Chapter 1
State Dependent Channels: Strong Converse and Bounds on Reliability Function

Himanshu Tyagi and Prakash Narayan

Abstract
We consider an information theoretic model of a communication channel with a time-varying probability law. Specifically, our model consists of a state dependent *discrete memoryless channel*, in which the underlying state process is independent and identically distributed with known probability distribution, and for which the channel output at any time instant depends on the inputs and states only through their current values. For this channel, we provide a strong converse result for its capacity, explaining the structure of optimal transmission codes. Exploiting this structure, we obtain upper bounds for the reliability function when the transmitter is provided channel state information causally and noncausally. Instrumental to our proofs is a new technical result which provides an upper bound on the rate of codes with codewords that are "conditionally typical over large *message dependent* subsets of a typical set of state sequences." This technical result is a nonstraightforward extension of an analogous result for a discrete memoryless channel without states; the latter provides a bound on the rate of a good code with codewords of a fixed composition.

1.1 Introduction

The information theoretic model of a communication channel for message transmission is described by the conditional probability law of the channel output given the input. For instance, the binary symmetric channel is a model for describing the com-

Himanshu Tyagi
Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, College Park, MD 20742, e-mail: tyagi@umd.edu

Prakash Narayan
Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, College Park, MD 20742, e-mail: prakash@umd.edu

munication of binary data in which noise may cause random bit-flips with a fixed probability. A reliable encoded transmission of a message generally entails multiple uses of the channel. In several applications, such as mobile wireless communication, digital fingerprinting, and storage memories, the probability law characterizing the channel can change with time. This time-varying behavior of the channel probability is described typically in terms of the evolution of the underlying channel condition, termed "state." The availability of *channel state information* (CSI) at the transmitter or receiver can enhance overall communication performance (cf. [7, 1, 6]).

We consider a state dependent *discrete memoryless channel* (DMC), in which the underlying state process is independent and identically distributed (i.i.d.) with known probability mass function (pmf), and for which the channel output at any time instant depends on the inputs and states only through their current values. We address the cases of *causal and noncausal* CSI at the transmitter. In the former case, the transmitter has knowledge of all the past channel states as well as the current state; this model was introduced by Shannon [8]. In the latter case, the transmitter is provided access at the outset to the entire state sequence prevailing during the transmission of a message; see Gelfand-Pinsker [5]. We restrict ourselves to the situation where the receiver has no CSI, for receiver CSI can be accommodated by considering the states, too, as channel outputs.

Two information theoretic performance measures are of interest: *Channel capacity* and *reliability function*. The channel capacity characterizes the largest rate of encoded transmission for reliable communication. The reliability function describes the best exponential rate of decay of decoding error probability with transmission duration for coding rates below capacity. The capacities of the models above with causal and noncausal CSI were characterized in classic papers by Shannon [8] and Gelfand-Pinsker [5]. The reliability function is not fully characterized even for a DMC without states; however, good upper and lower bounds are known, which coincide at rates close to capacity [9, 10, 3].

Our contributions are twofold. First, we provide a *strong converse* for the capacity of state dependent channels, which explains the structure of optimal codes. Second, exploiting this structure, we obtain upper bounds for the reliability functions of the causal and noncausal CSI models. Instrumental to our proofs is a new technical result which provides an upper bound on the rate of codes with codewords that are "conditionally typical over large *message dependent* subsets of a typical set of state sequences." This technical result is a nonstraightforward analog of [3, Lemma 2.1.4] for a DMC without states; the latter provides a bound on the rate of a good code with codewords of a fixed composition. A preliminary conference version of this work is in [11].

In the next section, we compile pertinent technical concepts and tools that will be used to prove our results. These standard staples can be found, for instance, in [3, 2]. The channel models are described in Section 1.3. Sections 1.4-1.6 contain our main results.

1.2 Preliminaries: Types, typical sets and image sets

Let $\mathscr{X}$ be a finite set. For a sequence $\mathbf{x} \in \mathscr{X}^n$, the type of $\mathbf{x}$, denoted by $Q_{\mathbf{x}}$, is a pmf on $\mathscr{X}$, where $Q_{\mathbf{x}}(x)$ is the relative frequency of x in $\mathbf{x}$. Similarly, *joint types* are pmfs on product spaces. For example, the joint type of two given sequences $\mathbf{x} \in \mathscr{X}^n$, $\mathbf{s} \in \mathscr{S}^n$ is a pmf Q on $\mathscr{X} \times \mathscr{S}$, where $Q_{\mathbf{x},\mathbf{s}}(x,s)$ is the relative frequency of the tuple (x,s) among the tuples (x_t, s_t), $t = 1, ..., n$. Joint types of several n-length sequences are defined similarly.

The number of types of sequences in $\mathscr{X}^n$ is bounded above by $(n+1)^{|\mathscr{X}|}$. Denoting by $\mathscr{T}_Q^{(n)}$ the set of all sequences in $\mathscr{X}^n$ of type Q, we note that

$$(n+1)^{-\|\mathscr{X}\|} \exp[nH(Q)] \leq \left\| \mathscr{T}_Q^{(n)} \right\| \leq \exp[nH(Q)]. \tag{1.1}$$

Next, for any pmf $\mathtt{P}$ on $\mathscr{X}$, and type Q on $\mathscr{X}^n$,

$$\begin{aligned} \mathtt{P}^n(\mathbf{x}) &= \prod_{t=1}^{n} \mathtt{P}(x_t) = \prod_{x \in \mathscr{X}} \mathtt{P}(x)^{nQ(x)} \\ &= \exp[-n(D(\mathtt{P}\|Q) + H(Q))], \quad \mathbf{x} \in \mathscr{T}_Q^{(n)}, \end{aligned}$$

from which, along with (1.1), it follows that

$$(n+1)^{-\|\mathscr{X}\|} \exp[-n(D(\mathtt{P}\|Q)] \leq \mathtt{P}^n \left(\mathscr{T}_Q^{(n)} \right) \leq \exp[-n(D(\mathtt{P}\|Q)].$$

Next, for a pmf $\mathtt{P}$ on $\mathscr{X}$ and $\delta > 0$, a sequence $\mathbf{x} \in \mathscr{X}^n$ is $\mathtt{P}$-typical with constant δ if

$$\max_{x \in \mathscr{X}} |Q_{\mathbf{x}}(x) - \mathtt{P}(x)| \leq \delta,$$

and $\mathtt{P}(x) = 0$ implies $Q_{\mathbf{x}}(x) = 0$. The set of all $\mathtt{P}$-typical sequences with constant δ, is called the $\mathtt{P}$-typical set, denoted $\mathscr{T}_{[\mathtt{P}]}^{(n)}$ (where the dependence on δ is not displayed explicitly). Thus,

$$\mathscr{T}_{[\mathtt{P}]}^{(n)} = \bigcup_{\substack{\text{types } Q: \\ \max_{x \in \mathscr{X}} |Q_{\mathbf{x}}(x) - \mathtt{P}(x)| \leq \delta}} \mathscr{T}_Q^{(n)}.$$

In general, $\delta = \delta_n$ and is assumed to satisfy the "δ-convention" [3], namely

$$\delta_n \to 0, \quad \sqrt{n}\delta_n \to \infty \text{ as } n \to \infty. \tag{1.2}$$

The typical set has large probability. Precisely, for $\delta = \delta_n$ as in (1.2),

$$\mathtt{P}^n \left(\mathscr{T}_Q^{(n)} \right) \geq 1 - \frac{\|\mathscr{X}\|}{4n\delta^2}. \tag{1.3}$$

Consider sequences $\mathbf{x} \in \mathscr{X}^n$, $\mathbf{y} \in \mathscr{Y}^n$ of joint type $Q_{\mathbf{x},\mathbf{y}}$. The sequence $\mathbf{y} \in \mathscr{Y}^n$ has conditional type V if $Q_{\mathbf{x},\mathbf{y}} = Q_{\mathbf{x}}V$, for some stochastic matrix $V : \mathscr{X} \to \mathscr{Y}$. Given a stochastic matrix $W : \mathscr{X} \to \mathscr{Y}$, and $\mathbf{x} \in \mathscr{X}^n$, a sequence $\mathbf{y} \in \mathscr{Y}^n$ of conditional type V is W-conditionally typical if for all $x \in \mathscr{X}$

$$\max_{y \in \mathscr{Y}} |V(y \mid x) - W(y \mid x)| \leq \delta,$$

and $W(y \mid x) = 0$ implies $V(y \mid x) = 0$. The set of all W-conditionally typical sequences conditioned on $\mathbf{x} \in \mathscr{X}^n$, denoted by $\mathscr{T}^{(n)}_{[W]}(\mathbf{x})$. In a manner similar to (1.3), it holds that

$$W^n\left(\mathscr{T}^{(n)}_{[W]}(\mathbf{x}) \mid \mathbf{x}\right) \geq 1 - \frac{\|\mathscr{X}\|\|\mathscr{Y}\|}{4n\delta^2}.$$

For a subset A of $\mathscr{X}$, we shall require also estimates of the minimum cardinality of sets in $\mathscr{Y}$ with significant W-conditional probability given $x \in A$. Precisely, a set $B \subseteq \mathscr{Y}$ is an ε-image $(0 < \varepsilon \leq 1)$ of $A \subseteq \mathscr{X}$ under $W : \mathscr{X} \to \mathscr{Y}$ if $W(B \mid x) \geq \varepsilon$ for all $x \in A$. The minimum cardinality of ε-images of A is termed the image size of A (under W), and is denoted by $g_W(A, \varepsilon)$. Coding theorems in information theory use estimates of the rates of the image size of $A \subseteq \mathscr{X}^n$ under W^n, i.e., $(1/n)\log g_{W^n}(A, \varepsilon)$. In particular, for multiterminal systems, we compare the rates of image sizes of $A \subseteq \mathscr{X}^n$ under two different channels W^n and V^n. Precisely, given stochastic matrices $W : \mathscr{X} \to \mathscr{Y}$ and $V : \mathscr{X} \to \mathscr{S}$, for every $0 < \varepsilon < 1$, $\delta > 0$, and for every $A \subseteq \mathscr{T}^{(n)}_{[\mathtt{P}_X]}$, there exists an auxiliary rv U and associated pmfs $\mathtt{P}_{UXY} = \mathtt{P}_{U|X}\mathtt{P}_X W$ and $\mathtt{P}_{UXZ} = \mathtt{P}_{U|X}\mathtt{P}_X V$ such that

$$\left|\frac{1}{n}\log \mathtt{g}_{\mathtt{W}^{\mathtt{n}}}(B(m_0), \varepsilon) - H(Y|U) - t\right| < \delta, \tag{1.4}$$
$$\left|\frac{1}{n}\log \mathtt{g}_{\mathtt{V}^{\mathtt{n}}}(B(m_0), \varepsilon) - H(S|U) - t\right| < \delta,$$

where $0 \leq t \leq \min\{I(U \wedge Y), I(U \wedge S)\}$.

1.3 Channels with States

Consider a state dependent DMC $W : \mathscr{X} \times \mathscr{S} \to \mathscr{Y}$ with finite input, state and output alphabets $\mathscr{X}$, $\mathscr{S}$ and $\mathscr{Y}$, respectively. The $\mathscr{S}$-valued state process $\{S_t\}_{t=1}^{\infty}$ is i.i.d. with known pmf $\mathtt{P}_S$. The probability law of the DMC is specified by

$$W^n(\mathbf{y} \mid \mathbf{x}, \mathbf{s}) = \prod_{t=1}^{n} W(y_t \mid x_t, s_t), \quad \mathbf{x} \in \mathscr{X}^n, \mathbf{s} \in \mathscr{S}^n, \mathbf{y} \in \mathscr{Y}^n.$$

An (M,n)-code with encoder CSI consists of the mappings (f,ϕ) where the encoder mapping $f=(f_1,\ldots,f_n)$ is either *causal*, i.e.,

$$f_t : \mathscr{M}\times\mathscr{S}^t \to \mathscr{X}, \quad t=1,\ldots,n,$$

or *noncausal*, i.e.,

$$f_t : \mathscr{M}\times\mathscr{S}^n \to \mathscr{X}, \quad t=1,\ldots,n.$$

with $\mathscr{M}=\{1,\ldots,M\}$ being the set of messages. The decoder ϕ is a mapping

$$\phi : \mathscr{Y}^n \to \mathscr{M}.$$

We restrict ourselves to the situation where the receiver has no CSI. When the receiver, too, has CSI, our results apply in a standard manner by considering an associated DMC with augmented output alphabet $\mathscr{Y}\times\mathscr{S}$.

The rate of the code is $(1/n)\log M$. The corresponding (maximum) probability of error is

$$e(f,\phi) = \max_{m\in\mathscr{M}} \sum_{\mathbf{s}\in\mathscr{S}^n} \mathtt{P}_S(\mathbf{s}) W^n((\phi^{-1}(m))^c \mid f(m,\mathbf{s}),\mathbf{s}) \tag{1.5}$$

where $\phi^{-1}(m)=\{\mathbf{y}\in\mathscr{Y}^n : \phi(\mathbf{y})=m\}$ and $(\cdot)^c$ denotes complement.

Definition 1. Given $0<\varepsilon<1$, a number $R>0$ is ε-achievable if for every $\delta>0$ and for all n sufficiently large, there exist (M,n)-codes (f,ϕ) with $(1/n)\log M > R-\delta$ and $e(f,\phi)<\varepsilon$. The supremum of all ε-achievable rates is denoted by $C(\varepsilon)$. The capacity of the DMC is

$$C = \lim_{\varepsilon\to 0} C(\varepsilon).$$

If $C(\varepsilon)=C$ for $0<\varepsilon<1$, the DMC is said to satisfy a strong converse [12]. This terminology reflects the fact that for rates $R>C$, $e(f,\phi)>\varepsilon$ for $n\geq N(\varepsilon)$, $0<\varepsilon<1$. (In contrast, a standard converse shows that for $R>C$, $e(f,\phi)$ cannot be driven to 0 as $n\to\infty$.)

For a given pmf $\mathtt{P}_{\tilde{X}\tilde{S}}$ on $\mathscr{X}\times\mathscr{S}$, and an rv U with values in a finite set $\mathscr{U}$, let $\mathscr{P}(\mathtt{P}_{\tilde{X}\tilde{S}},W)$ denote the set of all pmfs $\mathtt{P}_{UXSY}$ on $\mathscr{U}\times\mathscr{X}\times\mathscr{S}\times\mathscr{Y}$ with

$$X = h(U,S) \tag{1.6}$$

for some mapping h,

$$U -\!\circ\!- X,S -\!\circ\!- Y, \qquad P_{X,S,Y} = \mathtt{P}_{\tilde{X}\tilde{S}}W. \tag{1.7}$$

For $\gamma\geq 0$, let $\mathscr{P}_\gamma(\mathtt{P}_{\tilde{X}\tilde{S}},W)$ be the subset of $\mathscr{P}(\mathtt{P}_{\tilde{X}\tilde{S}},W)$ with $I(U\wedge S)\leq\gamma$; note that $\mathscr{P}_0(\mathtt{P}_{\tilde{X}\tilde{S}},W)$ corresponds to the subset of $\mathscr{P}(\mathtt{P}_{\tilde{X}\tilde{S}},W)$ with U independent of S.

The classical results on the capacity of a state-dependent channel are due to Shannon [8] when the encoder CSI is causal, and Gelfand and Pinsker [5] when the encoder CSI is noncausal.

Theorem 1. *For the case with causal CSI, the capacity is*

$$C_{Sh} = \max_{\mathrm{P}_{X|S}} \max_{\mathscr{P}_0\left(\mathrm{P}_{X|S}\mathrm{P}_S, W\right)} I(U \wedge Y),$$

and holds with the strong converse.

Remark: The capacity formula was derived by Shannon [8], and the strong converse was proved later by Wolfowitz [12].

Theorem 2. *(Gelfand-Pinsker [5]) For the case with noncausal CSI, the capacity is*

$$C_{GP} = \max_{\mathrm{P}_{X|S}} \max_{\mathscr{P}\left(\mathrm{P}_{X|S}\mathrm{P}_S, W\right)} I(U \wedge Y) - I(U \wedge S).$$

One main result below is to show that the previous result, too, holds with a strong converse.

Definition 2. The *reliability function* $E(R),\ R \geq 0$, of the DMC W is the largest number $E \geq 0$ such that for every $\delta > 0$ and for all sufficiently large n, there exist n-length block codes (f, ϕ) with causal or noncausal CSI as above of rate greater than $R - \delta$ and $e(f, \phi) \leq \exp\left[-n(E - \delta)\right]$ (see for instance [3]).

1.4 A Technical Lemma

For a DMC without states, the result in [3, Corollary 6.4] provides, in effect, an image size characterization of a good codeword set; this does not involve any auxiliary rv. In the same spirit, our key technical lemma below provides an image size characterization for good codeword sets for the causal and noncausal DMC models, which now involves an auxiliary rv.

Lemma 1. *Let $\varepsilon, \tau > 0$ be such that $\varepsilon + \tau < 1$. Given a pmf $\mathrm{P}_{\tilde{S}}$ on $\mathscr{S}$ and conditional pmf $\tilde{P}_{X|S}$, let (f, ϕ) be a (M, n)-code as above. For each $m \in \mathscr{M}$, let $A(m)$ be a subset of $\mathscr{S}^n$ which satisfies the following conditions*

$$A(m) \subseteq \mathscr{T}^n_{[\mathrm{P}_{\tilde{S}}]}, \tag{1.8}$$

$$\|A(m)\| \geq \exp\left[n\left(H(\mathrm{P}_{\tilde{S}}) - \frac{\tau}{6}\right)\right], \tag{1.9}$$

$$f(m, \mathbf{s}) \in \mathscr{T}^n_{[\mathrm{P}_{\tilde{X}|\tilde{S}}]}(\mathbf{s}), \quad \mathbf{s} \in A(m). \tag{1.10}$$

Furthermore, let (f, ϕ) satisfy one of the following two conditions

$$W^n(\phi^{-1}(m) \mid f(m, \mathbf{s}), \mathbf{s}) \geq 1 - \varepsilon, \quad \mathbf{s} \in A(m), \tag{1.11a}$$

$$\frac{1}{\|A(m)\|} \sum_{\mathbf{s} \in A(m)} W^n(\phi^{-1}(m) \mid f(m, \mathbf{s}), \mathbf{s}) \geq 1 - \varepsilon. \tag{1.11b}$$

(a) *In the causal CSI case, for $n \geq N(\|\mathscr{X}\|, \|\mathscr{S}\|, \|\mathscr{Y}\|, \tau, \varepsilon)$*[1]*, it holds that*

$$\frac{1}{n}\log M \leq I(U \wedge Y) + \tau,$$

for some $\mathtt{P}_{UXSY} \in \mathscr{P}_\tau(\mathtt{P}_{\tilde{X}|\tilde{S}}\mathtt{P}_{\tilde{S}}, W)$.
(b) *In the noncausal CSI case, for $n \geq N(\|\mathscr{X}\|, \|\mathscr{S}\|, \|\mathscr{Y}\|, \tau, \varepsilon)$, it holds that*

$$\frac{1}{n}\log M \leq I(U \wedge Y) - I(U \wedge S) + \tau,$$

for some $\mathtt{P}_{UXSY} \in \mathscr{P}(\mathtt{P}_{\tilde{X}|\tilde{S}}\mathtt{P}_{\tilde{S}}, W)$.

Furthermore, in both cases it suffices to restrict the rv U to take values in a finite set $\mathscr{U}$ with $\|\mathscr{U}\| \leq \|\mathscr{X}\|\|\mathscr{S}\| + 1$.

Proof: Our proof below is for the case when (1.11a) holds. The case when (1.11b) holds can be proved similarly with minor modifications; specifically, in the latter case, we can find subsets $A'(m)$ of $A(m)$, $m \in \mathscr{M}$, that satisfy (1.8)-(1.10) and (1.11a) for some $\varepsilon', \tau' > 0$ with $\varepsilon' + \tau' < 1$ for all n sufficiently large.

With (1.11a) holding, set

$$B(m) = \{(f(m, \mathbf{s}), \mathbf{s}) \in \mathscr{X}^n \times \mathscr{S}^n : \mathbf{s} \in A(m)\}, \quad m \in \mathscr{M}.$$

Let $\mathtt{P}_{\tilde{Y}} = \mathtt{P}_{\tilde{X}\tilde{S}}W$ be a pmf on $\mathscr{Y}$ defined by

$$\mathtt{P}_{\tilde{Y}}(y) = \sum_{s,x} \mathtt{P}_{\tilde{S}\tilde{X}}(s,x) W(y \mid x, s), \quad y \in \mathscr{Y}.$$

Consequently,

$$W^n(\mathscr{T}^n_{[\mathtt{P}_{\tilde{Y}}]} \mid f(m, \mathbf{s}), \mathbf{s}) > \varepsilon + \tau, \quad \mathbf{s} \in A(m), \tag{1.12}$$

for all $n \geq N(\|\mathscr{X}\|, |\mathscr{S}\|, |\mathscr{Y}\|, \tau, \varepsilon)$ (not depending on m and $\mathbf{s}$ in $A(m)$). Denoting

$$C(m) = \phi^{-1}(m) \cap \mathscr{T}^n_{[\mathtt{P}_{\tilde{Y}}]},$$

we see from (1.11a) and (1.12) that

$$W^n(C(m) \mid f(m, \mathbf{s}), \mathbf{s}) > \tau > 0, \quad , (f(m, \mathbf{s}), \mathbf{s}) \in B(m),$$

so that

$$\|C(m)\| \geq g_{W^n}(B(m), \tau),$$

where $g_{W^n}(B(m), \tau)$ denotes the smallest cardinality of a subset D of $\mathscr{Y}^n$ with

[1] In our assertions, we indicate the validity of a statement "for all $n \geq N(.)$" by showing the explicit dependency of N; however the standard picking of the "largest such N" from (finitely-many) such Ns is not indicated.

$$W^n(D \mid (f(m,\mathbf{s}),\mathbf{s})) > \tau, \quad (f(m,\mathbf{s}),\mathbf{s}) \in B(m). \tag{1.13}$$

With $m_0 = \mathtt{arg}\min_{1\leq m\leq M} \|C(m)\|$, we have

$$M\|C(m_0)\| \leq \sum_{m=1}^{M} \|C(m)\| = \|\mathscr{T}^n_{[\mathtt{P}_{\tilde{Y}}]}\| \leq \exp n\left(H(\mathtt{P}_{\tilde{Y}}) + \frac{\tau}{6}\right).$$

Consequently,

$$\frac{1}{n}\log M \leq H(\mathtt{P}_{\tilde{Y}}) + \frac{\tau}{6} - \frac{1}{n}\log \mathrm{g}_{\mathtt{W}^n}(B(m_0),\tau). \tag{1.14}$$

The remainder of the proof entails relating the "image size" of $B(m_0)$, i.e., $g_{W^n}(B(m_0),\tau)$, to $\|A(m_0)\|$, and is completed below separately for the cases of causal and noncausal CSI.

First consider the causal CSI case. For a rv $\hat{S}^n$ distributed uniformly over $A(m_0)$, we have from (1.9) that

$$\frac{1}{n}H(\hat{S}^n) = \frac{1}{n}\log\|A(m_0)\| \geq H(\mathtt{P}_{\tilde{S}}) - \frac{\tau}{6}. \tag{1.15}$$

Since

$$\frac{1}{n}H(\hat{S}^n) = \frac{1}{n}\sum_{i=1}^{n} H(\hat{S}_i \mid \hat{S}^{i-1}) = H(\hat{S}_I \mid \hat{S}^{I-1}, I),$$

where the rv I is distributed uniformly over the set $\{1,\ldots,n\}$ and is independent of all other rvs, the previous identity, together with (1.15), yields

$$H(\mathtt{P}_{\tilde{S}}) - H(\hat{S}_I \mid \hat{S}^{I-1}, I) \leq \frac{\tau}{3}. \tag{1.16}$$

Next, denote by $\hat{X}^n$ the rv $f(m_0,\hat{S}^n)$ and by $\hat{Y}^n$ the rv which conditioned on $\hat{X}^n$, $\hat{S}^n$, has (conditional) distribution W^n, i.e., $\hat{Y}^n$ is the random output of the DMC W when the input is set to $\left(\hat{X}^n,\hat{S}^n\right)$. Then, using [3, Lemma 15.2], we get

$$\frac{1}{n}\log \mathrm{g}_{\mathtt{W}^n}(B(m_0),\tau) \geq \frac{1}{n}H(\hat{Y}^n) - \frac{\tau}{6}, \tag{1.17}$$

for all n sufficiently large. Furthermore,

$$\begin{aligned}
\frac{1}{n}H(\hat{Y}^n) &= \frac{1}{n}\sum_{i=1}^{n} H(\hat{Y}_i \mid \hat{Y}^{i-1}) \\
&\geq H(\hat{Y}_I \mid \hat{X}^{I-1}, \hat{S}^{I-1}, \hat{Y}^{I-1}, I) \\
&= H(\hat{Y}_I \mid \hat{X}^{I-1}, \hat{S}^{I-1}, I) \\
&= H(\hat{Y}_I \mid \hat{S}^{I-1}, I),
\end{aligned}$$

where the last-but-one equality follows from the DMC assumption, and the last equality holds since $\hat{X}^{I-1} = f(m_0, \hat{S}^{I-1})$. The inequality above, along with (1.17) and (1.14) gives

$$\frac{1}{n}\log M \le H(\mathtt{P}_{\tilde{Y}}) - H(\hat{Y}_I \mid \hat{S}^{I-1}, I) + \frac{\tau}{3}. \tag{1.18}$$

Denote by $\hat{U}$ the rv $(\hat{S}^{I-1}, I)$ and note that the following Markov property holds:

$$\hat{Y}_I -\!\!\circ\!\!- \hat{X}_I, \hat{S}_I -\!\!\circ\!\!- \hat{U}.$$

Also, from the definition of $B(m_0)$,

$$\begin{aligned}
\mathtt{P}_{\hat{X}_I,\hat{S}_I}(x,s) &= \frac{1}{n}\sum_{i=1}^{n} \mathtt{P}_{\hat{X}_i,\hat{S}_i}(x,s) \\
&= \frac{1}{n}\sum_{i=1}^{n} \sum_{\mathbf{x},\mathbf{s}\in B(m_0)} \frac{\mathbf{1}(x_i = x, s_i = s)}{\|B(m_0)\|} \\
&= \frac{1}{\|B(m_0)\|} \sum_{\mathbf{x},\mathbf{s}\in B(m_0)} Q_{\mathbf{x},\mathbf{s}}(x,s),
\end{aligned}$$

where $Q_{\mathbf{x},\mathbf{s}}(x,s)$ is the joint type of $\mathbf{x},\mathbf{s}$, and the last equation follows upon interchanging the order of summation. It follows from (1.8) and (1.10) that $\|\mathtt{P}_{\hat{X}_I,\hat{S}_I} - \mathtt{P}_{\tilde{X}\tilde{S}}\| \le \delta_n$ for some $\delta_n \to 0$ satisfying the delta-convention. Furthermore,

$$\begin{aligned}
\|\mathtt{P}_{\hat{Y}_I} - \mathtt{P}_{\tilde{Y}}\| &= \sum_{y} \left| \sum_{x,s} W(y|x,s)\mathtt{P}_{\tilde{X}\tilde{S}}(x,s) - \sum_{x,s} W(y|x,s)\mathtt{P}_{\hat{X}_I\hat{S}_I}(x,s) \right| \\
&\le \sum_{x,s}\sum_{y} W(y|x,s) \left| \mathtt{P}_{\tilde{X}\tilde{S}}(x,s) - \mathtt{P}_{\hat{X}_I\hat{S}_I}(x,s) \right| \\
&= \|\mathtt{P}_{\tilde{X}\tilde{S}} - \mathtt{P}_{\hat{X}_I\hat{S}_I}\| \le \delta_n.
\end{aligned}$$

Let the rvs $\tilde{X}, \tilde{S}, \tilde{Y}$ have a joint distribution $\mathtt{P}_{\tilde{X}\tilde{S}\tilde{Y}}$. Define a rv U which takes values in the same set as $\hat{U}$, has $\mathtt{P}_{\hat{U}|\hat{X}_I\hat{S}_I}$ as its conditional distribution given X, S, and satisfies the Markov relation

$$Y -\!\!\circ\!\!- X, S -\!\!\circ\!\!- U.$$

Then using the continuity of the entropy function and the arguments above, (1.18) yields

$$\frac{1}{n}\log M \le I(U \wedge Y) + \tau,$$

and (1.16) yields

$$I(U \wedge S) \le \tau,$$

for all n sufficiently large, where $\mathtt{P}_{UXSY} \in \mathscr{P}_\tau(\mathtt{P}_{\tilde{X}\tilde{S}}, W)$.

Turning to the case with noncausal CSI, define a stochastic matrix $V : \mathscr{X} \times \mathscr{S} \to \mathscr{S}$ with

$$V(s' \mid x, s) = \mathbf{1}(s' = s),$$

and let $\mathtt{g}_{\mathtt{V}^\mathtt{n}}$ be defined in a manner analogous to $\mathtt{g}_{\mathtt{W}^\mathtt{n}}$ above with $\mathscr{S}^n$ in the role of $\mathscr{Y}^n$ in (1.13). For any $m \in \mathscr{M}$ and subset E of $\mathscr{S}^n$, observe that

$$V^n(E \mid f(m, \mathbf{s}), \mathbf{s}) = \mathbf{1}(s \in E), \quad \mathbf{s} \in \mathscr{S}^n.$$

In particular, if E satisfies

$$V^n(E \mid f(m, \mathbf{s}), \mathbf{s}) > \tau, \quad \mathbf{s} \in A(m), \tag{1.19}$$

it must be that $A(m) \subseteq E$, and since $E = A(m)$ satisfies (1.19), we get that

$$\|A(m)\| = \mathtt{g}_{\mathtt{V}^\mathtt{n}}(B(m), \tau) \tag{1.20}$$

using the definition of $B(m)$. Using the image size characterization in (1.4) [3, Theorem 15.11], there exists an auxiliary rv U and associated pmf $\mathtt{P}_{UXSY} = \mathtt{P}_{U|XS}\mathtt{P}_{\tilde{X}\tilde{S}}W$ such that

$$\begin{aligned} \left| \frac{1}{n} \log \mathtt{g}_{\mathtt{W}^\mathtt{n}}(B(m_0), \tau) - H(Y|U) - t \right| &< \frac{\tau}{6}, \\ \left| \frac{1}{n} \log \mathtt{g}_{\mathtt{V}^\mathtt{n}}(B(m_0), \tau) - H(S|U) - t \right| &< \frac{\tau}{6}, \end{aligned} \tag{1.21}$$

where $0 \leq t \leq \min\{I(U \wedge Y), I(U \wedge S)\}$. Then, using (1.14), (1.20) and (1.21) we get

$$\frac{1}{n} \log M \leq I(U \wedge Y) + H(S \mid U) - \frac{1}{n} \log \|A(m_0)\| + \frac{\tau}{2},$$

which, by (1.9), yields

$$\frac{1}{n} \log M \leq I(U \wedge Y) - I(U \wedge S) + \tau.$$

In (1.21), $\mathtt{P}_{UXSY} \in \mathscr{P}(\mathtt{P}_{\tilde{X}|\tilde{S}}\mathtt{P}_{\tilde{S}}, W)$ but need not satisfy (1.6). Finally, the asserted restriction to $\mathtt{P}_{UXSY} \in \mathscr{P}(\mathtt{P}_{\tilde{X}|\tilde{S}}\mathtt{P}_{\tilde{S}}, W)$ follows from the convexity of $I(U \wedge Y) - I(U \wedge S)$ in $\mathtt{P}_{X|US}$ for a fixed $\mathtt{P}_{US}$ (as observed in [5]).

Lastly, it follows from the support lemma [3, Lemma 15.4] that it suffices to consider those rvs U for which $\|\mathscr{U}\| \leq \|\mathscr{X}\|\|\mathscr{S}\| + 1$. □

1.5 The Strong Converse

Theorem 3. *(Strong Converse) Given* $0 < \varepsilon < 1$ *and a sequence of* (M_n, n) *codes* (f_n, ϕ_n) *with* $e(f_n, \phi_n) < \varepsilon$*, it holds that*

$$\limsup_n \frac{1}{n} \log M_n \le C,$$

where $C = C_{Sh}$ *and* C_{GP} *for the cases of causal and noncausal CSI, respectively.*

Proof: Given $0 < \varepsilon < 1$ and a (M,n)-code (f, ϕ) with $e(f,\phi) \le \varepsilon$, the proof involves the identification of sets $A(m)$, $m \in \mathscr{M}$, satisfying (1.8)-(1.10) and (1.11a). The assertion then follows from Lemma 1. Note that $e(f,\phi) \le \varepsilon$ implies

$$\sum_{\mathbf{s} \in \mathscr{S}^n} \mathtt{P}_S(\mathbf{s}) W^n(\phi^{-1}(m) \mid f(m,\mathbf{s}), \mathbf{s}) \ge 1 - \varepsilon$$

for all $m \in \mathscr{M}$. Since $\mathtt{P}_S\left(\mathscr{T}^n_{[\mathtt{P}_S]}\right) \to 1$ as $n \to \infty$, we get that for every $m \in \mathscr{M}$,

$$\mathtt{P}_S\left(\left\{\mathbf{s} \in \mathscr{T}^n_{[\mathtt{P}_S]} : W^n(\phi^{-1}(m) \mid f(m,\mathbf{s}), \mathbf{s}) > \frac{1-\varepsilon}{2}\right\}\right) \ge \frac{1-\varepsilon}{3} \tag{1.22}$$

for all $n \ge N(\|\mathscr{S}\|, \varepsilon)$. Denoting the set $\left\{\cdot\right\}$ in (1.22) by $\hat{A}(m)$, clearly for every $m \in \mathscr{M}$,

$$W^n(\phi^{-1}(m) \mid f(m,\mathbf{s}), \mathbf{s}) \ge \frac{1-\varepsilon}{2}, \quad \mathbf{s} \in \hat{A}(m),$$

and

$$\mathtt{P}_S\left(\hat{A}(m)\right) \ge \frac{1-\varepsilon}{3}$$

for all $n \ge N(\|\mathscr{S}\|, \varepsilon)$, whereby for an arbitrary $\delta > 0$, we get

$$\|\hat{A}(m)\| \ge \exp\left[n(H(\mathtt{P}_S) - \delta)\right]$$

for all $n \ge N(\|\mathscr{S}\|, \delta)$. Partitioning $\hat{A}(m)$, $m \in \mathscr{M}$, into sets according to the (polynomially many) conditional types of $f(m,\mathbf{s})$ given $\mathbf{s}$ in $\hat{A}(m)$, we obtain a subset $A(m)$ of $\hat{A}(m)$ for which

$$f(m,\mathbf{s}) \in \mathscr{T}^n_m(\mathbf{s}), \quad \mathbf{s} \in A(m),$$
$$\|A(m)\| \ge \exp\left[n(H(\mathtt{P}_S) - 2\delta)\right],$$

for all $n \ge N(\|\mathscr{S}\|, \|\mathscr{X}\|, \delta)$, where $\mathscr{T}^n_m(\mathbf{s})$ represents a set of those sequences in $\mathscr{X}^n$ that have the same conditional type (depending only on m).

Once again, the polynomial size of such conditional types yields a subset $\mathscr{M}'$ of $\mathscr{M}$ such that $f(m,\mathbf{s})$ has a fixed conditional type (not depending on m) given $\mathbf{s}$ in $A(m)$, and with

$$\frac{1}{n}\log\|\mathscr{M}'\| \geq \frac{1}{n}\log M - \delta$$

for all $n \geq N(\|\mathscr{S}\|,\|\mathscr{X}\|,\delta)$. Finally, the strong converse follows by applying Lemma 1 to the subcode corresponding to $\mathscr{M}'$ and noting that $\delta > 0$ is arbitrary. □

1.6 Outer Bound on Reliability Function

An upper bound for the reliability function $E(R)$, $0 < R < C$, of a DMC without states, is derived in [3] using a strong converse for codes with codewords of a fixed type. The key technical Lemma 1 above gives an upper bound on the rate of codes with codewords that are conditionally typical over large *message dependent* subsets of the typical set of state sequences, and serves, in effect, as an analog of [3, Corollary 6.4] for state dependent channels to derive an upper bound on the reliability function.

Theorem 4. *(Sphere Packing Bound) Given $\delta > 0$, for $0 < R < C$, it holds that*

$$E(R) \leq E_{SP}(1+\delta)+\delta,$$

where

$$E_{SP} = \min_{\mathrm{P}_{\tilde{S}}}\max_{\mathrm{P}_{\tilde{X}|\tilde{S}}}\min_{V\in\mathscr{V}(R,\mathrm{P}_{\tilde{X}\tilde{S}})}\left[D(\mathrm{P}_{\tilde{S}}\|\mathrm{P}_S)+D(V\|W\mid \mathrm{P}_{\tilde{X}\tilde{S}})\right] \tag{1.23}$$

with

$$\mathscr{V}(R,\mathrm{P}_{\tilde{X}\tilde{S}}) = \mathscr{V}_{Sh}(R,\mathrm{P}_{\tilde{X}\tilde{S}}) = \left\{V:\mathscr{X}\times\mathscr{S}\to\mathscr{Y}: \max_{P_{UXSY}\in\mathscr{P}_0(\mathrm{P}_{\tilde{X}\tilde{S}},V)} I(U\wedge Y) < R\right\}, \tag{1.24}$$

and

$$\mathscr{V}(R,\mathrm{P}_{\tilde{X}\tilde{S}}) = \mathscr{V}_{GP}(R,\mathrm{P}_{\tilde{X}\tilde{S}}) = \left\{V:\mathscr{X}\times\mathscr{S}\to\mathscr{Y}: \max_{P_{UXSY}\in\mathscr{P}(\mathrm{P}_{\tilde{X}\tilde{S}},V)} I(U\wedge Y)-I(U\wedge S) < R\right\}, \tag{1.25}$$

for the causal and noncausal CSI cases, respectively.

Remark: In (1.23), the terms $D(\mathrm{P}_{\tilde{S}}\|\mathrm{P}_S)$ and $D(V\|W\mid \mathrm{P}_{\tilde{S}}\mathrm{P}_{\tilde{X}|\tilde{S}})$ account, respectively, for the shortcomings of a given code for corresponding “bad” state pmf and “bad” channel.

Proof: Consider sequences of type $\mathrm{P}_{\tilde{S}}$ in $\mathscr{S}^n$. Picking $\hat{A}(m) = \mathscr{T}^n_{\mathrm{P}_{\tilde{S}}}$, $m \in \mathscr{M}$, in the proof of Theorem 3, and following the arguments therein to extract the subset $A(m)$ of $\hat{A}(m)$, we have for a given $\delta > 0$ that for all $n \geq N(\|\mathscr{S}\|, \|\mathscr{X}\|, \delta)$, there exists a subset $\mathscr{M}'$ of $\mathscr{M}$ and a fixed conditional type, say $\mathrm{P}_{\tilde{X}|\tilde{S}}$ (not depending on m), such that for every $m \in \mathscr{M}'$,

$$\begin{aligned} A(m) &\subseteq \hat{A}(m) = \mathscr{T}^n_{\mathrm{P}_{\tilde{S}}}, \\ \|A(m)\| &\geq \exp\left[n(H(\mathrm{P}_{\tilde{S}}) - \delta)\right], \\ f(m,\mathbf{s}) &\in \mathscr{T}^n_{\mathrm{P}_{\tilde{X}|\tilde{S}}}(\mathbf{s}), \qquad \mathbf{s} \in A(m), \\ \frac{1}{n}\log\|\mathscr{M}'\| &\geq R - \delta. \end{aligned}$$

Then for every $V \in \mathscr{V}(R, \mathrm{P}_{\tilde{X}\tilde{S}})$, we obtain using Lemma 1 (in its version with condition (1.11b)), that for every $\delta' > 0$, there exists $m \in \mathscr{M}'$ (possibly depending on δ' and V) with

$$\frac{1}{\|A(m)\|}\sum_{\mathbf{s}\in A(m)} V^n((\phi^{-1}(m))^c \mid f(m,\mathbf{s}),\mathbf{s}) \geq 1 - \delta'$$

for all $n \geq N(\|\mathscr{S}\|, \|\mathscr{X}\|, \|\mathscr{Y}\|, \delta')$. Since the average V^n-(conditional) probability of $\left(\phi^{-1}(m)\right)^c$ is large, its W^n-(conditional) probability cannot be too small. To that end, for this m, apply [3, Theorem 10.3, (10.21)] with the choices

$$\begin{aligned} Z &= \mathscr{Y}^n \times A(m), \\ S &= (\phi^{-1}(m))^c \times A(m), \\ Q_1(\mathbf{y},\mathbf{s}) &= \frac{V^n(\mathbf{y} \mid f(m,\mathbf{s}),\mathbf{s})}{\|A(m)\|}, \\ Q_2(\mathbf{y},\mathbf{s}) &= \frac{W^n(\mathbf{y} \mid f(m,\mathbf{s}),\mathbf{s})}{\|A(m)\|}, \end{aligned}$$

for $(\mathbf{y},\mathbf{s}) \in Z$, to obtain

$$\frac{1}{\|A(m)\|}\sum_{\mathbf{s}\in A(m)} W^n((\phi^{-1}(m))^c \mid f(m,\mathbf{s}),\mathbf{s}) \geq \exp\left(-\frac{nD(V\|W \mid \mathrm{P}_{\tilde{X}|\tilde{S}}\mathrm{P}_{\tilde{S}}) + 1}{1-\delta'}\right).$$

Finally,

$$\begin{aligned} e(f,\phi) &\geq \sum_{\mathbf{s}\in A(m)} \mathrm{P}_S(\mathbf{s}) W^n(\left(\phi^{-1}(m)\right)^c \mid f(m,\mathbf{s}),\mathbf{s}) \\ &\geq \exp[-n(D(\mathrm{P}_{\tilde{S}}\|\mathrm{P}_S) + D(V\|W \mid \mathrm{P}_{\tilde{X}|\tilde{S}}\mathrm{P}_{\tilde{S}})(1+\delta) + \delta)] \end{aligned}$$

for $n \geq N(\|\mathscr{S}\|, \|\mathscr{X}\|, \|\mathscr{Y}\|, \delta, \delta')$, whereby it follows for the noncausal CSI case that

$$\limsup_n -\frac{1}{n}\log e(f,\phi) \leq \min_{\mathrm{P}_{\tilde{S}}} \max_{\mathrm{P}_{\tilde{X}|\tilde{S}}} \min_{V \in \mathscr{V}(R, \mathrm{P}_{\tilde{X}\tilde{S}})} [D(\mathrm{P}_{\tilde{S}}\|\mathrm{P}_S) + D(V\|W \mid \mathrm{P}_{\tilde{X}|\tilde{S}}\mathrm{P}_{\tilde{S}})(1+\delta)+\delta]$$

for every $\delta > 0$. Similarly, for the case of causal CSI, for $\tau > 0$, letting

$$\mathscr{V}_\tau(R, \mathrm{P}_{\tilde{X}\tilde{S}}) = \left\{V : \mathscr{X} \times \mathscr{S} \to \mathscr{Y} : \max_{P_{UXSY} \in \mathscr{P}_\tau(\mathrm{P}_{\tilde{X}\tilde{S}}, V)} I(U \wedge Y) < R\right\}, \tag{1.26}$$

we get

$$\limsup_n -\frac{1}{n}\log e(f,\phi) \leq \min_{\mathrm{P}_{\tilde{S}}} \max_{\mathrm{P}_{\tilde{X}|\tilde{S}}} \min_{V \in \mathscr{V}_\tau(R, \mathrm{P}_{\tilde{X}\tilde{S}})} [D(\mathrm{P}_{\tilde{S}}\|\mathrm{P}_S) + D(V\|W \mid \mathrm{P}_{\tilde{X}|\tilde{S}}\mathrm{P}_{\tilde{S}})].$$

Finally, the continuity of the right side of (1.26), as shown in the Appendix, yields the claimed expression for E_{SP} in (1.23), (1.24). □

Acknowledgements The work of H. Tyagi and P. Narayan was supported by the U.S. National Science Foundation under Grants CCF0830697 and CCF1117546.

References

1. E. Biglieri, J. Proakis, and S. Shamai (Shitz). Fading channels: information-theoretic and communications aspects. *IEEE Trans. Inform. Theory*, 44(6):2619 – 2692, 1998.
2. I. Csiszár. The method of types. *IEEE Trans. Inform. Theory*, 44(6):2505–2523, 1998.
3. I. Csiszár and J. Körner. *Information theory: coding theorems for discrete memoryless channels*. 2nd Edition. Cambridge University Press, 2011.
4. A. A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's inequality. *IEEE Trans. Inform. Theory*, 49(6):1491–1498, 2003.
5. S. I. Gelfand and M. S. Pinsker. Coding for channels with random parameters. *Problem of Control and Information Theory*, 9(1):19–31, 1980.
6. G. Keshet, Y. Steinberg, and N. Merhav. Channel coding in presence of side information. *Foundations and Trends in Communication and Information Theory*, 4(6):445–586, 2008.
7. A. Lapidoth and P. Narayan. Reliable communication under channel uncertainty. *IEEE Trans. Inform. Theory*, 44(6):2148 – 2177, 1998.
8. C. E. Shannon. Channels with side information at the transmitter. *IBM Journal of Research and Development*, 2:289–293, 1958.
9. C. E. Shannon, R. G. Gallager, and E. R. Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels-i. *Information and Control*, pages 65–103, Dec 1966.
10. C. E. Shannon, R. G. Gallager, and E. R. Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels-ii. *Information and Control*, pages 522–552, May 1967.
11. H. Tyagi and P. Narayan The Gelfand-Pinsker channel: strong converse and upper bound for the reliability function. *Proceedings of the IEEE International Symposium on Information Theory*, Seoul, Korea, 2009.
12. J. Wolfowitz. *Coding theorems of information theory*. New York:Springer-Verlag, 1978.

Appendix
Continuity of the right side of (1.26)

Let

$$f(R, \mathsf{P}_{U\tilde{X}\tilde{S}}) = \min_{\substack{V: I(U \wedge Y) < R \\ \mathsf{P}_{Y|\tilde{X}\tilde{S}} = V}} D(\mathsf{P}_{\tilde{S}} \| \mathsf{P}_S) + D(V \| W \mid \mathsf{P}_{\tilde{X}|\tilde{S}} \mathsf{P}_{\tilde{S}}). \tag{1.27}$$

Further, let

$$g(\mathsf{P}_{\tilde{S}}, \tau) = \max_{\substack{\mathsf{P}_{U\tilde{X}|\tilde{S}}: I(U \wedge \tilde{S}) \leq \tau \\ U -\!\circ\!- \tilde{X}, \tilde{S} -\!\circ\!- Y}} f(R, \mathsf{P}_{U\tilde{X}\tilde{S}}), \tag{1.28}$$

and

$$g(\tau) = \min_{\mathsf{P}_{\tilde{S}}} g(\mathsf{P}_{\tilde{S}}, \tau). \tag{1.29}$$

To show the continuity of $g(\tau)$ at $\tau = 0$, first note that $g(\tau) \geq g(0)$ for all $\tau \geq 0$. Next, let $\mathsf{P}^0_{\tilde{S}}$ attain the minimum in (1.29) for $\tau = 0$. Clearly,

$$g(\mathsf{P}^0_{\tilde{S}}, \tau) \geq g(\tau). \tag{1.30}$$

Also, let $\mathsf{P}^\tau_{U\tilde{X}|\tilde{S}}$ attain the maximum of $g(\mathsf{P}^0_{\tilde{S}}, \tau)$ in (1.28). For the associated joint pmf $\mathsf{P}^0_{\tilde{S}} \mathsf{P}^\tau_{U\tilde{X}|\tilde{S}}$, let P^τ_U denote the resulting U-marginal pmf, and consider the joint pmf $\mathsf{P}^\tau_U \mathsf{P}^0_{\tilde{S}} \mathsf{P}^\tau_{\tilde{X}|U\tilde{S}}$. Then, using (1.28), (1.29) and the observations above,

$$f(R, \mathsf{P}^\tau_U \mathsf{P}^0_{\tilde{S}} \mathsf{P}^\tau_{\tilde{X}|U\tilde{S}}) \leq g(0) \leq g(\tau) \leq g(\mathsf{P}^0_{\tilde{S}}, \tau) = f(R, \mathsf{P}^0_{\tilde{S}} \mathsf{P}^\tau_{U\tilde{X}|\tilde{S}}).$$

The continuity of $g(\tau)$ at $\tau = 0$ will follow upon showing that

$$f(R, \mathsf{P}^0_{\tilde{S}} \mathsf{P}^\tau_{U\tilde{X}|\tilde{S}}) - f(R, \mathsf{P}^\tau_U \mathsf{P}^0_{\tilde{S}} \mathsf{P}^\tau_{\tilde{X}|U\tilde{S}}) \to 0 \text{ as } \tau \to 0.$$

The constraint on the mutual information (1.28) gives by Pinsker's inequality [3, 4] that,

$$\tau \geq D\left(\mathsf{P}^\tau_{U|\tilde{S}} \mathsf{P}^0_{\tilde{S}} \| \mathsf{P}^\tau_U \mathsf{P}^0_{\tilde{S}}\right) \geq 2 \left\| \mathsf{P}^\tau_{U|\tilde{S}} \mathsf{P}^0_{\tilde{S}} - \mathsf{P}^\tau_U \mathsf{P}^0_{\tilde{S}} \right\|^2,$$

i.e.,

$$\left\| \mathsf{P}^\tau_{U|\tilde{S}} \mathsf{P}^0_{\tilde{S}} - \mathsf{P}^\tau_U \mathsf{P}^0_{\tilde{S}} \right\| \leq \sqrt{\frac{\tau}{2}}. \tag{1.31}$$

For $\mathsf{P}_{U\tilde{X}\tilde{S}} = \mathsf{P}^\tau_U \mathsf{P}^0_{\tilde{S}} \mathsf{P}^\tau_{\tilde{X}|U\tilde{S}}$, let V^0 attain the minimum in (1.26), i.e.,

$$\mathrm{P}_{Y|\tilde{X}\tilde{S}} = V^0, \quad I(U \wedge Y) < R, \text{ and}$$
$$f(R, \mathrm{P}_U^\tau \mathrm{P}_{\tilde{S}}^0 \mathrm{P}_{\tilde{X}|U\tilde{S}}^\tau) = D(\mathrm{P}_{\tilde{S}} \| \mathrm{P}_S) + D(V^0 \| W \mid \mathrm{P}_{\tilde{X}|\tilde{S}} \mathrm{P}_{\tilde{S}}).$$

By (1.31), for $\mathrm{P}_{U\tilde{X}\tilde{S}} = \mathrm{P}_{\tilde{S}}^0 \mathrm{P}_{U\tilde{X}|U\tilde{S}}^\tau$ and $\mathrm{P}_{Y|\tilde{X}\tilde{S}} = V^0$, by standard continuity arguments we have

$$I(U \wedge Y) < R + \nu,$$

and

$$D(\mathrm{P}_{\tilde{S}} \| \mathrm{P}_S) + D(V^0 \| W \mid \mathrm{P}_{\tilde{X}|\tilde{S}} \mathrm{P}_{\tilde{S}}) \leq f(R, \mathrm{P}_U^\tau \mathrm{P}_{\tilde{S}}^0 \mathrm{P}_{\tilde{X}|U\tilde{S}}^\tau) + \nu,$$

where $\nu = \nu(\tau) \to 0$ as $\tau \to 0$. Consequently,

$$f(R, \mathrm{P}_{\tilde{S}}^0 \mathrm{P}_{U\tilde{X}|U\tilde{S}}^\tau) \leq D(\mathrm{P}_{\tilde{S}} \| \mathrm{P}_S) + D(V^0 \| W \mid \mathrm{P}_{\tilde{X}|\tilde{S}} \mathrm{P}_{\tilde{S}}) \leq f(R, \mathrm{P}_U^\tau \mathrm{P}_{\tilde{S}}^0 \mathrm{P}_{\tilde{X}|U\tilde{S}}^\tau) + \nu.$$

Finally, noting the continuity of $f(R, \mathrm{P}_{U\tilde{X}\tilde{S}})$ in R [3, Lemma 10.4], the proof is completed. □