part of a rotated summation is just the amplitude of the summation*. Per the above reasonings, ρ_{m_1,m_2} finds the largest inner product between $\mathbf{s}^{m_1}$ and $\mathbf{s}^{m_2}$ under unknown t_0 and θ.

The ρ in (19) then finds the pair (m_1,m_2) that are the closest to each other. Finally, ψ in (20) converts the inner product ρ to the KL divergence, assuming $\text{SNR}=1$, where $1-\rho$ is the KL-divergence between the closest (m_1,m_2) and 0.5 is the KL-divergence between message $\mathbf{s}^m$ and the idle all-zero waveform. Since one has to resolve the ambiguity of both m_1-versus-m_2 and m-vs-idle, we take the minimum when computing ψ. Scaling ψ to the true *per-sample* $\text{SNR}=\frac{P}{\sigma^2}$, the

[4] As will be seen shortly, the pdf formulas of (10) are used to construct the scheme $(T,\hat{M})$, which is then analyzed using the probability law described in (9). Note that the pdfs in (10) are *not* the pdfs of (9) because the distribution in (10) changes at $t=t_0$, but the distribution in (9) changes at $t=\lfloor t_0\rfloor_{NF}$. Also, the pdfs of (10) are *not* the pdfs of (2) either because of the transient behavior in (2). It is best to just view (10) as a building block of the QCD scheme, not the true pdf of any probability law being considered.

final KL-divergence then dictates the asymptotic performance of our scheme in the form of (21)–(23).

We now provide the converse results.

Proposition 2: Assume sufficiently large F, sufficiently large $H = |\Theta_d|$, and the global boundedness constraint (16). For any $c > 0$, define $\tilde{D}^*(c)$ as the optimal value of the following minimization problem:

$$\tilde{D}^*(c) \triangleq \inf_{\text{any } (T,\hat{M})} \tilde{D}(T) \tag{24}$$

$$\text{subject to} \quad \tilde{p}_e(T,\hat{M}) \le e^{-c} \tag{25}$$

$$\text{and } \tilde{T}_{\mathsf{FR}}(T) \ge e^{c}. \tag{26}$$

We then have

$$\tilde{D}^*(c) \ge c \cdot \frac{\sigma^2}{P \cdot \psi} \cdot (1 + o(1)). \tag{27}$$

The proof of Proposition 2 is by the reduction-based argument using the converse in [12]. We thus omit the details.

Comparing (21)–(23) and (24)–(27), our scheme achieves simultaneously the same exponential decay/increase of $\tilde{p}_e$ and $\tilde{T}_{\mathsf{FR}}$ and the same linear increase of $\tilde{D}$ with respect to c as any achievability scheme one can possibly design. Therefore, our scheme is asymptotically optimal when $c \to \infty$.

Remark: While being order-optimal, the coefficients of our achievability performance in (22)–(23) are quite loose when compared to (25)–(26). This is due to multiple relaxation steps in the proof of achievability, such as the union bounds. The actual performance of our scheme is much tighter as is evidenced in the numerical evaluation in Sec. IV.

C. Code Construction

By Proposition 1, the smaller the ρ, the larger the ψ, the better the performance of our scheme $(T(c), \hat{M}(c))$. We now discuss how to design a codebook $\{\mathbf{s}^m : m\}$ (or equivalently the corresponding $\{\mathbf{b}^m : m\}$) that minimizes ρ defined in (19).

Note that (1) and (17) jointly imply the power constraint:

$$\sum_{i=0}^{N-1} |b_i^m|^2 = 1, \ \forall m \in [1, 2^q]. \tag{28}$$

Based on (28), our construction is described as follows:

1: **for** $m \in [1, 2^q]$ **do**
2: Choose an N-dimensional amplitude vector $(a_0, a_1, \cdots, a_{N-1})$ independently and uniformly randomly from the *unit sphere*. Also see (28).
3: Choose N phases $\phi_0, \cdots, \phi_{N-1}$ independently and uniformly randomly from $[0, 2\pi)$.
4: Set $b_i^m = a_i \cdot e^{j\phi_i}$ for all $i \in [0, N)$.
5: **end for**
6: **while** within the max. number of allowable iterations **do**
7: Evaluate ρ in (19) using the latest $\{\mathbf{b}^m : m \in [1, 2^q]\}$. Let (m_1^*, m_2^*) denote the message pair that attains the maximum ρ value, i.e., $\rho = \rho_{m_1^*, m_2^*}$.
8: Choose two new N-dimensional vectors $(a_0, \cdots, a_{N-1})$ and $(\phi_0, \cdots, \phi_{N-1})$ in the same way as in Lines 2 and 3.
9: Update $b_i^{m_1^*} = a_i \cdot e^{j\phi_i}$ for all $i \in [0, N)$.
10: Evaluate ρ using the new $\mathbf{b}^{m_1^*}$ while keeping the rest $\{\mathbf{b}^{m'} : m' \ne m_1^*\}$.
11: **if** the new $\mathbf{b}^{m_1^*}$ results in a smaller ρ **then**
12: Keep the update.
13: **else**
14: Discard the update.
15: **end if**
16: **end while**

Basically, we incrementally improve the codebook performance by replacing the bottleneck $\mathbf{b}^{m_1^*}$ with a new one.

Table I summarizes the smallest ρ values obtained via the above construction for different (q, N) values. The number of bits being transmitted is $8 \le q \le 12$. The symbol period N is 32 or 40. As can be seen, our construction is quite effective since any ρ smaller than 0.5 will not have any meaningful impact to the performance due to $\psi = \min(1 - \rho, 0.5)$ in (20). All our ρ values are very close to 0.5. Furthermore, the smaller the ratio $\frac{q}{N}$, the better the ρ value. The intuition is that the scheme (codebook) only needs to send a smaller number of bits, in average, over the entire period N, which makes it easier for the Rx to distinguish between different messages m.

No. of bits q	8	10	10	12
Period N	32	40	32	40
The $\frac{q}{N}$ ratio	0.25	0.25	0.3125	0.3
ρ	0.4979	0.5128	0.5433	0.5423

TABLE I

D. Practical Implementation

The execution of our algorithm relies on computing the values of $g^m(t)$ defined in (10)–(12). We now elaborate how to compute $g^m(t)$ in an iterative CuSum fashion. Assuming H is sufficiently large, i.e., the discrete set Θ_d in (12) can be replaced by a continuous interval $[0, 2\pi)$, we define

$$g_i^m(t) \triangleq \max_{1 \le t_0 = kNF + i \le t+1} \ \max_{\theta \in [0, 2\pi)} \sum_{\tau=t_0}^{t} L^{m,t_0,\theta}(\tau), \tag{29}$$

i.e., the maximum operation is only over those $\{t_0 : t_0 \mod NF = i\}$. Then (12) and (29) immediately imply

$$g^m(t) = \begin{cases} \max_{i \in [0, NF)} g_i^m(t) & \text{if } m \in [1, 2^q] \\ 0 & \text{if } m = 0. \end{cases} \tag{30}$$

Proposition 3: $g_i^m(t)$ can be computed iteratively by:

$$g_i^m(t) = \left| \sum_{\tau=\lambda_i^m+1}^{t} \frac{\left(x_{\mathsf{ss},(\tau \bmod NF)}^{m,i}\right)^* \cdot Y_\tau}{\sigma^2} \right| - \sum_{\tau=\lambda_i^m+1}^{t} \frac{\left|x_{\mathsf{ss},(\tau \bmod NF)}^{m,i}\right|^2}{2\sigma^2} \tag{31}$$

where $(\tau \bmod NF)$ is the modulo of NF, $x_{\mathsf{ss},i}^{m,t_0}$ is the steady-state waveform defined in (8), and

$$\lambda_i^m \triangleq \sup\{\lambda \le t : \ \lambda = k \cdot NF + i - 1, \ k \in [0, \infty), \text{ satisfying } \min(g_i^m(\lambda), \lambda - i + 1) \le 0\}. \tag{32}$$

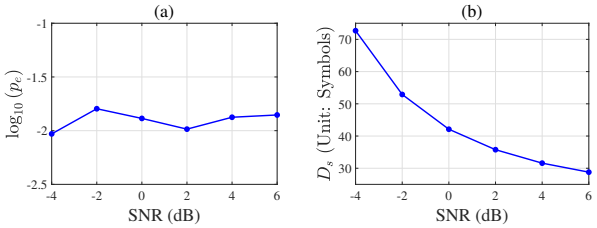

Fig. 1. Empirical error probability p_e and delay D_s versus different SNRs.

That is, we want λ_i^m to be the last time $g_i^m(\lambda)$ being ≤ 0, for all λ of the form $\lambda = k \cdot NF + i - 1$ with some integer k. If all those λ have $g_i^m(\lambda) > 0$, then we set $\lambda_i^m = i - 1$ (i.e., $k = 0$ and $\lambda = k \cdot NF + i - 1 = i - 1$).

The proof of Proposition 3 follows similar principles as the CuSum algorithm [6], [10], and is omitted due to space constraints.

Using Proposition 3, we can compute $g_i^m(t)$ efficiently for each (i, m) by tracking the two summations in (31) separately and by updating the starting index $\lambda_i^m + 1$ (effectively resetting the two summations) whenever $g_i^m(t) \leq 0$ for those $t = k \cdot NF + i - 1$. The complexity of this iterative computation is thus $O(MNF)$ since we have MNF different pairs of (i, m).

IV. Numerical Evaluation

While Propositions 1 and 2 characterize the asymptotic performance ($c \to \infty$) under the transient-free approximation model (probability law) $\tilde{\mathbb{P}}^{m,t_0,\theta}$ in Sec. II-A, we use numerical evaluation to evaluate the performance of finite c under the realistic model $\mathbb{P}^{m,t_0,\theta}$ described in the beginning of Sec. II.

Remark: Note that our algorithm uses the pdfs specified in (10), which are very different from the pdfs of the true $\mathbb{P}^{m,t_0,\theta}$ due to the transient behavior around time $t = t_0$ in (2). Our numerical results thus enable us to evaluate the impact of this mismatch between the steady-state-based design (10)–(15) versus the realistic, transient-rich environment $\mathbb{P}^{m,t_0,\theta}$ in (2). As can be seen in Fig. 1(b), our scheme achieves very short latency despite the fact that it applies the steady-state-based pdfs (10) to the transient-rich environment in (2), where the former is not the true pdf of the latter. We believe the reason is that the transient behavior of the $\mathrm{sinc}(\cdot)$ in (2) converges relatively swiftly to its steady state after just a few side lobes (usually less than 5 symbol durations or so).

For our evaluation, we fix $q = 8$, $N = 32$, and generate a period-32 codebook $\{\mathbf{s}^m : m \in [1, 2^q]\}$ with $\rho = 0.4979$ and $\psi = \max(1 - \rho, 0.5) = 0.5$, i.e. one of the four codebooks reported in Table I. We also fix $F = 30$ and $c = 3.6$. For each *per-symbol* SNR level $\frac{F \cdot P}{\sigma^2} = \frac{2P}{N_0}$, we run our CuSum algorithm for 3000 trials. During each trial, we choose the message m uniformly randomly from $[1, 2^q]$, the starting point t_0 (unit: samples) uniformly randomly from $[30F + 1, 60F]$ and the phase shift θ uniformly randomly from Θ_d with $H = 1000$. When implementing (2), we assumed that the $\mathrm{sinc}()$ function is exactly 0 outside $l_s = 20$ side lobes. To respect the causality of the transmission, we define the *empirical decoding delay* as $D_s = T - t_0 + l_s \cdot F$ (unit: samples) to include the time needed to transmit the l_s side lobs before the official starting time $t = t_0$. That is, D_s includes the $l_s \cdot F$ samples needed for the transmitter to "ramp up" the transmission before sending the first symbol s_0^m at time $t = t_0$. The D_s and the empirical error probability p_e for different SNRs are reported in Fig. 1.

The asymptotic analysis (22) in Proposition 1 suggests that the error probability p_e is decided mostly by the threshold value c. For the same $c = 3.6$, the variation of p_e versus SNR in Fig. 1(a) is indeed quite small. On the other hand, (21) in Proposition 1 suggests that the delay D is negatively correlated to the SNR $\frac{P}{\sigma^2}$, which is also verified in Fig. 1(b).

Note that the same codebook at the transmitter is used for all SNRs in Fig. 1. Since the underlying SNR can be estimated at the receiver (likely with some slight mismatch), it means that when facing a low-SNR channel, the receiver will automatically delay its decision so that it can accumulate more observations[5] to meet the predefined requirement on p_e.

For the same codebook at the transmitter, the receiver can easily adjust the reliability p_e requirement by selecting a different c value. For example, when we set $c = 5$ in our simulation, the resulting p_e is around 10^{-3}. Since we only simulated 3000 trials, we deliberately chose $c = 3.6$ to keep $p_e \approx 10^{-2}$ for a more accurate empirical estimation.

For comparison, we also examine the synchronization performance of the Zaddoff-Chu sequence with 139 symbols, a typical preamble choice in [4], [5]. The same transmission model in (2) and (3) is used to properly scale the preamble with the desired SNR. At SNR=4dB, for about 1.75% of the total 10^5 trials, the timing synchronization of the Zaddoff-Chu sequence is off by 2 samples (i.e., $2/F = 6.6\%$ of a symbol). Since large synchronization error will have catastrophic impact to the decoding error probability, we would expect that for 4dB, one may not want to use any preamble shorter than 139 symbols. Note that at 4dB the average delay of our preamble-free scheme is only 32 symbols, see Fig. 1(b). That is, our scheme can successfully deliver the 8-bit message when a preamble-based solution would have only finished sending 23% of the 139-symbol preamble. For SNRs between -4db to 6dB, the latency savings range from 47%–79%, which has not even accounted for the fact that the preamble-based scheme still has to send the 8-bit message as an error-control-coded payload, and our D_s is conservative and includes $l_s F$ ramp-up time (unit: samples) of transmitting l_s side lobes before time t_0.

V. Conclusion

This work has studied low-latency preamble-free transmission of short messages, and developed an asymptotically optimal solution. Numerical results show that our construction can shorten the delay by 47%–79% when compared to existing preamble-based solutions. The proposed scheme is inherently robust and can automatically adapt to different underlying channel conditions.

[5]The mathematical reason is that a low-SNR channel has less "upward drift" of the LLR $L^{m,t_0,\theta}(t)$, thus the longer hitting time $T(c)$.

REFERENCES

[1] H. Kim, "Ultra-reliable and low latency communication systems," in *Design and Optimization for 5G Wireless Communications*. Chichester, UK: John Wiley & Sons, Ltd, 2020, pp. 303–342.

[2] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk and scale," *arXiv.org*, 2018.

[3] M. A. Siddiqi, H. Yu, and J. Joung, "5g ultra-reliable low-latency communication implementation challenges and operational issues with iot devices," *Electronics (Basel)*, vol. 8, no. 9, pp. 981–, 2019.

[4] J. G. Andrews, "A primer on zadoff chu sequences," *arXiv.org*, 2023.

[5] T. A. Khan and X. Lin, "Random access preamble design for 3gpp non-terrestrial networks," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–5.

[6] A. Warner and G. Fellouris, "Cusum for sequential change diagnosis," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 486–491.

[7] O. Hadjiliadis, H. Zhang, and H. Poor, "One shot schemes for decentralized quickest change detection," *IEEE transactions on information theory*, vol. 55, no. 7, pp. 3346–3359, 2009.

[8] J. Z. Hare, L. Kaplan, and V. V. Veeravalli, "Toward uncertainty aware quickest change detection," in *2021 IEEE 24th International Conference on Information Fusion (FUSION)*. International Society of Information Fusion (ISIF), 2021, pp. 1–8.

[9] K. Liu and Y. Mei, "Discussion on "sequential detection/isolation of abrupt changes" by igor v. nikiforov," *Sequential analysis*, vol. 35, no. 3, pp. 316–319, 2016.

[10] I. Nikiforov, "A simple recursive algorithm for diagnosis of abrupt changes in random signals," *IEEE transactions on information theory*, vol. 46, no. 7, pp. 2740–2746, 2000.

[11] J. Unnikrishnan, V. V. Veeravalli, and S. P. Meyn, "Minimax robust quickest change detection," *IEEE transactions on information theory*, vol. 57, no. 3, pp. 1604–1614, 2011.

[12] I. Nikiforov, "A lower bound for the detection/isolation delay in a class of sequential tests," *IEEE transactions on information theory*, vol. 49, no. 11, pp. 3037–3047, 2003.

[13] A. G. Tartakovsky, "Multidecision quickest change-point detection: Previous achievements and open problems," *Sequential analysis*, vol. 27, no. 2, pp. 201–231, 2008.