Causal sampling, compressing, and channel coding of streaming data

Thesis by
Nian Guo

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended August 5, 2022

Nian Guo
ORCID: 0000-0003-4490-328X

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Professor Victoria Kostina for her endless support and patience. Her passion and earnestness for research have greatly inspired me. Without her, I could not have undertaken this journey.

I am grateful to my defense and candidacy committees, Professor Michelle Effros, Professor Jehoshua (Shuki) Bruck, Professor Babak Hassibi, and Professor John Doyle, for their invaluable insights, comments, and encouragement.

I would also like to thank my fellow labmates and friends at Caltech for the stimulating discussions and all the fun that we have had in the last five years.

Finally, I would like to thank my family for their unconditional love and support. Their belief in me has kept my spirit up during the journey.

ABSTRACT

With the emergence of the Internet of Things, communication systems, such as those employed in distributed control and tracking scenarios, are becoming increasingly dynamic, interactive, and delay-sensitive. The data in such real-time systems arrive at the encoder progressively in a streaming fashion. An intriguing question is: what codes can transmit streaming data with both high reliability and low latency? Classical non-causal (block) encoding schemes can transmit data reliably but under the assumption that the encoder knows the entire data block before the transmission. While this is a realistic assumption in delay-tolerant systems, it is ill-suited to real-time systems due to the delay introduced by collecting data into a block. This thesis studies causal encoding: the encoder transmits information based on the causally received data while the data is still streaming in and immediately incorporates the newly received data into a continuing transmission on the fly.

This thesis investigates causal encoding of streaming data in three scenarios: causal sampling, causal lossy compressing, and causal joint source-channel coding (JSCC). In the causal sampling scenario, a sampler observes a continuous-time source process and causally decides when to transmit real-valued samples of it under a constraint on the average number of samples per second; an estimator uses the causally received samples to approximate the source process in real time. We propose a causal sampling policy that achieves the best tradeoff between the sampling frequency and the end-to-end real-time estimation distortion for a class of continuous Markov processes. In the causal lossy compressing scenario, the sampling frequency constraint in the causal sampling scenario is replaced by a rate constraint on the average number of bits per second. We propose a causal code that achieves the best causal distortion-rate tradeoff for the same class of processes. In the causal JSCC scenario, the noiseless channel and the continuous-time process in the previous scenarios are replaced by a discrete memoryless channel with feedback and a sequence of streaming symbols, respectively. We propose a causal joint source-channel code that achieves the maximum exponentially decaying rate of the error probability compatible with a given rate. Remarkably, the fundamental limits in the causal lossy compressing and the causal JSCC scenarios achieved by our causal codes are no worse than those achieved by the best non-causal codes. In addition to deriving the fundamental limits and presenting the causal codes that achieve the limits, we also show that our codes apply to control systems, are resilient to system

deficiencies such as channel delay and noise, and have low complexities.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] N. Guo and V. Kostina. "Optimal causal rate-constrained sampling of the Wiener process". In: *IEEE Transactions on Automatic Control* 67.4 (Apr. 2022), pp. 1776–1791. DOI: 10.1109/TAC.2021.3071953.
Nian Guo participated in the conception of the project, derived the results, implemented the simulations, and wrote the manuscript.

[2] N. Guo and V. Kostina. "Reliability function for streaming over a DMC with feedback". In: *IEEE International Symposium on Information Theory*. June 2022, pp. 3204–3209. DOI: 10.1109/ISIT50566.2022.9834852.
Nian Guo participated in the conception of the project, derived the results, and wrote the manuscript.

[3] N. Guo and V. Kostina. "Reliability function for streaming over a DMC with feedback". In: *submitted to IEEE Transactions on Information Theory* (June 2022). URL: https://arxiv.org/abs/2202.05770.
Nian Guo participated in the conception of the project, derived the results, implemented the simulations, and wrote the manuscript.

[4] N. Guo and V. Kostina. "Instantaneous SED coding over a DMC". In: *IEEE International Symposium on Information Theory*. July 2021, pp. 148–153. DOI: 10.1109/ISIT45174.2021.9518087.
Nian Guo participated in the conception of the project, derived the results, implemented the simulations, and wrote the manuscript. This paper was a finalist for IEEE Jack Keil Wolf Student Paper Award.

[5] N. Guo and V. Kostina. "Optimal causal rate-constrained sampling for a class of continuous Markov processes". In: *IEEE Transactions on Information Theory* 67.12 (Sept. 2021), pp. 7876–7890. DOI: 10.1109/TIT.2021.3114142.
Nian Guo participated in the conception of the project, derived the results, and wrote the manuscript.

[6] N. Guo and V. Kostina. "Optimal causal rate-constrained sampling for a class of continuous Markov processes". In: *IEEE International Symposium on Information Theory*. June 2020, pp. 2456–2461. DOI: 10.1109/ISIT44484.2020.9174333.
Nian Guo participated in the conception of the project, derived the results, and wrote the manuscript.

[7] N. Guo and V. Kostina. "Optimal causal rate-constrained sampling of the Wiener process". In: *Allerton Conference on Communication, Control, and Computing*. Sept. 2019, pp. 1090–1097. DOI: 10.1109/ALLERTON.2019.8919710.
Nian Guo participated in the conception of the project, derived the results, implemented the simulations, and wrote the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Chapter 1

INTRODUCTION

This thesis lifts the conventional assumption of non-causal (block) encoding in classical information theory that allows the encoder to know the entire message before the transmission. This thesis focuses on causal encoding: the encoder transmits information based only on the causally received data, while the data is still streaming in. This thesis derives fundamental limits of causal encoding and designs causal codes that achieve the limits.

The investigation of causal encoding for streaming data is of great practical interest. From smart transportation to industrial automation, we are now marching into an era where numerous devices connect to each other sharing instant information. The communication systems that enable such real-time information sharing are extremely dynamic, interactive, and delay-sensitive. For such real-time communication systems, source messages arrive at the encoder progressively in a streaming fashion. For example, the height and the speed data of an unmanned aerial vehicle stream into the encoder in real time. To transmit such a source, the idea of causal encoding naturally arises and fits the streaming nature of the data. In contrast, non-causal (block) encoding schemes need to buffer the arriving data into a block and then transmit the data block, causing significant performance degradation. For example, for transmitting 16 i.i.d. equiprobable bits that arrive at the encoder one by one over a binary symmetric channel with feedback at crossover probability 0.05 and error probability 10^{-6}, the rate empirically achieved by the best non-causal (block) encoding scheme preceded by a buffer is only 60% of the rate empirically achieved by the best causal encoding scheme (Fig. 4.9). As a consequence, re-evaluating the fundamental limits in the causal encoding setting and designing novel causal encoding schemes for streaming data that attain fundamental limits are critical.

This thesis investigates causal encoding of streaming data in three operational scenarios: causal frequency-constrained sampling (Chapter 2), causal lossy compressing (Chapter 3), and causal joint source-channel coding over a DMC with feedback (Chapter 4). For each scenario, we analyze the causal encoding of streaming data via the steps below:

- Set up an information-theoretic framework: define causal codes, specify the performance measure, and establish the tradeoff between the communication rate (i.e., how *fast* the transmission is) and the communication fidelity (i.e., how *reliable* the transmission is) for causal encoding of streaming data.
- Derive the fundamental limit, i.e., the best tradeoff between the communication rate and the communication fidelity, and find the causal codes that achieve the limit.
- Demonstrate the robustness of our causal codes in non-ideal systems and demonstrate the applicability of our causal codes to multiple-input multiple-output systems or control systems.

Next, for each of three operational scenarios, we briefly introduce the basic setup, the recent advancements in prior literature, and the unsolved problems in prior literature that are tackled in this thesis.

1.1 Causal frequency-constrained sampling

In Chapter 2, we consider a causal frequency-constrained sampling problem, which is also known as the optimal scheduling or the remote causal estimation problem. The basic task of causal frequency-constrained sampling is to sample a source process based on the causally observed process under a constraint on the average number of samples transmitted per second so that the source process can be approximated from the samples in real time. The problem of causal frequency-constrained sampling arises due to the development of the wireless sensor networks and network control systems of the Internet of Things. In such systems, nodes are spatially dispersed, communication delays between nodes are undesirable, and communication between nodes is a limited resource. A sampling frequency constraint is commonly used as a communication constraint between nodes [1, 2, 3, 4, 5, 6, 7, 8, 9] as it reflects the transmission rate of data packets.

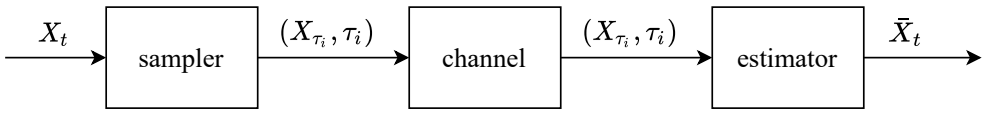


Figure 1.1: Causal frequency-constrained sampling.

In the basic setup of the causal frequency-constrained sampling (see Fig. 1.1), a sampler (i.e., a transmitting node) tracks a source process $\{X_t\}$ and decides a sequence of sampling times $\tau_1, \tau_2, \ldots$ based on the causally observed source process. At each sampling time τ_i, the sample X_{τ_i} and the sampling time τ_i are passed to an estimator without delay through a noiseless channel. At time t, the estimator (i.e., the receiving node) yields a real-time estimate $\bar{X}_t$ of the current value of the source process based on all the causally received samples and sampling times. The communication between the sampler and the estimator is subject to the sampling frequency constraint.

In this thesis, we measure the fundamental limit of the causal frequency-constrained sampling problem by a distortion-frequency function–the minimum end-to-end estimation distortion compatible with a given sampling frequency.

Causal sampling policies can be divided into two classes. One class comprises *signal-independent* (or signal-agnostic, time-triggered, deterministic) sampling policies whose sampling times do not depend on the source process, e.g., a uniform sampling policy that transmits samples at periodic times. The other class comprises *signal-dependent* (or signal-aware, event-triggered) sampling policies whose sampling times causally depend on the source process, e.g., a symmetric threshold sampling policy that transmits a sample if the process innovation falls outside a symmetric interval.

Causal signal-dependent sampling policies have attracted great research interest. One of the earliest works that analytically showed the advantage of signal-dependent sampling policy over signal-independent sampling policy was presented by Åström and Bernhardsson [10]. They considered two continuous-time scalar systems

$$\mathrm{d}X_t = U_t \mathrm{d}t + \mathrm{d}W_t, \tag{1.1}$$

$$\mathrm{d}X_t = aX_t \mathrm{d}t + U_t \mathrm{d}t + \mathrm{d}W_t, \tag{1.2}$$

driven by the Wiener process W_t and controlled by U_t. Under the same average control-injecting frequency (i.e., sampling frequency), they showed that injecting an impulse control that resets the state to zero once the state falls outside a symmetric interval leads to a smaller variance than injecting a minimum variance control [11] periodically. This work reveals that a good causal sampling policy should transmit *surprising*-enough samples that trigger certain uncommon events, while the deterministic time distance between two samples does not discriminate between surprising and non-surprising events and thus needs a higher sampling frequency to

achieve the same distortion. This work also motivates further research on signal-dependent sampling policies in the causal frequency-constrained sampling setting [1, 2, 3, 4, 5, 6, 7, 8, 9]. Causal frequency-constrained sampling policies have been studied for the i.i.d Gaussian random variables [1]; the Gauss-Markov process [2]; the partially observed Gauss-Markov process [3]; the first-order autoregressive Markov process $X_{t+1} = aX_t + V_t$ driven by an i.i.d. process $\{V_t\}$ with unimodal and even distribution [4][5]; the finite time-horizon Wiener and Ornstein-Uhlenbeck (OU) processes [6]; the infinite time-horizon multidimensional Wiener process [7]; the infinite-time horizon Wiener process [8], and the OU processes [9] with channel delay.

Remarkably, the causal sampling policies that achieve the best tradeoff between the sampling frequency and the estimation distortion in [1, 2, 3, 4, 5, 6, 7, 8, 9] all admit a symmetric structure, that is, a sample is taken if the process innovation crosses either of two symmetric thresholds. For example, in the infinite time horizon, under the average sampling frequency F and the mean-square error (MSE) distortion measure, the minimum MSE of causal sampling the Wiener process $\{W_t\}_{t=0}^{\infty}$ is achieved by [7]

$$\tau_{i+1} = \inf \left\{ t \geq \tau_i \colon |W_t - W_{\tau_i}| \geq \sqrt{\frac{1}{F}} \right\}. \tag{1.3}$$

We notice that all source processes considered in [1, 2, 3, 4, 5, 6, 7, 8, 9] have even and unimodal pdf and satisfy the Markov property. In Chapter 2, we abstract their similarities and show that for a class of continuous Markov processes satisfying symmetry and regularity conditions, the optimal causal frequency-constrained sampling policy is a symmetric threshold sampling policy. This extends the prior results in [1, 2, 3, 4, 5, 6, 7, 8, 9] to a wider class of stochastic processes.

While a sampling time carries information as it signifies an occurrence of an event, *silence* also carries information. Here, the silence refers to the duration in-between two consecutive sampling times. Take the symmetric threshold sampling policy in (1.3) as an example. The silence due to the next sample not having been taken implies that the source process still belongs to the symmetric interval. Yet, in [6, 7, 8, 9], a common assumption is that the estimator ignores the information carried by the *silence* and purely depends on the past samples and sampling times. The assumption allows one to solve the problem by applying Snell's envelope, a classical method for solving an optimal stopping problem that requires solving a stochastic differential equation (SDE) [6, 7, 8, 9]. In Chapter 2.3, we use a different set of

tools, namely, the majorization theory and the real induction, to find the fundamental limit-achieving causal sampling policies without using the simplifying assumption; we show that the silence is useless for a minimum mean-square error estimator due to the symmetric structure of the optimal sampling policy, confirming the assumption in prior literature.

In Chapter 2.4, we extend the point-to-point system to an n-sampler n-estimator system, where the k-th estimator causally listens to the first k samplers, $k = 1, 2, \ldots, n$, and each sampler is subject to a sampling frequency constraint. For such a system, we show that implementing symmetric threshold sampling policies at n samplers attains the minimum real-time estimation distortions at all n estimators. In Chapter 2.5, lifting the assumption that the channel in Fig. 1.1 is ideal, we show that causal frequency-constrained sampling policies that attain the minimum distortion for a channel with delay and a packet-drop channel remain symmetric threshold sampling policies.

1.2 Causal lossy data compression

While the prior works [1, 2, 3, 4, 5, 6, 7, 8, 9] on causal frequency-constrained sampling did not take the quantization effect into consideration, in almost all modern communications, a real-valued sample carrying an infinite amount of information is quantized before the transmission. In Chapter 3, we replace the sampling frequency constraint in Chapter 2 by a bitrate constraint, routinely considered in information theory, and we consider causal rate-constrained sampling–a causal lossy data compression problem for continuous-time processes. We first review the basics of causal lossy compression for discrete-time processes, we then set up the problem of causal lossy data compression for continuous-time processes, and we finally introduce the quantized event-triggered control as an important application.

Causal lossy data compression for discrete-time processes

The basic task of causal lossy data compression is to compress a discrete-time source sequence only based on the causally received source symbols under a rate constraint (bits per channel use) so that the source sequence can be reproduced via the causally received codewords with the minimum distortion.

In the basic setup of the causal lossy data compression, the symbols of a source sequence S^n (to be compressed) arrive at the encoder one by one at times $t = 1, 2, \ldots, n$. The alphabet of the source sequence $\mathcal{S}^n$, the alphabet of the decoded sequence $\hat{\mathcal{S}}^n$, the probability distribution of the source P_{S^n}, and a distortion measure

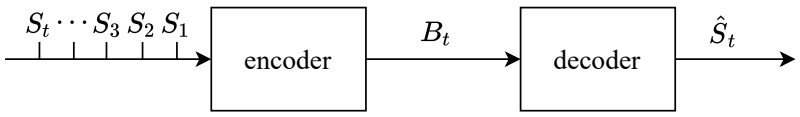


Figure 1.2: Causal compressing of streaming source symbols.

$\mathsf{d}_n \colon \mathcal{S}^n \times \hat{\mathcal{S}}^n \to \mathbb{R}_+$, which evaluates the fidelity of the estimation, are given. A zero-delay source code (see Fig. 1.2) operates as follows: at time t, the encoder uses the causally received symbols S^t as well as the past codewords B^{t-1} to form a new codeword B_t; the decoder uses the causally received codewords B^t as well as the past estimate $\hat{S}^{t-1}$ to form an estimate $\hat{S}_t$ of the source symbol S_t. Let L_t be the length of the codeword B_t. The rate of a zero-delay source code is often measured by the average number of bits transmitted per source symbol in the limit of large source length, i.e., by $R \triangleq \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{t=1}^n L_t\right]$.

The fundamental limit of the causal lossy data compression is given by the minimum achievable rate R compatible with a distortion D in the limit of large source length. While Shannon set forth the concept of rate-distortion tradeoff in the classical block encoding context [12, 13], the rate-distortion tradeoffs in the causal encoding setting are often measured by the operational causal rate-distortion function $R^{\mathrm{op}}(D)$, that is, the minimum R compatible with distortion D achieved by a zero-delay code; and its information-theoretic counterpart–the informational causal rate-distortion function $R^{\mathrm{it}}(D)$. The informational causal rate-distortion function, first introduced by Gorbunov and Pinsker [14], replaces the operational rate R by the normalized directed information [15] between the source sequence and the decoded sequence

$$\frac{1}{n} I(S^n \to \hat{S}^n) = \frac{1}{n}\sum_{t=1}^{n} I(S^t; \hat{S}_t | \hat{S}^{t-1}), \tag{1.4}$$

and serves as a lower bound to $R^{\mathrm{op}}(D)$.

Causal rate-constrained sampling

While most existing works on causal rate-distortion tradeoffs considered discrete-time source processes [14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25], in Chapter 3, we establish the problem of causal lossy compression of continuous-time processes, termed *causal rate-constrained sampling*.

In the basic setup of causal rate-constrained sampling (see Fig. 1.3), the encoder

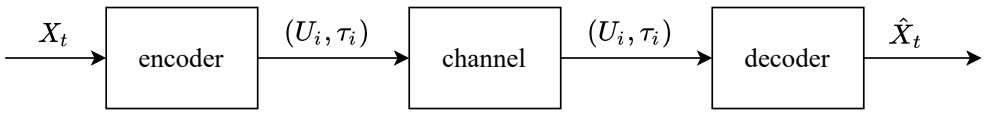


Figure 1.3: Causal compressing of a streaming continuous-time process.

observes the continuous-time source process $\{X_t\}$ and causally decides when and what to transmit about the process. The encoder consists of a causal sampling policy that decides the sampling times $\tau_1, \tau_2, \ldots$ and a causal compressing policy that decides the codewords $U_1, U_2, \ldots$ to transmit at each sampling time. The codeword U_i is passed to the decoder without delay through a noiseless channel. The real-valued sampling time τ_i is also immediately known by the decoder since the channel is delay-free. At time t, the decoder yields a real-time estimate $\hat{X}_t$ of the current value of the source process based on all the causally received codewords and sampling times.

The fundamental limit of the causal rate-constrained sampling problem is measured by a distortion-rate function, which quantifies the minimum end-to-end estimation distortion compatible with a communication rate (i.e., the average number of bits transmitted per second). The goal is to find the causal codes that achieve the distortion-rate function.

In Chapter 3.3, for a class of continuous Markov processes satisfying symmetry and regularity conditions, we present the causal code, termed the sign-of-innovation (SOI) code, that achieves the best tradeoff between the rate and the distortion. It transmits a bit representing the sign of the process innovation once the process innovation crosses either of two symmetric thresholds. Since a transmission occurs only if this event occurs, the sampling times carry information. Indeed, this is a significant difference between the classical causal lossy compressing [14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25] and our causal rate-constrained sampling: the sampling times are also parameters to be optimized and are allowed to causally depend on the source process. Due to the use of timing information, the distortion achieved by our SOI code can even *beat* (be smaller than) that achieved by the best non-causal source codes for the Wiener process (Chapter 3.8). Without using the timing information and restricting the sampling times to be deterministic, we show in Chapter 3.4 that the uniform sampling policy achieves the best rate-distortion tradeoff for the Wiener process, but the achieved distortion is 5-fold worse than that achieved by the SOI

code.

In Chapter 3.6, we re-consider the n-sampler n-estimator system in Chapter 2.4 with the n frequency constraints replaced by n bitrate constraints. For such a system, we show that appending an SOI compressor to each sampler gives n causal encoders that attain the minimum estimation distortions at all n decoders. In Chapter 3.7, we show that our SOI code is resilient to channel delay and noise.

Quantized event-triggered control

One important application of causal rate-constrained sampling schemes is quantized event-triggered control. The problem arises since the real-valued state of the plant is often quantized by the observer before it is received by the controller, and since the event-triggered (i.e., signal-dependent) control demonstrates a better performance than the time-triggered (i.e., signal-independent) control, e.g., [10, 26, 27]. Although the quantized event-triggered control is in a closed loop and the causal rate-constrained sampling is in an open loop, they are related in the sense that the controller can form a better control signal if it can causally estimate the plant more accurately.

In the basic setup of quantized event-triggered control (see Fig. 1.4), the encoder observes the plant Y_t driven by the noise X_t and causally decides the sampling times $\tau_1, \tau_2, \dots$ and the codewords $U_1, U_2, \dots$ to transmit at each sampling time. At each sampling time, the codeword and the sampling time are passed to the decoder via a channel. The decoder uses the causally received codewords and sampling times to form a control signal Z_t, aiming to stabilize the system.

Quantized event-triggered control schemes have been designed to stabilize different systems, however, existing works [28, 29, 30, 26, 31, 32, 33, 34, 27, 35] did not consider the optimality of the proposed schemes, that is, which control scheme can minimize the deviation of the plant to zero with the minimum possible rate remained unknown. In Chapter 3.5, we show that our SOI code, which achieves the best rate-distortion tradeoff in the causal rate-constrained sampling setting, also applies to quantized event-triggered control under regularity conditions. For the Wiener process disturbance, our SOI control scheme reduces to Åström and Bernhardsson [10]'s event-triggered control scheme.

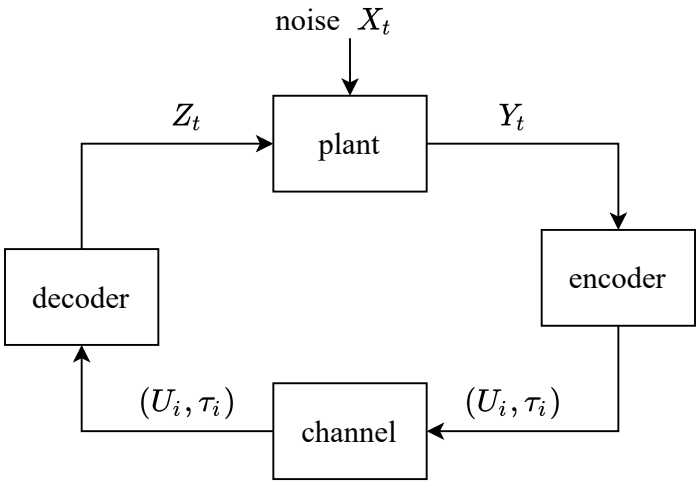


Figure 1.4: System model of quantized event-triggered control.

1.3 Causal joint source-channel coding with feedback

While in the basic setups of the causal frequency-constrained sampling and the causal rate-constrained sampling (Chapters 2 and 3), the channel is noiseless, in Chapter 4, we proceed to consider transmitting a sequence of streaming symbols over a noisy channel with feedback via causal joint source-channel coding (JSCC). We first review the basics of the classical non-causal channel coding with feedback. We then introduce our novel setup–causal JSCC with feedback.

Non-causal channel coding with feedback

The basic task of non-causal (block) channel coding with feedback is to transmit equiprobable messages that are fully accessible to the encoder before the transmission over a noisy channel with feedback, so that the transmitted message can be identified at the decoder. The problem arises since the feedback, though unable to increase the capacity of a memoryless channel [36], can simplify the design of capacity-achieving codes [37, 38, 39] and improve achievable delay-reliability tradeoffs [40, 41]. While the channel feedback can be used in a very limited way, e.g., the feedback only contains one bit and is transmitted only once [41], here we take the maximum advantage of the feedback by assuming that the feedback link noiselessly communicates the full channel output to the encoder at every time.

In the basic setup of non-causal channel coding with feedback (see Fig. 1.5), a variable-length channel code with block encoding operates as follows. At each time

t, the encoder uses the message $S \in [M]$ as well as the past channel outputs Y^{t-1} to form a channel input X_t; upon receiving Y_t, the decoder uses all the causally received channel outputs Y^t to adjust its belief about the message S and to decide a stopping time η to output its estimate $\hat{S} \in [M]$. The rate of the code is $R = \frac{\log M}{\mathbb{E}[\eta]}$ nats per channel use.

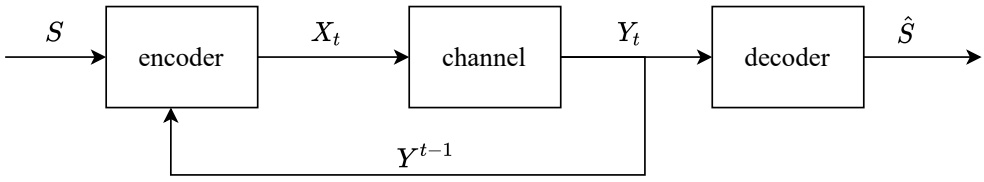


Figure 1.5: Communication over a channel with full feedback.

The classical fundamental limits of non-causal channel coding with feedback are the channel capacity and the reliability function. The channel capacity C represents the best (maximum) achievable rate below which the error probability vanishes in the limit of large delay (i.e., blocklength). The underlying principle behind capacity-achieving block encoding schemes with feedback [37, 42, 38, 39, 40, 43, 44, 45, 46], termed posterior matching [39], is to match the channel input distribution to the capacity-achieving input distribution using the posterior of the message and to transmit information that the decoder has not yet received. The reliability function (a.k.a. optimal error exponent) measures the best delay-reliability tradeoff, namely, it represents the maximum exponentially decaying rate of the error probability at code rate $R < C$ as the blocklength is taken to infinity. The reliability function for transmitting equiprobable messages over a DMC with full feedback using a variable-length channel code with block encoding is first shown by Burnashev [40]:

$$E(R) = C_1 \left(1 - \frac{R}{C}\right), \tag{1.5}$$

where C_1 is the maximum Kullback–Leibler divergence between channel transition probabilities of the DMC, C is the channel capacity, and R is the rate. Variable-length channel codes with block encoding that achieve Burnashev's reliability function have been proposed in [40, 43, 44, 45, 46].

Causal joint source-channel coding with feedback

While the classical non-causal channel coding schemes with feedback [37, 42, 38, 39, 40, 43, 44, 45, 47, 46] assume that the source symbols are equiprobable and

are entirely known by the encoder before the transmission, in Chapter 4, we consider a streaming source, which emits non-equiprobably distributed source symbols $S_1, S_2, \ldots$ at a sequence of times $t_1 \leq t_2 \leq \ldots$. The basic task of causal JSCC with feedback is to transmit a streaming source, based only on the causally received source symbols and channel feedback, over a noisy channel, so that the streaming source can be identified at the decoder.

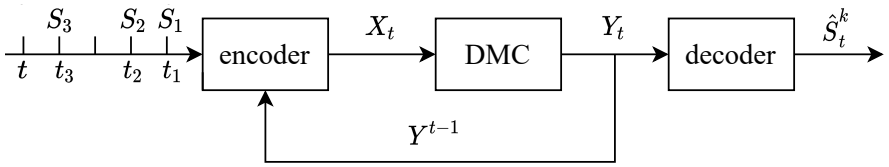


Figure 1.6: Real-time feedback communication system with a streaming source.

In the basic setup of causal JSCC with feedback (see Fig. 1.6), a code with instantaneous encoding operates as follows. At each time t, the encoder uses the causally received source symbols and channel feedback Y^{t-1} to form a channel input X_t; the decoder uses the causally received channel outputs Y^t to adjust its belief about the streaming source and to form an estimate $\hat{S}_t^k$ of k source symbols.

In Chapter 4, we mainly focus on a class of codes with instantaneous encoding that aims to recover a fixed number of source symbols k and outputs its estimate at a stopping time η_k adapted to the filtration generated by the channel outputs. The rate of the code is $R = \frac{k}{\mathbb{E}[\eta_k]}$ symbols per channel use.

Similar to the works [40, 43, 44, 45, 46] on non-causal channel coding with feedback, this thesis measures the fundamental limit of causal JSCC with feedback by the reliability function $E(R)$–the maximum exponentially decaying rate of the error probability compatible with rate R achieved by a sequence of codes with instantaneous encoding for transmitting k streaming symbols over a DMC with feedback as $k \to \infty$. The JSCC reliability function for streaming is an appropriate performance measure for causal JSCC since it displays the best delay-reliability tradeoff, critical for time-sensitive applications. The goal is to find a sequence of codes with instantaneous encoding that achieves the JSCC reliability function for streaming.

In Chapter 4.4, we show the JSCC reliability function for streaming, which extends Burnashev's reliability function (1.5) to JSCC and to the streaming sources. Surprisingly, the JSCC reliability function for streaming is equal to the JSCC reliability function for a classical fully accessible source, whose symbols are entirely known

by the encoder before the transmission. This means that revealing symbols only progressively to the encoder does not incur penalties on the reliability function. The achievability of the JSCC reliability function for streaming is supplied by our instantaneous encoding phase (Chapter 4.3). It operates during the symbol arriving period and serves as a building block that can be inserted before any reliability function-achieving block encoding scheme to attain the JSCC reliability function for streaming.

Although the class of codes with instantaneous encoding that we focus on in Chapter 4 has an appealing property–it overcomes the detrimental effect due to the streaming nature of the source on the reliability function–few works have designed such codes. Antonini et al.'s [47] proposed a code with instantaneous encoding that transmits k streaming bits over a binary symmetric channel with feedback, but they did not provide analytical results on the achievable rate and error exponent.

Yet, another class of codes with instantaneous encoding has already been investigated in the field of control [48, 49, 50, 51], termed *anytime* codes. The class of anytime codes is slightly different from the class of codes with instantaneous encoding that we focus on in that an anytime code can choose to decode any number of source symbols k at any time t with an error probability that decays exponentially with decoding delay $t - t_k$. Sahai and Mitter [48] showed that an anytime code can be used to stabilize a discrete-time unstable scalar linear system with bounded noise over a noisy channel with feedback. To do so, an anytime encoder is embedded in the observer and treats the evolving plant as a streaming source; an anytime decoder is embedded in the controller so that the control signals are formed based on decoder's causal estimations. In Chapter 4.5, we design an instantaneous small-enough difference (SED) code for symmetric binary-input DMCs. It empirically behaves like an anytime code, thus it can be used for system stabilization. Interestingly, a sequence of instantaneous SED codes for transmitting k symbols also achieves the JSCC reliability function for streaming as $k \to \infty$. We design low-complexity algorithms to implement our instantaneous encoding phase and our instantaneous SED code in Chapter 4.7.

In most existing works on causal encoding of a streaming source [47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58] with or without feedback, the decoder is assumed to know the exact symbol arriving times. This assumption becomes unrealistic if a streaming source emits symbols at random times. In Chapter 4.6, assuming that the decoder only knows the symbol arriving distribution rather than the exact symbol

arriving times of a streaming source, we generalize our instantaneous SED code to transmit such a source and derive the JSCC reliability function for such a source.

In Chapter 4.9, we consider a class of DMCs whose transition probability matrix contains zeros, meaning that some channel output is not reachable from a channel input, e.g., a binary erasure channel belongs to this class. Taking advantage of this property of the DMC, we design a sequence of codes with instantaneous encoding that attains exactly zero error at any rate asymptotically below Shannon's JSCC limit. This extends Burnashev's zero error code [40, Sec. 6] to JSCC and to the streaming scenarios.

Chapter 2

CAUSAL FREQUENCY-CONSTRAINED SAMPLING

2.1 Introduction

In this chapter, we consider the following communication scenario: a sampler observes a stochastic process and causally decides when to sample it under a constraint on the expected number of samples transmitted per second; an estimator uses the causally received samples to approximate the process in real time; the channel is delay-free and noiseless. As we have briefly discussed in Chapter 1.1, we refer to this scenario as *causal frequency-constrained sampling*.

For a class of continuous Markov processes (e.g., Wiener process, continuous Lévy process, and Ornstein-Uhlenbeck process) satisfying symmetry and regularity conditions, we find the optimal causal sampling and estimating policies that minimize the end-to-end estimation mean-square error (MSE) under a frequency constraint. We show that the optimal sampling policy transmits a real-valued sample once the process innovation passes one of two symmetric thresholds. The optimal estimator only uses the last sample and the last sampling time to decide the running estimate of the current process, until the next sample arrives.

Extending the single-sampler single-estimator system to an n-sampler n-estimator system, we consider a scenario termed *successive refinement via causal frequency-constrained sampling*: n samplers simultaneously track the source process and causally sample it under a sequence of frequency constraints; the k-th estimator receives the sample and the sampling times generated by the first k samplers, $k = 1, 2, \ldots, n$; n estimators use the causally received information to approximate the process in real time. We design n causal sampling policies that successively refine the estimation MSEs and attain the best distortion-frequency tradeoffs at all n estimators.

Lifting the assumption that the channel is perfect, we show the optimal causal frequency-constrained sampling policy of a continuous Lévy process for a channel with a delay and for a packet-drop channel. For a channel with a fixed delay, we show that the optimal causal sampling policy for a delay-free channel remains optimal. For a packet-drop channel that either transmits a sample noiselessly or drops it, we show that if the last sample is successfully delivered, then optimal causal sampling

policy follows a symmetric threshold policy; otherwise, it retransmits the dropped sample immediately.

Sections 2.2–2.3, which formulate the causal frequency-constrained sampling problem and present the optimal causal sampling and estimating policies, appear in the research papers [59, 60]. Sections 2.4–2.5, which investigate the best distortion-frequency tradeoffs for successive refinement and imperfect channels, appear for the first time.

Prior work

In wireless sensor networks and network control systems of the Internet of Things, nodes are spatially dispersed, communication between nodes is a limited resource, and delays are undesirable. We study the fundamental limits of the communication scenario: a transmitting node (sampler) observes a stochastic process, e.g., location, speed, temperature, and wants to communicate it in real-time to the receiving node (estimator); the receiving node (estimator) aims to recover the process in real time using the causally received information.

Related work includes [1, 2, 3, 4, 5, 6, 7, 8, 9], where it is assumed that the encoder transmits real-valued samples of the source process and that the communication is subject to a sampling frequency constraint or a transmission cost. Finding sampling policies at the encoder and estimation policies at the estimator to jointly minimize the end-to-end distortion under transmission constraints falls into the realm of optimal scheduling and remote sequential estimation problems. The causal sampling and estimation policies that achieve the optimal tradeoff between the sampling frequency and the distortion have been studied for the following *discrete-time processes*: the i.i.d process [1]; the Gauss-Markov process [2]; the partially observed Gauss-Markov process [3]; and, the first-order autoregressive Markov process $X_{t+1} = aX_t + V_t$ driven by an i.i.d. process $\{V_t\}$ with unimodal and even distribution [4][5]. Imer and Başar [1] considered causal estimation of i.i.d. processes under MSE and the constraint on the number of transmissions over a finite time horizon, and showed that the time-varying symmetric threshold sampling policy is optimal for i.i.d. Gaussian processes. Lipsa and Martins [2] proved that a time-varying symmetric threshold policy and a Kalman-like filter jointly minimize a discounted cost function consisting of MSE and a communication cost, for scalar discrete-time Gauss-Markov (GM) processes over a finite time horizon. For partially observed discrete-time GM processes, Wu et al. [3] fixed an event-triggered policy, where

the sampler transmits only if the L-infinity norm of the measurement innovation exceeds a constant, and derived both the accurate and an approximate minimum MSE (MMSE) estimator to combine with that sampling policy. Chakravorty and Mahajan [4] showed that a threshold sampling policy with two constant thresholds and an innovation-based filter jointly minimize a discounted cost function consisting of the MSE and a transmission cost in the infinite time horizon. Molin and Hirche [5] proposed an iterative algorithm to find the sampling policy that achieves the minimum of a cost function consisting of a linear combination of the MSE and the transmission cost in the finite time horizon, and showed that the algorithm converges to a two-threshold policy.

The optimal sampling policies for some *continuous-time processes* have also been studied: first-order stochastic systems with a Wiener process disturbance [10]; the finite time-horizon Wiener and Ornstein-Uhlenbeck (OU) processes [6]; the infinite time-horizon multidimensional Wiener process [7]; the infinite-time horizon Wiener process [8], and the OU processes [9] with channel delay. Åström and Bernhardsson [10] compared uniform and symmetric threshold sampling policies in first-order stochastic systems with a Wiener process disturbance. They showed that the symmetric threshold sampling policy gives a lower distortion than the uniform sampling under the same average sampling frequency. Rabi et al. [6] formulated the problem of causal estimation of the Wiener process and the OU under the constraint on the number of transmissions over a finite time horizon as an optimal stopping time problem. Rabi et al. [6] showed that the optimal deterministic sampling policy and the optimal event-triggered sampling policy for the Wiener process are a uniform policy and a symmetric threshold policy, respectively. Nar and Başar [7] extended the optimal stopping time problem in [6] to the multidimensional Wiener process, and proved that a symmetric threshold policy remains optimal over both finite and infinite time horizons. In particular, Nar and Başar [7] showed that the optimal threshold over the infinite horizon is a constant depending on the average sampling frequency. Sun et al. [8] proved that a symmetric threshold policy remains optimal even when the samples of the Wiener process experience an i.i.d. random transmission delay, but the threshold depends on the distribution of channel delay and is different from the one in [7]. The optimal causal sampling policies for the Wiener and the OU processes determined in [6, 7, 8, 9] are threshold sampling policies, whose thresholds are obtained by solving optimal stopping time problems via Snell's envelope. The proofs in [6, 7, 8, 9] rely on a conjecture about the form of the MMSE estimating policy, implying that the causal sampling policies in [6, 7, 8, 9]

are optimal with respect to the conjectured estimating policy, rather than the optimal estimating policy. Namely, Rabi et al. [6] conjectured that the MMSE estimating policy under the optimal sampling policy is equal to the MMSE estimating policy under deterministic (process-independent) sampling policies without a proof. Nar and Başar [7] arrived at the MMSE estimating policy for the Wiener process by referring to the results in [61], where the stochastic processes considered in [61] are in discrete-time and the increments of the discrete-time process are assumed to have finite support. Yet, the Wiener process is a continuous-time process with Gaussian increments having infinite support. Sun et al. [8] and Ornee and Sun [9] assumed that the estimating policy ignores the implied knowledge when no samples are received at the estimator, neglecting the possible influence of the sampling policy on the estimating policy. Nonparametric estimation of Lévy processes from uniform non-causal samples has been studied in [62, 63].

In contrast to the scenarios in [1, 2, 3, 4, 5, 6, 7, 8, 9], where the communication channel is assumed to be noiseless (perhaps with delays [8, 9]), [64, 65, 66] consider noisy communication channels. Using dynamic programming, Gao et al. in [64] derived the optimal sampling, encoding, and decoding policies for the event-triggered sampling of an i.i.d. Laplacian source with subsequent transmission over a channel with a Gamma additive noise, under an average power constraint. For discrete-time first-order autoregressive Markov processes considered in [4, 5], Ren et al. [65] introduced a fading channel between the sampler and the estimator, where a successful transmission depends on both the channel gains and the transmission power, and found the optimal encoding and decoding policies that minimize an infinite horizon cost function combining the MSE and the power usage. For discrete-time first-order autoregressive sources considered in [4, 5, 65], Chakravorty and Mahajan [66] further proved that the optimal estimation policy is a Kalman-like filter and that the optimal sampling policy is symmetric threshold policy when the communication channel is a packet-drop channel with Markovian states.

Chapter organization and contribution

In Section 2.2, we formulate a single-sampler single-estimator (point-to-point) causal frequency-constrained sampling problem, define causal frequency-constrained codes, and define the distortion-frequency function $\underline{D}(F)$ to quantify the tradeoffs between frequency F and MSE d .

In Section 2.3, we present the causal sampling policy that achieves the distortion-

frequency function for a class of continuous Markov processes (e.g., the Wiener process, Lévy process, and the OU process). We show that the optimal causal sampling policy transmits a sample of the source process if the process innovation exceeds one of two symmetric thresholds. Compared to the previous work on sampling of continuous-time processes [6, 7, 8, 9], our results apply to a wider class of processes, namely, the processes satisfying (P.1)–(P.3) in Section 2.2. Furthermore, we confirm the validity of the conjecture on the MMSE estimating policy in [6, 7]. To do so, we use a set of tools that differs from that in [6, 7]: where [6, 7] use Snell's envelope to find the optimal sampling policy under the conjecture on the form of the MMSE estimating policy, we apply majorization theory and real induction to find the jointly optimal sampling and estimating policies.

In Section 2.4, we extend the point-to-point communication system in Sections 2.2–2.3 to an n-sampler n-estimator system and consider the successive refinement in the causal frequency-constrained sampling setting. This problem arises if an estimator can choose the number of samplers it listens to depending on the accuracy of the estimate it aims to attain. This problem parallels the classical successive refinement problem [67], which is a data compression problem: an encoder successively compresses a source in n stages under n bitrate constraints R^n; a decoder refines the source estimate as it receives more information bits, so that the code pair achieves the distortion-rate function $D(R_1), D(R_1 + R_2), \ldots, D\left(\sum_{k=1}^{n} R_k\right)$ at each stage. In Section 2.4, we replace the encoder and the decoder that operate in n stages by n samplers and n estimators, where the k-th estimator receives samples and sampling times from the first k samplers, $k = 1, 2, \ldots, n$; we replace the sequence of rate constraints R^n by a sequence of frequency constraints F^n; we replace the distortion-rate function by the distortion-frequency function $\underline{D}(F)$. We show that n causal sampling policies that achieve the distortion-frequency functions $\underline{D}(F_1), \underline{D}(F_1 + F_2), \ldots \underline{D}(\sum_{k=1}^{n} F_k)$ at n estimators cooperate with each other, so that the optimal causal sampling policies at the first k samplers can be viewed as a single sampling policy that operates under frequency $\sum_{i=1}^{k} F_i$.

In Section 2.5, we replace the perfect channel in Sections 2.2–2.3 by imperfect ones and show how the distortion-frequency tradeoffs for a continuous Lévy process are affected. For a channel with a fixed delay, we show that the optimal causal sampling policy for a delay-free channel remains optimal. For a packet-drop channel with 1-bit feedback indicating whether or not the sample is dropped, we show that the optimal causal sampling policy operates as follows. If the last sample is not

dropped, it follows a symmetric threshold sampling policy; otherwise, it retransmits the dropped sample immediately. This sampling policy transmits new samples at a lower frequency than the optimal sampling policy for a noiseless channel, in exchange for the retransmission opportunities of the dropped samples.

Notations

We denote by $\{X_t\}_{t=s}^{r}$ the portion of the stochastic process within the time interval $[s, r]$, and denote by $\{X_t\}_{t>s}^{r}$ the portion of the stochastic process within the time interval $(s, r]$. For a possibly infinite sequence $x = \{x_1, x_2, \dots\}$, we write $x^i = \{x_1, x_2, \dots, x_i\}$ to denote the vector of its first i elements. For a continuous random variable X, we denote its pdf by f_X. We denote by $\mathrm{Supp}(f_X) \triangleq \{x\colon f_X(x) > 0\}$ the support of f_X. We use $\sigma(\cdot)$ to denote the σ-algebra of its argument. We use $X \leftarrow Y$ to represent a substitution of X by Y.

2.2 Problem statement

Consider the single-sampler single-estimator system in Fig. 2.1. A source outputs a real-valued continuous-time stochastic process $\{X_t\}_{t=0}^{T}$ with state space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, where $\mathcal{B}_{\mathbb{R}}$ is the Borel σ-algebra on $\mathbb{R}$.

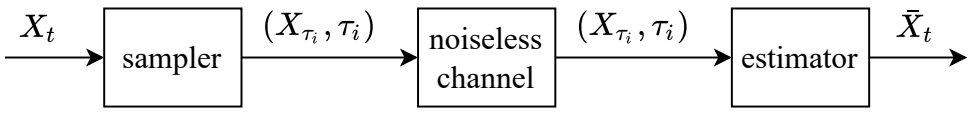


Figure 2.1: System Model. Sampling times τ_i, $i = 1, 2, \dots$ are determined by the sampling policies.

A sampler tracks the source process $\{X_t\}_{t=0}^{T}$ and causally decides to sample it at a sequence of stopping times

$$0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_N \le T \tag{2.1}$$

that are decided by a causal sampling policy. Thus, the total number of time stamps N can be random. The time horizon T can either be finite or infinite. At time τ_i, the sampler passes sample X_{τ_i} to the estimator without delay through a noiseless channel. At time t, $t \in [\tau_i, \tau_{i+1})$, the estimator estimates the source process X_t, yielding $\bar{X}_t$, based on all the received samples and the sampling time stamps, i.e., (X_{τ_j}, τ_j), $j = 1, 2, \dots, i$. Note that the sampler and the estimator can leverage

the timing information for free due to the clock synchronization and the zero-delay channel.

We formally define causal sampling and estimating policies.

Definition 1 ((F, d, T) causal frequency-constrained code)**.** *A time horizon-T causal frequency-constrained code for the stochastic process $\{X_t\}_{t=0}^T$ is a pair of causal sampling and estimating policies:*

1. *The causal* sampling *policy is a collection of stopping times $\tau_1, \tau_2, \ldots$ (2.1) adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^T$ at which samples are generated.*
2. *Given a causal sampling policy, the real-valued samples $\{X_{\tau_j}\}_{j=1}^i$ and sampling time stamps τ^i, the MMSE estimating policy is*

$$\bar{X}_t \triangleq \mathbb{E}[X_t | \{X_{\tau_j}\}_{j=1}^i, \tau^i, t < \tau_{i+1}],\ t \in [\tau_i, \tau_{i+1}). \tag{2.2}$$

In an (F, d, T) code, the average sampling frequency must satisfy

$$\frac{\mathbb{E}[N]}{T} \leq F\ (\textit{samples per sec}),\ (T < \infty), \tag{2.3a}$$

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N]}{T} \leq F\ (\textit{samples per sec}),\ (T = \infty), \tag{2.3b}$$

where N is the total number of stopping times in (2.1), while the MSE must satisfy

$$\frac{1}{T}\mathbb{E}\left[\int_0^T (X_t - \bar{X}_t)^2\right] \leq d,\ (T < \infty), \tag{2.4a}$$

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T (X_t - \bar{X}_t)^2\right] \leq d,\ (T = \infty). \tag{2.4b}$$

Allowing more freedom in designing the estimating policy will not lead to a lower MSE, since (2.2) is the MMSE estimator.

We present the assumptions on the source process and on the causal sampling policies below. Throughout, we impose the following assumptions on the source process $\{X_t\}_{t=0}^T$. Let $\{\mathcal{F}_t\}_{t=0}^T$ be the filtration generated by $\{X_t\}_{t=0}^T$.

(P.1) *(Strong Markov property)* $\{X_t\}_{t=0}^T$ satisfies the strong Markov property: For all almost surely finite stopping times $\tau \in [0, T]$ and all $t \in [0, T - \tau]$, $X_{t+\tau}$ is conditionally independent of $\mathcal{F}_\tau$ given X_τ.

(P.2) *(Continuous paths)* $\{X_t\}_{t=0}^T$ has continuous paths: X_t is almost surely continuous in t.

(P.3) *(Mean-square residual error properties)* For all almost surely finite stopping times $\tau \in [0,T]$ and all $t \in [\tau,T]$, the mean-square residual error $\tilde{X}_t = X_t - \mathbb{E}[X_t|X_\tau,\tau]$ satisfies:

(P.3-a) $\tilde{X}_t$ is independent of $\mathcal{F}_\tau$ and $\tilde{X}_t$ has the Markov property, i.e., for all $r \in [\tau,t]$, $\tilde{X}_t$ is conditionally independent of $\mathcal{F}_r$ given $\tilde{X}_r$.

(P.3-b) $\tilde{X}_t$ can be expressed as

$$\tilde{X}_t = q_t(s)\tilde{X}_s + R_t(s,\tau), \tag{2.5}$$

where $s \in [\tau,t]$, $q_t(s)$ is a deterministic function of (t,s), and $R_t(s,\tau)$ is a random process with continuous paths, i.e., $R_t(s,\tau)$ is almost surely continuous in t. Furthermore, the random variable $R_t(s,\tau)$ has an even and quasi-concave pdf, and $q_t(t) = 1$, $R_t(t,\tau) = 0$.

We assume that the initial state $X_0 = 0$ at time $\tau_0 \triangleq 0$ is known both at the sampler and the estimator. For example, the Wiener process satisfies (P.1)–(P.3), whose definition is given below.

Definition 2 (Wiener process, e.g., [68])**.** *A Wiener process $\{W_t\}_{t\geq 0}$ is a stochastic process characterized by the following three properties:*

- *For all non-negative s and t, W_s and $W_{s+t} - W_t$ have the same distribution ($W_0 = 0$);*
- *The increments $W_{t_i} - W_{s_i}$ $(i \geq 1)$ are independent whenever the intervals $(s_i, t_i]$ are disjoint;*
- *The random variable W_t follows the Gaussian distribution $\mathcal{N}(0,t)$.*

Any stochastic process of the form $X_t = g_1(t)W_{g_2(t)} + g_3(t)$ satisfies (P.1)–(P.3), where g_1, g_2, g_3 are continuous deterministic functions of the time t, and g_2 is positive and non-decreasing in t. The parameters in (2.5) for this example process are $q_t(s) = \frac{g_1(t)}{g_1(s)}$ and $R_t(s,\tau) = g_1(t)W_{g_2(t)-g_2(s)}$. The Wiener process, the Ornstein-Uhlenbeck (OU) process, and the continuous Lévy processes are special cases of this form. These processes are widely used in financial mathematics and physics. There are also other stochastic processes satisfying (P.1)–(P.3), e.g., $X_t = W_{t+c_1} + c_2 W_t$, where $c_1, c_2 \in \mathbb{R}$, which is expressed by (2.5) with $q_t(s) = 1$, $R_t(s,\tau) = (1+c_1)W_{t-s}$.

Definition 3 (time-homogeneous process)**.** *We say that a stochastic process* $\{X_t\}_{t=0}^T$ *is time-homogeneous, if for a stopping time* $\tau \in [0, T]$ *and a constant* $s \in [0, T-\tau]$, $X_{s+\tau} - \mathbb{E}[X_{s+\tau}|X_\tau]$ *follows a distribution that only depends on* s.

We focus on causal sampling policies that satisfy the following assumptions.

(S.1) The sampling interval between any two consecutive stopping times, $\tau_{i+1} - \tau_i$, satisfies

$$\mathbb{E}[\tau_{i+1} - \tau_i] < \infty,\ i = 0, 1, \ldots, \tag{2.6}$$

and the MSE within each interval satisfies

$$\mathbb{E}\left[\int_{\tau_i}^{\tau_{i+1}} (X_t - \bar{X}_t)^2 \mathrm{d}t\right] < \infty,\ i = 0, 1, \ldots \tag{2.7}$$

(S.2) The Markov chain $\tau_{i+1} - \tau_i - \{X_t\}_{t=0}^{\tau_i}$ holds for all $i = 0, 1, \ldots$

(S.3) For all $i = 0, 1, \ldots$, the conditional pdfs $f_{\tau_{i+1}|\tau_i}$ exist.

Note that (2.6) holds trivially if $T < \infty$. Sun et al. [8] and Ornee and Sun [9] also assumed (2.6) in their analyses of the infinite time horizon problems for the Wiener [8] and the OU [9] processes. We use (2.7) to obtain a simplified form of the distortion-frequency tradeoff for time-homogeneous processes (see (2.13) below). Furthermore, (2.7) allows us to prove that the optimal sampling intervals $\tau_{i+1} - \tau_i$ form an i.i.d. process (see (2.12) below). We use (S.2), (S.3) to show that the optimal sampling policy is a symmetric threshold sampling policy in the frequency-constrained setting. See Appendix A.1 for a sufficient condition on the stochastic process for the optimal sampling policy to satisfy (S.2). For example, in the infinite time horizon, stochastic processes of the form $X_t = cW_{at} + bt$ satisfy the sufficient condition. Assumption (S.2) implies that the stopping times form a Markov chain. In contrast, the sampling intervals of causal sampling policies are assumed to form a regenerative process in [8][9].

To quantify the tradeoffs between the sampling frequency (2.3) and the MSE (2.4), we introduce the distortion-frequency function.

Definition 4 (Distortion-frequency function (DFF))**.** *The DFF for causal frequency-constrained sampling of the process* $\{X_t\}_{t=0}^T$ *is the minimum MSE achievable by causal frequency-constrained codes,*

$$\underline{D}(F) \triangleq \inf\{d : \exists\ (F, d, T)\ \textit{causal frequency-constrained code satisfying (S.1), (S.2), (S.3)}\}. \tag{2.8}$$

In the causal frequency-constrained sampling scenario, we say that a causal sampling policy is *optimal* if, when succeeded by the MMSE estimating policy (2.2), it forms an (F, d, T) code with $d = \underline{D}(F)$.

2.3 Optimal causal frequency-constrained sampling

In Theorem 1 below, we show that the optimal sampling policy is a two-threshold policy that is symmetric with respect to the expected value of the process given the last sample and the last sampling time, henceforth referred to as a *symmetric threshold policy*. In Theorem 2, we show a simplified form of the policy for time-homogeneous processes.

Theorem 1. *The optimal causal sampling policy in either finite or infinite time horizon for a class of continuous Markov processes satisfying assumptions (P.1)–(P.3) in Section 2.2 is a symmetric threshold sampling policy of the form*

$$\tau_{i+1} = \inf\{t \geq \tau_i : X_t - \mathbb{E}[X_t | X_{\tau_i}, \tau_i] \notin (-a_i(t, \tau_i), a_i(t, \tau_i))\}, \tag{2.9}$$

where the threshold a_i is a non-negative deterministic function of (t, τ_i).

Proof sketch. In Appendix A.2, we first introduce Lemmas 2–5 that supply majorization and real induction tools. Then, fixing an arbitrary causal sampling policy, we construct a symmetric threshold sampling policy that has the same sampling frequency as the fixed policy. Using majorization and real induction tools, we show that the MSE achieved by the symmetric threshold sampling policy is no larger than that achieved by the fixed policy. □

Theorem 1 shows that the optimal sampling policy is found within a much smaller set of sampling policies than that allowed in Definition 4: the input stochastic process $\{X_t\}_{t=0}^T$ is sampled only if the process innovation passes one of two symmetric thresholds. The thresholds depend on $\{X_t\}_{t=0}^T$ only through the current time t, the last sampling time, and the number of samples taken until t. Using the form of the sampling policy (2.9), we show that the MMSE estimating policy (2.2) simplifies as follows.

Corollary 1.1. *In the setting of Theorem 1, under the optimal sampling policy (2.9), the MMSE estimating policy reduces to*

$$\bar{X}_t = \mathbb{E}[X_t | X_{\tau_i}, \tau_i], \ t \in [\tau_i, \tau_{i+1}). \tag{2.10}$$

Proof. Appendix A.3. □

In the frequency-constrained setting, the expectation in (2.10) can be calculated at the estimator even without the knowledge of the sampling policy, whereas the expectation in (2.2) depends on the sampling policy at the sampler through the conditioning on the event that the next sample has not been taken yet, i.e., $t < \tau_{i+1}$. Corollary 1.1 confirms the conjecture in [6, Eq.(3)] and [7, Eq.(5)] on the form of the MMSE estimating policy.

Corollary 1.2. *In the setting of Theorem 1, the optimal causal sampling policy satisfies* (2.3) *with equality.*

Proof. Appendix A.4. □

Corollary 1.2 indicates that the inequality in the sampling frequency constraint (2.3) can be simplified to an equality.

Corollary 1.3. *In the setting of Theorem 1, the threshold in* (2.9) *satisfies*

$$\lim_{\delta \to 0^+} a_i(t+\delta, \tau_i) \geq a_i(t, \tau_i), \ \forall t \in [\tau_i, \tau_{i+1}), i = 0, 1, \ldots \tag{2.11}$$

Proof. Appendix A.5. □

Corollary 1.3 implies that the threshold $a_i(t, \tau_i)$, at time $t \in [\tau_i, \tau_{i+1})$, is either right-continuous or has a jump to a larger value. Thus, the continuous-path process $X_t - \mathbb{E}[X_t|X_{\tau_i}, \tau_i]$ in (2.9) must hit one of the symmetric thresholds $\pm a_i(\tau_{i+1}, \tau_i)$ at $t = \tau_{i+1}$.

Theorem 2. *In the infinite time horizon, the optimal causal sampling policy for time-homogeneous continuous Markov processes satisfying assumptions (P.1)–(P.3) in Section 2.2 is a symmetric threshold sampling policy of the form*

$$\tau_{i+1} = \inf\{t \geq \tau_i : X_t - \mathbb{E}[X_t|X_{\tau_i}, \tau_i] \notin (-a(t-\tau_i), a(t-\tau_i))\}, \tag{2.12}$$

where the threshold a is a non-negative deterministic function of $t - \tau_i$. The optimal threshold of (2.12) *is the solution to the following optimization problem,*

$$\underline{D}(F) = \min_{\substack{\{a(t)\}_{t\geq 0}: \\ \mathbb{E}[\tau_1]=\frac{1}{F}}} \frac{\mathbb{E}\left[\int_0^{\tau_1} (X_t - \mathbb{E}[X_t])^2 \mathrm{d}t\right]}{\mathbb{E}[\tau_1]}. \tag{2.13}$$

Proof sketch. The time homogeneity and the infinite time horizon allow us to treat each sampling time τ_i as a *new start* as if the sampler forgot the past sampling times and was about to take the first sample of process $\{X_t - \mathbb{E}[X_t|X_{\tau_i}, \tau_i]\}_{t\geq\tau_i}$, which has the same distribution as the original process $\{X_t\}_{t\geq 0}$. Thus, the sampling thresholds reduce to (2.12). Since the sampling intervals are i.i.d. and the process is time-homogeneous, the MSE accumulated in each sampling interval can be considered as i.i.d. renewal rewards. Applying the renewal reward theory to the MSEs, the long-term average MSE (2.4b) reduces to the MSE in the first interval only, thus (2.13) holds. See details in Appendix A.6. □

Remark 1. *In the setting of Theorem 2, the sampling intervals $\tau_{i+1} - \tau_i$, $i = 0, 1, \ldots$ under a symmetric threshold sampling policy of the form (2.12) are i.i.d.*

Theorem 2 shows that the optimal sampling policy in Theorem 1 can be further simplified for time-homogeneous processes in the infinite time horizon. As a consequence of time homogeneity, thresholds in (2.12) only depend on the time elapsed since the last sampling time. In contrast, the thresholds in (2.9) depend on the last sampling time as well.

Next, we show examples of using (2.13) to solve for the optimal threshold in (2.12). Before we show the examples, we introduce Lemma 1 below, which displays useful properties of the Wiener process.

Lemma 1 (Theorem 2.40 [69], Theorem 2.44 [69], Lemma 3 [8], Corollary 2.42 [69])**.** *Consider the Wiener process $\{W_t\}_{t=0}^{\infty}$, and let $\tau' \leq \tau$ be stopping times such that $\mathbb{E}[\tau] < \infty$. Then,*

(a) (Wald's lemma) $\mathbb{E}[W_\tau] = 0$*;*

(b) (Wald's second lemma) $\mathbb{E}[W_\tau^2] = \mathbb{E}[\tau]$*;*

(c) $\mathbb{E}\left[\int_0^\tau W_t^2 \mathrm{d}t\right] = \frac{1}{6}\mathbb{E}[W_\tau^4]$*;*

(d) $\mathbb{E}[W_\tau^2] = \mathbb{E}[W_{\tau'}^2] + \mathbb{E}[(W_\tau - W_{\tau'})^2]$.

Example 1: Applying (2.13) to the Wiener process in Definition 2, we conclude that the sampling threshold that achieves (2.13) is equal to $a(t) = \sqrt{\frac{1}{F}}$, thus the optimal causal sampling policy in (2.9) is

$$\tau_{i+1} = \inf\left\{t \geq \tau_i : |W_t - W_{\tau_i}| \geq \sqrt{\frac{1}{F}}\right\}, \tag{2.14}$$

and the DFF is equal to

$$\underline{D}(F) = \frac{1}{6F}. \tag{2.15}$$

See Fig. 2.2 for the sampling policy for the Wiener process $\{W_t\}_{t=0}^{\infty}$. While Nar and Başar [7] showed the optimal sampling policy for the Wiener process via solving Snell's envelope which requires solving an SDE, we provide a much simpler method below using (2.13).

Proof of Example 1. Converse: Plugging $X_t \leftarrow W_t$, we lower bound the objective function of (2.13) as

$$\frac{\mathbb{E}\left[\int_0^{\tau_1} W_t^2 \mathrm{d}t\right]}{\mathbb{E}[\tau_1]} = \frac{1}{6}\frac{\mathbb{E}\left[W_{\tau_1}^4\right]}{\mathbb{E}[\tau_1]} \tag{2.16a}$$

$$\geq \frac{\mathbb{E}[W_{\tau_1}^2]^2}{6\mathbb{E}[\tau_1]} \tag{2.16b}$$

$$= \frac{\mathbb{E}[\tau_1]}{6} \tag{2.16c}$$

$$= \frac{1}{6F}, \tag{2.16d}$$

where (2.16a) holds due to Lemma 1 (c); (2.16b) holds by applying Jensen's inequality to lower bound (2.16a); (2.16c) holds due to Lemma 1 (b); (2.16d) holds by plugging the minimization constraint in (2.13) into (2.16c).

Achievability: Plugging the sampling threshold $a(t) = \sqrt{\frac{1}{F}}$ into (2.13), we obtain $\underline{D}(F) = \frac{1}{6F}$. □

Example 2: Applying (2.13) to the continuous Lévy process

$$X_t = cW_{at} + bt, \tag{2.17}$$

$a, b, c \in \mathbb{R}$, $a > 0$, we conclude that the sampling threshold that achieves (2.13) is equal to $a(t) = c\sqrt{\frac{a}{F}}$, thus the optimal sampling policy in (2.12) is

$$\tau_{i+1} = \inf\left\{t \geq \tau_i : |X_t - \bar{X}_t| \geq c\sqrt{\frac{a}{F}}\right\}, \tag{2.18}$$

and the DFF is equal to

$$\underline{D}(F) = \frac{ac^2}{6F}. \tag{2.19}$$

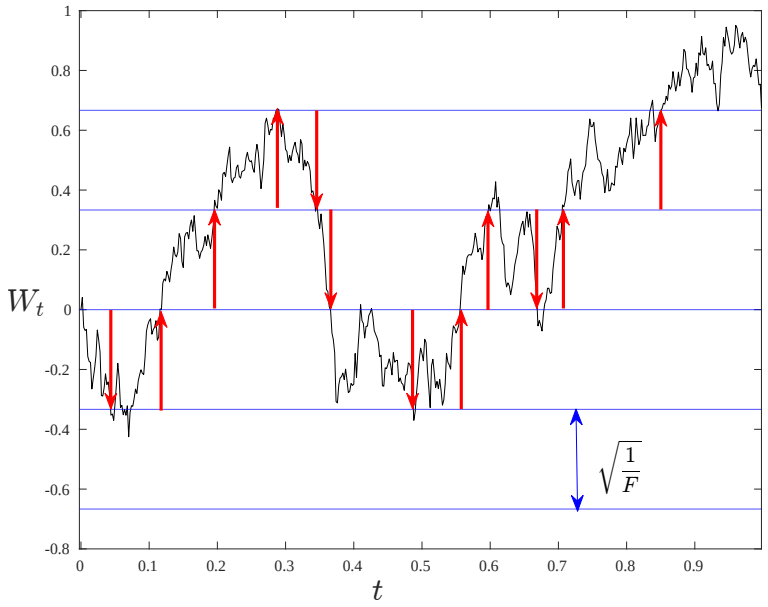


Figure 2.2: Symmetric threshold sampling of the Wiener process W_t. The black curve represents the Wiener process. The gap between blue horizontal lines represents the sampling threshold $a(t) = \sqrt{\frac{1}{F}} = \frac{1}{3}$. A down-arrow appears if the process innovation $W_t - W_{\tau_i}$ crosses the negative threshold. An up-arrow appears if the process innovation $W_t - W_{\tau_i}$ crosses the positive threshold.

Example 3: Applying (2.13) to the Ornstein-Uhlenbeck (OU) process

$$\mathrm{d}X_t = \theta(\mu - X_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \tag{2.20}$$

where μ, θ, σ are positive constants, we conclude that the sampling threshold that achieves (2.13) is $a(t) = \sqrt{R_1^{-1}\left(\frac{1}{F}\right)}$, thus the optimal causal sampling policy in (2.12) is

$$\tau_{i+1} = \inf\left\{t \geq \tau_i : |X_t - \bar{X}_t| \geq \sqrt{R_1^{-1}\left(\frac{1}{F}\right)}\right\}, \tag{2.21}$$

and the DFF is given by

$$\underline{D}(F) = F \cdot R_2\left(R_1^{-1}\left(\frac{1}{F}\right)\right), \tag{2.22}$$

where

$$R_1(v) \triangleq \frac{v}{\sigma^2}{}_2F_2\left(1, 1; \frac{3}{2}, 2; \frac{\theta}{\sigma^2}v\right), \tag{2.23}$$

$$R_2(v) \triangleq -\frac{v}{2\theta} + \frac{\sigma^2}{2\theta}R_1(v), \tag{2.24}$$

where ${}_2F_2$ is a generalized hypergeometric function. Under the assumption (S.1) in Section 2.2 and the assumption that the sampling intervals form a regenerative

process, Ornee and Sun [9] found the optimal sampling policy (2.21) for the OU process in the infinite horizon by forming an optimal stopping problem. They solved the optimal stopping problem via the Snell's envelope which requires solving an SDE. We provide an alternative method to find the optimal sampling policy for the OU process in Appendix A.7 using (2.13).

2.4 Successive refinement via causal frequency-constrained sampling

We extend the single-sampler single-estimator system in Sections 2.2–2.3 to an n-sampler n-estimator system, and we consider a successive refinement problem in a novel causal frequency-constrained sampling setting as shown in Fig. 2.3.

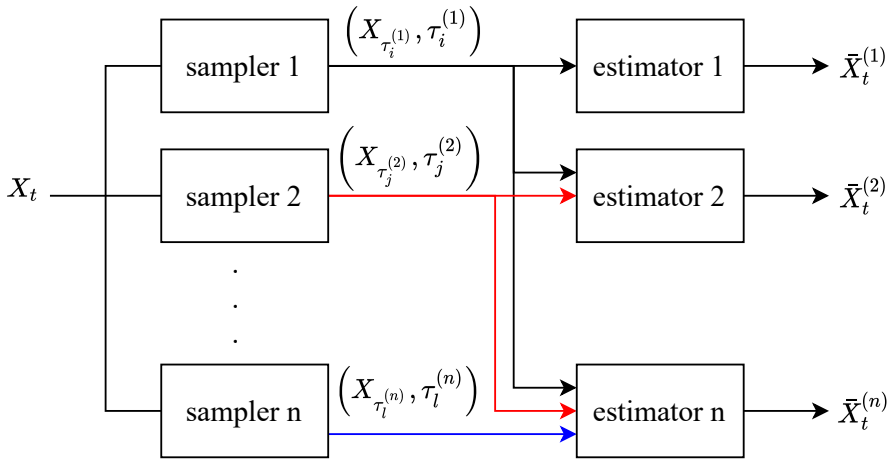


Figure 2.3: An n-sampler n-estimator system for successive refinement via causal frequency-constrained sampling.

All n samplers simultaneously track the source process $\{X_t\}_{t=0}^T$. Based on the causally observed process, the k-th sampler, $k = 1, 2, \ldots, n$, decides the sampling time stamps

$$0 \leq \tau_1^{(k)} \leq \tau_2^{(k)} \leq \cdots \leq \tau_{N_k}^{(k)} \leq T, \tag{2.25}$$

where N_k is a random variable that represents the total number of samples taken by the k-th sampler within time $[0, T]$. At time t, the k-th estimator uses all the samples and the sampling times generated by the first k samplers by time t, denoted by

$$M_t^{(k)} \triangleq \cup_{j=1}^k \left\{ \left(X_{\tau_i^{(j)}}, \tau_i^{(j)} \right) : \tau_i^{(j)} \leq t, i = 1, 2, \ldots \right\}, \tag{2.26}$$

to form a real-time estimate $\bar{X}_t^{(k)}$ of the source process.

We define n causal sampling policies and n causal estimating policies for successive refinement via causal frequency-constrained sampling below.

Definition 5 (An (F^n, d^n, T) causal frequency-constrained code for successive refinement)**.** *Fix a source process $\{X_t\}_{t=0}^T$. An (F^n, d^n, T) causal frequency-constrained code for successive refinement consists of n causal sampling policies and n causal estimating policies:*

1. *Each of n causal sampling policies, defined in Definition 1-1, is a collection of stopping times (2.25);*
2. *Given the real-valued samples and the sampling times generated by the first k causal sampling policies, i.e., $M_t^{(k)}$ (2.26), the k-th MMSE estimating policy is*

$$\bar{X}_t^{(k)} \triangleq \mathbb{E}\left[X_t \middle| M_t^{(k)}\right], \ k = 1, 2, \ldots, n. \tag{2.27}$$

The average sampling frequencies of all n causal sampling policies must satisfy (2.3) with $N \leftarrow N_k$, $F \leftarrow F_k$ $k = 1, 2, \ldots, n$, while all n MSEs must satisfy (2.4) with $\bar{X}_t \leftarrow \bar{X}_t^{(k)}$, $d \leftarrow d_k$, $k = 1, 2, \ldots, n$.

The estimating policy (2.27) ignores the knowledge that the next sample has not been taken, yet this will not incur any penalty on the achievable estimation MSEs for a class of source processes considered in Theorem 3 below due to Corollary 1.1.

To quantize the tradeoffs between the sampling frequencies F^n and the MSEs d^n, we introduce the *distortion-frequency region*. Fix a source process $\{X_t\}_{t=0}^T$. A frequency-distortion tuple (F^n, d^n) is said to be *achievable* if there exists an (F^n, d^n, T) causal frequency-constrained code for successive refinement whose first k causal sampling policies form a single causal sampling policy satisfying assumptions (S.1)–(S.3) in Section 2.2 for all $k = 1, 2, \ldots, n$. The distortion-frequency region $\mathcal{R}(F^n)$ is the closure of the set of distortions d^n such that (F^n, d^n) is achievable.

Given a sampling frequency F, we denote by $\pi(F)$ the optimal causal sampling policy (2.12) for an infinite-horizon, time-homogeneous source process $\{X_t\}_{t=0}^\infty$ satisfying (P.1)–(P.3). We denote by π_k the causal sampling policy of the k-th sampler in Fig. 2.3, $k = 1, 2 \ldots, n$. The distortion-frequency region $\mathcal{R}(F^n)$ for a class of source processes is shown below.

Theorem 3. *Consider an infinite-horizon, time-homogeneous source process* $\{X_t\}_{t=0}^{\infty}$ *satisfying (P.1)–(P.3) whose optimal causal sampling policy* $\pi(F)$ (2.12) *has a time-invariant sampling threshold, i.e.,* $\exists$ *function* $\theta(\cdot)\colon \mathbb{R}_+ \to \mathbb{R}_+$ *such that* $a(t-\tau_i) = \theta(F),\ \forall t \in [\tau_i, \tau_{i+1}),\ i = 0, 1, \ldots$ *If the sampling frequency constraints* F^n *satisfy*

$$\frac{\theta\left(\sum_{j=1}^{k} F_j\right)}{\theta\left(\sum_{j=1}^{k+1} F_j\right)} = z_k, \tag{2.28}$$

for some positive integer $z_k \in \mathbb{Z}_+$ *and for all* $k = 1, 2, \ldots, n-1$*, then the distortion-frequency region* $\mathcal{R}(F^n)$ *is*

$$d_k \geq \underline{D}\left(\sum_{j=1}^{k} F_j\right),\ k = 1, 2, \ldots, n. \tag{2.29}$$

Together with n *causal estimating policies in* (2.27)*,* n *causal sampling policies that achieve the right side of* (2.29) *for all* $k = 1, 2, \ldots, n$ *are*

$$\pi_1 = \pi(F_1),\ \textit{if } k = 1, \tag{2.30}$$

$$\pi_k = \pi\left(\sum_{j=1}^{k} F_j\right) \setminus \pi\left(\sum_{j=1}^{k-1} F_j\right),\ \textit{if } k = 2, \ldots, n. \tag{2.31}$$

Proof. Converse: Given any frequency constraints F^n, we show that the achievable distortions $d^n \in \mathcal{R}(F^n)$ are lower bounded as (2.29). For the k-th estimator, the samples and the sampling times that it receives from the first k samplers, i.e., $M_t^{(k)}$ (2.26), can be viewed as the samples and the sampling times generated by a single causal sampling policy satisfying (S.1)–(S.3) under sampling frequency $F \leq \sum_{j=1}^{k} F_j$. Since the DFF $\underline{D}(F)$ is a non-increasing function of F, the MSE d_k at the k-th estimator is lower bounded as (2.29).

Achievability: We show that for any F^n satisfying (2.28), the sampling policies in (2.30)–(2.31) achieve the converse bounds (2.29). The sampling policies in (2.30)–(2.31) imply that the samples and the sampling times received by the k-th estimator are equivalent to those generated by the causal sampling policy $\pi\left(\sum_{j=1}^{k} F_j\right)$. By definition, $\pi\left(\sum_{j=1}^{k} F_j\right)$ achieves (2.29) with equality. It remains to show that the sampling policies in (2.30)–(2.31) satisfy the frequency constraints F^n. Since (2.28) ensures that every sampling time in $\pi\left(\sum_{j=1}^{k-1} F_j\right)$ also belongs to $\pi\left(\sum_{j=1}^{k} F_j\right)$, the

sampling frequency of the k-th causal sampling policy is equal to

$$\lim_{T\to\infty}\frac{\mathbb{E}[N_k]}{T}=\sum_{j=1}^{k}F_j-\sum_{j=1}^{k-1}F_j=F_k, \tag{2.32}$$

for all $k=1,2,\ldots,n$. □

The source processes that satisfy the assumptions in Theorem 3 include the Wiener process, the continuous Lévy processes (2.17), and the OU processes (2.20), since, as we have shown below Lemma 1, they all have time-invariant sampling thresholds.

Theorem 3 shows that the first $1\leq k\leq n$ optimal causal sampling policies can be jointly viewed as a single causal sampling policy $\pi\left(\sum_{j=1}^{k}F_j\right)$ that operates under frequency $\sum_{j=1}^{k}F_j$. The symmetric threshold sampling policies (2.30)–(2.31) together with the estimating policies (2.27) successively refine the estimation MSEs at n estimators to $\underline{D}(F_1),\underline{D}(F_1+F_2),\ldots,\underline{D}(\sum_{k=1}^{n}F_k)$ (2.29). The achievability of the DFFs implies that the knowledge that the next sample has not been taken can indeed be ignored in the MMSE estimating policy (2.27) without loss of optimality.

2.5 Frequency-constrained sampling over imperfect channels

In this section, we show how the distortion-frequency tradeoffs for a continuous Lévy process (2.17) are affected if the channel is imperfect.

Channel with delay

We consider the communication scenario in Fig. 2.1 with the delay-free channel replaced by a channel that introduces a fixed channel delay $\delta\geq 0$ between the sampling time and the sample-delivery time: if the sampling time is τ_i, then the sample-delivery time is $\tau_i+\delta$. The sampler and the estimator are clock-synchronized, meaning that the estimator knows the sampling time from the delivery time and the fixed delay. We show that the optimal causal sampling policy for the continuous Lévy process $X_t=cW_{at}+bt$ (2.17) remains the symmetric threshold sampling policy in (2.18). We denote by Π the set of all causal sampling policies in the infinite time horizon.

The DFF for a channel with fixed delay δ is defined as

$$\underline{D}^{\mathrm{ch}}(F)=\inf_{\substack{\pi\in\Pi:\\ (2.3b)}}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i+\delta}^{\tau_{i+1}+\delta}(X_t-\bar{X}_t^{\mathrm{ch}})^2\mathrm{d}t\right], \tag{2.33}$$

where, similar to [6]–[9], we use the MMSE estimating policy

$$\bar{X}_t^{\mathrm{ch}} \triangleq \mathbb{E}[X_t|\{X_{\tau_j}\}_{j=1}^{i}, \tau^i] \tag{2.34a}$$

$$= cW_{a\tau_i} + bt,\ t \in [\tau_i + \delta, \tau_{i+1} + \delta), \tag{2.34b}$$

where (2.34b) is by the strong Markov property of the continuous Lévy process. Unlike Theorems 1–2 where we proved that the event $t < \tau_{i+1}$ in the estimating policy (2.2) can be ignored without loss of optimality, here we do not delve into the issue of whether ignoring the known event $t < \tau_{i+1}+\delta$ in the conditional expectation (2.34) is optimal.

We show the optimal causal sampling policy that achieves $\underline{D}^{\mathrm{ch}}(F)$.

Proposition 1. *In causal frequency-constrained sampling of the continuous Lévy process* (2.17) *with a fixed channel delay* δ *and the estimating policy* (2.34)*, the optimal causal sampling policy remains the symmetric threshold sampling policy in* (2.18) *and achieves*

$$\underline{D}^{\mathrm{ch}}(F) = \frac{ac^2}{6F} + ac^2\delta. \tag{2.35}$$

Proof. Plugging (2.34) into the (2.33), we obtain the objective function of $\underline{D}^{\mathrm{ch}}(F)$ as (Appendix A.8):

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i}^{\tau_{i+1}} (cW_{at} - cW_{a\tau_i})^2 \mathrm{d}t\right] + ac^2\delta. \tag{2.36}$$

Since the first term of (2.36) is the objective function of the DFF for a delay-free channel and the second term $ac^2\delta$ is a fixed number, we conclude that (2.36) is minimized by (2.18). The first term in (2.36) is given by (2.19). □

The first term on the right side of (2.35) is equal to the DFF (2.19) for a delay-free channel. The second term on the right side of (2.35) is the penalty on the achievable MSE due to the delay δ. The optimal sampling policy in the fixed-delay scenario coincides with the optimal sampling policy in the delay-free scenario. This differs from the result of [8], according to which the optimal causal sampling policy for the Wiener process (a special continuous Lévy process with $a = 1, b = 0, c = 1$) through a channel with an i.i.d. delay Y_i is a symmetric threshold sampling policy:

$$\tau_{i+1} = \inf\{t + \tau_i + Y_i : |W_{t+\tau_i+Y_i} - W_{\tau_i}| \geq \beta\}, \tag{2.37}$$

where β is a threshold that depends on the distribution of Y_i and the sampling frequency constraint. The setting in [8] is different from ours in Section 2.5, since the channel in [8] only serves one sample at a time. Because samples must wait in a queue before the previous sample is delivered, the optimal sampler in [8] takes a new sample after the previous sample is delivered, whereas in our setting, the sampler may take a new sample after or before the delivery of the previous sample. This results in the policy in [8] attaining an MSE in the constant-delay scenario no smaller than that indicated in (2.35).

Packet-drop channel with feedback

We replace the noiseless channel in Fig. 2.1 by a packet-drop channel, and we consider the system in Fig. 2.4. We denote a packet drop by symbol e. The channel transition probability of a packet-drop channel with packet-drop probability p is given by

$$P_{Y|X}(x|x) = 1 - p, \ \forall x \in \mathbb{R}, \tag{2.38a}$$

$$P_{Y|X}(e|x) = p, \ \forall x \in \mathbb{R}. \tag{2.38b}$$

We assume that the packet-drop channel is memoryless and a packet drop is independent of the source process. We denote by B_{τ_i} a 1-bit feedback sent from the

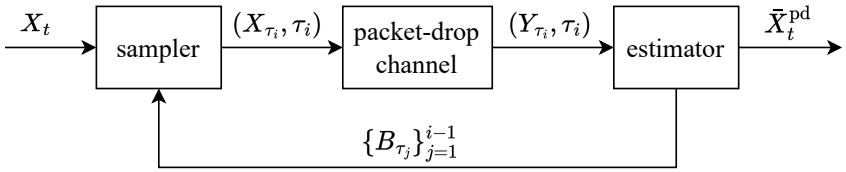


Figure 2.4: System model for causal frequency-constrained sampling over a packet-drop channel with feedback.

estimator to the sampler at sampling time τ_i. If $B_{\tau_i} = 1$, then the packet is not dropped, i.e., $Y_{\tau_i} = X_{\tau_i}$; otherwise, the packet is dropped, i.e., $Y_{\tau_i} = e$. Since by assumption, both the sampler and the estimator know $X_0 = 0$ at $\tau_0 = 0$, it holds that $B_{\tau_0} \triangleq 1$. Given feedback bits $\{B_{\tau_j}\}_{j=1}^{i}$, we denote the indices of all the successful transmissions by

$$\mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right) \triangleq \{j \colon B_{\tau_j} = 1, j = 1, 2, \ldots, i\}, \tag{2.39}$$

and we denote the time of the last successful transmission by time τ_i by

$$S_i \triangleq \tau_{\max \mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right)}. \tag{2.40}$$

We define an $\langle n, d, T\rangle$ causal code for a packet-drop channel that transmits n real-valued samples of a source process $\{X_t\}_{t=0}^{\infty}$ within an expected time horizon T at an MSE less than or equal to d.

Definition 6 (An $\langle n, d, T\rangle$ causal code for a packet-drop channel)**.** *Fix a source process $\{X_t\}_{t=0}^{\infty}$ and fix a packet-drop channel with a single-letter transition probability $P_{Y|X}\colon \mathbb{R} \to \mathbb{R} \cup \{e\}$. An $\langle n, d, T\rangle$ causal code for a packet-drop channel is a pair of causal sampling and estimating policies:*

1. *The causal sampling policy is a collection of n stopping times*

$$0 \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_n \tag{2.41}$$

   adapted to the filtration generated by the source process $\{X_t\}_{t=0}^{\infty}$ and the feedback process $\{B_{\tau_i}\}_{i=1}^{n}$.

2. *Given channel outputs $\{Y_{\tau_j}\}_{j=1}^{i}$ and sampling times τ^i, the estimating policy is*

$$\bar{X}_t^{\mathrm{pd}} \triangleq \mathbb{E}\left[X_t \middle| \{Y_{\tau_j}, \tau_j\}_{j\in\mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right)}\right]. \tag{2.42a}$$

The expectation of the n-th sampling time must satisfy

$$\mathbb{E}[\tau_n] = T, \tag{2.43}$$

while the MSE must satisfy

$$\frac{1}{T}\mathbb{E}\left[\int_0^{\tau_n} \left(X_t - \bar{X}_t^{\mathrm{pd}}\right)^2 \mathrm{d}t\right] \leq d. \tag{2.44}$$

The causal sampling policy is assumed to satisfy (S.1) in Section 2.2 as well as

(S.4) Given the number of unsuccessful transmissions $i - \max\mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right)$ after the last successful transmission, the sampling interval $\tau_{i+1} - \tau_i$ is independent of the source process $\{X_t\}_{t=0}^{\tau_i}$ and the feedback bits $\{B_{\tau_j}\}_{j=1}^{i}$ by time τ_i, $i = 1, 2, \ldots, n$.

The estimating policy in (2.42) can be suboptimal since it ignores the knowledge implied by the sampling times of the dropped packets. For a continuous Lévy process in (2.17), the estimating policy $\bar{X}_t^{\mathrm{pd}}$ (2.42) reduces to

$$\bar{X}_t^{\mathrm{pd}} = cW_{aS_i} + bt,\ t \in [\tau_i, \tau_{i+1}), \tag{2.45}$$

due to the strong Markov property of the Lévy process.

To quantify the tradeoffs among the number of samples n (2.41), the expected time horizon T (2.43), and the MSE (2.44), we introduce the distortion-sample-time function for a packet-drop channel.

Definition 7 (Distortion-sample-time function (DSTF) for a packet-drop channel)**.** *Fix a source process $\{X_t\}_{t=0}^{\infty}$. The DSTF for a packet-drop channel is the minimum MSE (2.44) achievable by causal codes with n samples and expected horizon T:*

$$\underline{D}^{\mathrm{pd}}(n,T) \triangleq \inf\{d\colon \exists\, \langle n,d,T\rangle \text{ causal code for a packet-drop channel satisfying (S.1), (S.4)}\}. \tag{2.46}$$

We present $\underline{D}^{\mathrm{pd}}(n,T)$ and the causal code for a packet-drop channel that achieves it for a continuous Lévy process.

Theorem 4. *Fix a continuous Lévy process $\{X_t\}_{t=0}^{\infty}$ in (2.17) and fix a packet-drop channel (2.38) with packet-drop probability $p < \frac{1}{5}$. Together with the estimating policy in (2.42), the causal sampling policy that operates as:*

- *if $B_{\tau_i} = 1$, then the next sampling time is*

$$\tau_{i+1} = \inf\left\{t \geq \tau_i\colon |X_t - \bar{X}_t^{\mathrm{pd}}| \geq c\sqrt{\frac{aT}{1+(n-1)(1-p)}}\right\}; \tag{2.47}$$

- *if $B_{\tau_i} = 0$, then the next sampling time is $\tau_{i+1} = \tau_i$;*

achieves the DSTF for a packet-drop channel

$$\underline{D}^{\mathrm{pd}}(n,T) = \frac{ac^2T}{6(1+(n-1)(1-p))}. \tag{2.48}$$

Proof sketch. We first show that the DSTF is lower bounded by a minimization problem achieved by the causal sampling policy in Theorem 4. Then, we show that under the causal sampling policy in Theorem 4, the DSTF coincides with its lower bound. See details Appendix A.9. □

The causal sampling policy in Theorem 4 operates as follows. If the last sample is not dropped, then the sampler follows the symmetric threshold sampling policy in (2.47) to determine the next sampling time; otherwise, the sampler immediately retransmits the dropped sample until it is successfully received by the estimator.

Considering $\frac{n}{T}$ as the sampling frequency, letting $F \triangleq \frac{n}{T}$, and taking $n \to \infty$, we rewrite the sampling policy (2.47) and $\underline{D}^{\mathrm{pd}}(n,T)$ in (2.48) in terms of frequency F as

$$\tau_{i+1} = \inf\left\{t \geq \tau_i \colon |X_t - \bar{X}_t^{\mathrm{pd}}| \geq c\sqrt{\frac{a}{(1-p)F}}\right\},\ i = 0,1,2,\ldots; \tag{2.49}$$

$$\underline{D}^{\mathrm{pd}}(F) = \frac{ac^2}{6(1-p)F}. \tag{2.50}$$

We observe that the causal sampling policy in (2.49) employs a sampling threshold $\sqrt{\frac{1}{1-p}}$-fold larger than the sampling threshold for a noiseless channel. The lower sampling frequency for transmitting new samples allows for the retransmission opportunities.

2.6 Conclusion

We study the optimal causal frequency-constrained sampling for a class of continuous processes satisfying regularity conditions (P.1)–(P.3) in Section 2.2. We show that the optimal causal frequency-constrained sampling policy is a symmetric threshold sampling policy (Theorems 1), where the sampler transmits a new sample once the process innovation crosses either of the two symmetric thresholds. As a result of the symmetric structure of the optimal sampling policy, we confirm the conjecture on the estimating policy in prior literature [6]–[7] for the Wiener process and the OU process (Corollary 1.1), that is, the knowledge that the next sample has not been taken is useless in reducing the estimation MSE. If the stochastic process is also time-homogeneous (Definition 3), we show that in the finite time horizon, the sampling threshold only depends on the time elapsed from the last sampling time (Theorem 2), and we simplify the DFF to the minimization problem in (2.13). As we show below Lemma 1, solving the simplified minimization problem is much easier than solving an SDE (Snell's envelope) in prior literature [6]–[9]. Extending the one-sampler one-estimator system to an n-sampler n-estimator system, we consider the successive refinement problem in the causal frequency-constrained sampling setting, which parallels the classical successive refinement problem in the non-causal lossy compressing setting. For a class of source processes that have a time-invariant sampling threshold, we show that the optimal causal sampling policies cooperate so that the first k sampling policies, $k = 1,2,\ldots,n$, can be viewed as a single optimal causal sampling policy under the summed frequency $\sum_{j=1}^{k} F_k$. The sampling policies successively refine the estimation MSEs and attain the DFFs (Theorem 3). Dropping the assumption that the channel is delay-free, we show that

the optimal causal sampling policy of a continuous Lévy process for a delay-free channel remains optimal for a channel with a fixed delay (Proposition 1), revealing the resilience of the optimal causal sampling policy to the channel delay. Dropping the assumption that the channel is noiseless, we show the optimal causal sampling policy of a continuous Lévy process for a packet-drop channel with feedback. It transmits new samples following a symmetric threshold sampling policy but with a threshold larger than that for a noiseless channel. The sampler thus transmits fewer new samples in exchange for the opportunities to retransmit the dropped ones.

2.7 Future research directions

Based on the findings in Sections 2.2–2.5, we list several interesting directions for future research.

Causal frequency-constrained sampling over a channel with a random delay

It would be interesting to find the optimal causal sampling policy for a channel with a random delay. There are two possible ways to introduce the random delay.

Additive random delay: Assuming that the sampling time is τ_i, an additive random delay D_i leads to a sample delivery time $\tau_i + D_i$. The difficulties of solving the optimal causal sampling policy for a channel with an additive random delay are in two aspects. First, if one assumes that the additive random delays are i.i.d., then the delay may mess up the order of the samples received by the estimator, e.g., it is possible that $\tau_i + D_i > \tau_{i+1} + D_{i+1}$. One needs to take the possibly permuted sample order into consideration when finding the optimal causal sampling policy. Second, if one drops the i.i.d. assumption on the random delays and assumes that $\tau_i + D_i \leq \tau_{i+1} + D_{i+1}$ holds almost surely for all $i = 1, 2, \ldots$, then the dependency between random delays makes solving the DFF difficult. These difficulties inspire the formulation below.

Channel as a first-in first-out (FIFO) queue with random service time: Assume that the channel is a FIFO queue with i.i.d. service time (i.e., random delay). The sample is not served by the channel until the previous sample is delivered. As a result, it is suboptimal to transmit a new sample when the queue is non-empty [8, 9]. This is the setting considered in [8, 9], where the optimal causal sampling policies for the Wiener process and the OU process are presented. In this setting, one can further find the optimal causal sampling policy for a wider class of stochastic processes. One possible method is to simplify the DFF to an optimization problem over just one sampling interval (similar to (2.13)), and to solve the simplified problem using

tools similar to those used in Theorem 1 and in [6, 7, 8, 9], e.g., the strong Markov and the martingale properties of the source process, majorization theory, and Snell's envelope.

Causal frequency-constrained sampling over a noisy channel

It would be interesting to find the optimal causal sampling policy for different noisy channels. The first difficulty is to solve the MMSE estimator (2.2). The MMSE estimator can be difficult to solve even for sampling the Wiener process over an AWGN channel that introduces a Gaussian noise Z_t. This is because the stopping time makes the random variable W_{τ_i} non-Gaussian, that is, $\mathbb{E}[W_t|W_{\tau_i}+Z_{\tau_i},\tau_i]$ may not be a linear function of $W_{\tau_i}+Z_{\tau_i}$. Yet, one can use the linear MMSE estimator as a potentially suboptimal estimator to solve for the causal sampling policy that attains the minimum MSE under a sampling frequency constraint.

Causal frequency-constrained sampling for a wider class of source processes

It would be interesting to find the optimal causal sampling policy (Theorem 1) for a wider class of stochastic processes.

Example 1: One can find the optimal causal sampling policy for a multidimensional stochastic process of which each dimension is a scalar stochastic process satisfying (P.1)–(P.3) in Section 2.2. One existing method for sampling a multidimensional Wiener process [7] with independent dimensions is to establish an optimal stopping problem for the multidimensional process and solving the problem via Snell's envelope [6, 7]. The optimal causal sampling policy of the multidimensional Wiener process is a symmetric threshold sampling policy that transmits a sample if the l^2 norm of the multidimensional process innovation crosses a threshold [7]. Alternatively, based on the proof for Theorem 1 in Appendix A.2, we conjecture that one can first generalize all majorization tools to vectors and then apply real inductions on the MSE over vectors. While this may only give a structural result on the optimal causal sampling policy, one may also need to leverage other tools to specify the structural result to an explicit policy. We conjecture that symmetric structure of the optimal causal sampling policy for the multidimensional Wiener process remains optimal for a wider class of multidimensional stochastic processes satisfying Markov and symmetric properties (P.1)–(P.3).

Example 2: One can find the optimal causal sampling policy for a stochastic process whose mean-square residual error does not have an even and quasi-concave pdf. The even and quasi-concave assumption allows us to use the majorization tools

in Lemmas 2–4. One can drop this assumption if the majorization tools can be generalized to other pdfs. We conjecture that the optimal causal sampling policy transmits a sample once the process innovation falls outside a *typical* interval in which the innovation lies with high probability. For a general pdf, the optimal causal sampling policy could have more than two sampling thresholds.

Causal frequency-constrained sampling for a partially observed system

While the source processes are assumed to be fully known by the sampler in Sections 2.2–2.5, it is practically important to find the optimal causal sampling policy for a partially observed stochastic process.

Problem: Given a source process $\{X_t\}_{t=0}^T$, we assume that the sampler only observes $\{Y_t\}_{t=0}^T$, where $Y_t = X_t + V_t$ for some noise process $\{V_t\}_{t=0}^T$. The sampler causally decides the sampling times of the observed process $\{Y_t\}_{t=0}^T$. At time $t \in [\tau_i, \tau_{i+1})$, $i = 0, 1, \ldots$ the estimator forms a real-time estimate of the source process using the causally received samples $\{Y_{\tau_j}\}_{j=1}^i$ and sampling times τ^i. One can try to find the optimal causal sampling policy that minimizes the end-to-end estimation MSE under a sampling frequency constraint.

Partially observed sampling vs. fully observed sampling: Assume that the estimator recovers the source process using the MMSE estimating policy $\bar{X}_t \triangleq \mathbb{E}[X_t|\{Y_{\tau_j}\}_{j=1}^i, \tau^i]$, $t \in [\tau_i, \tau_{i+1})$. Let $\bar{X}'_t \triangleq \mathbb{E}[X_t|\{Y_s\}_{s=0}^t]$ be the causal estimate of the source process using the partially observed process $\{Y_t\}_{t=0}^T$. One can show that the end-to-end estimation MSE (2.4) can be decomposed as

$$\frac{1}{T}\left(\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i}^{\tau_{i+1}}(X_t-\bar{X}'_t)^2\mathrm{d}t\right]+\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i}^{\tau_{i+1}}(\bar{X}'_t-\bar{X}_t)^2\mathrm{d}t\right]\right), \tag{2.51}$$

where the first term in (2.51) represents the MSE due to the partial observation independent of the sampling policy; the second term in (2.51) is the objective function to be minimized over causal sampling policies. Once one shows that $\bar{X}_t$ in the second term is equivalent to the estimate of $\bar{X}'_t$, i.e., $\bar{X}_t = \mathbb{E}[\bar{X}'_t|\{\bar{X}'_{\tau_j}\}_{j=1}^i, \tau^i]$, $t \in [\tau_i, \tau_{i+1})$, the second term can be viewed as the MSE due to causal sampling $\{\bar{X}'_t\}_{t=0}^T$ under a frequency constraint, and the problem reduces to the fully observed sampling problem stated in Section 2.2. The difficulties lie in evaluating $\bar{X}'_t$ at stopping times and showing that $\bar{X}'_t$ satisfies the regularity conditions in Section 2.2.

Causal frequency-constrained sampling for a wider class of distortion measures One can change the MSE to other distortion measures and evaluate the optimal causal sampling policy. For example, one can find the optimal causal sampling policy that minimizes the age of information. Age of information is defined as the time difference between the current time and the generation time of the last sample received by a receiver, measuring the freshness of a sample. Sun et al. [8] showed that minimizing the MSE for the Wiener process over a set of deterministic sampling policies is equivalent to minimizing the age of information. Ornee and Sun [9] showed that minimizing the MSE for the OU process over a set of deterministic sampling policies is equivalent to minimizing a non-linear function of the age of information. These two findings demonstrate the close connection between the operational distortion measure (i.e., MSE) and the age of information. Evaluating the optimal causal sampling policies under different measures helps to draw connections between them and gain insight into the operational meaning of informational measures like the age of information.

Chapter 3

CAUSAL RATE-CONSTRAINED SAMPLING

3.1 Introduction

In digital communications, real-valued samples are quantized before the transmission. In this chapter, we replace the sampling frequency constraint in Chapter 2 by a bitrate constraint on the expected number of bits transmitted per second, and we consider the following communication scenario: an encoder observes a stochastic process and causally decides when and what to transmit about it, under a rate constraint; a decoder uses the received codewords to causally estimate the process in real time; the channel is delay-free and noiseless. As we have briefly discussed in Chapter 1.2, we refer to this communication scenario as *causal rate-constrained sampling*.

For a class of continuous Markov processes (e.g., Wiener process, continuous Lévy process, and Ornstein-Uhlenbeck process) satisfying symmetry and regularity conditions, we find the optimal causal encoding and decoding policies that minimize the end-to-end estimation mean-square error under the rate constraint. We show that the optimal encoding policy transmits a 1-bit codeword once the process innovation passes one of two symmetric thresholds of the optimal causal sampling policy in Chapter 2. The optimal decoder noiselessly recovers the last sample from the 1-bit codewords and codeword-generating time stamps, and uses it to decide the running estimate of the current process, until the next codeword arrives. Since the 1-bit codewords represent the sign of the process innovations, we term the optimal causal rate-constrained code as the sign-of-innovation (SOI) code. The SOI code applies to rate-constrained control: it minimizes the mean-square cost of a continuous-time control system driven by a continuous Markov process and controlled by an additive control signal. The SOI code also applies to successive refinement in the causal rate-constrained sampling setting.

Replacing the perfect channel by imperfect ones, we show that the SOI code is resilient to channel delay and noise. For a channel with a fixed delay, the SOI code, as the optimal causal rate-constrained code for a delay-free channel, remains optimal. For a binary erasure channel (BEC), we show that the optimal causal rate-constrained code is essentially the optimal causal frequency-constrained sampling

policy for a packet-drop channel (Section 2.5) followed by an SOI compressor.

Surprisingly, for the Wiener process, the distortion-rate tradeoff achieved by the SOI code is significantly better than that achieved by the best non-causal code. This is because the SOI code leverages the free timing information supplied by the zero-delay channel between the encoder and the decoder. The key to unlocking that gain is the event-triggered nature of the SOI sampling policy. In contrast, the causal distortion-rate tradeoffs achieved with deterministic sampling policies are much worse. We show that the optimal deterministic sampling policy that achieves an informational causal distortion-rate function is a uniform sampling policy. In either signal-dependent or deterministic sampling, the optimal strategy is to sample the process as frequently as possible and to transmit 1-bit codewords to the decoder without delay.

Sections 3.2–3.5, 3.8, which formulate the causal rate-constrained sampling problem, present the optimal causal codes, show that the SOI code applies to rate-constrained control, and discuss delay-tolerant rate-constrained sampling, appear in the research papers [59, 60, 70, 71]. Sections 3.6–3.7, which investigate the best distortion-rate tradeoffs for successive refinement and imperfect channels, appear for the first time.

Prior work

Although the works [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 61, 64, 65, 66] on causal sampling in Chapter 2 did not consider quantization effects, in digital communication systems, real-valued numbers are quantized into bits before a transmission. In the field of information theory, researchers have investigated informational causal rate-distortion functions for different Gaussian processes, e.g., [14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25], which serve as lower bounds to the operational causal rate-distortion functions. Yet, most of them focused on discrete-time processes. The rate-distortion tradeoffs for continuous-time processes are largely studied in the non-causal setting. Berger [72] derived the non-causal distortion-rate function for the Wiener process using reverse water-filling over the power spectrum of the process. For the non-causal lossy source coding of the uniformly sampled Wiener process, Kipnis et al. [73] derived the tradeoffs among the sampling frequency, the communication bitrate, and the estimation MSE, achievable in the limit of infinite delay. However, the infinite delay introduced by classical rate-distortion theory in [72, 73] is unsuitable in many delay-sensitive applications.

Causal rate-constrained sampling is closely related to quantized event-triggered control, which has attracted significant research attention in recent years [27, 30, 28, 32, 26, 31, 34, 35, 29, 33, 74]. Kofman and Braslavsky [27] designed a quantized event-triggered controller for noiseless partially observed continuous-time LTI systems to ensure asymptotic convergence of the system to the origin with zero average rate, seemingly violating the *data-rate theorem* [75]. Similar to [27], the fact that sampling time stamps of event-triggered policies carry information is also exploited in [30, 28, 32, 26]. Pearson et al. [30] considered encoding the deterministic and possibly nonuniformly sampled states of noiseless continuous-time LTI systems into symbols in a finite alphabet with a *free* symbol representing the absence of transmission. For discrete-time linear systems with additive disturbances, Khina et al. [28] considered a setting where at each discrete-time instant, the encoder either transmits 1 bit or transmits the free symbol, and designed a quantizer with three bins using a Lloyd-Max algorithm with the quantization bin of the largest probability corresponding to the free symbol. Ling [32] designed a periodic event-triggered quantization policy to stabilize continuous-time LTI systems subject to i.i.d. feedback dropouts, bounded network delay, and bounded noise, which leads to a stabilizing rate that is lower than the one the data-rate theorem [75] requires for time-triggered policies. Khojasteh et al. [26] considered sampling noiseless continuous-time LTI systems where the state estimation error exceeds an exponentially decaying function. They found that for small enough delays, the information transmission rate required for stabilizing systems can be any positive value; it starts to increase once the delay exceeds a critical value. Quantized event-triggered control has also been studied for continuous-time LTI systems with bounded disturbances [31], for partially observed continuous-time LTI systems without noise [34] and with bounded noise [35], for discrete-time noiseless linear systems [29], and for partially observed continuous-time LTI systems with time-varying network delay [76]. Event-triggered control schemes to guarantee exponential stabilization were designed both for continuous-time LTI systems with bounded disturbances under a bounded rate constraint [33] and for noiseless continuous-time LTI systems under time-varying rates constraints and channel blackouts [74].

Chapter organization and contribution

We adopt an information-theoretic approach to continuous-time causal estimation by considering the optimal tradeoff between the achievable MSE and the average number of bits communicated. This is different from the models studied in [1, 2,

3, 4, 5, 61, 10, 6, 7, 8, 9, 64, 65, 66], where communication cost is measured by the number of transmissions, and each infinite-precision transmission can carry an infinite amount of information. For communication over digital channels, a bitrate constraint, routinely considered in information theory, is more appropriate. In contrast to [14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25] that compress discrete-time processes and transmit codewords at consecutive times, we compress continuous-time processes and allow the transmission times to causally depend on the process. Our setting is different from [72, 73] in that we do not ignore delay: our distortion at time t is measured with respect to the actual values of the process and the estimate at time t; whereas [72, 73] permit an infinite delay, following a standard assumption in information theory. In contrast to the works [27, 30, 28, 32, 26, 31, 34, 35, 29, 33, 74] that do not claim or consider the optimality of the proposed event-triggered policies, we show the optimality of the SOI code for a rate-constrained control problem.

In Section 3.2, we formulate a single-encoder single-decoder (point-to-point) causal rate-constrained sampling problem. We define the causal rate-constrained codes and the distortion-rate function $D(R)$ to quantify the rate R and the distortion d tradeoffs.

In Section 3.3, we show that the causal code that attains the optimal distortion-rate tradeoff is the SOI code: it generates a 1-bit codeword representing the sign of the process innovation once the innovation exceeds one of two symmetric thresholds. This surprisingly simple structure is a consequence of both the real-time distortion constraint, which penalizes coding delays, and the symmetry of the innovation distribution (P.3), which ensures the optimality of the two-threshold sampling policy. The SOI encoder can be implemented as a sampler followed by a compressor without loss of optimality. To study the tradeoffs between the sampling frequency and the rate per sample under a rate per second constraint R, we define the distortion-frequency-rate function. It is achieved by the maximum frequency R (samples per sec) and the minimum rate 1 (bit per sample), implying that transmitting 1 bit codewords as frequently as possible is optimal.

In Section 3.4, we show the optimal distortion-rate tradeoff attained by causal rate-constrained codes with deterministic sampling times. In the SOI code, the encoder continuously tracks the process and generates a bit once the process passes a pre-set threshold. To reconstruct the process, both those bits and their time stamps are required at the decoder. In the scenario where the sampler is process-agnostic,

or the decoder has no access to timing information, one has to adopt a process-independent sampling policy. We prove that a uniform sampling policy achieves the informational distortion-rate function (IDRF) for the Wiener process. To define the IDRF for the deterministic sampling policies, we change the operational rate constraint to a directed mutual information rate constraint, which serves as an information-theoretic lower bound. To confirm that the IDRF is a meaningful gauge of what is achievable in the zero-delay causal compression, we implement the greedy Lloyd-Max compressor [28] to compress the process innovations, and we verify that the performance of the resulting scheme is close to the IDRF.

In Section 3.5, we show that the SOI code remains optimal in a rate-constrained control scenario with a stochastic plant driven by a process satisfying assumptions (P.1)–(P.3) in Section 2.2. The SOI code minimizes the mean-square cost between the desirable state 0 and the state of the stochastic plant.

In Section 3.6, we extend the point-to-point communication system to an n-encoder n-decoder system and consider a successive refinement problem in the causal rate-constrained sampling setting. Similar to classical successive refinement [67], which is a non-causal data compression problem, the successive refinement in Section 3.6 also studies the tradeoffs between the rate and the distortion. Yet, the tradeoffs are studied in the causal encoding setting. The successive refinement in the causal rate-constrained sampling setting is equivalent to the successive refinement in the causal frequency-constrained sampling setting (Section 2.4) with the frequency constraints F^n replaced by the rate constraints R^n. We show that appending an SOI compressor to each of n optimal causal sampling policies in Section 2.4 gives n causal encoding policies that successively refine the estimation MSEs at n decoders to DRFs $D(R_1), D(R_1 + R_2), \ldots, D(\sum_{k=1}^{n} R_k)$.

In Section 3.7, we drop the assumption that the channel is perfect and show the optimal causal rate-constrained codes for imperfect channels. For a channel with a fixed delay, we show that the SOI code remains optimal. For a BEC with 1-bit feedback indicating whether or not the bit is erased, the optimal causal rate-constrained code can be obtained by appending an SOI compressor to the optimal causal frequency-constrained sampling policy for a packet-drop channel with feedback.

In Sections 3.8, we discuss how the achievable distortion-rate tradeoffs for the Wiener process are affected if a delay is tolerable. Surprisingly, the distortion achieved by the SOI code is smaller than that achieved by the best non-causal code. This is because, in the SOI code, the encoder and the decoder know the random

sampling times perfectly, whereas in the classical non-causal coding setting, the free timing information is not considered. We also show that if the decoder is allowed to wait for the next codeword before decoding, the achievable MSE can be further decreased.

3.2 Problem statement

Consider the system in Fig. 3.1. A source outputs a real-valued continuous-time stochastic process $\{X_t\}_{t=0}^{T}$ with state space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, where $\mathcal{B}_{\mathbb{R}}$ is the Borel σ-algebra on $\mathbb{R}$.

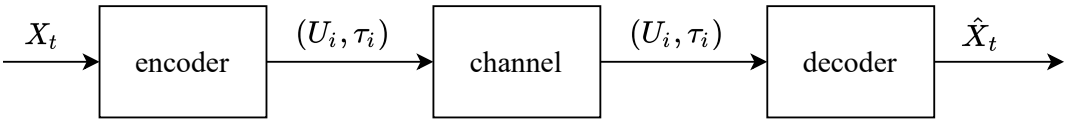


Figure 3.1: System Model. Sampling time τ_i and codeword U_i are chosen by the encoder's sampling and compressing policies, respectively.

An encoder tracks the input process $\{X_t\}_{t=0}^{T}$ and decides to disclose information about it at a sequence of stopping times (2.1) that are decided by a causal sampling policy. Thus, the total number of time stamps N can be random. The time horizon T can either be finite or infinite. At time τ_i, the encoder generates a codeword U_i according to a causal compressing policy, based on the process stopped at τ_i, $\{X_t\}_{t=0}^{\tau_i}$. Then, the codeword U_i is passed to the decoder without delay through a noiseless channel. At time t, $t \in [\tau_i, \tau_{i+1})$, the decoder estimates the input process X_t, yielding $\hat{X}_t$, based on all the received codewords and the codeword-generating time stamps, i.e., $(U_j, \tau_j),\ j = 1, 2, \ldots, i$. Similar to Chapter 2, the encoder and the decoder can leverage the timing information for free due to the clock synchronization and the zero-delay channel.

We formally define encoding and decoding policies, and define a distortion-rate function (DRF) to describe the tradeoffs between the estimation distortion and the communication rate.

Definition 8 ((R, d, T) causal rate-constrained codes)**.** *A time horizon-T causal rate-constrained code for the stochastic process $\{X_t\}_{t=0}^{T}$ is a pair of encoding and decoding policies. The encoding policy consists of a causal sampling policy and a causal compressing policy.*

1. *The causal sampling policy, defined in Definition 1-1, decides the stopping times (2.1) at which codewords are generated.*

2. *The causal compressing policy, characterized by the $\mathbb{Z}_+$-valued process $\{f_t\}_{t=0}^T$ adapted to $\{\mathcal{F}_t\}_{t=0}^T$, decides the codeword to transmit at time τ_i,*

$$U_i = f_{\tau_i}. \tag{3.1}$$

Given an encoding policy, the MMSE decoding *policy uses the received codewords and codeword-generating time stamps to estimate the process,*

$$\hat{X}_t = \mathbb{E}[X_t | U^i, \tau^i, t < \tau_{i+1}],\ t \in [\tau_i, \tau_{i+1}). \tag{3.2}$$

In an (R, d, T) code, the lengths of the codewords must satisfy the average communication rate constraint R bits per sec:

$$\frac{1}{T}\mathbb{E}\left[\sum_{i=1}^{N} \ell(U_i)\right] \leq R \text{ (bits per sec)},\ (T < \infty), \tag{3.3a}$$

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{i=1}^{N} \ell(U_i)\right] \leq R \text{ (bits per sec)},\ (T = \infty), \tag{3.3b}$$

where $\ell\colon \mathbb{Z}_+ \to \mathbb{Z}_+$ denotes the length of its argument in bits, $\ell(x) = \lfloor \log_2(x) \rfloor + 1$ for $x > 0$, $\ell(0) = 1$, while the MSE must satisfy

$$\frac{1}{T}\mathbb{E}\left[\int_0^T (X_t - \hat{X}_t)^2 \mathrm{d}t\right] \leq d,\ (T < \infty), \tag{3.4a}$$

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T (X_t - \hat{X}_t)^2 \mathrm{d}t\right] \leq d,\ (T = \infty). \tag{3.4b}$$

Allowing more freedom in designing the decoding policy will not lead to a lower MSE, because (3.2) is the MMSE estimator.

Definition 9 (Distortion-rate function (DRF))**.** *The DRF for causal rate-constrained sampling of the process $\{X_t\}_{t=0}^T$ is the minimum MSE achievable by causal rate-R codes:*

$$D(R) \triangleq \inf\{d : \exists\ (R, d, T) \text{ causal rate-constrained code satisfying (S.1), (S.2), (S.3)}\}. \tag{3.5}$$

We say that a causal (R, d, T) code is *optimal* if $d = D(R)$.

3.3 Optimal causal rate-constrained sampling

We first present the optimal causal code that achieves the DRF. We then show that the optimal encoder can be implemented as a causal sampler followed by a causal compressor.

Optimal causal code: Sign-of-innovation code

We introduce a class of causal codes, namely, the sign-of-innovation (SOI) codes. We prove that an SOI code is the optimal code as long as the process satisfies the assumptions (P.1)–(P.3) in Section 2.2.

Definition 10 (A Sign-of-innovation (SOI) code)**.** *The SOI code for a continuous-path process $\{X_t\}_{t=0}^T$ consists of an encoding and a decoding policy. Given a symmetric threshold sampling policy in (2.9) that satisfies (S.1)–(S.3), at each stopping time τ_i, $i = 1, 2, \ldots$, the SOI encoding policy generates a 1-bit codeword*

$$U_i = \begin{cases} 1 & \text{if} \quad X_{\tau_i} - \mathbb{E}[X_{\tau_i}|X_{\tau_{i-1}}, \tau_{i-1}] = a_{i-1}(\tau_i, \tau_{i-1}) \\ 0 & \text{if} \quad X_{\tau_i} - \mathbb{E}[X_{\tau_i}|X_{\tau_{i-1}}, \tau_{i-1}] = -a_{i-1}(\tau_i, \tau_{i-1}). \end{cases} \tag{3.6}$$

At time τ_i, the MMSE decoding policy noiselessly recovers X_{τ_i}, $i = 1, 2, \ldots$ via the received codewords U^i,

$$X_{\tau_i} = (2U_i - 1)a_{i-1}(\tau_i, \tau_{i-1}) + \mathbb{E}[X_{\tau_i}|X_{\tau_{i-1}}, \tau_{i-1}], \tag{3.7}$$

and uses (2.10) as the estimate of X_t until U_{i+1} arrives.

Theorem 5. *In either finite or infinite time horizon, for a process $\{X_t\}_{t=0}^T$ satisfying assumptions (P.1)–(P.3) in Section 2.2, the SOI code, whose stopping times are decided by the optimal symmetric threshold sampling policy (2.9) with average sampling frequency (2.3) $F = R$, is the optimal causal code.*

Proof. Converse: In Appendix B.1, we show that the DRF (3.5) is lower bounded by the DFF (2.8) as

$$D(R) \geq \underline{D}(R). \tag{3.8}$$

Achievability: We proceed to show that the equality in (3.8) is achievable by the SOI code. Corollary 1.3 implies that the 1-bit codeword in (3.6) together with the recovered samples $\{X_{\tau_j}\}_{j=1}^{i-1}$ suffices to recover $X_{\tau_i}, i = 1, 2, \ldots$ noiselessly at the decoder. Moreover, since $\ell(U_i) = 1$ under a 1-bit SOI compressor, the rate constraint (3.3) is equal the frequency constraint (2.3), i.e., $\mathbb{E}\left[\sum_{i=1}^N \ell(U_i)\right] = \mathbb{E}[N]$. Thus, (3.8) is achieved with equality under the SOI code. □

Theorem 5 shows that the optimal codeword-generating times are the sampling times of the optimal causal sampling policy. Furthermore, the optimal decoding policy only depends on the thresholds of the sampling policy and the sampling time stamps. Thus, finding the optimal causal code is simplified to finding the optimal causal sampling policy. Using the optimal causal sampling policy below Lemma 1 in Section 2.3 and Theorem 5, one can easily obtain the optimal causal code for the Wiener process, the Lévy process (2.17), and the OU process (2.20) in the infinite time horizon $T = \infty$:

- (Wiener process) The SOI code generates 1-bit codewords U_i (3.6) at

$$\tau_{i+1} = \inf \left\{ t \geq \tau_i : |W_t - W_{\tau_i}| \geq \sqrt{\frac{1}{R}} \right\}, \tag{3.9}$$

  and the DRF is equal to

$$D(R) = \frac{1}{6R}. \tag{3.10}$$

  Fig. 2.2 in Section 2.3 shows the SOI encoding policy for the Wiener process. The gap between horizontal lines represents the sampling threshold $\sqrt{\frac{1}{R}}$. A down-arrow appears if the process innovation $W_t - W_{\tau_i}$ crosses the negative threshold, and codeword $U_i = 0$ is transmitted. An up-arrow appears if the process innovation $W_t - W_{\tau_i}$ crosses the positive threshold, and codeword $U_i = 1$ is transmitted.

- (Continuous Lévy process (2.17)) The SOI code generates 1-bit codewords U_i (3.6) at

$$\tau_{i+1} = \inf \left\{ t \geq \tau_i : |X_t - \bar{X}_t| \geq c\sqrt{\frac{a}{R}} \right\}, \tag{3.11}$$

  and the DRF is equal to

$$D(R) = \frac{ac^2}{6R}. \tag{3.12}$$

- (OU process in (2.20)) The SOI code generates 1-bit codewords U_i (3.6) at

$$\tau_{i+1} = \inf \left\{ t \geq \tau_i : |X_t - \bar{X}_t| \geq \sqrt{R_1^{-1}\left(\frac{1}{R}\right)} \right\}, \tag{3.13}$$

  and the DRF is given by

$$D(R) = R \cdot R_2\left(R_1^{-1}\left(\frac{1}{R}\right)\right), \tag{3.14}$$

  where $R_1(v)$ and $R_2(v)$ are defined in (2.23)–(2.24).

As a consequence of Theorem 5, under the optimal causal encoding policy, i.e., the SOI encoding policy, the MMSE decoding policy in (3.2) reduces to

$$\hat{X}_t = \mathbb{E}[X_t | U^i, \tau^i],\ t \in [\tau_i, \tau_{i+1}), \tag{3.15}$$

which does not rely on the knowledge that the next codeword has not yet arrived, i.e., $t < \tau_{i+1}$.

Separation of sampling and compressing

Theorem 5 implies that the optimal encoding policy can be implemented as a sampler followed by a compressor, see Fig. 3.2. The sampler takes measurements of the

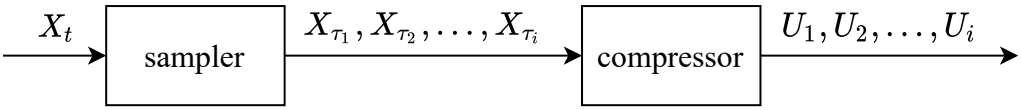


Figure 3.2: Decomposition of the encoder.

source process under the optimal causal sampling policy, i.e., the symmetric threshold sampling policy (2.12), and outputs samples without delay to the compressor. Upon receiving a new sample, the compressor immediately generates a codeword under the SOI compressing policy described in Definition 10.

To gain insight into the tradeoffs between the sampling frequency F at the sampler and the rate per sample R_s at the compressor, we define an (F, R_s, d, T) causal frequency- and rate-constrained code.

Definition 11 (An (F, R_s, d, T) causal frequency- and rate-constrained code)**.** *An (F, R_s, d, T) causal frequency- and rate-constrained code for the source process $\{X_t\}_{t=0}^T$ is a triplet of causal sampling, compressing and decoding policies:*

1. *The causal sampling policy defined in Definition 8-1 satisfies the sampling frequency constraint (2.3);*
2. *The causal compressing policy consists of a sequence of encoders $e_T \triangleq \{\mathsf{e}_1, \mathsf{e}_2, \ldots\}$ with $\mathsf{e}_i \colon \mathbb{R}^i \times \mathbb{Z}^{i-1} \times \mathbb{R}^i \to \mathbb{Z}$ that forms a codeword $U_i \in \mathbb{Z}$ at time τ_i via*

$$U_i = \mathsf{e}_i \left(\{X_{\tau_j}\}_{j=1}^i, U^{i-1}, \tau^i \right); \tag{3.16}$$

The codewords' lengths must satisfy

$$\frac{1}{\mathbb{E}[N]}\mathbb{E}\left[\sum_{i=1}^{N}\ell(U_i)\right] \leq R_s \text{ (bits per sample)}; \tag{3.17}$$

3. *The decoding policy causally maps the received codewords and the sampling times to a continuous-time estimate* (3.15).

The causal sampling, causal compressing, and decoding policies must satisfy the long-term MSE constraint in (3.4).

We quantify the tradeoffs among the sampling frequency F, the rate per sample R_s, and the achievable distortion d using the distortion-frequency-rate function (DFRF) defined below.

Definition 12 (Distortion-frequency-rate function(DFRF))**.** *The DFRF for causal frequency and rate-constrained sampling of the process $\{X_t\}_{t=0}^{T}$ is the minimum distortion achievable by causal frequency-F and rate-R_s codes:*

$$D(F,R_s) \triangleq \inf\{d : \exists\ (F,R_s,d,T) \text{ causal frequency- and rate-constrained code satisfying (S.1), (S.2), (S.3)}\}. \tag{3.18}$$

Theorem 5 indicates that the DRF and the DFRF are related as follows.

Corollary 5.1. *In causal coding of a process $\{X_t\}_{t=0}^{T}$ satisfying assumptions (P.1)–(P.3) in Section 2.2, the DRF* (3.5) *and the DFRF* (3.18) *satisfy*

$$D(R) = \min_{\substack{F>0,R_s\geq 1:\\ FR_s\leq R}} D(F,R_s) \tag{3.19a}$$

$$= D(R,1), \tag{3.19b}$$

where (3.19b) *is achieved by the SOI code in Definition 10.*

Proof. First, every pair of causal sampling policy and causal compressing policy in Definition 11 is a causal encoding policy in Definition 8 provided that $FR_s \leq R$. Thus, the DRF $D(R)$ is upper bounded as

$$D(R) \leq \min_{\substack{F>0,R_s\geq 1:\\ FR_s\leq R}} D(F,R_s). \tag{3.20}$$

Second, the SOI code can be implemented as the symmetric threshold sampling policy with sampling frequency R samples per sec followed by an SOI compressor with $R_s = 1$ bit per sample. Thus, it holds that

$$D(R) = D(R, 1). \tag{3.21}$$

From (3.20)–(3.21), we conclude (3.19). □

Corollary 5.1 illuminates the working principle of the optimal causal code for the stochastic processes considered in Section 2.2: the optimal encoding policy transmits 1-bit codewords, representing the signs of process innovations, as frequently as possible. In other words, the optimal causal code uses the minimum compression rate (1 bit per sample) in exchange for the maximum average sampling frequency R (samples per sec).

3.4 Optimal causal rate-constrained deterministic sampling

In this section, we first define the informational distortion-rate function (IDRF) and the information distortion frequency-rate function (IDFRF) under deterministic sampling policies for the Wiener process. We then show the optimal deterministic sampling policy that achieves the IDRF for the Wiener process and display the relation between the IDRFR and the IDRF.

A sampling policy is *deterministic* if its sampling times $\tau_1, \tau_2, \ldots, \tau_N$ (2.1) are deterministic. Under a deterministic sampling policy, the total number of samples N within the time horizon $[0, T]$ is a constant. We denote by $\mathcal{C}_T$ and Π_T^{DET} the set of all compressing policies in Definition 8 and the set of all deterministic sampling policies over the time horizon $[0, T]$, respectively.

We form an (R, d, T) causal rate-constrained code with deterministic sampling by restricting the causal sampling policy in an (R, d, T) causal rate-constrained code in Definition 8 to a deterministic sampling policy and simplifying $\hat{X}_t$ to (3.15).

We define the operational DRF $D_{\mathrm{DET}}^{\mathrm{op}}(R)$ for source process $\{X_t\}_{t=0}^T$ under deterministic sampling policies as:

$$D_{\mathrm{DET}}^{\mathrm{op}}(R) \triangleq \limsup_{T\to\infty} \inf\{d\colon (R, d, T) \text{ causal rate-constrained code with deterministic sampling}\}, \tag{3.22}$$

which can be decomposed as (Appendix B.3)

$$D_{\mathrm{DET}}^{\mathrm{op}}(R) = \limsup_{T\to\infty} \inf_{\pi_T \in \Pi_T^{\mathrm{DET}}} \frac{1}{T}\Bigg\{ \mathbb{E}\left[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (X_t - \bar{X}_t)^2 \mathrm{d}t\right] \tag{3.23a}$$

$$+ \inf_{\substack{\{f_t\}_{t=0}^{T} \in \mathcal{C}_T: \\ \text{(3.3a)}}} \mathbb{E}\left[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (\bar{X}_t - \hat{X}_t)^2 \mathrm{d}t\right]\Bigg\}, \tag{3.23b}$$

where $\bar{X}_t$ is defined in (2.10) and $\hat{X}_t$ is defined in (3.15). For the Wiener process, $\bar{X}_t = W_{\tau_i}$, $\hat{X}_t = \hat{W}_{\tau_i}$. The expectation on the right side of (3.23a) is the distortion due to causally estimating the source process from its samples. The expectation in (3.23b) is the distortion due to quantization.

The informational counterpart of $D_{\mathrm{DET}}^{\mathrm{op}}(R)$ for the Wiener process is defined below.

Definition 13 (Informational distortion-rate function (IDRF))**.** *The IDRF for the Wiener process under deterministic sampling policies is defined as*

$$D_{\mathrm{DET}}(R) \triangleq \limsup_{T\to\infty} \inf_{\pi_T \in \Pi_T^{\mathrm{DET}}} \frac{1}{T}\Bigg\{ \mathbb{E}\Bigg[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (W_t - W_{\tau_i})^2 dt\Bigg] \tag{3.24a}$$

$$+ \inf_{\substack{\bigotimes_{i=1}^{N} P_{\hat{W}_{\tau_i}|W^{\tau_i},\hat{W}^{\tau_{i-1}}}: \\ \frac{I(W^{\tau_N}\to\hat{W}^{\tau_N})}{T} \le R}} \mathbb{E}\Bigg[\sum_{i=1}^{N} (\tau_{i+1} - \tau_i)(W_{\tau_i} - \hat{W}_{\tau_i})^2\Bigg]\Bigg\}. \tag{3.24b}$$

The minimization problem (3.24b) in $D_{\mathrm{DET}}(R)$ is the causal IDRF for the discrete-time stochastic process formed by the samples. Note that (3.24b) is minimized over the directed information rate, which gives an information-theoretic lower bound to the rate in (3.3a). According to [77, Sec. II-C], we have

$$D_{\mathrm{DET}}^{\mathrm{op}}(R) \ge D_{\mathrm{DET}}(R). \tag{3.25}$$

To gain insight into the tradeoffs between the sampling frequency F at the sampler and the rate per sample R_s at the compressor, we introduce informational distortion-frequency-rate function below.

Definition 14 (Informational distortion-frequency-rate function (IDFRF))**.** *The IDFRF for the Wiener process under deterministic sampling policies is defined as*

$$D_{\mathrm{DET}}(F, R_s) \triangleq \limsup_{T\to\infty} \inf_{\substack{\pi_T \in \Pi_T^{\mathrm{DET}}: \\ \text{(2.3a)}}} \frac{1}{T}\Bigg\{ \mathbb{E}\Bigg[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (W_t - W_{\tau_i})^2 \mathrm{d}t\Bigg] \tag{3.26a}$$

$$+ \inf_{\substack{\bigotimes_{i=1}^{N} P_{\hat{W}_{\tau_i}|W^{\tau_i},\hat{W}^{\tau_{i-1}}}: \\ \frac{I(W^{\tau_N}\to\hat{W}^{\tau_N})}{N}\le R_s}} \mathbb{E}\Bigg[\sum_{i=1}^{N} (\tau_{i+1}-\tau_i)(W_{\tau_i} - \hat{W}_{\tau_i})^2\Bigg]\Bigg\}. \tag{3.26b}$$

Similar to (3.24b), the optimization problem in (3.26b) is the causal IDRF for the Guass-Markov (GM) process formed by the samples, but the rate in (3.26b) is the rate per sample R_s rather than the rate per second R in (3.24b).

We show the optimal deterministic sampling policy that achieves the IDRF.

Theorem 6. *In causal coding of the Wiener process, the uniform sampling policy with the sampling interval equal to*

$$\tau_{i+1} - \tau_i = \frac{1}{R},\ i = 0, 1, 2, \ldots, \tag{3.27}$$

achieves

$$D_{\mathrm{DET}}(R) = \min_{f>0, R_s\ge 1:\, fR_s\le R} D_{\mathrm{DET}}(F, R_s) \tag{3.28}$$

$$= D_{\mathrm{DET}}(R, 1) \tag{3.29}$$

$$= \frac{5}{6R}. \tag{3.30}$$

Proof sketch. See details in Appendix B.4. In Lemma 11, we write $D_{\mathrm{DET}}(F, R_s)$ in (3.26) as $\limsup_{N\to\infty} D_N(F, R_s)$ and write $D_N(F, R_s)$ as a minimization problem building on existing results on the causal IDRF (3.26b) of discrete-time GM processes. In Lemma 12, we provide a lower bound on $D_N(F, R_s)$. In Lemma 13, we provide an upper bound on $D_N(F, R_s)$ achieved by uniform sampling. In Lemma 14, we show that the lower bound and the upper bound coincide as $N \to \infty$ and obtain

$$D_{\mathrm{DET}}(F, R_s) = \frac{1}{2F} + \frac{1}{F(2^{2R_s} - 1)}. \tag{3.31}$$

In Lemma 15, we prove (3.28) by showing that the minimization in (3.28) can be interchanged with the limit in $D_{\mathrm{DET}}(F, R_s)$. To prove (3.29), it remains to minimize

$D_{\mathrm{DET}}(F, R_s)$ in (3.28) over feasible F and R_s:

$$\min_{F>0, R_s\geq 1:\, FR_s\leq R} D_{\mathrm{DET}}(F, R_s) = \min_{R_s\geq 1} D_{\mathrm{DET}}\left(\frac{R}{R_s}, R_s\right) \tag{3.32a}$$

$$= D_{\mathrm{DET}}(R, 1) \tag{3.32b}$$

$$= \frac{1}{2R} + \frac{1}{3R} = \frac{5}{6R}, \tag{3.32c}$$

where (3.32a) holds because $D_{\mathrm{DET}}(F, R_s)$ in (3.31) decreases monotonically in F for any given $R_s \geq 1$, and (3.32b) holds because $D_{\mathrm{DET}}\left(\frac{R}{R_s}, R_s\right)$ increases monotonically as R_s increases in the range $R_s \geq 1$. Thus, the minimum is achieved at $F = R$, $R_s = 1$. Note that $\frac{1}{2R}$ in (3.32c) comes from the sampling distortion and $\frac{1}{3R}$ comes from the causal IDRF for the discrete-time samples. □

Theorem 6 shows that the uniform sampling policy (3.27) operates at the maximum sampling frequency R. Proposition 5.1 and Theorem 6 indicate that the *working principle* of the optimal encoding policy is to transmit 1-bit codewords as frequently as possible.

In the setting of Theorem 6, although evaluating $D_{\mathrm{DET}}(R)$ does not give us an operational compressing policy, we know that the stochastic kernel that achieves the causal IDRF for discrete-time GM processes formed by the samples under uniform sampling policies has the form $\bigotimes_{i=1}^{\infty} P_{\hat{W}_{\tau_i}|W_{\tau_i}-\hat{W}_{\tau_{i-1}},\hat{W}_{\tau_{i-1}}}$ [23, Eq. (5.12)], suggesting that at the encoder, it is sufficient to compress the quantization innovation $W_{\tau_i}-\hat{W}_{\tau_{i-1}}$ only. The decoder computes the estimate $\hat{W}_{\tau_i}$ as $\hat{W}_{\tau_i} = \hat{W}_{\tau_{i-1}}+\mathsf{q}_i(W_{\tau_i}-\hat{W}_{\tau_{i-1}})$, where $\mathsf{q}_i = \mathsf{g}_i \circ \mathsf{f}_i$, $\mathsf{f}_i\left(W_{\tau_i} - \hat{W}_{\tau_{i-1}}\right)$ is the i-th binary codeword U_i, and $\mathsf{g}_i(\cdot) \in \mathbb{R}$ is the quantization representation point of its argument. In practice, one can use the *greedy Lloyd-Max quantizer* [28] that runs the Lloyd-Max algorithm for the quantization innovation in each step based on its prior pdf. Specifically, the prior pdf for the $(i+1)$-th step quantization innovation $W_{\tau_{i+1}} - \hat{W}_{\tau_i}$ can be computed by convolving the pdfs of the quantization error $W_{\tau_i} - \hat{W}_{\tau_i}$ and the process increment $W_{\tau_{i+1}} - W_{\tau_i}$. The globally optimal scheme has a negligible gain over the greedy Lloyd-Max algorithm even in the finite horizon [28].

Fig. 3.3 displays distortion-rate tradeoffs obtained in Theorems 5 and 6 for the Wiener process, as well as a numerical simulation of the uniform sampler in Theorem 6 with the greedy Lloyd-Max quantizer. The symmetric threshold sampling policy followed by the 1-bit SOI compressor leads to a much lower MSE than uniform sampling. Indeed, according to Theorems 5 and 6, $\frac{D_{\mathrm{DET}}(R)}{D(R)} = 5$, and $D^{\mathrm{op}}_{\mathrm{DET}}(R)$ for

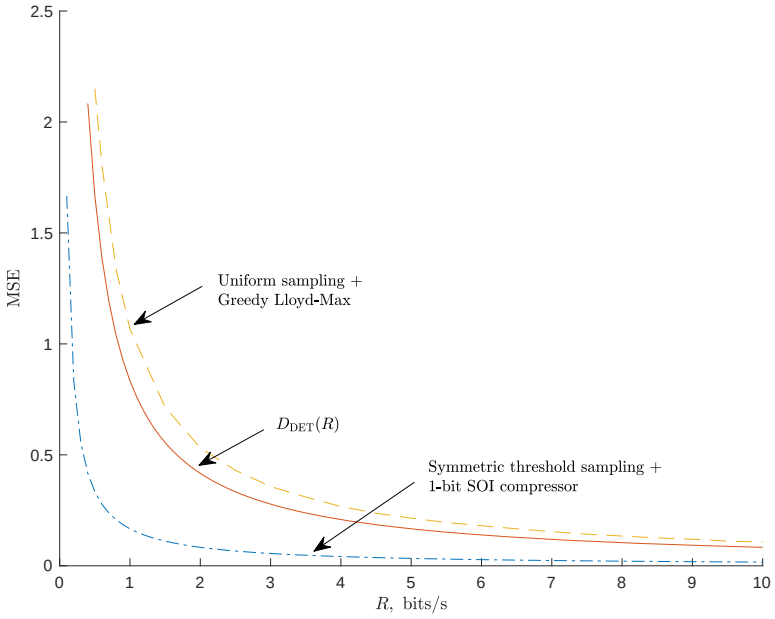


Figure 3.3: MSE versus rate.

the uniform sampling is even higher than $D_{\mathrm{DET}}(R)$ by (3.25). Note that the greedy Lloyd-Max curve is rather close to the $D_{\mathrm{DET}}(R)$ curve, indicating that the IDRF is a meaningful gauge of what is attainable in zero-delay continuous-time causal compression.

3.5 Rate-constrained control

The SOI code introduced in Definition 10 applies to the rate-constrained control scenario in Fig. 3.4. The stochastic plant evolves according to

$$Y_t = X_t + Z_t, \tag{3.33}$$

where X_t is a stochastic disturbance satisfying the assumptions (P.1)–(P.3) in Section 2.2, and Z_t is the additive control signal output from the controller. The encoder observes Y_t, causally decides the stopping times $\tau_1, \tau_2, \ldots$ adapted to the filtration generated by $\{Y_t\}_{t=0}^{T}$, and generates a codeword U_i at each stopping time τ_i based on its past observations $\{Y_t\}_{t=0}^{\tau_i}$. The controller collects the received codewords to causally form the control signal Z_t, with the goal to minimize the mean-square cost on Y_t deviating from the target state 0,

$$\frac{1}{T}\mathbb{E}\left[\int_0^T Y_t^2 \mathrm{d}t\right]. \tag{3.34}$$

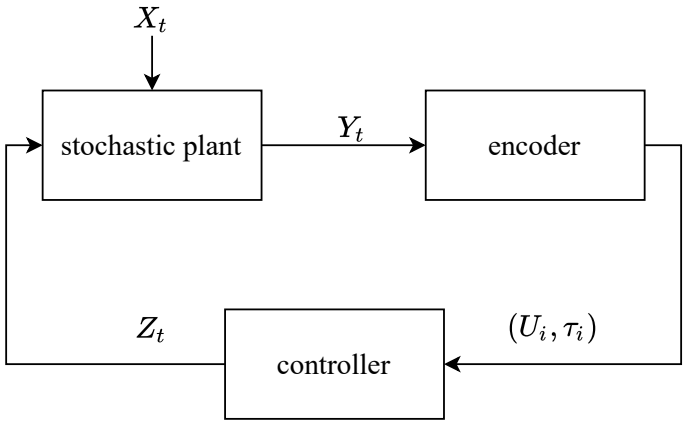


Figure 3.4: Control system.

We aim to find the encoding policy satisfying (S.1)–(S.3) and the control policy that jointly minimize the mean-square cost (3.34) under the communication rate constraint (3.3) between the encoder and the controller.

Proposition 2. *In the rate-constrained control system, the optimal encoding policy that minimizes the mean-square cost in* (3.34) *is the SOI code in Theorem 5, and the optimal control signal is*

$$Z_t = -\hat{X}_t. \tag{3.35}$$

Proof. Given the received codewords U^i and the fact that the next codeword has not been transmitted at $t < \tau_{i+1}$, the optimal control signal Z_t that minimizes (3.34) is indeed the optimal MMSE decoding policy $\hat{X}_t$ in (3.2). Substituting (3.33) and (3.35) into (3.34), we obtain the following MSE,

$$\frac{1}{T}\mathbb{E}\left[\int_0^T (X_t - \hat{X}_t)^2 \mathrm{d}t\right], \tag{3.36}$$

which is the same as (3.4). Thus, the problem of finding the optimal encoding policy in this rate-constrained control system reduces to the problem that we solved in Section 3.3, whose result is given by Theorem 5. □

Under the optimal control policy in Proposition 2, the optimal encoder does not rely on the control signal to decide the codeword generating times.

In the traditional stochastic differential equation (SDE) formulation, e.g., [10, 78, 79], the evolution of the plant is described as

$$\mathrm{d}Y_t = \mathrm{d}X_t + L_t \mathrm{d}t, \tag{3.37}$$

where L_t is the control signal. The state evolutions (3.33) and (3.37) are the same if and only if the control signals in (3.33) and (3.37) are related as

$$\int_0^t L_s \mathrm{d}s = Z_t, \ \forall t \in [0, T]. \tag{3.38}$$

In Appendix B.2, we show how to recover $\{L_t\}_{t=0}^T$ from $\{Z_t\}_{t=0}^T$ using (3.38). Åström and Bernhardsson [10] considered the controlled system in (3.37) with $X_t \leftarrow W_t$ (i.e., Wiener process) and proposed a control policy that injects an impulse signal to drive Y_t to zero once $|Y_t|$ exceeds a threshold. The control signal L_t corresponding to the optimal control $Z_t = -W_{\tau_i}$, $t \in [\tau_i, \tau_{i+1})$, $i = 1, 2, \ldots$ in (3.35) recovers Åström and Bernhardsson's impulse control policy [10] for the Wiener process disturbance.

3.6 Successive refinement via causal rate-constrained sampling

We extend the successive refinement problem in causal frequency-constrained sampling setting (Section 2.4) to causal rate-constrained sampling setting by replacing sampling frequency constraints F^n by bitrate constraints R^n. We show the optimal causal rate-constrained sampling policies at n encoders using the SOI code in Definition 10. We denote the sampling times of the k-th encoder by (2.25), we denote the codewords generated at the sampling times (2.25) by $U_1^{(k)}, U_2^{(k)}, \ldots$, and we denote the codewords and the sampling times generated by the first k encoders by time t by

$$Q_t^{(k)} = \cup_{j=1}^k \left\{ \left(U_i^{(j)}, \tau_i^{(j)} \right) : \tau_i^{(j)} \le t, i = 1, 2, \ldots \right\}. \tag{3.39}$$

We formally define n encoding policies and n decoding policies for successive refinement via causal rate-constrained sampling below.

Definition 15 (An (R^n, d^n, T) causal rate-constrained code for successive refinement)**.** *Fix a source process $\{X_t\}_{t=0}^T$. An (R^n, d^n, T) causal rate-constrained code for successive refinement consists of n causal encoding policies and n causal decoding policies:*

1. *The k-th causal encoding policy is defined in Definition 8 with $\tau_i \leftarrow \tau_i^{(k)}$, $U_i \leftarrow U_i^{(k)}$, $k = 1, 2, \ldots, n$;*

2. *Given the codewords and the sampling times generated by the first k causal encoding policies, i.e., $Q_t^{(k)}$ (3.39), the k-th MMSE decoding policy is*

$$\hat{X}_t^{(k)} \triangleq \mathbb{E}\left[X_t \middle| Q_t^{(k)} \right], \ k = 1, 2, \ldots, n. \tag{3.40}$$

In an (R^n, d^n, T) causal rate-constrained code for successive refinement, the average rates of all n causal encoding policies must satisfy (3.3) with $N \leftarrow N_k$, $R \leftarrow R_k$ $k = 1, 2, \ldots, n$, while all n MSEs must satisfy (3.4) with $\hat{X}_t \leftarrow \hat{X}_t^{(k)}$, $d \leftarrow d_k$, $k = 1, 2, \ldots, n$.

To quantize the tradeoffs between the rates R^n and the MSEs d^n, we introduce the *distortion-rate region*. Fix a source process $\{X_t\}_{t=0}^T$. A rate-distortion tuple (R^n, d^n) is said to be *achievable* if there exists an (R^n, d^n, T) causal rate-constrained code whose first k causal sampling policies form a single causal sampling policy satisfying assumptions (S.1)–(S.3) in Section 2.2 for all $k = 1, 2, \ldots, n$. The distortion-rate region $\mathcal{R}(R^n)$ for the rate vector R^n is the closure of the set of distortions d^n such that (R^n, d^n) is achievable.

Here we continue to use $\pi(F)$ and π_k defined right above Theorem 3 in Section 2.4 to denote the optimal causal sampling policy at frequency F and the causal sampling policy at the k-th encoder. The distortion-rate region $\mathcal{R}(R^n)$ for a class of source processes is shown below.

Theorem 7. *Consider an infinite-horizon, time-homogeneous source process $\{X_t\}_{t=0}^\infty$ satisfying (P.1)–(P.3) whose optimal causal sampling policy $\pi(F)$ (2.12) has a time-invariant sampling threshold, i.e., $\exists$ function $\theta(\cdot)\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $a(t - \tau_i) = \theta(F)$, $\forall t \in [\tau_i, \tau_{i+1})$, $i = 0, 1, \ldots$ If the rate constraints R^n satisfy*

$$\frac{\theta\left(\sum_{j=1}^k R_j\right)}{\theta\left(\sum_{j=1}^{k+1} R_j\right)} = z_k \tag{3.41}$$

for some positive $z_k \in \mathbb{Z}_+$ and for all $k = 1, 2, \ldots, n-1$, then the distortion-rate region $\mathcal{R}(R^n)$ is

$$d_k \geq D\left(\sum_{j=1}^k R_j\right), k = 1, 2, \ldots, n. \tag{3.42}$$

Together with n decoding policies in (3.40), n causal encoding policies that achieve the right side of (3.42) for all $k = 1, 2, \ldots, n$ operate as follows. The k-th encoding policy forms 1-bit codewords using the SOI compressor (3.6) compatible with the symmetric threshold sampling policy $\pi\left(\sum_{j=1}^k R_j\right)$ but only transmits the bits

generated at stopping times in π_k:

$$\pi_1 = \pi(R_1), \text{ if } k = 1, \tag{3.43}$$

$$\pi_k = \pi\left(\sum_{j=1}^{k} R_j\right) \setminus \pi\left(\sum_{j=1}^{k-1} R_j\right), \text{ if } k = 2, \ldots, n. \tag{3.44}$$

Proof. Converse: Given any rate constraints R^n, we show that the achievable distortions $d^n \in \mathcal{R}(R^n)$ are lower bounded as (3.42). For the k-th decoder, the codewords and the sampling times that it receives from the first k encoders, i.e., $Q_t^{(k)}$ (3.39), can be viewed as the codewords and the sampling times generated by a single causal encoding policy under rate $\leq \sum_{j=1}^{k} R_j$. Since the DRF $D(R)$ is a non-increasing function of R, we conclude that the MSE d_k at the k-th decoder is lower bounded as (3.42).

Achievability: We show that for any R^n satisfying (3.41), the causal encoding policies (3.43)–(3.44) together with the MMSE decoding policies (3.40) achieve (3.42). The codewords and the sampling times received by the k-th decoder are equivalent to those generated by the SOI code at rate $\sum_{j=1}^{k} R_j$. Theorem 5 shows that the SOI code achieves the DRF on the right side of (3.42). It remains to show that the rate constraints R^n are satisfied at all n encoders. Since at each sampling time, the SOI encoder only generates 1 bit, the rate of a causal encoding policy is equal to its sampling frequency. According to the proof of Theorem 3, as long as (3.41) holds, the causal sampling policies (3.43)–(3.44) satisfy all sampling frequency constraints R^n simultaneously. □

3.7 Rate-constrained sampling over imperfect channels

We replace the perfect channel in Section 3.2 by a channel with a fixed delay and a binary erasure channel (BEC), respectively. For a channel with a fixed delay, we show that the SOI code for a continuous Lévy process remains optimal. For a BEC, we show that appending the SOI compressor (3.6) to the optimal causal sampling policy for a packet-drop channel in Theorem 4 gives a code that attains the optimal distortion-rate tradeoff.

Channel with delay

We re-consider the communication scenario in Section 2.5 with the sampling frequency constraint (2.3b) in (2.33) replaced by the communication rate constraint (3.3b). We denote by Π and $\mathcal{C}$ the set of all causal sampling policies and the set of

all causal compressing policies in the infinite horizon, respectively. The DRF for a channel with fixed delay δ is defined as

$$D^{\mathrm{ch}}(R) = \inf_{\substack{\pi\in\Pi,\\ \{f_t\}_{t=0}^{\infty}\in\mathcal{C}:\\ \text{(3.3b)}}} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i+\delta}^{\tau_{i+1}+\delta}(X_t-\hat{X}_t^{\mathrm{ch}})^2\mathrm{d}t\right], \tag{3.45}$$

where $\hat{X}_t^{\mathrm{ch}}$ is the MMSE decoding policy of the continuous Lévy process (2.17)

$$\hat{X}_t^{\mathrm{ch}} \triangleq \mathbb{E}[X_t|U^i,\tau^i], t\in[\tau_i+\delta,\tau_{i+1}+\delta). \tag{3.46}$$

We present the optimal causal code that achieves $D^{\mathrm{ch}}(R)$ (3.45).

Proposition 3. *In causal rate-constrained sampling of the continuous Lévy process* (2.17) *with a fixed channel delay δ and the decoding policy* (3.46)*, the optimal causal code remains the SOI code in Theorem 5 and achieves*

$$D^{\mathrm{ch}}(R) = \frac{ac^2}{6R} + ac^2\delta. \tag{3.47}$$

Proof. Converse: We show that the DRF for a channel with fixed delay δ (3.45) is lower bounded by the DFF for a channel with fixed delay δ (2.33) at $F = R$, i.e.,

$$D^{\mathrm{ch}}(R) \geq \underline{D}^{\mathrm{ch}}(R). \tag{3.48}$$

We denote by $\bar{X}_t' \triangleq \mathbb{E}[X_t|\{X_t\}_{t=0}^{\tau_i}]$, $t\in[\tau_i+\delta,\tau_{i+1}+\delta)$ the MMSE estimator given all the past process by time τ_i. Since $\sigma(U^i,\tau_i)\subseteq\sigma(\{X_t\}_{t=0}^{\tau_i})$, the DRF with channel delay $D^{\mathrm{ch}}(R)$ is lower bounded by the right side of (3.45) with $\hat{X}_t^{\mathrm{ch}}\leftarrow\bar{X}_t'$, (3.3b)$\leftarrow$(2.3b), and $F\leftarrow R$. Since by the strong Markov property $\bar{X}_t' = \bar{X}_t^{\mathrm{ch}}$ (2.34), the lower bound reduces to $\underline{D}^{\mathrm{ch}}(R)$.

Achievability: We show that the SOI code achieves the converse bound (3.48). Since the decoder can noiselessly recover the samples from the 1-bit codewords, the MMSE decoding policy $\hat{X}_t^{\mathrm{ch}}$ in (3.46) is equal to $\bar{X}_t^{\mathrm{ch}}$ in (2.34). Since the length $\ell(U_i)=1$, the rate constraint (3.3b) is equal to the frequency constraint (2.3b). □

The first term in (3.47) is the DRF for a delay-free channel and the second term in (3.47) is the penalty due to the channel delay δ. Proposition 3 demonstrates that our SOI code is resilient to channel delay.

Binary erasure channel

We consider the communication scenario in Fig. 3.5, which is similar to that in Section 2.5 except that the packet-drop channel is replaced by a BEC and the sampling frequency constraint is replaced by a rate constraint. Here, we slightly abuse the notations to denote by $0 \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_n$ (2.41) a sequence of *bit*-generating times and to denote by $U_i \in \{0, 1\}$ a bit generated at time τ_i.

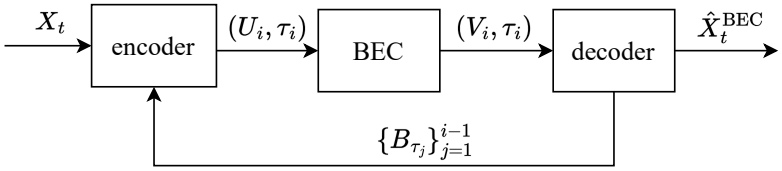


Figure 3.5: System model for causal rate-constrained sampling over a BEC with feedback.

In Fig. 3.5, the encoder transmits bit U_i over a BEC(p), whose channel transition probability is given by $P_{V|U} \colon \{0, 1\} \to \{0, 1, e\}$,

$$P_{V|U}(u|u) = 1 - p, \ \forall u \in \{0, 1\}, \tag{3.49a}$$

$$P_{V|U}(e|u) = p, \ \forall u \in \{0, 1\}. \tag{3.49b}$$

Upon receiving the channel output V_i at each sampling time τ_i, $i = 1, 2, \ldots$, the decoder sends a 1-bit feedback $B_{\tau_i} \in \{0, 1\}$ to inform the encoder whether bit U_i is erased or not. If $B_{\tau_i} = 1$, the bit is not erased, i.e., $V_i = U_i$; otherwise, the bit is erased, i.e., $V_i = e$. Since both the encoder and the decoder know $X_0 = 0$ at $\tau_0 = 0$, it holds that $B_{\tau_0} \triangleq 1$.

We define an $\langle n, d, T \rangle$ causal code for a BEC that transmits n bits of a source process $\{X_t\}_{t=0}^{\infty}$ within an expected time horizon T at an MSE less than or equal to d.

Definition 16 (An $\langle n, d, T \rangle$ causal code for a BEC)**.** *Fix a source process $\{X_t\}_{t=0}^{\infty}$ and fix a BEC with a single-letter transition probability $P_{V|U} \colon \{0, 1\} \to \{0, 1, e\}$. An $\langle n, d, T \rangle$ causal code for a BEC is a pair of encoding and decoding policies. The encoding policy consists of a causal sampling policy and a causal compressing policy.*

1. *The causal* sampling *policy is a collection of n stopping times* (2.41) *defined in Definition 6-1;*

2. *The causal* compressing *policy, characterized by the* $\{0,1\}$*-valued process* $\{f_t\}_{t\geq 0}$ *adapted to the filtration generated by the source process* $\{X_t\}_{t=0}^{\infty}$ *and the feedback bit process* $\{B_{\tau_i}\}_{i=1}^{n}$*, decides the bit* U_i *to transmit at time* τ_i*,* $U_i = f_{\tau_i}$*,* $i = 1, 2, \ldots, n$*.*

Given channel outputs $\{V_j\}_{j=1}^{i}$ *and sampling times* τ^i*, the decoding policy is*

$$\hat{X}_t^{\mathrm{BEC}} \triangleq \mathbb{E}\left[X_t \middle| \{V_j, \tau_j\}_{j\in\mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right)}\right],\ t \in [\tau_i, \tau_{i+1}), \tag{3.50}$$

where $\mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right)$ *is defined in* (2.39)*, which contains all the indices of the successful transmissions by time* τ_i*.*

The expectation of the n*-th sampling time must satisfy* (2.43)*, while the MSE must satisfy*

$$\frac{1}{T}\mathbb{E}\left[\int_0^{\tau_n} (X_t - \hat{X}_t^{\mathrm{BEC}})^2 \mathrm{d}t\right] \leq d. \tag{3.51}$$

The causal sampling policy is assumed to satisfy (S.1) in Section 2.2 and (S.4) in Section 2.5. The decoding policy in (3.50) can be suboptimal since it ignores the knowledge implied by the sampling times of the erased bits.

To quantify the tradeoffs among the number of bits n (2.41), the expected time horizon T (2.43), and the MSE (3.51), we introduce the distortion-sample-time function for a BEC.

Definition 17 (Distortion-sample-time function (DSTF) for a BEC)**.** *Fix a source process* $\{X_t\}_{t=0}^{\infty}$*. The DSTF for a BEC is the minimum MSE* (3.51) *achievable by causal codes with* n *bits and expected horizon* T*:*

$$D^{\mathrm{BEC}}(n, T) \triangleq \inf\{d\colon \exists\ \langle n, d, T\rangle \text{ causal code for a BEC satisfying (S.1), (S.4)}\}. \tag{3.52}$$

For a continuous Lévy process (2.17), we show that appending the SOI compressing policy (3.6) to the optimal causal sampling policy in Theorem 4 gives a causal code for a BEC that achieves $D^{\mathrm{BEC}}(n, T)$.

Theorem 8. *Fix a continuous Lévy process* $\{X_t\}_{t=0}^{\infty}$ *in* (2.17) *and fix a BEC*(p) *with erasure probability* $p < \frac{1}{5}$ (3.49)*. Together with the decoding policy in* (3.50)*, the causal encoding policy that operates as:*

- *if* $B_{\tau_i} = 1$*, then the next sampling time is* (2.47) *and the encoder transmits bit* U_{i+1} *using the SOI compressing policy* (3.6)*;*

- *if* $B_{\tau_i} = 0$*, then the next sampling time is* $\tau_{i+1} = \tau_i$ *and the encoder transmits* $U_{i+1} = U_i$*,* $i = 0, 1, \ldots, n-1$*;*

achieves the DSTF for a BEC

$$D^{\mathrm{BEC}}(n, T) = \frac{ac^2 T}{6(1 + (n-1)(1-p))}. \tag{3.53}$$

Proof. Converse: In Appendix B.10, we show that the DSTF for a BEC $D^{\mathrm{BEC}}(n, T)$ is lower bounded by the DSTF for a packet-drop channel $\underline{D}^{\mathrm{pd}}(n, T)$, i.e.,

$$D^{\mathrm{BEC}}(n, T) \geq \underline{D}^{\mathrm{pd}}(n, T). \tag{3.54}$$

Achievability: The causal encoding policy in Theorem 8 together with the decoding policy in (3.50) achieves the converse bound $\underline{D}^{\mathrm{pd}}(n, T)$ since the decoder can noiselessly recover the samples at the sampling times using the successfully received the 1-bit codewords, i.e., $\hat{X}_t^{\mathrm{BEC}} = \bar{X}_t^{\mathrm{pd}}$. □

The optimal causal encoding policy for a BEC in Theorem 8 operates as follows: if the last bit is not erased, then the encoder follows the symmetric threshold sampling policy (4) and transmits a 1-bit codeword that describes the sign of the process innovation at each sampling time; otherwise, the encoder retransmits the erased bit immediately.

Considering $\frac{n}{T}$ as the rate, letting $R \triangleq \frac{n}{T}$, and taking $n \to \infty$, we rewrite the sampling policy (2.47) and $D^{\mathrm{BEC}}(n, T)$ in (2.48) in terms of rate R as

$$\tau_{i+1} = \inf \left\{ t \geq \tau_i \colon |X_t - \hat{X}_t^{\mathrm{BEC}}| \geq c\sqrt{\frac{a}{(1-p)R}} \right\}, \; i = 0, 1, 2, \ldots; \tag{3.55}$$

$$D^{\mathrm{BEC}}(R) = \frac{ac^2}{6(1-p)R}. \tag{3.56}$$

3.8 Delay-tolerant rate-constrained sampling

In Sections 3.2–3.7, the distortion constraint (3.4) penalizes any delay at the encoder or the decoder. While this is a realistic assumption in some scenarios of remote tracking and control, in this section we consider how the achievable distortion-rate tradeoffs for the Wiener process are affected if the assumption is weakened.

Delay at the encoder and the decoder

In the scenario of encoding the entire process to preserve it for the future, a large delay is permissible. In the extreme, the encoder and the decoder may wait until the end of the Wiener process $\{W_t\}_{t=0}^T$ before coding. This corresponds to the classical scenario of non-causal (block) compression. The informational DRF for this scenario is given by

$$D_{\text{noncausal}}(R) = \lim_{T\to\infty} \inf_{\substack{P_{\{\hat{W}_t\}_{t=0}^T|\{W_t\}_{t=0}^T}: \\ \frac{1}{T}I(\{W_t\}_{t=0}^T;\{\hat{W}_t\}_{t=0}^T)\le R}} \mathbb{E}\left[\frac{1}{T}\int_0^T (W_t-\hat{W}_t)^2 \mathrm{d}t\right]. \tag{3.57}$$

Using reverse water-filling over the power spectrum of the process, Berger [72] derived the informational DRF for the Wiener process:

$$D_{\text{noncausal}}(R) = \frac{2\log_2 e}{\pi^2 R} \quad \text{bits/s}. \tag{3.58}$$

The informational DRF (3.58) is a lower bound to its operational counterpart. As for the achievability, Berger showed that given a rate $R \ge 0$, and $\epsilon > 0$, there exists a code with rate $R+\epsilon$ that achieves the distortion $D_{\text{noncausal}}(R)+\epsilon$. Berger's coding scheme operates as follows [72]: the Wiener process is divided into successive time intervals of a large enough length of T seconds. For each interval, the Karhunen-Loève (KL) coefficients of the process are calculated, and at most $2^{T(R+\epsilon)}$ codewords are used to jointly encode these coefficients with the resulting MSE per second equal to $D_{\text{noncausal}}(R)+\epsilon$. In parallel with encoding the KL expansion coefficients, an integrating delta modulator is employed to encode each endpoint of the length-T intervals with MSE ϵ using ϵ bits on average.

Comparing $D_{\text{noncausal}}(R)$ in (3.58) with $D(R)$ in Theorem 5, we see that, surprisingly, the optimal zero-delay policy outperforms the best infinite delay one:

$$\frac{D(R)}{D_{\text{noncausal}}(R)} \approx 0.57. \tag{3.59}$$

This is because in zero-delay causal coding the timing information is free. Indeed, the decoder time-stamps the arrival of each codeword, and since the channel is delay-free, it knows the codeword-generating times. In classical noncausal (block) lossy compression, no encoder and decoder synchronization is assumed, and thus the encoder is tasked with encoding both the values of the Wiener process and the times corresponding to these values. In many operational scenarios of remote tracking and control, the encoder and decoder are naturally synchronized, providing free timing information. Since Berger's distortion-rate function in (3.58) does not take that into

account, it cannot adequately characterize the fundamental information-theoretic limits in those scenarios.

Delay at the decoder

In the scenario of causal coding where a small delay is tolerable, e.g., speech communication, one can leverage both the free timing information and the coding delay to improve distortion-rate tradeoffs. A *one sample look-ahead decoder* waits for the next codeword U_{i+1} before estimating the source process at time t, $\tau_i \leq t < \tau_{i+1}$, thereby introducing a maximum average delay of $\mathbb{E}[\tau_{i+1} - \tau_i] = \frac{1}{R}$ at the decoder. We show that the one sample look-ahead decoder greatly reduces the MSE compared to the DRF of the Wiener process under the causal real-time estimation.

The one sample look-ahead decoder for the Wiener process W_t is given by

$$\hat{W}_t^{\text{look-ahead}} \triangleq \mathbb{E}[W_t | U^{i+1}, \tau^{i+1}],\ t \in [\tau_i, \tau_{i+1}). \tag{3.60}$$

We append the one sample look-ahead decoder to the SOI encoder in Theorem 5 and calculate the resulting MSE in the infinite horizon:

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (W_t - \hat{W}_t^{\text{look-ahead}})^2 \mathrm{d}t\right]. \tag{3.61}$$

Since the one sample look-ahead decoder (3.60) can noiselessly recover the samples $\{W_{\tau_j}\}_{j=1}^{i+1}$ via the causally received 1-bit codewords U^{i+1}, it reduces to

$$\hat{W}_t^{\text{look-ahead}} = \mathbb{E}[W_t | \{W_{\tau_j}\}_{j=1}^{i+1}, \tau^{i+1}] \tag{3.62a}$$

$$= \mathbb{E}[W_t | W_{\tau_i}, W_{\tau_{i+1}}, \tau_i, \tau_{i+1}],\ t \in [\tau_i, \tau_{i+1}), \tag{3.62b}$$

where (3.62b) holds because

$$W_t - (W_{\tau_i}, W_{\tau_{i+1}}\tau_i, \tau_{i+1}) - (\{W_{\tau_j}\}_{j=1}^{i-1}, \{W_{\tau_j}\}_{j=i+1}^{N}, \{\tau_j\}_{j=1}^{i-1}, \{\tau_j\}_{j=i+1}^{N}) \tag{3.63}$$

is a Markov chain. In particular, when the sampling times are deterministic, $(W_{\tau_i}, W_t, W_{\tau_{i+1}})$ is a Gaussian random vector, thus the estimate in (3.62b) is the linear interpolation between W_{τ_i} and $W_{\tau_{i+1}}$. Under symmetric threshold sampling policies, the samples are not Gaussian, and the linear interpolation is suboptimal. Yet, if in (3.61) we substitute for $\hat{W}_t^{\text{look-ahead}}$ the suboptimal estimate $\frac{W_{\tau_{i+1}}+W_{\tau_i}}{2}$, then the resulting the MSE is equal to $\frac{1}{12R}$, a two-fold improvement over (3.10).

3.9 Conclusion

We study the optimal causal rate-constrained code for a class of continuous Markov processes satisfying regularity conditions (P.1)–(P.3). Prior art on remote estimation and optimal scheduling mostly considered a sampling frequency constraint, whereas, in this work, we introduce a rate constraint. We show that the optimal causal code is the SOI code that transmits 1-bit codewords representing the signs of process innovations at the stopping times decided by the optimal frequency-constrained sampling policy (Theorem 5). The SOI encoder can be implemented as a sampler followed by a compressor. The SOI code can be easily obtained once we know the optimal sampling policy, revealing the close connection between the frequency-constrained and rate-constrained causal sampling problems. The performance of the SOI code for the Wiener process is better than that achieved by the best non-causal code (Section 3.8). This underscores the power of free timing information, which is not explored in the non-causal compressing setting. The key to transmitting information via timing is to use process-dependent rather than deterministic sampling times because the latter contains zero information. The optimal deterministic sampling policy for the Wiener process is the uniform sampling policy (Theorem 6). In either signal-dependent or deterministic sampling setting, the best strategy is to transmit 1-bit codewords as frequently as possible (Corollary 5.1, Theorem 6). For the rate-constrained control, we show that the SOI code minimizes the mean-square cost between the desirable state 0 and the state of the stochastic plant driven by a process satisfying conditions (P.1)–(P.3) (Proposition 2). For the successive refinement via causal rate-constrained sampling, we show that the optimal causal encoding policies can be obtained by appending an SOI compressor to each of the optimal causal sampling policies established in the frequency-constrained setting (Theorem 7). Dropping the assumption that the channel is delay-free, we show that the SOI code of a continuous Lévy process remains optimal for a channel with a fixed delay (Proposition 3). Dropping the assumption that the channel is noiseless, we show that the optimal causal code for a BEC with feedback can be obtained by implementing the optimal causal sampling policy for a packet-drop channel (Section 2.4) followed by an SOI compressor (Theorem 8): if the last bit is successfully delivered, then the SOI compressor generates a new bit once the process innovation crosses either of two symmetric thresholds; if the last bit is erased, the SOI compressor retransmits the erased bit immediately. This shows that the SOI code is resilient to channel delay and noise.

3.10 Future research directions

Based on the findings in Sections 3.2–3.8, we list several interesting directions for future research.

It would be interesting to replace the sampling frequency constraint by the bitrate constraint in open problems 1–4 listed in Section 2.7. The replacement turns the causal frequency-constrained sampling problems into causal rate-constrained sampling problems, which are potentially more challenging. We discuss these open problems one by one.

Causal rate-constrained sampling over a channel with a random delay

One can find the optimal rate-constrained code for a channel with a random delay modeled as a FIFO queue with a random service time. Although the optimal causal sampling policy of the Wiener process in this setting remains a symmetric threshold sampling policy (2.37), we cannot construct the optimal causal rate-constrained code by simply appending a 1-bit SOI compressor to the sampler. Note that the optimal causal sampling policy in (2.37) transmits a new sample after the previous sample is delivered. Since waiting for the delivery of the previous sample may cause the thresholds not to be hit with equality at the time $\tau_i, i = 1, 2, \ldots$, the process innovation $W_{\tau_{i+1}} - W_{\tau_i}$ may not be a binary random variable for all $i = 0, 1, \ldots$ If one insists on using the 1-bit SOI compressor, the error due to quantization will accumulate and blow up.

Causal rate-constrained sampling over a noisy channel

One can find the optimal causal rate-constrained code for a noisy channel (other than BEC in Section 3.7). This is a joint source-channel coding problem extremely sensitive to coding delay. Joint source-channel codes that can quickly incorporate newly arrived information into a continuing transmission like the one we developed in Chapter 4 will be instrumental for making progress in this direction.

Causal rate-constrained sampling for a wider class of source processes

One can find the optimal causal rate-constrained code for a wider class of stochastic processes. The optimal causal code for stochastic processes satisfying symmetry and regularity conditions admits the simple form because 1 bit is enough to describe the binary process innovation noiselessly. We notice that as long as the sampling thresholds and the process innovation all have continuous paths, the simple form of the optimal causal code pertains even for an optimal frequency-constrained sampling

policy with multiple sampling thresholds. At each time t, there must exist a narrowest sampling interval formed by two sampling thresholds that contains the current value of the process innovation. Due to the continuity of the sampling thresholds and the process innovation, one of two boundaries of the interval must be hit at each sampling time. Thus, a 1-bit SOI compressor describes the process innovation noiselessly, and the distortion-rate function coincides with its lower bound: the distortion-frequency function.

If a 1-bit SOI compressor fails to noiselessly describe the process innovation at each sampling time, one has to adopt a more complicated compression scheme. The possible difficulties include:

1) The operational DRF for the causal rate-constrained sampling of a general stochastic process is very hard to solve. One may need to define and solve its informational counterpart using directed information.

2) The difficulties of solving the IDRF are in three aspects. First, even if the IDRF is decomposed as the MSE due to sampling and the MSE due to quantization like (3.23), the compressor may need to compress blocks of a continuous-time process rather than samples. Second, even if compressing samples suffices, the stopping times make the probability distribution of samples intractable. For example, samples of the Wiener process under deterministic sampling times form a Gauss-Markov process, but the sample innovations of the Wiener process under the symmetric threshold sampling policy form a Bernoulli process. Third, solving IDRF for a discrete-time process itself is challenging. For example, for Gauss-Markov processes, the IDRF needs to be solved via semi-definite programming [20].

Causal rate-constrained sampling for a partially observed system

Open problem 4 in Section 2.7 reduces to open problem 3 if minimizing the second term of (2.51) can be considered as a causal frequency-constrained sampling problem. Thus, open problem 4 faces the same difficulties as open problem 3 under a bitrate constraint.

Value of timing information

It would also be interesting to study the timing information. As we have shown that the DRF achieved by the SOI code is even smaller than the DRF achieved by the best non-causal code due to the leverage of the free timing information. Anantharam and Verdú [80] showed that times carry information by evaluating the channel capacity

of a single-server queue with an exponentially distributed service time and designing capacity-achieving block codes. The codewords are independent realizations of a Poisson process, representing vectors of time intervals. Sun et al. [8] distorted the sample delivery times by a FIFO queue. Under the same sampling frequency, the MSE achieved by the optimal causal sampling policy for the queue [8] is an upper bound of the DRF achieved by the SOI code for the noiseless channel. One can quantify the value of timing information by employing a channel that introduces distortions to the times.

Example: One can consider a channel that only allows transmissions at discrete times. This is equivalent to quantizing the real-valued sampling times into discrete values $d, 2d, 3d, \ldots$. If we use the SOI code and assume that the decoder can distinguish the order of bits arrived at the same time, then the quantization of the times can be reviewed as the channel delay. Namely, the decoder can still noiselessly recover the samples but with delays.

Chapter 4

CAUSAL JOINT SOURCE-CHANNEL CODING WITH FEEDBACK

4.1 Introduction

Conventionally, posterior matching is investigated in channel coding and block encoding contexts–the source symbols are equiprobably distributed and are entirely known by the encoder before the transmission. In this chapter, we consider causal joint source-channel coding (JSCC) of streaming source, whose symbols progressively arrive at the encoder at a sequence of deterministic times.

We derive the joint source-channel coding reliability function for streaming over a discrete memoryless channel (DMC) with feedback. We propose a novel *instantaneous encoding phase* that operates during the symbol arriving period and that achieves the JSCC reliability function for streaming when followed by a block encoding scheme that achieves the JSCC reliability function for a classical source whose symbols are fully accessible before the transmission. During the instantaneous encoding phase, the evolving message alphabet is partitioned into groups whose priors are close to the capacity-achieving distribution, and the encoder determines the group index of the actual sequence of symbols arrived so far and applies randomization to exactly match the distribution of the transmitted index to the capacity-achieving one. Surprisingly, the JSCC reliability function for streaming is equal to that for a fully accessible source, implying that the knowledge of the entire symbol sequence before the transmission offers no advantage in terms of the reliability function.

For streaming over a symmetric binary-input DMC, we propose a one-phase *instantaneous small-enough difference (SED) code* that not only achieves the JSCC reliability function, but also, thanks to its single-phase time-invariant coding rule, can be used to stabilize an unstable linear system over a noisy channel. Furthermore, we show that the instantaneous SED code can be used to transmit a streaming source whose symbol arriving times are unknown to the decoder. Using the instantaneous SED code, we derive the JSCC reliability function for a streaming source whose symbol arriving times have limited randomness.

For equiprobably distributed source symbols, we design low complexity algorithms

to implement both the instantaneous encoding phase and the instantaneous SED code. The algorithms group the source sequences into sets we call types, which enable the encoder and the decoder to track the priors and the posteriors of source sequences jointly.

While the reliability function is derived for non-degenerate DMCs, i.e., DMCs whose transition probability matrix has all positive entries, for degenerate DMCs we design a code with instantaneous encoding that achieves zero error for all rates below the Shannon's joint source-channel coding limit.

Most results in this chapter appear in the research papers [81, 82, 83]. The JSCC reliability function for a streaming source whose symbol arriving times have limited randomness in Section 4.6 appears for the first time.

Prior art

Designing good channel block encoding schemes with feedback is a classical problem in information theory [37, 42, 38, 39, 40, 43, 44, 45, 47, 46]. Although feedback cannot increase the capacity of a memoryless channel [36], it renders the design of capacity-achieving codes simpler [37, 42, 38, 39] and improves the tradeoffs between the decoding delay and the reliability [40, 41]. The underlying principle behind capacity-achieving block encoding schemes with feedback [37, 42, 38, 39, 40, 43, 44, 45, 47, 46], termed posterior matching [39], is to transmit a channel input that has two features. First, the channel input is independent of the past channel outputs, representing the new information in the message that the decoder has not yet observed. Second, the probability distribution of the channel input is matched to the capacity-achieving one using the posterior of the message.

While asymptotically achieving the channel capacity ensures the best possible transmission rates in the limit of large delay, optimizing the tradeoff between delay and reliability is critical for time-sensitive applications. The delay-reliability tradeoff is often measured by the reliability function (a.k.a. optimal error exponent), which is defined as the maximum rate of the exponential decay of the error probability at a rate strictly below the channel capacity as the blocklength is taken to infinity. It is a classical fundamental limit that helps to gain insight into the finite blocklength performance of codes via large deviations theorems in probability. In the context of channel coding, the reliability function of a DMC with feedback is first shown by Burnashev [40]. Variable-length channel codes with block encoding that achieve Burnashev's reliability function are proposed in [40, 43, 44, 45, 46]. Burnashev's

[40] and Yamamoto and Itoh (Y-I)'s schemes [43] are structurally similar in that they both have two phases. In the communication phase, the encoder matches the distribution of its output to the capacity-achieving input distribution, while aiming to increase the decoder's belief about the true message. In the confirmation phase, the encoder repeatedly transmits one of two symbols indicating whether the decoder's estimate at the end of the communication phase is correct or not. Caire et al. [44] showed that the code transmitted in the communication phase of the Y-I scheme can be replaced by any non-feedback block channel code, provided that the error probability of the block code is less than a constant determined by the code rate as the blocklength goes to infinity. Naghshvar et al. [45] challenged the convention of using a two-phase code [40, 43, 44] to achieve Burnashev's reliability function by proposing the MaxEJS code, which searches for the deterministic encoding function that maximizes an extrinsic Jensen-Shannon (EJS) divergence at each time. Since the MaxEJS code has a double exponential complexity in the length of the message sequence k, Naghshvar et al. [45] proposed a simplified encoding function for symmetric binary-input DMCs that is referred to as the *small-enough difference* (SED) rule in [47]. The SED encoder partitions the message alphabet into two groups such that the difference between the Bernoulli$\left(\frac{1}{2}\right)$ capacity-achieving distribution and the group posteriors is small. While the SED rule still has an exponential complexity in the length of the message, Antontini et al. [47] designed a systematic variable-length code for transmitting k bits over a binary symmetric channel (BSC) with feedback that has complexity $O(k^2)$. The complexity reduction is realized by grouping messages with the same posterior. Yang et al. [46] generalized the Naghshvar et al.'s SED rule-based code [45] to binary-input binary-output asymmetric channels.

While the message in [40, 43, 44, 45, 47, 46] is equiprobable on its alphabet, the JSCC reliability function for transmitting a non-equiprobable discrete-memoryless source (DMS) over a DMC has also been studied [84, 85, 86, 87]. For fixed-length almost lossless coding without feedback, Gallager [84] derived an achievability bound on the JSCC reliability function, which indicates that JSCC leads to a strictly larger error exponent than separate source and channel coding in some cases; Csiszàr [85] provided achievability and converse bounds on the JSCC reliability function using random coding and type counting; Zhong et al. [86] showed that Csiszàr's achievability bound [85] is tighter than Gallager's bound [84] and provided sufficient conditions for the JSCC reliability function to be strictly larger than the separate source and channel coding reliability function. For variable-length lossy coding with feedback, Truong and Tan [87] derived the JSCC excess-distortion reliability

function under the assumption that 1 source symbol is transmitted per channel use on average. To achieve the excess-distortion reliability function, Truong and Tan [87] used separate source and channel codes: the source is compressed down to its rate-distortion function, and the compressed symbols are transmitted using the Y-I communication phase, while the Y-I confirmation phase is modified to compare the uncompressed source and its lossy estimate instead of the compressed symbol and the estimate thereof. Due to the modification, some channel coding errors bear no effect on the overall decoding error, and the overall decoding error is dominated by the decoding error of the repetition code in the confirmation phase.

While most feedback coding schemes in the literature considered block encoding of a source whose outputs are accessible in their entirety before the transmission, [37, 42, 38, 39, 40, 43, 44, 45, 47, 46, 87], several existing works considered instantaneous encoding of a streaming source [48, 88, 49, 50, 51, 52, 89, 81]. A large portion of them [48, 88, 49, 50, 51] explores instantaneous (causal) encoding schemes for stabilizing a control system. The evolving system state is considered as a streaming data source, the observer instantaneously transmits information about the state to the controller, and the controller injects control signals to the plant. Sahai and Mitter [48] defined the *anytime* capacity at anytime reliability α as the maximum transmission rate R (nats per channel use) such that the decoding error of the first k R-nat symbols at time t decays as $e^{-\alpha(t-k)}$ for any $k \leq t$; they showed that the scalar linear system can be stabilized provided that the logarithm of its unstable coefficient is less than the anytime capacity; they suggested that codes that lead to an exponentially decaying error have a natural tree structure (similar to Schulman's code [88] for interactive computing) that tracks the state evolution over time. Tree coding schemes for stabilizing control systems have been studied in [49, 50]. Assuming that the inter-arrival times of message bits are known by the decoder and that the channel is a BSC, Lalitha et al. [51] proposed an anytime code [48] that achieves a positive anytime reliability and derived a lower bound on the maximum rate that leads to an exponentially vanishing error probability. Instantaneous encoding schemes have also been studied in pure communication settings, where one may evaluate the error exponent [52, 81], consider a streaming source with finite length [89, 81], and allow non-periodic deterministic [51] or random [81] streaming times. Chang and Sahai [52] considered instantaneous encoding of i.i.d. message symbols that arrive at the encoder at consecutive times for the transmission over a binary erasure channel (BEC) with feedback, and showed the zero-rate JSCC error exponent of erroneously decoding the k-th message symbol at time t for fixed k and $t \to \infty$.

Antonini et al. [89] designed a causal encoding scheme for $k < \infty$ streaming bits with a fixed arrival rate over a BSC and showed by simulation that the code rate approaches the channel capacity as the bit arrival rate approaches the transmission rate. In our previous work [81], we proposed a code that uses an adapted SED rule [45] to instantaneously transmit $k < \infty$ randomly arriving bits and that leads to an achievability bound on the reliability function for binary-input DMCs with instantaneous encoding, and designed a polynomial-time version of it. While the instantaneous encoding schemes in [48, 88, 49, 50, 51, 52, 89, 81] employ feedback, transmission schemes for streaming data *without* feedback have been investigated for finite memory encoders [53, 57, 90], for distributed sources [54], and for point-to-point channels in the moderate deviations [56] and the central limit theorem [55, 58].

Streaming data has also been investigated in the field of computer and system sciences. While traditional database management systems only allow one-time queries over a static data set, in the past twenty years, researchers became interested in developing efficient data stream management systems that can handle continuous queries over streaming data [91, 92, 93, 94, 95, 96, 97, 98]. Due to the unbounded size of streaming data, the main challenge of continuous queries is the unbounded memory required to compute the exact answer [91, 96]. To overcome the challenge, various algorithms [95, 97, 98] have been designed to approximate the answer with low memory sizes.

Chapter organization and contribution

In Section 4.3, we propose a novel coding phase–the instantaneous encoding phase–for transmitting a sequence of k source symbols over a DMC with feedback. It performs instantaneous encoding during the arriving period of the symbols. At time t, the encoder and the decoder calculate the priors of all possible symbol sequences using the source distribution and the posteriors at time $t-1$. Then, they partition the evolving message alphabet into groups, so that the group priors are close to the capacity-achieving distribution. In contrast to Naghshvar et al.'s SED rule [45] for symmetric binary-input channels, our partitioning rule is applicable to any DMCs, and it uses group priors instead of group posteriors for the partitioning. Using group priors is necessary because if a new symbol arrives at time t, the posteriors at time $t-1$ are insufficient to describe the symbol sequences at time t. Feedback codes with block encoding [37, 42, 38, 39, 40, 43, 44, 45, 47, 46, 87] only need to consider the posteriors, since block encoding implies that the priors at time t are equal to the

posteriors at time $t-1$. Once the groups are partitioned, the encoder determines the index of the group that contains the true symbol sequence it received so far and applies randomization to match the distribution of the transmitted index to the capacity-achieving one.

In Section 4.4, we derive the JSCC reliability function for the almost lossless transmission of a discrete streaming source over a DMC with feedback. Since allowing the encoder to know the entire source sequence before the transmission will not decrease the reliability function, converse bounds for a classical fully accessible source pertain. We extend Berlin et al.'s converse bound [99] for Burnashev's reliability function to JSCC. For fully accessible sources, we show that the converse is achievable by a variable-length joint source-channel code with block encoding– the MaxEJS code [45]. For a source whose symbols arrive at the encoder with an infinite arriving rate (symbols per channel use) as the source length goes to infinity, we show that the converse is achievable by the buffer-then-transmit code that idles the transmissions and only buffers the arriving symbols during the symbol arriving period and implements a block encoding scheme that achieves the JSCC reliability function for a fully accessible source after the arriving period. For example, a classical fully accessible source has an infinite symbol arriving rate because its symbols arrive all at once. Yet, this buffer-then-transmit code fails to achieve the JSCC reliability function for streaming if the source symbols arrive at the encoder with a finite arriving rate of symbols per channel use. For streaming symbols with an arriving rate greater than $\frac{1}{\underline{H}}\left(H(P_Y^*)-\log\frac{1}{p_{\max}}\right)$, we show that preceding any code with block encoding that achieves the JSCC reliability function for a fully accessible source by our instantaneous encoding phase will make it achieve the block encoding error exponent as if the encoder knew the entire source sequence before the transmission. Here $\underline{H}$ is a lower bound on the information in the streaming source and is equal to the source entropy rate if the source is information stable, $H(P_Y^*)$ is the entropy of the channel output distribution induced by the capacity-achieving channel input distribution, and $p_{\max}$ is the maximum channel transition probability. Thus, surprisingly, the JSCC reliability function for streaming is simply equal to that for a fully accessible source. Furthermore, we show via simulations that the reliability function gives a surprisingly good approximation to the delay-reliability tradeoffs attained by the JSCC reliability function-achieving codes in the ultra-short blocklength regime.

The above discussion highlights the existence of a sequence of codes with instan-

taneous encoding indexed by the length of the source sequence k that achieves the JSCC reliability function as $k \to \infty$. However, in the remote tracking and control scenarios, a single code that can choose to decode any k symbols of a streaming source at any time t with an error probability that decays exponentially with the decoding delay (i.e., an anytime code [48]) is desired. To this end, in Section 4.5, we design the *instantaneous small-enough difference (SED) code*. The instantaneous SED code is similar to the instantaneous encoding phase except that it continues the transmissions after the symbol arriving period, drops the randomization step, and specifies the group partitioning rule to the instantaneous SED rule. The instantaneous SED code is also similar to the instantaneous encoding scheme in our previous work [81] designed for transmitting a streaming source with random symbol arriving times unknown to the decoder, except that [81] used an instantaneous *smallest-difference* rule. The instantaneous smallest-difference rule minimizes the difference between the group priors and the capacity-achieving probabilities, whereas the instantaneous SED rule only drives their difference small enough. The instantaneous SED rule reduces to Naghshvar et al.'s [45] SED rule if the source is fully accessible before the transmission. In contrast to the instantaneous encoding phase followed by a block encoding scheme, the instantaneous SED code only has one phase, namely, it follows the same transmission strategy at each time. For transmitting i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ bits that arrive at the encoder at consecutive times over a BSC(0.05), simulations of the instantaneous SED code show that the error probability of decoding the first $k = [4\colon 4\colon 16]$ bits at times $t \in [4, 70]$, $t \geq k$, decreases exponentially with an anytime reliability $\alpha \simeq 0.172$, outperforming the theoretical anytime reliability of Lalitha et al's anytime code [51]. This implies that the binary instantaneous SED code can be used to stabilize an unstable linear system with bounded noise [48, 88, 49, 50, 51]. Although the achievability of a positive anytime reliability is evidenced by the simulation, it is difficult to prove analytically since one cannot leverage the submartingales and the bounds on the expected decoding time of a block encoding scheme in [40, 45]. Nevertheless, we show that a sequence of instantaneous SED codes indexed by the length of the symbol sequence k achieves the JSCC reliability function for streaming over a Gallager-symmetric [84, p. 94] binary-input DMC. This result is based on our finding that after dropping the randomization step, the instantaneous encoding phase continues to achieve the JSCC reliability function when followed by a reliability function-achieving block encoding scheme, but a cost of increasing the lower bound on the symbol arriving rate to $\frac{1}{\log \frac{1}{p_{S,\max}}} \left(\log \frac{1}{p_{\min}} - \log \frac{1}{p_{\max}}\right)$. Here, $p_{S,\max}$ is the maximum symbol

arriving probability and $p_{\min}$ is the minimum channel transition probability.

In Section 4.6, we consider a streaming source whose symbol arriving times are random and are unknown to the decoder. We generalize the instantaneous SED code to transmit such a source. For a streaming source whose symbol arriving times have limited randomness, we show the JSCC reliability function using the instantaneous SED code. It is equal to the JSCC reliability function for a streaming source with deterministic symbol arriving times since the randomness in the symbol arriving times becomes negligible as the source length goes to infinity.

Since the size of the evolving source alphabet grows exponentially in time t, the complexities of the instantaneous encoding phase and the instantaneous SED code are exponential in time t. In Section 4.7, for the source symbols that are equiprobably distributed, we design low-complexity algorithms for both codes that we term *type-based* codes. The complexity reduction is achieved by judiciously partitioning the evolving source alphabet into *types*. The cardinality of the partition is $O(t)$ for deterministic symbol arriving times and is equal to $O(t^2)$ for random symbol arriving times, i.e., the cardinality is exponentially smaller than the size of the source alphabet. The type partitioning enables the encoder and the decoder to update the priors and the posteriors of the source sequences as well as to partition source sequences in terms of types rather than individual sequences. Since the prior and the posterior updates have a linear complexity in the number of types, and the type-based group partitioning rule has a log-linear complexity in the number of types due to type sorting, our type-based codes only have a log-linear complexity $O(t \log t)$ for deterministic symbol arriving times and have a polynomial complexity $O(t^2 \log t)$ for random symbol arriving times. Although Antonini et al.'s block encoding scheme for BSCs [47] attains a reduction in complexity also by grouping message sequences, the types in Antonini et al.'s scheme [47] are generated all at once by grouping the message sequences that have the same Hamming distance to the received channel outputs, while the types in our type-based instantaneous encoding phase evolve with the arrival of source symbols. The empirical performances of our type-based codes as well as their corresponding original codes are displayed in Section 4.8.

In Section 4.9, for the transmission over a degenerate DMC, i.e., a DMC whose transition matrix contains a zero, we propose a code with instantaneous encoding that achieves zero error for all rates asymptotically below Shannon's JSCC limit. While feedback codes in most prior literature [43, 44, 45, 47, 46, 87] are designed for non-

degenerate DMCs, i.e., a DMC whose transition probability matrix has all positive entries, Burnashev [40, Sec. 6] constructed a channel code for degenerate DMCs that achieves zero error for all rates asymptotically below the channel capacity. Our code extends Burnashev's code [40, Sec. 6] to JSCC and to the streaming source. Similar to [40, 43, 44, 87], our code is divided into blocks, and each block consists of a communication phase and a confirmation phase. Burnashev's [40, Sec. 6] communication phases use a block encoding scheme that can transmit reliably for all rates below the channel capacity. The communication phase in the first block of our scheme uses a code with instantaneous encoding that can transmit reliably for all rates below Shannon's JSCC limit; our ℓ-th communication phase transmits the uncompressed source sequence to avoid compression errors, and uses random coding to establish an analyzable probability distribution of the decoding time. Our confirmation phase is the same as that of Burnashev's code [40, Sec. 6]: the encoder repeatedly transmits a pre-selected symbol that never leads to channel output y if the decoder's estimate at the end of the communication phase is wrong, and transmits another symbol that can lead to y if the estimate is correct. The confirmation phases rely on the degenerate nature of the channel to ensure zero error: receiving a y secures an error-free estimate of the source.

Notations

$\log(\cdot)$ is the natural logarithm. For any positive integer q, we denote $[q] \triangleq \{1, 2, \ldots, q\}$. We denote by $[q]^k$ the set of all q-ary sequences of length equal to k. For a sequence of random variables X_k, $k = 1, 2, \ldots$ and a real number $a \in \mathbb{R}$, we write $X_k \xrightarrow{\text{i.p.}} a$ to denote that X_k converges to a in probability, i.e., $\lim_{k\to\infty} \mathbb{P}[|X_k - a| \geq \epsilon] = 0$, $\forall \epsilon > 0$. For any set $\mathcal{A}$, we denote by $\mathbb{1}_{\mathcal{A}}(x)$ an indicator function that is equal to 1 if and only if $x \in \mathcal{A}$. For two positive functions $f, g\colon \mathbb{Z}_+ \to \mathbb{R}_+$, we write $f(k) = o(g(k))$ to denote $\lim_{k\to\infty} \frac{f(k)}{g(k)} = 0$; we write $f(k) = O(g(k))$ to denote $\limsup_{k\to\infty} \frac{f(k)}{g(k)} < \infty$; we write $f(k) = \Omega(g(k))$ to denote $\limsup_{k\to\infty} \frac{f(k)}{g(k)} > 0$ [100].

4.2 Problem statement

Consider the setup in Fig. 4.1. We formally define the discrete source that streams into the encoder as follows.

Definition 18 (A $(q, \{t_n\}_{n=1}^{\infty})$ discrete streaming source (DSS))**.** *We say that a source is a DSS if it emits a sequence of discrete source symbols $S_n \in [q]$, $n = 1, 2, \ldots$ at times $t_1 \leq t_2 \leq \ldots$, where symbol S_n that arrives at the encoder at time t_n is*

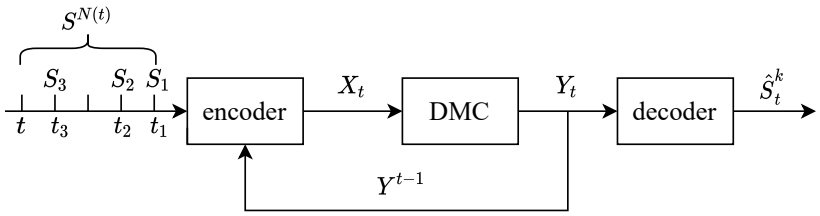


Figure 4.1: Real-time feedback communication system with a streaming source.

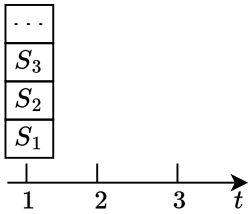


Figure 4.2: A fully accessible source: $1 = t_1 = t_2 = \dots$

distributed according to the source distribution

$$P_{S_n|S^{n-1}}, n = 1, 2, \dots \tag{4.1}$$

Throughout, we assume that the entropy rate of the DSS

$$H \triangleq \lim_{n\to\infty} \frac{H(S^n)}{n} \text{ (nats per symbol)} \tag{4.2}$$

is well-defined and positive; the first symbol S_1 arrives at the encoder at time $t_1 \triangleq 1$; both the encoder and the decoder know the symbol alphabet $[q]$, the arrival times $t_1, t_2, \dots$, and the source distribution (4.1). The DSS reduces to the classical *discrete source* (DS) that is fully accessible to the encoder before the transmission if

$$t_n = 1, \ \forall n = 1, 2, \dots \tag{4.3}$$

Fig. 4.2 and 4.3 display a fully accessible source and a streaming source. Operationally, symbol S_n represents a data packet. We denote the number of symbols that the encoder has received by time t by

$$N(t) \triangleq \max\{n \colon t_n \le t, n = 1, 2, \dots\}. \tag{4.4}$$

Given a DSS (Definition 18) with symbol arriving times $t_1, t_2, \dots$, we denote its *symbol arriving rate* by

$$f \triangleq \liminf_{n\to\infty} \frac{n}{t_n} \text{ (symbols per channel use)} \in [0, \infty]. \tag{4.5}$$

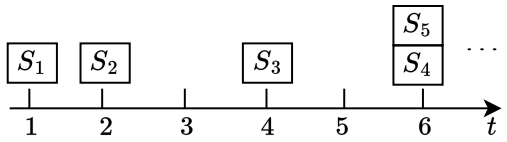


Figure 4.3: A streaming source: $t_1 = 1, t_2 = 2, t_3 = 4, t_4 = t_5 = 6, \ldots$

The symbol arriving rate $f = \infty$ implies that the source symbols arrive at the encoder so frequently that the number of channel uses increases slower than the source length. For example, the DS (4.3) has $f = \infty$. The symbol arriving rate $f < \infty$ implies that the number of channel uses goes to infinity as the source length goes to infinity. For example, if one source symbol arrives at the encoder every $\lambda \geq 1$ channel uses, $\lambda \in \mathbb{Z}_+$, i.e.,

$$t_n = \lambda(n-1) + 1, \tag{4.6}$$

then

$$f = \frac{1}{\lambda}. \tag{4.7}$$

We assume that the channel is a DMC with a single-letter transition probability distribution $P_{Y|X} \colon \mathcal{X} \to \mathcal{Y}$.

Definition 19 (Non-degenerate and degenerate DMCs)**.** *A DMC is non-degenerate if it satisfies*

$$P_{Y|X}(y|x) > 0, \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{4.8}$$

A DMC is degenerate if there exist $y \in \mathcal{Y}$, $x \in \mathcal{X}$, $x' \in \mathcal{X}$, such that

$$P_{Y|X}(y|x) > 0, \tag{4.9a}$$
$$P_{Y|X}(y|x') = 0. \tag{4.9b}$$

A non-degenerate DMC is considered in [40, 43, 44, 45, 47, 46] e.g., a BSC. A degenerate DMC is considered in [40, Sec.6], e.g., a BEC. Fig. 4.4 display examples of DMCs. We denote the capacity of the DMC by

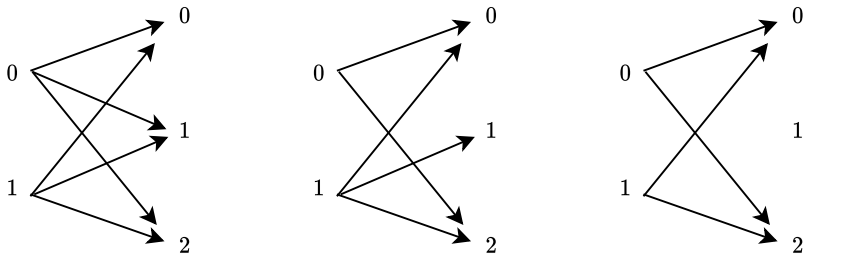
$$C \triangleq \max_{P_X} I(X;Y), \tag{4.10}$$

and we denote the maximum Kullback–Leibler (KL) divergence between its transition probabilities by

$$C_1 \triangleq \max_{x,x' \in \mathcal{X}} D(P_{Y|X=x} || P_{Y|X=x'}). \tag{4.11}$$

Assumption (4.8) posits that C_1 (4.11) is finite.

A DMC is *symmetric* (Gallager-symmetric [84, p. 94]) if the columns in its channel transition probability matrix can be partitioned so that within each partition, all rows are permutations of each other, and all columns are permutations of each other.



(a) non-degenerate (b) degenerate (c) neither non-degenerate nor degenerate

Figure 4.4: A DMC $P_{Y|X} \colon \{0,1\} \to \{0,1,2\}$. An arrow from channel input $x \in \{0,1\}$ to channel output $y \in \{0,1,2\}$ signifies $P_{Y|X}(y|x) > 0$. Channel (a) is a non-degenerate DMC that satisfies (4.8). Channel (b) is a degenerate DMC that satisfies (4.9) with $y = 1$, $x = 1$, $x' = 0$. Channel (c) does not satisfy (4.8)–(4.9) since $y = 1$ is not reachable.

We proceed to define the codes that we use to transmit a DSS over a DMC with feedback. All the codes in this chapter are *variable-length joint source-channel codes with feedback*. We distinguish two classes of codes, one is called a code with instantaneous encoding, and the other is called a code with block encoding. Next, we define the code with instantaneous encoding designed to recover the first k symbols of a DSS at rate R symbols per channel use and error probability ϵ.

Definition 20 (A (k, R, ϵ) code with instantaneous encoding)**.** *Fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS and fix a DMC with a single-letter transition probability distribution $P_{Y|X} \colon \mathcal{X} \to \mathcal{Y}$. An (k, R, ϵ) code with instantaneous encoding consists of:*
1. a sequence of (possibly randomized) encoding functions $\mathsf{f}_t \colon [q]^{N(t)} \times \mathcal{Y}^{t-1} \to \mathcal{X}$, $t = 1, 2, \ldots$ that the encoder uses to form the channel input

$$X_t \triangleq \mathsf{f}_t \left(S^{N(t)}, Y^{t-1} \right); \tag{4.12}$$

2. a sequence of decoding functions $\mathsf{g}_t \colon \mathcal{Y}^t \to [q]^k$, $t = 1, 2, \ldots$ *that the decoder uses to form the estimate*

$$\hat{S}_t^k \triangleq \mathsf{g}_t(Y^t); \tag{4.13}$$

3. a stopping time η_k *adapted to the filtration generated by the channel output* $Y_1, Y_2, \ldots$ *that determines when the transmission stops and that satisfies*

$$\frac{k}{\mathbb{E}[\eta_k]} \geq R \text{ (symbols per channel use)}, \tag{4.14}$$

$$\mathbb{P}[\hat{S}_{\eta_k}^k \neq S^k] \leq \epsilon. \tag{4.15}$$

For any rate $R > 0$, the minimum error probability achievable by rate-R codes with instantaneous encoding and message length k is given by

$$\epsilon^*(k, R) \triangleq \inf\{\epsilon \colon \exists \ (k, R, \epsilon) \text{ code with instantaneous encoding}\}. \tag{4.16}$$

For transmitting a DSS over a non-degenerate DMC with noiseless feedback via a code with instantaneous encoding, we define the *JSCC reliability function for streaming* as

$$E(R) \triangleq \lim_{k \to \infty} \frac{R}{k} \log \frac{1}{\epsilon^*(k, R)}. \tag{4.17}$$

If a DSS satisfies (4.3), i.e., a DS, a code with instantaneous encoding (i.e., causal code) in Definition 20 reduces to a code with block encoding (i.e., non-causal code), and the JSCC reliability function for streaming (4.17) reduces to the JSCC reliability function for a fully accessible source.

Similar to classical codes with block encoding, a (k, R, ϵ) code with instantaneous encoding in Definition 20 is designed to recover only the first k symbols of a DSS, and $E(R)$ (4.17) is achieved by a sequence of codes with instantaneous encoding indexed by the length of the symbol sequence k as $k \to \infty$. We proceed to define a code with instantaneous encoding that decodes the first k symbols at a time $t \geq t_k$ with an error probability that decays exponentially with delay $t - t_k$, for all k and t. Because the decoding time and the number of symbols to decode can be chosen on the fly, this code is referred to as an *anytime* code and can be used to stabilize an unstable linear system with bounded noise over a noisy channel with feedback [48]. We formally define anytime codes as follows.

Definition 21 (A (κ, α) anytime code)**.** *Fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS and fix a DMC with a single-letter transition probability distribution $P_{Y|X}\colon \mathcal{X} \to \mathcal{Y}$. A (κ, α) anytime code consists of:*

1. a sequence of (possibly randomized) encoding functions defined in Definition 20-1;

2. a sequence of decoding functions $\mathsf{g}_{t,k}\colon \mathcal{Y}^t \to [q]^k$ indexed both by the decoding time t and the length of the decoded symbol sequence k that the decoder uses to form an estimate $\hat{S}_t^k \triangleq \mathsf{g}_{t,k}(Y^t)$ of the first k symbols at time t.

For all $k = 1, 2, \ldots$, $t = 1, 2, \ldots$, $t \geq t_k$, the error probability of decoding the first k symbols at time t must satisfy

$$\mathbb{P}[\hat{S}_t^k \neq S^k] \leq \kappa e^{-\alpha(t - t_k)} \tag{4.18}$$

for some $\kappa, \alpha \in \mathbb{R}_+$.

The exponentially decaying rate α of the error probability in (4.18) is referred to as the anytime reliability. While Sahai and Mitter's anytime code in [48, Definition 3.1] is defined to transmit a DSS that emits source symbols one by one at consecutive times, Definition 21 slightly extends [48, Definition 3.1] to a general DSS in Definition 18.

In this chapter, we aim to find $E(R)$ (4.17), the codes with instantaneous encoding that achieve $E(R)$, and an anytime code.

4.3 Instantaneous encoding phase

With the aim of transmitting the first k source symbols of a DSS, we present our instantaneous encoding phase, which specifies the encoding functions $\{\mathsf{f}_t\}_{t=1}^{t_k}$ in Definition 20. We fix a DMC with a single-letter transition probability distribution $P_{Y|X}\colon \mathcal{X} \to \mathcal{Y}$ and capacity-achieving distribution P_X^*, and we fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS with distribution (4.1). We denote the following functions of the channel outputs,

$$\rho_i(Y^t) \triangleq P_{S^{N(t)}|Y^t}(i|Y^t), \tag{4.19}$$

$$\theta_i(Y^{t-1}) \triangleq P_{S^{N(t)}|Y^{t-1}}(i|Y^{t-1}), \tag{4.20}$$

$$\pi_x(Y^{t-1}) \triangleq \sum_{i \in \mathcal{G}_x(Y^{t-1})} \theta_i(Y^{t-1}), \tag{4.21}$$

where we refer to $\rho_i(Y^t)$ and $\theta_i(Y^t)$ as the posterior and the prior of source sequence $i \in [q]^{N(t)}$, respectively; we refer to $\pi_x(Y^{t-1})$ as the prior of the group $\mathcal{G}_x(Y^{t-1})$

corresponding to channel input $x \in \mathcal{X}$ that we specify in (4.24) below. The probability distributions $P_{S^{N(t)}|Y^t}$ and $P_{S^{N(t)}|Y^{t-1}}$ are determined by the code below.

Algorithm: The instantaneous encoding phase operates during times $t = 1, 2, \ldots, t_k$.

At each time t, the encoder and the decoder first update the priors $\theta_i(y^{t-1})$ for all $i \in [q]^{N(t)}$. At symbol arriving times $t = t_n$, $n = 1, 2, \ldots, k$ the prior $\theta_i(y^{t-1})$, $i \in [q]^{N(t)}$ is updated using the posterior $\rho_{i^{N(t-1)}}(y^{t-1})$ and the source distribution (4.1), i.e.,

$$\theta_i(y^{t-1}) = P_{S^{N(t)}|S^{N(t-1)}}\left(i|i^{N(t-1)}\right) \rho_{i^{N(t-1)}}(y^{t-1}), \tag{4.22}$$

where $i^{N(t-1)}$ is the length-$N(t-1)$ prefix of sequence i. At times in-between arrivals, i.e., at $t \in (t_n, t_{n+1})$, $n = 1, 2, \ldots, k-1$, the prior $\theta_i(y^{t-1})$ is equal to the posterior $\rho_i(y^{t-1})$ for all $i \in [q]^{N(t)}$, i.e.,

$$\theta_i(y^{t-1}) = \rho_i(y^{t-1}). \tag{4.23}$$

At each time t, once the priors are updated, the encoder and the decoder partition the message alphabet $[q]^{N(t)}$ into $|\mathcal{X}|$ disjoint groups $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$ such that for all $x \in \mathcal{X}$,

$$\pi_x(y^{t-1}) - P_X^*(x) \leq \min_{i\in\mathcal{G}_x(y^{t-1})} \theta_i(y^{t-1}). \tag{4.24}$$

The partitioning rule (4.24) ensures that the group priors $\{\pi_x(y^{t-1})\}_{x\in\mathcal{X}}$ are close enough to the capacity-achieving distribution $\{P_X^*(x)\}_{x\in\mathcal{X}}$. There always exists a partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$ of $[q]^{N(t)}$ that satisfies the partitioning rule (4.24), since the partition given by the *greedy heuristic* algorithm [101] satisfies it, see the algorithm and the proof in Appendix C.1.

Using the partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$, the encoder and the decoder construct two sets by comparing the group priors $\{\pi_x(y^{t-1})\}_{x\in\mathcal{X}}$ with the capacity-achieving distribution $\{P_X^*(x)\}_{x\in\mathcal{X}}$:

$$\underline{\mathcal{X}}(y^{t-1}) \triangleq \{x \in \mathcal{X} : \pi_x(y^{t-1}) \leq P_X^*(x)\}, \tag{4.25}$$

$$\overline{\mathcal{X}}(y^{t-1}) \triangleq \{x \in \mathcal{X} : \pi_x(y^{t-1}) > P_X^*(x)\}. \tag{4.26}$$

Then, the encoder and the decoder determine a set of probabilities $\{p_{\overline{x}\to\underline{x}}\}_{\overline{x}\in\overline{\mathcal{X}}(y^{t-1}),\underline{x}\in\underline{\mathcal{X}}(y^{t-1})}$ for randomizing the channel input, such that for all $\overline{x} \in \overline{\mathcal{X}}(y^{t-1})$, $\underline{x} \in \underline{\mathcal{X}}(y^{t-1})$, it

holds that

$$\pi_{\overline{x}}(y^{t-1}) - \sum_{\underline{x}\in\underline{\mathcal{X}}(y^{t-1})} p_{\overline{x}\to\underline{x}} = P_X^*(\overline{x}), \tag{4.27}$$

$$\pi_{\underline{x}}(y^{t-1}) + \sum_{\overline{x}\in\overline{\mathcal{X}}(y^{t-1})} p_{\overline{x}\to\underline{x}} = P_X^*(\underline{x}). \tag{4.28}$$

The output of the encoder is formed via randomization as follows. The encoder first determines the group that contains the sequence $S^{N(t)}$ it received so far:

$$Z_t \triangleq \sum_{x\in\mathcal{X}} x \mathbb{1}_{\mathcal{G}_x(y^{t-1})}\left(S^{N(t)}\right). \tag{4.29}$$

Then, the encoder outputs X_t according to

$$P_{X_t|Z_t,Y^{t-1}}(x|z,y^{t-1}) = \begin{cases} \frac{P_X^*(z)}{\pi_z(y^{t-1})}, & \text{if } x = z, z \in \overline{\mathcal{X}}(y^{t-1}), \\ \frac{p_{z\to x}}{\pi_z(y^{t-1})}, & \text{if } x \in \underline{\mathcal{X}}(y^{t-1}), z \in \overline{\mathcal{X}}(y^{t-1}) \\ \mathbb{1}_{\{z\}}(x), & \text{if } z \in \underline{\mathcal{X}}(y^{t-1}). \end{cases} \tag{4.30}$$

The decoder also knows the randomization distribution $P_{X_t|Z_t,Y^{t-1}}$ (4.30), since it knows group priors $\{\pi_x(y^{t-1})\}_{x\in\mathcal{X}}$ (4.24), sets $\overline{\mathcal{X}}(y^{t-1})$ and $\underline{\mathcal{X}}(y^{t-1})$ (4.25)–(4.26), and probabilities $\{p_{\overline{x}\to\underline{x}}\}_{\overline{x}\in\overline{\mathcal{X}}(y^{t-1}),\underline{x}\in\underline{\mathcal{X}}(y^{t-1})}$ (4.27)–(4.28). Due to (4.25)–(4.30), the channel input distribution at time $t = 1, 2, \ldots, t_k$, is equal to the capacity-achieving channel input distribution, i.e., for all $y^{t-1} \in \mathcal{Y}^{t-1}$,

$$P_{X_t|Y^{t-1}}(x|y^{t-1}) = P_X^*(x). \tag{4.31}$$

See the proof of (4.31) in Appendix C.2. Fig. 4.5 below provides an example of group partitioning and channel input randomization.

Upon receiving the channel output $Y_t = y_t$ at time t, the encoder and the decoder update the posteriors $\rho_i(y^t)$ for all possible sequences of source symbols $i \in [q]^{N(t)}$ using the prior $\theta_i(y^{t-1})$, the channel output y_t, and the randomization probability (4.30), i.e.,

$$\rho_i(y^t) = \frac{\sum_{x\in\mathcal{X}} P_{Y|X}(y_t|x) P_{X_t|Z_t,Y^{t-1}}(x|z(i),y^{t-1})}{P_Y^*(y_t)} \theta_i(y^{t-1}), \tag{4.32}$$

where $z(i)$ is the index of the group that contains sequence i, i.e., it is equal to the right side of (4.29) with $S^{N(t)} \leftarrow i$; P_Y^* is the channel output distribution induced by the capacity-achieving distribution P_X^*; (4.32) holds due to (4.31) and the Markov chain $Y_t - X_t - (Z_t, Y^{t-1}) - S^{N(t)}$.

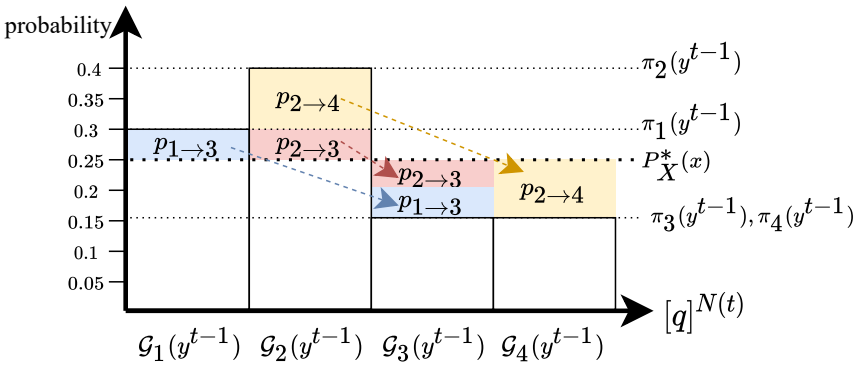


Figure 4.5: An example of group partitioning and channel input randomization for a DMC with uniform capacity-achieving distribution $P_X^*(x) = 0.25$, $\mathcal{X} = [4]$. The horizontal axis represents a partition of 4 groups. The vertical axis represents the prior probabilities of the groups. The source alphabet $[q]^{N(t)}$ is partitioned into $\{\mathcal{G}_x(y^{t-1})\}_{x\in[4]}$ such that the partitioning rule (4.24) is satisfied. Groups $\mathcal{G}_x(y^{t-1}), x \in \{1,2\}$ constitute $\overline{\mathcal{X}}(y^{t-1})$ (4.26) and groups $\mathcal{G}_x(y^{t-1}), x \in \{3,4\}$ constitute $\underline{\mathcal{X}}(y^{t-1})$ (4.25). The probabilities $\{p_{\overline{x}\to\underline{x}}\}_{\overline{x}\in\{1,2\},\underline{x}\in\{3,4\}}$ (4.27)–(4.28) used to randomize transmitted group indices are colored. The randomization matches the probability of transmitting group index $x \in [4]$ to $P_X^*(x)$.

We conclude the presentation of the instantaneous encoding phase with several remarks.

The randomization (4.25)–(4.30) of the instantaneous encoding phase is only used for analysis: Theorem 9 in Section 4.4 continues to hold if the randomization step (4.25)–(4.30) is dropped and the deterministic group index Z_t (4.29) is transmitted, but at a cost of imposing assumptions on the DSS that are stricter than assumptions (a)–(b) in Theorem 9. See Remark 2 in Section 4.4 for details. From the perspective of encoding, the randomization (4.30) turns the encoding function f_t into a stochastic kernel $P_{X_t|S^{N(t)},Y^{t-1}}$. From the perspective of the channel, the randomization $P_{X_t|Z_t,Y^{t-1}}$ (4.30) together with the DMC $P_{Y|X}$ can be viewed as a cascaded DMC with channel input (Z_t, Y^{t-1}). The randomness in (4.29) is not common with the decoder as it only needs to know the distribution $P_{X_t|Z_t,Y^{t-1}}$ to update posterior $\rho_i(y^t)$ in (4.32).

The complexity of the instantaneous encoding phase is $O\left(q^{N(t)} \log q^{N(t)}\right)$ if the classical greedy heuristic algorithm (Appendix C.1) is used for group partitioning (4.24). For equiprobably distributed source symbols, we design an efficient algorithm that reduces the complexity down to $O(t \log t)$ in Section 4.7.

4.4 Joint source-channel coding reliability function

In this section, we show the JSCC reliability function for streaming $E(R)$ (4.17) using the instantaneous encoding phase introduced in Section 4.3. For brevity, we denote the maximum and the minimum channel transition probabilities of a DMC $P_{Y|X}\colon \mathcal{X} \to \mathcal{Y}$ by

$$p_{\max} \triangleq \max_{x\in\mathcal{X},y\in\mathcal{Y}} P_{Y|X}(y|x), \tag{4.33}$$

$$p_{\min} \triangleq \min_{x\in\mathcal{X},y\in\mathcal{Y}} P_{Y|X}(y|x), \tag{4.34}$$

and we denote the maximum symbol arriving probability of the DSS (4.1) by

$$p_{S,\max} \triangleq \max_{n\in\mathbb{N},s\in[q],s'\in[q]^{n-1}} P_{S_n|S^{n-1}}(s|s'). \tag{4.35}$$

Theorem 9. *Fix a non-degenerate DMC with capacity C (4.10), maximum KL divergence C_1 (4.11), and maximum channel transition probability $p_{\max}$ (4.33). Fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS with entropy rate $H > 0$ (4.2) and symbol arriving rate f (4.5). If the DSS has $f < \infty$ (4.5), then we further assume that*

(a) *the information in the DSS is asymptotically lower bounded as*

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} \geq \underline{H}\right] = 1 \tag{4.36}$$

for some $\underline{H} \in (0,\infty)$;

(b) *the symbol arriving rate is large enough:*

$$f > \frac{1}{\underline{H}}\left(H(P_Y^*) - \log\frac{1}{p_{\max}}\right). \tag{4.37}$$

Then, the JSCC reliability function for streaming (4.17) is equal to

$$E(R) = C_1\left(1 - \frac{H}{C}R\right),\ 0 < R < \frac{C}{H}. \tag{4.38}$$

Proof Sketch. The converse proof is in Appendix C.3: allowing the encoder to know the entire source sequence before the transmission will not reduce the JSCC reliability function, therefore converse bounds for a (fully accessible) DS apply. Namely, we lower bound the expected decoding time for any code with block encoding to attain a target error probability using Fano's inequality and a binary hypothesis test. This extends Berlin et al.'s [99] converse bound on Burnashev's reliability function applicable to the channel coding setting to the JSCC setting.

In the achievability proof, we fix a sequence of codes with instantaneous encoding for transmitting the first k symbols of a DSS, $k = 1, 2, \ldots$, over a non-degenerate DMC with feedback, evaluate the asymptotic behavior of the code sequence as $k \to \infty$, and conclude the achievability of $E(R)$ (4.38).

For any (fully accessible) DS, the JSCC reliability function (4.38) is achievable by the MaxEJS code [45, Sec. IV-C], and is achievable by the SED code [45, Sec. V-B] if the channel is a symmetric binary-input DMC (Appendix C.7).

For any DSS with $f = \infty$, including the DS (4.3), the buffer-then-transmit code for k source symbols that achieves $E(R)$ (4.38) operates as follows. It waits until the k-th symbol arrives at time t_k, and at times $t \geq t_k + 1$, applies a JSCC reliability function-achieving code with block encoding for k symbols S^k of a (fully accessible) DS with prior P_{S^k} (4.1) (e.g., the MaxEJS code or the SED code [45]). The buffer-then-transmit code achieves (see details in Appendix C.8)

$$E(R) \geq C_1 \left(1 - \left(\frac{H}{C} + \frac{1}{f}\right) R\right), \tag{4.39}$$

which reduces to $E(R)$ (4.38) for $f = \infty$. Indeed, $f = \infty$ means that the arrival time t_k is negligible compared to the blocklength. The buffer-then-transmit code fails to achieve $E(R)$ (4.38) if $f < \infty$.

For any DSS with $f < \infty$ that satisfies the assumptions (a)–(b) in Theorem 9, the code with instantaneous encoding for k source symbols that achieves $E(R)$ (4.38) implements the instantaneous encoding phase (Section 4.3) at times $t = 1, 2, \ldots, t_k$ and operates as a JSCC reliability function-achieving code with block encoding for k symbols S^k of a (fully accessible) DS with prior $P_{S^k|Y^{t_k}}$ at times $t \geq t_k + 1$, where $Y_1, \ldots, Y_{t_k}$ are the channel outputs generated in the instantaneous encoding phase. For example, we can insert the instantaneous encoding phase before the MaxEJS code (or the SED code for symmetric binary-input DMCs). See Appendix C.9. □

Assumption (a) holds with $\underline{H} = H$ for any information stable source since such sources satisfy $\frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} \xrightarrow{\text{i.p.}} H$ [102]. For example, $\underline{H} = H(S)$ if the source emits i.i.d. symbols. Assumption (b) in Theorem 9 implies

$$f \geq \frac{C}{H} \tag{4.40}$$

since $H(Y|X) \geq \log_2 \frac{1}{p_{\max}}$ and $H \geq \underline{H}$. The symbol arriving rate constraint (4.40) ensures that all coding rates $R < \frac{C}{H}$ are achievable. Otherwise, if (4.40)

is not satisfied and the DSS has $p_{S,\max} < 1$, the rate region achievable by any code with instantaneous encoding is limited to $R \leq f$. The limitation arises because decoding S^k before the final arrival time t_k results in a non-vanishing error probability (Appendix C.12). For example, if the DSS emits i.i.d. symbols with entropy rate $H = 1$ nat per symbol arriving at the encoder every 1000 channel uses, and the DMC has capacity $C = 1$ nat per channel use, then the achievable rate is limited by $\frac{1}{1000}$ symbols per channel use, which is far less than Shannon's JSCC limit $\frac{C}{H} = 1$ symbol per channel use.

Since the (fully accessible) DS (4.3) is a special DSS, Theorem 9 gives the JSCC reliability function (4.38) for a fully accessible source. It generalizes Burnashev's reliability function [40] to the classical JSCC context, and generalizes Truong and Tan's excess-distortion reliability function [87] at zero distortion to the DS with memory and to all rates $R < \frac{C}{H}$.

Remarkably, Theorem 9 establishes that the JSCC reliability function for a streaming source (satisfying assumptions (a)–(b)) is equal to that for a fully accessible source. This is surprising as this means that revealing source symbols only causally to the encoder has no detrimental effect on the reliability function.

While the instantaneous encoding phase in Section 4.3 achieves $E(R)$ (4.38), in fact, any coding strategy during the symbol arriving period that satisfies

$$\lim_{k\to\infty} \frac{I(S^k; Y^{t_k})}{t_k} = C \tag{4.41}$$

achieves $E(R)$ (4.38) when followed by a JSCC reliability function-achieving code with block encoding. This is because (C.42b)–(C.42c) in the achievability proof in Appendix C.9 always hold for such a coding strategy. For equiprobably distributed q-ary source symbols that arrive at the encoder one by one at consecutive times $t = 1, 2, \ldots, k$ and a symmetric q-input DMC, uncoded transmission during the symbol arriving period $t = 1, 2, \ldots, k$ satisfies (4.41) and thus constitutes an appropriate instantaneous encoding phase for that scenario. If $q = 2$, this corresponds to the systematic transmission phase in [47]. Furthermore, even if the instantaneous encoding phase in Section 4.3 drops the randomization (4.25)–(4.30) and transmits Z_t (4.29) as the channel input, it continues to satisfy the sufficient condition (4.41) under a more conservative condition than (4.37) (see Remark 2 below).

Remark 2. *Fix a non-degenerate DMC with the maximum and the minimum channel transition probabilities $p_{\max}$ and $p_{\min}$, and fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS with maximum*

symbol arriving probability $p_{S,\max} < 1$ *and symbol arriving rate* $f < \infty$*. If the DSS satisfies*

(b′) *the symbol arriving rate is large enough:*

$$f > \frac{1}{\log \frac{1}{p_{S,\max}}} \left(\log \frac{1}{p_{\min}} - \log \frac{1}{p_{\max}} \right), \tag{4.42}$$

then the instantaneous encoding phase in Section 4.3 that transmits the non-randomized Z_t (4.29) *as the channel input at each time* $t = 1, 2, \ldots, t_k$ *satisfies* (4.41)*, which means that it achieves* $E(R)$ (4.38)*, the JSCC reliability function for streaming, when followed by a JSCC reliability function-achieving code with block encoding.*

Proof sketch. We show that under assumption (b′), all source priors $\theta_i(y^{t-1})$, $i \in [q]^{N(t)}$, converge pointwise to zero in t during the symbol arriving period $t \in [1, t_k]$ as $k \to \infty$. The convergent source priors and the partitioning rule (4.24) imply that the group priors converge pointwise to the capacity-achieving distribution P_X^*. Since the encoder transmits a group index without randomization as the channel input, the channel input distribution converges to the capacity-achieving distribution, yielding (4.41). See Appendix C.13 for details. □

Note that the result of Remark 2 does not require assumption (a) since $P_{S^n}(s^n) \leq (p_{S,\max})^n$, $\forall s^n \in [q]^n$, already implies that it holds with $\underline{H} \leftarrow \log \frac{1}{p_{S,\max}}$.

Since $\underline{H} \geq \log \frac{1}{p_{S,\max}}$ and $\log \frac{1}{p_{\min}} \geq H(P_Y^*)$, assumption (b′) is stricter than assumption (b). The increase of the threshold is because 1) the channel output distribution P_Y^* in (C.52b) is replaced by $P_{Y_t|Y^{t-1}}(\cdot|\cdot) \geq p_{\min}$ (C.64); 2) in the proof of Remark 2, we show that all the source priors converge *pointwise* to zero (C.67) during the symbol arriving period as $k \to \infty$ using the upper bound $P_{S_n|S^{n-1}}(\cdot|\cdot) \leq p_{S,\max}$, whereas in Theorem 9, we only need that the source prior of the true symbol sequence converges *in probability* to zero (C.55).

4.5 Instantaneous SED code

While the JSCC reliability function-achieving codes with instantaneous encoding in Section 4.3 are designed to transmit the first k symbols of a DSS, and a sequence of such codes indexed by the source length k achieves $E(R)$ (4.38) as $k \to \infty$, we now show an anytime code (Definition 21) termed the instantaneous SED code. In

Section 4.5, we present the algorithm of the instantaneous SED code for a symmetric binary-input DMC. In Section 4.5, we show by simulations that the instantaneous SED code empirically achieves a positive anytime reliability, and thus can be used to stabilize an unstable linear system with bounded noise over a noisy channel. In Section 4.5, we show that if the instantaneous SED code is restricted to transmit the first k symbols of a DSS, a sequence of instantaneous SED codes indexed by the length of the symbol sequence k also achieves $E(R)$ (4.38) for streaming over a symmetric binary-input DMC.

Algorithm of the instantaneous SED code

The instantaneous SED code is almost the same as the instantaneous encoding phase in Section 4.3, expect that 1) it particularizes the partitioning rule (4.24) to the instantaneous SED rule in (4.43)–(4.44) below; 2) its encoder does not randomize the channel input and transmits Z_t (4.29) at each time t; 3) it continues to operate after the symbol arriving period. Fixing a symmetric binary-input DMC $P_{Y|X}\colon \{0,1\} \to \mathcal{Y}$ and fixing a $(q, \{t_n\}_{n=1}^{\infty})$ DSS, we present the algorithm of the instantaneous SED code.

Algorithm: The instantaneous SED code operates at times $t = 1, 2, \ldots$

At each time t, the encoder and the decoder first update the priors $\theta_i\left(y^{t-1}\right)$ for all possible sequences $i \in [q]^{N(t)}$ that the source could have emitted by time t. If $t = t_n$, $n = 1, 2, \ldots$, the prior is updated using (4.22); otherwise, the prior is equal to the posterior (4.23).

Once the priors are updated, the encoder and the decoder partition the source alphabet $[q]^{N(t)}$ into 2 disjoint groups $\{\mathcal{G}_x\}_{x\in\{0,1\}}$ according to the *instantaneous SED rule*, which says the following: if $x, x' \in \{0,1\}$ satisfy

$$\pi_x(y^{t-1}) \geq \pi_{x'}(y^{t-1}), \tag{4.43}$$

then they must also satisfy

$$\pi_x(y^{t-1}) - \pi_{x'}(y^{t-1}) \leq \min_{i\in\mathcal{G}_x(y^{t-1})} \theta_i(y^{t-1}). \tag{4.44}$$

There always exists a partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\{0,1\}}$ that satisfies the instantaneous SED rule (4.43)–(4.44) since the partition that attains the smallest difference $|\pi_0(y^{t-1}) - \pi_1(y^{t-1})|$ satisfies it [45, Appendix III-E].

Once the source alphabet is partitioned, the encoder transmits the index Z_t (4.29) of the group that contains the true source sequence $S^{N(t)}$ as the channel input.

Upon receiving the channel output $Y_t = y_t$ at time t, the encoder and the decoder update the posteriors $\rho_i\left(y^t\right)$ for all $i \in [q]^{N(t)}$ using the priors $\theta_i\left(y^{t-1}\right)$ and the channel output y_t, i.e.,

$$\rho_i\left(y^{t-1}\right) = \frac{P_{Y|X}(y_t|z(i))}{\sum_{x\in\mathcal{X}} P_{Y|X}(y|x)\pi_x(y^{t-1})}\theta_i(y^{t-1}), \tag{4.45}$$

where $z(i)$ is the index of the group that contains sequence i, i.e., it is equal to the right side of (4.29) with $S^{N(t)} \leftarrow i$.

The maximum a posteriori (MAP) decoder estimates the first k symbols at time t as

$$\hat{S}_t^k \triangleq \arg\max_{i\in[q]^k} P_{S^k|Y^t}(i|Y^t). \tag{4.46}$$

We conclude the presentation of the algorithm with several remarks.

We call the group partitioning rule in (4.43)–(4.44) instantaneous small-enough difference (SED) rule since it reduces to Naghshvar et al.'s SED rule [45] if the source is fully accessible to the encoder before the transmission. The rule ensures that the difference between a group prior $\pi_x(y^{t-1})$ and its corresponding capacity-achieving probability $P_X^*(x) = \frac{1}{2}$, $x \in \{0, 1\}$ is bounded by the source prior on the right side of (4.44).

Instantaneous SED code is an anytime code

We first provide numerical evidence showing that the instantaneous SED code is an anytime code: it empirically attains an error probability that decreases exponentially as (4.18). We then determine which unstable scalar linear systems can be stabilized by the instantaneous SED code.

In Fig. 4.6, we display the error probability (4.18) of the instantaneous SED code, where the y-axis corresponds to the error probability of decoding the length-k prefix of a DSS at time t (4.18). At each time t, we generate a Bernoulli$\left(\frac{1}{2}\right)$ source bit and a realization of a BSC(0.05), run these experiments for 10^5 trials, and obtain the error probability (4.18) by dividing the total number of errors by the total number of trials. To reduce the implementation complexity, we simulate the type-based version of the instantaneous SED code in Section 4.7, which has a log-linear complexity. The type-based version is an approximation of the exact instantaneous SED code since it uses an approximating instantaneous SED rule and an approximating decoding rule to mimic the instantaneous SED rule (4.43)–(4.44) and the MAP decoder (4.46), respectively, however, it performs remarkably close to the original instantaneous

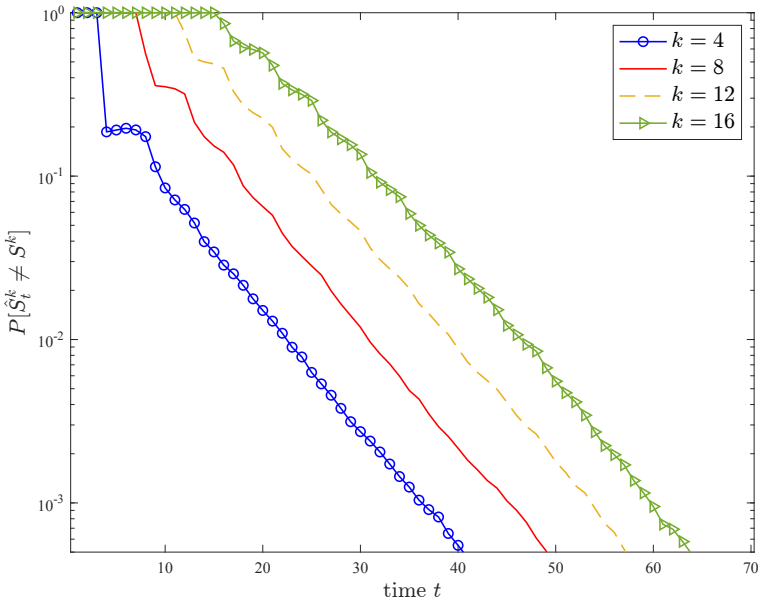


Figure 4.6: The error probability $\mathbb{P}[\hat{S}_t^k \neq S^k]$ of decoding the first k symbols of a DSS at time t achieved by the type-based instantaneous SED code (Section 4.7). The DSS emits a Bernoulli$\left(\frac{1}{2}\right)$ bit at times $t = 1, 2, \ldots$. The channel is a BSC(0.05).

SED code. See Section 4.7 for details. The slope of the curves corresponds to the anytime reliability α (4.18) of the instantaneous SED code. The anytime reliability for the source and the channel in Fig. 4.6 is approximately equal to $\alpha \approx 0.172$. The simulation results in Fig. 4.6 align with our expectation: the error probability decays exponentially with delay $t - k$ (4.18), implying that the instantaneous SED code is an anytime code.

We proceed to display the unstable scalar linear system that can be stabilized by the instantaneous SED code. Consider the scalar linear system in Fig. 4.7, $Z_{t+1} = \lambda Z_t + U_t + W_t$, where $\lambda > 1$, Z_t is the real-valued state, U_t is the real-valued control signal, $|W_t| \leq \frac{\Omega}{2}$ is the bounded noise, and the initial state is $Z_1 \triangleq 0$. At time t, the observer uses the observed states Z^t as well as the past channel feedback Y^{t-1} to form a channel input X_t; the controller uses the received channel outputs Y^t to form a control signal U_t. For a $(q, \{t_n\}_{n=1}^{\infty})$ DSS that emits source symbols one by one at consecutive times $t_n = n$, $n = 1, 2, \ldots$, the anytime rate of a (κ, α) anytime code in Definition 21 is defined as $R_{\text{any}} = \log q$ nats per channel use, e.g., for the DSS in Fig. 4.6, $R_{\text{any}} = \log 2$; the α-anytime capacity $C_{\text{any}}(\alpha)$ is defined as the least upper bound on the anytime rates R_{any} such that the anytime reliability α is achievable [48]. For such a DSS, Sahai and Mitter [48, Lemma 4.1 in Sec. IV-D] showed that the unstable scalar linear system with bounded noise in Fig. 4.7 can be stabilized so

that η-th moment $\mathbb{E}[|Z_t|^\eta]$ stays finite at all times, provided that $C_{\text{any}}(\alpha) > \log \lambda$, $\alpha > \eta \log \lambda$. Thus, the instantaneous SED code can be used to stabilize the η-th moment of the unstable scalar linear system in Fig. 4.7 over a BSC(0.05) for any coefficient

$$\lambda < e^{\min\{R_{\text{any}}, \frac{\alpha}{\eta}\}} \tag{4.47}$$

$$= \min\left\{2, e^{\frac{0.172}{\eta}}\right\}. \tag{4.48}$$

E.g., if $\eta = 2$, then $\lambda < 1.09$. In comparison, the theoretical results in [51, Corollary 1, Fig. 2] with $n = 1$ show that, for the control over a BSC(0.05) in Fig. 4.7, Lalitha et al.'s anytime code is only guaranteed to stabilize the η-th moment of a linear system with $\lambda = 1$.

The control scheme [48, Sec. IV] that stabilizes the system in Fig. 4.7 employs an anytime code and operates as follows. At each time t, the observer computes an R_{any}-nat virtual control signal $\bar{U}_t$ and acts as an anytime encoder to transmit $\bar{U}_t$ as the t-th symbol of a DSS over a noisy channel with feedback. Here, $\bar{U}_t$ controls a virtual state $\bar{Z}_{t+1} = \lambda \bar{Z}_t + W_t + \bar{U}_t$, and is equal to the negative of the R_{any}-nat quantization of $\lambda \bar{Z}_t$. It ensures the boundedness of $\bar{Z}_{t+1}$. Upon receiving the channel output, the controller acts as an anytime decoder to refresh its estimate $\hat{\bar{U}}_t^t$ of $\bar{U}^t$ and forms a control signal U_t that compensates the past estimation errors of the virtual control signals as if the plant $\{Z_s\}_{s=1}^{t+1}$ was controlled by $\hat{\bar{U}}_t^t$ heretofore. As a result of applying U_t, the actual state Z_{t+1} is forced close to the bounded virtual state $\bar{Z}_{t+1}$ with the difference $|Z_{t+1} - \bar{Z}_{t+1}|$ governed by the difference between $\bar{U}_t^t$ and $\hat{\bar{U}}_t^t$. The exponentially decaying with $t-k$ error probability of decoding $\bar{U}_t^k$ achieved by the anytime code together with the bounded $\bar{Z}_{t+1}$ ensures a finite $\mathbb{E}[|Z_{t+1}|^\eta]$. In fact, the full feedback channel in Fig. 4.7 can be replaced by a channel that only feeds the control signal from the controller to the observer, since Z_t, Z_{t-1}, U_{t-1} suffice to compute W_{t-1} and thereby to compute $\bar{U}_t$ at each time t.

Instantaneous SED code achieves $E(R)$

We first restrict the instantaneous SED code in Section 4.5 to transmit only the first k source symbols of a DSS, and we form a sequence of instantaneous SED codes indexed by the length of the symbol sequence k. We then show that the code sequence achieves the JSCC reliability function (4.38) for streaming over a symmetric binary-input DMC as $k \to \infty$.

We restrict the instantaneous SED code in Section 4.5 to transmit the first k symbols of a $(q, \{t_n\}_{n=1}^\infty)$ DSS as follows.

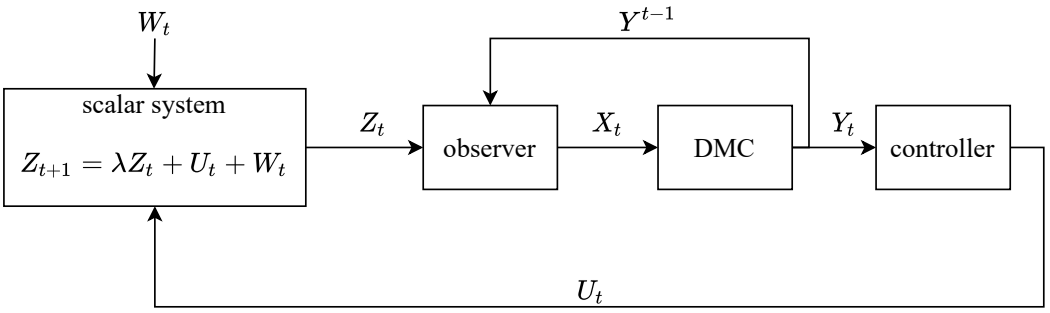


Figure 4.7: A scalar linear system controlled over a noisy channel with noiseless feedback.

1) The alphabet $[q]^{N(t)}$ that contains all possible sequences that could have arrived by time t is replaced by the alphabet $[q]^{\min\{N(t),k\}}$ that stops evolving and reduces to $[q]^k$ after all k symbols arrive at time t_k. As a consequence, for $t \geq t_k + 1$ and all $i \in [q]^k$, the priors $\theta_i(y^{t-1})$ are equal to the corresponding posteriors $\rho_i(y^{t-1})$, the encoder and the decoder partition $[q]^k$ to obtain $\{\mathcal{G}_x(y^{t-1})\}_{x\in\{0,1\}}$, and only the posteriors $\rho_i(y^t)$ are updated.

2) The transmission is stopped and the MAP estimate (4.46) of S^k is produced at the stopping time

$$\eta_k \triangleq \min\left\{t\colon \max_{i\in[q]^k} P_{S^k|Y^t}(i|Y^t) \geq 1-\epsilon\right\}, \epsilon \in (0,1). \tag{4.49}$$

The MAP decoder (4.46) together with the stopping rule (4.49) ensures the error constraint in (4.15), since the MAP decoder (4.46) implies $\mathbb{P}[\hat{S}^k_{\eta_k} = S^k] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{\hat{S}^k_{\eta_k}\}}(S^k)\middle|Y^{\eta_k}\right]\right] = \mathbb{E}\left[\max_{i\in[q]^k} P_{S^k|Y^{\eta_k}}(i|Y^{\eta_k})\right]$, which is lower bounded by $1-\epsilon$ due to the stopping time (4.49).

Theorem 10. *Fix a non-degenerate symmetric binary-input DMC and a $(q, \{t_n\}_{n=1}^{\infty})$ DSS satisfying assumption (b') in Remark 2. The sequence of instantaneous SED codes for transmitting the first k symbols of the DSS achieves $E(R)$ (4.38) as $k \to \infty$.*

Proof. First, we observe that after the symbol arriving period $t \geq t_k + 1$, the instantaneous SED code reduces to the SED code [45, Sec. V-B] because the instantaneous SED rule (4.43)–(4.44) reduces to the SED rule [45, Eq. (50)] if all k source symbols are fully accessible (4.3) to the encoder. The SED code [45] achieves the JSCC reliability function (4.38) for transmitting a fully accessible source over a non-degenerate symmetric binary-input DMC (Appendix C.7).

Second, we observe that during the symbol arriving period $t = 1, 2 \ldots, t_k$, the instantaneous SED code corresponds to dropping the randomization step of the instantaneous encoding phase in Section 4.3. This is because for a symmetric binary-input DMC, (4.43) implies $\pi_{x'}(y^{t-1}) \leq P_X^*(x') = \frac{1}{2}$, thus any partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\{0,1\}}$ that satisfies the instantaneous SED rule (4.43)–(4.44) also satisfies the partitioning rule in (4.24). Therefore, Remark 2 implies that the instantaneous SED code at times $t = 1, 2, \ldots, t_k$ satisfies the sufficient condition (4.41) under assumption (b').

As we have discussed in the proof sketch of Theorem 9, a JSCC reliability function-achieving code with instantaneous encoding can be obtained by preceding a JSCC reliability function-achieving code with block encoding by an instantaneous encoding phase that satisfies (4.41). The two observations above imply that the instantaneous SED code achieves $E(R)$ (4.38) in the setting of Theorem 10. □

4.6 Streaming with random arrivals

Problem statement

While the DSS in Sections 4.2–4.5 emits symbols at a sequence of deterministic arriving times $\{t_n\}_{n=1}^{\infty}$, in this section, we proceed to consider a DSS that emits symbols at random symbol arriving times $\{\tau_n\}_{n=1}^{\infty}$. We continue to denote by $N(t)$ the number of source symbols that have arrived at the encoder by time t. Here, $N(t)$ is a random variable since the symbol arriving times are random. We denote by $\mathcal{Q}_t$ the set of all q-ary sequences of length less than or equal to t, i.e.,

$$\mathcal{Q}_t \triangleq \cup_{1\leq s\leq t}[q]^s. \tag{4.50}$$

We denote by $\boxplus$ the concatenation operation between two strings, e.g., $101 \boxplus 1 = 1011$. We denote by $\boxminus$ the truncation operation that deletes the last bit of a string, e.g., $101\boxminus = 10$.

We say that a source is a q-ary *DSS with random arrivals*, if it emits a sequence of source symbols $S_n \in [q]$, $n = 1, 2, \ldots$ that streams into the encoder at random times

$$\tau_1 < \tau_2 < \ldots \tag{4.51}$$

The strict inequality in (4.51) means that at each time $t = 1, 2, \ldots$, the DSS only emits one symbol or no symbols. As a result, the length of the source sequence arrived by time t is less than or equal to t, i.e.,

$$S^{N(t)} \in \mathcal{Q}_t, t = 1, 2, \ldots \tag{4.52}$$

A new source symbol arrives at time $t+1$ according to the probability distribution

$$P_{S^{N(t+1)}|S^{N(t)}}. \tag{4.53}$$

Since at most one symbol can arrive at any time, the conditional probability distribution $P_{S^{N(t+1)}|S^{N(t)}}(\cdot|s)$ can only place non-zero masses at $S^{N(t+1)} = s$, $S^{N(t+1)} = s \boxplus s'$ for $s' \in [q]$. We assume that both the encoder and the decoder know the symbol arriving probability distribution (4.53), the decoder does not know the exact realizations of the symbol arriving times, and the first symbol arrives at $\tau_1 \triangleq 1$.

We define a code with instantaneous encoding that we use to transmit a DSS with random arrivals over a non-degenerate DMC with feedback.

Definition 22 (A (k, R, ϵ) code with instantaneous encoding for random arrivals)**.** *Fix a q-ary DSS with random arrivals and fix a non-degenerate DMC (4.8) with a single-letter channel transition probability $P_{Y|X}\colon \mathcal{X} \to \mathcal{Y}$. A (k, R, ϵ) code with instantaneous encoding for random arrivals consists of:*
1. A sequence of encoding functions $\mathsf{f}_t\colon \mathcal{Q}_t \times \mathcal{Y}^{t-1} \to \mathcal{X}$, $t = 1, 2, \ldots$ that the encoder uses to form the channel inputs X_t (4.12).
2. A sequence of decoding functions g_t, $t = 1, 2, \ldots$ defined in Definition 20-2.
3. A stopping time η_k defined in Definition 20-3, which satisfies both the rate constraint R (4.14) and the error constraint ϵ (4.15).

For any $R > 0$, the minimum error probability achievable by rate-R codes with instantaneous encoding for random arrivals and message length k is given by

$$\tilde{\epsilon}^*(k, R) \triangleq \min\{\epsilon\colon \exists\ (k, R, \epsilon) \text{ code with instantaneous encoding for random arrivals}\}.$$

For transmitting a DSS with random arrivals over a non-degenerate DMC with noiseless feedback via a code with instantaneous encoding for random arrivals, we define the JSCC reliability function for *random* streaming as

$$\tilde{E}(R) \triangleq \lim_{k\to\infty} \frac{R}{k} \log \frac{1}{\tilde{\epsilon}^*(k, R)}. \tag{4.54}$$

If the symbol arriving times are deterministic, a (k, R, ϵ) code with instantaneous encoding for random arrivals reduces to a (k, R, ϵ) code with instantaneous encoding in Definition 20, and the JSCC reliability function for random streaming $\tilde{E}(R)$ reduces to the JSCC reliability function for streaming $E(R)$ (4.17).

Instantaneous SED code for random arrivals

We generalize the instantaneous SED code in Section 4.5 to a DSS with random arrivals. The key is to allow the encoder and the decoder to track the priors and the posteriors of all possible sequences that could have arrived by time t. We fix a q-ary DSS with random arrivals.

To generalize the anytime instantaneous SED code in Section 4.5 to a DSS with random arrivals, we replace alphabet $[q]^{N(t)}$ in Section 4.5 by alphabet $\mathcal{Q}_t$ that contains all possible source sequences that could have arrived at the encoder by time t. As a consequence, at times $t = 1, 2, \ldots$, for all sequences $i \in \mathcal{Q}_t$, the priors are updated as

$$\theta_i(y^{t-1}) = \sum_{j \in \mathcal{Q}_{t-1}} P_{S^{N(t)}|S^{N(t-1)}}(i|j)\rho_j(y^{t-1}); \tag{4.55}$$

the encoder and the decoder partition $\mathcal{Q}_t$ into groups $\{\mathcal{G}_x(y^{t-1})\}_{x\in\{0,1\}}$ that satisfy the instantaneous SED rule (4.43)–(4.44); the posteriors of all sequences in $\mathcal{Q}_t$ are updated as (4.45).

To restrict the anytime instantaneous SED code for random arrivals described above to transmit only the first k symbols, we replace the alphabet $\mathcal{Q}_t$ that contains all possible sequences that could have arrived by time t by the alphabet $\mathcal{Q}_{\min\{t,k\}}$ that stops evolving at times $t > k$; we equip the code with the stopping time η_k in (4.49) and the MAP decoder in (4.46).

Joint source-channel coding reliability function for random streaming

We derive the JSCC reliability function for random streaming $\tilde{E}(R)$ (4.54) using the instantaneous SED code for random arrivals. Similar to (4.35), we denote

$$\tilde{p}_{S,\max} \triangleq \max_{t\in\mathbb{N}, s\in\mathcal{Q}_t} P_{S^{N(t)}|S^{N(t-1)}}(s|s) + P_{S^{N(t)}|S^{N(t-1)}}(s|s\boxminus), \tag{4.56}$$

where $s\boxminus$ is the sequence after truncating the last (newest) symbol of sequence s.

Theorem 11. *Fix a non-degenerate symmetric binary-input DMC with channel capacity C (4.10), maximum KL divergence C_1 (4.11), maximum channel transition probability $p_{\max}$ (4.33), and minimum channel transition probability $p_{\min}$ (4.34). Fix a DSS with random arrivals that emits symbols $S_1, S_2, \ldots$ at a sequence of random symbol arriving times $\tau_1 < \tau_2 < \ldots$ with entropy rate $H > 0$ (4.2) and $\tilde{p}_{S,\max} < 1$. If the DSS with random arrivals satisfies*

(c) *the symbol arriving times* $\tau_1, \tau_2, \ldots$ *are bounded:* $\exists$ *function* $\mathsf{h}(\cdot)\colon \mathbb{Z}_+ \to \mathbb{Z}_+$, *such that* $\mathsf{h}(n) = o(n)$ *and*

$$\tau_n \leq n + \mathsf{h}(n),\ n = 1, 2, \ldots; \tag{4.57}$$

(d) *the entropy rate of the symbol arriving times is zero, i.e.,*

$$\lim_{n\to\infty} \frac{H(\tau^n)}{n} = 0; \tag{4.58}$$

(e) *assumption* (b′) *in Remark 2 is satisfied with* $f \leftarrow 1$, $p_{S,\max} \leftarrow \tilde{p}_{S,\max}$;

then, the JSCC reliability function for random streaming is equal to

$$\tilde{E}(R) = C_1 \left(1 - \frac{H}{C} R\right),\ 0 < R < \frac{C}{H}. \tag{4.59}$$

Proof sketch. The converse proof is in Appendix C.16: we show that converse bounds on the JSCC reliability function for a fully accessible source apply to $\tilde{E}(R)$. The achievability proof is in Appendix C.17: we show that the detrimental effect on the reliability function due to the randomness in the symbol arriving times vanishes as the source length $k \to \infty$. □

Assumptions (c)–(d) posit that the symbol arriving times of the DSS have limited randomness. An example of such a source emits symbols as follows: among the first n source symbols, $n = 1, 2, \ldots$, there are $\mathsf{h}'(n) \leq \mathsf{h}(n)$ symbols that can randomly select their symbol arriving times within $\mathsf{h}(n)$ time options, and the remaining symbols arrive at deterministic times so that the n-th symbol arriving time is bounded as (4.57). The symbol arriving times in the example satisfies (4.58), see Appendix C.20. Such a source could appear in a time-slotted communication scenario: a source emits packets with most packets arriving at the encoder at deterministic times and a few packets arriving with random and bounded delays due to system deficiencies.

Theorem 11 establishes that the JSCC reliability function for a DSS with random symbol arriving times satisfying assumptions (c)–(e) is equal to that for a DSS with deterministic symbol arriving times, i.e., $\tilde{E}(R) = E(R)$. This means that even though the decoder does not know the exact symbol arriving times, the instantaneous SED code for random arrivals achieves $\tilde{E}(R)$ (4.38) as if the decoder knew the symbol arriving times.

While the instantaneous SED code for random arrivals achieves $\tilde{E}(R)$, in fact, any coding strategy at times $t = 1, 2, \ldots, k + \mathsf{h}(k)$ that satisfies (c.f. (4.41))

$$\lim_{k\to\infty} \frac{I\left(S^k; Y^{k+\mathsf{h}(k)}\right)}{k + \mathsf{h}(k)} = C \tag{4.60}$$

achieves $\tilde{E}(R)$ when followed by the SED code. This is because plugging (4.60) into (C.96) gives $\tilde{E}(R)$ (4.59).

Relaxing assumption (c) and dropping assumptions (d)–(e) in Theorem 11, we obtain an achievability bound on $\tilde{E}(R)$.

Proposition 4. *Fix a non-degenerate symmetric binary-input DMC with channel capacity C (4.10) and maximum KL divergence C_1 (4.11), and fix a DSS with random arrivals that emits symbols $S_1, S_2, \ldots$ at a sequence of random symbol arriving times $\tau_1 < \tau_2 < \ldots$ with entropy rate $H > 0$ (4.2). If the symbol arriving times satisfies assumption (c) with the right side of (4.57) relaxed to $\mathbb{E}[\tau_n] + \mathsf{h}(n)$, then, the JSCC reliability function for random streaming is lower bounded as*

$$\tilde{E}(R) \geq C_1 \left(1 - \limsup_{k\to\infty} \left(\frac{H\left(S^k | Y^{\mathbb{E}[\tau_k]+\mathsf{h}(k)}\right)}{kC} + \frac{\mathbb{E}[\tau_k]}{k}\right) R\right), \tag{4.61}$$

where $Y_1, Y_2, \ldots$ are the channel outputs in response to the channel inputs generated by the encoder of the instantaneous SED code for random arrivals.

Proof. Appendix C.21. □

In the setting of Proposition 4, a buffer-then-transmit code that idles the transmissions at times $t = 1, \ldots, \tau_k$ and operates as a code with block encoding at time $t \geq \tau_k + 1$ only achieves (c.f. (4.39))

$$\tilde{E}(R) \geq C_1 \left(1 - \left(\frac{H}{C} + \limsup_{k\to\infty} \frac{\mathbb{E}[\tau_k]}{k}\right) R\right). \tag{4.62}$$

The achievability bound in (4.61) is larger than or equal to the achievability bound in (4.62) since S^k and $Y^{\mathbb{E}[\tau_k]+\mathsf{h}(k)}$ are not independent. This means that in terms of achievable error exponent, the instantaneous SED code for random arrivals performs no worse than the best buffer-then-transmit code.

4.7 Low-complexity codes with instantaneous encoding

We present the type-based algorithms for the instantaneous encoding phase in Section 4.3, for the instantaneous SED code as an anytime code in Section 4.5, for the instantaneous SED code restricted to transmit k symbols only in Section 4.5, and for the instantaneous SED code for random arrivals in Section 4.6. The type-based instantaneous encoding phase is the exact phase in Section 4.3, whereas the type-based instantaneous SED codes (for random arrivals) are approximations of the original codes in Sections 4.5 and 4.6. All the type-based algorithms for deterministic arrivals have a log-linear complexity $O(t \log t)$ in time t. The type-based instantaneous SED code for random arrivals has a polynomial complexity $O(t^2 \log t)$ in time t. The latter complexity is larger due to the decoder's unawareness of the random symbol arriving times.

We assume that the source symbols of the DSS are equiprobably distributed, i.e., the source distribution (4.1) satisfies

$$P_{S_n|S^{n-1}}(a|b) = \frac{1}{q}, \tag{4.63}$$

for all $a \in [q]$, $b \in [q]^{n-1}$, $n = 1, 2, \ldots$

In our type-based codes, the evolving source alphabet is judiciously divided into disjoint sets that we call *types*, so that the source sequences in each type share the same prior and the same posterior. Here, the same prior is guaranteed by the equiprobably distributed symbols (4.63), and the same posterior is guaranteed by moving a whole type to a group during the group partitioning process (see step (iii) below). As a consequence of classifying source sequences into types, the prior update, the group partitioning, and the posterior update can be implemented in terms of types rather than individual source sequences, which results in an exponential reduction of complexity.

We denote by $\mathcal{S}_1, \mathcal{S}_2, \ldots$ a sequence of types. We slightly abuse the notation to denote by $\theta_{\mathcal{S}_j}(Y^{t-1})$ and $\rho_{\mathcal{S}_j}(Y^t)$ the prior and the posterior of a single source sequence in type $\mathcal{S}_j$ at time t rather than the prior and the posterior of the whole type. We fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS that satisfies (4.63) and fix a DMC with a single-letter transition probability $P_{Y|X} \colon \mathcal{X} \to \mathcal{Y}$.

Type-based instantaneous encoding phase

The type-based instantaneous encoding phase operates at times $t = 1, 2, \ldots, t_k$, where k is the number of source symbols of a DSS that we aim to transmit.

(i) *Type update:* At each time t, the algorithm first updates the types. At $t = 1$, the algorithm is initialized with one type $\mathcal{S}_1 \triangleq [q]^{N(1)}$. At $t = t_n$, $n = 2, \ldots, k$, the algorithm updates all the existing types by appending every sequence in $[q]^{N(t)-N(t-1)}$ to every sequence in the type. After the update, the length of the source sequences in each type is equal to $N(t)$; the cardinality of each type is multiplied by $q^{N(t)-N(t-1)}$; the total number of types remains unchanged. At $t \neq t_n$, $n = 1, 2, \ldots, k$, the algorithm does not update the types.

(ii) *Prior update:* Once the types are updated, the algorithm proceeds to update the prior of the source sequences in each existing type. The prior $\theta_{\mathcal{S}_j}(y^{t-1})$, $j = 1, 2, \ldots$ of the source sequences in type $\mathcal{S}_j$ is fully determined by (4.22) with $\theta_i(y^{t-1}) \leftarrow \theta_{\mathcal{S}_j}(y^{t-1})$, $P_{S^{N(t)}|S^{N(t-1)}}(\cdot|\cdot) \leftarrow \left(\frac{1}{q}\right)^{N(t)-N(t-1)}$, and $\rho_{i^{N(t-1)}}(y^{t-1}) \leftarrow \rho_{\mathcal{S}_j}(y^{t-1})$. If the types are not updated, the priors are equal to the posteriors, i.e., $\theta_{\mathcal{S}_j}(y^{t-1}) \leftarrow \rho_{\mathcal{S}_j}(y^{t-1})$, $j = 1, 2, \ldots$

(iii) *Group partitioning:* Using all the existing types and their priors, the algorithm determines a partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$ that satisfies the partitioning rule (4.24) via a *type-based greedy heuristic algorithm*. It operates as follows. It initializes all the groups $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$ by empty sets and initializes the group priors $\{\pi_x(y^{t-1})\}_{x\in\mathcal{X}}$ by zeros. It forms a queue by sorting all the existing types according to priors $\theta_{\mathcal{S}_j}(y^{t-1})$, $j = 1, 2, \ldots$ in a descending order. It moves the types in the queue one by one to one of the groups $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$. Before each move, it first determines a group $\mathcal{G}_{x^*}(y^{t-1})$ whose current prior $\pi_{x^*}(y^{t-1})$ has the largest gap to the corresponding capacity-achieving probability $P_X^*(x^*)$,

$$x^* \triangleq \arg\max_{x\in\mathcal{X}} P_X^*(x) - \pi_x(y^{t-1}). \tag{4.64}$$

Suppose the first type in the sorted queue, i.e., the type whose sequences have the largest prior, is $\mathcal{S}_j$. It then proceeds to determine the number of sequences that are moved from type $\mathcal{S}_j$ to group $\mathcal{G}_{x^*}(y^{t-1})$ by calculating

$$n \triangleq \left\lceil \frac{P_X^*(x^*) - \pi_{x^*}(y^{t-1})}{\theta_{\mathcal{S}_j}(y^{t-1})} \right\rceil. \tag{4.65}$$

If $n \geq |\mathcal{S}_j|$, then it moves the whole type $\mathcal{S}_j$ to group $\mathcal{G}_{x^*}(y^{t-1})$; otherwise, it splits $\mathcal{S}_j$ into two types by keeping the smallest or the largest n consecutive[1] (in lexicographic order) sequences in $\mathcal{S}_j$ and transferring the rest into a new type, and

[1]This step ensures that all sequences in a type are consecutive. Thus, as we will discuss in the last paragraph in Section 4.7, it is sufficient to store two sequences, one with the smallest and one with the largest lexicographic orders, in a type to fully specify that type.

it moves type $\mathcal{S}_j$ to group $\mathcal{G}_{x^*}(y^{t-1})$ and moves the new type to the beginning of the queue. It updates the prior $\pi_{x^*}(y^{t-1})$ after each move.

(iv) *Randomization:* The type-based instantaneous encoding algorithm implements the randomization in (4.25)–(4.30) with respect to a partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$.

(v) *Posterior update:* Upon receiving the channel output $Y_t = y_t$, the algorithm updates the posterior of the source sequences in each existing type. The posterior $\rho_{\mathcal{S}_j}(y^t)$, $j = 1, 2, \ldots$ of the source sequences in type $\mathcal{S}_j$ is fully determined by (4.32) with $\rho_i(y^t) \leftarrow \rho_{\mathcal{S}_j}(y^t)$, $\theta_i(y^{t-1}) \leftarrow \theta_{\mathcal{S}_j}(y^{t-1})$.

Using (4.65) and Appendix C.1, we conclude that the type-based greedy heuristic algorithm ensures (4.24).

We show that the complexity of the type-based instantaneous encoding phase is log-linear $O(t\log t)$ at times $t = 1, 2, \ldots, t_k$. We first show that the number of types grows linearly, i.e., $O(t)$. Since the type update in step (i) does not add new types, the number of types increases only due to the split of types during group partitioning in step (iii). At most $|\mathcal{X}|$ types are split at each time. This is because the ceiling in (4.65) ensures that the group that receives the n sequences from a split type will have a group prior no smaller than the corresponding capacity-achieving probability, thus the group will no longer be the solution to the maximization problem (4.64) and will not cause the split of other types. We proceed to analyze the complexity of each step of the algorithm. Step (i) (type update) has a linear complexity in the number of types, i.e., $O(t)$. This is because the methods of updating and splitting a type in steps (i) and (iii) ensure that the sequences in any type are consecutive, thus it is sufficient to store the starting and the ending sequences in each type to fully specify all the sequences in that type. As a result, updating a type is equivalent to updating the starting and the ending sequences of that type. Step (ii) (prior update) and step (v) (posterior update) have a linear complexity in the number of types, i.e., $O(t)$. Step (iii) (group partitioning) has a log-linear complexity in the number of types due to type sorting, i.e., $O(t\log t)$. This is because the average complexity of sorting a sequence of numbers is log-linear in the size of the sequence [103]. Step (iv) (randomization) has complexity $O(1)$ due to determining $\{p_{\overline{x}\to\underline{x}}\}_{\overline{x}\in\overline{\mathcal{X}}(y^{t-1}),\underline{x}\in\underline{\mathcal{X}}(y^{t-1})}$ in (4.27)–(4.28).

Type-based instantaneous SED codes for deterministic arrivals

We present type-based codes for the anytime instantaneous SED code in Section 4.5 and for the instantaneous SED code restricted to transmit k symbols in Section 4.5,

respectively.

The type-based anytime instantaneous SED code for a symmetric binary-input DMC operates at times $t = 1, 2, \ldots$:

(i′) *Type update:* At each time t, the algorithm updates types as in step (i) with $k = \infty$.

(ii′) *Prior update:* The algorithm updates the prior of the source sequences in each existing type as in step (ii) with $k = \infty$.

(iii′) *Group partitioning:* Using all the existing types and their priors, the algorithm determines a partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\{0,1\}}$ using an *approximating* instantaneous SED rule that mimics the exact rule in (4.43)–(4.44) as follows. It forms a queue by sorting all the existing types according to priors $\theta_{\mathcal{S}_j}(y^{t-1})$, $j = 1, 2, \ldots$ in a descending order. It moves the types in the queue one by one to $\mathcal{G}_0(y^{t-1})$ until $\pi_0(y^{t-1}) \geq P_X^*(0) = 0.5$ for the first time. Suppose the last type moved to $\mathcal{G}_0(y^{t-1})$ is $\mathcal{S}_j$. To make the group priors more even, it then calculates the number of sequences n to be moved away from $\mathcal{S}_j$ as

$$n \triangleq \arg\min_{n\in\{\underline{n},\bar{n}\}} \left|\left(\pi_0(y^{t-1}) - n\theta_{\mathcal{S}_j}(y^{t-1})\right) - \left(\pi_1(y^{t-1}) + n\theta_{\mathcal{S}_j}(y^{t-1})\right)\right|, \tag{4.66a}$$

$$\underline{n} \triangleq \left\lfloor \frac{\pi_0(y^{t-1}) - 0.5}{\theta_{\mathcal{S}_j}(y^{t-1})} \right\rfloor, \tag{4.66b}$$

$$\bar{n} \triangleq \left\lceil \frac{\pi_0(y^{t-1}) - 0.5}{\theta_{\mathcal{S}_j}(y^{t-1})} \right\rceil. \tag{4.66c}$$

It splits $\mathcal{S}_j$ into two types by transferring the first or the last n (4.66a) lexicographically ordered sequences in $\mathcal{S}_j$ to a new type. It moves the new type and all the remaining types in the queue to $\mathcal{G}_1(y^{t-1})$.

(iv′) The randomization step in (iv) is dropped.

(v′) *Posterior update:* The algorithm updates the posteriors of the source sequences in each existing type. The posterior $\rho_{\mathcal{S}_j}(y^t)$, $j = 1, 2, \ldots$, is fully determined by (4.45) with $\rho_i(y^t) \leftarrow \rho_{\mathcal{S}_j}(y^t)$, $\theta_i(y^{t-1}) \leftarrow \theta_{\mathcal{S}_j}(y^{t-1})$.

(vi′) *Decoding at time t:* To decode the first k symbols at time t, where k can be any integer that satisfies $t_k \leq t$, the algorithm first finds the type whose source sequences have the largest posterior. Then, it searches for the most probable length-k prefix in that type by relying on the fact that sequences in the same type share the same posterior; thus, the prefix shared by the maximum number of sequences

is the most probable one. Namely, the algorithm extracts the length-k prefixes of the starting and the ending sequences, denoted by i^k_{start} and i^k_{end}, respectively. If $i^k_{\text{start}} = i^k_{\text{end}}$ (Fig. 4.8-a), then the decoder outputs $\hat{S}^k_t = i^k_{\text{start}}$. If i^k_{start} and i^k_{end} are not lexicographically consecutive (Fig. 4.8-b), then the decoder outputs a length-k prefix in between the two prefixes. If i^k_{start} and i^k_{end} are lexicographically consecutive (Fig. 4.8-c), then the algorithm computes the number of sequences in the type that have prefix i^k_{start} and the number of sequences in the type that have prefix i^k_{end} using the last $N(t) - k$ symbols of the starting and the ending sequences; the decoder outputs the prefix that is shared by more source sequences.

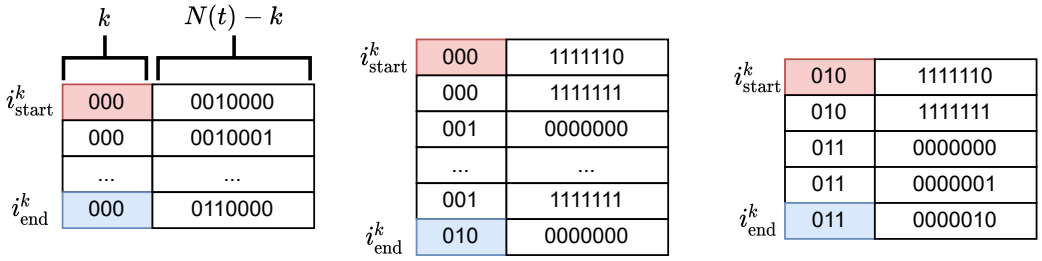


Figure 4.8: Tables (a), (b), (c) represent three types at time t. Each row represents a source sequence in the type. The first row and the last row in each type represent the starting sequence and the ending sequence in that type, respectively. The first column represents the length-k prefix of sequences in the type. The source sequences in a type are lexicographically consecutive due to the methods of updating and splitting a type in steps (i′) and (iii′). In (a), since $i^k_{\text{start}} = i^k_{\text{end}} = 000$, the most probable sequence is 000. In (b), since $i^k_{\text{start}} = 000$ and $i^k_{\text{end}} = 010$ are not lexicographically consecutive, the most probable prefix is 001. In (c), since $i^k_{\text{start}} = 010$ and $i^k_{\text{end}} = 011$ are lexicographically consecutive, the number of sequences with prefix i^k_{start} can be computed by subtracting 1111110, the last $N(t)-k$ symbols of the starting sequence, from 1111111 and adding 1; the number of sequences with prefix i^k_{end} is equal to the last $N(t) - k$ symbols of the ending sequence plus 1. Since (c) contains more sequences with prefix 011, this is the most probable prefix.

We proceed to show that the complexity of the type-based anytime instantaneous SED code is $O(t \log t)$. Similar to the type-based instantaneous encoding phase in Section 4.7, the number of types grows linearly with time t since the number of types increases only if a type is split in step (iii′), and at most 1 type is split at each time t. The complexities of steps (i′), (ii′), (v′) are all linear in the number of types $O(t)$ due to the discussion at the end of Section 4.7. The complexity of step (iii′) is log-linear in the number of types $O(t \log t)$ due to sorting the types. Since the

sequences in a type are lexicographically consecutive due to the updating and the splitting methods in steps (i′) and (iii′), it suffices to use the starting and the ending sequences in a type to determine the most probable prefix in that type. Thus, the complexity of step (vi′) is linear in the number of types due to searching for the type whose sequences have the largest posterior.

Restricting the type-based anytime instantaneous SED code described above to transmit only the first k symbols of a DSS is equivalent to implementing steps (i), (ii), (iii′), (v) one by one, and performing decoding as follows.

(vi″) *Decoding and stopping:* If there exists a type $\mathcal{S}_j$ that satisfies $\rho_{\mathcal{S}_j}(y^t) \geq 1-\epsilon$ and contains a source sequence of length k, then the decoder stops and outputs a sequence in that type as the estimate $\hat{S}^k_{\eta_k}$.

The complexity of the type-based instantaneous SED code for transmitting k symbols remains log-linear, $O(t \log t)$, since the complexity of step (vi″) is $O(t)$ due to searching for the type that satisfies the requirements.

While the type-based instantaneous encoding phase in Section 4.7 is the exact algorithm of the instantaneous encoding phase in Section 4.3, the type-based anytime instantaneous SED code and the type-based instantaneous SED code for transmitting k symbols are *approximations* of the original algorithms in Sections 4.5 and 4.5 due to two reasons below:

First, in step (iii′) (group partitioning), we use the approximating instantaneous SED rule to mimic the exact rule in (4.43)–(4.44). The minimum of the objective function in (4.66a) is equal to the difference $|\pi_0(y^{t-1}) - \pi_1(y^{t-1})|$ between the group priors of the partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\{0,1\}}$ obtained by the approximating rule in step (iii′). The difference is upper bounded as (Appendix C.14)

$$|\pi_0(y^{t-1}) - \pi_1(y^{t-1})| \leq \theta_{\mathcal{S}_j}(y^{t-1}), \tag{4.67}$$

where $\mathcal{S}_j$ is the last type moved to $\mathcal{G}_0(y^{t-1})$ so that its group prior exceeds 0.5 for the first time. If $\pi_0(y^{t-1}) \geq \pi_1(y^{t-1})$, (4.67) recovers (4.44) since $\theta_{\mathcal{S}_j}(y^{t-1})$ is the smallest prior in $\mathcal{G}_0(y^{t-1})$, thus the approximating instantaneous SED rule recovers the exact rule. If $\pi_0(y^{t-1}) < \pi_1(y^{t-1})$, $\theta_{\mathcal{S}_j}(y^{t-1})$ on the right side of (4.67) is the largest prior in $\mathcal{G}_1(y^{t-1})$, violating the right side of (4.44).

We use the approximating algorithm of the instantaneous SED rule (4.43)–(4.44) since it is unclear how to implement the exact instantaneous SED rule with polynomial complexity. In the worst case, the complexity of the latter is as high as double

exponential $O\left(2^{q^{N(t)}}\right)$ due to solving a minimization problem via an exhaustive search [45, Algorithm 1]. An exact algorithm for the SED rule with exponential complexity in the source length is given by [45, Algorithm 2].

Second, in step (vi′) (decoding at time t) of the type-based anytime instantaneous SED code, we only find the most likely length-k prefix in the type that achieves $\max_j \rho_{\mathcal{S}_j}(y^t)$, yet it is possible that this prefix is not the one that has the globally largest posterior (4.46). To search for the most probable length-k prefix, one needs to compute the posteriors for all q^k prefixes of length k using $O(t)$ types, resulting in an exponential complexity $O(q^k t)$ in the length of the prefix k, whereas the complexity of step (vi′) is only $O(t)$ independent of k.

Although the type-based instantaneous SED code is an approximation, as we are about to see in Fig. 4.10 Section 4.8, it is almost as good as the exact code.

Type-based instantaneous SED code for random arrivals

We generalize the type-based instantaneous SED code for deterministic arrivals in Section 4.7 to random arrivals. We assume that a DSS with random arrivals emits a sequence of equiprobable bits $S_n \in \{0, 1\}$, $n = 1, 2, \ldots$ following a Bernoulli-δ process[2], i.e., $\forall b \in \{0, 1\}$, $s \in \mathcal{Q}_t$,

$$P_{S_1}(0) = P_{S_1}(1) = 0.5, \tag{4.68a}$$

$$P_{S^{N(t+1)}|S^{N(t)}}(s \boxplus b|s) = \frac{\delta}{2},\ t \geq 1. \tag{4.68b}$$

We call a binary sequence s_1 the *parent* of a binary sequence s_2 if $s_1 = s_2\boxminus$, and call a type $\mathcal{S}_i$ the *parent* of a type $\mathcal{S}_j$ if all the parents of the strings in $\mathcal{S}_j$ are in $\mathcal{S}_i$. We denote by $p(j)$ the index of the parent of $\mathcal{S}_j$, e.g., $p(j) = i$.

The type-based code for random arrivals differs from the type-based codes for deterministic arrivals in that it creates new types at each time while keeping the parent types. The type-based codes for deterministic arrivals in Section 4.7 need not introduce the concept of parent types since the deterministic symbol arriving times imply that the lengths of all the source sequences at time t are the same.

The type-based instantaneous SED code for random arrivals operates at time $t = 1, 2, \ldots$ as follows.

[2]The symbol arriving probability distribution must satisfy that $P_{S^{N(t+1)}|S^{N(t)}}(s \boxplus 0|s) = P_{S^{N(t+1)}|S^{N(t)}}(s \boxplus 1|s)$, otherwise, the binary sequences in a type will not share the same prior probability. The type-based instantaneous SED code continues to apply if δ in (4.68) is time-varying and/or has memory.

(i‴) *Update types:* At $t = 1$, the encoder and the decoder create two types $\mathcal{S}_1 \triangleq \{0\}$, $\mathcal{S}_2 \triangleq \{1\}$. At $t \geq 2$, each type created at time $t-1$ generates a new type by appending a 0 and a 1 to every binary sequence in that type, and the types that have already been created by time $t-1$ are kept. If the goal is to transmit the first k symbols only, then the algorithm stops creating new types at times $t \geq k+1$; otherwise, the algorithm keeps creating new types at each time t.

(ii‴) *Prior update:* Once the types are updated, the algorithm proceeds to update the prior of the binary sequences in each existing type. The prior $\theta_{\mathcal{S}_i}(y^{t-1})$ (4.55), $i = 1, 2, \ldots$ is fully determined by $\rho_{\mathcal{S}_i}(y^{t-1})$, $\rho_{\mathcal{S}_{p(i)}}(y^{t-1})$ and the symbol arrival probability distribution (4.68).

(iii‴) *Group partitioning:* The algorithm implements the approximating instantaneous SED rule in step (iii′). Note that the types whose parent type is split may have two parents. To ensure that each type has one valid parent, we recursively search for the types whose binary sequences have parents from more than one type and split them accordingly. This guarantees that the posterior $\rho_{p(k)}(t)$, $k = 1, 2, \ldots$, $t = 1, 2, \ldots$ is deterministic.

(iv‴) *Updating posteriors:* The posterior $\rho_{\mathcal{S}_i}(y^t)$ (4.45), $i = 1, 2, \ldots$ is fully determined by $\theta_{\mathcal{S}_i}(y^{t-1})$, the channel transition probability, $Y_t = y_t$, and the group priors $\{\pi_x(y^{t-1})\}_{x\in\{0,1\}}$.

(v‴) The randomization step is dropped.

(vi‴) *Decoding:* The algorithm implements step (vi′) as an anytime code, or it implements step (vi″) for transmitting k symbols.

For a BSC, a heuristic analysis in Appendix C.15 shows that the number of types at time t is $O(t^2)$. Since steps (i‴)(ii‴)(iv‴)(vi‴) have a linear complexity in the number of types, i.e., $O(t^2)$, and step (iii‴) has a log-linear complexity in the number of types, i.e., $O(t^2 \log t)$, the complexity of the type-based instantaneous SED code for random arrivals is $O(t^2 \log t)$.

Similar to the type-based instantaneous SED code for deterministic arrivals, the type-based instantaneous SED code for random arrivals is an approximation of the original code in Section 4.6 due to the use of the approximating instantaneous SED rule in step (iii‴). Yet, Fig. 4.11 below shows that the rate gap between the instantaneous SED code for random arrivals and its corresponding type-based code is negligible.

4.8 Simulations

Fig. 4.9 shows the performance of our new instantaneous encoding schemes. Namely, we fix an error probability $\epsilon = 10^{-6}$, a BSC(0.05), and a DSS that emits i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ bits one by one at consecutive times. We display the rate $R_k \triangleq \frac{k}{\mathbb{E}[\eta_k]}$ as a function of source length k empirically attained by the instantaneous encoding phase followed by the SED code [45, Algorithm 2] and the instantaneous SED code in Section 4.5, and we compare achievable rates to that of the SED code for a fully accessible source, as well as to that of a buffer-then-transmit code that implements the SED code during the block encoding phase. We also plot the rate R_k obtained from the reliability function approximation (4.17):

$$E(R_k) \simeq \frac{R_k}{k} \log \frac{1}{\epsilon}. \tag{4.69}$$

Due to the discussions in the proof sketch of Theorem 9, the instantaneous encoding phase followed either by the MaxEJS code or by the SED code achieves the JSCC reliability function for streaming (4.38). For the simulations in Fig. 4.9, we choose the SED code since it applies to a BSC and its complexity, exponential in the source length, is lower than the double-exponential complexity of the MaxEJS code. To obtain the empirical rate in Fig. 4.9, at each source length k, we run the experiments for every code for 10^5 trials, and we obtain the denominator $\mathbb{E}[\eta_k]$ of the empirical rate by averaging the stopping times in all the experiments.

We observe from Fig. 4.9 that the achievable rate of the instantaneous encoding phase followed by the SED code is significantly larger than that of the buffer-then-transmit code, and approaches that of the SED code as k increases even though the SED encoder knows the entire source sequence before the transmission. The instantaneous SED code demonstrates an even better performance: it is essentially as good as the SED code. The rate obtained from reliability function approximation (4.69) is remarkably close to the empirical achievable rates of our codes with instantaneous encoding even for very short source length $k \simeq 16$. For example, at $k = 16$, the rate obtained from approximation (4.69) is 0.58 (symbols per channel use) and the empirical rate of the instantaneous SED code is 0.59 (symbols per channel use). This means that the reliability function (4.17), an inherently asymptotic notion, accurately reflects the delay-reliability tradeoffs attained by the JSCC reliability function-achieving codes in the ultra-short blocklength regime. The achievable rate corresponding to the buffer-then-transmit code is limited by (4.39).

Fig. 4.10 shows the performance of the type-based instantaneous SED code. We fix an error probability $\epsilon = 10^{-6}$ (4.15), a BSC(p) with $p = 0.05, 0.03, 0.01$, and a

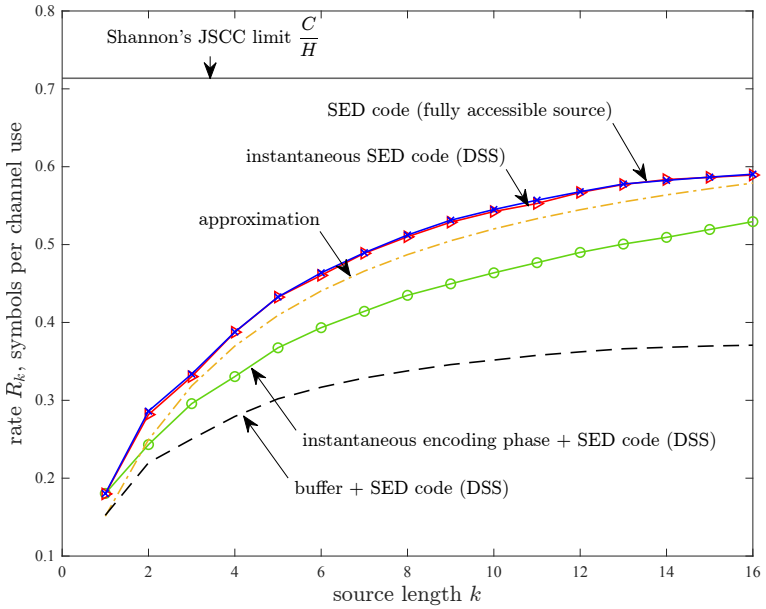


Figure 4.9: Rate R_k (symbols per channel use) vs. source length k. The error probability is constrained by $\epsilon = 10^{-6}$ (4.15). The DMC is a BSC(0.05). Naghshvar et al.'s SED code [45, Algorithm 2] operates on a fully accessible block S^k of independent Bernoulli$\left(\frac{1}{2}\right)$ bits. The instantaneous encoding phase followed by the SED code, the instantaneous SED code, and the buffer-then-transmit code operate on k i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ source bits emitted one by one at times $t = 1, 2, \ldots, k$. The curves are displayed for the range of k's where the complexities of the SED code and the instantaneous SED code are not prohibitive.

DSS that emits i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ bits one by one at consecutive times. We plot rate $R_k = \frac{k}{\mathbb{E}[\eta_k]}$ as a function of source length k empirically achieved by the instantaneous SED code in Section 4.5 and its corresponding type-based code in Section 4.7, as well as the rate obtained from the reliability function approximation (4.69). At each source length k, we run the experiments using the same method as in Fig. 4.9. The rate gap between the instantaneous SED code and the type-based instantaneous SED code is negligible, meaning that the type-based instantaneous SED code with only log-linear complexity is a good approximation to the exact code in Section 4.5. Furthermore, it is interesting to see that even though the DSS has symbol arriving rate $f = 1$ symbol per channel use, which is far less than that required in assumption (b'), the achievable rates of the instantaneous SED code stay very close to the rates obtained from the reliability function approximation. This suggests that assumption (b') on the symbol arriving rate, sufficient for the instantaneous SED code to achieve $E(R)$, could be conservative.

Fig. 4.11 shows the performance of the type-based instantaneous SED code for

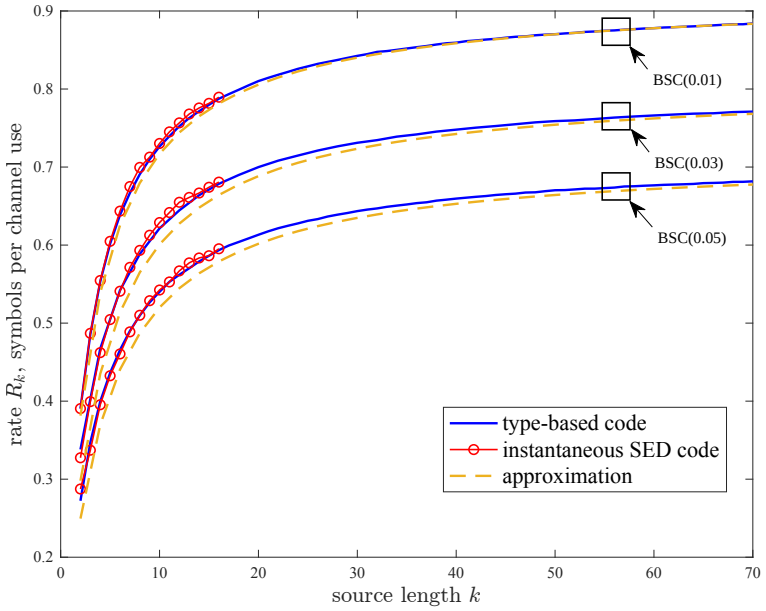


Figure 4.10: Rate R_k (symbols per channel use) vs. source length k. The error probability is constrained by $\epsilon = 10^{-6}$ (4.15). The type-based instantaneous SED code in Section 4.7 and the instantaneous SED code in Section 4.5 operate on k i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ source bits emitted one by one at times $t = 1, 2, \ldots, k$.

random arrivals. We fix an error probability $\epsilon = 10^{-3}$, a BSC(0.02), a DSS that emits i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ bits one by one at consecutive times, and a DSS with random arrivals that emits i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ bits following (4.68) with $\delta = 0.98$, i.e., it emits a new bit with probability 0.98 at each time t. The achievable rates of the instantaneous SED code for random arrivals and the achievable rates of its corresponding type-based code have a negligible gap, meaning that the type-based code for random arrivals with only a polynomial complexity is a good approximation of the original code. The achievable rates for the DSS with random arrivals are smaller than those for the DSS with deterministic and consecutive symbol arriving times. This suggests that the instantaneous SED code for random arrivals might not incorporate *enough* information in the random arriving times into its channel inputs, and that the JSCC reliability function for random arrivals could depend on the distribution for the random arriving times. Nevertheless, the instantaneous SED codes for both deterministic and random arrivals significantly outperform the buffer-then-transmit codes.

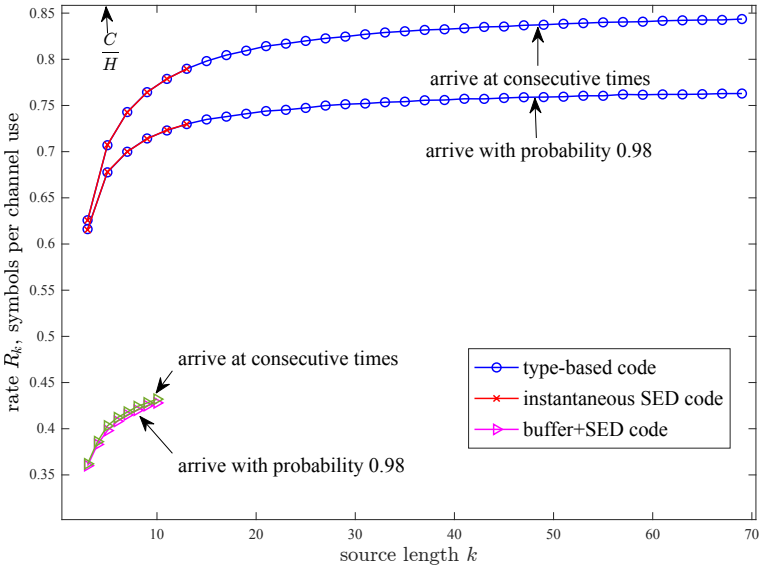


Figure 4.11: Rate R_k (symbols per channel use) vs. source length k. The error probability is constrained by $\epsilon = 10^{-3}$ (4.15). The type-based instantaneous SED code in Section 4.7 and the instantaneous SED code in Section 4.5 operate on k i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ source bits emitted one by one at times $t = 1, 2, \ldots, k$. The type-base instantaneous SED code for random arrivals in Section 4.7 and the instantaneous SED code for random arrivals in Section 4.6 operate on k i.i.d. Bernoulli$\left(\frac{1}{2}\right)$ source bits emitted according to (4.68) with $\delta = 0.98$. A buffer-then-transmit code starts to implement the SED code [45] right after the arrival of the k-th bit.

4.9 Streaming over a degenerate DMC with zero error

In this section, we propose a code with instantaneous encoding for a degenerate DMC (4.9) that achieves zero decoding error at any rate asymptotically below $\frac{C}{H}$. Here, our code does not exactly follow Definition 20 since it generalizes the code with instantaneous encoding in Definition 20 by allowing *common randomness* $U \in \mathcal{U}$, which is a random variable that is revealed to the encoder and the decoder before the transmission. With common randomness U, the encoder f_t (4.12) can use U to form X_t, and the decoder g_t (4.13) can use U to decide the stopping time η_k and the estimate $\hat{S}^k_{\eta_k}$. We refer to such a code as a $\langle k, R, \epsilon \rangle$ *code with instantaneous encoding and common randomness* if it achieves rate R (4.14) and error probability ϵ (4.15) for transmitting k symbols of a DSS. Common randomness is widely used to specify a random codebook in the scenario where multiple constraints on expectations of quantities that depend on the codebook must be satisfied simultaneously and where Shannon's probabilistic method is not sufficient to claim the existence of a deterministic codebook satisfying all constraints, e.g., [43][41][87][104]. Since for a fixed k, we seek to satisfy two constraints, on the rate and on the error probability,

the cardinality of $\mathcal{U}$ can be restricted as $|\mathcal{U}| \leq 2$ (Appendix C.22).

Theorem 12, stated next, establishes the existence of zero-error codes for the transmission over a degenerate DMC at any rate asymptotically below $\frac{C}{H}$.

Theorem 12. *Fix a degenerate DMC with capacity C (4.10), fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS with entropy entropy rate $H > 0$ (4.2) satisfying assumptions* (a)–(b) *in Theorem 9, and fix any $R < \frac{C}{H}$. There exists a sequence of $\langle k, R_k, 0\rangle$ codes with instantaneous encoding and common randomness that satisfies*

$$\lim_{k\to\infty} R_k = R. \tag{4.70}$$

Proof sketch. Our zero-error code for degenerate DMCs extends Burnashev's scheme [40, Sec. 6] to JSCC and to streaming sources: to achieve Shannon's JSCC limit $\frac{C}{H}$, a Shannon limit-achieving code is used in the first communication phase to compress the source; to transmit streaming sources, we combine an instantaneous encoding phase that satisfies (4.41) with a Shannon limit-achieving block encoding scheme to form a Shannon limit-achieving instantaneous encoding scheme. To achieve zero error, we employ confirmation phases similar to those in Burnashev's scheme [40]. We say that a $\langle k, R, \epsilon_k\rangle$ code with instantaneous encoding and common randomness achieves Shannon's JSCC limit $\frac{C}{H}$ if for all $R < \frac{C}{H}$, a sequence of such codes indexed by k satisfies $\epsilon_k \to 0$ as $k \to \infty$. Our zero-error code includes such Shannon limit-achieving codes as a building block. Note that in contrast to the discussions in Sections 4.4–4.5 focused on the exponential rate of decay of ϵ_k to 0 (4.17) over non-degenerate DMCs, here merely having ϵ_k decrease to 0 suffices. The following argument shows the existence of such codes for the class of channels that includes both non-degenerate and degenerate DMCs.

We employ the joint source-channel code in [104, Theorem 2] due to the simplicity of the error analysis it affords. The code in [104, Theorem 2] is a $\langle k, R, \epsilon_k\rangle$ Shannon limit-achieving code with block encoding and common randomness because its expected decoding time to attain error probability ϵ is upper bounded as (C.30) in Appendix C.7 with $C_1 \leftarrow C$ [104, Eq. (16)], implying that it achieves a positive error exponent that is equal to (4.38) with $C_1 \leftarrow C$ for all $R < \frac{C}{H}$. The block encoding scheme in [104, Theorem 2] is a stop-feedback code, meaning that the encoder uses channel feedback only to decide whether to stop the transmission but not to form channel inputs. If the DSS has an infinite symbol arriving rate $f = \infty$ (4.5), a buffer-then-transmit code using the block encoding scheme in [104, Theorem 2] achieves the Shannon limit since it achieves the same error exponent as the code in

[104, Theorem 2]. To see this, one can simply invoke (C.30) in Lemma 20 with $C_1 \leftarrow C$ and follow the proofs in Appendix C.8. By the same token, if the DSS has a finite symbol arriving rate $f < \infty$ (4.5), a code implementing an instantaneous encoding phase that satisfies (4.41) followed by the block encoding scheme in [104, Theorem 2] for k source symbols with prior $P_{S^k|Y^{t_k}}$ achieves the Shannon limit with the same error exponent as the code in [104, Theorem 2].

Our zero-error code with instantaneous encoding and common randomness for transmitting k symbols over a degenerate DMC operates as follows (details in Appendix C.23). Similar to [40]–[44], [87], our code is divided into blocks. Each block contains a communication phase and a confirmation phase. In the first block, the communication phase uses a $\langle k, R, \epsilon_k \rangle$ Shannon limit-achieving code with instantaneous encoding and common randomness. The confirmation phase selects two symbols x (4.9a) and x' (4.9b) as the channel inputs (i.e., x' never leads to channel output y); the encoder repeatedly transmits x if the decoder's estimate of the source sequence at the end of the communication phase is correct, and transmits x' otherwise. If the decoder receives a y in the confirmation phase, meaning that the encoder communicated its knowledge that the decoder's estimate is correct with zero error, then it outputs its estimate, otherwise, the next block is transmitted. The ℓ-th block, $\ell \geq 2$, differs from the first block in that it does not compress the source to avoid errors due to an atypical source realization and in that it uses random coding whereas the first block can employ any Shannon-limit achieving code.

We proceed to discuss the error and the rate achievable by our code (details in Appendix C.23).

Our code achieves zero error by employing confirmation phases that rely on the degenerate nature of the channel: receiving a y in the confirmation phase guarantees a correct estimate.

Our code achieves all rates asymptotically below $\frac{C}{H}$ because 1) the first block employs a Shannon limit-achieving code in the communication phase, 2) the length of the confirmation phase is made negligible compared to the length of the communication phase as the source length $k \to \infty$, meaning that the length of the first block asymptotically equals the length of its communication phase, and 3) subsequent blocks asymptotically do not incur a penalty on rate, as we discuss next. Since the length of each block is comparable to the length of the first block, it is enough to show that the expected number of blocks T_k transmitted after the first block converges to zero. The refreshing of random codebook for all uncompressed source

sequences in every block after the first block ensures that the channel output vectors in these subsequent blocks are i.i.d. and are independent of the channel outputs in the first block. Conditioned on $T_k > 0$, the i.i.d. vectors give rise to a geometric distribution of T_k with failure probability converging to 0, which implies $\mathbb{E}[T_k] \to 0$ as $k \to \infty$. □

A stop-feedback code with block encoding that retransmits blocks with the overall rate asymptotically equal to the rate of the first block is also used by Forney [105, p. 213] for deriving a lower bound on the reliability function of a DMC.

4.10 Conclusion

We have derived the reliability function for transmitting a discrete streaming source over a DMC with feedback using variable-length joint source-channel coding with instantaneous encoding under regularity conditions (Theorem 9). Since a classical fully accessible DS is a special DSS (see (4.3)), Theorem 9 extends Burnashev's reliability function to the classical JSCC scenario with block encoding, as well as to a streaming scenario. The most surprising observation is that the JSCC reliability function for a streaming source is equal to that for a fully accessible source. A naive buffer-then-transmit code that idles the transmission during the symbol arriving period does not achieve the JSCC reliability function for a non-trivial streaming source (see (4.39)). To achieve the JSCC reliability function for such sources, we have proposed a novel instantaneous encoding phase (Section 4.3). We have shown that preceding a JSCC reliability function-achieving code with block encoding, e.g., the MaxEJS code or the SED code [45], by our instantaneous encoding phase (Section 4.3) will make it overcome the detrimental effect due to the streaming nature of the source and make it achieve the same error exponent as if the encoder knew the entire source sequence before the transmission. The instantaneous encoding phase (Section 4.3) achieves the JSCC reliability function because it satisfies the sufficient condition (4.41) on the statistics of the encoder outputs during the symbol arriving period, for example, the instantaneous encoding phase continues to achieve the sufficient condition (4.41) after it drops the randomization step, but at a cost of increasing the threshold for the symbol arriving rate (Remark 2). While our JSCC reliability function-achieving codes are designed to transmit k symbols of a streaming source and stop, we have also designed an instantaneous SED code (Section 4.5) that can choose the decoding time and the number of symbols to decode on the fly. It empirically attains a positive anytime reliability (Fig. 4.6),

thus it can be used to stabilize an unstable scalar linear system with a bounded noise over a noisy channel. A sequence of such codes indexed by the source length to decode also achieves the JSCC reliability function for streaming in the limit of large source length (Theorem 12). Furthermore, the instantaneous SED code can also be used to transmit source symbols with random symbol arriving times that are unknown to the decoder. For a DSS whose symbol arriving times have limited randomness, we derive the JSCC reliability function for random streaming (Theorem 11) using the instantaneous SED code. It is equal to the JSCC reliability function for deterministic streaming, meaning that the instantaneous SED code performs as if the decoder knew the random times. For practical implementations, we have designed type-based algorithms for the instantaneous encoding phase and the instantaneous SED code (Section 4.7) with a log-linear complexity, and we have designed a type-based algorithm for the instantaneous SED code for random arrivals (Section 4.6) with a polynomial complexity. While the codes that achieve the JSCC reliability function are designed for non-degenerate DMCs, we have also designed zero-error codes with instantaneous encoding for degenerate DMCs (Section 4.9), extending Burnashev's zero-error channel code to the JSCC and to the streaming scenarios.

4.11 Future research directions

Based on the findings in Sections 4.2–4.9, we list several interesting directions for future research.

JSCC reliability function for a wider class of channels

It would be interesting to find the JSCC reliability function for a wider class of channels.

Converse: As we have discussed in the converse proof of Theorem 9 (Appendix C.3), converse bounds on the JSCC reliability function for a fully accessible source continue to hold for the JSCC reliability function for a streaming source. This observation simplifies the converse proof. If the reliability function does not exist (e.g., AWGN channels), one can still use the achievable error probability of a block encoding scheme as the baseline (e.g., Schalkwijk-Kailath (S-K)'s scheme [38] for AWGN channels) and compare the error probability of a proposed code with instantaneous encoding scheme with the baseline.

Achievability: One can design a code with instantaneous encoding for an AWGN channel by transmitting the process innovation scaled to satisfy a power constraint

at each time. Alternatively, discretizing the channel input of an AWGN channel, one can construct a code with instantaneous encoding by implementing the instantaneous encoding phase during the symbol arriving period $1, 2, \ldots, t_k$ and a modified S-K's [38] block encoding scheme for transmitting k source symbols with prior $P_{S^k|Y^{t_k}}$ after the symbol arriving period. The modification is to establish a pulse-amplitude modulation with non-uniform gaps in the first step according to source prior $P_{S^k|Y^{t_k}}$, so that the amplitudes achieve the minimum error while satisfying a power constraint.

JSCC reliability function for a wider class of streaming sources

It would be interesting to find the JSCC reliability function for a wider class of streaming sources.

1) One can try to find the JSCC reliability function for streaming symbols whose symbol alphabet size $q = \infty$. Our current analysis is not compatible with such streaming symbols since we make use of the fact $\frac{\log q}{H} < \infty$ both in (C.12) and in showing that the right side of (C.33) is equal to (C.30). For symbol alphabet size $q = \infty$, the group partitioning complexity (4.24) of using the greedy heuristic algorithm (Appendix C.1) becomes infinite. It might be helpful to map every source sequence in the source alphabet to a real number, and leverage a partitioning rule similar to Horstein's scheme [37] to partition an interval on the real line into $|\mathcal{X}|$ disjoint sub-intervals.

2) One can try to relax assumption (b) on the symbol arriving rate of the streaming source. While (4.40) gives a converse bound on the symbol arriving rate and (4.37) guarantees achievability of $E(R)$ (4.38), the existence of a critical symbol arriving rate f_{cr} such that for all $f > f_{\text{cr}}$, $E(R)$ (4.38) is achievable, and for all $f < f_{\text{cr}}$, $E(R)$ (4.38) is not achievable, remains open. While $E(R)$ (4.38) is not a function of f, it is conceivable that for $f < f_{\text{cr}}$, the reliability function (4.17) will depend on f. This is reminiscent of the channel reliability function for transmitting over a DMC without feedback via a fixed-length block code, which is known only for rates greater than a critical value where its converse bound (sphere-packing exponent [106]) coincides with its achievability bound (random-coding exponent [107]).

Our current method prevents us to relax the threshold in assumption (b) to $\frac{C}{H}$. The key to close the gap is to show that the logarithm of the numerator in (C.52a) can be upper bounded by $H(Y|X)$ in probability. Yet, in the current proof, we upper bound it by $\log \frac{1}{p_{\max}}$ to show that the source prior $\theta_{S^{N(t)}}(Y^{t-1})$ converges to zero in probability during the symbol arriving period. There is hope to close the gap since if

we think inversely, i.e., given that the source prior converges to zero in probability, the partitioning rule (4.24) drives the randomization probability $P_{X_t|Z_t,Y^{t-1}}(x|x,Y^{t-1})$ to 1 for any $x \in \mathcal{X}$, and the logarithm of the numerator converges in probability to $H(Y|X)$. The difficulty of showing a sharper bound lies in the *intermixed* relation between the randomization distribution and the source prior: the convergence of the randomization distribution $P_{X_t|Z_t,Y^{t-1}}$ relies on the convergence of the source prior while it in turn dominates whether the source prior converges (see (C.52a)). One needs other tools to analyze the instantaneous encoding phase or needs a judicious modification to the instantaneous encoding phase.

Low-complexity code for non-equiprobable symbols

While our type-based codes (Section 4.7) are designed for equiprobable symbols, it is practically important to design low-complexity instantaneous encoding schemes for a wider class of streaming sources. Type-based codes in Section 4.7 are suboptimal for non-equiprobable symbols because the source sequences in a type will no longer have the same prior after the code appends every symbol in $[q]$ to every source sequence in a type. As a suboptimal method, for symbols with a non-uniform distribution on the alphabet $[q]$, the encoder and the decoder can ignore the true symbol distribution and assume that the symbols are equiprobable.

Analytical proof for an anytime code

It would be interesting to prove that the instantaneous SED code is an anytime code analytically. It is difficult to extend our analysis for $E(R)$ (4.38) to show that the instantaneous SED code satisfies (4.18). The submartingales in [40][45] used to compute the upper bound on the expected decoding time for a block encoding scheme to attain a target error probability fail to hold if the encoder keeps incorporating newly arrived symbols after time t_k. Therefore, we cannot directly use Lemma 20 in Appendix C.7 to upper bound the expected decoding time, and different tools are needed to analyze the anytime reliability.

Reliability function for lossy JSCC of streaming sources

While we focus on almost lossless coding, it would be interesting to derive the JSCC reliability function for transmitting a streaming source using a variable-length lossy joint source-channel code over a DMC with feedback. Truong and Tan [87] showed such a reliability function (a.k.a. excess-distortion exponent) for a fully accessible and memoryless discrete source at an average rate of 1 symbol per channel use.

The reliability function can be easily extended to R symbols per channel use by modifying the length of each confirmation phase in Truong and Tan's code [87] to $\frac{k}{R} - \frac{kR(D)}{C}$:

$$E_{\text{excess}}(D, R) = C_1 \left(1 - \frac{R(D)}{C} R\right), \tag{4.71}$$

where D is the distortion level, $R(D)$ is the distortion-rate function, and R is the rate.

To extend (4.71) to our streaming scenario, we first notice that (4.71) serves as a converse bound. As for achievability, one can try to design an instantaneous encoding phase that achieves the converse bound (4.71) when followed by a reliability function-achieving lossy block encoding scheme. Inspired by the underlying principle (4.41) and the achievability proof in Appendix C.9, we conjecture that an appropriate instantaneous encoding phase should judiciously shape the joint probability distribution of the source symbols S^k and the channel outputs Y^{t_k} during the symbol arriving period, so that

1) after the symbol arriving period, $R(D)$ for the source distribution $P_{S^k|Y^{t_k}}$ decreases compared to that for the initial source distribution P_{S^k};

2) the decreased amount exactly compensates the detrimental effect due to symbol arriving time t_k as $k \to \infty$.

Streaming with limited feedback

While we assume that the encoder receives full feedback at each time in Sections 4.2–4.7, it is practically important to design good codes with instantaneous encoding over a DMC with limited feedback. *Example 1:* One can consider that the feedback link is used every d times, i.e., the encoder only knows the channel output vector $\{Y_t\}_{t=md+1}^{(m+1)d}$ at time $(m+1)d+1$, $m = 0, 1, 2, \ldots$ We conjecture that for small d, a *good* code with instantaneous encoding acts similarly to a code with full feedback at each time, except that the encoder tries to 1) align its belief with decoder's belief by guessing the channel output via the channel transition probability when the feedback is unavailable, and 2) re-synchronize its belief with decoder's belief when the feedback is available. We conjecture that for large d, a *good* code with instantaneous encoding acts similarly to block encoding schemes.

Example 2: One can consider a scenario where the decoder is only allowed to feedback a symbol Y'_t from alphabet $\mathcal{Y}'$ at each time, and the size of $\mathcal{Y}'$ is smaller

than the size of the channel output's alphabet $|\mathcal{Y}'| < |\mathcal{Y}|$. If $|\mathcal{Y}'| = 2$, the feedback is often known as the ACK/NACK feedback. Allowing the encoder to know the probability distribution $P_{Y_t'|Y_t}$ of transmitting Y_t' given the true channel output Y_t, one can optimize the probability distribution $P_{Y_t'|Y_t}$ to minimize the error probability.

Example 3: One can also consider a scenario where the decoder can choose the feedback times online, so that the error probability can be minimized with the minimum number of feedback times. We conjecture that the decoder chooses to feedback when it becomes uncertain about the source, in other words, a feedback occurs if the entropy of the source posterior is larger than a threshold.

Streaming with bounded memory

It would be interesting to see whether a code with bounded memory can attain the JSCC reliability function for streaming. To implement our type-based codes, both the encoder and the decoder need to keep track of all the types by storing the starting and the ending sequences in each type. Thus, the required memory size grows linearly with time. The linear growth in memory size is practical for transmitting a finite and fixed number of streaming symbols (Sections 4.3, 4.4, 4.5) as the memory size remains bounded. Yet, it leads to an unbounded memory size if the number of streaming symbols to decode is infinite. Since memory is a limited resource, it is practically important to design an instantaneous encoding scheme with bounded memory for transmitting an infinite number of streaming symbols. To do so, one can constantly discard types with very low posteriors and only keep the most likely types. Alternatively, one can use the sliding window technique developed for transmitting streaming data without feedback [53, 57, 90]: At each time t, the encoder only transmits symbols within a window of a fixed length, and the window keeps dropping *old* symbols and incorporating *fresh* symbols with time.

Chapter 5

CONCLUSION

This thesis exploits an encoding method known as causal encoding, critical for transmitting streaming data in real-time communication scenarios such as remote tracking and distributed control. While classical non-causal (block) encoding schemes introduce communication delay due to buffering the streaming data into a block before the transmission, causal encoding transmits information based on the causally received data while the data is still streaming in, circumventing the undesirable delay. Causal encoding is investigated in three operational scenarios: causal frequency-constrained sampling, causal rate-constrained sampling (compression), and causal joint source-channel coding with feedback. In these operational scenarios of causal encoding, we derive the fundamental limits, namely, the distortion-frequency function, the distortion-rate function, and the JSCC reliability function for streaming. We design causal encoding schemes that achieve the limits, apply to control systems, adapt to system deficiencies such as delay and noise, and have low complexities. Our causal encoding schemes demonstrate surprisingly good performance. For example, in causal compressing of the Wiener process, we show that the distortion achieved by our SOI code is even smaller than the distortion achieved by the best non-causal code due to the leverage of free timing information. Timing information, rarely used in the design of classical non-causal encoding schemes, is commonly available in real-time systems. It opens the door to designing causal codes that outperform the best non-causal codes. Even not using the timing information, in causal JSCC with feedback, we show that the JSCC reliability function for a streaming source is equal to the JSCC reliability function for a fully accessible source. Our findings suggest that causal encoding can save communication time without sacrificing communication fidelity. It is conceivable that causal encoding will become one promising method to satisfy the ever-growing demand for ultra-reliable low latency communications in the near future.

BIBLIOGRAPHY

[1] O. C. Imer and T. Basar. “Optimal estimation with limited measurements”. In: *Proceedings of the 44th IEEE Conference on Decision and Control*. Dec. 2005, pp. 1029–1034.

[2] G. M. Lipsa and N. C. Martins. “Remote state estimation with communication costs for first-order LTI systems”. In: *IEEE Transactions on Automatic Control* 56.9 (Apr. 2011), pp. 2013–2025.

[3] J. Wu et al. “Event-based sensor data scheduling: trade-off between communication rate and estimation quality”. In: *IEEE Transactions on Automatic Control* 58.4 (Aug. 2013), pp. 1041–1046.

[4] J. Chakravorty and A. Mahajan. “Fundamental limits of remote estimation of autoregressive Markov processes under communication constraints”. In: *IEEE Transactions on Automatic Control* 62.3 (June 2017), pp. 1109–1124.

[5] A. Molin and S. Hirche. “Event-triggered state estimation: an iterative algorithm and optimality properties”. In: *IEEE Transactions on Automatic Control* 62.11 (May 2017), pp. 5939–5946.

[6] M. Rabi, G. V. Moustakides, and J. S. Baras. “Adaptive sampling for linear state estimation”. In: *SIAM Journal on Control and Optimization* 50.2 (Mar. 2012), pp. 672–702.

[7] K. Nar and T. Başar. “Sampling multidimensional Wiener processes”. In: *53rd IEEE Conference on Decision and Control*. Dec. 2014, pp. 3426–3431.

[8] Y. Sun, Y. Polyanskiy, and E. Uysal. “Sampling of the Wiener process for remote estimation over a channel with random delay”. In: *IEEE Transactions on Information Theory* 66.2 (Aug. 2020), pp. 1118–1135.

[9] T. Z. Ornee and Y. Sun. “Sampling and remote estimation for the Ornstein-Uhlenbeck process through queues: Age of information and beyond”. In: *IEEE/ACM Transactions on Networking* 29.5 (May 2021), pp. 1962–1975.

[10] K. Åström and B. Bernhardsson. “Comparison of Riemann and Lebesgue sampling for first order stochastic systems”. In: *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.* Vol. 2. Dec. 2002, pp. 2011–2016.

[11] K. J. Åström. *Introduction to stochastic control theory*. New York: Academic Press, 1970.

[12] C. E. Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (July 1948), pp. 379–423.

[13] C. E. Shannon. “Coding theorems for a discrete source with a fidelity criterion”. In: *Institute of Radio Engineers, International Convention Record* 4 (Mar. 1959), pp. 142–163.

[14] A. Gorbunov and M. S. Pinsker. “Nonanticipatory and prognostic epsilon entropies and message generation rates”. In: *Problemy Peredachi Informatsii* 9.3 (1973), pp. 12–21.

[15] J. Massey. “Causality, feedback and directed information”. In: *Proceedings International Symposium on Information Theory and its Applications*. Nov. 1990, pp. 303–305.

[16] A. Gorbunov and M. S. Pinsker. “Prognostic epsilon entropy of a Gaussian message and a Gaussian source”. In: *Problemy Peredachi Informatsii* 10.2 (1974), pp. 5–25.

[17] S. Tatikonda, A. Sahai, and S. Mitter. “Stochastic linear control over a communication channel”. In: *IEEE Transactions on Automatic Control* 49.9 (Sept. 2004), pp. 1549–1561.

[18] M. S. Derpich and J. Østergaard. “Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources”. In: *IEEE Transactions on Information Theory* 58.5 (Jan. 2012), pp. 3131–3152.

[19] C. D. Charalambous, P. A. Stavrou, and N. U. Ahmed. “Nonanticipative rate distortion function and relations to filtering theory”. In: *IEEE Transactions on Automatic Control* 59.4 (Nov. 2013), pp. 937–952.

[20] T. Tanaka et al. “Semidefinite programming approach to Gaussian sequential rate-distortion trade-offs”. In: *IEEE Transactions on Automatic Control* 62.4 (Aug. 2016), pp. 1896–1910.

[21] P. A. Stavrou, T. Charalambous, and C. D. Charalambous. “Filtering with fidelity for time-varying Gauss-Markov processes”. In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE. Dec. 2016, pp. 5465–5470.

[22] P. A. Stavrou, T. Charalambous, and C. D. Charalambous. “Finite-time nonanticipative rate distortion function for time-varying scalar-valued Gauss-Markov sources”. In: *IEEE Control Systems Letters* 2.1 (Nov. 2017), pp. 175–180.

[23] P. A. Stavrou et al. “Optimal estimation via nonanticipative rate distortion function and applications to time-varying Gauss–Markov processes”. In: *SIAM Journal on Control and Optimization* 56.5 (Oct. 2018), pp. 3731–3765.

[24] P. A. Stavrou et al. “An upper bound to zero-delay rate distortion via Kalman filtering for vector Gaussian sources”. In: *2017 IEEE Information Theory Workshop (ITW)*. Nov. 2017, pp. 534–538.

[25] P. A. Stavrou, T. Tanaka, and S. Tatikonda. “The time-invariant multidimensional Gaussian sequential rate-distortion problem revisited”. In: *IEEE Transactions on Automatic Control* 65.5 (Sept. 2019), pp. 2245–2249.

[26] M. J. Khojasteh et al. “The value of timing information in event-triggered control”. In: *IEEE Transactions on Automatic Control* 65.3 (May 2019), pp. 925–940.

[27] E. Kofman and J. H. Braslavsky. “Level crossing sampling in feedback stabilization under data-rate constraints”. In: *Proceedings of the 45th IEEE Conference on Decision and Control*. IEEE. Dec. 2006, pp. 4423–4428.

[28] A. Khina et al. “Algorithms for optimal control with fixed-rate feedback”. In: *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE. Dec. 2017, pp. 6015–6020.

[29] S. Yoshikawa, K. Kobayashi, and Y. Yamashita. “Quantized event-triggered control of discrete-time linear systems with switching triggering conditions”. In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 101.2 (Feb. 2018), pp. 322–327.

[30] J. Pearson, J. P. Hespanha, and D. Liberzon. “Control with minimal cost-per-symbol encoding and quasi-optimality of event-based encoders”. In: *IEEE Transactions on Automatic Control* 62.5 (Aug. 2016), pp. 2286–2301.

[31] D. Lehmann and J. Lunze. “Event-based control using quantized state information”. In: *IFAC Proceedings Volumes* 43.19 (Jan. 2010), pp. 1–6.

[32] Q. Ling. “Periodic event-triggered quantization policy design for a scalar LTI system with iid feedback dropouts”. In: *IEEE Transactions on Automatic Control* 64.1 (Apr. 2018), pp. 343–350.

[33] P. Tallapragada and J. Cortés. “Event-triggered stabilization of linear systems under bounded bit rates”. In: *IEEE Transactions on Automatic Control* 61.6 (Sept. 2015), pp. 1575–1589.

[34] A. Tanwani, C. Prieur, and M. Fiacchini. “Observer-based feedback stabilization of linear systems with event-triggered sampling and dynamic quantization”. In: *Systems & Control Letters* 94 (Aug. 2016), pp. 46–56.

[35] M. Abdelrahim, V. Dolk, and W. Heemels. “Input-to-state stabilizing event-triggered control for linear systems with output quantization”. In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE. Dec. 2016, pp. 483–488.

[36] C. E. Shannon. “The zero error capacity of a noisy channel”. In: *IRE Transactions on Information Theory* 2.3 (Sept. 1956), pp. 8–19.

[37] M. Horstein. “Sequential transmission using noiseless feedback”. In: *IEEE Transactions on Information Theory* 9.3 (July 1963), pp. 136–143.

[38] J. Schalkwijk and T. Kailath. “A coding scheme for additive noise channels with feedback–I: No bandwidth constraint”. In: *IEEE Transactions on Information Theory* 12.2 (Apr. 1966), pp. 172–182.

[39] O. Shayevitz and M. Feder. “Optimal feedback communication via posterior matching”. In: *IEEE Transactions on Information Theory* 57.3 (Feb. 2011), pp. 1186–1222.

[40] M. V. Burnashev. “Data transmission over a discrete channel with feedback. Random transmission time”. In: *Problemy peredachi informatsii* 12.4 (1976), pp. 10–30.

[41] Y. Polyanskiy, H. V. Poor, and S. Verdú. “Feedback in the non-asymptotic regime”. In: *IEEE Transactions on Information Theory* 57.8 (July 2011), pp. 4903–4925.

[42] M. V. Burnashev and K. Zigangirov. “An interval estimation problem for controlled observations”. In: *Problemy Peredachi Informatsii* 10.3 (1974), pp. 51–61.

[43] H. Yamamoto and K. Itoh. “Asymptotic performance of a modified schalkwijk-barron scheme for channels with noiseless feedback (corresp.)” In: *IEEE Transactions on Information Theory* 25.6 (Nov. 1979), pp. 729–733.

[44] G. Caire, S. Shamai, and S. Verdú. “Propagation, feedback and belief”. In: *4th International Symposium on Turbo Codes & Related Topics; 6th International ITG-Conference on Source and Channel Coding*. VDE. Apr. 2006, pp. 1–6.

[45] M. Naghshvar, T. Javidi, and M. Wigger. “Extrinsic Jensen–Shannon divergence: Applications to variable-length coding”. In: *IEEE Transactions on Information Theory* 61.4 (Feb. 2015), pp. 2148–2164.

[46] H. Yang et al. “Sequential transmission over binary asymmetric channels with feedback”. In: *IEEE Transactions on Information Theory (Early Access)* (June 2022).

[47] A. Antonini, H. Yang, and R. D. Wesel. “Low complexity algorithms for transmission of short blocks over the BSC with full feedback”. In: *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE. June 2020, pp. 2173–2178.

[48] A. Sahai and S. Mitter. “The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link—Part I: Scalar systems”. In: *IEEE Transactions on Information Theory* 52.8 (July 2006), pp. 3369–3395.

[49] R. T. Sukhavasi and B. Hassibi. “Linear error correcting codes with anytime reliability”. In: *2011 IEEE International Symposium on Information Theory Proceedings*. IEEE. July 2011, pp. 1748–1752.

[50] A. Khina, W. Halbawi, and B. Hassibi. “(Almost) practical tree codes”. In: *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE. July 2016, pp. 2404–2408.

[51] A. Lalitha et al. “Real-time binary posterior matching”. In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE. July 2019, pp. 2239–2243.

[52] C. Chang and A. Sahai. “Error exponents for joint source-channel coding with delay-constraints”. In: *Allerton Conference*. Citeseer. Sept. 2006.

[53] S. C. Draper and A. Khisti. “Truncated tree codes for streaming data: Infinite-memory reliability using finite memory”. In: *2011 8th International Symposium on Wireless Communication Systems*. IEEE. Nov. 2011, pp. 136–140.

[54] S. C. Draper, C. Chang, and A. Sahai. “Lossless coding for distributed streaming sources”. In: *IEEE Transactions on Information Theory* 60.3 (Dec. 2013), pp. 1447–1474.

[55] S.-H. Lee, V. Y. Tan, and A. Khisti. “Streaming data transmission in the moderate deviations and central limit regimes”. In: *IEEE Transactions on Information Theory* 62.12 (Oct. 2016), pp. 6816–6830.

[56] S.-H. Lee, V. Y. Tan, and A. Khisti. “Exact moderate deviation asymptotics in streaming data transmission”. In: *IEEE Transactions on Information Theory* 63.5 (Mar. 2017), pp. 2726–2736.

[57] P.-W. Su et al. “Random linear streaming codes in the finite memory length and decoding deadline regime – Part I: exact analysis”. In: *IEEE Transactions on Information Theory* (May 2022).

[58] P.-W. Su et al. “Sequentially mixing randomly arriving packets improves channel dispersion over block-based designs”. In: *IEEE International Symposium on Information Theory (ISIT)*. IEEE. June 2022.

[59] N. Guo and V. Kostina. “Optimal causal rate-constrained sampling for a class of continuous Markov processes”. In: *IEEE Transactions on Information Theory* 67.12 (Sept. 2021), pp. 7876–7890.

[60] N. Guo and V. Kostina. “Optimal causal rate-constrained sampling for a class of continuous Markov processes”. In: *IEEE International Symposium on Information Theory*. June 2020, pp. 2456–2461.

[61] A. Nayyar et al. “Optimal strategies for communication and remote estimation with an energy harvesting sensor”. In: *IEEE Transactions on Automatic Control* 58.9 (Mar. 2013), pp. 2246–2260.

[62] J. E. Figueroa-López. “Nonparametric estimation for Lévy models based on discrete-sampling”. In: *Lecture Notes- Monograph Series of the 3rd Erich L. Lehmann Symposium* 57 (Jan. 2009), pp. 117–146.

[63] J. Kappus. “Adaptive nonparametric estimation for Lévy processes observed at low frequency”. In: *Stochastic Processes and their Applications* 124.1 (Jan. 2014), pp. 730–758.

[64] X. Gao, E. Akyol, and T. Başar. “Optimal estimation with limited measurements and noisy communication”. In: *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE. Dec. 2015, pp. 1775–1780.

[65] X. Ren et al. “Infinite horizon optimal transmission power control for remote state estimation over fading channels”. In: *IEEE Transactions on Automatic Control* 63.1 (May 2017), pp. 85–100.

[66] J. Chakravorty and A. Mahajan. “Remote estimation over a packet-drop channel with Markovian state”. In: *IEEE Transactions on Automatic Control* 65.5 (July 2019), pp. 2016–2031.

[67] W. H. Equitz and T. M. Cover. “Successive refinement of information”. In: *IEEE Transactions on Information Theory* 37.2 (Mar. 1991), pp. 269–275.

[68] G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford University Press, July 2020.

[69] P. Mörters and Y. Peres. *Brownian motion*. Vol. 30. Cambridge University Press, Mar. 2010.

[70] N. Guo and V. Kostina. “Optimal causal rate-constrained sampling of the Wiener process”. In: *IEEE Transactions on Automatic Control* 67.4 (Apr. 2021), pp. 1776–1791.

[71] N. Guo and V. Kostina. “Optimal causal rate-constrained sampling of the Wiener process”. In: *Allerton Conference on Communication, Control, and Computing*. Sept. 2019, pp. 1090–1097.

[72] T. Berger. “Information rates of Wiener processes”. In: *IEEE Transactions on Information Theory* 16.2 (Mar. 1970), pp. 134–139.

[73] A. Kipnis, Y. C. Eldar, and A. J. Goldsmith. “Information rates of sampled Wiener processes”. In: *2016 IEEE International Symposium on Information Theory (ISIT)*. July 2016, pp. 740–744.

[74] P. Tallapragada, M. Franceschetti, and J. Cortés. “Event-triggered control under time-varying rates and channel blackouts”. In: *IFAC Journal of Systems and Control* 9 (Sept. 2019), p. 100064.

[75] G. N. Nair et al. “Feedback control under data rate constraints: An overview”. In: *Proceedings of the IEEE* 95.1 (Mar. 2007), pp. 108–137.

[76] M. Sun et al. “Quantized control of event-triggered networked systems with time-varying delays”. In: *Journal of the Franklin Institute* 356.17 (Nov. 2019), pp. 10368–10392.

[77] V. Kostina and B. Hassibi. “Rate-cost tradeoffs in control”. In: *IEEE Transactions on Automatic Control* 64.11 (Apr. 2019), pp. 4525–4540.

[78] T. E. Duncan and B. Pasik-Duncan. “A direct method for solving stochastic control problems”. In: *Communications in Information and Systems* 12.1 (2012), pp. 1–14.

[79] R. Munos and P. Bourgine. “Reinforcement learning for continuous stochastic control problems”. In: *Advances in neural information processing systems* 10 (1997).

[80] V. Anantharam and S. Verdú. “Bits through queues”. In: *IEEE Transactions on Information Theory* 42.1 (Jan. 1996), pp. 4–18.

[81] N. Guo and V. Kostina. “Instantaneous SED coding over a DMC”. In: *IEEE International Symposium on Information Theory*. IEEE. July 2021, pp. 148–153.

[82] N. Guo and V. Kostina. “Reliability function for streaming over a DMC”. In: *IEEE International Symposium on Information Theory*. June 2022, pp. 3204–3209.

[83] N. Guo and V. Kostina. “Reliability function for streaming over a DMC”. In: *submitted to IEEE Transactions on Information Theory* (June 2022).

[84] R. G. Gallager. *Information theory and reliable communication*. Vol. 588. New York: Wiley, Jan. 1968.

[85] I. Csiszár. “Joint source-channel error exponent”. In: *Problems of Control and Information Theory* 9.5 (1980), pp. 315–328.

[86] Y. Zhong, F. Alajaji, and L. L. Campbell. “On the joint source-channel coding error exponent for discrete memoryless systems”. In: *IEEE Transactions on Information theory* 52.4 (Apr. 2006), pp. 1450–1468.

[87] L. V. Truong and V. Y. Tan. “The reliability function of variable-length lossy joint source-channel coding with feedback”. In: *IEEE Transactions on Information Theory* 65.8 (Apr. 2019), pp. 5028–5042.

[88] L. J. Schulman. “Coding for interactive communication”. In: *IEEE Transactions on Information Theory* 42.6 (Nov. 1996), pp. 1745–1756.

[89] A. Antonini, R. Gimelshein, and R. D. Wesel. “Causal (progressive) encoding over binary symmetric channels with noiseless feedback”. In: *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE. July 2021, pp. 142–147.

[90] R. Gelles, R. Ostrovsky, and A. Roytman. “Efficient error-correcting codes for sliding windows”. In: *International Conference on Current Trends in Theory and Practice of Informatics*. Springer. Jan. 2014, pp. 258–268.

[91] B. Babcock et al. “Models and issues in data stream systems”. In: *Proceedings of ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. June 2002, pp. 1–16.

[92] S. Chandrasekaran and M. J. Franklin. “Streaming queries over streaming data”. In: *Proceedings of International Conference on Very Large Databases*. Elsevier. Aug. 2002, pp. 203–214.

[93] S. Babu and J. Widom. “Continuous queries over data streams”. In: *ACM Sigmod Record* 30.3 (Sept. 2001), pp. 109–120.

[94] D. Terry et al. “Continuous queries over append-only databases”. In: *ACM Sigmod Record* 21.2 (June 1992), pp. 321–330.

[95] N. Alon, Y. Matias, and M. Szegedy. “The space complexity of approximating the frequency moments”. In: *Journal of Computer and System Sciences* 58.1 (Feb. 1999), pp. 137–147.

[96] A. Arasu et al. “Characterizing memory requirements for queries over continuous data streams”. In: *ACM Transactions on Database Systems* 29.1 (Mar. 2004), pp. 162–194.

[97] B. Babcock, M. Datar, and R. Motwani. “Sampling from a moving window over streaming data”. In: *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*. Jan. 2002, pp. 633–634.

[98] S. Acharya, P. B. Gibbons, and V. Poosala. “Congressional samples for approximate answering of group-by queries”. In: *Proceedings of ACM SIGMOD International Conference on Management of Data*. June 2000, pp. 487–498.

[99] P. Berlin et al. “A simple converse of Burnashev’s reliability function”. In: *IEEE Transactions on Information Theory* 55.7 (June 2009), pp. 3074–3080.

[100] G. H. Hardy and J. E. Littlewood. “Some problems of Diophantine approximation”. In: *Acta mathematica* 37.1 (Dec. 1914), pp. 155–191.

[101] R. E. Korf. “From approximate to optimal solutions: A case study of number partitioning”. In: *14th International Joint Conference on Artificial Intelligence*. Aug. 1995, pp. 266–272.

[102] T. Kadota. “On the information stability of stationary ergodic processes”. In: *SIAM Journal on Applied Mathematics* 26.1 (Jan. 1974), pp. 176–182.

[103] T. H. Cormen et al. *Introduction to algorithms, 3rd-edition*. Cambridge, MA: The MIT Press, 2009.

[104] V. Kostina, Y. Polyanskiy, and S. Verdú. “Joint source-channel coding with feedback”. In: *IEEE Transactions on Information Theory* 63.6 (Feb. 2017), pp. 3502–3515.

[105] G. Forney. “Exponential error bounds for erasure, list, and decision feedback schemes”. In: *IEEE Transactions on Information Theory* 14.2 (1968), pp. 206–220.

[106] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp. “Lower bounds to error probability for coding on discrete memoryless channels. I”. In: *Information and Control* 10.1 (1967), pp. 65–103.

[107] R. Gallager. “A simple derivation of the coding theorem and some applications”. In: *IEEE Transactions on Information Theory* 11.1 (Jan. 1965), pp. 3–18.

[108] B. Hajek, K. Mitzel, and S. Yang. “Paging and registration in cellular networks: Jointly optimal policies and an iterative algorithm”. In: *IEEE Transactions on Information Theory* 54.2 (Jan. 2008), pp. 608–622.

[109] P. L. Clark. “The instructor’s guide to real induction”. In: *Mathematics Magazine* 92.2 (Mar. 2019), pp. 136–150.

[110] M. Brown and S. M. Ross. *Renewal reward processes*. Tech. rep. 93. Cornell University, Department of Operations Research, Sept. 1969.

[111] H. V. Poor. *An introduction to signal detection and estimation*. Springer Science & Business Media, Mar. 2013.

[112] E. Talvila. “Necessary and sufficient conditions for differentiating under the integral sign”. In: *The American Mathematical Monthly* 108.6 (June 2001), pp. 544–548.

[113] R. A. Gordon. *The integrals of lebesgue, denjoy, perron, and henstock*. 4. American Mathematical Soc., 1994.

[114] J. L. Lions. “Free boundary problems and impulse control”. In: *IFIP Technical Conference on Optimization Techniques*. Springer. May 1973, pp. 116–123.

[115] A. Bensoussan. “Impulse control and quasi-variational inequalities”. In: (1984).

[116] A. Friedman. “Optimal stopping problems in stochastic control”. In: *SIAM Review* 21.1 (Jan. 1979), pp. 71–80.

[117] Z. G. Li, C. Y. Wen, and Y. C. Soh. “Analysis and design of impulsive control systems”. In: *IEEE Transactions on Automatic Control* 46.6 (June 2001), pp. 894–897.

[118] A. Braides et al. *Gamma-convergence for beginners*. Vol. 22. Clarendon Press, 2002.

[119] D. Williams. *Probability with martingales*. Cambridge University Press, Feb. 1991.

[120] M. Fréchet. “Généralisation du théoreme des probabilités totales”. In: *Fundamenta mathematicae* 1.25 (1935), pp. 379–387.

[121] Y. Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

[122] H. G. Eggleston. *Convexity*. Cambridge University Press, 1958.

Appendix A

CAUSAL FREQUENCY-CONSTRAINED SAMPLING: PROOFS

A.1 Sufficient condition for (S.2)

Before we show the sufficient condition in Proposition 5 below, we first characterize the causal sampling policy in Definition 1.

Any causal sampling policy in Definition 1 can be characterized by a set-valued process we term *sampling-decision process*. It is a $\mathcal{B}_{\mathbb{R}}$-valued process $\{\mathcal{P}_t\}_{t=0}^T$ adapted to $\{\mathcal{F}_t\}_{t=0}^T$, which decides the stopping times

$$\tau_{i+1} = \inf\{t \geq \tau_i \colon \tilde{X}_t \notin \mathcal{P}_t\}, \tag{A.1}$$

where the mean-square residual error process $\{\tilde{X}_t\}_{t=0}^T$ in (A.1) is defined as

$$\tilde{X}_t \triangleq X_t - \mathbb{E}[X_t|X_{\tau_i}, \tau_i],\ \forall t \in [\tau_i, \tau_{i+1}). \tag{A.2}$$

Given any sampling policy $\tau_1, \tau_2, \ldots$ and a realization of the process up to time t, we can set

$$\mathcal{P}_t = \begin{cases} \mathcal{A}_t, & t \neq \tau_i,\ i = 1, 2, \ldots, \\ \mathcal{A}_t^c, & t = \tau_i,\ i = 1, 2, \ldots, \end{cases} \tag{A.3}$$

where $\mathcal{A}_t$ is any Borel set the realization of $\tilde{X}_t$ belongs to. Without assumption (S.2), $\mathcal{P}_t$ for $t \in [\tau_i, \tau_{i+1})$ can depend on the input process $\{X_s\}_{s=0}^t$ up to time t. Under assumption (S.2), $\mathcal{P}_t$ for $t \in [\tau_i, \tau_{i+1})$ can only depend on the stopping time τ_i and $\{\tilde{X}_s\}_{s=\tau_i}^t$ (A.2).

We proceed to present the sufficient condition on the stochastic process under which the optimal sampling policy satisfies (S.2). We define notations that will be used in Proposition 5 below. Consider a sampling-decision process $\{\mathcal{P}_t\}_{t=\tau_k}^T$ with stopping times $\tau_k, \tau_{k+1}, \ldots$, the mean-square residual error $\tilde{X}_t$ (A.2), and the MMSE decoding policy $\bar{X}_t$ (2.2). The value of $\{\mathcal{P}_t\}_{t=\tau_k}^T$ at time $t \in [\tau_k, T]$ only depends on $\{X_s - \mathbb{E}[X_s|X_{\tau_k}, \tau_k]\}_{s=\tau_k}^t$ and τ_k, i.e.,

$$\mathcal{P}_t = \mathcal{P}_t(\{X_s - \mathbb{E}[X_s|X_{\tau_k}, \tau_k]\}_{s=\tau_k}^t, \tau_k),\ t \in [\tau_k, T]. \tag{A.4}$$

Denote by $\Pi_{[\tau_k, T]}$ the set of all sampling-decision processes of the form (A.4). As a result, the stopping times associated with $\{\mathcal{P}_t\}_{\tau_k}^T \in \Pi_{[\tau_k, T]}$ only satisfy (S.2) at

$i = k$. Let $N\left(\{\mathcal{P}_t\}_{t=\tau_k}^T\right)$ represent the number of samples taken between $[\tau_k, T]$ under $\{\mathcal{P}_t\}_{t=\tau_k}^T$. We denote

$$\underline{D}_r(\phi) \triangleq \min_{\substack{\{\mathcal{P}_t\}_{t=r}^T \in \Pi_{[r,T]}: \\ \frac{1}{T}\mathbb{E}[N\left(\{\mathcal{P}_t\}_{t=r}^T\right)|\tau_k=r] \leq \phi}} \frac{1}{T}\mathbb{E}\left[\int_{\tau_k}^T (X_t - \bar{X}_t)^2 \mathrm{d}t \middle| \tau_k = r\right]. \tag{A.5}$$

Consider an arbitrary sampling-decision process $\{\mathcal{P}'_t\}_{t=0}^T$ (A.1) with stopping times $\tau'_1, \tau'_2, \ldots$, the mean-square residual error $\tilde{X}'_t$, and the MMSE decoding policy $\bar{X}'_t$. The value of the sampling-decision process $\{\mathcal{P}'_t\}_{t=0}^T$ at time t can depend on $\{X_s\}_{s=0}^t$, i.e., for all $t \in [\tau'_k, T]$,

$$\mathcal{P}'_t = \mathcal{P}'_t\left(\{X_s\}_{s=0}^{\tau'_k}, \left\{X_s - \mathbb{E}[X_s|X_{\tau'_k}, \tau'_k]\right\}_{s=\tau'_k}^t, \tau'_k\right). \tag{A.6}$$

Denote by $\Pi'_{[\tau'_k,T]}$ the set of all sampling-decision processes of the form (A.6).

Proposition 5. *For a stochastic process $\{X_t\}_{t=0}^T$ satisfying (P.1)–(P.3), if $\underline{D}_r(\phi)$ in (A.5) is a convex function in ϕ for all $k = 0, 1, \ldots$ and $r \in [0, T]$, then the optimal sampling policy satisfies (S.2).*

Proof. Fix an arbitrary sampling-decision process $\{\mathcal{P}'_t\}_{t=\tau'_k}^T \in \Pi'_{[\tau'_k,T]}$ at $\tau'_k = r$. To show that the optimal sampling policy of $\{X_t\}_{t=0}^T$ satisfies (S.2), it suffices to show that for all $k = 0, 1, \ldots$, $\underline{D}_r\left(\frac{1}{T}\mathbb{E}[N(\{\mathcal{P}'_t\}_{t=r}^T)|\tau'_k = r]\right)$ is no larger than the MSE achieved by $\{\mathcal{P}'_t\}_{t=r}^T$, i.e.,

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{T}\int_{\tau'_k}^T (X_t - \bar{X}'_t)^2 \mathrm{d}t \middle| \tau'_k = r\right] \\ \geq\ &\underline{D}_r\left(\frac{1}{T}\mathbb{E}[N(\{\mathcal{P}'_t\}_{t=r}^T)|\tau'_k = r]\right). \end{aligned} \tag{A.7}$$

We fix an arbitrary realization of $\{X_s\}_{s=0}^r = x$ that leads to $\tau'_k = r$, and we construct $\{\mathcal{P}_t\}_{t=r}^T$ as

$$\mathcal{P}_t = \mathcal{P}'_t\left(x, \{X_s - \mathbb{E}[X_s|X_r, r]\}_{s=r}^t, r\right). \tag{A.8}$$

The sampling-decision process $\{\mathcal{P}_t\}_{t=r}^T$ (A.8) satisfies the minimization constraint in (A.5) with

$$\phi = \frac{1}{T}\mathbb{E}[N(\{\mathcal{P}'_t\}_{t=r}^T)|\{X_s\}_{s=0}^r = x, \tau'_k = r] \tag{A.9}$$

due to the reasons that follow. The process $\{\mathcal{P}_t\}_{t=r}^T$ (A.8) belongs to $\Pi_{[r,T]}$ since it samples the input process after time r as if it has observed $\{X_s\}_{s=0}^r = x$ regardless of the actual realization of $\{X_s\}_{s=0}^r$. Since $\{\tilde{X}_t\}_{t\geq\tau_k}$, at $\tau_k = r$, is independent of $\mathcal{F}_r$ by (P.3-a), and τ_{i+1}, $i \geq k$, is conditionally independent of $\{X_s\}_{s=0}^r$ given $\tau_k = r$ due to $\{\mathcal{P}_t\}_{t=r}^T \in \Pi_{[r,T]}$, we conclude that under $\{\mathcal{P}_t\}_{t=r}^T$, the random process $\{X_t - \bar{X}_t\}_{t=r}^T$ conditioned on $\tau_k = r$ has the same probability distribution as $\{X_t - \bar{X}'_t\}_{t=r}^T$ under $\{\mathcal{P}'_t\}_{r=0}^T$ conditioned on $\{X_s\}_{s=0}^r = x, \tau'_k = r$. This implies that $\{\mathcal{P}_t\}_{t=r}^T$ (A.8) achieves average sampling frequency ϕ (A.9), and that

$$\begin{aligned}
&\mathbb{E}\left[\int_{\tau'_k}^T (X_t - \bar{X}'_t)^2 \mathrm{d}t \middle| \{X_s\}_{s=0}^r = x, \tau'_k = r\right] \\
&= \mathbb{E}\left[\int_{\tau_k}^T (X_t - \bar{X}_t)^2 \mathrm{d}t \middle| \{X_s\}_{s=0}^r, \tau_k = r\right] && \text{(A.10a)} \\
&= \mathbb{E}\left[\int_{\tau_k}^T (X_t - \bar{X}_t)^2 \mathrm{d}t \middle| \tau_k = r\right] && \text{(A.10b)} \\
&\geq \underline{D}_r\left(\frac{1}{T}\mathbb{E}[N(\{\mathcal{P}'_t\}_{t=r}^T) | \{X_s\}_{s=0}^r = x, \tau'_k = r]\right), && \text{(A.10c)}
\end{aligned}$$

where (A.10c) holds because $\{\mathcal{P}_t\}_{t=\tau_k}^T \in \Pi_{[\tau_k,T]}$. Since (A.10c) holds for an arbitrary realization of $\{X_s\}_{s=0}^r$ compatible with $\tau'_k = r$, it holds almost surely that

$$\begin{aligned}
&\mathbb{E}\left[\int_{\tau'_k}^T (X_t - \bar{X}'_t)^2 \mathrm{d}t \middle| \{X_s\}_{s=0}^r, \tau'_k = r\right] && \text{(A.11)} \\
&\geq \underline{D}_r\left(\frac{1}{T}\mathbb{E}[N(\{\mathcal{P}'_t\}_{t=r}^T) | \{X_s\}_{s=0}^r, \tau'_k = r]\right).
\end{aligned}$$

Taking an expectation of (A.11), we conclude

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\int_{\tau'_k}^T (X_t - \bar{X}'_t)^2 \mathrm{d}t \middle| \tau'_k = r\right] && \text{(A.12)} \\
&\geq \mathbb{E}\left[\underline{D}_r\left(\frac{1}{T}\mathbb{E}[N(\{\mathcal{P}'_t\}_{t=r}^T) | \{X_s\}_{s=0}^r, \tau'_k = r]\right) \middle| \tau'_k = r\right], && \text{(A.13)}
\end{aligned}$$

and (A.7) follows via Jensen's inequality. □

A.2 Proof of Theorem 1

Tools

We first introduce Lemmas 2–5 that supply majorization and real induction tools for proving Theorem 1.

Function f *majorizes* g, $f \succ g$, if and only if for any Borel measurable set $\mathcal{B} \in \mathcal{B}_{\mathbb{R}}$ with finite Lebesgue measure, there exits a Borel measurable set $\mathcal{A} \in \mathcal{B}_{\mathbb{R}}$ with the

same Lebesgue measure, such that [2]

$$\int_{\mathcal{B}} g(x)dx \leq \int_{\mathcal{A}} f(x)dx. \tag{A.14}$$

Function $f : \mathbb{R} \to \mathbb{R}$ is *even* if $f(x) = f(-x)$ for all $x \in \mathbb{R}$.

Function $f : \mathbb{R} \to \mathbb{R}$ is *quasi-concave* if for all $x, y \in \mathbb{R}$, $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\}. \tag{A.15}$$

We denote by $\mathbb{1}_{(a,b)}(x)$ an indicator function that is equal to 1 if and only if $x \in (a, b)$.

Lemmas 2–4, stated next, show several majorization properties of pdfs.

Lemma 2. *([2, Lemma 2]) Fix two pdfs f_X and g_X, such that f_X is even and quasi-concave and $f_X \succ g_X$. Fix a scalar $c > 0$, and a function $h\colon \mathbb{R} \to [0, 1]$, such that*

$$\int_{\mathbb{R}} f_X(x)\mathbb{1}_{(-c,c)}(x)dx = \int_{\mathbb{R}} g_X(x)h(x)dx, \tag{A.16}$$

Then,

$$f_{X|X\in(-c,c)} \succ g'_X, \tag{A.17}$$

where the pdfs $f_{X|X\in(-c,c)}$ and g'_X are given by,

$$\begin{aligned} f_{X|X\in(-c,c)}(x) &= \frac{f_X(x)\mathbb{1}_{(-c,c)}(x)}{\int_{\mathbb{R}} f_X(x)\mathbb{1}_{(-c,c)}(x)dx} \\ g'_X(x) &= \frac{g_X(x)h(x)}{\int_{\mathbb{R}} g_X(x)h(x)dx}. \end{aligned} \tag{A.18}$$

Lemma 3. *([108, Lemma 6.7]) Fix two pdfs f_X and g_X, such that f_X is even and quasi-concave and that f_X majorizes g_X, $f_X \succ g_X$. Fix an even and quasi-concave pdf r_Y. Then, the convolution of f_X and r_Y majorizes the convolution of g_X and r_Y,*

$$f_X * r_Y \succ g_X * r_Y, \tag{A.19}$$

Furthermore, $f_X * r_Y$ is even and quasi-concave.

Lemma 4. *([2, Lemma 4]) Fix two pdfs f_X and g_X such that f_X is even and quasi-concave and that f_X majorizes g_X, $f_X \succ g_X$. Then,*

$$\int_{\mathbb{R}} x^2 f_X(x)dx \leq \int_{\mathbb{R}} (x-y)^2 g_X(x)dx, \ \forall y \in \mathbb{R}. \tag{A.20}$$

Lemma 5, stated next, provides a mathematical proof technique called *real induction*. We will use it to prove that the assertions in Lemma 6, stated below, hold on a continuous interval.

Lemma 5. *(Real induction [109, Thm. 2]) A subset $S \subset [a,b]$, $a < b$ is called inductive if*

1) $a \in S$;

2) If $a \leq x < b$, $x \in S$, *then there exists* $y > x$ *such that* $[x,y] \in S$;

3) If $a \leq x < b$, $[a,x) \in S$, *then* $x \in S$.

If a subset $S \subset [a,b]$ is inductive, then $S = [a,b]$.

A technical lemma

We define the following notations for two sampling-decision processes $\{\mathcal{P}_t\}_{t=0}^T$ and $\{\mathcal{P}_t^{\text{sym}}\}_{t=0}^T$ (see Appendix A.1). Fix an arbitrary sampling-decision process $\{\mathcal{P}_t\}_{t=0}^T$ (A.1) satisfying (S.1)–(S.2). It gives rise to a sampling policy with stopping times $\tau_1, \tau_2, \ldots$ via (A.1). We recall the definition of the mean-square residual error (MSRE) process $\{\tilde{X}_t\}_{t=0}^T$ in (P.3) and denote the MSRE process under $\{\mathcal{P}_t\}_{t=0}^T$ as

$$\begin{aligned} \tilde{X}_t &= \tilde{X}_t(\{\mathcal{P}_s\}_{s=0}^T) && \text{(A.21a)} \\ &\triangleq X_t - \mathbb{E}[X_t | X_{\tau_i}, \tau_i], t \in [\tau_i, \tau_{i+1}). && \text{(A.21b)} \end{aligned}$$

We define the residual error estimate (REE) process $\{\bar{\tilde{X}}_t\}_{t=0}^T$ under $\{\mathcal{P}_t\}_{t=0}^T$ as

$$\begin{aligned} \bar{\tilde{X}}_t &= \bar{\tilde{X}}_t(\{\mathcal{P}_s\}_{s=0}^T) && \text{(A.22a)} \\ &\triangleq \bar{X}_t - \mathbb{E}[X_t | X_{\tau_i}, \tau_i] && \text{(A.22b)} \\ &= \mathbb{E}[\tilde{X}_t | \{X_{\tau_j}\}_{j=1}^i, \tau^i, t < \tau_{i+1}] && \text{(A.22c)} \\ &= \mathbb{E}[\tilde{X}_t | \tau_i, t < \tau_{i+1}],\ t \in [\tau_i, \tau_{i+1}), && \text{(A.22d)} \end{aligned}$$

where $\bar{X}_t = \bar{X}_t(\{\mathcal{P}_s\}_{s=0}^T)$ is the MMSE decoding policy defined in (2.2); the equality in (A.22c) holds since $\mathbb{E}[X_t | X_{\tau_i}, \tau_i] \in \sigma(\{X_{\tau_j}\}_{j=1}^i, \tau^i, t < \tau_{i+1})$; (A.22d) holds because $\tilde{X}_t$ is independent of $\{X_{\tau_j}\}_{j=1}^i, \tau^i$ due to (P.3-a), and the event $\{t < \tau_{i+1}\}$ is independent of $\{X_{\tau_j}\}_{j=1}^i, \tau^{i-1}$ given τ_i due to (S.2). We recall that $N(\{\mathcal{P}_t\}_{t=0}^T)$ defined above Proposition 5 in Appendix A.1 represents the number of stopping times in $[0,T]$, and we simplify this notation as

$$N \triangleq N(\{\mathcal{P}_t\}_{t=0}^T). \tag{A.23}$$

We denote the left-closed continuous interval

$$\Omega_{\tau_{i+1}}(s) \triangleq \{t \in [s,T] \colon \mathbb{P}[\tau_{i+1} > t | \tau_i = s] > 0\}, \tag{A.24}$$

for all $s \in \mathrm{Supp}(f_{\tau_i})$, and the left-open continuous interval

$$\bar{\Omega}_{\tau_{i+1}}(s) \triangleq \Omega_{\tau_{i+1}}(s) \setminus \{s\}. \tag{A.25}$$

Given $\{\mathcal{P}_t\}_{t=0}^T$, we construct a sampling-decision process $\{\mathcal{P}_t^{\mathrm{sym}}\}_{t=0}^T$ (A.1) of the form (2.9), which via (A.1) is associated with a sampling policy with stopping times $\tau_1', \tau_2', \dots$, such that the symmetric thresholds $\{a_i(r,s)\}_{r=s}^T$ of $\{\mathcal{P}_t^{\mathrm{sym}}\}_{t=0}^T$ satisfy for all $s \in \mathrm{Supp}(f_{\tau_i})$, $t \in [s,T]$,

$$\begin{aligned} & \mathbb{P}[\tilde{X}_r' \in (-a_i(r,s), a_i(r,s)), \forall r \in [s,t] | \tau_i' = s] \\ = \; & \mathbb{P}[\tau_{i+1} > t | \tau_i = s]. \end{aligned} \tag{A.26}$$

This is possible since by adjusting the thresholds, the left side of (A.26) can be equal to any non-increasing function in t bounded between $[0,1]$. Under $\{\mathcal{P}_t^{\mathrm{sym}}\}_{t=0}^T$ (A.26), for all $s \in \mathrm{Supp}(f_{\tau_i})$, $i = 1, 2, \dots$, it holds that

$$\Omega_{\tau_i}(s) = \Omega_{\tau_i'}(s), \tag{A.27}$$

$$\bar{\Omega}_{\tau_i}(s) = \bar{\Omega}_{\tau_i'}(s). \tag{A.28}$$

We denote the MSRE and the REE processes and the number of stopping times on $[0,T]$ under $\{\mathcal{P}_t^{\mathrm{sym}}\}_{t=0}^T$ respectively by

$$\tilde{X}_t' = \tilde{X}_t(\{\mathcal{P}_s^{\mathrm{sym}}\}_{s=0}^T), \tag{A.29}$$

$$\bar{\tilde{X}}_t' = \bar{\tilde{X}}_t(\{\mathcal{P}_s^{\mathrm{sym}}\}_{s=0}^T) = 0, \tag{A.30}$$

$$N' = N(\{\mathcal{P}_s^{\mathrm{sym}}\}_{s=0}^T), \tag{A.31}$$

where (A.30) holds since we can write $\bar{\tilde{X}}_t'$ as (A.22d) with τ_i replaced by τ_i' using the argument that justifies (A.22d); $\tilde{X}_t'$ has an even and quasi-concave pdf due to the assumption (P.3-b), and the pdf of $\tilde{X}_t$ conditioned on $\tau_i', t < \tau_{i+1}'$ under a symmetric threshold sampling-decision process of the form (2.9) is still even and quasi-concave.

We denote the following probabilities

$$\mathbb{Q}_i(a,b,c,d) \triangleq \mathbb{P}[\tau_{i+1} > a | \tau_{i+1} > b, \tau_i = c, \tilde{X}_a = d] \tag{A.32a}$$

$$\mathbb{Q}_i'(a,b,c,d) \triangleq \mathbb{P}[\tau_{i+1}' > a | \tau_{i+1}' > b, \tau_i' = c, \tilde{X}_a' = d]. \tag{A.32b}$$

We proceed to introduce Lemma 6 using the notations defined in (A.21)–(A.32b). We will use the assertions in Lemma 6 to compare the MSEs achieved by $\{\mathcal{P}_t\}_{t=0}^T$ and $\{\mathcal{P}_t^{\text{sym}}\}_{t=0}^T$.

Lemma 6. *The pdfs $f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>t}$ and $f_{\tilde{X}'_t|\tau'_i=s,\tau'_{i+1}>t}$ exist for all $s \in \text{Supp}(f_{\tau_i})$, $t \in \bar{\Omega}_{\tau_{i+1}}(s)$. Furthermore, for all $s \in \text{Supp}(f_{\tau_i})$, $t \in \bar{\Omega}_{\tau_{i+1}}(s)$, it holds that*

$$f_{\tilde{X}'_t|\tau'_i=s,\tau'_{i+1}>t} \succ f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>t}, \tag{A.33}$$

$$f_{\tilde{X}'_t|\tau'_i=s,\tau'_{i+1}>t} \text{ is even and quasi-concave.} \tag{A.34}$$

Proof of Lemma 6. We prove that $f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>t}$ exists. The proof that $f_{\tilde{X}'_t|\tau'_i=s,\tau'_{i+1}>t}$ exists is similar. Since $\tilde{X}_t$ at $t \geq \tau_i = s$, is independent of $\mathcal{F}_s$ by (P.3-a) and is equal to $R_t(s,s)$ by (P.3-b), we compute $f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>s}$ using (2.5),

$$f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>s} = f_{R_t(s,s)}. \tag{A.35}$$

Thus, $f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>s}$ exists since $f_{R_t(s,s)}$ is a valid pdf by (P.3-b). To establish that $f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>t}(y)$ exists, we compute

$$f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>t}(y) = f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>s,\tau_{i+1}>t}(y) \tag{A.36a}$$

$$= \frac{\mathbb{Q}_i(t,s,s,y) f_{\tilde{X}_t|\tau_i=s,\tau_{i+1}>s}(y)}{\mathbb{P}[\tau_{i+1}>t|\tau_i=s,\tau_{i+1}>s]}, \tag{A.36b}$$

where (A.36a) holds since $\tau_{i+1} > t$ implies $\tau_{i+1} > s$. In (A.36b), we observe that for all $t \in \bar{\Omega}_{\tau_{i+1}}(s)$, the pdf $f_{\tilde{X}_t|\tau_{i+1}>s,\tau_i=s}$ exists by (A.35); the denominator of (A.36b) is nonzero. We conclude that the pdf $f_{\tilde{X}_t|\tau_i=s,\tau_i>t}$ exists for all $s \in \text{Supp}(f_{\tau_i})$, $t \in \bar{\Omega}_{\tau_{i+1}}(s)$.

The assertion (A.33) holds if and only if

(a) for all $s \in \text{Supp}(f_{\tau_i})$, $t \in \bar{\Omega}_{\tau_{i+1}}(s)$ and for any Borel measurable set $\mathcal{B} \in \mathcal{B}_{\mathbb{R}}$ with finite Lebesgue measure, there exists a Borel measurable set $\mathcal{A} \in \mathcal{B}_{\mathbb{R}}$ with the same Lebesgue measure, such that

$$\begin{aligned} &\mathbb{P}[\tilde{X}'_t \in \mathcal{A}|\tau'_i = s, \tau'_{i+1} > t] \\ \geq\ &\mathbb{P}[\tilde{X}_t \in \mathcal{B}|\tau_i = s, \tau_{i+1} > t], \end{aligned} \tag{A.37}$$

holds. This is because (A.37) is a rewrite of (A.33) using the definition of majorization (A.14).

The assertion (A.34) holds if and only if for all $s \in \text{Supp}(f_{\tau_i})$, $t \in \bar{\Omega}_{\tau_{i+1}}(s)$, all of the following hold:

(b) the conditional cdf $\mathbb{P}[\tilde{X}'_t \le y|\tau'_i = s, \tau'_{i+1} > t]$ is convex for $y < 0$ and is concave for $y > 0$;

(c) for any $y > 0$,

$$\begin{aligned}&\mathbb{P}[\tilde{X}'_t \in (0, y]|\tau'_i = s, \tau'_{i+1} > t]\\ =&\mathbb{P}[\tilde{X}'_t \in [-y, 0)|\tau'_i = s, \tau'_{i+1} > t].\end{aligned} \tag{A.38}$$

This is because $f_{\tilde{X}'_t|\tau'_i=s,\tau'_{i+1}>t}$ is quasi-concave if and only if (b) holds, and $f_{\tilde{X}'_t|\tau'_i=s,\tau'_{i+1}>t}$ is even if and only if (c) holds.

Items (a)–(c) facilitate proving that the assertions (A.33)–(A.34) hold on the left-open interval $\bar{\Omega}_{\tau_{i+1}}(s)$. Real induction, which must be used on a left-closed interval, does not apply to show (A.33)–(A.34) directly, since the densities in (A.33)–(A.34) do not exist at $t = s$. Instead, we apply real induction to show (a)–(c). Using real induction in Lemma 5, we verify that conditions 1), 3), 2) in Lemma 5 hold for (a)–(c) in on $t \in \Omega_{\tau_{i+1}}(s)$ one by one.

To verify that the condition 1) in Lemma 5 holds, we need to show that (a)–(c) hold for $t = s$. This is trivial since

$$\begin{aligned}&\mathbb{P}[\tilde{X}'_s = 0|\tau'_i = s, \tau'_{i+1} > s]\\ =\ &\mathbb{P}[\tilde{X}_s = 0|\tau_i = s, \tau_{i+1} > s]\\ =\ &1.\end{aligned} \tag{A.39}$$

Next, we show that condition 3) in Lemma 5 holds, that is, assuming that (a)–(c) hold for all $t \in [s, r)$, $r \in \bar{\Omega}_{\tau_{i+1}}(s)$, we prove that (a)–(c) hold for $t = r$. Equivalently, we show that (A.33)–(A.34) hold for $t = r$. Let $\delta \in (0, r - s]$. At time $t = r$, we calculate the left side of (A.33) as

$$\begin{aligned}&f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r}(y)\\ =\ &\lim_{\delta\to 0^+} f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta,\tau'_{i+1}>r}(y) &\text{(A.40a)}\\ =\ &\lim_{\delta\to 0^+} \frac{\mathbb{Q}'_i(r, r-\delta, s, y) f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta}(y)}{\int_{\mathbb{R}} \mathbb{Q}'_i(r, r-\delta, s, y) f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta}(y)dy} &\text{(A.40b)}\\ =\ &\lim_{\delta\to 0^+} \frac{\mathbb{1}_{(-a_i(r,s),a_i(r,s))}(y) f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta}(y)}{\int_{\mathbb{R}} \mathbb{1}_{(-a_i(r,s),a_i(r,s))}(y) f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta}(y)dy}, &\text{(A.40c)}\end{aligned}$$

where (A.40a) holds since the event $\tau'_{i+1} > r$ implies the event $\tau'_{i+1} > r - \delta$; the pdf $f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta}$ in (A.40b) exists since (A.36) holds with $\tilde{X}_t$, $\tau_i = s$, $\tau_{i+1} > s$,

$\tau_{i+1} > t$ replaced by $\tilde{X}'_r$, $\tau'_i = s$, $\tau'_{i+1} > s$, $\tau'_{i+1} > r - \delta$, respectively; (A.40c) holds since

$$\lim_{\delta \to 0^+} \mathbb{Q}'_i(r, r-\delta, s, y) = \mathbb{1}_{(-a_i(r,s), a_i(r,s))}(y). \tag{A.41}$$

Similarly, replacing $\mathbb{Q}'_i$ in (A.40b) by $\mathbb{Q}_i$, we calculate the right side of (A.33) as

$$\begin{aligned} & f_{\tilde{X}_r|\tau_i=s,\tau_{i+1}>r}(y) \\ = & \lim_{\delta \to 0^+} \frac{\mathbb{Q}_i(r, r-\delta, s, y) f_{\tilde{X}_r|\tau_i=s,\tau_{i+1}>r-\delta}(y)}{\int_{\mathbb{R}} \mathbb{Q}_i(r, r-\delta, s, y) f_{\tilde{X}_r|\tau_i=s,\tau_{i+1}>r-\delta}(y) dy}, \end{aligned} \tag{A.42}$$

where the pdf $f_{\tilde{X}_r|\tau_i=s,\tau_{i+1}>r-\delta}(y)$ exists since (A.36) holds with $\tilde{X}_t$, $\tau_{i+1} > t$ replaced by $\tilde{X}_r$, $\tau_{i+1} > r - \delta$ respectively.

To check that (A.33) holds at $t = r$, we first prove that $f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta}$ majorizes $f_{\tilde{X}_r|\tau_i=s,\tau_{i+1}>r-\delta}$. Note that $R_r(r-\delta, s)$ is independent of $\{\tilde{X}_t\}_{t=0}^{r-\delta}$ due to (P.3-a), and thus is independent of the event $\{\tau'_{i+1} > r - \delta, \tau'_i = s\}$. We obtain $\tilde{X}'_r$ using (2.5),

$$f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta} = f_{q_r(r-\delta)\tilde{X}'_{r-\delta}|\tau'_i=s,\tau'_{i+1}>r-\delta} * f_{R_r(r-\delta,s)}. \tag{A.43}$$

By (A.43) and the inductive hypothesis that (a)–(c) holds for $t \in [s, r)$, the assumptions in Lemma 3 are satisfied with $f_X \leftarrow f_{q_r(r-\delta)\tilde{X}'_{r-\delta}|\tau'_i=s,\tau'_{i+1}>r-\delta}$, $g_X \leftarrow f_{q_r(r-\delta)\tilde{X}_{r-\delta}|\tau_i=s,\tau_{i+1}>r-\delta}$, $r_Y \leftarrow f_{R_r(r-\delta,s)}$. We conclude that

$$f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta} \succ f_{\tilde{X}_r|\tau_i=s,\tau_{i+1}>r-\delta}, \tag{A.44}$$

$$f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta} \text{ is even and quasi-concave.} \tag{A.45}$$

Due to (A.45) and the fact that the indicator function in (A.40c) is over an interval symmetric about zero, we conclude (A.34) holds for $t = r$. By (A.26), (A.44) and (A.45), the assumptions in Lemma 2 are satisfied with $f_X \leftarrow f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r-\delta}$, $g_X \leftarrow f_{\tilde{X}_r|\tau_i=s,\tau_{i+1}>r-\delta}$, $f_{X|X\in(-c,c)} \leftarrow f_{\tilde{X}'_r|\tau'_i=s,\tau'_{i+1}>r}$, and $g'_X \leftarrow f_{\tilde{X}_r|\tau_i=s,\tau_{i+1}>r}$, $c \leftarrow a_i(r,s)$, $h \leftarrow \mathbb{Q}_i(r, r-\delta, s, y)$. Thus, we conclude that (A.33) holds for $t = r$. Therefore, (A.33)–(A.34) hold for $t = r$, i.e., (a)–(c) hold for $t = r$.

To prove that the condition 2) in Lemma 5 holds, we assume (a)–(c) hold for $t = r$, and prove that the following holds:

$$\lim_{\delta \to 0^+} f_{\tilde{X}'_{r+\delta}|\tau'_i=s,\tau'_{i+1}>r+\delta} \succ \lim_{\delta \to 0^+} f_{\tilde{X}_{r+\delta}|\tau_i=s,\tau_{i+1}>r+\delta}, \tag{A.46a}$$

$$\lim_{\delta \to 0^+} f_{\tilde{X}'_{r+\delta}|\tau'_i=s,\tau'_{i+1}>r+\delta} \text{ is even and quasi-concave.} \tag{A.46b}$$

The right and the left sides of (A.46a) are equal to (A.40c) and (A.42) respectively with r replaced by $r+\delta$. It is easy to see that (A.43)–(A.45) and the assumptions in Lemma 2 hold with r replaced by $r+\delta$. Thus, we conclude that (A.46) holds.

Using the real induction in Lemma 5, we have shown that (a)–(c) hold for all $s \in \mathrm{Supp}(f_{\tau_i})$, $t \in \Omega_{\tau_{i+1}}(s)$. Thus, (A.33)–(A.34) hold for all $s \in \mathrm{Supp}(f_{\tau_i})$, $t \in \bar{\Omega}_{\tau_{i+1}}(s)$. □

Proof of Theorem 1

The sampling-decision process $\{\mathcal{P}_t^{\mathrm{sym}}\}_{t=0}^T$ leads to the same average sampling frequency as $\{\mathcal{P}_t\}_{t=0}^T$. This is because (A.26) implies that for all $s \in \mathrm{Supp}(f_{\tau_i})$, $t \in [s, T]$,

$$\mathbb{P}[\tau_{i+1} > t | \tau_i = s] = \mathbb{P}[\tau'_{i+1} > t | \tau'_i = s]. \tag{A.47}$$

Together with the Markov property of the stopping times (assumption (S.2)), (A.47) implies that the joint distribution of $\tau_1, \tau_2, \ldots$ is equal to the joint distribution of $\tau'_1, \tau'_2, \ldots$ We conclude that $\{\mathcal{P}_t\}_{t=0}^T$ and $\{\mathcal{P}_t^{\mathrm{sym}}\}_{t=0}^T$ lead to the same average sampling frequency

$$\mathbb{E}[N] = \mathbb{E}[N']. \tag{A.48}$$

Next, we show $\{\mathcal{P}_t^{\mathrm{sym}}\}_{t=0}^T$ achieves an MSE no larger than that achieved by $\{\mathcal{P}_t\}_{t=0}^T$. Due to (A.22d), (A.30), and (A.33)–(A.34) in Lemma 6, we can apply Lemma 4 with $f_X \leftarrow f_{\tilde{X}'_t | \tau'_i = s, \tau'_{i+1} > t}$ and $g_X \leftarrow f_{\tilde{X}_t | \tau_i = s, \tau_{i+1} > t}$, yielding

$$\mathbb{E}\left[(\tilde{X}_t - \bar{\tilde{X}}_t)^2 | \tau_i = s, \tau_{i+1} > t\right] \geq \mathbb{E}\left[\tilde{X}'^2_t | \tau'_i = s, \tau'_{i+1} > t\right]. \tag{A.49}$$

Combining (A.47) and (A.49), we conclude by law of total expectation that $\{\mathcal{P}_t^{\mathrm{sym}}\}_{t=0}^T$ achieves an MSE no larger than that achieved by $\{\mathcal{P}_t\}_{t=0}^T$.

A.3 Proof of Corollary 1.1

Under a symmetric threshold sampling policy (2.9), the MMSE decoding policy in (2.2) can be expanded as, for $\tau_i \leq t < \tau_{i+1}$,

$$\begin{aligned}
\bar{X}_t =& \mathbb{E}[X_t | \{X_{\tau_j}\}_{j=1}^i, \tau^i, t < \tau_{i+1}] && \text{(A.50a)}\\
=& \bar{\tilde{X}}_t + \mathbb{E}[X_t | X_{\tau_i}, \tau_i] && \text{(A.50b)}\\
=& \mathbb{E}[X_t | X_{\tau_i}, \tau_i], && \text{(A.50c)}
\end{aligned}$$

where $\bar{\tilde{X}}_t$ in (A.50b) is equal to $\bar{\tilde{X}}'_t$ in (A.30), thus is equal to zero.

A.4 Proof of Corollary 1.2

Given any causal sampling policy such that (2.3) is satisfied with a strict inequality, we construct a causal sampling policy that satisfies (2.3) with equality and leads to an MSE no worse than that achieved by the given causal sampling policy.

Given an arbitrary symmetric threshold sampling policy (2.9) with stopping times $\tau_1, \tau_2, \ldots$, we denote by N_t the number of samples taken in $[0, t]$. Let t', $t' \in (0, T)$ be a dummy deterministic time. We decompose the MSE under the given sampling policy as

$$\mathbb{E}\left[\sum_{i=0}^{N_{t'}-1} \int_{\tau_i}^{\tau_{i+1}} (X_t - \mathbb{E}[X_t|X_{\tau_i}, \tau_i])^2 \mathrm{d}t\right] \tag{A.51a}$$

$$+\mathbb{E}\left[\int_{\tau_{N_{t'}}}^{t'} (X_t - \mathbb{E}[X_t|X_{\tau_{N_{t'}}}, \tau_{N_{t'}}])^2 \mathrm{d}t\right] \tag{A.51b}$$

$$+\mathbb{E}\left[\int_{t'}^{\tau_{N_{t'}+1}} (X_t - \mathbb{E}[X_t|X_{\tau_{N_{t'}}}, \tau_{N_{t'}}])^2 \mathrm{d}t\right] \tag{A.51c}$$

$$+\mathbb{E}\left[\sum_{i=N_{t'}+1}^{N_T} \int_{\tau_i}^{\tau_{i+1}} (X_t - \mathbb{E}[X_t|X_{\tau_i}, \tau_i])^2 \mathrm{d}t\right], \tag{A.51d}$$

where $\tau_{N_T+1} \triangleq T$.

Under the given sampling policy $\tau_1, \tau_2, \ldots$, we construct a sampling policy by inserting an extra deterministic sampling time t'. The resultant MSE is the same as (A.51) with (A.51c) replaced by

$$\mathbb{E}\left[\int_{t'}^{\tau_{N_{t'}+1}} (X_t - \mathbb{E}[X_t|X_{t'}])^2 \mathrm{d}t\right], \tag{A.52}$$

since a sample is taken at time t' under the constructed sampling policy. Since

$$\sigma(X_{\tau_{N_{t'}}}, \tau_{N_{t'}}) \subseteq \sigma(\mathcal{F}_{t'}) \tag{A.53a}$$

$$\mathbb{E}[X_t|\mathcal{F}_{t'}] = \mathbb{E}[X_t|X_{t'}], \tag{A.53b}$$

where (A.53b) is due to the strong Markov process (P.1) in Section 2.2, we conclude that (A.51c) $\geq$ (A.52).

Thus, by introducing extra sampling times, we can achieve the same or a lower MSE. We can express the difference between the frequency constraint F and the average sampling frequency under the given sampling policy as

$$FT - \mathbb{E}[N_T] = I + D, \tag{A.54}$$

where $I \in \mathbb{N}$ represents the non-negative integer part, and $D \in (0,1)$ represents the decimal part. By introducing I different deterministic sampling times, we can compensate the integer part I. By introducing a random sampling time stamp t with probability D to sample and probability $1-D$ not to sample, we can compensate the decimal part. Therefore, for any sampling policy whose average sampling frequency is strictly less than F, we can always construct a sampling policy that achieves the maximum sampling frequency F and leads to an MSE no worse than that achieved by the arbitrarily fixed sampling policy.

A.5 Proof of Corollary 1.3

We show that symmetric thresholds $\{a_i(r,s)\}_{r=s}^{T}$ in (A.26) must satisfy (2.11) for all $s \in \mathrm{Supp}(f_{\tau_i})$.

Due to (S.3), the probability on the right side of (A.26) is continuous in $t \in [s,T]$ for all $s \in \mathrm{Supp}(f_{\tau_i})$. Thus, for all $s \in \mathrm{Supp}(f_{\tau_i})$, $t \in [s,T)$,

$$\lim_{\delta \to 0^+} \mathbb{P}\left[\tilde{X}'_r \in (-a_i(r,s), a_i(r,s)), \forall r \in [s, t+\delta] \middle| \tau'_i = s\right] \tag{A.55a}$$

$$= \mathbb{P}\left[\tilde{X}'_r \in (-a_i(r,s), a_i(r,s)), \forall r \in [s,t] \middle| \tau'_i = s\right]. \tag{A.55b}$$

By the continuity of $\tilde{X}'_r$ in (P.3-b), (A.55) implies (2.11).

A.6 Proof of Theorem 2

First, we introduce Lemma 7, stated next, that will be helpful in proving (2.13). Second, we prove that symmetric threshold sampling policies (2.9) in Theorem 1 can be reduced to (2.12) in the setting of Theorem 2, i.e., under the assumption that $\{X_t\}_{t\geq 0}$ has time-homogeneous property in Definition 3 and $T = \infty$. Then, we show that Remark 1 holds and prove that (2.13) holds using Lemma 7.

Lemma 7. *(e.g., [110, Proposition 1(ii)]) Suppose that $Z_0, Z_1, \ldots$ are i.i.d. Let $Q_t \triangleq \sum_{i=0}^{\infty} \mathbb{1}_{[0,t]}\left(\sum_{k=0}^{i} Z_k\right)$. Let $R_0, R_1, \ldots$ be i.i.d rewards, and let $S_t \triangleq \sum_{i=0}^{Q_t} R_i$ be the renewal reward process. If $0 < \mathbb{E}[Z_i] < \infty$, $\mathbb{E}[|R_i|] < \infty$, then*

$$\lim_{T\to\infty} \frac{\mathbb{E}[S_T]}{T} = \frac{\mathbb{E}[R_0]}{\mathbb{E}[Z_0]}. \tag{A.56}$$

Since the stochastic process considered in Theorem 2 is infinitely long, we use the DFF in the infinite time horizon:

$$\underline{D}^{\infty}(F) = \inf_{\substack{\{\mathcal{P}_t\}_{t\geq 0}\in\Pi: \\ \text{(2.3b)}}} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T (\tilde{X}_t - \bar{\tilde{X}}_t)^2\right], \tag{A.57}$$

where Π is the set of all sampling-decision processes (A.1) of the form (2.9) satisfying (S.1) and (S.2) in Section 2.2 over the infinite time horizon. Note that for any stopping time τ, and for any $t \geq \tau$, we have

$$\{\tilde{X}_t\}_{t\geq\tau} \text{ and } \{\tilde{X}_{t-\tau}\}_{t-\tau\geq 0} \text{ have the same distribution,} \tag{A.58}$$

$$\{\tilde{X}_t\}_{t\geq\tau} \text{ is independent of } \{\tilde{X}_t\}_{t=0}^{\tau}, \tag{A.59}$$

where (A.58) is due to the time-homogeneity of $\{X_t\}_{t\geq 0}$ in Definition 3, and (A.59) is due to (P.3-a) in Section 2.2.

Using (A.58)–(A.59) and assumption (S.1), we will prove that the sampling-decision process that achieves the $\underline{D}^{\infty}(F)$ for time-homogeneous continuous Markov processes satisfying assumptions (P.1)–(P.3) is of the form (2.12).

Given an arbitrary sampling-decision process $\{\mathcal{P}_t\}_{t\geq 0}$ of the form (2.9), we define its MSRE (A.21) and REE (A.22) processes as

$$\begin{aligned} \tilde{X}_t &\triangleq \tilde{X}_t(\{\mathcal{P}_s\}_{s\geq 0}), \\ \bar{\tilde{X}}_t &\triangleq \bar{\tilde{X}}_t(\{\mathcal{P}_s\}_{s\geq 0}). \end{aligned} \tag{A.60}$$

Denote by $\tau_1, \tau_2, \ldots$ the stopping times of the causal sampling policy characterized by $\{\mathcal{P}_t\}_{t\geq 0}$. Assume that the sampling-decision process that achieves $\underline{D}^{\infty}(F)$ (A.57) is $\left\{\mathcal{P}_t^{(\mathrm{a})}\right\}_{t\geq 0}$. We have,

$$\underline{D}^{\infty}(F) \tag{A.61a}$$

$$= \inf_{\substack{\{\mathcal{P}_t\}_{t\geq 0}\in\Pi: \\ \mathcal{P}_t=\mathcal{P}_t^{(\mathrm{a})},t\leq\tau_i, \\ \text{(2.3b)}}} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_{\tau_i}^{T} (\tilde{X}_t - \bar{\tilde{X}}_t)^2 \mathrm{d}t\right] \tag{A.61b}$$

$$= \inf_{\substack{\{\mathcal{P}_t\}_{t\geq 0}\in\Pi: \\ \text{(2.3b)}}} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_{0}^{T-\tau_i} (\tilde{X}_t - \bar{\tilde{X}}_t)^2 \mathrm{d}t\right] \tag{A.61c}$$

$$= \underline{D}^{\infty}(F), \tag{A.61d}$$

where (A.61b) is due to assumption (S.1); (A.61c) is due to (A.58); the equality in (A.61d) is achieved since (A.61c) is upper-bounded by (A.61d) and is equal to (A.61a) simultaneously. Suppose that the sampling-decision processes that achieve (A.61b)–(A.61c) are $\{\mathcal{P}_t^{(\mathrm{b})}\}_{t\geq 0}$ and $\{\mathcal{P}_t^{(\mathrm{c})}\}_{t\geq 0}$, respectively. From (A.61a) and (A.61b), we observe that

$$\left\{\mathcal{P}_t^{(\mathrm{a})}\right\}_{t\geq\tau_i} = \left\{\mathcal{P}_t^{(\mathrm{b})}\right\}_{t\geq\tau_i}, \ i = 0, 1, \ldots \tag{A.62}$$

We prove that under sampling-decision processes satisfying assumption (S.1), it holds that

$$\mathbb{E}\left[\int_{T-\tau_i}^{T}(\tilde{X}_t-\bar{\tilde{X}}_t)^2\right]<\infty, \tag{A.63}$$

so that using (A.61c), (A.61d), and (A.63), we conclude

$$\left\{\mathcal{P}_t^{(\mathrm{c})}\right\}_{t\geq 0}=\left\{\mathcal{P}_t^{(\mathrm{a})}\right\}_{t\geq 0}. \tag{A.64}$$

By assumption (S.1) we know that there exist sampling-decision processes that lead to

$$\mathbb{E}\left[\int_{0}^{\tau_i}(\tilde{X}_t-\bar{\tilde{X}}_t)^2\mathrm{d}t\right]<\infty. \tag{A.65}$$

Thus, there exist sampling-decision processes such that (A.63) holds. Since the goal is to minimize the MSE, it suffices to consider sampling-decision processes that lead to (A.63).

Due to (A.58), the probability distributions of $\tilde{X}_t, t\in[0,T-\tau_i]$ in (A.61b) and $\tilde{X}_t, t\in[\tau_i,T]$ (A.61c) are the same. Thus, the sampling-decision process $\{\mathcal{P}_t\}_{t\geq\tau_i}=\left\{\mathcal{P}_{t-\tau_i}^{(\mathrm{a})}\right\}_{t-\tau_i\geq 0}$ achieves the infimum in (A.61b). We conclude

$$\left\{\mathcal{P}_t^{(\mathrm{b})}\right\}_{t\geq\tau_i}=\left\{\mathcal{P}_{t-\tau_i}^{(\mathrm{a})}\right\}_{t-\tau_i\geq 0},\ i=0,1,\ldots \tag{A.66}$$

Using (A.62) and (A.66), we conclude that $\left\{\mathcal{P}_{t-\tau_i}^{(\mathrm{a})}\right\}_{t-\tau_i\geq 0}=\left\{\mathcal{P}_t^{(\mathrm{a})}\right\}_{t\geq\tau_i}$, $i=0,1,\ldots$, i.e.,

$$a_0(s,0)=a_i(s+\tau_i,\tau_i). \tag{A.67}$$

Thus, (2.12) follows.

Next, we show Remark 1 using (2.12). We conclude that the sampling intervals $T_i\triangleq\tau_{i+1}-\tau_i$, $i=0,1,\ldots$, are independent due to (A.59) and the fact that the sampling-decision process (2.12) is independent of the process prior to the last stopping time; the sampling intervals T_i, $i=0,1,\ldots$, are identically distributed due to (A.58) and the fact that the sampling-decision process (2.12) only takes into account the time elapsed from the last sampling time $t-\tau_i$, $t\in[\tau_i,\tau_{i+1})$, $i=0,1,\ldots$

We proceed to show that the optimization problem associated with $\underline{D}^{\infty}(F)$ can be reduced to (2.13) by Lemma 7. The assumptions in Lemma 7 are satisfied with $Z_i\leftarrow T_i$, $R_i\leftarrow\int_{\tau_i}^{\tau_{i+1}}(X_t-\mathbb{E}[X_t|X_{\tau_i},\tau_i])^2\mathrm{d}t$. The sampling intervals $T_0,T_1,\ldots$

are i.i.d. due to Remark 1. The expectation of T_i is finite by assumption (S.1). The reward random variables R_i are i.i.d. due to (A.58)–(A.59) and Remark 1. Furthermore, the expectation of the reward is finite by assumption (S.1). Therefore, using (A.56), we simplify the DFF in (2.8) to (2.13).

A.7 Optimal sampling policy for the OU process

Using (2.10), we calculate that for $t \in [\tau_i, \tau_{i+1})$,

$$X_t - \bar{X}_t = O_{t-\tau_i} \triangleq \frac{\sigma}{\sqrt{2\theta}} e^{-\theta(t-\tau_i)} W_{e^{2\theta(t-\tau_i)}-1}. \tag{A.68}$$

Before we solve the optimization problem (2.13), we show several useful properties: By solving Dynkin's formula for $R_1(O^2_{\tau_1})$ and $R_2(O^2_{\tau_1})$ in (2.23)–(2.24), we obtain [9, Eq.(44)]:

$$\mathbb{E}\left[\int_0^{\tau_1} O_t^2 \mathrm{d}t\right] = \mathbb{E}\left[R_2(O^2_{\tau_1})\right], \tag{A.69a}$$

$$\mathbb{E}[\tau_1] = \mathbb{E}[R_1(O^2_{\tau_1})]. \tag{A.69b}$$

Two functions R_1 and R_2 are related as follows:

$$\mathbb{E}\left[R_2(O^2_{\tau_1})\right] = \frac{\sigma^2}{2\theta}\mathbb{E}[R_1(O^2_{\tau_1})] - \frac{1}{2\theta}\mathbb{E}[O^2_{\tau_1}], \tag{A.70a}$$

$$R_2(\mathbb{E}[O^2_{\tau_1}])] = \frac{\sigma^2}{2\theta} R_1(\mathbb{E}[O^2_{\tau_1}]) - \frac{1}{2\theta}\mathbb{E}[O^2_{\tau_1}]. \tag{A.70b}$$

We proceed to solve (2.13). We lower bound the objective function of (2.13) as

$$\frac{\mathbb{E}\left[\int_0^{\tau_1} O_t^2 \mathrm{d}t\right]}{\mathbb{E}[\tau_1]} \tag{A.71a}$$

$$= \frac{\mathbb{E}\left[R_2(O^2_{\tau_1})\right]}{\mathbb{E}[R_1(O^2_{\tau_1})]} \tag{A.71b}$$

$$= \frac{\frac{\sigma^2}{2\theta}\mathbb{E}[R_1(O^2_{\tau_1})] - \frac{1}{2\theta}\mathbb{E}[O^2_{\tau_1}]}{\mathbb{E}[R_1(O^2_{\tau_1})]} \tag{A.71c}$$

$$\geq \frac{\frac{\sigma^2}{2\theta}\frac{1}{F} - \frac{1}{2\theta} R_1^{-1}(\frac{1}{F})}{\frac{1}{F}} \tag{A.71d}$$

$$= F \cdot R_2\left(R_1^{-1}\left(\frac{1}{F}\right)\right), \tag{A.71e}$$

where (A.71a) is obtained by plugging (A.68) into (2.13); (A.71b) holds by plugging (A.69) into (A.71a); (A.71c) holds by plugging (A.70a) into (A.71b); (A.71d) holds since 1) the minimization constraint of (2.13) together with (A.69b) implies that

$\mathbb{E}[R_1(O^2_{\tau_1})] = \frac{1}{F}$, 2) R_1 is a convex function such that $R_1(\mathbb{E}[O^2_{\tau_1}]) \leq \mathbb{E}[R_1(O^2_{\tau_1})] \leq \frac{1}{F}$, 3) R_1 is an monotonically increasing function such that 2) implies $\mathbb{E}[O^2_{\tau_1}] \leq R_1^{-1}\left(\frac{1}{F}\right)$; (A.71e) holds by (A.70b).

Plugging (2.21) into (A.71b), we verify that the lower bound in (A.71e) is achieved by the symmetric threshold sampling policy in (2.21).

A.8 Proof of Proposition 1

We show that the objective function of $\underline{D}_{\text{ch}}(F)$ in (2.33) can be decomposed as (2.36). Plugging (2.34) into (2.33), we expand the objective function as

$$\frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i+\delta}^{\tau_{i+1}+\delta}(cW_{at}-cW_{a\tau_i})^2\mathrm{d}t\right] \tag{A.72a}$$

$$= \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i}^{\tau_{i+1}}(cW_{at}-cW_{a\tau_i})^2\mathrm{d}t\right] - \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i}^{\tau_i+\delta}(cW_{at}-cW_{a\tau_i})^2\mathrm{d}t\right]$$
$$+ \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_{t+1}}^{\tau_{i+1}+\delta}(cW_{at}-cW_{a\tau_i})^2\mathrm{d}t\right], \tag{A.72b}$$

where (A.72b) holds by splitting the integral in (A.72a). Replacing $cW_{at}-cW_{a\tau_i} \leftarrow cW_{at}-cW_{a\tau_{i+1}}+cW_{a\tau_{i+1}}-cW_{a\tau_i}$ in the last term of (A.72b), we conclude that the last term is equal to

$$\frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_{i+1}}^{\tau_{i+1}+\delta}(cW_{at}-cW_{a\tau_{i+1}})^2\mathrm{d}t\right] + \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\delta(cW_{a\tau_{i+1}}-cW_{a\tau_i})^2\right], \tag{A.73}$$

The time-homogeneity of the Wiener process implies that the second term in (A.72b) is equal to the first term in (A.73). Lemma 1 (b) implies that the second term in (A.73) is equal to $ac^2\delta$. Plugging (A.73) into (A.72b), we obtain (2.33).

A.9 Proof of Theorem 4

We first introduce Lemmas 8–10 that are useful for lower bounding the objective function of the DSTF in (7). Using Lemmas 8–10, we establish a lower (converse) bound on the DSTF (7). Finally, we show that under the causal sampling policy in Theorem 4, the DSTF coincides with its lower bound.

Tools

In this section, we present Lemmas 8–10. We denote by $T_i \triangleq \tau_{i+1}-\tau_i$ the i-th sampling interval, $i = 0, 1, 2\ldots, n-1$. We denote by $\mathcal{B}(k)$ the set that contains all

feedback sequences b^i of length $i = k, k+1, \ldots, n-1$ that have k unsuccessful transmissions after the last successful transmission:

$$\mathcal{B}(k) \triangleq \cup_{i=k}^{n-1} \left\{ b^i \in \{0,1\}^i \colon i - \max \mathcal{N}(b^i) = k \right\}, \tag{A.74}$$

$k = 0, 1, \ldots, n-1$. Thus, if $b^i \in \mathcal{B}(k)$, then the time of the last successful transmission (2.40) reduces to

$$S_i = \tau_{i-k}. \tag{A.75}$$

Lemma 8, stated next, shows a conditional expectation of a sampling interval.

Lemma 8. *Given any integers $i, \ell \geq k$ and any bit sequences $b^i, b^\ell \in \mathcal{B}(k)$, $k = 0, 1, \ldots, n-1$, it holds that*

$$\mathbb{E}[T_i | \{B_{\tau_j}\}_{j=1}^i = b^i] = \mathbb{E}[T_\ell | \{B_{\tau_j}\}_{j=1}^\ell = b^\ell]. \tag{A.76}$$

Proof. First, assumption (S.4) implies that the sampling interval T_i is determined by the innovation of the source process $X_t - \mathbb{E}[X_t | X_{\tau_i}, \tau_i]$ after time τ_i and the number of unsuccessful transmissions $i - \max \mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^i\right)$ after the last successful transmission. Second, the continuous Lévy process is time-homogeneous (Definition 3), meaning that the distribution of $X_t - \mathbb{E}[X_t | X_{\tau_i}, \tau_i]$ only depends on the time elapsed from the last sampling time. Third, the assumption $b^i, b^\ell \in \mathcal{B}(k)$ means that $i - \max \mathcal{N}(b^i) = \ell - \max \mathcal{N}(b^\ell) = k$. Therefore, the three observations imply (A.76). □

Lemma 9, stated next, shows the probability distribution of the channel feedback.

Lemma 9. *The pmf of the bit sequence $\{B_{\tau_j}\}_{j=1}^i$ is given by*

$$\mathbb{P}\left[\{B_{\tau_j}\}_{j=1}^i \in \mathcal{B}(0)\right] = (1-p)^{\mathbb{1}_{(0,n-1]}(i)} \tag{A.77}$$

$$\mathbb{P}\left[\{B_{\tau_j}\}_{j=1}^i \in \mathcal{B}(k)\right] = (1-p)^{\mathbb{1}_{(0,i)}(k)} p^k, \ k = 1, \ldots, i. \tag{A.78}$$

Proof. The pmf is obtained by the facts that 1) packet drops are i.i.d.; 2) packet drops are independent of the source process; 3) $B_0 \triangleq 1$ at $\tau_0 = 0$. □

Lemma 10, stated next, provides a lower bound on the MSE over one sampling interval.

Lemma 10. *Given feedback bits $b^i \in \mathcal{B}(k)$, it holds that*

$$\mathbb{E}\left[\int_{\tau_i}^{\tau_{i+1}} (W_t - W_{S_i})^2 \mathrm{d}t \middle| \{B_{\tau_j}\}_{j=1}^i = b^i\right] \tag{A.79a}$$

$$\geq \frac{1}{6}\mathbb{E}[T_i|\{B_{\tau_j}\}_{j=1}^i = b^i]^2 + \mathbb{E}[T_i|\{B_{\tau_j}\}_{j=1}^i = b^i] \sum_{j=1}^{k} \mathbb{E}[T_{i-j}|\{B_{\tau_j}\}_{j=1}^{i-j} = b^{i-j}]. \tag{A.79b}$$

Proof. Plugging (A.75) into (A.79a), we obtain

$$\mathbb{E}\left[\int_{\tau_i}^{\tau_{i+1}} (W_t - W_{\tau_{i-k}})^2 \mathrm{d}t \middle| \{B_{\tau_j}\}_{j=1}^i = b^i\right] \tag{A.80a}$$

$$= \mathbb{E}\Big[\int_{\tau_i}^{\tau_{i+1}} (W_t - W_{\tau_i})^2 + (W_{\tau_i} - W_{\tau_{i-k}})^2$$
$$+ 2(W_t - W_{\tau_i})(W_{\tau_i} - W_{\tau_{i-k}})\mathrm{d}t | \{B_{\tau_j}\}_{j=1}^i = b^i\Big] \tag{A.80b}$$

$$= \mathbb{E}\left[\int_{\tau_i}^{\tau_{i+1}} (W_t - W_{\tau_i})^2 \mathrm{d}t \middle| \{B_{\tau_j}\}_{j=1}^i = b^i\right]$$
$$+ \mathbb{E}[T_i|\{B_{\tau_j}\}_{j=1}^i = b^i]\mathbb{E}\left[(W_{\tau_i} - W_{\tau_{i-k}})^2 \middle| \{B_{\tau_j}\}_{j=1}^i = b^i\right] \tag{A.80c}$$

$$= \frac{1}{6}\mathbb{E}[W_{T_i}^4|\{B_{\tau_j}\}_{j=1}^i = b^i] + \mathbb{E}[T_i|\{B_{\tau_j}\}_{j=1}^i = b^i] \sum_{j=1}^{k} \mathbb{E}[W_{T_{i-j}}^2|\{B_{\tau_j}\}_{j=1}^i = b^i] \tag{A.80d}$$

$$\geq \frac{1}{6}\mathbb{E}[T_i|\{B_{\tau_j}\}_{j=1}^i = b^i]^2 + \mathbb{E}[T_i|\{B_{\tau_j}\}_{j=1}^i = b^i] \sum_{j=1}^{k} \mathbb{E}[T_{i-j}|\{B_{\tau_j}\}_{j=1}^{i-j} = b^{i-j}], \tag{A.80e}$$

where (A.80b) holds by replacing $W_t - W_{\tau_{i-k}} \leftarrow W_t - W_{\tau_i} + W_{\tau_i} - W_{\tau_{i-k}}$ in (A.80a) and rearranging terms; the second term in (A.80b) is equal to the second term in (A.80c) since assumption (S.4) implies that given the past feedback bits, the sampling interval T_i is independent of $W_{\tau_i} - W_{\tau_{i-k}}$; the third term in (A.80b) is zero by the orthogonal principle of the MMSE estimator [111, Prop. V.C.2]; the first term in (A.80d) holds by applying Lemma 1 (c) to the first term of (A.80c); the second term of (A.80d) holds by applying Lemma 1 (d) to expand the last conditional expectation in (A.80c); the first term in (A.80e) holds by first applying Jensen's inequality to lower bound the first term in (A.80d) by $\mathbb{E}[W_{T_i}^2|\{B_{\tau_j}\}_{j=1}^i = b^i]^2$ and then applying Lemma 1 (b); the second term in (A.80e) holds by applying Lemma 1 (b) to the summands in the second term of (A.80d), and using the fact that the sampling interval T_{i-j} is independent of the future feedback bits $\{B_{\tau_j}\}_{j=i-j+1}^i$. □

Converse

In this section, we lower bound the DSTF (7) by a minimization problem and show that the minimum is achieved by the causal sampling policy in Theorem 4. We denote by t_k the expected sampling time in (A.76) conditioned on a sequence of feedback bits in $\mathcal{B}(k)$.

The objective function of the DSTF $\underline{D}^{\mathrm{pd}}(n,T)$ is lower bounded as

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{n-1}\int_{\tau_i}^{\tau_{i+1}}(X_t-\bar{X}_t^{\mathrm{pd}})^2\mathrm{d}t\right]\\
&=\frac{ac^2}{T}\mathbb{E}\left[\sum_{i=0}^{n-1}\int_{\tau_i}^{\tau_{i+1}}(W_t-W_{S_i})^2\mathrm{d}t\right] \qquad \text{(A.81)}\\
&\geq\frac{ac^2}{T}\sum_{i=0}^{n-1}\left(\frac{1}{6}t_0^2(1-p)^{\mathbb{1}_{(0,n-1]}(i)}+\sum_{k=1}^{i}\left(\frac{1}{6}t_k^2+t_k\sum_{j=1}^{k}t_{k-j}\right)(1-p)^{\mathbb{1}_{(0,i)}(k)}p^k\right), \qquad \text{(A.82)}
\end{aligned}$$

where (A.81) holds by plugging the continuous Lévy process (2.17) and the estimate (2.45) into the MSE (2.44); (A.82) holds by plugging Lemmas 8–10 into (A.81). Rewriting the lower bound in (A.82) in a matrix form and minimizing it under the time constraint (2.43), we obtain a lower bound on the DSTF as

$$\underline{D}^{\mathrm{pd}}(n,T)\geq\frac{ac^2}{T}\min_{\substack{\mathbf{t}=[t_0,t_1,\ldots,t_{n-1}]^T\in\mathbb{R}_+^n:\\ d(p)^T\mathbf{t}=T}}\mathbf{t}^TA(p)\mathbf{t}, \qquad \text{(A.83)}$$

where

$$c_i\triangleq p^i(1+(1-p)(n-i-1)), \qquad \text{(A.84)}$$

$$A(p)\triangleq\begin{bmatrix}\frac{1}{6}c_0 & & & & \\ c_1 & \frac{1}{6}c_1 & & \mathbf{0} & \\ c_2 & c_2 & \frac{1}{6}c_2 & & \\ \vdots & \vdots & & \ddots & \\ c_{n-1} & c_{n-1} & c_{n-1} & \cdots & \frac{1}{6}c_{n-1}\end{bmatrix}. \qquad \text{(A.85)}$$

$$d(p)^T\triangleq[c_0\ c_1\ \cdots\ c_{n-1}]. \qquad \text{(A.86)}$$

We proceed to show that the minimum on the right side of (A.83) is achieved by the causal sampling policy in Theorem 4. Fix any packet-drop probability $p<\frac{1}{5}$, fix t_k

for all $k \neq i, j$, and let $C \triangleq T - \sum_{k \neq i,j} c_k t_k$. Without loss of generality, we assume $i < j$. We find t_i, t_j that minimize (A.83) by solving

$$\min_{t_i, t_j \in \mathbb{R}_+ :\, c_i t_i + c_j t_j = C} \frac{1}{6} c_i t_i^2 + \frac{1}{6} c_j t_j^2 + c_j t_i t_j + l_i t_i + l_j t_j, \tag{A.87}$$

where the coefficients l_i and l_j are given by

$$l_i \triangleq \sum_{k=0}^{i-1} c_i t_k + \sum_{k=i+1}^{j-1} c_k t_k + \sum_{k=j+1}^{n-1} c_k t_k \tag{A.88}$$

$$l_j \triangleq \sum_{k=0}^{i-1} c_j t + k + \sum_{k=i+1}^{j-1} c_j t_k + \sum_{k=j+1}^{n-1} c_k t_k. \tag{A.89}$$

Let $x \triangleq c_i t_i$, $x \in [0, C]$, we write the objective function in (A.87) as a function of x as

$$f(x) \triangleq \frac{x^2}{6c_i} + \frac{(C-x)^2}{6c_j} + \frac{x(C-x)}{c_i} + l_i \frac{x}{c_i} + l_j \frac{(C-x)}{c_j}. \tag{A.90}$$

The quadratic function $f(x)$ is minimized at

$$x^* = \frac{(C + 3l_i)c_i - 3(C + l_i)c_j}{c_i - 5c_j}, \tag{A.91}$$

which lies in $[C, \infty]$ since

$$C - x^* = \frac{-2Cc_j - 3(l_j c_i - l_i c_j)}{c_i - 5c_j} \tag{A.92a}$$

$$= \frac{-2Cc_j - 3(c_j \sum_{k=i+1}^{j-1} (c_i - c_k) t_k + (c_i - c_j) \sum_{k=j+1}^{n} c_k t_k)}{c_i - 5c_j} \tag{A.92b}$$

$$\leq 0, \tag{A.92c}$$

where (A.92b) holds by plugging (A.88)–(A.89) into (A.92a); (A.92c) holds since 1) $c_k \geq c_{k'}$ for any $k \leq k'$ implies that the numerator of (A.92b) is negative, and 2) $p < \frac{1}{5}$ implies that the denominator of (A.92b) is positive. Therefore, the minimum of $f(x)$ for $x \in [0, C]$ is attained at $x = C$ and the corresponding t_i, t_j are

$$t_i = \frac{C}{c_i} \tag{A.93a}$$

$$t_j = 0, \; i < j. \tag{A.93b}$$

Using (A.93), we proceed to show that the minimum on the right side of (A.83) is achieved at

$$t_0 = \frac{T}{c_0} = \frac{T}{1 + (1-p)(n-1)} \tag{A.94a}$$

$$t_1 = \cdots = t_{n-1} = 0 \tag{A.94b}$$

using the algorithm below. We start with an arbitrary vector $\mathbf{t} \in \mathbb{R}_+^n$ that satisfies the constraint in (A.83), and we implement:

1. Initialize $k \leftarrow n-1$;
2. If $t_k \neq 0$, then let $t_{k-1} \leftarrow t_{k-1} + \frac{c_k t_k}{c_{k-1}}$, $t_k \leftarrow 0$;
3. If $k > 0$, let $k \leftarrow k-1$, and go back to step 2; Otherwise, exit the loop.

The minimum of (A.83) is achieved at (A.94) since step 2 yields a new vector $\mathbf{t}$ that satisfies the time constraint in (A.83) and leads to an MSE no larger than the MSE before step 2, and the output of this program is always (A.94) regardless of the starting vector.

From the definition of t_k (A.76) and Lemma 1 (b), we conclude that the causal sampling policy in Theorem 4 satisfies (A.94) and thus achieves the lower bound (A.83) on the DSTF.

Achievability

Plugging the causal sampling policy in Theorem 4 into the objective function of the DSTF (A.81), we conclude that the DSTF is equal to its lower bound (A.83). Therefore, the causal sampling policy in Theorem 4 achieves the DSTF.

Appendix B

CAUSAL RATE-CONSTRAINED SAMPLING: PROOFS

B.1 Proof of Theorem 5

We show the converse (3.8). Denote by Π_T the set of all sampling-decision processes (A.1) that satisfy (S.1)–(S.3) on $[0,T]$. Denote by $\mathcal{C}_T$ the set of all causal compressing policies on $[0,T]$. We lower bound the DRF in (3.5) as

$$D(R) = \inf_{\substack{\{\mathcal{P}_t\}_{t=0}^T \in \Pi_T,\\ \{f_t\}_{t=0}^T \in \mathcal{C}_T:\\ \text{(3.3a)}}} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i}^{\tau_{i+1}} (X_t - \mathbb{E}[X_t|U^i,\tau^i,t<\tau_{i+1}])^2 \mathrm{d}t\right] \tag{B.1a}$$

$$\geq \inf_{\substack{\{\mathcal{P}_t\}_{t=0}^T \in \Pi_T:\\ \frac{\mathbb{E}[N]}{T}\leq R}} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i}^{\tau_{i+1}} (X_t - \mathbb{E}[X_t|\{X_s\}_{s=0}^{\tau_i},\tau^i,t<\tau_{i+1}])^2 \mathrm{d}t\right] \tag{B.1b}$$

$$= \inf_{\substack{\{\mathcal{P}_t\}_{t=0}^T \in \Pi_T:\\ \frac{\mathbb{E}[N]}{T}\leq R}} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{N}\int_{\tau_i}^{\tau_{i+1}} (\tilde{X}_t - \mathbb{E}[\tilde{X}_t|\tau_i,t<\tau_{i+1}])^2 \mathrm{d}t\right] \tag{B.1c}$$

$$= \underline{D}(R), \tag{B.1d}$$

where (B.1b) holds since $\mathbb{E}[N] \leq \mathbb{E}\left[\sum_{i=1}^{N}\ell(U_i)\right]$, and U^i belongs to the σ-algebra generated by the stochastic process $\{X_s\}_{s=0}^{\tau_i}$. The equality in (B.1c) is obtained by subtracting and adding $\mathbb{E}[X_t|X_{\tau_i},\tau_i]$ to X_t in (B.1b), where

$$\begin{aligned}&\mathbb{E}[\tilde{X}_t|\tau_i,t<\tau_{i+1}]\\ =\;&\mathbb{E}[X_t|\{X_s\}_{s=0}^{\tau_i},\tau^i,t<\tau_{i+1}] - \mathbb{E}[X_t|X_{\tau_i},\tau_i]\end{aligned} \tag{B.2}$$

holds due to the argument that justifies (A.22d) with $\{X_{\tau_j}\}_{j=1}^{i} \leftarrow \{X_s\}_{s=0}^{\tau_i}$.

While (B.1) shows that the converse (3.8) holds for the finite horizon ($T<\infty$), the converse also holds for the infinite horizon ($T=\infty$). This is because (B.1) continues to hold with the minimization constraints $\{\mathcal{P}_t\}_{t=0}^T \in \Pi_T$, $\{f_t\}_{t=0}^T \in \mathcal{C}_T$, (3.3a), and $\frac{\mathbb{E}[N]}{T} \leq R$ replaced by $\{\mathcal{P}_t\}_{t\geq 0} \in \Pi_\infty$, $\{f_t\}_{t\geq 0} \in \mathcal{C}_\infty$, (3.3b), and $\limsup_{T\to\infty}\frac{\mathbb{E}[N]}{T} \leq R$, respectively, and with $\limsup_{T\to\infty}$ inserted right before the objective functions in (B.1a)–(B.1c).

B.2 Recovering L_t from Z_t

The two formulations (3.33) and (3.37) are related as follows. Any state evolution described by (3.37) can be written in the form of (3.33) by setting Z_t as in (3.38). Conversely, a state evolution described by (3.33) can be written as (3.37) if and only if Z_t, when viewed as a function of t, is almost surely *generalized absolutely continuous in the restricted sense* (ACG_*) between any consecutive discontinuous points of $\{Z_t\}_{t=0}^T$ [112][113]. This is because control signal L_s in (3.38) is well-defined if and only if Z_t^* satisfies the ACG_* property. The function $f\colon [a,b] \to \mathbb{R}$ is said to be ACG_* [112][113] over set $\mathcal{E} \subset [a,b]$ if f is continuous, and $\mathcal{E}$ is a countable union of sets $\mathcal{E}_n$ on each of which f satisfies the following: for each $\epsilon > 0$, there exists $\delta > 0$ such that $\sum_{i=1}^k \sup_{x,y\in[x_i,y_i]} |F(x) - F(y)| < \epsilon$ for all finite sets of disjoint open intervals $\{(x_i, y_i)\}_{i=1}^k$ with endpoints in $\mathcal{E}_n$ and $\sum_{i=1}^k |x_i - y_i| < \delta$. For example, for stochastic processes of the form $X_t = g_1(t)W_{g_2(t)} + g_3(t)$, the optimal control signal $\{Z_t\}_{t=0}^T$ (3.35) almost surely satisfies the ACG_* property. Here, $g_1(\cdot), g_3(\cdot)$ are continuous and differentiable except perhaps on a countable set, $g_2(\cdot)$ is continuous, positive, and non-decreasing, and $\{W_t\}_{t=0}^T$ is the Wiener process.

We show how to recover L_t (3.37) from Z_t (3.33), provided that $\{Z_t\}_{t=0}^T$ satisfies the ACG_* property. Denote by $\delta(\cdot)$ the Dirac-delta function. Let ν_i be the i-th discontinuous point of $\{Z_t\}_{t=0}^T$. For $\{Z_t\}_{t=0}^T$ in (3.35), ν_i is simply equal to the sampling times τ_i, $i = 1, 2, \ldots$ Without loss of generality, we assume that $\{Z_t\}_{t=0}^T$ is right-continuous at the discontinuous point ν_i, since the mean-square cost in (3.34) is not affected by the assumption. Denote by ν_i^- the time just before time ν_i, where $Z_{\nu_i^-} \neq Z_{\nu_i}$.

Proposition 6. *Assume that $\{Z_t\}_{t=0}^T$ is almost surely ACG_* on $[\nu_i, \nu_{i+1})$ and is right-continuous at the discontinuous point ν_i. Then, control signal $\{L_t\}_{t=0}^T$ in (3.37) for $t \in [\nu_i, \nu_{i+1})$, $i = 1, 2, \ldots$, is given by*

$$L_t = \begin{cases} \left(Z_{\nu_i} - Z_{\nu_i^-}\right)\delta(t - \nu_i), & t = \nu_i, \\ \lim_{\delta\to 0^+} \frac{Z_t - Z_{t-\delta}}{\delta}, & t \in (\nu_i, \nu_{i+1}). \end{cases} \tag{B.3}$$

Proof. For $t \in [\nu_i, \nu_{i+1})$, we rewrite (3.38) as

$$\int_{\nu_i}^{t} L_s \mathrm{d}s = Z_t - \lim_{\delta \to 0^+} \int_0^{\nu_i - \delta} L_s \mathrm{d}s \tag{B.4a}$$

$$= Z_t - Z_{\nu_i^-} \tag{B.4b}$$

$$= (Z_t - Z_{\nu_i}) + (Z_{\nu_i} - Z_{\nu_i^-}), \tag{B.4c}$$

which is equivalent to (B.3). □

Note that L_{ν_i} is an impulse control at $t = \nu_i$ [10, 114, 115, 116, 117], and L_t, $t \in (\nu_i, \nu_{i+1})$ is equal to the left-derivative of Z_t. This is because Z_t may not be differentiable at t, but its left-derivative exists since the ACG_* property of Z_t implies that it is differentiable almost everywhere on (ν_i, ν_{i+1}) [112]. For example, if $X_t = W_t$, the optimal control signal (3.35) is $Z_t = -W_{\tau_i}$, $t \in [\tau_i, \tau_{i+1})$, and the corresponding control signal in (3.37) is $L_t = -(W_{\tau_i} - W_{\tau_{i-1}})\delta(t - \tau_i)$ for $t \in [\tau_i, \tau_{i+1})$.

B.3 Decomposition of $D_{\mathrm{DET}}^{\mathrm{op}}(R)$

We show that $D_{\mathrm{DET}}^{\mathrm{op}}(R)$ (3.22) can be decomposed as (3.23). We write $D_{\mathrm{DET}}^{\mathrm{op}}(R)$ as follows.

$$D_{\mathrm{DET}}^{\mathrm{op}}(R) = \limsup_{T\to\infty} \inf_{\substack{\pi_T \in \Pi_T^{\mathrm{DET}} \\ \{f_t\}_{t=0}^T \in \mathcal{C}_T : \\ \text{(3.3a)}}} \frac{1}{T} \mathbb{E}\left[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (X_t - \hat{X}_t)^2 \right] \tag{B.5a}$$

$$= \limsup_{T\to\infty} \inf_{\substack{\pi_T \in \Pi_T^{\mathrm{DET}} \\ \{f_t\}_{t=0}^T \in \mathcal{C}_T : \\ \text{(3.3a)}}} \frac{1}{T} \mathbb{E}\left[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (X_t - \bar{X}_t)^2 + (\bar{X}_t - \hat{X}_t)^2 \mathrm{d}t \right]$$

$$+ 2\mathbb{E}\left[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (X_t - \bar{X}_t)(\bar{X}_t - \hat{X}_t) \mathrm{d}t \right] \tag{B.5b}$$

$$= \limsup_{T\to\infty} \inf_{\substack{\pi_T \in \Pi_T^{\mathrm{DET}} \\ \{f_t\}_{t=0}^T \in \mathcal{C}_T : \\ \text{(3.3a)}}} \frac{1}{T} \mathbb{E}\left[\sum_{i=0}^{N} \int_{\tau_i}^{\tau_{i+1}} (X_t - \bar{X}_t)^2 + (\bar{X}_t - \hat{X}_t)^2 \mathrm{d}t \right], \tag{B.5c}$$

where (B.5a) holds by definition; (B.5b) holds by applying $X_t - \hat{X}_t \leftarrow X_t - \bar{X}_t + \bar{X}_t - \hat{X}_t$ to (B.5a); (B.5c) holds since the last term in (B.5b) is zero by the orthogonality principle of the MMSE estimator [111, Prop. V.C.2]. Since the causal compressing policy only influences the second term in (B.5c), we move the

minimization over the compressing policies $\mathcal{C}_T$ directly in front of the second term in (B.5c) to obtain (3.23).

B.4 Proof of Theorem 6

To obtain the IDRF $D_{\mathrm{DET}}(R)$ (3.24) and the optimal deterministic sampling policy that achieves it, we first compute the IDFRF (3.26). We denote by $T^N \triangleq \{T_0, T_1, \ldots, T_N\}$ a vector of sampling intervals that determines a deterministic sampling policy, where

$$\begin{aligned} T_i &= \tau_{i+1} - \tau_i, \ i = 0, \ldots, N-1 \\ T_N &= T - \tau_N. \end{aligned} \tag{B.6}$$

We denote by $D^N \triangleq \{D_1, \ldots, D_N\}$ a vector of quantization distortions, where $D_i \triangleq \mathbb{E}[(W_{\tau_i} - \hat{W}_{\tau_i})^2]$. Since the samples taken under a deterministic sampling policy form a GM process, Lemma 11, stated next, expands $D_{\mathrm{DET}}(F, R_s)$ by building on existing results on the causal IDRF (3.26b) of discrete-time GM processes.

Lemma 11. *The IDFRF under deterministic sampling policies can be written as*

$$D_{\mathrm{DET}}(F, R_s) = \limsup_{N \to \infty} D_N(F, R_s), \tag{B.7a}$$

$$D_N(F, R_s) = \inf_{\substack{T^N \geq 0: \\ \text{(B.8)}}} \frac{F}{N} \left(\sum_{i=0}^{N} \frac{T_i^2}{2} + \min_{\substack{D^N \geq 0: \\ \text{(B.9)}}} \sum_{i=1}^{N} T_i D_i \right), \tag{B.7b}$$

where the minimization constraints in (B.7) *are*

$$\frac{1}{N} \sum_{i=0}^{N} T_i = \frac{1}{F}, \tag{B.8}$$

$$z\left(D^N\right) \triangleq \frac{1}{N} \left(\sum_{i=1}^{N-1} \log_2 \left(1 + \frac{T_i}{D_i} \right) + \log_2 \left(\frac{T_0}{D_N} \right) \right) \leq 2R_s, \tag{B.9a}$$

$$D_{i-1} + T_{i-1} \geq D_i, i = 1, \ldots, N, \ D_0 = 0. \tag{B.9b}$$

Proof. Appendix B.5. □

Lemma 12, stated next, provides a lower bound on the right side of (B.7a).

Lemma 12. *$D_N(F,R_s)$ in (B.7b) is lower-bounded as*

$$D_N(F,R_s) \geq \underline{D}_N(F,R_s), \tag{B.10a}$$

$$\triangleq \inf_{\substack{T_0\geq 0, T_N\geq 0\\ T_0+T_N\leq \frac{N}{F}}} \frac{F}{2}\left(\frac{T_0^2+T_N^2+2\log_2 e\,\lambda^*(F,R_s,N)}{N}+\right.$$

$$\left.\frac{N-1}{N}T^*(F,N)\sqrt{T^*(F,N)^2+4\log_2 e\,\lambda^*(F,R_s,N)}\right), \tag{B.10b}$$

where $T^*(F,N)$ is given by

$$T^*(F,N) \triangleq \frac{N}{F(N-1)} - \frac{T_0+T_N}{N-1}, \tag{B.11}$$

and $\lambda^*(F,R_s,N)\geq 0$ is the unique solution to

$$z\left(D^{N*}\right) = 2R_s \tag{B.12}$$

with D^N in (B.9a) replaced by

$$D_i^* = \frac{-T_i+\sqrt{T_i^2+4\log_2 e\,\lambda^*(F,R_s,N)}}{2},\ i=1,\ldots,N-1, \tag{B.13a}$$

$$D_N^* = \frac{\lambda^*(F,R_s,N)\log_2 e}{T_N}, \tag{B.13b}$$

and T_i, $i=1,\ldots,N-1$ in (B.9a) replaced by $T^*(F,N)$ (B.11).

Proof. Appendix B.6. □

Lemma 13 provides an upper bound on the right side of (B.7a).

Lemma 13. *$D_N(F,R_s)$ in (B.7b) is upper-bounded as*

$$D_N(F,R_s) \leq \bar{D}_N(F,R_s) \tag{B.14a}$$

$$\triangleq \frac{N}{F(N+1)^2} + \frac{\log_2 e\,\lambda^*(F,R_s,N)F}{N}+$$

$$\frac{N-1}{2(N+1)}\sqrt{\left(\frac{N}{F(N+1)}\right)^2+4\log_2 e\,\lambda^*(F,R_s,N)}, \tag{B.14b}$$

where $\lambda^*(F,R_s,N)\geq 0$ is the unique solution to (B.12) with D^N in (B.9a) replaced by (B.13), and with T^N in (B.9a) replaced by

$$T_0 = T_1 = \cdots = T_N = \frac{N}{F(N+1)}. \tag{B.15}$$

Proof. Appendix B.7. □

Using Lemmas 12 and 13, we obtain the IDFRF $D_{\mathrm{DET}}(F, R_s)$.

Lemma 14. *$D_{\mathrm{DET}}(F, R_s)$ in (B.7a) is given by (3.31), where (3.31) can be achieved by a uniform sampling policy with sampling intervals equal to $T_i = \frac{1}{F},\ i = 0, 1, \ldots$*

Proof. Appendix B.8. □

We proceed to show the IDRF $D_{\mathrm{DET}}(R)$ using the IDFRF $D_{\mathrm{DET}}(F, R_s)$ in (3.31). Lemma 15, stated next, displays the relation between $D_{\mathrm{DET}}(R)$ and $D_{\mathrm{DET}}(F, R_s)$.

Lemma 15. *The IDRF (3.26) under deterministic sampling policies satisfies (3.28).*

Proof. The IDRF $D_{\mathrm{DET}}(R)$ is related to the IDFRF $D_{\mathrm{DET}}(F, R_s)$ in (B.7) as follows,

$$D_{\mathrm{DET}}(R) = \limsup_{N\to\infty} \inf_{\substack{F>0, R_s\geq 1:\\ FR_s\leq R}} D_N(F, R_s), \tag{B.16}$$

which does not directly imply (3.28), since the right side of (3.28) switches the order of $\limsup$ and $\inf$ in (B.16). In Appendix B.9, we show that $\limsup$ and $\inf$ in (B.16) are interchangeable. □

B.5 Proof of Lemma 11

We denote by $\tilde{D}_N(R_s)$ (3.26b) the IDRF for discrete-time samples of the Wiener process

$$W_{\tau_{i+1}} = W_{\tau_i} + V_{\tau_i},\ V_{\tau_i} \sim \mathcal{N}(0, T_i). \tag{B.17}$$

Using the representation of its dual in [20, Eq. (18)] derived using a semi-definite programming approach, we write

$$\tilde{D}_N(R_s) = \inf_{\substack{D_i\geq 0,\ i=1,\ldots,N:\\ D_{i-1}+T_{i-1}\geq D_i,\quad i=1,2,\cdots,N,\\ \frac{1}{N}\left(\sum_{i=1}^{N}\frac{1}{2}\log_2(D_{i-1}+T_{i-1})-\frac{1}{2}\log_2 D_i\right)\leq R_s.}} \sum_{i=1}^{N} T_i D_i. \tag{B.18}$$

Since the sampling intervals T^N are deterministic, we calculate the summand in (3.26a) as

$$\mathbb{E}\left[\int_{\tau_i}^{\tau_{i+1}} (W_t - W_{\tau_i})^2 \mathrm{d}t\right] = \mathbb{E}\left[\int_0^{T_i} W_t^2 \mathrm{d}t\right] = \frac{T_i^2}{2}. \tag{B.19}$$

Plugging (B.18) and (B.19) into (3.26), we write

$$D_{\mathrm{DET}}(F,R_s)=\limsup_{T\to\infty}\inf_{\substack{\pi_T\in\Pi_T^{\mathrm{DET}}\\ \text{(2.3a)}}}\frac{1}{T}\left(\sum_{i=0}^{N}\frac{T_i^2}{2}+\tilde{D}_N(R_s)\right). \tag{B.20}$$

Note that as $T\to\infty$, the number of samples N must increase no slower than $\sqrt{T}$. Since the largest sampling interval satisfies $\max_{i=0,\dots,N}T_i\geq\frac{T}{N+1}$, the summand in (B.20) $\frac{\max_i T_i^2}{2T}\geq\frac{T}{2(N+1)^2}$ will blow up to infinity if N increases slower than $\sqrt{T}$. Thus, we can replace the $\limsup_{T\to\infty}$ in (B.20) by $\limsup_{N\to\infty}$, replace T in (B.20) by its equivalent $\frac{F}{N}$ in (3.3a), and replace the minimization constraint (3.3a) in (B.20) by its equivalent (B.8).

B.6 Proof of Lemma 12

We split $D_N(F,R_s)$ into layered optimization problems:

$$D_N(F,R_s)\triangleq\inf_{\substack{T_0\geq0,T_N\geq0:\\ T_0+T_N\leq\frac{N}{F}}}D_N(F,R_s,T_0,T_N), \tag{B.21a}$$

$$D_N(F,R_s,T_0,T_N)\triangleq\min_{\substack{T_1,\dots,T_{N-1}\geq0:\\ \frac{1}{N}\sum_{i=1}^{N-1}T_i=\frac{1}{F}-\frac{T_0+T_N}{N}}}\frac{F}{N}\left(\sum_{i=0}^{N}\frac{T_i^2}{2}+D_N\left(F,R_s,T^N\right)\right), \tag{B.21b}$$

$$D_N\left(F,R_s,T^N\right)\triangleq\min_{\substack{D^N\geq0:\\ \text{(B.9)}}}\sum_{i=1}^{N}T_iD_i. \tag{B.21c}$$

We denote by $\underline{D}_N(F,R_s,T^N)$ the lower bound to $D_N(F,R_s,T^N)$ obtained by deleting the minimization constraint (B.9b) in (B.21c), i.e.,

$$\underline{D}_N(F,R_s,T^N)\triangleq\min_{\substack{D^N\geq0:\\ \text{(B.9a)}}}\sum_{i=1}^{N}T_iD_i, \tag{B.22}$$

We denote by $\underline{D}_N(F,R_s,T_0,T_N)$ the corresponding lower bound to $D_N(F,R_s,T_0,T_N)$ in (B.21b):

$$\underline{D}_N(F,R_s,T_0,T_N)\triangleq\min_{\substack{T_1,\dots,T_{N-1}\geq0:\\ \frac{1}{N}\sum_{i=1}^{N-1}T_i=\frac{1}{F}-\frac{T_0+T_N}{N}}}\frac{F}{N}\left(\sum_{i=0}^{N}\frac{T_i^2}{2}+\underline{D}_N\left(F,R_s,T^N\right)\right). \tag{B.23}$$

We calculate the corresponding lower bound to $D_N(F,R_s)$:

$$\underline{D}_N(F,R_s)\triangleq\min_{\substack{T_0\geq0,T_N\geq0:\\ T_0+T_N\leq\frac{N}{F}}}\underline{D}_N(F,R_s,T_0,T_N). \tag{B.24}$$

We first show that the optimization problem in the right side of (B.22) is a convex optimization problem that satisfies Slater's condition, i.e., strong duality holds. Then, we solve its Lagrangian dual problem to get the optimal $D_1^* \ldots, D_N^*$ in (B.13) that achieve the minimum in the right side of (B.22), where $\lambda^*(F, R_s, N) \geq 0$ is the unique solution to (B.12).

The objective function $\sum_{i=1}^{N} T_i D_i$ (B.22) is an affine function in D^N. Furthermore, $z\left(D^N\right)$ is a convex function since

$$\frac{\partial^2 z\left(D^N\right)}{\partial D_i^2} = \frac{\log_2 e T_i (2D_i + T_i)}{N(D_i^2 + D_i T_i)^2} \geq 0, \ \forall i = 1, \ldots, N-1, \tag{B.25a}$$

$$\frac{\partial^2 z\left(D^N\right)}{\partial D_N^2} = \frac{\log_2 e}{N D_N^2} \geq 0, \tag{B.25b}$$

$$\frac{\partial^2 z\left(D^N\right)}{\partial D_i \partial D_j} = 0, \ \forall i, j = 1, \ldots, N. \tag{B.25c}$$

Therefore, the minimization problem in the right side of (B.22) is convex. Notice that $z(D, D, \ldots, D)$ decreases from $+\infty$ to $-\infty$ as D increases from 0 to ∞. Thus, there exists a $\tilde{D} \geq 0$ such that Slater's condition is satisfied, i.e.,

$$z\left(\tilde{D}, \tilde{D}, \ldots, \tilde{D}\right) < 2R_s. \tag{B.26}$$

We conclude that 1) the strong duality holds, 2) $\underline{D}(F, R_s, T^N)$ can be obtained via its Lagrangian dual problem, and 3) there must exist an optimal Lagrangian multiplier $\lambda^*(F, R_s, N) \geq 0$ that satisfies the complementary slackness (B.12) in the KKT conditions. Indeed, (B.12) always has a non-negative solution $\lambda^*(F, R_s, N)$, since as a function of $\lambda^*(F, R_s, N)$, $z\left(D^{N*}\right)$ is continuous and monotonically decreasing from $+\infty$ to $-\infty$ as $\lambda^*(F, R_s, N)$ increases from 0 to $+\infty$.

D^{N*} in (B.13) is the solution to the Lagrangian function:

$$\min_{D^N} \sum_{i=1}^{N} T_i D_i + \lambda^*(F, R_s, N) \left(\sum_{i=1}^{N-1} \log_2 \left(1 + \frac{T_i}{D_i}\right) + \log_2 \left(\frac{T_0}{D_N}\right) - 2R_s \right), \tag{B.27}$$

where (B.13) is obtained by taking derivatives of the objective function of (B.27) with respect to each $D_i, i = 1, 2 \ldots, N$.

Plugging D^{N*} (B.13) into (B.22), we obtain $\underline{D}_N\left(F, R_s, T^N\right)$ and proceed to evaluate $\underline{D}_N(F, R_s, T_0, T_N)$ in (B.23), which is

$$\underline{D}_N(F, R_s, T_0, T_N) = \min_{\substack{T_1, \ldots, T_{N-1} \geq 0: \\ \frac{1}{N}\sum_{i=1}^{N-1} T_i = \frac{1}{F} - \frac{T_0 + T_N}{N}}} g(T_1, \ldots, T_{N-1}), \tag{B.28}$$

where

$$g(T_1,\ldots,T_{N-1}) \triangleq \frac{F}{2N}\Bigg[T_0^2+T_N^2+2\log_2 e\,\lambda^*(F,R_s,N) + \sum_{i=1}^{N-1} T_i\sqrt{T_i^2+4\log_2 e\,\lambda^*(F,R_s,N)}\Bigg]. \tag{B.29}$$

We make use of the *Schur-convexity* of (B.29) to calculate $\underline{D}_N(F,R_s,T_0,T_N)$. Assume that a function $f(x^d)$ is symmetric, and its first partial derivative with respect to each x_i, $i=1,\ldots,d$ exists. Then, $f(x^d)$ is Schur-convex if and only if

$$(x_i-x_j)\left(\frac{\partial f(x^d)}{\partial x_i}-\frac{\partial f(x^d)}{\partial x_j}\right)\geq 0,\ \forall\, i,j=1,\ldots,d. \tag{B.30}$$

It is clear that $g(T_1,\ldots,T_{N-1})$ is symmetric since it is invariant to the permutations of $T_1,\ldots,T_{N-1}$. To calculate the partial derivatives of (B.29), we first compute the implicit differentiation $\frac{\partial\lambda^*(F,R_s,N)}{\partial T_i}$ by taking the derivative with respect to T_i on the both sides of (B.12), yielding

$$\frac{\partial\lambda^*(F,R_s,N)}{\partial T_i}=\frac{1}{\sqrt{T_i^2+4\log_2 e\,\lambda^*(F,R_s,N)}}\cdot \frac{2\lambda^*(F,R_s,N)}{1+\sum_{k=1}^{N-1}\frac{T_k}{\sqrt{T_k^2+4\log_2 e\,\lambda^*(F,R_s,N)}}}. \tag{B.31}$$

Using (B.31) to compute the first partial derivative, we obtain

$$\frac{\partial g(T_1,\ldots,T_{N-1})}{\partial T_i}=\frac{F}{N}\sqrt{T_i^2+4\log_2 e\,\lambda^*(F,R_s,N)}. \tag{B.32}$$

Using (B.32), we can verify that $g(T_1,\ldots,T_{N-1})$ satisfies (B.30). Therefore, $g(T_1,\ldots,T_{N-1})$ is a Schur-convex function.

Let $x=(x_1,\ldots,x_d)\in\mathbb{R}^d$, $y=(y_1,\ldots,y_d)\in\mathbb{R}^d$ be two non-increasing sequences of real numbers. Recall that x is *majorized* by y if for each $k=1,\ldots,d$, $\sum_{i=1}^k x_i\leq\sum_{i=1}^k y_i$ with equality if $k=d$. For a Schur-convex function f, if x is majorized by y, then $f(x)\leq f(y)$. In our case, the feasible T_i's must satisfy the constraint in (B.28). Any sequence $T_1,\ldots,T_{N-1}$ that satisfies the constraint in (B.28) majorizes the sequence in (B.11). Thus, the infimum in (B.28) is achieved by the sequence $T_1^*,\ldots,T_{N-1}^*$ in (B.11). Finally, plugging $T_1^*,\ldots,T_{N-1}^*$ (B.11) into (B.28), and further plugging (B.28) into the right side of (B.24), we complete the proof.

B.7 Proof of Lemma 13

Plugging (B.15) into (B.13), we obtain:

$$D_1^* = \cdots = D_{N-1}^* = \frac{-\frac{N}{F(N+1)} + \sqrt{\left(\frac{N}{F(N+1)}\right)^2 + 4\log_2 e\, \lambda^*(F, R_s, N)}}{2}, \tag{B.33a}$$

$$D_N^* = \frac{F(N+1)}{N} \log_2 e\, \lambda^*(F, R_s, N), \tag{B.33b}$$

where $\lambda^*(F, R_s, N)$ is defined in Lemma 13. We first show that the T^N in (B.15) and the D^N in (B.33) satisfy the deleted constraint (B.9b). Then, we plug T^N (B.15) and D^N (B.33) as feasible solutions into the minimization problem associated with $D_N(F, R_s)$ in (B.7b) to obtain the upper bound in (B.14).

For $i = 2, \ldots, N-1$, the deleted constraint (B.9b) is satisfied trivially, since $D_{i-1} = D_i$ and $T_{i-1} \geq 0$. To prove that the deleted constraint (B.9b) also holds at $i = 1$ and N, we upper bound $\lambda^*(F, R_s, N)$ for every $N > 2$. If

$$T_1 = \cdots = T_{N-1}, \tag{B.34}$$

we can rearrange terms in the complementary slackness condition (B.12) and conclude $x = \lambda^*(F, R_s, N) \log_2 e$ is the unique solution to the following equation,

$$h_N(T_0, T_N, T_1, R_s, x) - x = 0, \tag{B.35}$$

where $h_N(T_0, T_N, T_1, R_s, x)$ is defined to be equal to

$$\frac{T_1^2}{2^{2R_s + \frac{2}{N-1}R_s - \frac{\log_2 T_0 + \log_2 T_N}{N-1} + \frac{\log_2 x}{N-1}} - 1} + \left(\frac{T_1}{2^{2R_s + \frac{2}{N-1}R_s - \frac{\log_2 T_0 + \log_2 T_N}{N-1} + \frac{\log_2 x}{N-1}} - 1}\right)^2. \tag{B.36}$$

The left side of (B.35) monotonically decreases as x increases.

Given R_s and plugging (B.15) into (B.35), we conclude that the $\lambda^*(F, R_s, N)$ in Lemma 13 is the unique solution to

$$h_N\left(\frac{N}{F(N+1)}, \frac{N}{F(N+1)}, \frac{N}{F(N+1)}, R_s, x\right) - x = 0, \tag{B.37}$$

Plugging

$$x = \frac{N^2}{2F^2(N+1)^2} \tag{B.38}$$

into (B.37), we observe that the left side of (B.37) is ≤ 0 for all $N > 2$. Thus, we conclude

$$\lambda^*(F, R_s, N) \log_2 e \leq \frac{N^2}{2F^2(N+1)^2}, \ \forall\, N > 2. \tag{B.39}$$

Plugging (B.39) into (B.33), we obtain

$$D_1^* \leq \sqrt{\lambda^*(F, R_s, N) \log_2 e} \leq \frac{N}{F(N+1)}, \tag{B.40a}$$

$$D_N^* \leq \frac{N}{2F(N+1)}. \tag{B.40b}$$

Substituting (B.15) and (B.40) into (B.9b), we conclude that (B.9b) holds for $i = 1$ and $i = N$. Now, we can plug (B.15) and (B.33) as feasible solutions into (B.7b) to obtain the right side of (B.14).

B.8 Proof of Lemma 14

From Lemmas 12 and 13, and (B.7a), we have

$$\liminf_{N\to\infty} \underline{D}_N(F, R_s) \leq D_{\mathrm{DET}}(F, R_s) \leq \limsup_{N\to\infty} \bar{D}_N(F, R_s). \tag{B.41}$$

We show that both bounds are equal to the right side of (3.31).

To compute the lower bound in (B.41), we need to understand the behavior of $T^*(F, N)$, $\lambda^*(F, R_s, N)$, and T_0^*, T_N^* as N goes to infinity, where T_0^*, T_N^* achieve the minimum of the left side of (B.41). T_0^* and T_N^* must increase as

$$T_0^* + T_N^* = O\left(\sqrt{N}\right), \tag{B.42}$$

or $\frac{T_0^{*2}+T_N^{*2}}{N}$ in (B.10b) will blow up to infinity as $N \to \infty$. Substituting (B.42) to (B.11), we obtain

$$T^*(F, N) = \frac{1}{F} + O\left(\frac{1}{\sqrt{N}}\right). \tag{B.43}$$

We proceed to compute

$$\lambda^* \triangleq \lim_{N\to\infty} \lambda^*(F, R_s, N). \tag{B.44}$$

For given T_0^*, T_N^*, and R_s, $x = \lambda^*(F, R_s, N) \log_2 e$ is the unique solution to (B.35) with T_0, T_N, and $T(N)$ replaced by T_0^*, T_N^*, and $T^*(F, N)$ in (B.11). We prove that

$$\lambda^* \log_2 e \geq \frac{1}{2^{2R_s} F^2}, \tag{B.45a}$$

$$\lambda^* \log_2 e \leq \frac{1}{2F^2}. \tag{B.45b}$$

We substitute (B.42) and (B.43) into the left side of (B.35) and take $\lim_{N\to\infty}$ to conclude that

$$\lim_{N\to\infty} h_N\left(T_0^*, T_N^*, T^*(F, N), R_s, \frac{1}{2F^2}\right) - \frac{1}{2F^2} \leq 0. \tag{B.46}$$

Since the left side of (B.35) is monotonically decreasing in x, we conclude (B.45a) holds. To prove (B.45b), we similarly compute

$$\lim_{N\to\infty} h_N\left(T_0^*, T_N^*, T^*(F,N), R_s, \frac{1}{2^{2R_s}F^2}\right) - \frac{1}{2^{2R_s}F^2} \geq 0. \tag{B.47}$$

Via the squeeze theorem, (B.45) implies

$$\lambda^*(F, R_s, N) = O(1). \tag{B.48}$$

Plugging (B.42), (B.43), and (B.48) into (B.35), and taking $N \to \infty$ on both sides of (B.35), we obtain

$$\lambda^* \log_2 e = \frac{1}{F^2(2^{2R_s}-1)^2} + \frac{1}{F^2(2^{2R_s}-1)}. \tag{B.49}$$

Plugging (B.42), (B.43), and (B.49) into (B.10b) and taking $\lim_{N\to\infty}$, we compute that $\lim_{N\to\infty} \underline{D}_N(F, R_s)$ is equal to

$$\frac{1}{2F} + \frac{1}{F(2^{2R_s}-1)} + \lim_{N\to\infty} \inf_{\substack{T_0\geq 0, T_N \geq 0 \\ T_0+T_N \leq \frac{N}{F}}} \frac{F}{2}\left(\frac{T_0^2+T_N^2}{N}\right) \tag{B.50a}$$

$$= \frac{1}{2F} + \frac{1}{F(2^{2R_s}-1)}, \tag{B.50b}$$

where 0 is achieved in the last term of (B.50a) by choosing any pair of $T_0, T_N \geq 0$ that satisfies

$$T_0 + T_N = o\left(\sqrt{N}\right). \tag{B.51}$$

We choose T_0 and T_N in (B.15) that satisfy (B.51), such that together with $T_1, \ldots, T_{N-1}$ in (B.15), the lower bound of $D_{\text{DET}}(F, R_s)$ in (B.41) is achieved.

Now, we compute the upper bound in the right side of (B.41). $\lambda^*(F, R_s, N)\log_2 e$ in (B.14b) is the unique solution to (B.35). Note that (B.49) holds for any T_0 and T_N that satisfy (B.42). Since T_0 and T_N in (B.15) satisfy (B.42), we conclude that the $\lim_{N\to\infty}$ of $\lambda^*(F, R_s, N)\log_2 e$ in (B.14b) is also equal to (B.49). Plugging (B.49) into (B.14b) and taking $\limsup_{N\to\infty}$, we calculate that the upper bound of $D_{\text{DET}}(F, R_s)$ in (B.41) is equal to (B.50b).

Furthermore, we observe that the uniform sampling intervals (B.15) achieving both the upper and the lower bound of $D_{\text{DET}}(F, R_s)$, converge to $\frac{1}{F}$ asymptotically. We conclude that the uniform sampling policy with the sampling interval $\frac{1}{F}$ achieves $D_{\text{DET}}(F, R_s)$.

B.9 Proof of Lemma 15

The max-min inequality and (B.16) imply that

$$D_{\mathrm{DET}}(R) \leq \min_{\substack{F>0, R_s \geq 1: \\ FR_s \leq R}} \limsup_{N\to\infty} \bar{D}_N(F, R_s). \tag{B.52}$$

On the other hand,

$$D_{\mathrm{DET}}(R) \geq \lim_{N\to\infty} \inf_{\substack{F>0, R_s \geq 1: \\ FR_s \leq R}} \underline{D}_N(F, R_s) \tag{B.53a}$$

$$= \inf_{\substack{F>0, R_s \geq 1: \\ FR_s \leq R}} \lim_{N\to\infty} \underline{D}_N(F, R_s), \tag{B.53b}$$

where (B.53a) is by (B.16), and (B.53b) will be proved in the sequel. Using (B.41) with both bounds equal to each other, (B.52), and (B.53), we complete the proof of Lemma 15.

We proceed to prove (B.53b) via *the fundamental theorem of* Γ*-convergence*. Let $\mathcal{X}$ be a topological space and $G_N : \mathcal{X} \to [0, +\infty]$, $N = 1, 2, \ldots$, be a sequence of functions defined on $\mathcal{X}$. A sequence of functions G_N, $N = 1, 2, \ldots$ Γ*-converges* [118] to its Γ*-limit* $G\colon \mathcal{X} \to [0, +\infty]$ if:

(i) For every $x \in \mathcal{X}$, and for every sequence $x_N \in \mathcal{X}, N = 1, 2, \ldots$ converging to x,

$$G(x) \leq \liminf_{N\to\infty} G_N(x_N). \tag{B.54}$$

(ii) For every $x \in \mathcal{X}$, there exists a sequence $x_N \in \mathcal{X}, N = 1, 2, \ldots$ converging to x such that

$$G(x) \geq \limsup_{N\to\infty} G_N(x_N). \tag{B.55}$$

A sequence of functions G_N, $N = 1, 2, \ldots$ is *equicoercive* [118] if there exists a compact set $\mathcal{K}$ independent of N s.t.

$$\inf_{x\in\mathcal{X}} G_N(x) = \inf_{x\in\mathcal{K}} G_N(x). \tag{B.56}$$

The fundamental theorem of Γ-convergence [118] says that if G_N is equicoercive and Γ-converges to $G\colon \mathcal{X} \to [0, +\infty]$, then

$$\min_{x\in\mathcal{X}} G(x) = \lim_{N\to\infty} \inf_{x\in\mathcal{X}} G_N(x). \tag{B.57}$$

We will show that for any scalars $F > 0$, $R_s \geq 1$ and for any sequences $F_{(N)} \to F$, $R_{s(N)} \to R_s$, we have

$$\lim_{N\to\infty} \underline{D}_N(F_{(N)}, R_{s(N)}) = D_{\mathrm{DET}}(F, R_s), \tag{B.58}$$

which means in particular that $D_{\mathrm{DET}}(\cdot,\cdot)$ is the Γ-limit of $\underline{D}_N(\cdot,\cdot)$. We will also prove that $\underline{D}_N(F,R_s)$ is equicoercive, and (B.53b) will follow via the fundamental theorem of Γ-convergence. By verifying that the reasoning in (B.42)-(B.50) goes through replacing F and R_s by $F_{(N)}$ and $R_{s(N)}$ respectively, we conclude that (B.58) holds.

It remains to prove that $\underline{D}_N(F,R_s)$ is equicoercive. Ignoring the two non-negative $\lambda^*(F,R_s,N)$ terms in (B.10b), we observe that $\underline{D}_N(F,R_s)$ is lower bounded by

$$\inf_{\substack{T_0\geq 0,T_N\geq 0\\ T_0+T_N\leq \frac{N}{F}}} \frac{F}{2}\left(\frac{T_0^2+T_N^2}{N}+\frac{N-1}{N}T^*(F,N)^2\right) \tag{B.59a}$$

$$=\inf_{\substack{T_0\geq 0,T_N\geq 0\\ T_0+T_N\leq \frac{N}{F}}} \frac{1}{2}\left(F\frac{T_0^2+T_N^2}{N}+\frac{N}{F(N-1)}\left(1-\frac{F(T_0+T_N)}{N}\right)^2\right), \tag{B.59b}$$

where (B.59b) is obtained by plugging (B.11) into (B.59a). We denote the objective function in (B.59b) by $q(T_0,T_N)$. We prove that $q(T_0,T_N)$ is a Schur-convex function: 1) $q(T_0,T_N)$ is symmetric, since it is invariant to the permutations of T_0 and T_N; 2) the first-order partial derivatives of $q(T_0,T_N)$ with respect to T_0 and T_N are

$$\frac{\partial q}{\partial T_0}=\frac{F}{N}T_0+\frac{F}{N(N-1)}(T_0+T_N)-\frac{1}{N-1}, \tag{B.60a}$$

$$\frac{\partial q}{\partial T_N}=\frac{F}{N}T_N+\frac{F}{N(N-1)}(T_0+T_N)-\frac{1}{N-1}, \tag{B.60b}$$

where (B.60) satisfies (B.30). Using the property of Schur-convex functions stated in Lemma 12 after (B.32), we know that the minimum of $q(T_0,T_N)$ is achieved by

$$T_0=T_N=a,\ \text{for some } 0\leq a\leq\frac{N}{2F}. \tag{B.61}$$

Plugging (B.61) into $q(T_0,T_N)$, we find that the optimal a that minimizes $q(a,a)$ is given by

$$a=\frac{N}{(N+1)F}. \tag{B.62}$$

Plugging (B.61) and (B.62) into (B.59b), we obtain

$$\underline{D}_N(F,R_s)\geq\frac{N^2}{2F(N+1)^2}. \tag{B.63}$$

On the other hand, plugging (B.39) into the right side of (B.10), we obtain

$$\bar{D}_N(F,R_s)\leq\frac{3N}{2F(N+1)^2}+\frac{\sqrt{3}N(N-1)}{2F(N+1)^2}. \tag{B.64}$$

Choosing $F = R$ in (B.64), we conclude that

$$\inf_{\substack{F>0, R_s \geq 1 \\ FR_s \leq R}} \underline{D}_N(F, R_s) \leq \frac{3N}{2R(N+1)^2} + \frac{\sqrt{3}N(N-1)}{2R(N+1)^2}. \tag{B.65}$$

For any $F \in \left(0, \frac{R}{3+\sqrt{3}}\right)$, the right side of (B.63) is larger than the right side of (B.65). Thus, the infimum is attained in the following compact set:

$$F \in \left[\frac{R}{3+\sqrt{3}}, R\right], \tag{B.66}$$

where the upper bound of F is obtained by lower-bounding R_s by 1. Correspondingly, R_s lies in the following compact set:

$$R_s \in \left[1, 3+\sqrt{3}\right], \tag{B.67}$$

Using (B.66)–(B.67), we conclude that $\underline{D}_N(F, R_s)$ is equicoercive.

B.10 Converse proof of Theorem 8

We show that the DSTF for a BEC is lower bounded by the DSTF for a packet-drop channel (3.54). We denote by Π^{BEC} and $\mathcal{C}^{\mathrm{BEC}}$ the set of all causal sampling policies and the set of all causal compressing policies in Definition 16. We lower bound the DSTF for a BEC as

$$\begin{aligned}
& D^{\mathrm{BEC}}(n, T) \\
= & \inf_{\substack{\pi \in \Pi^{\mathrm{BEC}} \\ \{f_t\}_{t=0}^{\infty} \in \mathcal{C}^{\mathrm{BEC}}: \\ \text{(2.43)}}} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} (X_t - \hat{X}_t^{\mathrm{BEC}})^2\right] && \text{(B.68a)} \\
\geq & \inf_{\substack{\pi \in \Pi^{\mathrm{BEC}}: \\ \text{(2.43)}}} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} \left(X_t - \mathbb{E}\left[X_t \middle| \{X_s\}_{s=0}^{S_i}, \{\tau_j\}_{j \in \mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right)}\right]\right)^2\right] && \text{(B.68b)} \\
= & \inf_{\substack{\pi \in \Pi^{\mathrm{BEC}}: \\ \text{(2.43)}}} \frac{1}{T}\mathbb{E}\left[\sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} (X_t - \bar{X}_t^{\mathrm{pd}})^2\right] && \text{(B.68c)} \\
= & \underline{D}^{\mathrm{pd}}(n, T). && \text{(B.68d)}
\end{aligned}$$

where (B.68b) holds since

$$\sigma\left(\{V_j, \tau_j\}_{j \in \mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right)}\right) \subseteq \sigma\left(\{X_s\}_{s=0}^{S_i}, \{\tau_j\}_{j \in \mathcal{N}\left(\{B_{\tau_j}\}_{j=1}^{i}\right)}\right), \tag{B.69}$$

where S_i is the time of the last successful transmission defined in (2.40); (B.68c) holds since by the strong Markov property of the Lévy process, the estimate in (B.68b) is equal to $\bar{X}_t^{\mathrm{pd}}$ in (2.42).

Appendix C

CAUSAL JOINT SOURCE-CHANNEL CODING WITH FEEDBACK: PROOFS

C.1 A partition that satisfies (4.24)

For any $t = 1, 2, \ldots$ and any $y^{t-1} \in \mathcal{Y}^{t-1}$, we show that the greedy heuristic algorithm [101] yields a partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$ that satisfies the partitioning rule (4.24).

The greedy heuristic algorithm operates as follows. At time t, it initializes all the groups $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$ by empty sets and initializes all the group priors $\{\pi_x(y^{t-1})\}_{x\in\mathcal{X}}$ by zeros. It sorts all the source sequences in $[q]^{N(t)}$ according to their priors $\theta_i(y^{t-1})$, $i \in [q]^{N(t)}$ in a descending manner. Starting from the sequence with the largest prior, it moves the sequence in the sorted list to the group $\mathcal{G}_{x^*}(y^{t-1})$ whose current group prior has the largest gap to the corresponding capacity-achieving probability, i.e.,

$$x^* \triangleq \arg\max_{x\in\mathcal{X}} P_X^*(x) - \pi_x(y^{t-1}). \tag{C.1}$$

The group prior $\pi_{x^*}(y^{t-1})$ is updated after each move. The partitioning process repeats until all the source sequences have been classified.

We show that the resulting partition $\{\mathcal{G}_x(y^{t-1})\}_{x\in\mathcal{X}}$ satisfies (4.24). We first notice that the maximization problem on the right side of (C.1) must be strictly larger than zero before all source sequences have been classified. If moving sequence i to group $\mathcal{G}_{x^*}(y^{t-1})$ leads to

$$\pi_{x^*}(y^{t-1}) - P_X^*(x^*) < 0, \tag{C.2}$$

then (4.24) is obviously satisfied. If moving sequence i to group $\mathcal{G}_{x^*}(y^{t-1})$ leads to

$$\pi_{x^*}(y^{t-1}) - P_X^*(x^*) \geq 0, \tag{C.3}$$

then (4.24) is satisfied since (C.2) holds before the move, and sequence i has the smallest prior in $\mathcal{G}_{x^*}(y^{t-1})$ after the move. Furthermore, if the group $\mathcal{G}_{x^*}(y^{t-1})$ satisfies (C.3), it will no longer be the solution to the maximization problem in (C.1) and thus will no longer accept new sequences. This means that (4.24) holds for all $x \in \mathcal{X}$ at the end of the greedy heuristic partitioning.

C.2 Channel input distribution is equal to the capacity-achieving distribution

We show that (4.31) holds, i.e., the channel input distribution is equal to the capacity-achieving distribution. For any $x \in \mathcal{X}$ and $y^{t-1} \in \mathcal{Y}^{t-1}$, we expand right side of (4.31) as

$$P_{X_t|Y^{t-1}}(x|y^{t-1}) = \sum_{z\in\mathcal{X}} P_{X_t|Z_t,Y^{t-1}}(x|z,y^{t-1}) P_{Z_t|Y^{t-1}}(z|y^{t-1}) \tag{C.4a}$$

$$= \sum_{z\in\mathcal{X}} P_{X_t|Z_t,Y^{t-1}}(x|z,y^{t-1}) \pi_z(y^{t-1}) \tag{C.4b}$$

$$= P_{X_t|Z_t,Y^{t-1}}(x|x,y^{t-1}) \pi_x(y^{t-1}) \tag{C.4c}$$

$$+ \sum_{z\neq x} P_{X_t|Z_t,Y^{t-1}}(x|z,y^{t-1}) \pi_z(y^{t-1}), \tag{C.4d}$$

where (C.4a) holds by the law of total probability and (C.4b) holds by the definition of Z_t in (4.29).

By the randomization distribution in (4.30), if $x \in \overline{\mathcal{X}}(y^{t-1})$, then (C.4c) is equal to $P_X^*(x)$ and (C.4d) is equal to 0, and if $x \in \underline{\mathcal{X}}(y^{t-1})$, then (C.4c) is equal to $\pi_x(y^{t-1})$ and (C.4d) is equal to

$$\sum_{z\in\overline{\mathcal{X}}(y^{t-1})} \frac{p_{z\to x}}{\pi_x(y^{t-1})} \pi_x(y^{t-1}) = P_X^*(x) - \pi_x(y^{t-1}), \tag{C.5}$$

where (C.5) uses (4.28).

C.3 Converse proof of Theorem 9

Inspired by Berlin et al.'s converse proof [99] for Burnashev's reliability function, we provide a converse bound on the JSCC reliability function for a fully accessible source by lower bounding the expected stopping time of an arbitrary code with block encoding using the error probability at the stopping time. The converse bound continues to apply for the JSCC reliability function for streaming, since given a DMC, every code with instantaneous encoding for transmitting the first k symbols of a $(q, \{t_n\}_{n=1}^{\infty})$ DSS in Definition 20 is a special code with block encoding for transmitting the first k symbols $S^k \in [q]^k$ of a DS (4.3).

We consider k symbols $S^k \in [q]^k$ of a DS with source distribution P_{S^k}, and we fix a non-degenerate DMC with a single-letter transition probability $P_{Y|X} \colon \mathcal{X} \to \mathcal{Y}$. We fix an arbitrary code with block encoding with a stopping time η_k for transmitting S^k over the non-degenerate DMC with feedback. We assume that the decoder is a MAP decoder (4.46), since given any encoding function and any stopping time

in Definition 20, the MAP decoder (4.46) achieves the minimum error probability (4.15). For brevity, we denote the error probability of a MAP decoder given channel outputs $y^t \in \mathcal{Y}^t$ by

$$P_e(y^t) \triangleq 1 - \max_{s \in [q]^k} P_{S^k|Y^t}(s|y^t), \tag{C.6}$$

and we denote the error probability of a MAP decoder at the stopping time η_k by

$$P_e \triangleq \mathbb{E}[P_e(Y^{\eta_k})]. \tag{C.7}$$

We define stopping time τ_δ as

$$\tau_\delta \triangleq \min\{t\colon P_e(y^t) \leq \delta \text{ or } t = \eta_k\}. \tag{C.8}$$

To obtain the converse bound on the JSCC reliability function for a fully accessible source, we establish a lower bound on the expected decoding time $\mathbb{E}[\eta_k]$ using the error probability P_e and and source distribution P_{S^k}. To this end, we lower bound $\mathbb{E}[\tau_\delta]$ and $\mathbb{E}[\eta_k - \tau_\delta]$, respectively. The lower bound on $\mathbb{E}[\tau_\delta]$ is stated below.

Lemma 16 (Modified Lemma 2 in [99])**.** *Consider k symbols $S^k \in [q]^k$ of a DS with source distribution P_{S^k} (4.1) and fix a non-degenerate DMC with capacity C (4.10). For any $\delta \in \left(0, \frac{1}{2}\right]$, it holds that*

$$\mathbb{E}[\tau_\delta] \geq \frac{H(S^k)}{C}\left(1 - \left(\delta + \frac{P_e}{\delta}\right)\frac{\log q^k}{H(S^k)}\right) - \frac{h(\delta)}{C}. \tag{C.9}$$

Proof. Appendix C.4. □

The lower bound on $\mathbb{E}[\eta_k - \tau_\delta]$ is stated below.

Lemma 17 (Modified Eq. (17) [99])**.** *Consider k symbols $S^k \in [q]^k$ of a DS with source distribution P_{S^k} and fix a non-degenerate DMC with transition probability $P_{Y|X}\colon \mathcal{X} \to \mathcal{Y}$ and maximum KL divergence C_1 (4.11). For any $\delta \in \left(0, \frac{1}{2}\right]$, it holds that*

$$\mathbb{E}[\eta_k - \tau_\delta] \geq \frac{\log \frac{1}{P_e} - \log 4 + \log(\min\left\{p_{\min}\delta, 1 - \max_{s \in [q]^k} P_{S^k}(s)\right\})}{C_1}, \tag{C.10}$$

where $p_{\min}$ in (C.10) is the minimum channel transition probability (4.34).

Proof. Appendix C.5. □

Summing up the right sides of (C.9) and (C.10), we obtain the following lower bound on the expected decoding time η_k of an arbitrary code with block encoding:

$$\mathbb{E}[\eta_k] \geq \frac{H(S^k)}{C}\left(1-\left(\delta+\frac{P_e}{\delta}\right)\frac{\log q^k}{H(S^k)}\right)+\frac{\log\frac{1}{P_e}}{C_1} + \frac{-\log 4+\log(\min\left\{p_{\min}\delta, 1-\max_{s\in[q]^k} P_{S^k}(s)\right\})}{C_1}-\frac{h(\delta)}{C}. \tag{C.11}$$

The asymptotic performance of the lower bound (C.11) relies on two properties of the DS in Lemma 18, stated next.

Lemma 18. *Consider a DS with a well-defined and positive entropy rate H (4.2) and a finite single-letter alphabet $[q]$. Then,*

$$\lim_{k\to\infty}\frac{\log q^k}{H(S^k)}=\frac{\log q}{H}<\infty, \tag{C.12}$$

$$\liminf_{k\to\infty}\left(1-\max_{s\in[q]^k} P_{S^k}(s)\right)>0. \tag{C.13}$$

Proof. The proof of (C.13) is in Appendix C.6. □

Plugging (C.13)–(C.12) and $\delta=-\frac{1}{\log P_e}$ into the right side of (C.11), we obtain

$$\mathbb{E}[\eta_k]\geq\left(\frac{H(S^k)}{C}+\frac{\log\frac{1}{P_e}}{C_1}\right)(1-o(1)), \tag{C.14}$$

where $o(1)$ in (C.14) is a positive term that converges to 0 as both $P_e\to 0$ and $k\to\infty$. Rearranging terms of (C.14), we conclude that $E(R)$ is upper bounded by the right side of (4.38). Similar to [40], [99, Eq. (5)], [87, Proposition 9], here we need not consider the case where P_e does not converge to zero since this means $E(R)=0$.

C.4 Proof of Lemma 16

We follow Berlin et al.'s notations [99]: we denote by $\mathcal{H}(S^k|Y^t)$ a random variable that satisfies $\mathcal{H}(S^k|y^t)=H(S^k|Y^t=y^t)$. Note that $\mathbb{E}[\mathcal{H}(S^k|Y^t)]=H(S^k|Y^t)$. If any step below has already been proved in [99], we avoid repeated reasoning by referring to the proof in [99]. Compared to Berlin et al.'s proof in [99, Sec. IV], the proof below does not assume that the source is equiprobably distributed–it keeps the generic form of the source prior P_{S^k}.

The sequence $\{\mathcal{H}(S^k|Y^t) + tC\}_{t=0,1,\dots}$ is a submartingale ([99, Lemma 2]). Using Doob's optional stopping theorem [119], the initial state of the submartingale $\{\mathcal{H}(S^k|Y^t) + tC\}_{t=0,1,\dots}$ is upper bounded as

$$H(S^k) \leq H(S^k|Y^{\tau_\delta}) + \mathbb{E}[\tau_\delta]C. \tag{C.15}$$

For $\delta \in \left(0, \frac{1}{2}\right]$, the conditional entropy on the right side of (C.15) is upper bounded as

$$H(S^k|Y^{\tau_\delta}) \leq h(\delta) + \left(\delta + \frac{P_e}{\delta}\right) \log q^k \tag{C.16}$$

using Fano's inequality (in the same manner as in [99, Eq. (14)–(16)]). Plugging (C.16) to the right side of (C.15) and rearranging terms, we obtain (C.9).

C.5 Proof of Lemma 17

We obtain the lower bound on $\mathbb{E}[\eta_k - \tau_\delta]$ in Lemma 17 by constructing a binary hypothesis test performed over a non-degenerate DMC with feedback. We first state a lower bound on the error probability of such a test. Consider a binary hypothesis test (H_0, H_1) performed over a DMC with feedback via a variable-length code with block encoding. The encoder sends a sequence of symbols $X_1, X_2, \dots$, over the given DMC with feedback, such that at the stopping time T, if H_0 is true, then the channel output vector Y^T is distributed according to Q_{H_0}, otherwise, the channel output vector Y^T is distributed according to Q_{H_1}. At the stopping time, the decoder uses the decoding function $\hat{W}\colon \mathcal{Y}^T \to \{H_0, H_1\}$ to form a decoded hypothesis. We denote the set of channel outputs y^T that leads to decoded hypothesis H_i, $i \in \{0, 1\}$ by

$$\mathcal{Y}_{H_i} \triangleq \{y^T \in \mathcal{Y}^T\colon \hat{W}(y^T) = H_i\}, i \in \{0, 1\}. \tag{C.17}$$

We denote by p_{H_i}, $i \in \{0, 1\}$, the prior probability of hypothesis H_i, $i \in \{0, 1\}$ before the transmission. We denote the error probability of the binary hypothesis test at the stopping time T by

$$P_b \triangleq p_{H_0} Q_{H_0}(\mathcal{Y}_{H_1}) + p_{H_1} Q_{H_1}(\mathcal{Y}_{H_0}). \tag{C.18}$$

Lemma 19 (Lemma 1 in [99])**.** *Consider a binary hypothesis test with hypotheses H_0 and H_1 performed over a non-degenerate DMC with feedback that has maximum KL divergence C_1 (4.11) via a variable-length code with block encoding. The error probability of the binary hypothesis test P_b at stopping time T is lower bounded as*

$$P_b \geq \frac{\min\{p_{H_0}, p_{H_1}\}}{4} e^{-C_1 \mathbb{E}[T]}, \tag{C.19}$$

where p_{H_0} and p_{H_1} are the prior probabilities of the hypotheses.

We employ the same hypothesis test as that in [99, Section V]. Compared to Berlin et al.'s proof [99, Sec. V], the proof below lower bounds the priors of the hypotheses differently since we consider a generic source distribution whereas Berlin at al.'s [99] considered equiprobable source symbols.

The binary hypothesis test (cf. [99, Section V]) starts at time $\tau_\delta + 1$ and operates as follows. Given any Y^{τ_δ}, we partition the alphabet $[q]^k$ into two sets $\mathcal{G}(Y^{\tau_\delta})$ and $[q]^k \setminus \mathcal{G}(Y^{\tau_\delta})$ (we will specify $\mathcal{G}$ in the sequel). The two hypotheses are $H_0 \colon S^k \in \mathcal{G}(Y^{\tau_\delta})$ and $H_1 \colon S^k \in [q]^k \setminus \mathcal{G}(Y^{\tau_\delta})$. At the stopping time η_k, the MAP decoder outputs the estimate of the source $\hat{S}^k_{\eta_k}$ using the channel outputs Y^{η_k}. If the estimate satisfies $\hat{S}^k_{\eta_k} \in \mathcal{G}(Y^{\tau_\delta})$, then we declare H_0, otherwise, we declare H_1. The error probability of decoding S^k is lower bounded by the error probability of the binary hypothesis test ([99, the second paragraph below Prop. 2]), i.e., given any $t \geq 0$, $y^t \in \mathcal{Y}^t$,

$$\begin{aligned}\mathbb{P}[\hat{S}^k_{\eta_k} \neq S^k | Y^{\tau_\delta} = y^t] \geq \mathbb{P}[\hat{S}^k_{\eta_k} \notin \mathcal{G}(Y^{\tau_\delta}), S^k \in \mathcal{G}(Y^{\tau_\delta}) | Y^{\tau_\delta} = y^t] \\ + \mathbb{P}[\hat{S}^k_{\eta_k} \in \mathcal{G}(Y^{\tau_\delta}), S^k \notin \mathcal{G}(Y^{\tau_\delta}) | Y^{\tau_\delta} = y^t].\end{aligned} \tag{C.20}$$

We invoke Lemma 19 with $p_{H_0} \leftarrow \mathbb{P}[H_0 | Y^{\tau_\delta} = y^t]$, $p_{H_1} \leftarrow \mathbb{P}[H_1 | Y^{\tau_\delta} = y^t]$, $\mathbb{E}[T] \leftarrow \mathbb{E}[\eta_k - \tau_\delta | Y^{\tau_\delta} = y^t]$ to further lower bound the left side of (C.20) and obtain

$$\mathbb{P}[\hat{S}^k_{\eta_k} \neq S^k | Y^{\tau_\delta} = y^t] \geq \frac{\min\{\mathbb{P}[H_0 | Y^{\tau_\delta} = y^t], \mathbb{P}[H_1 | Y^{\tau_\delta} = y^t]\}}{4} e^{-C_1 \mathbb{E}[\eta_k - \tau_\delta | Y^{\tau_\delta} = y^t]}. \tag{C.21}$$

To lower bound the minimization function on the right side of (C.21), we show that alphabet $[q]^k$ can always be partitioned into two groups $\mathcal{G}(Y^{\tau_\delta})$ and $[q^k] \setminus \mathcal{G}(Y^{\tau_\delta})$ such that for all $t \geq 0$, $y^t \in \mathcal{Y}^t$, the priors of the hypotheses are lower bounded as

$$\mathbb{P}[H_0 | Y^{\tau_\delta} = y^t] \geq \min\left\{p_{\min}\delta, 1 - \max_{s \in [q]^k} P_{S^k}(s)\right\}, \tag{C.22a}$$

$$\mathbb{P}[H_1 | Y^{\tau_\delta} = y^t] \geq \min\left\{p_{\min}\delta, 1 - \max_{s \in [q]^k} P_{S^k}(s)\right\}, \tag{C.22b}$$

where $p_{\min}$ is defined in (4.34). The priors of the hypotheses are both lower bounded by $p_{\min}\delta$ for any $\delta \in \left(0, \frac{1}{2}\right]$ if either event $A_1 \triangleq \{\tau_\delta \geq 1\}$ or event $A_2 \triangleq \{\tau_\delta = 0, \max_{s \in [q]^k} P_{S^k}(s) \leq 0.5\}$ occurs, see [99, Section V]. The threshold 0.5 defining event A_2 corresponds to Berlin et al.'s reasoning in [99, the second case in the third paragraph after Prop. 2], which says at time τ_δ, if the posteriors[1]

[1]The source posterior at time 0 is equal to the source prior P_{S^k}.

of all source sequences in $[q]^k$ are upper bounded by $1-\delta \in [0.5, 1]$, then $[q]^k$ can be divided into two groups with the priors of both hypotheses lower bounded by $p_{\min}\delta$. In Berlin et al.'s [99] channel coding context where the source symbols are equiprobably distributed, the union of the events $A_1 \cup A_2$ occurs almost surely, since $P_{S^k}(s) = \frac{1}{q^k}$. Yet, in the JSCC context, it is possible that event $A_3 \triangleq \{\tau_\delta = 0, \max_{s\in[q]^k} P_{S^k}(s) > 0.5\}$ occurs. When A_3 occurs, we move the sequence $s \in [q]^k$ that attains the maximum in event A_3 to $\mathcal{G}(Y^{\tau_\delta})$, and move the remaining sequences to the other group. This group partitioning rule implies (C.22). Plugging (C.22) into (C.21), taking an expectation of (C.21) over Y^{τ_δ}, and applying Jensen's inequality to e^{-x} on the right side of (C.21), we obtain

$$P_e \geq \frac{\min\{p_{\min}\delta, 1-\max_{s\in[q]^k} P_{S^k}(s)\}}{4} e^{-C_1 \mathbb{E}[\eta_k - \tau_\delta]}. \tag{C.23}$$

Rearranging terms in (C.23), we obtain (C.10).

C.6 Proof of Lemma 18

We show that (C.13) holds. We upper bound the entropy rate as

$$\lim_{k\to\infty} \frac{H(S^k)}{k} \leq \liminf_{k\to\infty} \frac{1}{k} \left(h\left(\max_{s\in[q]^k} P_{S^k}(s) \right) + \left(1 - \max_{s\in[q]^k} P_{S^k}(s)\right) k \log q \right) \tag{C.24}$$

$$= \liminf_{k\to\infty} \left(1 - \max_{s\in[q]^k} P_{S^k}(s)\right) \log q, \tag{C.25}$$

where (C.24) holds since fixing the probability of the source sequence that attains $\max_{s\in[q]^k} P_{S^k}(s)$, the equiprobable distribution on the rest of $q^k - 1$ sequences maximizes the concave entropy function; (C.25) holds since the binary entropy function in (C.24) is bounded between $[0, 1]$. Finally, (C.13) holds since the entropy rate is positive by assumption.

C.7 Achievability proof of Theorem 9: A (fully accessible) DS

We show that both the MaxEJS code for all non-degenerate DMCs [45, Sec. IV-C] and the SED code for non-degenerate symmetric binary-input DMCs [45, Sec. V-B] achieve $E(R)$ (4.38) for a DS. We denote a deterministic encoding function at time t by

$$\gamma_t \colon [q]^k \to \mathcal{X}, \tag{C.26}$$

we denote the vector of the message posteriors at time t by

$$\boldsymbol{\rho}(Y^t) \triangleq [P_{S^k|Y^t}(1|Y^t), P_{S^k|Y^t}(2|Y^t), \ldots, P_{S^k|Y^t}(q^k|Y^t)], \tag{C.27}$$

and we denote the extrinsic Jensen-Shannon (EJS) divergence [45] at time t by

$$\mathrm{EJS}(\boldsymbol{\rho}(Y^{t-1}), \gamma_t) \triangleq \tag{C.28}$$
$$\sum_{i=1}^{q^k} P_{S^k|Y^{t-1}}(i|Y^{t-1}) D\left(P_{Y|X=\gamma_t(i)} \middle\| \sum_{j\neq i} \frac{P_{S^k|Y^{t-1}}(j|Y^{t-1})}{1-P_{S^k|Y^{t-1}}(i|Y^{t-1})} P_{Y|X=\gamma_t(j)}\right).$$

The MaxEJS code [45, Section IV.C] sets its encoding function γ_t^* at time t by solving the maximization problem:

$$\gamma_t^* \triangleq \arg\max_{\gamma_t \in \mathcal{E}} \mathrm{EJS}(\boldsymbol{\rho}(Y^{t-1}), \gamma_t), \tag{C.29}$$

where $\mathcal{E}$ is the set of all possible deterministic functions γ_t (C.26). The SED code [45] corresponds to the instantaneous SED code in Section 4.5 for a fully accessible source.

Lemma 20, stated next, will be used to examine whether a code with block encoding achieves the JSCC reliability function for a fully accessible source.

Lemma 20. *Consider k symbols $S^k \in [q]^k$ of a DS with prior probability P_{S^k} and fix a non-degenerate DMC with capacity C (4.10) and the maximum KL divergence C_1 (4.11). A code with block encoding achieves the JSCC reliability function (4.38) for the fully accessible source if and only if its stopping time η_k and its error probability ϵ (4.15) at the stopping time η_k satisfy*

$$\mathbb{E}[\eta_k] \leq \left(\frac{H(P_{S^k})}{C} + \frac{\log\frac{1}{\epsilon}}{C_1}\right)(1+o(1)), \tag{C.30}$$

where $o(1) \to 0$ as $k \to \infty$.

Proof. If a code with block encoding satisfies (C.30), then it achieves $E(R)$ (4.38) because plugging (C.30) into (4.17) gives (4.38). Conversely, if a code with block encoding achieves $E(R)$ (4.38), then $\mathbb{E}[\eta_k]$ is upper bounded by the right side of (C.30). This is because any achievability bound on $\mathbb{E}[\eta_k]$ that is asymptotically larger than the right side of (C.30) cannot achieve (4.38). □

We show that the MaxEJS code and the SED code both satisfy (C.30). While [45, Eq. (32)] in [45, Theorem 1] is obtained by plugging a uniform prior of the message to the entropy function in [45, Appendix II, Eq. (71)], we leave the prior in its generic form and obtain a modified version of [45, Theorem 1] as follows.

Lemma 21 (Modified Theorem 1 in [45]). *Fix a non-degenerate DMC with capacity* C (4.10) *and maximum KL divergence* C_1 (4.11)*, and consider* k *symbols* $S^k \in [q]^k$ *of a DS with source distribution* P_{S^k}*. If the encoding functions* γ_t*,* $t = 1, \ldots, \eta_k$ *of a code with block encoding with the MAP decoder* (4.46) *and the* ϵ*-stopping rule* (4.49) *satisfy*

$$\mathrm{EJS}(\boldsymbol{\rho}(Y^{t-1}), \gamma_t) \geq C, \tag{C.31}$$

$$\mathrm{EJS}(\boldsymbol{\rho}(Y^{t-1}), \gamma_t) \geq \left(1 - \frac{1}{1 + \max\{\log q^k, \log \frac{1}{\epsilon}\}}\right) C_1,$$

$$\textit{if } \max_{i \in [q]^k} \rho_i(Y^{t-1}) \geq 1 - \frac{1}{1 + \max\{\log q^k, \log \frac{1}{\epsilon}\}}, \tag{C.32}$$

then the expected decoding time of the code with block encoding is upper bounded as

$$\mathbb{E}[\eta_k] \leq \frac{H(P_{S^k}) + \log\log \frac{q^k}{\epsilon}}{C} + \frac{\log \frac{1}{\epsilon} + 1}{C_1} + \frac{6(4C_2)^2}{CC_1}, \tag{C.33}$$

where $C_2 \triangleq \max_{y \in \mathcal{Y}} \frac{\max_{x \in \mathcal{X}} P_{Y|X}(y|x)}{\min_{x \in \mathcal{X}} P_{Y|X}(y|x)}$.

Since the MaxEJS code satisfies (C.31)–(C.32) for all non-degenerate DMCs by [45, Proposition 2], and the SED code [45, Sec. V-B] satisfies (C.31)–(C.32) for non-degenerate symmetric binary-input DMCs by [45, Proposition 4], we conclude from (C.33) and (C.12) that they satisfy (C.30).

C.8 Achievability proof of Theorem 9: A DSS with $f = \infty$

Fixing a $(q, \{t_n\}_{n=1}^{\infty})$ DSS with $f = \infty$, we show that $E(R)$ (4.38) is achievable by a buffer-then-transmit code that buffers the arriving symbols at times $t = 1, \ldots, t_k$ and operates as a JSCC reliability function (4.38)-achieving code with block encoding for k symbols S^k of a (fully accessible) DS with prior P_{S^k} at times $t \geq t_k + 1$ (e.g., the MaxEJS code [45]). To this end, we show an achievability (upper) bound on the expected stopping time of the buffer-then-transmit code.

We denote by η'_k the stopping time of the buffer-then-transmit code. We denote by η_k the stopping time of a code with block encoding that achieves the JSCC reliability function (4.38) for a fully accessible source, and we denote by ϵ_k its error probability at η_k (4.15). Since the decoding starts after time t_k, we have

$$\eta'_k = t_k + \eta_k. \tag{C.34}$$

We invoke Lemma 20 with $\epsilon \leftarrow \epsilon_k$ to upper bound $\mathbb{E}[\eta_k]$ on the right side of (C.34) and obtain an achievability bound on the expected decoding time $\mathbb{E}[\eta'_k]$ of the buffer-then-transmit code:

$$\mathbb{E}[\eta'_k] \leq \left(\frac{H(P_{S^k})}{C} + \frac{\log \frac{1}{\epsilon_k}}{C_1} \right) (1 + o(1)) + t_k. \tag{C.35}$$

Plugging (C.35) into (4.17), we obtain (4.39). Since $f = \infty$, the achievability bound (4.39) is equal to (4.38).

C.9 Achievability proof of Theorem 9: A DSS with $f < \infty$

Fixing a $(q, \{t_n\}_{n=1}^{\infty})$ DSS with $f < \infty$, we show that $E(R)$ (4.38) is achievable by a code with instantaneous encoding that implements the instantaneous encoding phase at times $t = 1, 2, \ldots, t_k$ and operates as a JSCC reliability function (4.38)-achieving code with block encoding for k symbols S^k of a (fully accessible) DS with prior $P_{S^k|Y^{t_k}}$ at times $t \geq t_k + 1$, where $Y_1, \ldots, Y_{t_k}$ are the channel outputs generated in the instantaneous encoding phase. To this end, we will use Lemmas 22–24, stated below, together with Lemma 20 in Appendix C.7 to obtain an achievability (upper) bound on the expected stopping time of the code.

We fix an error probability ϵ_k (4.15). We denote by η_k the stopping time that ensures ϵ_k of a JSCC reliability function-achieving code with block encoding for k symbols with prior $P_{S^k|Y^{t_k}}$. The stopping time η'_k of the code with instantaneous encoding is

$$\eta'_k = t_k + \eta_k \tag{C.36}$$

and its error probability is ϵ_k.

To upper bound the expected decoding time $\mathbb{E}[\eta'_k]$, it suffices to upper bound $\mathbb{E}[\eta_k]$ (C.36). Lemmas 22–24, stated below, show the behavior of the mutual information $I(S^k; Y^{t_k})$ as $k \to \infty$ generated by the instantaneous encoding phase in Section 4.3.

Lemma 22. *Fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS, and fix a non-degenerate DMC with capacity C (4.10) and the maximum KL divergence C_1 (4.11). The instantaneous encoding phase that operates at times $t = 1, 2, \ldots, t_k$ in Section 4.3 gives rise to*

$$I(S^k; Y^{t_k}) = t_k C - I(X^{t_k} \to Y^{t_k} | S^k). \tag{C.37}$$

Proof. Appendix C.10. □

Lemma 23, stated next, displays the implications of assumption (b) in Theorem 9. Given a DSS, we extract all the *distinct* symbol arriving times from $t_1 \le t_2 \le \dots$, and we denote the sequence of distinct symbol arriving times by

$$d_1 < d_2 < \dots \tag{C.38}$$

For example, if a DSS emits a source symbol every $\lambda \ge 1$ channel uses (4.6), then the symbol arriving times are equal to the distinct symbols arriving times, i.e., $t_n = d_n$, and Lemma 23 below trivially holds.

Lemma 23. *Fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS with $f < \infty$ and f satisfying assumption (b) in Theorem 9. Then,*

(i) The time interval between consecutive symbol arriving times satisfies

$$t_{n+1} - t_n = o(n), n = 1, 2, \dots; \tag{C.39}$$

(ii) The DSS has an infinite number of distinct symbol arriving times $d_{n'}$, $n' = 1, 2, \dots$.

Proof. (i) Assumption (b) and $f < \infty$ ensure that $f \in \left(\frac{1}{\underline{H}}\left(H(P_Y^*) - \log\frac{1}{p_{\max}}\right), \infty\right)$ and that

$$f = \liminf_{n\to\infty} \frac{n+1}{t_{n+1}} \tag{C.40a}$$

$$= \frac{1}{\frac{1}{f} + \limsup_{n\to\infty} \frac{t_{n+1}-t_n}{n+1}}, \tag{C.40b}$$

where (C.40b) is by rewriting (C.40a). Now, (C.39) follows from (C.40a)=(C.40b).

(ii) The DSS has an infinite number of distinct symbol arriving times since $0 < f < \infty$ implies that there exist two positive functions g_1, g_2 with $g_1(n) = \Omega(n)$ and $g_2(n) = O(n)$ such that the symbol arriving time is bounded between $g_1(n) \le t_n \le g_2(n)$, and the symbol arriving interval is constrained by (C.39). □

Lemma 24, stated next, shows the asymptotic behavior of $I(X^{t_k} \to Y^{t_k}|S^k)$ in (C.37).

Lemma 24. *Fix a $(q, \{t_n\}_{n=1}^{\infty})$ DSS that satisfies (a)–(b) and $f < \infty$, and fix a non-degenerate DMC with capacity C (4.10) and the maximum KL divergence C_1*

(4.11). *The instantaneous encoding phase that operates at times* $t = 1, 2, \ldots, t_k$ *in Section 4.3 satisfies*

$$I(X^{t_k} \to Y^{t_k} | S^k) = o(t_k), \tag{C.41}$$

where $\lim_{k\to\infty} \frac{o(t_k)}{t_k} = 0$.

Proof. Appendix C.11. □

Using Lemmas 20, 22–24, we obtain an achievability bound on the expected decoding time $\mathbb{E}[\eta_k]$ (C.36):

$$\mathbb{E}[\eta_k'] \leq \left(\frac{H(S^k|Y^{t_k})}{C} + \frac{\log \frac{1}{\epsilon_k}}{C_1} \right) (1 + o(1)) + t_k \tag{C.42a}$$

$$= \left(\frac{H(S^k) - I(S^k; Y^{t_k})}{C} + \frac{\log \frac{1}{\epsilon_k}}{C_1} \right) (1 + o(1)) + t_k \tag{C.42b}$$

$$= \left(\frac{H(S^k)}{C} + \frac{\log \frac{1}{\epsilon_k}}{C_1} \right) (1 + o(1)) \tag{C.42c}$$

where (C.42a) holds by upper bounding $\mathbb{E}[\eta_k]$ in (C.36) using (C.30) with $P_{S^k} \leftarrow P_{S^k|Y^{t_k}=y^{t_k}}$ and taking an expectation with respect to Y^{t_k}; (C.42b) holds by expanding $H(S^k|Y^{t_k})$ in (C.42a); (C.42c) holds by plugging Lemmas 22 and 24 into $I(S^k; Y^{t_k})$ in (C.42b) and using the fact that $\frac{o(t_k)}{H(S^k)} \leq \frac{o(t_k)}{t_k} \frac{1}{fH} = o(1)$, true due to the assumptions that the entropy rate H and the symbol arriving rate f are both positive. Plugging the achievability bound (C.42) into (4.17), we conclude that the code with instantaneous encoding achieves (4.38).

C.10 Proof of Lemma 22

We first write the mutual information $I(S^k, X^t; Y_t | Y^{t-1})$ in two ways:

$$I(S^k, X^t; Y_t | Y^{t-1}) = I(S^k; Y_t | Y^{t-1}) + I(X^t; Y_t | Y^{t-1}, S^k) \tag{C.43a}$$

$$= I(X^t; Y_t | Y^{t-1}) + I(S^k; Y_t | Y^{t-1}, X^t), \tag{C.43b}$$

where the second term on the right side of (C.43b) is equal to 0 since $Y_t - (Y^{t-1}, X^t) - S^k$ is a Markov chain. Thus,

$$I(S^k; Y_t | Y^{t-1}) = I(X^t; Y_t | Y^{t-1}) - I(X^t; Y_t | Y^{t-1}, S^k). \tag{C.44}$$

We expand $I(S^k; Y^{t_k})$ on the left side of (C.37) as

$$I(S^k; Y^{t_k}) = \sum_{t=1}^{t_k} I(S^k; Y_t | Y^{t-1}) \tag{C.45a}$$

$$= \sum_{t=1}^{t_k} I(X^t; Y_t | Y^{t-1}) - I(X^t; Y_t | Y^{t-1}, S^k) \tag{C.45b}$$

$$= t_k C - I(X^{t_k} \to Y^{t_k} | S^k), \tag{C.45c}$$

where (C.45a) is by the chain rule; (C.45b) is by plugging (C.44) into (C.45a); (C.45c) is by applying the definition of the directed information to the second term of (C.45b) and plugging (4.31) and the fact that Y_i, $i = 1, \ldots, t_k$ are i.i.d. according to P_Y^* into the first term of (C.45b). The channel outputs $Y_1, Y_2, \ldots$ are independent since $Y_t - X_t - Y^{t-1}$ is a Markov chain and X_t is independent of Y^{t-1} (4.31). The channel outputs $Y_1, Y_2, \ldots$ are identically distributed according to P_Y^* since $X_1, X_2, \ldots$ follow the capacity-achieving distribution P_X^* (4.31).

C.11 Proof of Lemma 24

To show (C.41), we first upper bound the conditional directed information in (C.41) as a sum of conditional entropies, and upper bound each conditional entropy by a function of the source prior $\theta_{S^{N(t)}}(Y^{t-1})$. Then, we show that $\theta_{S^{N(t)}}(Y^{t-1})$ converges in probability to zero in time t for $t \in [1, t_k]$ as $k \to \infty$. Finally, we show that the convergence of the source prior leads to the convergence of the entropy sequence and conclude (C.41).

The conditional directed information in (C.41) can be upper bounded as

$$I(X^{t_k} \to Y^{t_k} | S^k) = \sum_{t=1}^{t_k} I(X^t; Y_t | Y^{t-1}, S^k) \tag{C.46a}$$

$$\leq \sum_{t=1}^{t_k} H(X_t | Y^{t-1}, S^k) \tag{C.46b}$$

$$= \sum_{t=1}^{t_k} H(X_t | Z_t, Y^{t-1}), \tag{C.46c}$$

where (C.46a) is by the chain rule, and (C.46c) holds since Z_t is a deterministic function of (Y^{t-1}, S^k) and $X_t - (Z_t, Y^{t-1}) - S^k$ is a Markov chain.

We upper bound each term in the sum of (C.46c) using $\theta_{S^{N(t)}}(Y^{t-1})$. Given that $Z_t = z$, $Y^{t-1} = y^{t-1}$, if $z \in \underline{\mathcal{X}}(y^{t-1})$, we use (4.30) to conclude

$$H(X_t | Z_t = z, Y^{t-1} = y^{t-1}) = 0. \tag{C.47}$$

If $z \in \overline{\mathcal{X}}(y^{t-1})$, we rearrange terms in (4.24) to obtain

$$1 - \frac{P_X^*(z)}{\pi_z(y^{t-1})} \leq 1 - \frac{P_X^*(z)}{P_X^*(z) + \min_{i \in \mathcal{G}_z(y^{t-1})} \theta_i(y^{t-1})} \tag{C.48a}$$

$$\leq \frac{\min_{i \in \mathcal{G}_z(y^{t-1})} \theta_i(y^{t-1})}{\min_{x \in \mathcal{X}} P_X^*(x)}. \tag{C.48b}$$

We upper bound $H(X_t | Z_t = z, Y^{t-1} = y^{t-1})$, $z \in \overline{\mathcal{X}}(y^{t-1})$ by

$$\begin{aligned}
& H(X_t | Z_t = z, Y^{t-1} = y^{t-1}) \\
= \; & \frac{P_X^*(z)}{\pi_z(y^{t-1})} \log \frac{\pi_z(y^{t-1})}{P_X^*(z)} + \sum_{x \in \underline{\mathcal{X}}(y^{t-1})} p_{z \to x} \log \frac{1}{p_{z \to x}} && \text{(C.49a)} \\
\leq \; & \frac{P_X^*(z)}{\pi_z(y^{t-1})} \log \frac{\pi_z(y^{t-1})}{P_X^*(z)} + \left(1 - \frac{P_X^*(z)}{\pi_z(y^{t-1})}\right) \log \frac{|\mathcal{X}| - 1}{1 - \frac{P_X^*(z)}{\pi_z(y^{t-1})}} && \text{(C.49b)} \\
= \; & \left(1 - \frac{P_X^*(z)}{\pi_z(y^{t-1})}\right) \log(|\mathcal{X}| - 1) + h\left(1 - \frac{P_X^*(z)}{\pi_z(y^{t-1})}\right) && \text{(C.49c)} \\
\leq \; & \frac{\min_{i \in \mathcal{G}_z(y^{t-1})} \theta_i(y^{t-1})}{\min_{x \in \mathcal{X}} P_X^*(x)} \log(|\mathcal{X}| - 1) + 2\sqrt{\frac{\min_{i \in \mathcal{G}_z(y^{t-1})} \theta_i(y^{t-1})}{\min_{x \in \mathcal{X}} P_X^*(x)}}, && \text{(C.49d)}
\end{aligned}$$

where (C.49a) holds by (4.27) and (4.30); (C.49b) holds since the sum in the second term on the right side of (C.49a) is maximized if $p_{z \to x}$ is equiprobable on $\underline{\mathcal{X}}(y^{t-1})$, and $|\underline{\mathcal{X}}(y^{t-1})| \leq |\mathcal{X}| - 1$; (C.49c) holds by rearranging terms; (C.49d) holds by applying the upper bound $h(p) \leq 2\sqrt{p}$ to the binary entropy function in (C.49c) and plugging (C.48) into (C.49c). Therefore, each term in (C.46c) is upper bounded as

$$\begin{aligned}
& H(X_t | Z_t, Y^{t-1}) \\
\leq \; & \frac{\log(|\mathcal{X}| - 1)}{\min_{x \in \mathcal{X}} P_X^*(x)} \mathbb{E}\left[\min_{i \in \mathcal{G}_{Z_t}(Y^{t-1})} \theta_i(Y^{t-1})\right] \\
& + \frac{2}{\sqrt{\min_{x \in \mathcal{X}} P_X^*(x)}} \mathbb{E}\left[\sqrt{\min_{i \in \mathcal{G}_{Z_t}(Y^{t_k - 1})} \theta_i(Y^{t-1})}\right] && \text{(C.50a)} \\
\leq \; & \frac{\log(|\mathcal{X}| - 1)}{\min_{x \in \mathcal{X}} P_X^*(x)} \mathbb{E}\left[\theta_{S^{N(t)}}(Y^{t-1})\right] + \frac{2}{\sqrt{\min_{x \in \mathcal{X}} P_X^*(x)}} \mathbb{E}\left[\sqrt{\theta_{S^{N(t)}}(Y^{t-1})}\right] && \text{(C.50b)} \\
\leq \; & \alpha \mathbb{E}\left[\sqrt{\theta_{S^{N(t)}}(Y^{t-1})}\right], && \text{(C.50c)}
\end{aligned}$$

where

$$\alpha \triangleq \max\left\{\frac{\log(|\mathcal{X}| - 1)}{\min_{x \in \mathcal{X}} P_X^*(x)}, \frac{2}{\sqrt{\min_{x \in \mathcal{X}} P_X^*(x)}}\right\}; \tag{C.51}$$

(C.50a) holds by (C.47) and (C.49); (C.50b) holds since $S^{N(t)} \in \mathcal{G}_{Z_t}(Y^{t_k-1})$.

To obtain the asymptotic behavior of $H(X_t|Z_t, Y^{t-1})$ in (C.50), we proceed to analyze the asymptotic behavior of $\theta_{S^{N(t)}}(Y^{t-1})$. The source prior $\theta_{S^{N(t)}}(Y^{t-1})$ in (C.50b) is upper bounded as

$$\theta_{S^{N(t)}}(Y^{t-1}) = P_{S^{N(t)}}(S^{N(t)}) \prod_{j=1}^{t-1} \frac{\sum_{x\in\mathcal{X}} P_{Y|X}(Y_j|x) P_{X_j|Z_j,Y^{j-1}}(x|Z_j, Y^{j-1})}{P_Y^*(Y_j)} \tag{C.52a}$$

$$\leq P_{S^{N(t)}}(S^{N(t)}) \prod_{j=1}^{t-1} \frac{p_{\max}}{P_Y^*(Y_j)}, \tag{C.52b}$$

where (C.52a) holds by (4.22) and (4.32); (C.52b) holds since the numerator in the product term of (C.52a) is upper bounded by $p_{\max}$ (4.33). Given a DSS in Lemma 24 with distinct symbol arriving times $d_{n'}$, $n' = 1, 2, \ldots$ (C.38) (n' is not bounded due to Lemma 23 (ii)), we denote the gap between the symbol arriving rate f and the threshold on the right side of (4.37) by

$$\gamma \triangleq f - \frac{1}{\underline{H}} \left(H(P_Y^*) - \log \frac{1}{p_{\max}} \right) \in (0, \infty). \tag{C.53}$$

For any $t \in [d_{n'}, d_{n'+1})$, $n' = 1, 2, \ldots$, the source prior $\theta_{S^{N(t)}}(Y^{t-1})$ (C.52) satisfies

$$\mathbb{P}\left[\frac{1}{t} \log \theta_{S^{N(t)}}(Y^{t-1}) \leq -\gamma \underline{H} \right] \tag{C.54a}$$

$$\geq \mathbb{P}\left[-\frac{1}{t} \left(\log \frac{1}{P_{S^{N(d_{n'})}}(S^{N(d_{n'})})} \right) + \frac{t-1}{t} \log p_{\max} + \frac{1}{t} \sum_{j=1}^{t-1} \log \frac{1}{P_Y^*(Y_j)} \leq -\gamma \underline{H} \right] \tag{C.54b}$$

$$\geq \mathbb{P}\left[\frac{N(d_{n'})}{d_{n'+1}-1} \left(\frac{1}{N(d_{n'})} \log \frac{1}{P_{S^{N(d_{n'})}}(S^{N(d_{n'})})} \right) \geq \log p_{\max} + H(P_Y^*) + \gamma \underline{H}, \right. \tag{C.54c}$$

$$\left. \frac{t-1}{t} \log p_{\max} + \frac{1}{t} \sum_{j=1}^{t-1} \log \frac{1}{P_Y^*(Y_j)} = \log p_{\max} + H(P_Y^*) \right]$$

$$\geq \mathbb{P}\left[\frac{N(d_{n'})}{d_{n'+1}-1} \left(\frac{1}{N(d_{n'})} \log \frac{1}{P_{S^{N(d_{n'})}}(S^{N(d_{n'})})} \right) \geq \log p_{\max} + H(P_Y^*) + \gamma \underline{H} \right] \tag{C.54d}$$

$$+ \mathbb{P}\left[\frac{t-1}{t} \log p_{\max} + \frac{1}{t} \sum_{j=1}^{t-1} \log \frac{1}{P_Y^*(Y_j)} = \log p_{\max} + H(P_Y^*) \right] - 1 \tag{C.54e}$$

$$\to 1, \tag{C.54f}$$

as $n' \to \infty$, where (C.54b) holds by plugging (C.52b) into (C.54a) and by replacing $N(t) \leftarrow N(d_{n'})$ since $t \in [d_{n'}, d_{n'+1})$; (C.54c) holds since $t \leq d_{n'+1} - 1$ and the event in (C.54b) is implied by the events in (C.54c); (C.54d)–(C.54e) hold by applying Fréchet inequalities [120] to the probability in (C.54c); (C.54f) holds since both probabilities in (C.54d)–(C.54e) converge to 1 as $n' \to \infty$: the probability in (C.54d) converges to 1 as $n' \to \infty$ by Lemma 23 (i), the fact that $\liminf_{n' \to \infty} \frac{N(d_{n'})}{d_{n'}} \geq f$ since $\left\{ \frac{N(d_{n'})}{d_{n'}} \right\}_{n'=1}^{\infty}$ is a subsequence of $\left\{ \frac{n}{t_n} \right\}_{n=1}^{\infty}$, the lower bound on the symbol arriving rate (assumption (b)), the lower bound on the information in $S^{N(d_{n'})}$ (assumption (a)), and the fact that $N(d_{n'}) \to \infty$ as $n' \to \infty$ since $N(d_{n'}) \geq n'$; the probability in (C.54e) converges to 1 since the sum over the logarithms of i.i.d. random variables $Y_1, Y_2, \ldots$ (they are i.i.d. by the argument below (C.45c)) in (C.54e) converges to $H(P_Y^*)$ by the law of large numbers, and $t \to \infty$ as $n' \to \infty$ due to $t \geq d_{n'}$. Rearranging terms in (C.54a), we we conclude that for any $\delta \in (0, 1)$, there exists $n_\delta \in \mathbb{Z}_+$, such that for all $n' \geq n_\delta$, $t \in [d_{n'}, d_{n'+1})$, the probability in (C.56) satisfies

$$\mathbb{P}[\theta_{S^{N(t)}}(Y^{t-1}) \leq e^{-\gamma \underline{H} t}] > 1 - \delta. \tag{C.55}$$

We analyze the asymptotic behavior of $H(X_t|Z_t, Y^{t-1})$ in (C.50) using (C.55). Using the boundedness of the source prior $\theta_{S^{N(t)}}(Y^{t-1}) \in [0, 1]$, we upper bound the expectations in the right side of (C.50c) as

$$\mathbb{E}\left[\sqrt{\theta_{S^{N(t)}}(Y^{t-1})}\right] \leq \mathbb{P}\left[\theta_{S^{N(t)}}(Y^{t-1}) > e^{-\gamma \underline{H} t}\right] + e^{-\frac{\gamma \underline{H}}{2} t} \mathbb{P}\left[\theta_{S^{N(t)}}(Y^{t-1}) \leq e^{-\gamma \underline{H} t}\right] \tag{C.56}$$

$$< \delta + e^{-\frac{\gamma \underline{H}}{2} d_{n_\delta}} (1 - \delta), \ \forall t \in [d_{n'}, d_{n'+1}), \tag{C.57}$$

where (C.57) holds due to (C.55) and the fact that the function $f(p) = p + \beta(1 - p)$, $\beta < 1$, is monotonically increasing on $p \in [0, 1]$.

Plugging (C.57) into (C.50c), we conclude that for all $n' \geq n_\delta$, it holds that

$$H(X_t|Z_t, Y^{t-1}) < \alpha \left(\delta + e^{-\frac{\gamma \underline{H}}{2} d_{n_\delta}} (1 - \delta) \right), \ \forall t \in [d_{n'}, d_{n'+1}). \tag{C.58}$$

We proceed to show (C.41) using (C.58). Dividing both sides of (C.46) by t_k and

taking $k \to \infty$, we upper bound the left side of (C.46) as

$$\begin{aligned}
&\limsup_{k\to\infty} \frac{1}{t_k} I(X^{t_k} \to Y^{t_k}|S^k) \\
\leq\ &\limsup_{k\to\infty} \frac{1}{t_k} \sum_{t=1}^{t_k} H(X_t|Z_t, Y^{t-1}) && \text{(C.59a)} \\
<\ &\limsup_{k\to\infty} \frac{1}{t_k} \left(|t_k - d_{n_\delta}| \alpha \left(\delta + e^{-\frac{\gamma_H}{2} d_{n_\delta}} (1-\delta) \right) + d_{n_\delta} \log |\mathcal{X}| \right) && \text{(C.59b)} \\
=\ &\alpha \left(\delta + e^{-\frac{\gamma_H}{2} d_{n_\delta}} (1-\delta) \right), && \text{(C.59c)}
\end{aligned}$$

where (C.59a) holds by (C.46c); (C.59b) holds by upper bounding $H(X_t|Z_t, Y^{t-1}) \leq \log|\mathcal{X}|$ for $t \leq d_{n_\delta}$ and upper bounding $H(X_t|Z_t, Y^{t-1})$ by (C.58) for $t > d_{n_\delta}$; (C.59c) holds since Lemma 23 (i) implies that $d_{n_\delta} < \infty$ for some $n_\delta \in \mathbb{Z}_+$, and $f < \infty$ implies that $t_k \to \infty$ as $k \to \infty$.

Since δ can be made arbitrarily small while d_{n_δ} can be made arbitrarily large, we conclude (C.41).

C.12 Decoding before the final arrival time

For transmitting the first k source symbols of a $(q, \{t_n\}_{n=1}^\infty)$ DSS with $p_{S,\max} < 1$, we show that if we decode before the final arrival time t_k, then the error probability $\mathbb{P}[S^k \neq \hat{S}_t^k]$, $t < t_k$ will not vanish with k for any code with instantaneous encoding.

For any $t < t_k$, $y^t \in \mathcal{Y}^t$, we lower bound the conditional error probability as

$$\mathbb{P}[S^k \neq \hat{S}_t^k | Y^t = y^t] \geq 1 - \max_{i \in [q]^k} P_{S^k|Y^t}(i|y^t), \tag{C.60}$$

where the equality is attained by the MAP decoder. Taking an expectation of both sides of (C.60), we obtain

$$\begin{aligned}
\mathbb{P}[S^k \neq \hat{S}_t^k] &\geq 1 - \mathbb{E}\left[\max_{i \in [q]^k} P_{S^k|Y^t}(i|Y^t) \right] && \text{(C.61a)} \\
&= 1 - \mathbb{E}\left[\max_{i \in [q]^k} \sum_{j \in [q]^{N(t)}} P_{S^k|S^{N(t)}}(i|j) P_{S^{N(t)}|Y^t}(j|Y^t) \right] && \text{(C.61b)} \\
&\geq 1 - \max_{i \in [q]^k, j \in [q]^{N(t)}} P_{S^k|S^{N(t)}}(i|j) && \text{(C.61c)} \\
&\geq 1 - \prod_{n=N(t)+1}^{k} \max_{s \in [q], s' \in [q]^{n-1}} P_{S_n|S^{n-1}}(s|s') && \text{(C.61d)} \\
&\geq 1 - (p_{S,\max})^{k-N(t)} && \text{(C.61e)}
\end{aligned}$$

$$> 0, \tag{C.61f}$$

where (C.61b) holds since $S^k - S^{N(t)} - Y^t$ is a Markov chain; (C.61c) holds by upper bounding $P_{S^k|S^{N(t)}}(i|j)$ in (C.61b) by its maximum; (C.61d) holds by writing $P_{S^k|S^{N(t)}}(\cdot|\cdot)$ as a product of probabilities $\{P_{S_n|S^{n-1}}(\cdot|\cdot)\}_{n=N(t+1)}^{k}$ and maximizing each term in the product; (C.61e) holds by upper bounding each term in the product by $p_{S,\max}$ (4.35); (C.61f) holds by the assumption $p_{S,\max} < 1$.

C.13 Proof of Remark 2

We show that after the instantaneous encoding phase drops the randomization step (4.25)–(4.30) and only transmits Z_t (4.29) as the channel input, it continues to satisfy the sufficient condition in (4.41) under assumption (b'). To this end, we first write $I(S^k; Y^{t_k})$ in (4.41) as a sum of mutual informations. Then, we show that all source priors converge pointwise to zero in time during the symbol arriving period $[1, t_k]$ as $k \to \infty$; this implies that group priors converge pointwise to the capacity-achieving probabilities. Finally, we show that the convergence of the group priors implies that the summands of $I(S^k; Y^{t_k})$ converge to the capacity C and conclude (4.41). Given channel outputs $y^{t-1} \in \mathcal{Y}^{t-1}$, we denote the source sequence in $[q]^{N(t)}$ that has the maximum source prior by

$$i^* \triangleq \underset{i \in [q]^{N(t)}}{\arg\max}\, \theta_i(y^{t-1}). \tag{C.62}$$

To expand $I(S^k; Y^{t_k})$ in (4.41), we first notice that (C.43)–(C.45b) continue to hold, thus $I(S^k; Y^{t_k})$ is equal to (C.45b). The second term on the right side of (C.45b) is equal to zero since X_t is a deterministic function of (Y^{t-1}, S^k), thus,

$$I(S^k; Y^{t_k}) = \sum_{t=1}^{t_k} I(X_t; Y_t | Y^{t-1}). \tag{C.63}$$

We proceed to analyze the asymptotic behavior of $\theta_{i^*}(y^{t-1})$ (C.62). Since the encoder drops the randomization step (4.25)–(4.30) and only transmits Z_t (4.29) as the channel input, the posterior update (4.32) becomes (4.45). Upper bounding $P_{S^{N(t)}|S^{N(t-1)}}(\cdot|\cdot)$ in the prior update (4.22) by the maximum symbol arriving probability $p_{S,\max}^{N(t)-N(t-1)}$ (4.35), and upper bounding the numerator by $p_{\max}$ and the denominator by $p_{\min}$ in the fraction on the right side of (4.45), we obtain an upper bound on the source prior $\theta_{i^*}(y^{t-1})$ as

$$\theta_{i^*}(y^{t-1}) \le p_{S,\max}^{N(t)} \left(\frac{p_{\max}}{p_{\min}}\right)^{t-1} \tag{C.64}$$

for all $i \in [q]^{N(t)}$, $y^{t-1} \in \mathcal{Y}^{t-1}$. Given a DSS that satisfies assumption (b') with $f < \infty$ and distinct symbol arriving times $d_{n'}$, $n' = 1, 2, \ldots$ (C.38) (n' is not bounded due to Lemma 23), similar to (C.53), we denote the gap between the symbol arriving rate f and the threshold in assumption (b') by

$$\gamma' \triangleq f - \frac{1}{\log \frac{1}{p_{S,\max}}} \left(\log \frac{1}{p_{\min}} - \log \frac{1}{p_{\max}} \right). \tag{C.65}$$

For any $t \in [d_{n'}, d_{n'+1})$, $n' = 1, 2, \ldots$, the source prior $\theta_{i^*}(y^{t-1})$ for any $i^* \in [q]^{N(t)}$, $y^{t-1} \in \mathcal{Y}^{t-1}$ in (C.64) satisfies

$$\begin{aligned}
\limsup_{n' \to \infty} \frac{1}{t} \log \theta_{i^*}(y^{t-1}) &\leq -\left(\liminf_{n' \to \infty} \frac{N(t)}{t} \log \frac{1}{p_{S,\max}} \right) + \log \frac{p_{\max}}{p_{\min}} && \text{(C.66a)} \\
&\leq -\left(\liminf_{n' \to \infty} \frac{N(d_{n'})}{d_{n'+1} - 1} \log \frac{1}{p_{S,\max}} \right) + \log \frac{p_{\max}}{p_{\min}} && \text{(C.66b)} \\
&\leq -f \log \frac{1}{p_{S,\max}} + \log \frac{p_{\max}}{p_{\min}} && \text{(C.66c)} \\
&= -\gamma' \log \frac{1}{p_{S,\max}}, && \text{(C.66d)}
\end{aligned}$$

where (C.66a) is by taking the logarithm, dividing by t, and taking n' to infinity on both sides of (C.64); (C.66b) holds since $\frac{N(d_{n'})}{d_{n'+1}-1} \leq \frac{N(t)}{t}$ for all $t \in [d_{n'}, d_{n'+1})$; (C.66c) holds due to Lemma 23 (i) and the fact that $\left\{ \frac{N(d_{n'})}{d_{n'}} \right\}_{n'=1}^{\infty}$ is a subsequence of $\left\{ \frac{n}{t_n} \right\}_{n=1}^{\infty}$; (C.66d) holds by plugging (C.65) into (C.66c). Rearranging terms of (C.66), we conclude that the maximum source prior (C.62) converges pointwise: for any $y^{t-1} \in \mathcal{Y}^{t-1}$,

$$\lim_{n' \to \infty} \theta_{i^*}(y^{t-1}) = 0, \ \forall t \in [d_{n'}, d_{n'+1}), \tag{C.67}$$

where $t \to \infty$ for any $t \in [d_{n'}, d_{n'+1})$ as $n' \to \infty$.

The convergence of the source prior (C.67) implies the convergence of the group prior. The partitioning rule in (4.24) ensures that the group prior $\pi_x(y^{t-1})$, $\forall x \in \mathcal{X}$ is simultaneously upper and lower bounded as

$$\begin{aligned}
P_X^*(x) + \theta_{i^*}(y^{t-1}) &\geq \pi_x(y^{t-1}) && \text{(C.68a)} \\
&\geq P_X^*(x) - |\mathcal{X}| \theta_{i^*}(y^{t-1}), && \text{(C.68b)}
\end{aligned}$$

where the upper bound (C.68a) holds by (4.24) and (C.62); the lower bound (C.68b) holds since all $|\mathcal{X}|$ group priors are upper bounded by (C.68a). From (C.67) and (C.68), we conclude that for all $x \in \mathcal{X}$, $y^{t-1} \in \mathcal{Y}^{t-1}$,

$$\lim_{n' \to \infty} \pi_x(y^{t-1}) = P_X^*(x), \ t \in [d_{n'}, d_{n'+1}). \tag{C.69}$$

Next, we show the convergence of the group prior (C.69) implies the convergence of the mutual information $I(X_t; Y_t|Y^{t-1})$ in the sum of (C.63). We expand the mutual information $I(X_t; Y_t|Y^{t-1})$ as

$$I(X_t; Y_t|Y^{t-1}) = \tag{C.70}$$
$$\sum_{y^{t-1}\in\mathcal{Y}^{t-1}} P_{Y^{t-1}}(y^{t-1}) \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)\pi_x(y^{t-1}) \log \frac{P_{Y|X}(y|x)}{\sum_{x'\in\mathcal{X}} P_{Y|X}(y|x')\pi_{x'}(y^{t-1})},$$

which achieves the channel capacity C if $\pi_x(y^{t-1}) = P_X^*(x)$ for all $x \in \mathcal{X}$, $y^{t-1} \in \mathcal{Y}^{t-1}$. Using (C.69) and (C.70), we conclude

$$\lim_{n'\to\infty} I(X_t; Y_t|Y^{t-1}) = C,\ t \in [d_{n'}, d_{n'+1}). \tag{C.71}$$

Since $I(X_t; Y_t|Y^{t-1}) \le C$, one can write the equivalent of (C.71) as: for all $\epsilon > 0$, there exists an $n_\epsilon \in \mathbb{N}$, such that for all $n' \ge n_\epsilon$, it holds that

$$I(X_t; Y_t|Y^{t-1}) > C - \epsilon,\ \forall t \in [d_{n'}, d_{n'+1}). \tag{C.72}$$

We proceed to show (4.41) using (C.63) and (C.72). Dividing both sides of (C.63) by t_k and taking $k \to \infty$, we lower bound the left side of (C.63) as

$$\lim_{k\to\infty} \frac{1}{t_k} I(S^k; Y^{t_k}) = \lim_{k\to\infty} \frac{1}{t_k} \sum_{t=1}^{t_k} I(X_t; Y_t|Y^{t-1}) \tag{C.73a}$$
$$> \lim_{k\to\infty} \frac{1}{t_k}(t_k - d_{n_\epsilon})(C - \epsilon) \tag{C.73b}$$
$$= C - \epsilon, \tag{C.73c}$$

where (C.73b) holds by lower bounding $I(X_t; Y_t|Y^{t-1})$ by (C.72) for $t > d_{n_\epsilon}$, and lower bounding $I(X_t; Y_t|Y^{t-1})$ by zero for $t \le d_{n_\epsilon}$.

Since ϵ in (C.73c) can be made arbitrarily small, and $\lim_{k\to\infty} \frac{1}{t_k} I(S^k; Y^{t_k}) \le C$ by data processing, we conclude by the squeeze theorem that under assumption (b'), the instantaneous encoding phase satisfies (4.41) even if it does not randomize the channel input.

C.14 The approximating instantaneous SED rule ensures (4.67)

We show that the approximating instantaneous SED rule in step (iii′) ensures (4.67). Since the left side of (4.67) is equal to the minimum value on the right side of (4.66a), it suffices to show that the latter is upper bounded by $\theta_{S_j}(y^{t-1})$.

We denote

$$c_n \triangleq (\pi_0(y^{t-1}) - n\theta_{\mathcal{S}_j}(y^{t-1})) - (\pi_1(y^{t-1}) + n\theta_{\mathcal{S}_j}(y^{t-1})) \tag{C.74a}$$

$$= 2\pi_0(y^{t-1}) - 1 - 2n\theta_{\mathcal{S}_j}(y^{t-1}), \tag{C.74b}$$

and we rewrite the minimization problem in (4.66a) as

$$\min_{n\in\{\underline{n},\bar{n}\}} |c_n|. \tag{C.75}$$

By definitions of $\underline{n}$ (4.66b) and $\bar{n}$ (4.66c), it holds that $\bar{n} - \underline{n} = 1$. Thus

$$c_{\underline{n}} - c_{\bar{n}} = 2\theta_{\mathcal{S}_j}(y^{t-1}). \tag{C.76}$$

Since $c_{\underline{n}} \geq 0$ and $c_{\bar{n}} \leq 0$, we conclude from (C.76) that

$$\min\{c_{\underline{n}}, |c_{\bar{n}}|\} \leq \theta_{\mathcal{S}_j}(y^{t-1}), \tag{C.77}$$

which means that (C.75) is upper bounded by $\theta_{\mathcal{S}_j}(y^{t-1})$.

C.15 Number of types for random arrivals

We aim to show that the number of types at time t is $O(t^2)$. We define the following notations. We use B or A as the index of a random variable to signify that the random variable is obtained *before* or *after* the group partitioning (step (iii‴)). We denote by $\Delta_B(t)$ and $\Delta_A(t)$ the number of existing types at time t before and after the encoder and the decoder partition $\mathcal{G}_0(y^{t-1})$ and $\mathcal{G}_1(y^{t-1})$, respectively. We denote by $\mathcal{S}^*(t)$ the last type moved to $\mathcal{G}_0(y^{t-1})$ at time t so that after $\mathcal{S}^*(t)$ is moved to $\mathcal{G}_0(y^{t-1})$, the group prior $\tilde{\pi}_0(y^{t-1})$ exceeds 0.5 for the first time (step (iii‴)). We denote by $\mathcal{W}_B(t)$ the set that contains all the new types created at time t right before the encoder and the decoder partition $\mathcal{G}_0(y^{t-1})$ and $\mathcal{G}_1(y^{t-1})$ (see step (i‴), e.g., $\mathcal{W}_B(1) = \{\mathcal{S}_1, \mathcal{S}_2\}$, $\mathcal{W}_B(2) = \{\mathcal{S}_3, \mathcal{S}_4\}$). After the group partitioning (step (iii‴)), some type in $\mathcal{W}_B(t)$ may be split. Let $\mathcal{W}_A(t)$ be the set that consists of all subsets of the split types in $\mathcal{W}_B(t)$ and all unsplit types in $\mathcal{W}_B(t)$ after the group partitioning.

We show by heuristic analysis that the average number of types at time $t+1$ is upper bounded as

$$\mathbb{E}[\Delta_B(t+1)] \leq \frac{2-\delta}{2}t^2 + \left(3 - \frac{\delta}{2}\right)t + \delta,\ t \geq 1, \tag{C.78}$$

$$\mathbb{E}[\Delta_A(t+1)] \leq \mathbb{E}[\Delta_B(t+1)] + (1-\delta)t + 1, \tag{C.79}$$

where δ is the bit arrival probability in (4.68).

We define a sequence of events $\mathcal{E}_t$, $t = 1, 2, 3, \ldots$ as

$$\mathcal{E}_1 \triangleq \{\mathcal{S}^*(1) \text{ is not split}\}. \tag{C.80}$$

$$\mathcal{E}_t \triangleq \{\mathcal{S}^*(t) \text{ is split}\} \cap \{\text{sequences in } \mathcal{S}^*(t) \text{ are of length } \min(\lfloor \delta t \rfloor, k)\},$$
$$t > 1. \tag{C.81}$$

Since at $t = 1$, $|\mathcal{S}_1| = |\mathcal{S}_2| = 1$, we have $\mathbb{P}[\mathcal{E}_1] = 1$. Extensive simulations on the evolution of types show that

$$\mathbb{P}[\mathcal{E}_t^c] \ll 1, t = 2, 3, \ldots \tag{C.82}$$

In the heuristic analysis that follows, we assume

$$\mathbb{P}[\mathcal{E}_t] = 1, t = 1, 2, \ldots \tag{C.83}$$

We will use rigorous analyses a)–d) below together with the assumption in (C.83) to justify (C.78)–(C.79). The analyses b)–d) below solely follow the construction of type-set instantaneous SED codes. The type update method in step (i‴) Section 4.7 implies:

a) The binary sequences in a type are of the same length and can be ordered in a consecutive lexicographic order, e.g., $\mathcal{S}_3 = \{00, 01\}$. To ensure that each type has only one parent, once a type is fixed to be split, among all its child types[2], at most one of them needs to be split accordingly. This is due to the reason that follows. Without loss of generality, we assume that we split an arbitrary type $\mathcal{S}_i$ that contains m binary sequences $s_1, s_2, \ldots, s_m$ sorted in a lexicographic order. The type-based SED rule (step (iii‴)) cuts $\mathcal{S}_i$ between s_{n^*} and s_{n^*+1} to split it. Among all child types of $\mathcal{S}_i$, only the type that contains both $s_{n^*} \boxplus 1$ and $s_{n^*+1} \boxplus 0$ will need to be split accordingly, and at most one type $\mathcal{S}_j$ contains both two sequences simultaneously. Recursively, due to the split of $\mathcal{S}_j$, the encoder and the decoder at most further split one child type of $\mathcal{S}_j$. The recursion stops if the split type has no child types, or if after the split of a type $\mathcal{S}_i$, no child types of $\mathcal{S}_i$ contain $s_{n^*} \boxplus 1$ and $s_{n^*+1} \boxplus 0$ simultaneously.

[2]We call $\mathcal{S}_j$ a child type of $\mathcal{S}_i$ if $\mathcal{S}_i$ is the parent of $\mathcal{S}_j$. According to type update method in step (i‴), any type $\mathcal{S}_i$ at most generates 1 type $\mathcal{S}_j$. If $\mathcal{S}_j$ is never split, $\mathcal{S}_i$ has only one child type $\mathcal{S}_j$. Yet, if $\mathcal{S}_j$ is split during the group partitioning, $\mathcal{S}_i$ has multiple child types.

b) For $t+1 \leq k$, only the types in $\mathcal{W}_A(t)$ generate new types at time $t+1$ in step (i‴) right before the group partitioning (step (iii‴)). These new types form $\mathcal{W}_{t+1,B}$. For $t+1 > k$, $\mathcal{W}_{t+1,B}$ and $\mathcal{W}_{t+1,A}$ are empty since new types are no longer generated at each time $t+1$ before the group partitioning.

c) By the definition of $\mathcal{W}_A(t)$, at time t, if a type to be split is originally in $\mathcal{W}_{t,B}$, the two split subsets are in $\mathcal{W}_A(t)$. The binary sequences of any types in $\mathcal{W}_B(t)$ are of length $\max(t, n)$.

Using (C.81) and a), we know:

d) Given $\mathcal{E}_t$ (C.81) occurs, after the split of $\mathcal{S}^*(t)$, the encoder and decoder at most further split $t - \lfloor \delta t \rfloor$ types. This is because (a) implies that given the split of $\mathcal{S}^*(t)$, the encoder and the decoder further split recursively at most 1 type of string length equal to $\lfloor \delta t \rfloor + 1, \lfloor \delta t \rfloor + 2, \ldots, \max(t, k)$.

Using b)–d), we conclude that the encoder and the decoder at most split one type in $\mathcal{W}_B(t)$, and the average cardinality of $\mathcal{W}_B(t), t = 1, 2, \ldots$ evolves as

$$\mathbb{E}[|\mathcal{W}_B(t+1)||\mathcal{E}_t] \leq \mathbb{E}[|\mathcal{W}_B(t)||\mathcal{E}_t] + 1, \ t \geq 2, \tag{C.84a}$$

$$|\mathcal{W}_B(1)| = |\mathcal{W}_B(2)| = 2, \tag{C.84b}$$

where (C.84b) holds since no type is split at $t = 1$. Using b) and d), we conclude that given $\mathcal{E}_t$, the average number of types evolves as

$$\mathbb{E}[N_B(t+1)|\mathcal{E}_t] \leq \mathbb{E}[\Delta_B(t)|\mathcal{E}_t] + 1 + t - \lfloor qt \rfloor + \mathbb{E}[|\mathcal{W}_B(t+1)||\mathcal{E}_t], \tag{C.85a}$$

$$\Delta_B(1) = 2. \tag{C.85b}$$

Replacing $\lfloor \delta t \rfloor$ by δt in (C.85a), plugging (C.84) into (C.85a), and using (C.83), we obtain (C.78). Using (C.83), d), and (C.78), we obtain (C.79).

Simulation results confirm our heuristic analysis. The fitting curves (C.78) in Fig. C.1 increase at a similar speed as the simulated curves, indicating that the heuristic expressions in (C.78)–(C.79) are meaningful gauges of the average number of types. The fitting curves in Fig. C.1 are slightly larger than the simulated curves since (C.78)–(C.79) are upper bounds of $\mathbb{E}[\Delta_B(t+1)]$ and $\mathbb{E}[\Delta_A(t+1)]$.

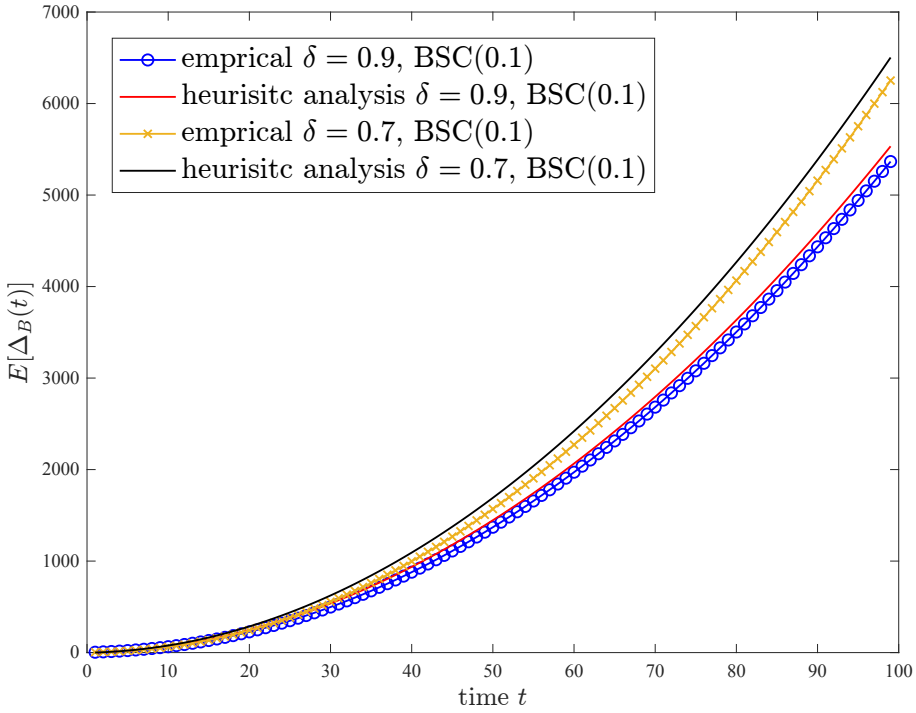


Figure C.1: Average number of types $\mathbb{E}[\Delta_B(t)]$ versus time t over a BSC(0.9). The bit arrival probability is δ in (4.68). The source length $k = 99$. The curves by heuristic analysis are plotted as (C.78). We only present curves for BSC(0.9), since according to our heuristic analysis, the upper bounds to the average number of types (C.78)–(C.79) are not functions of the crossover probability of the BSC.

C.16 Converse proof of Theorem 11

In the converse proof, we show that the converse bounds on the JSCC reliability function for a fully accessible source apply, i.e.,

$$\tilde{E}(R) \leq E(R). \tag{C.86}$$

Every code with instantaneous encoding for random arrivals (Definition 22) whose decoder does not know the symbol arriving times is a special code with instantaneous encoding whose decoder knows the random symbol arriving times, since the decoder in the latter code can choose to use the arriving times or not. Thus, converse bounds on the error exponent of the latter code applies to $\tilde{E}(R)$. We denote by η_k the stopping time of the latter code that ensures error probability ϵ_k. The latter code is indeed a set of codes with instantaneous encoding (Definition 20) for deterministic arrivals, each corresponds to a sequence of realizations of stopping times $\tau^k = t^k \in \mathbb{Z}_+^k$. For deterministic arriving times t^k, we denote by $\eta_k(t^k)$ the stopping time of a code with instantaneous encoding, and we denote by $\epsilon_k(t^k)$ the corresponding error probability

at the stopping time. It holds that

$$\mathbb{E}[\eta_k] = \mathbb{E}[\eta_k(\tau^k)] \tag{C.87}$$

$$\epsilon_k = \mathbb{E}[\epsilon_k(\tau^k)]. \tag{C.88}$$

Theorem 9 implies that the error exponent of every code with instantaneous encoding for deterministic arriving times t^k is upper bounded as

$$-\lim_{k\to\infty} \frac{\log \epsilon_k(t^k)}{\mathbb{E}[\eta_k(t^k)]} \leq E(R). \tag{C.89}$$

Therefore, the error exponent of the code whose decoder knows the random symbol arriving times satisfies (C.89)

$$-\lim_{k\to\infty} \frac{\log \epsilon_k}{\mathbb{E}[\eta_k]} = -\lim_{k\to\infty} \frac{\log \mathbb{E}[\epsilon_k(\tau^k)]}{\mathbb{E}[\eta_k(\tau^k)]} \tag{C.90}$$

$$\leq -\lim_{k\to\infty} \frac{\mathbb{E}[\log \epsilon_k(\tau^k)]}{\mathbb{E}[\eta_k(\tau^k)]} \tag{C.91}$$

$$\leq E(R), \tag{C.92}$$

where (C.91) holds by applying Jensen's inequality to the numerator of (C.90) since $\log(\cdot)$ is a concave function; (C.92) holds by plugging the upper bound on $-\log \epsilon_k(t^k)$ in (C.89) into the numerator of (C.91).

C.17 Achievability proof of Theorem 11

Fixing a DSS with random arrivals that satisfies assumptions (c)–(e), we show that $\tilde{E}(R)$ is achievable by the instantaneous SED code for random arrivals in Section 4.6. To this end, we show that the instantaneous SED code leads to Lemmas 25–26 stated below, and we use Lemmas 25–26 together with Lemma 20 to establish an achievability (upper) bound on the expected stopping time of the instantaneous SED code. We denote

$$\mathsf{d}(k) \triangleq k + \mathsf{h}(k). \tag{C.93}$$

Lemma 25. *Fix a non-degenerate symmetric binary-input DMC, and fix a DSS with random arrivals that satisfies* (c)*. The instantaneous SED code for random arrivals in Section 4.6 satisfies*

$$H(S^k|Y^{\mathsf{d}(k)}) = H(S^k) - I(X^{\mathsf{d}(k)} \to Y^{\mathsf{d}(k)}) + I(\tau^k; Y^{\mathsf{d}(k)}|S^k). \tag{C.94}$$

Proof. Appendix C.18. □

Lemma 26. *Fix a non-degenerate symmetric binary-input DMC with channel capacity C, and fix a DSS with random arrivals that satisfies assumptions* (c) *and* (e). *The instantaneous SED code for random arrivals in Section 4.6 satisfies*

$$\lim_{k\to\infty} \frac{1}{k} I(X^{\mathsf{d}(k)} \to Y^{\mathsf{d}(k)}) = C. \tag{C.95}$$

Proof. Appendix C.19. □

Since the k-th symbol must have arrived at the encoder by time $\mathsf{d}(k)$ by assumption (c), the instantaneous SED code in Section 4.6 that operates at time $t > \mathsf{d}(k)$ reduces to the SED code [45, Sec. V-B] for transmitting k symbols S^k of a (fully accessible) DS in $[q]^k$ with prior $P_{S^k|Y^{\mathsf{d}(k)}}$. Since the SED code is a JSCC reliability function (4.38)-achieving code for symmetric binary-input DMCs, we invoke Lemma 20 to conclude that the stopping time η_k of the instantaneous SED code and the error probability ϵ at the stopping time satisfy

$$\mathbb{E}[\eta_k] \le \left(\frac{H(S^k|Y^{\mathsf{d}(k)})}{C} + \frac{\log \frac{1}{\epsilon}}{C_1} \right) (1 + o(1)) + \mathsf{d}(k). \tag{C.96}$$

Plugging (C.96) into (4.54) and rearranging terms, we obtain an achievability bound on $\tilde{E}(R)$,

$$\tilde{E}(R) \ge C_1 \left(1 - \left(\limsup_{k\to\infty} \frac{H(S^k|Y^{\mathsf{d}(k)})}{kC} + \frac{\mathsf{d}(k)}{k} \right) R \right). \tag{C.97}$$

Using Lemmas 25–26 and (C.96), we proceed to calculate the conditional mutual information in (C.97) as

$$\lim_{k\to\infty} \frac{H(S^k|Y^{\mathsf{d}(k)})}{kC} = \lim_{k\to\infty} \frac{1}{kC} \left(H(S^k) - I(X^{\mathsf{d}(k)} \to Y^{\mathsf{d}(k)}) + I(\tau^k; Y^{\mathsf{d}(k)}|S^k) \right) \tag{C.98a}$$

$$= \frac{H}{C} - 1 + \lim_{k\to\infty} \frac{1}{kC} I(\tau^k; Y^{\mathsf{d}(k)}|S^k) \tag{C.98b}$$

$$\le \frac{H}{C} - 1 + \lim_{k\to\infty} \frac{1}{kC} H(\tau^k) \tag{C.98c}$$

$$\le \frac{H}{C} - 1, \tag{C.98d}$$

where (C.98a) holds by Lemma 25; (C.98b) holds by the definition of the entropy rate (4.2) and Lemma 26; (C.98c) holds by upper bounding the conditional mutual information in (C.98b) by the entropy in (C.98c); (C.98d) holds by assumption (d). Plugging the lower bound in (C.98) into the right side of (C.97) and using the fact $\lim_{k\to\infty} \frac{\mathsf{d}(k)}{k} = 1$, we obtain (4.59).

C.18 Proof of Lemma 25

The left side of (C.94) is equal to

$$H(S^k|Y^{\mathsf{d}(k)}) = H(S^k) - I(S^k; Y^{\mathsf{d}(k)}) \tag{C.99a}$$
$$= H(S^k) - I(S^k, \tau^k; Y^{\mathsf{d}(k)}) - I(\tau^k; Y^{\mathsf{d}(k)}|S^k), \tag{C.99b}$$

where (C.99b) holds by the chain rule of mutual information. We proceed to show that the second term in (C.99b) is equal to the second term on the right side of (C.94). We write the following conditional mutual information in two ways

$$I(S^k, \tau^k, X_t; Y_t|Y^{t-1}) \tag{C.100a}$$
$$= I(S^k, \tau^k; Y_t|Y^{t-1}) + I(X_t; Y_t|Y^{t-1}, S^k, \tau^k) \tag{C.100b}$$
$$= I(X_t; Y_t|Y^{t-1}) + I(S^k, \tau^k; Y_t|Y^{t-1}, X_t), \tag{C.100c}$$

where the second term in (C.100b) is equal to 0 since X_t is a deterministic function of (Y^{t-1}, S^k, τ^k) by (4.29); the second term in (C.100c) is equal to 0 since $Y_t - (Y^{t-1}, X_t) - (S^k, \tau^k)$ is a Markov chain. Thus, we conclude

$$I(S^k, \tau^k; Y_t|Y^{t-1}) = I(X_t; Y_t|Y^{t-1}). \tag{C.101}$$

Using (C.101), we write the second term in (C.99b) as

$$I(S^k, \tau^k; Y^{\mathsf{d}(k)}) = \sum_{t=1}^{\mathsf{d}(k)} I(S^k, \tau^k; Y_t|Y^{t-1}) \tag{C.102}$$
$$= I(X^{\mathsf{d}(k)} \to Y^{\mathsf{d}(k)}). \tag{C.103}$$

Plugging (C.102) into (C.99b), we obtain (C.94).

C.19 Proof of Lemma 26

To show (C.95), we first expand the left side of (C.95) as a sum of mutual informations. We show that the source prior converges pointwise to zero in time $t \in [1, k]$ as $k \to \infty$ under assumption (e), and conclude that each term in the sum converges as the source prior converges. Finally, we show that the convergence of the summands implies the convergence of (C.95).

The left side of (C.95) can be expanded as

$$I(X^{\mathsf{d}(k)} \to Y^{\mathsf{d}(k)}) = \sum_{t=1}^{\mathsf{d}(k)} I(X_t; Y_t|Y^{t-1}). \tag{C.104}$$

At time $t \leq k$, the prior $\theta_i(y^{t-1})$ (4.55) for all $i \in \mathcal{Q}_t$ is upper bounded as

$$\theta_i(y^{t-1}) = P_{S^{N(t)}|S^{N(t-1)}}(i|i)\rho_i(y^{t-1}) + P_{S^{N(t)}|S^{N(t-1)}}(i|i\boxminus)\rho_{i\boxminus}(y^{t-1}) \tag{C.105a}$$

$$\leq \tilde{p}_{S,\max} \max\{\rho_i(y^{t-1}), \rho_{i\boxminus}(y^{t-1})\}, \tag{C.105b}$$

where (C.105b) holds by the definition of $\tilde{p}_{S,\max}$ in (4.56). The posterior $\rho_i(y^{t-1})$ (4.32) for all $i \in \mathcal{Q}_t$ is upper bounded as

$$\rho_i(y^{t-1}) \leq \frac{p_{\max}}{p_{\min}} \theta_i(y^{t-1}). \tag{C.106}$$

From (C.105) and (C.106), we conclude that at time $t \leq k$, the prior is upper bounded as

$$\theta_i(y^{t-1}) \leq \left(\tilde{p}_{S,\max} \frac{p_{\max}}{p_{\min}}\right)^t, \ \forall i \in \mathcal{Q}_t. \tag{C.107}$$

Plugging the upper bound (C.107) into the right side of the instantaneous SED rule (4.44), we obtain for all $y^{t-1} \in \mathcal{Y}^{t-1}$ and $x \in \{0, 1\}$,

$$\left|\pi_x(y^{t-1}) - P_X^*(x)\right| \leq \left(\tilde{p}_{S,\max} \frac{p_{\max}}{p_{\min}}\right)^t. \tag{C.108}$$

Since each term $I(X_t; Y_t|Y^{t-1})$ in (C.104) can be written as (C.70) and is equal to C if $\pi_x(y^{t-1}) = P_X^*(x)$ for all $y^{t-1} \in \mathcal{Y}^{t-1}$, we conclude from (C.108) that there exists a function $f\colon \mathbb{Z}_+ \to \mathbb{R}_+$ of time t that satisfies $f(t) \to 0$ and

$$I(X_t; Y_t|Y^{t-1}) \geq C - f(t). \tag{C.109}$$

We proceed to show (C.95). Dividing both sides of (C.104) by k and taking $k \to \infty$, we lower bound the left side of (C.104) as

$$\lim_{k\to\infty} \frac{1}{k} I(X^{\mathsf{d}(k)}; Y^{\mathsf{d}(k)}) = \lim_{k\to\infty} \frac{1}{k} \sum_{t=1}^{\mathsf{d}(k)} I(X_t; Y_t|Y^{t-1}) \tag{C.110a}$$

$$\geq \lim_{k\to\infty} \frac{1}{k} \sum_{t=1}^{k} I(X_t; Y_t|Y^{t-1}) \tag{C.110b}$$

$$\geq \lim_{k\to\infty} \frac{1}{k} \sum_{t=1}^{k} C - f(t) \tag{C.110c}$$

$$= C, \tag{C.110d}$$

where (C.110a) holds by (C.104); (C.110b) holds since the mutual information is non-negative and $\mathsf{d}(k) \geq k$; (C.110c) holds by (C.109); (C.110d) holds by applying the Cesàro mean [121] to the converging sequence $f(t)$. Since the left side of (C.110a) is also upper bounded by C, we conclude (C.95).

C.20 Zero entropy rate of symbol arriving times

The symbol arriving times in the example satisfy (4.58) since the entropy is upper bounded as

$$H\left(\tau^n\right) = \log \binom{\mathsf{h}(n)}{\mathsf{h}'(n)} \tag{C.111}$$

$$\leq \log \left(e\frac{\mathsf{h}(n)}{\mathsf{h}'(n)}\right)^{\mathsf{h}'(n)} \tag{C.112}$$

$$= \mathsf{h}'(n)\left(1 + \log \frac{\mathsf{h}(n)}{\mathsf{h}'(n)}\right) \tag{C.113}$$

$$\leq \mathsf{h}'(n) + \mathsf{h}(n), \tag{C.114}$$

where (C.111) holds since there are in total $\binom{\mathsf{h}(n)}{\mathsf{h}'(n)}$ possible realizations of the random symbol arriving times and the entropy function is concave; (C.112) holds by applying the inequality $\binom{n}{k} \leq \left(e\frac{n}{k}\right)^k$ to (C.111); (C.114) holds by applying the inequalities $\log x < \log(1+x) < x$ for all $x > 0$ to the logarithm in (C.113).

C.21 Proof of Proposition 4

We show that the achievability bound in (4.61) holds. Since (C.97) in Appendix C.17 holds for any DSS whose random symbol arriving times are bounded as $\tau_n \leq \mathsf{d}(n), n = 1, 2, \ldots$, we replace $\mathsf{d}(k) \leftarrow \mathbb{E}[\tau_k] + \mathsf{h}(k)$ in (C.97) and obtain

$$\tilde{E}(R) \geq C_1 \left(1 - \left(\limsup_{k\to\infty} \frac{H\left(S^k | Y^{\mathbb{E}[\tau_k]+\mathsf{h}(k)}\right)}{kC} + \frac{\mathbb{E}[\tau_k]}{k}\right) R\right), \tag{C.115}$$

where $Y_1, Y_2, \ldots$ are the channel outputs in response to the channel inputs generated by the encoder of the instantaneous SED code in Section 4.6.

C.22 Cardinality of common randomness

We adapt the proof in [41, Theorem 19] to our codes with instantaneous encoding to show that for any $\langle k, R, \epsilon\rangle$ code with instantaneous encoding that allows $|\mathcal{U}| = \infty$, there exists a $\langle k, R, \epsilon\rangle$ code with instantaneous encoding that allows $|\mathcal{U}| \leq 2$. Fixing a source length k, for $u = 1, 2, \ldots, \infty$, we define $\mathcal{G}_u \subseteq \mathbb{R}^2$ as

$$\mathcal{G}_u \triangleq \{(R, \epsilon) : \exists\ \langle k, R, \epsilon\rangle \text{ code with instantaneous encoding that allows } |\mathcal{U}| \leq u\}. \tag{C.116}$$

We show that $\mathcal{G}_1$ is a connected set. To see this, we arbitrarily select two elements in $\mathcal{G}_1$, denoted by $\Lambda_1 \triangleq (R_1, \epsilon_1)$ and $\Lambda_2 \triangleq (R_2, \epsilon_2)$. We denote $\Lambda_3 \triangleq (\min\{R_1, R_2\}, \max\{\epsilon_1, \epsilon_2\})$. According to the rate and the error constraints in

(4.14)–(4.15), $\Lambda_i \in \mathcal{G}_1$, $i \in \{1, 2\}$, indicates that all elements (R, ϵ) that simultaneously satisfy $R \leq R_i$ and $\epsilon \geq \epsilon_i$ belong to $\mathcal{G}_1$ (see the shaded region in Fig. C.2). As a result, the line segments $L_i \triangleq \{\lambda\Lambda_i + (1-\lambda)\Lambda_3, \lambda \in [0, 1]\}$, $i = 1, 2$, belong to $\mathcal{G}_1$, and the arc $L_1 \cup L_2$ joins Λ_1 and Λ_2.

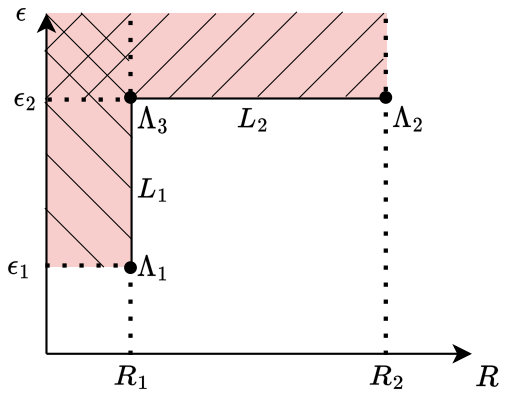


Figure C.2: Elements Λ_1 and Λ_2 are jointed by the arc $L_1 \cup L_2$.

Since $\mathcal{G}_1 \subseteq \mathbb{R}^2$, $\mathcal{G}_1$ is a connected set, and $\mathcal{G}_\infty$ is a convex hull of $\mathcal{G}_1$, by Fenchel-Eggleston-Carathéodory's theorem for connected sets [122, Theorem 18(ii)], any element in $\mathcal{G}_\infty$ can be represented as a convex combination of 2 elements in $\mathcal{G}_1$, in other words, $\mathcal{G}_2 = \mathcal{G}_\infty$.

C.23 Zero-error code for degenerate DMCs

In Appendix C.23, we present our zero-error code with instantaneous encoding and common randomness for transmitting k symbols of a DSS over a degenerate DMC. In Appendix C.23, we present the proof that the code in Appendix C.23 achieves zero error for any rate asymptotically below $\frac{C}{H}$.

For a degenerate DMC (4.9) in Theorem 12, we denote by $P_{Y|X} \colon \mathcal{X} \to \mathcal{Y}$ its single-letter transition probability and denote by P_X^* its capacity-achieving distribution. We relabel x in (4.9a) by ACK, and relabel x' in (4.9b) by NACK. We denote by $E_G(P_{Y|X}, R_c)$ Gallager's error exponent [107], where R_c is the channel coding rate in nats per channel use[3]. We denote by $R(\ell)$ the rate of the code used in the communication phase of the ℓ-th block, and we denote by $\hat{S}^k(\ell)$ the estimate formed at the end of the communication phase of the ℓ-th block.

[3]For the consistency of notation, we use the same unit (i.e., nats per channel use) for R_c as that in Gallager's paper [107]. The unit of all other rates in this chapter is symbols per channel use.

Zero-error code with instantaneous encoding and common randomness

Similar to [40, 43, 44, 87], our code is divided into blocks. Each block contains a communication phase and a confirmation phase. The first block is different from the blocks after it, since it uses a Shannon limit-achieving code in the communication phase, whereas the blocks after the first block use random coding for all source sequences in alphabet $[q]^k$. We introduce the first block and the ℓ-th block, $\ell \geq 2$, respectively.

The first block is transmitted according to steps i)–ii) below. See Fig. C.3 (a) below for the diagram of the time division of transmitted blocks. See Fig. C.3 (b)–(c) for the diagram of the first block.

i) Communication phase. The first k symbols S^k of the DSS in Theorem 12 is transmitted via a Shannon limit-achieving code with instantaneous encoding and common randomness at rate $R(1) < \frac{C}{H}$ symbols per channel use. (Such a code has been presented in the proof sketch of Theorem 12. Namely, if $f = \infty$, we use a buffer-then-transmit code that implements the block encoding scheme in [104, Theorem 2]; if $f < \infty$, we precede the block encoding scheme in [104, Theorem 2] by an instantaneous encoding phase that satisfies (4.41).) At the end of the communication phase, the decoder yields an estimate $\hat{S}^k(1)$ of the source S^k using the channel outputs that it has received in this phase.

ii) Confirmation phase. The encoder knows $\hat{S}^k(1)$ since it knows the channel outputs through the noiseless feedback. The encoder repeatedly transmits ACK if $S^k = \hat{S}^k(1)$, and transmits NACK if $S^k \neq \hat{S}^k(1)$, for n_k channel uses. We pick n_k as

$$n_k = \delta k, \tag{C.117}$$

where $\delta \in (0, 1)$ can be made arbitrarily small. At the end of the confirmation phase, if the decoder receives a y, then it terminates the transmission and output $\hat{S}^k_{\eta_k} = \hat{S}^k(1)$; otherwise, the encoder transmits the next block.

The ℓ-th block, $\ell \geq 2$, is transmitted according to steps iii)–iv) below.

iii) Communication phase. For every sequence in the alphabet $[q]^k$ of S^k, the encoder generates a codeword via random coding according to the capacity-achieving distribution P_X^* at rate $R(2) < \frac{C}{\log q}$ symbols per channel use. At the end of the communication phase, the maximum likelihood (ML) decoder yields an estimate $\hat{S}^k(\ell)$ of the source symbols S^k using the channel outputs that it has received in this phase.

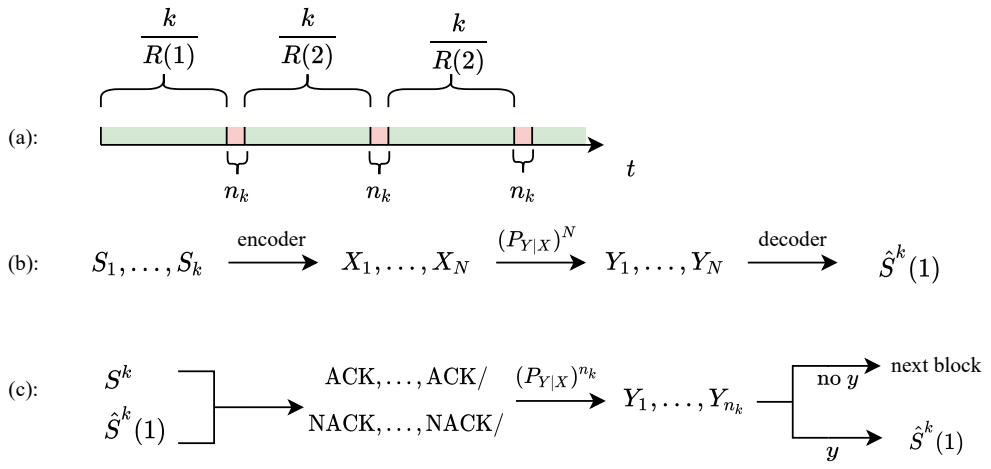


Figure C.3: (a) Time division of the transmitted blocks. The green regions represent the communication phases, and the red regions represent the confirmation phases. The *expected* length of the first communication phase is $\frac{k}{R(1)}$. The length of the ℓ-th communication phase, $\ell \geq 2$, is $\frac{k}{R(2)}$ since the random coding scheme has a fixed length. The length of the confirmation phase is n_k (C.117). (b) Communication phase of the first block. The codeword length N can be random with expectation $\mathbb{E}[N] = \frac{k}{R(1)}$. (c) Confirmation phase of the first block.

iv) Confirmation phase. The encoder, the decoder, and the stopping rule are the same as those in the first block with $\hat{S}^k(1) \leftarrow \hat{S}^k(\ell)$.

The random codebook is refreshed in every retransmitted block and is known by the decoder. This gives rise to the following observations:

1) The codewords transmitted in the communication phases of the $\ell = 1, 2, \ldots$ blocks are independent from each other;

2) As a result of 1), the channel outputs of the $\ell = 1, 2, \ldots$ blocks are independent from each other;

3) The codewords transmitted in the communication phase of the $\ell = 2, 3, \ldots$ blocks are i.i.d. random vectors. (The codeword in the first block is excluded since the first block need not use random coding in the communication phase);

4) As a result of 3), the channel outputs of the $\ell = 2, 3, \ldots$ blocks are i.i.d. random vectors.

We will use observations 2) and 4) in the proof below.

Proof of Theorem 12

Fix any $R < \frac{C}{H}$. We show that by adjusting $R(1)$, the rate of the Shannon limit-achieving code in the communication phase of the first block, to R, the code in

Appendix C.23 achieves zero error with rate converging to R (4.70).

We denote by η_k and T_k the stopping time and the number of blocks transmitted after the first block until the stopping time, respectively. We denote by A_ℓ the event that no y is received in the confirmation phase of the ℓ-th block.

Since the decoder will never receive y if ACK is transmitted in the confirmation phase, the error probability of the code in Appendix C.23 is zero, i.e.,

$$\mathbb{P}[S^k \neq \hat{S}^k(1+T_k)] = 0, \tag{C.118}$$

where $1+T_k$ represents the total number of blocks transmitted until the stopping time, and T_k is almost surely finite as a result of Lemmas 27 and 28 below. This confirms that the code in Section C.23 achieves zero error (4.15).

To analyze the behavior of the rate $R_k = \frac{k}{\mathbb{E}[\eta_k]}$, we first observe that since the expected length of the first block is $\frac{k}{R(1)} + \delta k$ and the (fixed) length of the ℓ-th block, $\ell \geq 2$, is $\frac{k}{R(2)} + \delta k$, the expected decoding time $\mathbb{E}[\eta_k]$ is equal to

$$\mathbb{E}[\eta_k] = \frac{k}{R(1)} + \delta k + \mathbb{E}[T_k]\left(\frac{k}{R(2)} + \delta k\right). \tag{C.119}$$

We bound the expected number of blocks T_k transmitted after the first block using Lemmas 27 and 28, stated next.

Lemma 27. *The number of blocks T_k transmitted after the first block satisfies*

$$\mathbb{E}[T_k] \leq \frac{\mathbb{P}[S^k \neq \hat{S}^k(1)] + (1 - P_{Y|X}(y|\mathrm{ACK}))^{\delta k}}{1 - \mathbb{P}[S^k \neq \hat{S}^k(\ell)] - (1 - P_{Y|X}(y|\mathrm{ACK}))^{\delta k}}. \tag{C.120}$$

Proof. Appendix C.23. □

Lemma 28. *Given a DSS with entropy rate $H > 0$ satisfying assumptions (a)–(b) in Theorem 9, the probability of erroneously decoding S^k at the end of the communication phase of the ℓ-th block is upper bounded as*

$$\mathbb{P}[S^k \neq \hat{S}^k(1)] \leq e^{-\frac{k}{R(1)}\left(\frac{C}{1+o(1)} - \frac{H(S^k)}{k} R(1)\right)}, \tag{C.121a}$$

$$\mathbb{P}[S^k \neq \hat{S}^k(\ell)] \leq e^{-\frac{k}{R(2)} E_G\left(P_{Y|X}, R(2)\log q\right)}, \ \ell = 2, 3, \ldots \tag{C.121b}$$

Proof. Since the block encoding scheme [104, Theorem 2] satisfies Lemma 20 with $C_1 \leftarrow C$, one can follow Appendices C.8–C.9 with $C_1 \leftarrow C$ to upper bound the

expected decoding time of the Shannon limit-achieving code in [104, Theorem 2] and thereby obtain (C.121a). The error probability (C.121b) holds since the random encoder together with the ML decoder attains Gallager's error exponent [107] for channel coding rate (nats per channel use) below C. This holds regardless of the distribution of the message because Gallager's error exponent holds under the maximum error probability criterion. □

Plugging (C.120) and (C.121) into the right side of (C.119), we obtain the asymptotic behavior of the rate as

$$\lim_{k\to\infty} R_k = \lim_{k\to\infty} \frac{k}{\mathbb{E}[\eta_k]} \tag{C.122a}$$

$$\geq R(1)\frac{1}{1+R(1)\delta}. \tag{C.122b}$$

Letting $R(1)$ be arbitrarily close to $\frac{C}{H}$ and taking δ to an arbitrarily small number, we conclude (4.70).

Proof of Lemma 27

We establish the pmf of T_k using the probabilities $\mathbb{P}[A_\ell]$, $\ell = 1, 2, \ldots$ The complementary cdf of T_k is given by

$$\mathbb{P}[T_k > 0] \ = \mathbb{P}[A_1], \tag{C.123}$$

where (C.123) holds by the definition of A_1 and the stopping rule of the code. We proceed to show the pmf at $T_k = t$, $t \geq 1$ conditioned on $T_k > 0$:

$$\mathbb{P}[T_k = t|T_k > 0] = \mathbb{P}[A_2 \cap \cdots \cap A_t \cap A_{t+1}^c|A_1] \tag{C.124a}$$

$$= \left(\prod_{i=2}^{t} \mathbb{P}[A_i|A_1, \ldots, A_{i-1}]\right) \mathbb{P}[A_{t+1}^c|A_1, \ldots, A_t] \tag{C.124b}$$

$$= (\mathbb{P}[A_2])^{t-1}(1 - \mathbb{P}[A_2]), \tag{C.124c}$$

where (C.124a) is by the stopping rule of the code; (C.124b) is by expanding (C.124a); (C.124c) is by observations 2) and 4) in Appendix C.23: observation 2) implies that event A_i and its complementary event A_i^c are both independent of $A_1, \ldots, A_{i-1}$, $i \geq 2$, observation 4) implies that $\mathbb{P}[A_i] = \mathbb{P}[A_2]$, $i \geq 2$. Since the conditional pmf $\mathbb{P}[T_k = t|T_k > 0]$ in (C.124) follows a geometric distribution with success probability $1 - \mathbb{P}[A_2]$, its mean is given by

$$\mathbb{E}[T_k|T_k > 0] = \frac{1}{1 - \mathbb{P}[A_2]}. \tag{C.125}$$

Using (C.123) and (C.125), we obtain $\mathbb{E}[T_k]$ as

$$\mathbb{E}[T_k] = \frac{\mathbb{P}[A_1]}{1 - \mathbb{P}[A_2]}. \tag{C.126}$$

It remains to compute the probability of event A_ℓ in (C.126) to conclude (C.120). In the confirmation phase of the ℓ-th block, $\ell = 1, 2, \ldots$, conditioned on $S^k = \hat{S}^k(\ell)$, the probability of event A_ℓ is given by[4]

$$\mathbb{P}[A_\ell | S^k = \hat{S}^k(\ell)] = (1 - P_{Y|X}(y|\text{ACK}))^{\delta k}. \tag{C.127a}$$

The probability of event A_ℓ is upper bounded as

$$\mathbb{P}[A_\ell] = \mathbb{P}[A_\ell | S^k \neq \hat{S}^k(\ell)]\mathbb{P}[S^k \neq \hat{S}^k(\ell)] + \mathbb{P}[A_\ell | S^k = \hat{S}^k(\ell)]\mathbb{P}[S^k = \hat{S}^k(\ell)] \tag{C.128a}$$

$$\leq \mathbb{P}[S^k \neq \hat{S}^k(\ell)] + \mathbb{P}[A_\ell | S^k = \hat{S}^k(\ell)], \tag{C.128b}$$

where (C.128b) holds by upper bounding the first and the last probabilities on the right side of (C.128a) by 1. Plugging the upper bound in (C.128b) into the right side of (C.126), we obtain (C.120).

[4]For practical implementations, one can choose ACK as the channel input that achieves the maximum transition probability $\max_{x \in \mathcal{X}} P_{Y|X}(y|x)$ to increase the probability of receiving a y.