A New Method for Employing Feedback to Improve Coding Performance

Aaron B. Wagner, *Fellow, IEEE*, Nirmal V. Shende, and Yücel Altuğ

***Abstract*— We introduce a novel mechanism, called timid/bold coding, by which feedback can be used to improve coding performance. For a certain class of DMCs, called *compound-dispersion channels*, we show that timid/bold coding allows for an improved second-order coding rate compared with coding without feedback. For DMCs that are not compound dispersion, we show that feedback does not improve the second-order coding rate. Thus we completely determine the class of DMCs for which feedback improves the second-order coding rate. An upper bound on the second-order coding rate is provided for compound-dispersion DMCs. We also show that feedback does not improve the second-order coding rate for very noisy DMCs. The main results are obtained by relating feedback codes to certain controlled diffusions.**

***Index Terms*— Channel coding, diffusions, feedback communications, second-order coding rate, stochastic control.**

I. INTRODUCTION

CONSIDER the canonical communication model consisting of a single encoder sending bits to a single decoder over a discrete memoryless channel (DMC). We assume the alphabets are finite, the channel law is completely known, and the transmission rate is fixed, i.e., the decoding of the entire message must occur at a prespecified time.

In practice, point-to-point communication links are usually paired with a feedback link from the decoder to the encoder, which can communicate messages in the reverse direction but can also be used to facilitate communication along the forward link. Although such feedback links are common in practice, it is not well understood theoretically how they can be most effectively used. We consider how unfettered use of a perfect feedback link can improve asymptotic coding performance across the forward channel. It is well known that feedback does not improve the capacity of a DMC [1]. We shall consider how feedback can be used to improve the more-refined *second-order coding rate* of the channel (see Definition 2 to follow).

Manuscript received September 30, 2019; revised February 19, 2020; accepted March 26, 2020. Date of publication May 25, 2020; date of current version October 21, 2020. This work was supported by the National Science Foundation under Grant CCF-1513858 and Grant CCF-1956386. This article was presented in part at the 2014 and 2018 IEEE International Symposia on Information Theory. *(Corresponding author: Aaron B. Wagner.)*

Aaron B. Wagner is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA (e-mail: wagner@cornell.edu).

Nirmal V. Shende is with Marvell Semiconductor, Santa Clara, CA 95054 USA (e-mail: nshende@marvell.com).

Yücel Altuğ is with Natera Inc., San Carlos, CA 94070 USA (e-mail: yucelaltug@gmail.com).

Communicated by L. Wang, Associate Editor for Shannon Theory.

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

Digital Object Identifier 10.1109/TIT.2020.2997385

A priori, it is not clear that feedback improves the second-order coding rate at all. Indeed, none of the mechanisms by which feedback is known to improve coding performance obtains for the setup under study. The channel has no memory, so feedback cannot be used to anticipate future channel disturbances (as in, e.g., [2]). The channel law is known, so feedback is not useful for learning the channel statistics (as in, e.g., [3]). The blocklength is fixed, so feedback does not allow the code to outwait unfavorable noise realizations (cf. [4]). There is no cost constraint, so the encoder cannot use feedback to opportunistically consume resources (cf. [5], [6]). Since the second-order coding rate focuses on a "high-rate" regime, the increase in the effective minimum distance of the code afforded by feedback is not useful (cf. [7]). Since the channel is point-to-point, none of the various ways that feedback can enable coordination in networks (e.g., [8]) can be applied. Indeed, a negative result is available showing that feedback does not increase the second-order coding rate for DMCs satisfying a certain symmetry condition [9, Theorem 15].

We introduce a novel mechanism by which feedback can improve coding performance for some DMCs, even when the coding is high-rate and fixed-blocklength and the channel is known and memoryless. The idea is the following. Suppose a player may flip one of two fair coins in each of n rounds. If the player chooses to flip the first (resp. second) coin, then she wins \$1 (resp. \$2) with probability half and loses \$1 (resp. \$2) with probability half. We assume that each flip of each coin is independent of everything else and that the initial wealth is $w\sqrt{n}$ with $w > 0$. The player wins the overall game if her wealth after n rounds is positive. How should the player decide which coin to flip in a given round in order to maximize her chance of winning? If the player is required to choose her strategy before the start of the game, i.e., she is not allowed to update her choice after seeing the previous flips, one can verify that playing the first coin in all of the rounds is asymptotically her best strategy. Indeed, under this strategy the central limit theorem (CLT) implies that the probability of losing converges to $\Phi(-w)$, where Φ is the distribution of the standard Gaussian random variable. If she plays the second coin in all rounds, then this probability is $\Phi(-w/2)$, which is worse. If she timeshares the two coins, the probability will be in between. Essentially, because she is expecting to win, she minimizes the probability of losing by minimizing the variance of her wealth after round n. Conversely, if she starts with $w < 0$, then she should play the second coin for all time. Since she is expecting to lose, she minimizes the probability of losing by maximizing the variance of her wealth after round n.

If the player can select the coin for each round using knowledge of the outcomes of the previous rounds, then she can do better by utilizing both coins. Consider, for simplicity, the scenario in which the player flips the first coin for the first $n/2$ rounds and then selects one coin to flip for all of the $n/2$ remaining rounds. A reasonable strategy is the following: if the wealth after the first half is positive, play "timid," i.e., flip the coin that pays $\pm\$1$. Otherwise, play "bold," i.e., flip the coin that pays $\pm\$2$. The justification is that if her wealth is positive after $n/2$ rounds, then the player is expecting to win, so she should minimize the variance of her wealth after round n. If her wealth is negative after round $n/2$, then she is expecting to lose, so she seeks to maximize the variance after n rounds. Another view is that if her wealth is negative after round $n/2$, then she needs to have more wins than losses during the second half in order to win overall; she needs to be lucky. Quoting Cover and Thomas [10, p. 391]: "If luck is required to win, we might as well assume that we will be lucky and play accordingly." Under the assumption that the player will have more wins than losses, playing the coin that pays $\pm\$2$ provides more wealth.

The connection to channel coding is provided by Lemmas 14 and 15 in the Appendix, which relate the design of feedback codes to the design of controllers for a particular controlled random walk. For channels with multiple capacity-achieving input distributions that give rise to information densities with different variances, which we call *compound-dispersion* channels (see Definition 1), the controlled random walk that arises through Lemmas 14 and 15 admits the timid/bold play mechanism described above, and this in turn yields feedback codes that asymptotically outperform the best non-feedback codes. In channel-coding terms, the idea is that, with compound-dispersion channels, the encoder can use codewords with symbols drawn from multiple input distributions such that the mean rate of information conveyance across the channel is the same under all of these distributions (namely, the Shannon capacity), but the variance is different. The encoder then monitors the progress of transmission via the feedback link and uses a "bold" input distribution when a decoding error is expected and a "timid" input distribution when it is not. We call this *timid/bold coding*.

Of course, it is desirable to update the strategy at each time during the block, instead of only halfway through. This, however, comes at the expense of more technical arguments. In particular, we use convergence results for Itô diffusion processes. An inspiration for this scheme is a result of McNamara on the optimal control of the diffusion coefficient of a diffusion process [11]. Consider the following stochastic differential equation (SDE):

$$\xi_t = \xi_0 + \int_0^t \psi_s(\xi_s)\, dB_s$$

where ξ_0 is a constant, $0 < \psi_s(x) \in [\psi_{\min}, \psi_{\max}]$ for all s and x, and $\{B_t\}$ is a Brownian motion. If the goal is to maximize $P(\xi_1 \geq 0)$ by choosing the function $\psi_s(\cdot)$, then McNamara shows that the bang-bang scheme

$$\psi^{\text{opt}}(u) = \begin{cases} \psi_{\min} & u > 0, \\ \psi_{\max} & u \leq 0. \end{cases} \tag{1}$$

is an optimal controller. If we view this as a gambling problem then, in words, the gambler should play maximally timid when she is expecting to win and maximally bold when she is expecting to lose.

McNamara [11] notes that animals have been observed to follow more-risky foraging strategies when near starvation and less-risky strategies when food reserves are high. Similar behavior is observed in sports, where, e.g., a hockey team will leave its goal unprotected in order to field an extra offensive player if it is losing late in the game. In the context of feedback communication, we show that timid/bold coding improves the second-order coding rate compared with the best non-feedback code for all compound-dispersion channels. We also show a matching converse result, namely that feedback does not improve the second-order coding rate of simple (i.e., non-compound) dispersion channels, improving upon [9, Theorem 15]. Thus, timid/bold coding provides a second-order coding rate improvement whenever such an improvement is possible.[1] The converse is obtained by using the code modification technique of Fong and Tan [12] along with a "Berry-Esseen"-type martingale CLT and large deviations results for martingales. In particular, this settles the problem of determining whether feedback improves the second-order coding rate for a given DMC.

For compound-dispersion channels, it is not clear if timid/bold coding is an optimal feedback signaling scheme. To shed some light on this question, we provide the first nontrivial impossibility result for the second-order coding rate of feedback communication over general DMCs. The technical challenge in proving such a result is that standard martingale central limit theorems do not provide useful bounds. Instead, we obtain the result using tools from stochastic calculus, namely, martingale embeddings, change-of-time methods, and McNamara's solution to the above-mentioned SDE. The bound on the second-order coding rate that we obtain is functionally identical to the second-order coding rate achieved by timid/bold coding, although evaluated at different channel parameters. The two bounds coincide for some channels but not in general.

Finally, we show that feedback does not improve the second-order coding rate for a class of DMCs called *very noisy channels (VNCs)*. Reiffen [13] introduced VNCs to model physical channels that operate at a very low signal-to-noise ratio.[2] VNCs are useful for modeling channels in which a resource (such as power) is spread over many degrees of freedom (such as bandwidth) [15]. We show that DMCs behave as simple-dispersion channels in the very noisy limit, and that feedback does not improve the second-order rate in this asymptotic regime. However, since DMCs only satisfy

[1] We assume throughout that the channel satisfies $V_{\min} > 0$ as explained in the next section.

[2] The VNCs introduced by Reiffen are called Class I VNCs by Majani [14], where he also defined Class II VNCs. In this paper, we focus on Class I VNCs and refer to them simply as VNCs.

the simple-dispersion property in the limit, our converse for simple-dispersion channels is not directly applicable. Hence, we use a different proof technique.

The balance of the paper is organized as follows. The next section describes the problem formulation more precisely and states all five of our results. The remaining five sections then provide the proofs of these five theorems in order. As described earlier, the Appendix provides two lemmas that relate the design of feedback codes to the design of controllers for controlled random walks. Although these lemmas have strong precedents in the literature, the connection between feedback signaling and controlled random walks seems to be novel.

II. Notation, Definitions and Statement of the Results

A. Notation

$\mathbb{R}, \mathbb{R}^+$, $\mathbb{R}^-$ and $\mathbb{R}_+$ denote the set of real, positive real, negative real and non-negative real numbers, respectively. $\mathbb{Z}^+$ denotes the set of positive integers. We assume the input alphabet, $\mathcal{X}$, and the output alphabet, $\mathcal{Y}$, of the channel are finite. For a finite set $\mathcal{A}$, $\mathcal{P}(\mathcal{A})$ denotes the set of all probability measures on $\mathcal{A}$. Similarly, for two finite sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{P}(\mathcal{B}|\mathcal{A})$ denotes the set of all stochastic matrices from $\mathcal{A}$ to $\mathcal{B}$. Given any $P \in \mathcal{P}(\mathcal{A})$ and $W \in \mathcal{P}(\mathcal{B}|\mathcal{A})$, $P \circ W$ denotes the distribution

$$(P \circ W)(a,b) = P(a)W(b|a).$$

Given any $P \in \mathcal{P}(\mathcal{A})$, $\mathcal{S}(P) := \{a \in \mathcal{A} : P(a) > 0\}$. $\Phi(\cdot)$ and $\phi(\cdot)$ denote the CDF and PDF of the standard Gaussian random variable, respectively. $\mathbf{1}\{\cdot\}$ denotes the standard indicator function. For a random variable Z, $\|Z\|_\infty$ denotes its essential supremum (that is, the infimum of those numbers z such that $P(Z \leq z) = 1$). Boldface letters will denote vectors (e.g., $\mathbf{y}^k = [y_1, \ldots, y_k]$) and continuous-time process (e.g., $\mathbf{N} = (N_t : t \geq 0)$). We follow the notation of Csiszár and Körner [16] for standard information-theoretic quantities. See Karatzas and Shreve [17] for standard definitions and notations used in stochastic calculus. Unless otherwise stated, all logarithms and exponentiations are base e.

B. Definitions

Given a DMC $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$, C denotes the capacity of the channel, and

$$\Pi_W^* := \{Q \in \mathcal{P}(\mathcal{X}) \colon I(Q;W) = C(W)\} \tag{2}$$

denotes the set of capacity-achieving input distributions. There exists a distribution q^* over $\mathcal{Y}$ such that for any $P \in \Pi_W^*$,

$$q^*(y) := \sum_{x \in \mathcal{X}} P(x)W(y|x). \tag{3}$$

and q^* can be assumed to satisfy $q^*(y) > 0$ for all $y \in \mathcal{Y}$ [18, Corollaries 1 and 2 to Theorem 4.5.1].[3] Define

$$\begin{aligned}
\mathrm{i}^*(X,Y) &:= \log \frac{W(Y|X)}{q^*(Y)}, \\
\nu_x &:= \mathrm{Var}[\mathrm{i}^*(X,Y)|X = x], \\
V_{\min} &:= \min_{P \in \Pi_W^*} \sum_{x \in \mathcal{X}} P(x)\nu_x, \\
V_{\max} &:= \max_{P \in \Pi_W^*} \sum_{x \in \mathcal{X}} P(x)\nu_x, \\
\nu_{\min} &:= \min_{x \in \mathcal{X}} \nu_x, \\
\nu_{\max} &:= \max_{x \in \mathcal{X}} \nu_x, \\
\mathrm{i}_{\max} &:= \max_{x \in \mathcal{X}, y \in \mathcal{Y} : W(y|x) > 0} |\mathrm{i}^*(x,y)|
\end{aligned}$$

Let $V_{\min}$ and $V_{\max}$ denote V_ε for an arbitrary $\varepsilon \in (0, \frac{1}{2})$ and $\varepsilon \in [\frac{1}{2}, 1)$, respectively, for notational convenience.

Definition 1: We will call a DMC with[4] $V_{\min} > 0$ *simple-dispersion* if $V_{\min} = V_{\max}$. Otherwise, it is called *compound-dispersion*.

Remark 1: The set of compound-dispersion DMCs is not empty. As an example, consider[5] $p \in (0,1)$ such that

$$h(p) + (1-p)\log 2 = h(q), \tag{4}$$

for some $q \in (0, 1/2)$, where $h(\cdot)$ denotes the *binary entropy function*, i.e., for any $r \in [0,1]$, $h(r) := -r \log r - (1 - r)\log(1-r)$. Define $\mathcal{X} := \{0,1,2,3,4,5\}$, $\mathcal{Y} := \{0,1,2\}$ and $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ as

$$W(y|x) := \begin{bmatrix} p & 0.5(1-p) & 0.5(1-p) \\ 0.5(1-p) & p & 0.5(1-p) \\ 0.5(1-p) & 0.5(1-p) & p \\ q & 1-q & 0 \\ 0 & q & 1-q \\ 1-q & 0 & q \end{bmatrix}. \tag{5}$$

One can numerically verify that if $p = 0.8$, then $q \approx 0.337$ satisfies (4) and the channel defined in (5) has $V_{\min} \approx 0.102$, which is attained by the uniform input distribution over the set of input symbols $\{3,4,5\}$, and $V_{\max} \approx 0.692$, which is attained by the uniform input distribution over the set of input symbols $\{0,1,2\}$. Note that for this channel $\nu_{\min} = V_{\min}$ and $\nu_{\max} = V_{\max}$. See Strassen [19, Sec. 5(ii)] for a similar example.

Outside of the realm of DMCs, there are less-contrived examples of compound dispersion channels [20]. In principle, one can apply timid/bold coding to such channels whenever feedback is available. Whether timid/bold coding provides sufficient gains on such channels to merit practical implementation is an interesting question that is not addressed in the present paper, which focuses on the theoretical development of the idea. $\Diamond$

[3] We assume without loss of generality that W does not contain an all-zero column.

[4] Note that if $V_{\min} > 0$, then the capacity of the channel is positive.

[5] One can verify that any $p \in [0.8, 1)$ satisfies the following.

An (n, R) code with ideal feedback for a DMC consists of an encoder f, which at the kth time instant $(1 \leq k \leq n)$ chooses an input $x_k = f(m, y_1 \ldots, y_{k-1}) \in \mathcal{X}$, where $m \in \{1, \ldots, \lceil \exp(nR) \rceil\}$ denotes the message to be transmitted, and a decoder g, which maps outputs $(y_1, \ldots, y_n)$ to $\hat{m} \in \{1, \ldots, \lceil \exp(nR) \rceil\}$. Given $\varepsilon \in (0, 1)$, define

$$M^*_{\mathrm{fb}}(n, \varepsilon) := \max\left\{\lceil \exp(nR) \rceil \in \mathbb{R}_+ : \bar{\mathrm{P}}_{\mathrm{e,fb}}(n, R) \leq \varepsilon\right\}, \quad (6)$$

where $\bar{\mathrm{P}}_{\mathrm{e}}(n, R)$ denotes the minimum average error probability attainable by any (n, R) code with feedback. Similarly,

$$M^*(n, \varepsilon) := \max\left\{\lceil \exp(nR) \rceil \in \mathbb{R}_+ : \bar{\mathrm{P}}_{\mathrm{e}}(n, R) \leq \varepsilon\right\}, \quad (7)$$

where $\bar{\mathrm{P}}_{\mathrm{e}}(n, R)$ denotes the minimum average error probability attainable by any (n, R) code (without feedback).

Definition 2: The *second-order coding rate* of a DMC $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ at the average error probability ε is defined as

$$\liminf_{n\to\infty} \frac{\log M^*(n, \varepsilon) - nC}{\sqrt{n}}. \quad (8)$$

The second-order coding rate with feedback is defined analogously.

C. Statement of Results

Before we state our results, we recall the following result of Strassen [19]. For any $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ and $\varepsilon \in (0, 1)$, Strassen shows[6]

$$\lim_{n\to\infty} \frac{\log M^*(n, \varepsilon) - nC}{\sqrt{n}} = \sqrt{V_\varepsilon}\Phi^{-1}(\varepsilon). \quad (9)$$

That is, the second-order coding rate without feedback is $\sqrt{V_\varepsilon}\Phi^{-1}(\varepsilon)$. Using timid/bold coding, we shall show that this can be strictly improved with feedback for any compound-dispersion channel, for any $0 < \varepsilon < 1$.

We begin with a preliminary result to this effect, which only holds for $0 < \varepsilon < 1/2$ and which does not provide as large of an improvement as the subsequent result, Theorem 2. The advantage is that its proof does not require any of the stochastic calculus used in the proofs that follow.

Theorem 1 (Coarse Achievability for Compound-Dispersion Channels): Fix an arbitrary $\varepsilon \in (0, 0.5)$ and consider a compound-dispersion channel W with $V_{\min} > 0$. Let $\beta = \sqrt{V_{\min}/V_{\max}} < 1$. Then there exists $1 < \alpha < 1/(2\varepsilon)$ such that

$$\begin{aligned} f(\alpha) = \varepsilon(\alpha - 1) - (1-\beta)\phi(2\sqrt{2}\Phi^{-1}(\alpha\varepsilon)) \cdot \\ \left(\frac{1}{\sqrt{2\pi}} - \phi(\sqrt{2}\Phi^{-1}(\alpha\varepsilon))\right) < 0, \end{aligned} \quad (10)$$

and for any such α,

$$\begin{aligned} \liminf_{n\to\infty} \frac{\log M^*_{\mathrm{fb}}(n, \varepsilon) - nC}{\sqrt{n}} &\geq \sqrt{V_\varepsilon}\Phi^{-1}(\alpha\varepsilon) \quad (11) \\ &> \sqrt{V_\varepsilon}\Phi^{-1}(\varepsilon). \quad (12) \end{aligned}$$

Proof: Please see Section III. □

[6]Strassen provides a more-refined result, which was corrected by Polyanskiy *et al.* [21]. No correction is needed for the weaker result quoted here, however. Strassen states his result for the maximal error probability criterion then extends the analysis to the average error probability criterion in Section 5(iii).

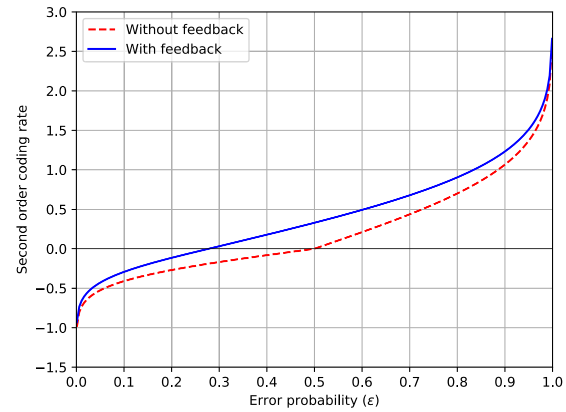

Fig. 1. Second-order coding rate with and without feedback for the channel in (5) with $p = 0.8$. For this channel, the lower bound in Theorem 2 and the upper bound in Theorem 4 coincide, determining the second-order coding rate with feedback.

The proof proceeds by switching between timid and bold coding at most once, halfway through the transmission. The next result improves upon this by allowing for a potential switch between timid and bold coding after each time step.

Theorem 2 (Refined Achievability for Compound-Dispersion Channels): Consider any $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ with $0 < V_{\min}$ and let $\beta := \sqrt{V_{\min}/V_{\max}}$. Thus

$$\liminf_{n\to\infty} \frac{\log M^*_{\mathrm{fb}}(n, \varepsilon) - nC}{\sqrt{n}} \geq \begin{cases} \sqrt{V_{\min}}\Phi^{-1}\left(\frac{1}{2\beta}\varepsilon(1+\beta)\right), & \varepsilon \in \left(0, \frac{\beta}{1+\beta}\right], \\ \sqrt{V_{\max}}\Phi^{-1}\left(\frac{1}{2}[\varepsilon(1+\beta) + (1-\beta)]\right), & \varepsilon \in \left(\frac{\beta}{1+\beta}, 1\right). \end{cases} \quad (13)$$

Proof: Please see Section IV. □

Note that the theorem applies to any DMC with $V_{\min} > 0$, but if $\beta = 1$ (i.e., the channel is simple dispersion), then (13) reduces to the achievability half of (9). The right-hand-side of (13) is shown in Fig. 1, alongside the second-order coding rate without feedback, for the channel in (5) with $p = 0.8$ and q selected to satisfy (4). Note that the range of ε over which one can approach the capacity from above, i.e., for which the second-order coding rate is positive, is enlarged by the presence of feedback. The right-hand-side of (13) is easily verified to exceed $\sqrt{V_\varepsilon}\Phi^{-1}(\varepsilon)$ for all ε if the channel is compound-dispersion (i.e., $\beta < 1$). The next result shows that if the channel is not compound-dispersion then feedback does not improve the second-order coding rate.

Theorem 3 (Feedback Does Not Improve the Second-Order Coding Rate for Simple-Dispersion Channels): For any $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ with $0 < V_{\min} = V_{\max}$ (i.e., simple-dispersion) and any $\varepsilon \in (0, 1)$,

$$\begin{aligned} \lim_{n\to\infty} \frac{\log M^*_{\mathrm{fb}}(n, \varepsilon) - nC}{\sqrt{n}} &= \lim_{n\to\infty} \frac{\log M^*(n, \varepsilon) - nC}{\sqrt{n}} \\ &= \sqrt{V_{\min}}\Phi^{-1}(\varepsilon) \\ &= \sqrt{V_\varepsilon}\Phi^{-1}(\varepsilon). \end{aligned}$$

Proof: Please see Section V. □

The proof of Theorem 3 uses a method of making feedback codes "constant-composition," which is inspired by Fong and Tan's work on parallel Gaussian channels [12]. Fong and Tan have also noted that their techniques can be applied to DMCs to obtain something like Theorem 3 [22].

If the channel is compound dispersion, then feedback improves the second-order coding rate, and Theorem 2 (along with (9)) provides a lower bound on the size of the improvement. The next theorem provides a comparable upper bound.

Theorem 4 (Impossibility for Compound-Dispersion Channels): Consider any $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ with $0 < \nu_{\min}$ and let $\lambda := \sqrt{\nu_{\min}/\nu_{\max}}$. Then

$$\limsup_{n\to\infty} \frac{\log M^*_{\text{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \tag{14}$$

$$\leq \begin{cases} \sqrt{\nu_{\min}}\Phi^{-1}\left(\frac{1}{2\lambda}\varepsilon(1+\lambda)\right), & \varepsilon \in \left(0, \frac{\lambda}{1+\lambda}\right], \\ \sqrt{\nu_{\max}}\Phi^{-1}\left(\frac{1}{2}[\varepsilon(1+\lambda)+(1-\lambda)]\right), & \varepsilon \in \left(\frac{\lambda}{1+\lambda}, 1\right). \end{cases} \tag{15}$$

Proof: Please see Section VI. □

The upper bound in Theorem 4 equals the achievability result in Theorem 2 but with $\nu_{\min}$ and $\nu_{\max}$ replacing $V_{\min}$ and $V_{\max}$, respectively. Thus the two results are similar in spirit. Both, in fact, use McNamara's scheme in (1). However, the range of values that the diffusion coefficient can assume is larger for the upper bound ($[\sqrt{\nu_{\min}}, \sqrt{\nu_{\max}}]$) than for the lower bound ($[\sqrt{V_{\min}}, \sqrt{V_{\max}}]$). For the channel in (5), $\nu_{\max} = V_{\max}$ and $\nu_{\min} = V_{\min}$, so the upper bound and lower bound coincide and the second-order coding rate with feedback is determined (and is depicted in Fig. 1). The two bounds do not coincide in general, however.

Finally, we consider very noisy channels (VNCs). For our purposes, a very noisy channel is one of the form

$$W_\zeta(y|x) = \Gamma(y)\left(1+\zeta\lambda(x,y)\right), \tag{16}$$

where Γ is a probability distribution on the output alphabet $\mathcal{Y}$ such that $\Gamma(y) > 0$ for all y, $\lambda(x,y)$ satisfies

$$\sum_{y\in\mathcal{Y}} \Gamma(y)\lambda(x,y) = 0 \tag{17}$$

for all $x \in \mathcal{X}$, and ζ is infinitesimally small. In the very noisy limit, i.e., as ζ tends to zero, $V_{\min}$ and $V_{\max}$ converge together and the channel behaves as one with simple dispersion. In light of Theorem 3, one therefore expects feedback not to improve the second-order coding rate in the very noisy limit. Since $V_{\min}$ and $V_{\max}$ are only equal in the limit (when suitably scaled), the result does not follow from Theorem 3, however. Since $\sqrt{\nu_{\min}}$ and $\sqrt{\nu_{\max}}$ do not necessarily converge together, the result does not follow from Theorem 4 either.

Theorem 5 (Feedback Does Not Improve the Second-Order Coding Rate in the Very Noisy Limit): Consider a channel family $W_\zeta \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ of the form $W_\zeta(y|x) = \Gamma(y)\left(1+\zeta\lambda(x,y)\right)$, with $\Gamma \in \mathcal{P}(\mathcal{Y})$. Let C_ζ, $V_{\min,\zeta}$, $V_{\max,\zeta}$, and $\log M^*_{\text{fb},\zeta}(n,\varepsilon)$ denote C, $V_{\min}$, $V_{\max}$, and $M^*_{\text{fb}}(n,\varepsilon)$, respectively, for the channel $W_\zeta \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$. If there exists $P \in \mathcal{P}(\mathcal{X})$ such that the quantity

$$\sum_{y\in\mathcal{Y}} \Gamma(y)\left(\sum_{x\in\mathcal{X}} P(x)\lambda^2(x,y) - \left(\sum_{x\in\mathcal{X}} P(x)\lambda(x,y)\right)^2\right)$$

is positive, which ensures that $C_\zeta > 0$ for all sufficiently small ζ, then

$$\limsup_{\zeta\to 0}\limsup_{n\to\infty} \frac{\log M^*_{\text{fb},\zeta}(n,\varepsilon) - nC_\zeta}{\sqrt{nV_{\min,\zeta}}} \leq \Phi^{-1}(\varepsilon),$$

for $\varepsilon \in \left(0, \frac{1}{2}\right]$ and

$$\limsup_{\zeta\to 0}\limsup_{n\to\infty} \frac{\log M^*_{\text{fb},\zeta}(n,\varepsilon) - nC_\zeta}{\sqrt{nV_{\max,\zeta}}} \leq \Phi^{-1}(\varepsilon),$$

for $\varepsilon \in \left(\frac{1}{2}, 1\right)$.

Proof: Please see Section VII. □

One can also show that feedback does not improve the high-rate error exponent or moderate deviations performance of VNCs [23]. Note that very noisy channels are unusual in that their reliability function is known at all rates [18], [24].

The next five sections contain the proofs of Theorems 1 through 5, respectively.

III. Proof of Theorem 1

Note that $f(\cdot)$ is continuous on $[1,\infty)$ and $f(1) < 0$. Hence there exists $1 < \alpha < 1/(2\varepsilon)$ with $f(\alpha) < 0$ and we fix any such α in what follows. Define

$$\nu = \sqrt{2}\Phi^{-1}(\alpha\varepsilon) < 0. \tag{18}$$

We shall use Lemma 14 in the Appendix. Note that we only require that (144) holds with the limit superior taken along the even integers. Accordingly, suppose that n is even. Let $Q_{\max}$ denote a distribution on $\mathcal{P}(\mathcal{X})$ that attains $V_{\max}$, and define $Q_{\min}$ similarly. Select the controller F as follows

$$F(x^k, y^k) = \begin{cases} Q_{\min} & \text{if } k \leq n/2 \\ Q_{\min} & \text{if } k > n/2,\ \log\frac{W(y^{n/2}|x^{n/2})}{q^*(y^{n/2})} > \frac{nC}{2} + \nu\sqrt{\frac{nV_{\min}}{2}} \\ Q_{\max} & \text{if } k > n/2,\ \log\frac{W(y^{n/2}|x^{n/2})}{q^*(y^{n/2})} \leq \frac{nC}{2} + \nu\sqrt{\frac{nV_{\min}}{2}}. \end{cases} \tag{19}$$

Note that $FW = q^* \times q^* \times \cdots q^*$. For convenience we define

$$\Gamma_n = (F\circ W)\left(\sum_{k=1}^n \log\frac{W(Y_k|X_k)}{q^*(Y_k)} \leq nC + \sqrt{nV_{\min}}\Phi^{-1}(\alpha\varepsilon)\right).$$

Let $\underline{G}_n$ denote the CDF of

$$\frac{1}{\sqrt{(n/2)V_{\min}}}\sum_{i=1}^{n/2}\left[\log\frac{W(Y_i|X_i)}{q^*(Y_i)} - C\right]$$

when

$$\left\{\log\frac{W(Y_i|X_i)}{q^*(Y_i)}\right\}_{i=1}^{n/2}$$

are i.i.d. with distribution $Q_{\min} \circ W$. Similarly, let $\overline{G}_n$ denote the distribution of

$$\frac{1}{\sqrt{(n/2)V_{\min}}} \sum_{i=1}^{n/2} \left[\log \frac{W(Y_i|X_i)}{q^*(Y_i)} - C \right]$$

when $\left\{ \log \frac{W(Y_i|X_i)}{q^*(Y_i)} \right\}_{i=1}^{n/2}$ are i.i.d. with distribution $Q_{\max} \circ W$. We have

$$\begin{aligned} \Gamma_n &= \int_{\nu}^{\infty} \underline{G}_n\left(\nu - x\right) d\underline{G}_n(x) + \int_{-\infty}^{\nu} \overline{G}_n\left(\nu - x\right) d\underline{G}_n(x) \\ &= \underline{G}_{2n}\left(\Phi^{-1}(\alpha\varepsilon)\right) \\ &\quad - \int_{-\infty}^{\nu} \left[\underline{G}_n\left(\nu - x\right) - \overline{G}_n\left(\nu - x\right)\right] d\underline{G}_n(x). \end{aligned} \tag{20}$$

From the Berry-Esseen theorem[7] [26], [27], along with a first-order Taylor series approximation, we deduce that

$$\underline{G}_{2n}\left(\Phi^{-1}(\alpha\varepsilon)\right) \leq \alpha\varepsilon + \frac{\underline{\kappa}}{2\sqrt{n}}, \tag{21}$$

where $\underline{\kappa} := \mathbb{E}_{Q_{\min} \circ W}\left[|\log W(Y|X)/q^*(Y) - C|^3\right]/V_{\min}^{3/2} + 1$. Another application of the Berry-Esseen theorem implies that for any $x \in \mathbb{R}$,

$$\left|\underline{G}_n\left(\nu - x\right) - \Phi\left(\nu - x\right)\right| \leq \frac{\underline{\kappa}}{\sqrt{2n}}, \tag{22}$$

$$\left|\overline{G}_n\left(\nu - x\right) - \Phi\left(\beta\left[\nu - x\right]\right)\right| \leq \frac{\overline{\kappa}}{\sqrt{2n}}, \tag{23}$$

where

$$\overline{\kappa} := \mathbb{E}_{Q_{\max} \circ W}\left[|\log W(Y|X)/q^*(Y) - C|^3\right]/V_{\max}^{3/2} + 1.$$

Equations (22) and (23) imply that

$$\int_{-\infty}^{\nu} \left[\underline{G}_n(\nu - x) - \overline{G}_n(\nu - x)\right] d\underline{G}_n(x) \tag{24}$$

$$\geq \int_{-\infty}^{\nu} \left[\Phi(\nu - x) - \Phi\left(\beta\left[\nu - x\right]\right)\right] d\underline{G}_n(x) - \frac{\underline{\kappa} + \overline{\kappa}}{\sqrt{2n}} \tag{25}$$

$$= \int_{-\infty}^{\nu} \underline{G}_n(x)\left[\phi(\nu - x) - \beta\phi\left(\beta\left[\nu - x\right]\right)\right] dx - \frac{\underline{\kappa} + \overline{\kappa}}{\sqrt{2n}} \tag{26}$$

$$\geq \int_{-\infty}^{\nu} \Phi(x)\left[\phi(\nu - x) - \beta\phi\left(\beta\left[\nu - x\right]\right)\right] dx - \frac{3\underline{\kappa} + \overline{\kappa}}{\sqrt{2n}} \tag{27}$$

$$= \int_{-\infty}^{\nu} \phi(x)\left[\Phi(\nu - x) - \Phi(\beta[\nu - x])\right] dx - \frac{3\underline{\kappa} + \overline{\kappa}}{\sqrt{2n}}, \tag{28}$$

[7]For the sake of notational convenience, we take the universal constant in the theorem as $1/2$, although this is not the best known constant for the case of i.i.d. random variables. See [25] for a survey of the best known constants in the Berry-Esseen theorem.

where (26) and (28) follow from integration by parts and (27) follows from the Berry-Esseen theorem. We continue as follows

$$\int_{-\infty}^{\nu} \phi(x)\left[\Phi(\nu - x) - \Phi(\beta[\nu - x])\right] dx \tag{29}$$

$$= \int_{0}^{\infty} \phi(\nu - z) \int_{\beta z}^{z} \phi(\zeta) d\zeta dz \tag{30}$$

$$\geq (1 - \beta) \int_{0}^{\infty} \phi(\nu - z) z \phi(z) dz \tag{31}$$

$$\geq (1 - \beta)\phi(2\nu) \int_{0}^{-\nu} z\phi(z) dz \tag{32}$$

$$= (1 - \beta)\phi(2\nu)\left(\frac{1}{\sqrt{2\pi}} - \phi(\nu)\right). \tag{33}$$

By plugging (33) into (28), and recalling (20) and (21), we deduce that

$$\Gamma_n \leq f(\alpha) + \varepsilon + \frac{4\underline{\kappa} + \overline{\kappa}}{\sqrt{2n}}. \tag{34}$$

Thus for all sufficiently large (and even) n, we have

$$\Gamma_n < \varepsilon. \tag{35}$$

So by Lemma 14,

$$\liminf_{n \to \infty} \frac{\log M_{\mathrm{fb}}^*(n, \varepsilon) - nC}{\sqrt{n}} \geq \sqrt{V_{\min}} \Phi^{-1}(\alpha\varepsilon). \tag{36}$$

□

Remark 2: Although Theorem 1 uses feedback only at a single epoch, it still provides a strict improvement over the best non-feedback code. It is possible to prove a version of Theorem 1 for large ε (for which one begins the transmission using $Q_{\max}$ instead of $Q_{\min}$). But we shall not pursue this here because our aim with Theorem 1 is only to elucidate the idea behind timid/bold coding while avoiding the diffusion machinery used in our main achievability result, Theorem 2. Theorem 2 takes timid/bold coding to its natural limit by allowing the encoder to switch between timid and bold signaling schemes after each time-step. ◊

IV. Proof of Theorem 2

Define the right-hand side of (13) as

$$r(\varepsilon) := \begin{cases} \sqrt{V_{\min}} \Phi^{-1}\left(\frac{\varepsilon(1+\beta)}{2\beta}\right), & 0 < \varepsilon \leq \frac{\beta}{1+\beta}, \\ \sqrt{V_{\max}} \Phi^{-1}\left(\frac{\varepsilon(1+\beta)+(1-\beta)}{2}\right), & \frac{\beta}{1+\beta} < \varepsilon < 1. \end{cases} \tag{37}$$

We would like to invoke Lemma 14 with the controller

$$F(x^{k-1}, y^{k-1}) = \begin{cases} Q_{\max} & \text{if } \sum_{j=1}^{k-1} \left[\log \frac{W(y_j|x_j)}{q^*(y_j)} - C\right] \leq \sqrt{n} \cdot r(\varepsilon) \\ Q_{\min} & \text{if } \sum_{j=1}^{k-1} \left[\log \frac{W(y_j|x_j)}{q^*(y_j)} - C\right] > \sqrt{n} \cdot r(\varepsilon). \end{cases} \tag{38}$$

We would then compute the key quantity, namely the limit superior as $n \to \infty$ of

$$(F \circ W)\left(-r(\varepsilon) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\log \frac{W(Y_i|X_i)}{q^*(Y_i)} - C\right] \leq 0\right), \tag{39}$$

in Lemma 14 by showing that the discrete-time process therein converges to the solution of the stochastic differential equation

$$\xi_t = -\frac{r(\varepsilon)}{\sqrt{V_{\max}}} + \int_0^t \bar{\sigma}(\xi_s)\, dB_s \tag{40}$$

where $\boldsymbol{B}$ is a standard Brownian motion and

$$\bar{\sigma}(x) := \mathbf{1}\{x \leq 0\} + \beta \mathbf{1}\{x > 0\}, \tag{41}$$

for which the relevant probability can be computed [28]. The main obstacle to this approach is that the diffusion coefficient in (41) is not Lipschitz, and standard results for showing weak convergence to diffusions require the limiting process to have Lipschitz coefficients.

To circumvent this, in place of (38) we use a controller that switches from $Q_{\max}$ to $Q_{\min}$ in a continuous way. This requires showing that the resulting continuous-time limit is close in distribution to that of (40), which we show in Lemma 1 to follow.

Following Øksendal (e.g., [29, Def. 7.1.1]), we define a one-dimensional, time-homogeneous *Itô diffusion* as follows.

Definition 3 (Itô Diffusion): A time-homogeneous Itô diffusion is a stochastic process $\mathbf{X}$ satisfying a stochastic differential equation of the form

$$X_t = x_0 + \int_0^t b(X_s)\, ds + \int_0^t \sigma(X_s)\, dB_s, \tag{42}$$

for some one-dimensional Brownian motion $\mathbf{B}$ defined on the same sample space, where $b\colon \mathbb{R} \to \mathbb{R}$ and $\sigma\colon \mathbb{R} \to \mathbb{R}$ are measurable functions that satisfy

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq D|x - y|, \forall\, x, y \in \mathbb{R}, \tag{43}$$

for some constant $D \in \mathbb{R}^+$.

Remark 3: Since (43) ensures that the conditions in [29, Theorem 5.2.1] are satisfied, (42) has a unique solution.

A. A Convergence Result

Let $\{Z_{i,k}\}_{k=1}^{\infty}$, $i \in \{0, 1\}$ denote i.i.d. sequences of bounded random variables, which are also independent of each other, such that for any $k \in \mathbb{Z}^+$,

$$\mathbb{E}[Z_{0,k}] = \mathbb{E}[Z_{1,k}] = 0, \tag{44}$$

$$\mathbb{E}[Z_{1,k}^2] = 1, \tag{45}$$

$$\mathbb{E}[Z_{0,k}^2] = \beta^2, \tag{46}$$

with $\beta \in (0, 1)$. Given any $\delta \in (0, 1]$ and $x \in [0, \delta]$, define

$$\alpha_\delta(x) := \frac{1}{1 - \beta^2}\left(\left[1 - x\left(\frac{1-\beta}{\delta}\right)\right]^2 - \beta^2\right). \tag{47}$$

Via direct computation, one can verify that

$$\alpha_\delta(x) \in [0, 1], \tag{48}$$

for the given range of δ and x. Let μ_i denote the law of $Z_{i,1}$ for $i \in \{0, 1\}$. Define the probability measure

$$\mu_{\delta,x} := (1 - \alpha_\delta(x))\mu_0 + \alpha_\delta(x)\mu_1. \tag{49}$$

For any $\varepsilon \in (0, 1)$, define

$$s(\varepsilon) := \begin{cases} -\beta\Phi^{-1}\left(\frac{1}{2\beta}\varepsilon(1+\beta)\right), & \varepsilon \in (0, \frac{\beta}{1+\beta}], \\ -\Phi^{-1}\left(\frac{1}{2}[\varepsilon(1+\beta) + (1-\beta)]\right), & \varepsilon \in (\frac{\beta}{1+\beta}, 1). \end{cases} \tag{50}$$

For any $\varepsilon \in (0, 1)$ and $n \in \mathbb{Z}^+$,

$$S_0^{\delta,\varepsilon,n} := s(\varepsilon)\sqrt{n}, \tag{51}$$

$$\begin{aligned} S_{k+1}^{\delta,\varepsilon,n} := S_k^{\delta,\varepsilon,n} &+ \mathbf{1}\left\{S_k^{\delta,\varepsilon,n} \leq 0\right\} Z_{1,k+1} \\ &+ \mathbf{1}\left\{S_k^{\delta,\varepsilon,n} > \delta\sqrt{n}\right\} Z_{0,k+1} \\ &+ \mathbf{1}\left\{0 < S_k^{\delta,\varepsilon,n} \leq \delta\sqrt{n}\right\} Z_{2,k+1}, \end{aligned} \tag{52}$$

for all $k \in \mathbb{Z}^+$, where $Z_{2,k+1}$ has distribution

$$\mu_{\delta, S_k^{\delta,\varepsilon,n}/\sqrt{n}}$$

and is independent of $\{Z_{i,j}\}_{j=1}^{\infty}$, $i \in \{0, 1\}$ and $\{Z_{2,j}\}_{j=1}^{k}$.

Proposition 1: Consider any $\varepsilon \in (0, 1)$. For any $\kappa \in \mathbb{R}^+$, there exist $\delta_o \in (0, 1)$ and $n_o \in \mathbb{Z}^+$ such that for all $n \geq n_o$,

$$\Pr\left(\frac{1}{\sqrt{n}} S_n^{\delta_o,\varepsilon,n} \leq 0\right) \leq \varepsilon + \kappa. \tag{53}$$

Proof: Similar to [30, p. 43], we interpolate the discrete-time Markov process defined in (51) and (52) as follows

$$\xi_t^{\varepsilon,\delta,n} := \frac{1}{\sqrt{n}} S_{[nt]}^{\delta,\varepsilon,n}, \tag{54}$$

for any $t \in \mathbb{R}_+$, where $[nt]$ denotes the integer part of nt. We prove the claim by investigating the limiting behavior of $\xi_t^{\varepsilon,\delta,n}$ as $\delta \to 0$ and $n \to \infty$. To this end, we use several stochastic processes, which are defined next.

For any $\delta \in (0, 1]$, define $\sigma_\delta : \mathbb{R} \to \mathbb{R}$ as

$$\sigma_\delta(x) := \begin{cases} 1, & x \leq 0, \\ 1 - x\left(\frac{1-\beta}{\delta}\right), & 0 < x < \delta, \\ \beta, & x \geq \delta. \end{cases} \tag{55}$$

Clearly, $\sigma_\delta(\cdot)$ is Lipschitz continuous, positive and bounded. For any $\varepsilon \in (0, 1)$, we use (55) to define an Itô diffusion $\xi_t^{\varepsilon,\delta}$ that is the solution of the following stochastic differential equation:

$$\xi_t^{\varepsilon,\delta} = \xi_0^{\varepsilon,\delta} + \int_0^t \sigma_\delta(\xi_s^{\varepsilon,\delta})\, dB_s, \tag{56}$$

with $\xi_0^{\varepsilon,\delta} := s(\varepsilon)$. Further, define $\bar{\sigma} : \mathbb{R} \to \mathbb{R}$ as in (41):

$$\bar{\sigma}(x) := \mathbf{1}\{x \leq 0\} + \beta\mathbf{1}\{x > 0\}, \tag{57}$$

and let $\xi_t^{\varepsilon,0}$ be the solution of the following stochastic differential equation (cf. (40)):

$$\xi_t^{\varepsilon,0} = \xi_0^{\varepsilon,0} + \int_0^t \bar{\sigma}(\xi_s^{\varepsilon,0})\, dB_s, \tag{58}$$

with $\xi_0^{\varepsilon,0} := s(\varepsilon)$. Existence of a (weak) solution of (58) can be verified by using [31, Theorem 23.1]. Further, an expression for the transition probabilities of the Markov process $\xi_t^{\varepsilon,0}$, denoted by $P_t(x,y)$, is known [28],

$$P_t(x,y) = \frac{1}{\sqrt{2\pi t}} \begin{cases} \frac{1}{\beta} e^{-(x-y)^2/2\beta^2 t} - \frac{(\beta-1)}{\beta(\beta+1)} e^{-(x+y)^2/2\beta^2 t} \\ \frac{2\beta}{(\beta+1)} e^{-(x-\beta y)^2/2\beta^2 t} \\ \frac{2}{\beta(\beta+1)} e^{-(\beta x-y)^2/2\beta^2 t} \\ e^{-(x-y)^2/2t} + \frac{(\beta-1)}{(\beta+1)} e^{-(x+y)^2/2t} \end{cases} \tag{59}$$

for the four cases $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$, $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^-$, $(x,y) \in \mathbb{R}^- \times \mathbb{R}^+$, and $(x,y) \in \mathbb{R}^- \times \mathbb{R}^-$, respectively.

In Lemmas 1 and 2 to follow, the mode of convergence is the weak convergence of probability measures in the space of right-continuous functions with left limits defined on $[0,1]$, i.e., $D[0,1]$, endowed with the Skorohod topology (e.g., [32, Section 12]).

Lemma 1:

$$\boldsymbol{\xi}^{\varepsilon,\delta} \xrightarrow{w.} \boldsymbol{\xi}^{\varepsilon,0}, \text{ as } \delta \to 0. \tag{60}$$

Proof: The claim follows from a convergence result due to Kulinich [33, Theorem 2]. To verify the conditions of this theorem for our case, we note that the function f_δ in [33, p. 856] can be taken to be $f_\delta(x) = x$, either by direct calculation or by noticing the fact that the Itô diffusion $\xi_t^{\varepsilon,\delta}$ is in its natural scale. The condition regarding $f'_\delta(\cdot)\sigma_\delta(\cdot)$ is satisfied, since

$$\beta \le f'_\delta(x)\sigma_\delta(x) \le 1, \tag{61}$$

for all $\delta \in (0,1]$ and $x \in \mathbb{R}$. Further, the condition

$$\lim_{K\to\infty} \lim_{\delta\to 0} \Pr(|f_\delta(\xi_0^{\varepsilon,\delta})| > K) = 0, \tag{62}$$

is also clearly satisfied since

$$f_\delta(\xi_0^{\varepsilon,\delta}) = s(\varepsilon) \in \mathbb{R}. \tag{63}$$

Finally, the condition regarding the function G_δ, which is defined in [33, p. 857], can be verified to hold for our case, since for any $x \in \mathbb{R}$, we have

$$\lim_{\delta\to 0} G_\delta(x) = \lim_{\delta\to 0} \int_0^x \frac{du}{\sigma_\delta^2(u)} \tag{64}$$

$$= \frac{x}{\bar{\sigma}^2(x)}, \tag{65}$$

via direct calculation. Hence, we can apply [33, Theorem 2] to deduce the assertion, since the generalized diffusion used in this theorem, which is defined in [33, Eq. (3)], reduces to ξ_t^0 in our case. □

Lemma 2: For any $\delta \in (0,1]$,

$$\boldsymbol{\xi}^{\varepsilon,\delta,n} \xrightarrow{w.} \boldsymbol{\xi}^{\varepsilon,\delta} \text{ as } n \to \infty. \tag{66}$$

Proof: The claim follows from a convergence result of Kushner [30, Theorem 1]. Specifically, we apply this theorem with the Markov chain

$$\left\{ \frac{1}{\sqrt{n}} S_k^{\delta,\varepsilon,n} \right\}_{k=0}^{\infty}, \tag{67}$$

$\mathcal{F}_{k,n}$ denoting the sigma-algebra generated by $\frac{S_i^{\delta,\varepsilon,n}}{\sqrt{n}}$ for all $i \le k$, and the sequence of positive real numbers $\delta t_i^n = \frac{1}{n}$. The definition of $S_k^{\delta,\varepsilon,n}$, along with (55) and elementary algebra, ensures that for any $n \in \mathbb{Z}^+$, we have

$$\mathbb{E}\left[\left(S_{k+1}^{\delta,\varepsilon,n} - S_k^{\delta,\varepsilon,n} \right)^2 \middle| \mathcal{F}_{k,n} \right] = \sigma_\delta^2 \left(\frac{S_k^{\delta,\varepsilon,n}}{\sqrt{n}} \right) \text{ (a.s.)}, \tag{68}$$

for all $t \in \mathbb{R}_+$ and $k \in \{0, \ldots, [nt]\}$, and hence the condition in [30, Eq. (1)] is satisfied. The proof will be complete if we can verify that the six assumptions of Kushner [30, pg. 42] are satisfied for our case. Indeed, except (A4) and (A6), these assumptions trivially hold with the aforementioned choices. (A6) is evidently true since $\xi_t^{\varepsilon,\delta}$ is the unique (strong) solution of (56), whereas (A6) only requires (56) to possess a unique weak solution (e.g., [29, Chapter 5.3]). To verify (A4), let $K \in \mathbb{R}^+$ be a constant such that

$$\max\{|Z_{0,1}|, |Z_{1,1}|\} \le K \text{ (a.s.)}, \tag{69}$$

whose existence is ensured by the boundedness of the random variables. From the definition of $S_k^{\delta,\varepsilon,n}$, one can verify that for any $t \in \mathbb{R}^+$,

$$0 \le \mathbb{E}\left[\sum_{k=0}^{[nt]} \left| \frac{S_{k+1}^{\delta,\varepsilon,n} - S_k^{\delta,\varepsilon,n}}{\sqrt{n}} \right|^3 \right] \tag{70}$$

$$\le \frac{1}{n^{3/2}} K^3([nt]+1) \to 0, \text{ as } n \to \infty. \tag{71}$$

Evidently, (71) implies (A4) and hence we can apply [30, Theorem 1] to infer the assertion. □

In order to conclude the proof, it suffices to note that

$$\lim_{\delta\to 0} \Pr(\xi_1^{\varepsilon,\delta} \le 0) = \Pr(\xi_1^{\varepsilon,0} \le 0), \tag{72}$$

$$\lim_{n\to\infty} \Pr(\xi_1^{\varepsilon,\delta,n} \le 0) = \Pr(\xi_1^{\varepsilon,\delta} \le 0), \ \forall\, \delta \in (0,1], \tag{73}$$

$$\Pr(\xi_1^{\varepsilon,0} \le 0) = \varepsilon, \tag{74}$$

where (72) and (73) follow from Lemmas 1 and 2, respectively, along with [32, Theorem 12.5], whereas (74) follows from an elementary calculation by using (59). □

B. Proof of Theorem 2

Fix any $\varepsilon \in (0,1)$. If $\beta = 1$ then the result is implied by (9). Otherwise, assume that

$$\beta = \sqrt{\frac{V_{\min}}{V_{\max}}} \in (0,1). \tag{75}$$

Choose some $0 < \kappa < \frac{\varepsilon}{2}$ that also satisfies

$$\kappa \le \frac{\left[\varepsilon - \frac{\beta}{1+\beta}\right]}{4} \tag{76}$$

if $\varepsilon > \frac{\beta}{1+\beta}$. Recall the function $r : (0,1) \mapsto \mathbb{R}$ from (37)

$$r(\varepsilon) := \begin{cases} \sqrt{V_{\min}}\Phi^{-1}\left(\frac{\varepsilon(1+\beta)}{2\beta}\right), & 0 < \varepsilon \le \frac{\beta}{1+\beta}, \\ \sqrt{V_{\max}}\Phi^{-1}\left(\frac{\varepsilon(1+\beta)+(1-\beta)}{2}\right), & \frac{\beta}{1+\beta} < \varepsilon < 1. \end{cases} \tag{77}$$

Using $r(\cdot)$, define

$$R_n(\cdot) := C + \frac{r(\cdot)}{\sqrt{n}}. \tag{78}$$

As promised, we shall use Lemma 14 in the Appendix. To this end, define the controller F_ℓ via (cf. (38))

$$F_\ell(x^{k-1}, y^{k-1}) \tag{79}$$
$$= \begin{cases} Q_{\max} & \text{if} \sum_{j=1}^{k-1} \left[\log \frac{W(y_j|x_j)}{q^*(y_j)} - C\right] \le \sqrt{n} r(\varepsilon - \kappa), \\ Q_{\min} & \text{if} \sum_{j=1}^{k-1} \left[\log \frac{W(y_j|x_j)}{q^*(y_j)} - C\right] > \sqrt{n} r(\varepsilon - \kappa) \\ & \qquad\qquad + \frac{1}{\ell}\sqrt{nV_{\max}}, \\ \overline{Q}_{\ell,k} & \text{otherwise,} \end{cases}$$

where

$$\overline{Q}_{\ell,k} = \alpha_{\ell,k} Q_{\max} + (1 - \alpha_{\ell,k}) Q_{\min}$$

and, using the function defined in (47),

$$\alpha_{\ell,k} := \alpha_{1/\ell}\left(-\frac{r(\varepsilon-\kappa)}{\sqrt{V_{\max}}} + \frac{1}{\sqrt{nV_{\max}}} \sum_{j=1}^{k-1} \left[\log \frac{W(y_j|x_j)}{q^*(y_j)} - C\right]\right),$$

where we use the convention

$$\sum_{j=1}^{0} \left[\log \frac{W(y_j|x_j)}{q^*(y_j)} - C\right] = 0. \tag{80}$$

By Proposition 1, there exists ℓ in $\mathbb{Z}^+$ and n_0 in $\mathbb{Z}^+$ such that if $n \ge n_0$,

$$(F_\ell \circ W)\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left(\log \frac{W(Y_k|X_k)}{q^*(Y_k)} - C\right) \le r(\varepsilon - \kappa)\right) \le \varepsilon - \frac{\kappa}{2}.$$

Lemma 14 then implies that

$$\liminf_{n\to\infty} \frac{\log M^*_{\mathrm{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \ge r(\varepsilon - \kappa). \tag{81}$$

Since $r(\cdot)$ is continuous and $\kappa > 0$ is arbitrary, the result follows. □

V. Proof of Theorem 3

In light of (9) and the fact that $V_\varepsilon = V_{\min} = V_{\max}$, it suffices to show that

$$\limsup_{n\to\infty} \frac{\log M^*_{\mathrm{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \le \sqrt{V_{\min}} \Phi^{-1}(\varepsilon). \tag{82}$$

Our approach will be to show that, for the code to have rate approaching capacity and error probability diminishing to zero, then, with high probability, the empirical distribution of $\mathbf{X}^n$ needs to be near the set of capacity-achieving input distributions. Since $V_{\min} = V_{\max}$, if the empirical distribution of $\mathbf{X}^n$ is nearly capacity-achieving, then the sum of the conditional variances of $\mathrm{i}^*(X_k, Y_k)$ given the past is close to $nV_{\min}$ a.s., and a martingale central limit theorem [34] can be applied. We begin with a few definitions needed for the reduction to codes with empirically-capacity-achieving $\mathbf{X}^n$.

Definition 4: The *type* of a sequence $\mathbf{x}^n$ is the distribution $P_{\mathbf{x}^n}$ on $\mathcal{X}$ defined as

$$P_{\mathbf{x}^n}(a) := \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{x_k = a\}.$$

Definition 5: For a sequence $\mathbf{x}^n \in \mathcal{X}^n$,

$$\phi_W(\mathbf{x}^n) := \inf_{P \in \Pi^*_W} d_{\mathrm{TV}}(P, P_{\mathbf{x}^n}),$$

where $d_{\mathrm{TV}}(P,Q)$ denotes the total variation distance between distributions P and Q.

Definition 6: Let $\mathcal{T}^n$ denote the set of all probability distributions on $\mathcal{X}$ that are types of some length-n sequence, and define

$$\mathcal{T}^n_\gamma := \left\{T \in \mathcal{T}^n, \inf_{P \in \Pi^*_W} d_{\mathrm{TV}}(P,T) > \gamma\right\},$$
$$\mathcal{T}^{c,n}_\gamma := \left\{T \in \mathcal{T}^n, \inf_{P \in \Pi^*_W} d_{\mathrm{TV}}(P,T) \le \gamma\right\}.$$

Let $\mathbf{f}(m, \mathbf{y}^i) := [f(m, \mathbf{y}^0), f(m, \mathbf{y}^1), \ldots, f(m, \mathbf{y}^i)] \in \mathcal{X}^{i+1}$ with the convention that both $\mathbf{y}^0$ and $\mathbf{f}(m, \mathbf{y}^i)$ for $i \le -1$ are empty strings.

Definition 7: If Q is a probability distribution on $\mathcal{X}$ and $A \subset \mathcal{X}$ is such that $Q(A) > 0$, then $Q|_A$ is the probability measure

$$Q_A(x) = \begin{cases} \frac{Q(x)}{Q(A)} & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases} \tag{83}$$

Definition 8: Given a controller $F : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{P}(\mathcal{X})$, the $(*, \gamma)$-modified controller $\tilde{F}$ is defined as follows. For $k < n$ and $x^k \in \mathcal{X}^k$, let

$$\mathcal{X}_{x^k} = \{x : (x^k, x) \text{ is a prefix of some } x^n \in \mathcal{T}^{c,n}_\gamma\}. \tag{84}$$

Fix some $x_0 \in \mathcal{X}$ arbitrarily. Let $\tilde{F}(x^k, y^k)$ be a point-mass on x_0 if either $k \ge n$ or $k < n$ but $F(x^k, y^k)(\mathcal{X}_{x^k}) = 0$ (note that the latter includes the case in which $\mathcal{X}_{x^k}$ is empty). Otherwise, let

$$\tilde{F}(x^k, y^k) = F(x^k, y^k)|_{\mathcal{X}_{x^k}}. \tag{85}$$

Definition 9: Given a controller $F : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{P}(\mathcal{X})$, the (T, γ)-modified controller is defined as in the previous definition but with the type T in place of $\mathcal{T}^{c,n}_\gamma$.

Lemma 15 in the Appendix states for any $\rho_n > 0$

$$\log M^*_{\mathrm{fb}}(n,\varepsilon) \le \sup_F \inf_{q \in \mathcal{P}(\mathcal{Y}^n)} \log \rho_n - \log\left(\left(1 - \varepsilon - (F \circ W)\left(\frac{W(\mathbf{Y}^n|\mathbf{X}^n)}{q(\mathbf{Y}^n)} > \rho_n\right)\right)^+\right). \tag{86}$$

where F is a controller: $F : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{P}(\mathcal{X})$. Let P denote the distribution $F \circ W$. We will choose

$$q(\mathbf{y}^n) = \frac{1}{2} \prod_{k=1}^{n} q^*(y_k) + \frac{1}{2|\mathcal{T}^n_\gamma|} \sum_{T \in \mathcal{T}^n_\gamma} \prod_{k=1}^{n} q_T(y_k), \tag{87}$$

where

$$q_T(y) := \sum_{x \in \mathcal{X}} T(x) W(y|x).$$

This choice is inspired by an analogous choice by Fong and Tan [12, (37)], who in turn credit Hayashi [35].

Let $K_W := \max\left(2|\mathcal{X}|\nu_{\max}, \frac{8|\mathcal{X}|\nu_{\max}}{V_{\min}}\right)$ and χ_W denote the constant in [34, Corollary to Theorem 2] when γ in that result is taken to be $2\mathfrak{i}_{\max}$ here. Fix $0 < \gamma \leq \frac{V_{\min}}{4|\mathcal{X}|\nu_{\max}}$, and define

$$\delta_n := \chi_W \cdot \left(\frac{\log n}{\sqrt{n}(V_{\min} - \gamma K_W)^{3/2}} + \sqrt{\gamma K_W}\right),$$

$$r_n := \begin{cases} \sqrt{V_{\min} - \gamma K_W} \cdot \Phi^{-1} \cdot \Delta_n + \frac{\log 2}{\sqrt{n}} & \varepsilon \in \left(0, \frac{1}{2} - 3\delta_n\right], \\ \sqrt{V_{\min} + \gamma K_W} \cdot \Phi^{-1} \cdot \Delta_n + \frac{\log 2}{\sqrt{n}} & \varepsilon \in \left(\frac{1}{2} - 3\delta_n, 1\right), \end{cases} \tag{88}$$

$$\rho_n := \exp(nC + \sqrt{n} r_n),$$

where we have used the shorthand

$$\Delta_n = \varepsilon + 3\delta_n.$$

We now analyze the probability term in (86).

$$P\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n\right) \tag{89}$$

$$= P\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n \bigcap \phi_W(\mathbf{X}^n) \leq \gamma\right) \tag{90}$$

$$+ P\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n \bigcap \phi_W(\mathbf{X}^n) > \gamma\right)$$

$$= P\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n \bigcap \phi_W(\mathbf{X}^n) \leq \gamma\right)$$

$$+ \sum_{T \in \mathcal{T}_\gamma^n} P\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n \bigcap P_{\mathbf{X}^n} = T\right). \tag{91}$$

We will now apply the code modification technique of Fong and Tan [12]. Let P_* (resp. P_T) denote the distribution induced by the $(*, \gamma)$-modified (resp. (T, γ)-modified) code.

Lemma 3: For an event $\mathcal{E} \in \sigma(\mathbf{X}^n, \mathbf{Y}^n)$

$$P\left(\mathcal{E} \bigcap \phi_W(\mathbf{X}^n) \leq \gamma\right) \leq P_*(\mathcal{E}),$$
$$P\left(\mathcal{E} \bigcap P_{\mathbf{X}^n} = T\right) \leq P_T(\mathcal{E}).$$

Proof: For any $(\mathbf{x}^n, \mathbf{y}^n)$ such that $\phi_W(\mathbf{x}^n) \leq \gamma$,

$$P_*((\mathbf{x}^n, \mathbf{y}^n)) = \prod_{k=1}^n \tilde{F}(x_k|\mathbf{x}^{k-1}, \mathbf{y}^{k-1}) W(y_k|x_k) \tag{92}$$

$$= \prod_{k=1}^n \frac{F(x_k|\mathbf{x}^{k-1}, \mathbf{y}^{k-1}) W(y_k|x_k)}{F(\mathcal{X}_{x^{k-1}}|\mathbf{x}^{k-1}, \mathbf{y}^{k-1})} \tag{93}$$

$$\geq \prod_{k=1}^n F(x_k|\mathbf{x}^{k-1}, \mathbf{y}^{k-1}) W(y_k|x_k) \tag{94}$$

$$= P(\mathbf{x}^n, \mathbf{y}^n). \tag{95}$$

The proof of the second part is analogous. □

Application of the above lemma to (91) yields

$$P\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n\right) \tag{96}$$

$$\leq P_*\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n\right)$$

$$+ \sum_{T \in \mathcal{T}_\gamma^n} P_T\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n\right). \tag{97}$$

We will now upper bound the first term on the right-hand side of the above equation using a martingale central limit theorem. Let $\mathcal{F}_k = \sigma(M, Y_1, \ldots, Y_k)$, and

$$Z_k := \mathfrak{i}^*(X_k, Y_k) - \mathbb{E}_*[\mathfrak{i}^*(X_k, Y_k)|\mathcal{F}_{k-1}], \tag{98}$$

$$S_k := \sum_{j=1}^k Z_j.$$

Then,

$$P_*\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log \rho_n\right) \tag{99}$$

$$\stackrel{(a)}{\leq} P_*\left(\log \frac{\prod_{k=1}^n W(Y_k|X_k)}{(1/2)\prod_{k=1}^n q^*(Y_k)} \geq \log \rho_n\right)$$

$$= P_*\left(\sum_{k=1}^n \left(\log \frac{W(Y_k|X_k)}{q^*(Y_k)} - C\right) \geq \sqrt{n} r_n - \log 2\right)$$

$$\stackrel{(b)}{=} P_*\left(\sum_{k=1}^n (\mathfrak{i}^* - \mathbb{E}_*[\mathfrak{i}^*|\mathcal{F}_{k-1}]) \geq \sqrt{n} r_n - \log 2\right)$$

$$= P_*\left(\sum_{k=1}^n Z_k \geq \sqrt{n} r_n - \log 2\right), \tag{100}$$

where in (a), we have used the definition of $q(\mathbf{Y}^n)$ in (87), and in (b), we have used the fact that

$$\mathbb{E}_*[\mathfrak{i}^*(X_k, Y_k)|X_k] = \sum_{y \in \mathcal{Y}} W(y|X_k) \log \frac{W(y|X_k)}{Q^*(Y_k)} \leq C$$

[18, Theorem 4.5.1] and written $\mathfrak{i}^*$ for $\mathfrak{i}^*(X_k, Y_k)$.

Lemma 4: Let $\mathcal{G}_k = \sigma(S_1, \ldots, S_k)$ for $1 \leq k \leq n$, with $\mathcal{G}_0$ being the trivial σ-algebra. Then with $K_W = \max\left(2|\mathcal{X}|\nu_{\max}, \frac{8|\mathcal{X}|\nu_{\max}}{V_{\min}}\right)$, we have P_*-a.s.,

$$V_{\min} - \gamma K_W \leq \frac{1}{n}\sum_{k=1}^n \mathbb{E}_*[Z_k^2|\mathcal{G}_{k-1}] \leq V_{\min} + \gamma K_W$$

$$\left\|\frac{\sum_{k=1}^n \mathbb{E}_*[Z_k^2|\mathcal{G}_{k-1}]}{\sum_{k=1}^n \mathbb{E}_*[Z_k^2]} - 1\right\|_\infty \leq \gamma K_W.$$

Proof: The following chain of equalities holds P_*-a.s.,

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}_*[Z_k^2|\mathcal{F}_{k-1}] = \frac{1}{n}\sum_{k=1}^n \mathbb{E}_*[Z_k^2|X_k]$$

$$= \frac{1}{n}\sum_{k=1}^n \mathrm{Var}[\mathfrak{i}(X_k, Y_k)|X_k]$$

$$= \frac{1}{n}\sum_{k=1}^n \sum_{x \in \mathcal{X}} \mathbf{1}\{X_k = x\}\nu_x$$

$$= \sum_{x \in \mathcal{X}} P_{\mathbf{X}^n}(x)\nu_x.$$

Since $\phi_W(\mathbf{X}^n) \leq \gamma$, there exists a $\tilde{P} \in \Pi_W^*$ such that $d_{\rm TV}(\tilde{P}, P_{\mathbf{X}^n}) \leq 2\gamma$. Thus we have for each $x \in \mathcal{X}$

$$|\tilde{P}(x) - P_{\mathbf{X}^n}(x)| \leq d_{\rm TV}(\tilde{P}, P_{\mathbf{X}^n}) \leq 2\gamma.$$

Thus

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^n \mathbb{E}_*[Z_k^2|\mathcal{F}_{k-1}] &= \sum_{x\in\mathcal{X}} P_{\mathbf{X}^n}(x)\nu_x \\ &\leq \sum_{x\in\mathcal{X}} \left(\tilde{P}(x) + 2\gamma\right)\nu_x \\ &= \sum_{x\in\mathcal{X}} \tilde{P}(x)\nu_x + 2\gamma\sum_{x\in\mathcal{X}}\nu_x \\ &\leq V_{\min} + 2\gamma|\mathcal{X}|\nu_{\max},\end{aligned}$$

where the last step follows since for any $\tilde{P} \in \Pi_W^*$, $\sum_{x\in\mathcal{X}} \tilde{P}(x)\nu_x = V_{\min}$.

Similarly

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}_*[Z_k^2|\mathcal{F}_{k-1}] \geq V_{\min} - 2\gamma|\mathcal{X}|\nu_{\max}.$$

Since $\mathcal{G}_{k-1} \subseteq \mathcal{F}_{k-1}$, taking the conditional expectation with respect to $\mathcal{G}_{k-1}$, we get,

$$\begin{aligned}V_{\min} - 2\gamma|\mathcal{X}|\nu_{\max} &\leq \frac{1}{n}\sum_{k=1}^n \mathbb{E}_*[Z_k^2|\mathcal{G}_{k-1}] \\ &\leq V_{\min} + 2\gamma|\mathcal{X}|\nu_{\max}.\end{aligned}$$

To prove the second part, we note that P_*-a.s.,

$$\begin{aligned}\left|\frac{\sum_{k=1}^n \mathbb{E}_*[Z_k^2|\mathcal{G}_{k-1}]}{\sum_{k=1}^n \mathbb{E}_*[Z_k^2]} - 1\right| &\leq \left|\frac{V_{\min} + 2\gamma|\mathcal{X}|\nu_{\max}}{V_{\min} - 2\gamma|\mathcal{X}|\nu_{\max}} - 1\right| \\ &= \frac{4\gamma|\mathcal{X}|\nu_{\max}}{V_{\min} - 2\gamma|\mathcal{X}|\nu_{\max}} \\ &\leq \frac{8\gamma|\mathcal{X}|\nu_{\max}}{V_{\min}},\end{aligned}$$

provided $\gamma \leq \frac{V_{\min}}{4|\mathcal{X}|\nu_{\max}}$.

The statement of the lemma now follows since $K_W = \max\left(2|\mathcal{X}|\nu_{\max}, \frac{8|\mathcal{X}|\nu_{\max}}{V_{\min}}\right)$. □

Continuing the chain of expressions in (100),

$$\begin{aligned}&P_*\left(\log\frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log\rho_n\right) \qquad (101)\\ &\leq P_*\left(\sum_{k=1}^n Z_k \geq \sqrt{n}r_n - \log 2\right), \\ &\overset{(a)}{\leq} P_*\left(\frac{1}{\sqrt{\sum_{k=1}^n \mathbb{E}_*[Z_k^2]}}\sum_{k=1}^n Z_k \geq \Phi^{-1}(\varepsilon + 3\delta_n)\right) \\ &\overset{(b)}{\leq} 1 - \varepsilon - 3\delta_n + \chi_W \cdot \\ &\left(\frac{n\log n}{(\sum_{k=1}^n \mathbb{E}_*[Z_k^2])^{3/2}} + \left\|\frac{\sum_{k=1}^n \mathbb{E}_*[Z_k^2|\mathcal{G}_{k-1}]}{\sum_{k=1}^n \mathbb{E}_*[Z_k^2]} - 1\right\|_\infty^{1/2}\right) \\ &\overset{(c)}{\leq} 1 - \varepsilon - 3\delta_n + \chi_W \cdot \left(\frac{\log n}{\sqrt{n}(V_{\min} - \gamma K_W)^{3/2}} + \sqrt{\gamma K_W}\right) \\ &= 1 - \varepsilon - 2\delta_n, \qquad (102)\end{aligned}$$

where, for (a) we have used

$$n(V_{\min} - \gamma K_W) \leq \sum_{k=1}^n \mathbb{E}_*[Z_k^2] \leq n(V_{\min} + \gamma K_W)$$

from Lemma 4, for (b), we have used the martingale central limit theorem [34, Corollary to Theorem 2], taking the constant as χ_W (which only depends upon $\mathrm{i}_{\max}$ since $|Z_k| \leq 2\mathrm{i}_{\max}$ a.s.), and for (c), we have used Lemma 4.

Moving to the second term in (97), and noting that $q(\mathbf{Y}^n) \geq \frac{1}{2|\mathcal{T}_\gamma^n|}\prod_{k=1}^n q_T(Y_k)$, we get

$$\begin{aligned}&\sum_{T\in\mathcal{T}_\gamma^n} P_T\left(\log\frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log\rho_n\right) \\ &\leq \sum_{T\in\mathcal{T}_\gamma^n} P_T\left(\log\frac{\prod_{k=1}^n W(Y_k|X_k)}{\frac{1}{2|\mathcal{T}_\gamma^n|}\prod_{k=1}^n q_T(Y_k)} \geq \log\rho_n\right) \\ &= \sum_{T\in\mathcal{T}_\gamma^n} P_T\left(\sum_{k=1}^n \log\frac{W(Y_k|X_k)}{q_T(Y_k)} \geq \log\rho_n - \log 2|\mathcal{T}_\gamma^n|\right).\end{aligned}$$

Consider

$$\begin{aligned}&\sum_{k=1}^n \mathbb{E}_T\left[\log\frac{W(Y_k|X_k)}{q_T(Y_k)}\middle|\mathcal{F}_{k-1}\right] \\ &= \sum_{x\in\mathcal{X}}\sum_{k=1}^n \mathbb{E}_T\left[\log\frac{W(Y_k|X_k)}{q_T(Y_k)}\middle|X_k = x\right]\mathbf{1}\{X_k = x\} \\ &= \sum_{x\in\mathcal{X}}\sum_{k=1}^n\sum_{y\in\mathcal{Y}} W(y|x)\log\frac{W(y|x)}{q_T(y)}\mathbf{1}\{X_k = x\} \\ &= n\sum_{x\in\mathcal{X}} T(x)\sum_{y\in\mathcal{Y}} W(y|x)\log\frac{W(y|x)}{q_T(y)} \\ &= nI(T;W).\end{aligned}$$

Recall that for any $P \in \Pi_W^*$ and $T \in \mathcal{T}_\gamma^n$, $d_{\rm TV}(P,T) > \gamma > 0$, hence $I(T;W) < C$. Let $K_T := C - I(T;W) > 0$, and $\tilde{\imath}_{\max,T} := \max_{x,y:W(y|x)q_T(y)>0}\left|\log\frac{W(y|x)}{q_T(y)}\right|$.

We now show that $\tilde{\imath}_{\max,T} \leq 2\log n$ P_T-a.s., for all sufficiently large n. Let $W_{\min} := \min_{x,y:W(y|x)>0} W(y|x)$ and $q_{T,\min} := \min_{q_T(y)>0} q_T(y)$. Then

$$\begin{aligned}q_{T,\min} &:= \min_{q_T(y)>0}\sum_x T(x)W(y|x) \\ &\geq \min_{x,y:W(y|x)>0} W(y|x)\min_{x:T(x)>0} T(x) = \frac{W_{\min}}{n},\end{aligned}$$

where the last equality follows since T is the type of a sequence. Thus

$$\begin{aligned}\tilde{\imath}_{\max,T} &= \max_{x,y:W(y|x)q_T(y)>0}\left|\log\frac{W(y|x)}{q_T(y)}\right| \\ &\leq \max_{x,y:W(y|x)q_T(y)>0}|\log W(y|x)| \\ &\qquad + \max_{y:q_T(y)>0}|\log q_T(y)| \\ &\leq |\log W_{\min}| + \left|\log\frac{W_{\min}}{n}\right| \\ &= \log\frac{n}{W_{\min}^2} \\ &\leq 2\log n,\end{aligned}$$

for all sufficiently large n.

Defining $\tilde{Z}_k := \log \frac{W(Y_k|X_k)}{q_T(Y_k)} - E_T\left[\log \frac{W(Y_k|X_k)}{q_T(Y_k)}\middle|\mathcal{F}_{k-1}\right]$, we have

$$\begin{aligned}
&\sum_{T\in\mathcal{T}_\gamma^n} P_T\left(\sum_{k=1}^n \log\frac{W(Y_k|X_k)}{q_T(Y_k)} \geq \log\rho_n - \log 2|\mathcal{T}_\gamma^n|\right)\\
&= \sum_{T\in\mathcal{T}_\gamma^n} P_T\Bigg(\sum_{k=1}^n\Bigg(\log\frac{W(Y_k|X_k)}{q_T(Y_k)} - E_T\left[\log\frac{W(Y_k|X_k)}{q_T(Y_k)}\middle|F_{k-1}\right]\Bigg) \geq nK_T + \sqrt{n}r_n - \log 2|\mathcal{T}_\gamma^n|\Bigg)\\
&= \sum_{T\in\mathcal{T}_\gamma^n} P_T\left(\sum_{k=1}^n \tilde{Z}_k \geq nK_T + \sqrt{n}r_n - \log 2|\mathcal{T}_\gamma^n|\right)\\
&\overset{(a)}{\leq} \sum_{T\in\mathcal{T}_\gamma^n} P_T\left(\sum_{k=1}^n \tilde{Z}_k \geq nK_T + \sqrt{n}r_n - |\mathcal{X}|\log 2(n+1)\right)\\
&\overset{(b)}{\leq} \sum_{T\in\mathcal{T}_\gamma^n} P_T\left(\sum_{k=1}^n \tilde{Z}_k \geq \frac{nK_T}{2}\right)\\
&\overset{(c)}{\leq} \sum_{T\in\mathcal{T}_\gamma^n} \exp\left(-\frac{nK_T^2}{128\log^2 n}\right)\\
&\overset{(d)}{\leq} \sum_{T\in\mathcal{T}_\gamma^n} \exp\left(-\frac{nK}{\log^2 n}\right)\\
&= |\mathcal{T}_\gamma^n|\exp\left(-\frac{nK}{\log^2 n}\right)\\
&\leq (n+1)^{|\mathcal{X}|}\exp\left(-\frac{nK}{\log^2 n}\right)\\
&\overset{(e)}{\leq} \delta_n, \qquad (103)
\end{aligned}$$

where, (a) follows since $|\mathcal{T}_\gamma^n| \leq |\mathcal{T}^n| \leq (n+1)^{|\mathcal{X}|}$, (b) follows since $\sqrt{n}r_n - |\mathcal{X}|\log 2(n+1) \geq -\frac{nK_T}{2}$ for all sufficiently large n, (c) follows from Azuma's inequality [36, (3.3), p. 61], and noting that $|\tilde{Z}_k| \leq 2\tilde{i}_{\max,T} \leq 4\log n$, (d) follows from defining $K := \min_{T\in\mathcal{T}_\gamma^n} \frac{K_T^2}{128}$, and (e) holds for all sufficiently large n.

From (97), (102), and (103), we get

$$P\left(\log\frac{\prod_{k=1}^n W(Y_k|X_k)}{q(\mathbf{Y}^n)} \geq \log\rho_n\right) \leq 1-\varepsilon-\delta_n.$$

Plugging the above inequality in (86),

$$\log M^*_{\text{fb}}(n,\varepsilon) \leq \log\rho_n - \log\delta_n,$$

i.e.,

$$\frac{\log M^*_{\text{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \leq r_n - \frac{\log\delta_n}{\sqrt{n}}. \qquad (104)$$

Using the definition of r_n in (88) and taking the limit gives

$$\limsup_{n\to\infty}\frac{\log M^*_{\text{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \leq \sqrt{V_{\min} - \gamma K_W}\,\Phi^{-1}\left(\varepsilon + \chi_W\sqrt{\gamma K_W}\right),$$

if $\varepsilon \in \left(0, \frac{1}{2} - \chi_W\sqrt{\gamma K_W}\right]$, and

$$\limsup_{n\to\infty}\frac{\log M^*_{\text{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \leq \sqrt{V_{\min} + \gamma K_W}\,\Phi^{-1}\left(\varepsilon + \chi_W\sqrt{\gamma K_W}\right).$$

if $\varepsilon \in \left(\frac{1}{2} - \chi_W\sqrt{\gamma K_W}, 1\right)$. Now taking $\gamma \to 0$ gives

$$\limsup_{n\to\infty}\frac{\log M^*_{\text{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \leq \sqrt{V_{\min}}\,\Phi^{-1}(\varepsilon),$$

proving (82) as desired. □

VI. Proof of Theorem 4

We begin with a few definitions from stochastic calculus. Throughout we assume that the filtration under consideration is right-continuous and complete (via e.g. [31, Lemma 7.8, p. 124]).

Definition 10: A process $\mathbf{N}$ is called a *local martingale* with respect to a filtration $(\mathcal{F}_t : t \geq 0)$ if N_t is $\mathcal{F}_t$-measurable for each t and there exists an increasing sequence of stopping times T_n, such that $T_n \to \infty$ and the stopped and shifted processes $\mathbf{N}^{T_n} := (N_{\min\{t,T_n\}} - N_0 : t \geq 0)$ are $(\mathcal{F}_t : t \geq 0)$-martingales for each n.

Definition 11: The *quadratic variation* of a continuous local martingale $\mathbf{N}$ is an a.s. unique continuous process of locally finite variation, $[\mathbf{N}]$, such that $\mathbf{N}^2 - [\mathbf{N}]$ is a local martingale. The existence and uniqueness of such process is guaranteed by [31, Theorem 17.5, p. 332].

Definition 12: A stochastic process is said to be $\mathcal{F}_t$*-predictable* if it is measurable with respect to the σ-algebra generated by all left-continuous $\mathcal{F}_t$-adapted processes.

By taking $q(\mathbf{y}^n) = \prod_{i=1}^n q^*(y_i)$ in (153) in Lemma 15 in the Appendix (which is almost certainly a source of looseness in the bound), we get, for any $\rho_n > 0$,

$$\log M^*_{\text{fb}}(n,\varepsilon) \leq \sup_F \log\rho_n - \log\left(\left(1-\varepsilon-P\left(\sum_{i=1}^n \mathfrak{i}^*(X_k,Y_k) > \log\rho_n\right)\right)^+\right), \qquad (105)$$

where the supremum is over controllers: $F : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{P}(\mathcal{X})$, and P denotes the distribution $F\circ W$. We use (153) over (154)-(155) in Lemma 15 because it yields a finite-n result ((129) to follow). Fix an arbitrary $\kappa > 0$, let $K_W := 16\mathfrak{i}^2_{\max}\nu_{\max}/\nu_{\min}$, and define

$$\delta_n := \frac{K_W}{\kappa^2\sqrt{n}}, \qquad (106)$$

$$\rho_n := \exp(nC + \sqrt{n}r_n), \qquad (107)$$

$$r_n := \begin{cases} \sqrt{\nu_{\min}}\,\Phi^{-1}\left(\frac{(1+\lambda)}{2\lambda}(\varepsilon + 2\delta_n)\right) + \kappa, \\ \sqrt{\nu_{\max}}\,\Phi^{-1}\left(\frac{(\varepsilon+2\delta_n)(1+\lambda)+(1-\lambda)}{2}\right) + \kappa, \end{cases} \qquad (108)$$

for the cases $0 < \varepsilon \leq \frac{\lambda}{1+\lambda} - 2\delta_n$ and $\frac{\lambda}{1+\lambda} - 2\delta_n < \varepsilon < 1$, respectively.

The proof will consist of the following steps:

1) We will define a martingale sequence $(S_k, 1 \le k \le n)$ such that $P\left(\sum_{k=1}^n \mathfrak{i}^*(X_k, Y_k) \ge \log \rho_n\right) \le P(S_n \ge r_n)$.
2) We will embed the martingale sequence $(S_k, 1 \le k \le n)$ in a Brownian motion $\mathbf{B}$ such that $S_k = B_{T_k}, 1 \le k \le n$, where $(T_k, 1 \le k \le n)$ are stopping times.
3) We will construct a process $\psi_t \in [\sqrt{\nu_{\min}}, \sqrt{\nu_{\max}}]$ and a Brownian motion $\mathbf{W}$ such that $\int_0^1 \psi_s \, dW_s \approx B_{T_n}$.
4) Applying a theorem from stochastic calculus, we will "mimic" the above Itô process by a solution of a SDE $\hat{\boldsymbol{\xi}}$.
5) Using McNamara's result on the optimal control of diffusion processes [11], we will upper bound the probability $P\left(\hat{\xi}_1 \ge 0\right)$ which will yield an upper bound on $P\left(\int_0^1 \psi_s \, dW_s \ge r_n\right)$.

Proceeding, define

$$\begin{aligned}
\mathcal{F}_k &:= \sigma(M, Y_1, \ldots, Y_k), \\
Z_k &:= \frac{1}{\sqrt{n}} \left(\mathfrak{i}^*(X_k, Y_k) - \mathbb{E}[\mathfrak{i}^*(X_k, Y_k)|\mathcal{F}_{k-1}]\right) \\
S_k &:= \sum_{j=1}^k Z_j, \\
\mathcal{G}_k &:= \sigma(S_1, \ldots, S_k)
\end{aligned}$$

We note that

$$|Z_k| \le \frac{2}{\sqrt{n}} \mathfrak{i}_{\max} \quad P-\text{a.s.} \tag{109}$$

Lemma 5: The sequence $(S_k, 1 \le k \le n)$ is a martingale with respect to the filtration $(\mathcal{G}_k, 1 \le k \le n)$ such that

$$\mathbb{E}[Z_k^2|\mathcal{G}_{k-1}] \in \left[\frac{\nu_{\min}}{n}, \frac{\nu_{\max}}{n}\right], \tag{110}$$

and

$$P\left(\sum_{k=1}^n \mathfrak{i}^*(X_k, Y_k) \ge \log \rho_n\right) \le P(S_n \ge r_n).$$

Proof: Since $\mathcal{G}_k \subseteq \mathcal{F}_k$ and

$$\mathbb{E}[Z_k|\mathcal{F}_{k-1}] = 0,$$

taking the conditional expectation with respect to $\mathcal{G}_{k-1}$, we get

$$\mathbb{E}[Z_k|\mathcal{G}_{k-1}] = 0.$$

Thus the sequence $(S_k, 1 \le k \le n)$ is a martingale with respect to the filtration $(\mathcal{G}_k, 1 \le k \le n)$. Moreover

$$\mathbb{E}[Z_k^2|\mathcal{F}_{k-1}] = \frac{1}{n} \sum_{x \in \mathcal{X}} \mathbf{1}\{X_k = x\} \nu_x \in \left[\frac{\nu_{\min}}{n}, \frac{\nu_{\max}}{n}\right]. \tag{111}$$

Once again taking the conditional expectation with respect to $\mathcal{G}_{k-1}$, we get

$$\mathbb{E}[Z_k^2|\mathcal{G}_{k-1}] \in \left[\frac{\nu_{\min}}{n}, \frac{\nu_{\max}}{n}\right]. \tag{112}$$

Now consider

$$\begin{aligned}
&P\left(\sum_{k=1}^n \mathfrak{i}^*(X_k, Y_k) \ge \log \rho_n\right) \\
&= P\left(\frac{1}{\sqrt{n}} \sum_{k=1}^n (\mathfrak{i}^*(X_k, Y_k) - C) \ge r_n\right) \\
&\le P\left(\frac{1}{\sqrt{n}} \sum_{k=1}^n \left(\mathfrak{i}^*(X_k, Y_k) - \mathbb{E}[\mathfrak{i}^*(X_k, Y_k)|\mathcal{F}_{k-1}]\right) \ge r_n\right) \\
&= P(S_n \ge r_n),
\end{aligned} \tag{113}$$

where in the middle step we have used the fact that [18, Theorem 4.5.1]

$$\begin{aligned}
\mathbb{E}[\mathfrak{i}^*(X_k, Y_k)|\mathcal{F}_{k-1}] &= \mathbb{E}[\mathfrak{i}^*(X_k, Y_k)|X_k] \\
&= \sum_{y \in \mathcal{Y}} W(y|X_k) \log \frac{W(y|X_k)}{Q^*(Y_k)} \\
&\le C.
\end{aligned} \tag{114}$$

□

Lemma 6: There exists a Brownian motion $\mathbf{B}$, and a sequence of non-decreasing stopping times $T_1, \ldots, T_n$ such that

$$S_k = B_{T_k} \text{ a.s.} \quad k \in \{1, \ldots, n\},$$

and if $\tilde{\mathcal{G}}_k = \sigma(S_1, T_1 \ldots, S_k, T_k)$, and $\tau_k = T_k - T_{k-1}$ (with $T_0 = 0$), then

$$E[\tau_k|\tilde{\mathcal{G}}_{k-1}] = \mathbb{E}[Z_k^2|\mathcal{G}_{k-1}], \tag{115}$$

$$E[\tau_k^2|\tilde{\mathcal{G}}_{k-1}] \le 4\mathbb{E}[Z_k^4|\mathcal{G}_{k-1}]. \tag{116}$$

Proof: The lemma is a straightforward application of [31, Theorem 14.16, p. 279] to the martingale sequence $(S_k, 1 \le k \le n)$. □

We are now at step 3) of the above program. The martingale $(S_k, 1 \le k \le n)$ can be viewed as a Brownian motion $\boldsymbol{B}$ observed at different stopping times. In particular, we have $S_n = B_{T_n}$. We next perform a stochastic change-of-time so that S_n can be viewed as an Itô process evaluated at a nearly deterministic time.

Lemma 7: There exists a filtration $\mathcal{H}_t$, an $\mathcal{H}_t$-predictable process $\boldsymbol{\psi}$, an $\mathcal{H}_t$ Brownian motion $\mathbf{W}$, and an $\mathcal{H}_t$-stopping time T_n^* such that

1) $\sqrt{\nu}_{\min} \le \psi_t \le \sqrt{\nu}_{\max}$ a.s.
2) $\int_0^{T_n^*} \psi_t \, dW_t = B_{T_n} = S_n$.
3) $\mathbb{E}[(T_n^* - 1)^2] \le \frac{K_W^{(1)}}{n}$, where $K_W^{(1)} := 64 \mathfrak{i}_{\max}^4 / \nu_{\min}^2$.

Proof: Define the increasing random times $\{T_k^*\}_{k=0}^n$ via $T_0^* = 0$ and

$$T_k^* = \sum_{j=1}^k \frac{\tau_j}{n \mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}]}, \quad 1 \le k \le n.$$

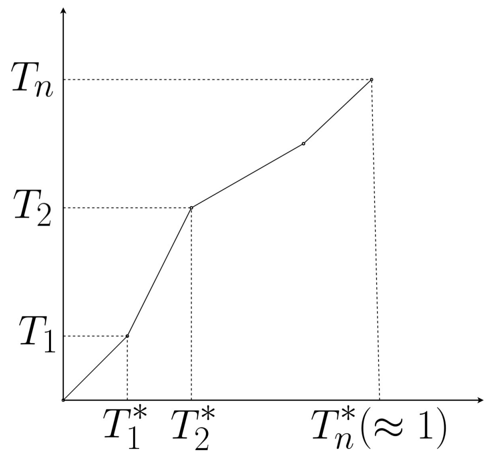

Fig. 2. Plot of A_t vs t for a fixed ω in the sample space.

Then define

$$\psi_t = \begin{cases} \sqrt{n\mathbb{E}[\tau_1|\tilde{\mathcal{G}}_0]} & T_0^* \le t \le T_1^* \\ \sqrt{n\mathbb{E}[\tau_2|\tilde{\mathcal{G}}_1]} & T_1^* < t \le T_2^* \\ \quad\vdots & \quad\vdots \\ \sqrt{n\mathbb{E}[\tau_n|\tilde{\mathcal{G}}_{n-1}]} & T_{n-1}^* < t \le T_n^* \\ \sqrt{\nu}_{\text{min}} & t > T_n^*. \end{cases} \tag{117}$$

Then, from the above definition, (115), and (110), it is clear that $\sqrt{\nu}_{\text{min}} \le \psi_t \le \sqrt{\nu}_{\text{max}}$ a.s.

We now employ the change-of-time method (see [37]). Let $A_t := \int_0^t \psi_s^2\, ds$. We note that $\mathbf{A}$ is continuous and strictly increasing, and we define the following time-changed process $\mathbf{N} := \mathbf{B} \circ \mathbf{A}$, i.e.,

$$N_t = B_{A_t} = B_{\int_0^t \psi_s^2\, ds},$$

and

$$\mathcal{H}_t := \sigma(B_{A_s}, 0 \le s \le t).$$

We have that (see Figure 2)

$$A_{T_k^*} = \int_0^{T_k^*} \psi_t^2\, dt = \sum_{j=1}^{k} \tau_j = T_k, \quad 1 \le k \le n.$$

Hence, $T_n^* = A_{T_n}^{-1}$, where $A_t^{-1}(\omega)$ is the inverse of $A_t(\omega)$ for each ω in the given sample space. We can write

$$T_n = \inf\{t > 0; A_t^{-1} > T_n^*\}.$$

Noting that A_t^{-1} is continuous and T_n is a $\sigma(B_s, 0 \le s \le t)$-stopping time, applying [31, Proposition 7.9, p. 124], we conclude that $A_{T_k}^{-1} = T_k^*$ is an $\mathcal{H}_t$-stopping time for each k (the role of process X_t in [31, Proposition 7.9, p. 124] is played by A_t^{-1} here).

Now applying [31, Theorem 17.24, p. 344] we get that $\mathbf{N}$ is a continuous local martingale with respect to the filtration $\mathcal{H}_t$ with quadratic variation

$$[\mathbf{N}] = [\mathbf{B}] \circ \mathbf{A} = \mathbf{A}, \tag{118}$$

since $[B]_t = t$ [31, Theorem 18.3, p. 352]. Now we follow the proof of [17, Theorem 4.2, p. 170]. Define $\mathbf{W}$ as

$$W_t = \int_0^t \frac{1}{\psi_s}\, dN_s.$$

Then $\mathbf{W}$ is a continuous local martingale with quadratic variation ([31, Lemma 17.10, p.335], noting that $1/\psi_s$ is a step process)

$$[W]_t = \int_0^t \frac{1}{\psi_s^2}\, d[N]_s = \int_0^t \frac{1}{\psi_s^2} \psi_s^2\, ds = t,$$

where we have used [31, Proposition 17.14, p. 338] for the middle equality. Hence $\mathbf{W}$ is a standard Brownian motion with respect to the filtration $\mathcal{H}_t$ [31, Theorem 18.3, p. 352].

Noting that there exists a (random) partition $0 = t_0 < t_1, \ldots, < t_l = t$ such that $\boldsymbol{\psi}$ is constant on $(t_k, t_{k+1}]$ for $0 \le k \le l-1$, we can write

$$\begin{aligned} \int_0^t \psi_s\, dW_s &= \sum_{k=0}^{l-1} \psi_{t_k}(W_{t_{k+1}} - W_{t_k}) \\ &= \sum_{k=0}^{l-1} \psi_{t_k} \frac{1}{\psi_{t_k}}(N_{t_{k+1}} - N_{t_k}) = N_t. \end{aligned}$$

Thus

$$\int_0^{T_n^*} \psi_s\, dW_s = N_{T_n^*} = B_{A_{T_n^*}} = B_{T_n} = S_n. \tag{119}$$

Since T_k^* is an $\mathcal{H}_t$ stopping time for each k, ψ is adapted to $\mathcal{H}_t$. Since it is left continuous, it is also predictable.

Now we bound $\mathbb{E}[(T_n^* - 1)^2]$:

$$\begin{aligned} \mathbb{E}[(T_n^* - 1)^2] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} \frac{\tau_j}{n\mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}]} - 1\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{j=1}^{n} \frac{\tau_j - \mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}]}{n\mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}]}\right)^2\right] \\ &\overset{(a)}{\le} \frac{1}{\nu_{\text{min}}^2} \mathbb{E}\left[\left(\sum_{j=1}^{n} \tau_j - \mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}]\right)^2\right] \\ &\overset{(b)}{=} \frac{1}{\nu_{\text{min}}^2} \mathbb{E}\left[\sum_{j=1}^{n} \left(\tau_j - \mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}]\right)^2\right] \\ &\overset{(c)}{\le} \frac{1}{\nu_{\text{min}}^2} \mathbb{E}\left[\sum_{j=1}^{n} \mathbb{E}[\tau_j^2|\tilde{\mathcal{G}}_{j-1}]\right] \\ &\overset{(d)}{\le} \frac{4}{\nu_{\text{min}}^2} \mathbb{E}\left[\sum_{j=1}^{n} \mathbb{E}[Z_j^4|\mathcal{G}_{j-1}]\right] \\ &\overset{(e)}{\le} \frac{4}{\nu_{\text{min}}^2} \mathbb{E}\left[\sum_{j=1}^{n} \frac{16\mathrm{i}_{max}^4}{n^2}\right] \\ &= \frac{64\mathrm{i}_{\text{max}}^4}{n\nu_{\text{min}}^2} \\ &\overset{(f)}{=} \frac{K_W^{(1)}}{n}. \end{aligned}$$

Here, (a) follows from (112) and (115), (b) follows from noting that the sequence $(\tau_j - \mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}], 1 \leq j \leq n)$ is a martingale difference sequence with respect to the filtration $(\tilde{\mathcal{G}}_j, 1 \leq j \leq n)$, making $(\sum_{j=1}^k \tau_j - \mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}], 1 \leq k \leq n)$ a martingale and the orthogonal increment property of martingales [38, Theorem 5.4.6], (c) follows from $\mathbb{E}[(\tau_j - \mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}])^2|\mathcal{G}_{j-1}] = \mathbb{E}[\tau_j^2|\tilde{\mathcal{G}}_{j-1}] - \left(\mathbb{E}[\tau_j|\tilde{\mathcal{G}}_{j-1}]\right)^2$, (d) follows from (116), (e) follows since $|Z_j| \leq \frac{2}{\sqrt{n}}\mathrm{i}_{\max}$ a.s. from (109), and (f) follows from defining $K_W^{(1)} := 64\mathrm{i}_{\max}^4/\nu_{\min}^2$. □

Now define

$$\xi_t := -(r_n - \kappa) + \int_0^t \psi_s \, dW_s. \tag{120}$$

We have the following lemma.

Lemma 8:

$$P\left(\int_0^{T_n^*} \psi_s \, dW_s \geq r_n\right) \leq P\left(\xi_1 \geq 0\right) + \delta_n.$$

Proof:

$$P\left(\int_0^{T_n^*} \psi_s \, dW_s \geq r_n\right) = P\left(\int_0^1 \psi_s \, dW_s + \theta_n \geq r_n\right),$$

where we have defined θ_n as

$$\theta_n := \int_0^\infty \mathbf{1}\{1 < s \leq T_n^*\}\psi_s \, dW_s - \int_0^\infty \mathbf{1}\{T_n^* \leq s < 1\}\psi_s \, dW_s.$$

The second moment of θ_n can be bounded as

$$\begin{aligned}
\mathbb{E}[\theta_n^2] &\overset{(a)}{\leq} 2\mathbb{E}\left[\left(\int_0^\infty \mathbf{1}\{1 < s \leq T_n^*\}\psi_s \, dW_s\right)^2\right] \\
&\quad + 2\mathbb{E}\left[\left(\int_0^\infty \mathbf{1}\{T_n^* \leq s < 1\}\psi_s \, dW_s\right)^2\right] \\
&\overset{(b)}{=} 2\mathbb{E}\left[\int_0^\infty \mathbf{1}\{1 < s \leq T_n^*\}\psi_s^2 \, ds\right] \\
&\quad + 2\mathbb{E}\left[\int_0^\infty \mathbf{1}\{T_n^* \leq s < 1\}\psi_s^2 \, ds\right] \\
&= 2\mathbb{E}\left[\mathbf{1}\{1 < T_n^*\}\int_1^{T_n^*} \psi_s^2 \, ds\right] \\
&\quad + 2\mathbb{E}\left[\mathbf{1}\{T_n^* < 1\}\int_{T_n^*}^1 \psi_s^2 \, ds\right] \\
&\leq 2\nu_{\max}\mathbb{E}[|T_n^* - 1|] \\
&\leq 2\nu_{\max}\sqrt{\mathbb{E}[(T_n^* - 1)^2]} \\
&\overset{(c)}{\leq} \frac{K_W}{\sqrt{n}}.
\end{aligned}$$

Here, for (a) we have used the inequality $(a-b)^2 \leq 2a^2 + 2b^2$, for (b) we have used [17, Problem 2.18, p. 144], and for (c) we have used Lemma 7, and recalling $K_W = 16\mathrm{i}_{\max}^2\nu_{\max}/\nu_{\min} = 2\nu_{\max}\sqrt{K_W^{(1)}}$.

Thus

$$\begin{aligned}
&P\left(\int_0^{T_n^*} \psi_s \, dW_s \geq r_n\right) \\
&= P\left(\int_0^1 \psi_s \, dW_s + \theta_n \geq r_n\right) \\
&= P\left(\int_0^1 \psi_s \, dW_s + \theta_n \geq r_n \bigcap |\theta_n| \leq \kappa\right) \\
&\quad + P\left(\int_0^1 \psi_s \, dW_s + \theta_n \geq r_n \bigcap |\theta_n| > \kappa\right) \\
&\leq P\left(\int_0^1 \psi_s \, dW_s \geq r_n - \kappa \bigcap |\theta_n| \leq \kappa\right) \\
&\quad + P\left(\int_0^1 \psi_s \, dW_s + \theta_n \geq r_n \bigcap |\theta_n| > \kappa\right) \\
&\leq P\left(\int_0^1 \psi_s \, dW_s \geq r_n - \kappa\right) + P\left(|\theta_n| > \kappa\right) \\
&\leq P\left(\int_0^1 \psi_s \, dW_s \geq r_n - \kappa\right) + \frac{E[\theta_n^2]}{\kappa^2} \\
&\leq P\left(\int_0^1 \psi_s \, dW_s \geq r_n - \kappa\right) + \frac{K_W}{\kappa^2\sqrt{n}} \\
&= P\left(\int_0^1 \psi_s \, dW_s \geq r_n - \kappa\right) + \delta_n \\
&= P\left(\xi_1 \geq 0\right) + \delta_n.
\end{aligned}$$

□

Note that $\boldsymbol{\xi}$ in (120) is an Itô *process*, for which $\boldsymbol{\psi}$ is permitted to be quite general. McNamara's stochastic control formulation only allows stochastic differential equations, where $\boldsymbol{\psi}$ must be a deterministic function of the present value of the process (and of time). But we can reduce the former to the latter [39, Corollary 3.7] (see also [40]): there exists a probability space with a measure $\hat{P}$ that supports a process $\hat{\boldsymbol{\xi}}$ and a Brownian motion $\hat{\mathbf{W}}$ such that

$$\hat{\xi}_t = -(r_n - \kappa) + \int_0^t \hat{\psi}_s(\hat{\xi}_s) \, d\hat{W}_s, \tag{121}$$

$$P\left(\xi_t \geq a\right) = \hat{P}\left(\hat{\xi}_t \geq a\right), \qquad a \in \mathbb{R}, \, t \geq 0, \tag{122}$$

and $\hat{\psi}_t(\cdot)$ satisfies

$$\hat{\psi}_t^2(u) = \mathbb{E}[\psi_t^2|\xi_t = u] \qquad P\text{-a.s.}, \, t \in \mathcal{N}^c,$$

where $\mathcal{N}$ is a Lebesgue-null set. In particular, we can take $\hat{\psi}_t(u) = \sqrt{\mathbb{E}[\psi_t^2|\xi_t = u]}$ [41, Section 5.3]. Note that the process $\hat{\boldsymbol{\xi}}$ has a deterministic function $\hat{\boldsymbol{\psi}}(\cdot)$ as the variance coefficient and the same one-dimensional law as that of $\boldsymbol{\xi}$ for each t.

Since $\hat{\psi}_t \in [\sqrt{\nu_{\min}}, \sqrt{\nu_{\max}}]$, (120) has a unique solution in distribution [41, Exercise 7.3.3] (see also the discussion after [39, Corollary 3.13]). Thus the setup in (121) is *admissible* as defined by McNamara in [11]. McNamara [11, Remark 8] shows that if the goal is to maximize $\hat{P}\left(\bar{\xi}_1 \geq 0\right)$ where

$$\bar{\xi}_t = -(r_n - \kappa) + \int_0^t \bar{\psi}_s(\bar{\xi}_s) \, d\hat{W}_s,$$

by choosing the optimal diffusion coefficient $\bar{\psi}_s(\cdot)$, then such optimal diffusion control is given by

$$\bar{\psi}^{\text{opt}}(u) := \sqrt{\nu_{\min}}\mathbf{1}\{u > 0\} + \sqrt{\nu_{\max}}\mathbf{1}\{u \le 0\}. \tag{123}$$

Let the corresponding SDE be

$$\bar{\xi}_t^{\text{opt}} := -(r_n - \kappa) + \int_0^t \bar{\psi}^{\text{opt}}(\bar{\xi}_s^{\text{opt}})\, d\hat{W}_s. \tag{124}$$

Thus

$$\hat{P}\left(\hat{\xi}_1 \ge 0\right) \le \hat{P}\left(\bar{\xi}_1^{\text{opt}} \ge 0\right). \tag{125}$$

Using the distribution function of the solution to (123) and (124) (see (59)), we get

$$\hat{P}\left(\bar{\xi}_1^{\text{opt}} \ge 0\right) = 1 - \frac{2\lambda}{1+\lambda}\Phi\left(\frac{r_n - \kappa}{\sqrt{\nu_{\min}}}\right), \tag{126}$$

when $r_n - \kappa \le 0$, and

$$\hat{P}\left(\bar{\xi}_1^{\text{opt}} \ge 0\right) = \frac{2}{1+\lambda} - \frac{2}{1+\lambda}\Phi\left(\frac{r_n - \kappa}{\sqrt{\nu_{\max}}}\right), \tag{127}$$

when $r_n - \kappa > 0$. For our choice of r_n in (108), we get

$$\hat{P}\left(\bar{\xi}_1^{\text{opt}} \ge 0\right) = 1 - \varepsilon - 2\delta_n. \tag{128}$$

Summarizing the chain of inequalities so far, we have

$$\begin{aligned} &P\left(\sum_{k=1}^n \mathrm{i}^*(X_k, Y_k) \ge \log \rho_n\right) \\ &\le P(S_n \ge r_n) \\ &= P\left(\int_0^{T_n^*} \psi_s\, dW_s \ge r_n\right) \\ &\le P\left(\xi_1 \ge 0\right) + \delta_n \\ &= \hat{P}\left(\hat{\xi}_1 \ge 0\right) + \delta_n \\ &\le \hat{P}\left(\bar{\xi}_1^{\text{opt}} \ge 0\right) + \delta_n \\ &= 1 - \varepsilon - \delta_n. \end{aligned}$$

Thus from (105)

$$\log M_{\text{fb}}^*(n, \varepsilon) \le nC + \sqrt{n} r_n - \log \frac{K_W}{\kappa^2 \sqrt{n}}, \tag{129}$$

and hence

$$\frac{\log M_{\text{fb}}^*(n,\varepsilon) - nC}{\sqrt{n}} \le r_n - \frac{1}{\sqrt{n}} \log \frac{K_W}{\kappa^2\sqrt{n}}.$$

From the definition of r_n in (108), and taking $n \to \infty$,

$$\begin{aligned} &\limsup_{n\to\infty} \frac{\log M_{\text{fb}}^*(n,\varepsilon) - nC}{\sqrt{n}} - \kappa \\ &\le \begin{cases} \sqrt{\nu_{\min}}\Phi^{-1}\left(\frac{1}{2\lambda}\varepsilon(1+\lambda)\right), & \varepsilon \in (0, \frac{\lambda}{1+\lambda}], \\ \sqrt{\nu_{\max}}\Phi^{-1}\left(\frac{1}{2}[\varepsilon(1+\lambda) + (1-\lambda)]\right), & \varepsilon \in (\frac{\lambda}{1+\lambda}, 1). \end{cases} \end{aligned}$$

Since κ is arbitrary, we may take $\kappa \to 0$ to prove the theorem. $\square$

VII. Very Noisy Channels

We first derive the scaling behavior of various channel parameters (C_ζ, $V_{\min,\zeta}$, etc.) with respect to ζ. Recall that the VNC is given by

$$W_\zeta(y|x) = \Gamma(y)\left(1 + \zeta\lambda(x,y)\right),$$

where Γ is a probability distribution on the output alphabet $\mathcal{Y}$, which we may assume, without loss of generality, has full support, $\lambda(x,y)$ satisfies

$$\sum_{y\in\mathcal{Y}} \Gamma(y)\lambda(x,y) = 0 \tag{130}$$

for all $x \in \mathcal{X}$, and ζ is infinitesimally small. Let

$$\lambda_{\max} := \max_{x\in\mathcal{X}, y\in\mathcal{Y}} |\lambda(x,y)|.$$

We will denote by $K(\Lambda)$ any non-negative constant which depends only on $(\lambda_{\max}, |\mathcal{X}|, |\mathcal{Y}|)$. The quantity represented by $K(\Lambda)$ will in general change from line to line in the derivation.

We will use the following approximation throughout the proof:

Lemma 9: For all u sufficiently close to zero,

$$|\log(1+u) - u| \le u^2,$$
$$\left|\log(1+u) - \left(u - \frac{u^2}{2}\right)\right| \le u^3.$$

The following lemma gives the scaling of the capacity C_ζ of the above channel.

Lemma 10: Let C_ζ denote the capacity of W_ζ. Then, for all sufficiently small ζ,

$$\left|C_\zeta - \zeta^2 \mathbf{C}\right| \le \zeta^3\ K(\Lambda),$$

where

$$\mathbf{C} := \max_{P\in\mathcal{P}(\mathcal{X})} \frac{1}{2}\sum_{y\in\mathcal{Y}} \Gamma(y)\left(\sum_{x\in\mathcal{X}} P(x)\lambda^2(x,y) - \lambda_P^2(y)\right), \tag{131}$$

and $\lambda_P(y) = \sum_{x\in\mathcal{X}} P(x)\lambda(x,y)$.

Proof: The channel capacity at ζ is given by

$$\begin{aligned} C_\zeta &= \max_{P\in\mathcal{P}(\mathcal{X})} I(P; W_\zeta) \\ &= \max_{P\in\mathcal{P}(\mathcal{X})} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P(x)\Gamma(y)\left(1+\zeta\lambda(x,y)\right) \cdot \\ &\qquad\qquad \log\frac{1+\zeta\lambda(x,y)}{1+\zeta\lambda_P(y)} \\ &\le \zeta^3 K(\Lambda) + \max_{P\in\mathcal{P}(\mathcal{X})} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P(x)\Gamma(y)\left(1+\zeta\lambda(x,y)\right) \cdot \\ &\qquad \left(\zeta\lambda(x,y) - \frac{\zeta^2\lambda^2(x,y)}{2} - \zeta\lambda_P(y) + \frac{\zeta^2\lambda_P^2(y)}{2}\right) \\ &\overset{(a)}{\le} \zeta^3 K(\Lambda) + \\ &\qquad \max_{P\in\mathcal{P}(\mathcal{X})} \sum_{y\in\mathcal{Y}} \Gamma(y)\left(\frac{\zeta^2}{2}\sum_{x\in\mathcal{X}} P(x)\lambda^2(x,y) - \frac{\zeta^2}{2}\lambda_P^2(y)\right) \\ &= \zeta^2 \mathbf{C} + \zeta^3 K(\Lambda). \end{aligned}$$

Here for (a), we note that $\sum_{y\in\mathcal{Y}}\Gamma(y)\lambda(x,y) = 0$, hence all the terms involving ζ disappear. The terms involving ζ^3 have been absorbed in $\zeta^3 K(\Lambda)$. Similarly, we can show $C_\zeta \geq \zeta^2\mathbf{C} - \zeta^3 K(\Lambda)$. □

Let q_ζ^* denote the output distribution corresponding to a capacity-achieving input distribution P_ζ^*, i.e.,

$$q_\zeta^*(y) = \Gamma(y)(1+\zeta\lambda_\zeta^*(y)),$$

where

$$\lambda_\zeta^*(y) := \sum_{x\in\mathcal{X}} P_\zeta^*(x)\lambda(x,y).$$

Here, we note that $|\lambda_\zeta^*(y)| \leq \lambda_{\max}$, and

$$\sum_{y\in\mathcal{Y}}\Gamma(y)\lambda_\zeta^*(y) = 0. \tag{132}$$

Also, since q_ζ^* is unique, λ_ζ^* is unique as well. Define

$$\mathcal{X}_\zeta^* := \{x : \mathbb{E}[i^*(X,Y)|X=x] = C_\zeta\}.$$

For $x \notin \mathcal{X}_\zeta^*$, let

$$\rho_{\zeta,x} := C_\zeta - \mathbb{E}[i^*(X,Y)|X=x],$$

where we note that $\rho_{\zeta,x} > 0$ [18, Theorem 4.5.1].

Define, for each $x\in\mathcal{X}$,

$$\nu_{x,\zeta} := \mathrm{Var}\left[i^*(X,Y)|X=x\right].$$

Lemma 11: For all sufficiently small ζ, the conditional expectation and variance of $i^*(X,Y)$ satisfy, for each x,

$$\left|\mathbb{E}[i^*(X,Y)|X=x] - \zeta^2\Psi_{\zeta,x}\right| \leq \zeta^3\ K(\Lambda),$$
$$\left|\nu_{x,\zeta} - 2\zeta^2\Psi_{\zeta,x}\right| \leq \zeta^3\ K(\Lambda),$$

where

$$\Psi_{\zeta,x} := \frac{1}{2}\sum_{y\in\mathcal{Y}}\Gamma(y)\left(\lambda(x,y)-\lambda_\zeta^*(y)\right)^2.$$

Hence for $x\in\mathcal{X}_\zeta^*$,

$$|\nu_{x,\zeta} - 2C_\zeta| \leq \zeta^3\ K(\Lambda),$$

and for $x\notin\mathcal{X}_\zeta^*$,

$$|\nu_{x,\zeta} - 2(C_\zeta - \rho_{\zeta,x})| \leq \zeta^3\ K(\Lambda).$$

Proof: We first note that since $|\lambda_\zeta^*(y)| \leq \lambda_{\max}$, we have $\Psi_{\zeta,x} \leq K(\Lambda)$. Now consider,

$$\begin{aligned}
&\mathbb{E}[i^*(X,Y)|X=x] \\
&= \sum_{y\in\mathcal{Y}} W_\zeta(y|x)\log\frac{W_\zeta(y|x)}{q_\zeta^*(y)} \\
&= \sum_{y\in\mathcal{Y}}\Gamma(y)\left(1+\zeta\lambda(x,y)\right)\log\frac{1+\zeta\lambda(x,y)}{1+\zeta\lambda_\zeta^*(y)} \\
&\leq \sum_{y\in\mathcal{Y}}\Gamma(y)\left(1+\zeta\lambda(x,y)\right)\left(\zeta\lambda(x,y) - \frac{\zeta^2\lambda^2(x,y)}{2}\right) \\
&\qquad - \sum_{y\in\mathcal{Y}}\Gamma(y)\left(1+\zeta\lambda(x,y)\right)\left(\zeta\lambda_\zeta^*(y) - \frac{\zeta^2\lambda_\zeta^{*2}(y)}{2}\right) \\
&\qquad + \zeta^3 K(\Lambda) \\
&\overset{(a)}{\leq} \frac{\zeta^2}{2}\sum_{y\in\mathcal{Y}}\Gamma(y)\left(\lambda(x,y)-\lambda_\zeta^*(y)\right)^2 + \zeta^3K(\Lambda) \\
&= \zeta^2\Psi_{\zeta,x} + \zeta^3K(\Lambda).
\end{aligned}$$

Here, (a) follows from (130), (132), and combining all terms involving ζ^3 with $\zeta^3K(\Lambda)$.

Similarly, one can show that

$$\mathbb{E}[i^*(X,Y)|X=x] \geq \zeta^2\Psi_{\zeta,x} - \zeta^3K(\Lambda).$$

Using Taylor's theorem one can show for all sufficiently small ζ,

$$\begin{aligned}
&\left|\left(\log\frac{1+\zeta\lambda(x,y)}{1+\zeta\lambda_\zeta^*(y)}\right)^2 - \zeta^2(\lambda(x,y)-\lambda_\zeta^*(y))^2\right| \\
&\leq \zeta^3\ K(\Lambda).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[(i^*(X,Y))^2|X=x\right] \\
&= \sum_{y\in\mathcal{Y}}\Gamma(y)\left(1+\zeta\lambda(x,y)\right)\left(\log\frac{1+\zeta\lambda(x,y)}{1+\zeta\lambda_\zeta^*(y)}\right)^2 \\
&\leq \sum_{y\in\mathcal{Y}}\Gamma(y)\left(1+\zeta\lambda(x,y)\right) \\
&\qquad \cdot\left(\zeta^2(\lambda(x,y)-\lambda_\zeta^*(y))^2\right) + \zeta^3K(\Lambda) \\
&\leq \zeta^2\sum_{y\in\mathcal{Y}}\Gamma(y)(\lambda(x,y)-\lambda_\zeta^*(y))^2 + \zeta^3K(\Lambda) \\
&= 2\zeta^2\Psi_{\zeta,x} + \zeta^3K(\Lambda).
\end{aligned}$$

Hence,

$$\begin{aligned}
\nu_{x,\zeta} &= \mathrm{Var}\left[i^*(X,Y)|X=x\right] \\
&= \mathbb{E}\left[(i^*(X,Y))^2|X=x\right] - (\mathbb{E}[i^*(X,Y)|X=x])^2 \\
&\leq 2\zeta^2\Psi_{\zeta,x} + \zeta^3K(\Lambda).
\end{aligned}$$

Note that $\mathbb{E}[i^*(X,Y)|X=x]^2 \leq \zeta^4K(\Lambda)$. This gives,

$$\nu_{x,\zeta} \geq 2\zeta^2\Psi_{\zeta,x} - \zeta^3K(\Lambda).$$

Since for $x\in\mathcal{X}_\zeta^*$, $\mathbb{E}\left[i^*(X,Y)|X=x\right] = C_\zeta$, for $x\in\mathcal{X}_\zeta^*$ we get

$$|\nu_{x,\zeta} - 2C_\zeta| \leq \zeta^3\ K(\Lambda),$$

and for $x \notin \mathcal{X}_\zeta^*$,

$$|\nu_{x,\zeta} - 2(C_\zeta - \rho_{\zeta,x})| \leq \zeta^3 \ K(\Lambda).$$

□

Recall that $V_{\min,\zeta}$ and $V_{\max,\zeta}$ are defined as

$$V_{\min,\zeta} := \min_{P \in \Pi^*_{W_\zeta}} \sum_{x \in \mathcal{X}} P(x)\nu_{x,\zeta}$$
$$V_{\max,\zeta} := \max_{P \in \Pi^*_{W_\zeta}} \sum_{x \in \mathcal{X}} P(x)\nu_{x,\zeta},$$

where $\Pi^*_{W_\zeta}$ is the set of capacity-achieving probability distributions.

Lemma 12: For all sufficiently small ζ, $V_{\min,\zeta}$ and $V_{\max,\zeta}$ satisfy

$$|V_{\min,\zeta} - 2C_\zeta| \leq \zeta^3 K(\Lambda),$$
$$\left|V_{\min,\zeta} - 2\zeta^2 \mathbf{C}\right| \leq \zeta^3 K(\Lambda)$$
$$|V_{\max,\zeta} - 2C_\zeta| \leq \zeta^3 K(\Lambda),$$
$$\left|V_{\max,\zeta} - 2\zeta^2 \mathbf{C}\right| \leq \zeta^3 K(\Lambda).$$

Proof: Note that if $P \in \Pi^*_{W_\zeta}$, then the support of P is contained in $\mathcal{X}^*_\zeta$. Thus from Lemma 11 we get $|V_{\min,\zeta} - 2C_\zeta| \leq \zeta^3 K(\Lambda)$. Moreover since from Lemma 10, $|C_\zeta - \zeta^2 \mathbf{C}| \leq \zeta^3 K(\Lambda)$, the inequality $\left|V_{\min,\zeta} - 2\zeta^2\mathbf{C}\right| \leq \zeta^3 K(\Lambda)$ follows. The second set of inequalities for $V_{\max,\zeta}$ can be deduced similarly. □

From Lemma 12, we can conclude that $V_{\max,\zeta} \approx V_{\min,\zeta}$. Thus taking a hint from Theorem 3, we expect that feedback will not improve the performance of VNCs with respect to the second-order coding rate. However, since we have not shown that $V_{\max,\zeta} = V_{\min,\zeta}$, Theorem 3 cannot be directly applied here. Since $\nu_{x,\zeta}$ is not constant over x, even asymptotically, Theorem 4 cannot be applied either. Thus we prove the converse with a different strategy.

Since $\nu_{x,\zeta} \approx 2C_\zeta$ for $x \in \mathcal{X}^*_\zeta$, and for $x \notin \mathcal{X}^*_\zeta$, we have that $\nu_{x,\zeta} \lesssim 2C_\zeta$, to obtain the converse we will add non-negative random variables whenever the input $X_k \notin \mathcal{X}^*_\zeta$ to "equalize" the conditional variance. The following lemma shows the existence of such random variables with desirable properties so that we can apply martingale convergence results. This will yield a proper upper on bound on the maximum possible message set size for sufficiently small ζ.

Lemma 13: We can extend the given probability space to define a sequence of non-negative random variables $\{\xi_k\}_{k=1}^n$, such that with $Z_k = i^*(X_k, Y_k) + \xi_k - C_\zeta$, $\mathcal{F}_k = \sigma(Z_1, \ldots, Z_k)$, and for all sufficiently small ζ,

$$|Z_k| \leq 3 \quad \text{a.s.},$$
$$\mathbb{E}[Z_k|\mathcal{F}_{k-1}] = 0 \quad \text{a.s.},$$
$$V_{\min,\zeta} - \zeta^3 K(\Lambda) \leq \mathbb{E}[Z_k^2] \leq V_{\min,\zeta} + \zeta^3 K(\Lambda),$$
$$V_{\max,\zeta} - \zeta^3 K(\Lambda) \leq \mathbb{E}[Z_k^2] \leq V_{\max,\zeta} + \zeta^3 K(\Lambda),$$
$$\left\| \frac{\sum_{k=1}^n \mathbb{E}[Z_k^2|\mathcal{F}_{k-1}]}{\sum_{k=1}^n \mathbb{E}[Z_k^2]} - 1 \right\|_\infty^{1/2} \leq \sqrt{\zeta} K(\Lambda).$$

Proof: For each $x \notin \mathcal{X}^*_\zeta$, define $\{\xi_{x,k}\}_{k=1}^n$ to be a sequence of i.i.d. random variables, independent of all other random variables such that

$$P(\xi_{x,k} = \rho_{\zeta,x} + 2) = 1 - P(\xi_{x,k} = 0) = \frac{\rho_{\zeta,x}}{\rho_{\zeta,x} + 2}.$$

The variance of the above random variable is

$$\begin{aligned}\mathrm{Var}[\xi_{x,k}] &= \mathbb{E}[(\xi_{x,k})^2] - (\mathbb{E}[\xi_{x,k}])^2 \\ &= \frac{\rho_{\zeta,x}}{(\rho_{\zeta,x}+2)}(\rho_{\zeta,x}+2)^2 \\ &\quad - \left(\frac{\rho_{\zeta,x}}{(\rho_{\zeta,x}+2)}(\rho_{\zeta,x}+2)\right)^2 \\ &= 2\rho_{\zeta,x}.\end{aligned}$$

Let

$$\xi_k = \sum_{x \notin \mathcal{X}^*_\zeta} \xi_{x,k} \mathbf{1}\{X_k = x\}.$$

Then,

$$\begin{aligned}|Z_k| &\leq |i^*(X_k, Y_k)| + \xi_k + C_\zeta \\ &\leq |i^*(X_k, Y_k)| + \max_{x \notin \mathcal{X}^*_\zeta} \rho_{\zeta,x} + 2 + C_\zeta \\ &\leq 3 \quad \text{a.s.},\end{aligned}$$

for all sufficiently small ζ. Let $\mathcal{G}_k = \sigma(M, Y_1, \xi_1, \ldots, Y_k, \xi_k)$. We note that X_k is $\mathcal{G}_{k-1}$ measurable (since the message M and past outputs $(Y_1, \ldots, Y_{k-1})$ determine the input X_k) and $\mathcal{F}_k \subseteq \mathcal{G}_k$. Thus,

$$\begin{aligned}\mathbb{E}[i^*(X_k, Y_k)|\mathcal{G}_{k-1}] &= \mathbb{E}[i^*(X_k, Y_k)|X_k] \\ &= C_\zeta - \rho_{\zeta,X_k}\mathbf{1}\{X_k \notin \mathcal{X}^*_\zeta\}.\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}[Z_k|\mathcal{G}_{k-1}] &= C_\zeta - \rho_{\zeta,X_k}\mathbf{1}\{X_k \notin \mathcal{X}^*_\zeta\} \\ &\quad + \rho_{\zeta,X_k}\mathbf{1}\{X_k \notin \mathcal{X}^*_\zeta\} - C_\zeta \\ &= 0.\end{aligned}$$

Taking the conditional expectation with respect to $\mathcal{F}_{k-1}$, and since $\mathcal{F}_{k-1} \subseteq \mathcal{G}_{k-1}$,

$$\mathbb{E}[Z_k|\mathcal{F}_{k-1}] = 0.$$

Also,

$$\begin{aligned}\mathbb{E}[Z_k^2|\mathcal{G}_{k-1}] &= \mathrm{Var}[Z_k|\mathcal{G}_{k-1}] \\ &= \mathrm{Var}[i^*(X_k, Y_k) + \xi_k|\mathcal{G}_{k-1}] \\ &\overset{(a)}{=} \mathrm{Var}[i^*(X_k, Y_k)|\mathcal{G}_{k-1}] + \mathrm{Var}[\xi_k|\mathcal{G}_{k-1}] \\ &\overset{(b)}{\leq} 2C_\zeta - 2\rho_{\zeta,X_k}\mathbf{1}\{X_k \notin \mathcal{X}^*_\zeta\} + \zeta^3 K(\Lambda) \\ &\qquad + 2\rho_{\zeta,X_k}\mathbf{1}\{X_k \notin \mathcal{X}^*_\zeta\} \\ &= 2C_\zeta + \zeta^3 K(\Lambda).\end{aligned} \tag{133}$$

Here (a) follows since given X_k, $i^*(X_k, Y_k)$ and ξ_k are conditionally independent, and (b) follows from Lemma 11 and noting that $\mathrm{Var}[\xi_{x,k}] = 2\rho_{\zeta,x}$.

Similarly,

$$\mathbb{E}[Z_k^2|\mathcal{G}_{k-1}] \geq 2C_\zeta - \zeta^3 K(\Lambda). \tag{134}$$

Thus from Lemma 12, (133) and (134),

$$V_{\min,\zeta} - \zeta^3 K(\Lambda) \le \mathbb{E}[Z_k^2|\mathcal{G}_{k-1}] \le V_{\min,\zeta} + \zeta^3 K(\Lambda),$$
$$V_{\max,\zeta} - \zeta^3 K(\Lambda) \le \mathbb{E}[Z_k^2|\mathcal{G}_{k-1}] \le V_{\max,\zeta} + \zeta^3 K(\Lambda).$$

Once again taking the conditional expectation with respect to $\mathcal{F}_{k-1}$,

$$V_{\min,\zeta} - \zeta^3 K(\Lambda) \le \mathbb{E}[Z_k^2|\mathcal{F}_{k-1}] \le V_{\min,\zeta} + \zeta^3 K(\Lambda), \quad (135)$$
$$V_{\max,\zeta} - \zeta^3 K(\Lambda) \le \mathbb{E}[Z_k^2|\mathcal{F}_{k-1}] \le V_{\max,\zeta} + \zeta^3 K(\Lambda).$$

Now consider

$$\frac{\sum_{k=1}^n \mathbb{E}[Z_k^2|\mathcal{F}_{k-1}]}{\sum_{k=1}^n \mathbb{E}[Z_k^2]} - 1 \le \frac{V_{\min,\zeta} + \zeta^3 K(\Lambda)}{V_{\min,\zeta} - \zeta^3 K(\Lambda)} - 1 = \frac{2\zeta^3 K(\Lambda)}{V_{\min,\zeta} - \zeta^3 K(\Lambda)}.$$

Here, we note that for the last equality to hold, the constants $K(\Lambda)$ appearing in the left and right terms of (135) should be equal. If they are not, we simply replace each by the maximum of the two constants. Similarly,

$$\frac{\sum_{k=1}^n \mathbb{E}[Z_k^2|\mathcal{F}_{k-1}]}{\sum_{k=1}^n \mathbb{E}[Z_k^2]} - 1 \ge -\frac{2\zeta^3 K(\Lambda)}{V_{\min,\zeta} + \zeta^3 K(\Lambda)}.$$

Thus,

$$\left\| \frac{\sum_{k=1}^n \mathbb{E}[Z_k^2|\mathcal{F}_{k-1}]}{\sum_{k=1}^n \mathbb{E}[Z_k^2]} - 1 \right\|_\infty^{1/2} \le \left(\frac{2\zeta^3 K(\Lambda)}{V_{\min,\zeta} - \zeta^3 K(\Lambda)} \right)^{1/2} \le \sqrt{\zeta} K(\Lambda),$$

where the last inequality is due to Lemma 12. □

Now we give the proof of Theorem 5. Define

$$\kappa_{\zeta,n} := K(\Lambda)\left(\frac{\log(n)}{\zeta^3\sqrt{n}} + \sqrt{\zeta}\right) \quad (136)$$
$$\rho_{\zeta,n} := \exp(nC_\zeta + \sqrt{n}r_n) \quad (137)$$
$$r_{\zeta,n} := \begin{cases} \sqrt{(V_{\min,\zeta} - \zeta^3 K(\Lambda))}\Phi^{-1}(\varepsilon + \kappa_{\zeta,n}) \\ \sqrt{(V_{\max,\zeta} + \zeta^3 K(\Lambda))}\Phi^{-1}(\varepsilon + \kappa_{\zeta,n}), \end{cases} \quad (138)$$

for the cases $0 < \varepsilon \le \frac{1}{2} - \kappa_{\zeta,n}$ and $\frac{1}{2} - \kappa_{\zeta,n} < \varepsilon < 1$, respectively.

Now defining $\{\xi_k\}_{k=1}^n$ as a sequence of random variables as in Lemma 13, consider the following chain of inequalities:

$$P\left(\sum_{k=1}^n i^*(X_k, Y_k) \ge \log \rho_{\zeta,n}\right) \quad (139)$$
$$\overset{(a)}{\le} P\left(\sum_{k=1}^n i^*(X_k, Y_k) + \xi_k \ge \log \rho_{\zeta,n}\right)$$
$$\overset{(b)}{=} P\left(\sum_{k=1}^n Z_k \ge \sqrt{n} r_{\zeta,n}\right)$$
$$\overset{(c)}{\le} P\left(\frac{1}{\sqrt{\sum_{k=1}^n \mathbb{E}[Z_k^2]}} \sum_{k=1}^n Z_k \ge \Phi^{-1}(\varepsilon + 2\kappa_{\zeta,n})\right)$$
$$\overset{(d)}{\le} 1 - \varepsilon - 2\kappa_{\zeta,n} \quad (140)$$
$$+ \chi \cdot \left(\frac{n\log(n)}{\left(\sum_{k=1}^n \mathbb{E}[Z_k^2]\right)^{3/2}} + \left\| \frac{\sum_{k=1}^n \mathbb{E}[Z_k^2|\mathcal{F}_{k-1}]}{\sum_{k=1}^n \mathbb{E}[Z_k^2]} - 1 \right\|_\infty^{1/2} \right)$$
$$\overset{(e)}{\le} 1 - \varepsilon - 2\kappa_{\zeta,n} + \left(K(\Lambda)\frac{\log(n)}{\zeta^3\sqrt{n}} + K(\Lambda)\sqrt{\zeta}\right)$$
$$= 1 - \varepsilon - \kappa_{\zeta,n}. \quad (141)$$

Here, (a) follows since ξ_k is a non-negative random variable, (b) follows from setting Z_k as in Lemma 13, (c) follows since $n(V_{\min,\zeta} - \zeta^3 K(\Lambda)) \le \sum_{k=1}^n \mathbb{E}[Z_k^2] \le n(V_{\max,\zeta} + \zeta^3 K(\Lambda))$ due to Lemma 13, (d) follows from the martingale central limit theorem [34, Corollary to Theorem 2], and taking the constant as χ (which does not depend upon the channel or n), and (e) follows from noting that $\left(\sum_{k=1}^n \mathbb{E}[Z_k^2]\right)^{3/2} \ge n\sqrt{n}(2\zeta^2\mathbf{C} - \zeta^3 K(\Lambda))^{3/2}$, and then absorbing χ into $K(\Lambda)$.

Invoking Lemma 15 from the Appendix with $q_\zeta(\mathbf{y}^n) = \prod_{i=1}^n q_\zeta^*(y_i)$, we get

$$\log M^*_{\text{fb},\zeta}(n,\varepsilon) \le \log \rho_{\zeta,n} - \log \kappa_{\zeta,n} \le nC_\zeta + \sqrt{n}r_{\zeta,n} - \log \kappa_{\zeta,n}.$$

If $0 < \varepsilon < \frac{1}{2}$,

$$\limsup_{n\to\infty} \frac{\log M^*_{\text{fb},\zeta}(n,\varepsilon) - nC_\zeta}{\sqrt{nV_{\min,\zeta}}} \le \sqrt{1 - \frac{\zeta^3\ K(\Lambda)}{V_{\min,\zeta}}}\Phi^{-1}(\varepsilon + K(\Lambda)\sqrt{\zeta}),$$

and hence,

$$\limsup_{\zeta\to 0}\limsup_{n\to\infty} \frac{\log M^*_{\text{fb},\zeta}(n,\varepsilon) - nC_\zeta}{\sqrt{nV_{\min,\zeta}}} \le \Phi^{-1}(\varepsilon).$$

Similarly, when $\frac{1}{2} \le \varepsilon < 1$,

$$\limsup_{\zeta\to 0}\limsup_{n\to\infty} \frac{\log M^*_{\text{fb},\zeta}(n,\varepsilon) - nC_\zeta}{\sqrt{nV_{\max,\zeta}}} \le \Phi^{-1}(\varepsilon).$$

Since $V_{\min,\zeta}/V_{\max,\zeta} \to 1$ as $\zeta \to 0$ by Lemma 12, the conclusion follows. □

Appendix

As noted in the introduction, the problem of maximizing the second-order coding rate with feedback is related to the design of controlled random walks.

Definition 13: A *controller* is a function $F : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{P}(\mathcal{X})$.

We shall sometimes write $F(\cdot|\mathbf{x}^k, \mathbf{y}^k)$ for $F(\mathbf{x}^k, \mathbf{y}^k)(\cdot)$. Given a controller F, let $F \circ W$ denote the distribution

$$(F \circ W)(\mathbf{x}^n, \mathbf{y}^n) = \prod_{k=1}^n F(x_k|\mathbf{x}^{k-1}, \mathbf{y}^{k-1})W(y_k|x_k) \quad (142)$$

and let $FW(\mathbf{y}^n)$ denote the marginal over $\mathbf{Y}^n$ induced by $F \circ W$.

The following lemma shows that any controller gives rise to an achievable second-order coding rate. The idea is to use the controller to generate a random ensemble of feedback codes and then use a technique that dates back to Feinstein [42] and Shannon [43] to bound the error probability of this ensemble.

Lemma 14 (Achievability): For any controller F and any n, θ, and rate R,

$$\bar{\text{P}}_{\text{e,fb}}(n,R) \le (F \circ W)\left(\frac{1}{n}\log\frac{W(\mathbf{Y}^n|\mathbf{X}^n)}{FW(\mathbf{Y}^n)} \le R + \theta\right) + e^{-n\theta}. \quad (143)$$

Thus, if for some α and ε,

$$\limsup_{n\to\infty} \inf_F (F \circ W)\left(\log \frac{W(\mathbf{Y}^n|\mathbf{X}^n)}{FW(\mathbf{Y}^n)} \le nC + \alpha\sqrt{n}\right) < \varepsilon, \tag{144}$$

then

$$\liminf_{n\to\infty} \frac{\log M^*_{\mathrm{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \ge \alpha. \tag{145}$$

Proof: We begin by showing (143). Consider a random code in which, for each message, the channel input at time k when the past inputs are $\mathbf{x}^{k-1}$ and the past outputs are $\mathbf{y}^{k-1}$ is chosen according to $F(\cdot|\mathbf{x}^{k-1}, \mathbf{y}^{k-1})$. That is, $f(m, \mathbf{y}^{k-1})$ is chosen randomly according to

$$F(\cdot|(f(m,\emptyset), f(m,y_1), \dots, f(m,\mathbf{y}^{k-2})), \mathbf{y}^{k-1}). \tag{146}$$

Given $\mathbf{y}^n$, the decoder selects the message with the lowest index that achieves the minimum over m of

$$\prod_{k=1}^n W(y_k|f(m,\mathbf{y}^{k-1})). \tag{147}$$

By the union bound and other standard steps, the ensemble average error probability of this code is upper bounded by

$$\begin{aligned}
&\sum_{\mathbf{x}^n,\mathbf{y}^n} (F\circ W)(\mathbf{x}^n,\mathbf{y}^n)\mathbf{1}\left\{\frac{1}{n}\log\frac{W(\mathbf{y}^n|\mathbf{x}^n)}{FW(\mathbf{y}^n)} \le R+\theta\right\} \qquad (148)\\
&\quad + e^{nR}\sum_{\mathbf{x}^n,\mathbf{y}^n}(F\circ W)(\mathbf{x}^n,\mathbf{y}^n)\cdot \\
&\quad\quad \sum_{\substack{\tilde{\mathbf{x}}^n: W(\mathbf{y}^n|\tilde{\mathbf{x}}^n)\\ \ge W(\mathbf{y}^n|\mathbf{x}^n)}} \prod_{k=1}^n F(\tilde{x}_k|\tilde{\mathbf{x}}^{k-1},\mathbf{y}^{k-1})\cdot \\
&\quad\quad\quad \mathbf{1}\left\{\frac{1}{n}\log\frac{W(\mathbf{y}^n|\mathbf{x}^n)}{FW(\mathbf{y}^n)} > R+\theta\right\} \\
&\le (F\circ W)\left(\frac{1}{n}\log\frac{W(\mathbf{Y}^n|\mathbf{X}^n)}{FW(\mathbf{Y}^n)} \le R+\theta\right) \qquad (149)\\
&\quad + e^{nR}\sum_{\mathbf{y}^n,\tilde{\mathbf{x}}^n} FW(\mathbf{y}^n)\prod_{k=1}^n F(\tilde{x}_k|\tilde{\mathbf{x}}^{k-1},\mathbf{y}^{k-1})\cdot \\
&\quad\quad \mathbf{1}\left\{\frac{1}{n}\log\frac{W(\mathbf{y}^n|\tilde{\mathbf{x}}^n)}{FW(\mathbf{y}^n)} > R+\theta\right\} \\
&\le (F\circ W)\left(\frac{1}{n}\log\frac{W(\mathbf{Y}^n|\mathbf{X}^n)}{FW(\mathbf{Y}^n)} \le R+\theta\right) \qquad (150)\\
&\quad + e^{nR}e^{-n(R+\theta)}\sum_{\tilde{\mathbf{x}}^n}\sum_{\mathbf{y}^n}\prod_{k=1}^n F(\tilde{x}_k|\tilde{\mathbf{x}}^{k-1},\mathbf{y}^{k-1})W(y_k|\tilde{x}_k),
\end{aligned}$$

which implies (143). Now suppose (144) holds and in (143), select $R = C + \alpha'/\sqrt{n}$ and $\theta = n^{-\beta}$ for some $1/2 < \beta < 1$ and $\alpha' < \alpha$. Then we have

$$\begin{aligned}
&\limsup_{n\to\infty} \bar{\mathrm{P}}_{\mathrm{e,fb}}\left(n, C+\frac{\alpha'}{\sqrt{n}}\right) \\
&\le \limsup_{n\to\infty}\inf_F (F\circ W)\Bigg(\frac{1}{n}\log\frac{W(\mathbf{Y}^n|\mathbf{X}^n)}{FW(\mathbf{Y}^n)} \le C \\
&\qquad\qquad + \frac{\alpha'}{\sqrt{n}} + \frac{1}{n^\beta}\Bigg).
\end{aligned}$$

Thus if (144) holds we have

$$\limsup_{n\to\infty} \bar{\mathrm{P}}_{\mathrm{e,fb}}\left(n, C+\frac{\alpha'}{\sqrt{n}}\right) < \varepsilon, \tag{151}$$

since $\alpha' < \alpha$. This implies that eventually,

$$\log M^*_{\mathrm{fb}}(n,\varepsilon) \ge nC + \alpha'\sqrt{n}. \tag{152}$$

Since this holds for any $\alpha' < \alpha$, (145) follows. □

The next result is used repeatedly in the paper as a starting point in proving converses. A similar inequality to (153) can be found in [12, (42)]. Observe that (154) and (155), which are consequences of (153), are nearly a converse of (144) and (145) above.

Lemma 15 (Converse): For any n, $\rho > 0$, and $\varepsilon > 0$,

$$\begin{aligned}
\log M^*_{\mathrm{fb}}(n,\varepsilon) &\le \sup_F \inf_{q\in\mathcal{P}(\mathcal{Y}^n)} \log\rho \\
&- \log\left(\left(1-\varepsilon-(F\circ W)\left(\frac{W(\mathbf{Y}^n|\mathbf{X}^n)}{q(\mathbf{Y}^n)} > \rho\right)\right)^+\right). \qquad (153)
\end{aligned}$$

In particular, if for some α and ε,

$$\begin{aligned}
&\liminf_{n\to\infty}\inf_F\sup_q (F\circ W)\left(\log\frac{W(\mathbf{Y}^n|\mathbf{X}^n)}{q(\mathbf{Y}^n)} \le nC+\alpha\sqrt{n}\right) \\
&> \varepsilon, \qquad (154)
\end{aligned}$$

then

$$\limsup_{n\to\infty}\frac{\log M^*_{\mathrm{fb}}(n,\varepsilon) - nC}{\sqrt{n}} \le \alpha. \tag{155}$$

Proof: Consider an (n, R) feedback code (f, g) with average error probability at most ε. We will denote this code by $\mathfrak{C}$ and its average error probability by $\varepsilon_{\mathfrak{C}}$. Define

$$M^*_{\mathrm{fb},\mathfrak{C}}(n) := \lceil \exp(nR) \rceil.$$

Then

$$M^*_{\mathrm{fb}}(n,\varepsilon) = \sup_{\mathfrak{C}:\varepsilon_{\mathfrak{C}}\le\varepsilon} M^*_{\mathrm{fb},\mathfrak{C}}(n).$$

The code $\mathfrak{C}$ induces a controller F via

$$\begin{aligned}
&F(x_k|\mathbf{x}^{k-1},\mathbf{y}^{k-1}) \\
&\quad := \frac{1}{M^*_{\mathrm{fb},\mathfrak{C}}(n)}\sum_{m=1}^{M^*_{\mathrm{fb},\mathfrak{C}}(n)} \mathbf{1}\{f(m,\mathbf{y}^{k-1}) = x_k\},
\end{aligned}$$

which, in fact, does not depend on $\mathbf{x}^{k-1}$. Now consider the problem of hypothesis testing where a random variable U taking values in $\mathcal{U}$ can have probability measure P or Q. Upon observing U, the goal is to declare either $U \sim P$ (hypothesis H_1) or $U \sim Q$ (hypothesis H_2). Let $\beta_\alpha(P,Q)$ denote the minimum attainable error probability under Q when the error probability under P does not exceed $1-\alpha$. Then the Neyman-Pearson lemma [44, Proposition II.D.1, p. 33] guarantees that there exists a (possibly randomized) test $T : \mathcal{U} \to \{0,1\}$ (where 0 corresponds to the test selecting Q) such that

$$\sum_{u\in\mathcal{U}} P(u)T(1|u) \ge \alpha,\ \sum_{u\in\mathcal{U}} Q(u)T(1|u) = \beta_\alpha(P,Q).$$

Then for any $\rho > 0$

$$\begin{aligned}
&\alpha - \rho\beta_\alpha(P,Q) \qquad (156)\\
&\leq \sum_{u\in\mathcal{U}} T(1|u)(P(u) - \rho Q(u))\\
&\leq \sum_{u\in\mathcal{U}} T(1|u)(P(u) - \rho Q(u))\mathbf{1}\{P(u) > \rho Q(u)\}\\
&= P\left(\frac{P(u)}{Q(u)} > \rho, T = 1\right) - \rho Q\left(\frac{P(u)}{Q(u)} > \rho, T = 1\right)\\
&\leq P\left(\frac{P(u)}{Q(u)} > \rho\right), \qquad (157)
\end{aligned}$$

which is a trivial strengthening of [21, Eq. (102)].

Fix a $q \in \mathcal{P}(\mathcal{Y}^n)$. Applying [21, Theorem 26] (with $Q_{Y|X} = q$, $\varepsilon' = 1 - 1/M^*_{\mathrm{fb},\mathfrak{C}}(n)$; the assertion there is without feedback but one can verify that it applies to the feedback case as well), we get

$$\beta_{1-\varepsilon_{\mathfrak{C}}}\left(F\circ W, F\circ q\right) \leq \frac{1}{M^*_{\mathrm{fb},\mathfrak{C}}(n)}.$$

Moreover, from (157)

$$\alpha \leq (F\circ W)\left(\frac{d(F\circ W)}{d(F\circ q)} > \rho\right) + \rho\beta_\alpha\left(F\circ W, F\circ q\right),$$

i.e.,

$$\begin{aligned}
&\beta_{1-\varepsilon_{\mathfrak{C}}}\left(F\circ W, F\circ q\right)\\
&\geq \frac{1}{\rho}\left(1 - \varepsilon_{\mathfrak{C}} - (F\circ W)\left(\frac{d(F\circ W)}{d(F\circ q)} > \rho\right)\right)^+.
\end{aligned}$$

Thus

$$\begin{aligned}
&\log M^*_{\mathrm{fb},\mathfrak{C}}(n)\\
&\leq \log\rho - \log\left[\left(1 - \varepsilon_{\mathfrak{C}} - (F\circ W)\left(\frac{d(F\circ W)}{d(F\circ q)} > \rho\right)\right)^+\right].
\end{aligned}$$

Using the fact that $\varepsilon_{\mathfrak{C}} \leq \varepsilon$ and that q was arbitrary, we obtain

$$\begin{aligned}
&\log M^*_{\mathrm{fb},\mathfrak{C}}(n)\\
&\leq \inf_{q\in\mathcal{P}(\mathcal{Y}^n)} \log\rho\\
&\quad - \log\left[\left(1 - \varepsilon - (F\circ W)\left(\frac{d(F\circ W)}{d(F\circ q)} > \rho\right)\right)^+\right].
\end{aligned}$$

Taking the supremum over all controllers F and noting that

$$\frac{d(F\circ W)}{d(F\circ q)} = \prod_{k=1}^n \frac{W(y_k|x_k)}{q(y_k|\mathbf{y}^{k-1})},$$

gives (153). (155) follows directly from (153) and (154). □

Acknowledgment

The authors would like to thank the associate editor and the anonymous reviewers for their attentive reading and helpful suggestions.

References

[1] C. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.

[2] Y.-H. Kim, "Feedback capacity of stationary Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 57–85, Jan. 2010.

[3] A. Tchamkerten and I. E. Telatar, "Variable length coding over an unknown channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2126–2145, May 2006.

[4] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," (in Russian), *Probl. Peredachi Inf.*, vol. 12, no. 4, pp. 10–30, 1976.

[5] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback—I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. 12, no. 2, pp. 172–182, Apr. 1966.

[6] J. Schalkwijk, "A coding scheme for additive noise channels with feedback—II: Band-limited signals," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 183–189, Apr. 1966.

[7] E. R. Berlekamp, "Block coding with noiseless feedback," Ph.D. dissertation, MIT, Cambridge, MA, USA, 1964.

[8] N. Gaarder and J. Wolf, "The capacity region of a multiple-access discrete memoryless channel can increase with feedback (corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 1, pp. 100–102, Jan. 1975.

[9] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

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

[11] J. M. McNamara, "Optimal control of the diffusion coefficient of a simple diffusion process," *Math. Oper. Res.*, vol. 8, no. 3, pp. 373–380, Aug. 1983.

[12] S. L. Fong and V. Y. F. Tan, "A tight upper bound on the second-order coding rate of the parallel Gaussian channel with feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6474–6486, Oct. 2017.

[13] B. Reiffen, "A note on 'very noisy channels,'" *Inf. Control*, vol. 6, no. 2, pp. 126–130, Jun. 1963.

[14] E. E. Majani, "A model for the study of very noisy channels, and applications," Ph.D. dissertation, California Inst. Technol., Pasadena, CA, USA, 1987.

[15] A. Lapidoth, I. E. Telatar, and R. Urbanke, "On wide-band broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3250–3258, Dec. 2003.

[16] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[17] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*. New York, NY, USA: Springer-Verlag, 1988.

[18] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[19] V. Strassen, "Asymptotische abschatzugen in Shannon's informations-theorie," in *Proc. Trans. 3rd Prague Conf Inf. Theory*, Prague, Czech, 1962, pp. 689–723.

[20] A. Collins and Y. Polyanskiy, "Orthogonal designs optimize achievable dispersion for coherent MISO channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 2524–2528.

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

[22] S. L. Fong and V. Y. F. Tan, "Private communication," 2017.

[23] N. V. Shende and A. B. Wagner, "On very noisy channels with feedback," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 1–8.

[24] S. Lee and K. A. Winick, "Very noisy channels, reliability functions, and exponentially optimum codes," *IEEE Trans. Inf. Theory*, vol. 40, no. 3, pp. 647–661, May 1994.

[25] V. Y. Korolev and I. G. Shevtsova, "A new moment-type estimate of convergence rate in the Lyapunov theorem," *Theory Probab. Appl.*, vol. 55, no. 3, pp. 505–509, Jan. 2011.

[26] A. C. Berry, "The accuracy of the Gaussian approximation to the sum of independent variates," *Trans. Amer. Math. Soc.*, vol. 49, no. 1, pp. 122–136, Jan. 1941.

[27] C.-G. Esseen, "Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law," *Acta Math.*, vol. 77, pp. 1–125, 1945.

[28] G. L. Kulinich, "On the limit behaviour of the solution of a stochastic diffusion equation," *Theory Probab. Appl.*, vol. 12, no. 3, pp. 497–799, Jul. 1967.

[29] B. Øksendal, *Stochastic Differential Equations: An Introduction With Applications*, 5th ed. Berlin, Germany: Springer-Verlag, 1998.

[30] H. J. Kushner, "On the weak convergence of interpolated Markov chains to a diffusion," *Ann. Probab.*, vol. 2, no. 1, pp. 40–50, Feb. 1974.

[31] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.

[32] P. Billingsley, *Convergence of Probability Measures*, 2nd ed. Hoboken, NJ, USA: Wiley, 1999.

[33] G. L. Kulinich, "On necessary and sufficient conditions for convergence of solutions to one-dimensional stochastic diffusion equations with a nonregular dependence of the coefficients on a parameter," *Theory Probab. Appl.*, vol. 27, no. 4, pp. 856–862, Dec. 1983.

[34] E. Bolthausen, "Exact convergence rates in some martingale central limit theorems," *Ann. Probab.*, vol. 10, no. 3, pp. 672–688, Aug. 1982.

[35] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[36] B. Bercu, B. Delyon, and E. Rio, *Concentration Inequalities for Sums and Martingales*. Cham, Switzerland: Springer, 2015.

[37] O. Barndorff-Nielsen and A. Shiryaev, *Change of Time and Change of Measure*, vol. 13. Singapore: World Scientific, 2010.

[38] R. Durrett, *Probability: Theory and Examples*, 4th ed. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[39] G. Brunick and S. Shreve, "Mimicking an Itô process by a solution of a stochastic differential equation," *Ann. Appl. Probab.*, vol. 23, no. 4, pp. 1584–1628, Aug. 2013.

[40] I. Gyöngy, "Mimicking the one-dimensional marginal distributions of processes having an Itô differential," *Probab. Theory Rel. Fields*, vol. 71, no. 4, pp. 501–516, 1986.

[41] D. Stroock and S. Varadhan, *Multidimensional Diffusion Processes*. Berlin, Germany: Springer, 2007.

[42] A. Feinstein, "A new basic theorem of information theory," *Trans. IRE Prof. Group Inf. Theory*, vol. 4, no. 4, pp. 2–22, Sep. 1954.

[43] C. E. Shannon, "Certain results in coding theory for noisy channels," *Inf. Control*, vol. 1, no. 1, pp. 6–25, Sep. 1957.

[44] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York, NY, USA: Springer-Verlag, 1994.

Aaron B. Wagner (Fellow, IEEE) received the B.S. degree in electrical engineering from the University of Michigan, Ann Arbor, in 1999, and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California at Berkeley in 2002 and 2005, respectively.

From 2005 to 2006, he was a Post-Doctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, and a Visiting Assistant Professor with the School of Electrical and Computer Engineering, Cornell University. Since 2006, he has been with the School of Electrical and Computer Engineering at Cornell, where he is currently a Professor. He has received the NSF CAREER Award, the David J. Sakrison Memorial Prize from the U.C. Berkeley EECS Department, the Bernard Friedman Memorial Prize in Applied Mathematics from the U.C. Berkeley Department of Mathematics, the James L. Massey Research and Teaching Award for Young Scholars from the IEEE Information Theory Society, and teaching awards at the department, college, and university level at Cornell. He was a Distinguished Lecturer of the IEEE Information Theory Society from 2018 to 2019.

Nirmal V. Shende received the B.Tech. degree in electronics and communication engineering from the Visvesvaraya National Institute of Technology, Nagpur, India, the M.E. degree in telecommunication engineering from IIT Bangalore, Bengaluru, India, and the Ph.D. degree from the School of Electrical and Computer Engineering, Cornell University, Ithaca, USA. He is currently with Marvell Semiconductor, Santa Clara, CA. His research interests are in the areas of information theory, quantum information theory, and wireless communication. He was a recipient of the Irwin M. and Joan K. Jacobs Fellowship from the School of Electrical and Computer Engineering, Cornell University.

Yücel Altuğ received the B.S. and M.S. degrees in electrical and electronics engineering from Boğaziçi University, Turkey, in 2006 and 2008, respectively, and the Ph.D. degree in electrical and computer engineering from Cornell University in 2013. After post-doctoral appointments at Cornell University and Princeton University, he is currently the Lead Data Scientist with Natera Inc. His research interests include Shannon theory, feedback communications, and stochastic modeling and algorithm design for next-generation DNA sequencing and genetic testing. He has been awarded the ECE Director's Ph.D. Thesis Research Award from Cornell University.