Channels with cost constraints: strong converse and dispersion

Victoria Kostina
California Institute of Technology
Pasadena, CA 91125
vkostina@caltech.edu

Sergio Verdú
Princeton University
Princeton, NJ 08544
verdu@princeton.edu

***Abstract*—This paper shows the strong converse and the dispersion of memoryless channels with cost constraints and performs refined analysis of the third order term in the asymptotic expansion of the maximum achievable channel coding rate, showing that it is equal to $\frac{1}{2}\frac{\log n}{n}$ in most cases of interest. The analysis is based on a non-asymptotic converse bound expressed in terms of the distribution of a random variable termed the b-tilted information density, which plays a role similar to that of the d-tilted information in lossy source coding. We also analyze the fundamental limits of lossy joint-source-channel coding over channels with cost constraints.**

***Index Terms*—Converse, finite blocklength regime, channels with cost constraints, joint source-channel coding, strong converse, dispersion, memoryless sources, memoryless channels, Shannon theory.**

I. Introduction

This paper is concerned with the maximum channel coding rate achievable at average error probability $\epsilon > 0$ where the cost of each codeword is constrained. The capacity-cost function $C(\beta)$ of a channel specifies the maximum achievable channel coding rate compatible with vanishing error probability and with codeword cost not exceeding β in the limit of large blocklengths.

A channel is said to satisfy the strong converse if $\epsilon \to 1$ as $n \to \infty$ for any code operating at a rate above the capacity. For memoryless channels without cost constraints, the strong converse was first shown by Wolfowitz: [1] treats the discrete memoryless channel (DMC), while [2] generalizes the result to memoryless channels whose input alphabet is finite while the output alphabet is the real line. Arimoto [3] showed a new converse bound stated in terms of Gallager's random coding exponent, which also leads to the strong converse for the DMC. Dueck and Körner [4] found the reliability function of DMC for rates above capacity, a result which implies a strong converse. Kemperman [5] showed that the strong converse holds for a DMC with feedback. A simple proof of strong converse for memoryless channels that does not invoke measure concentration inequalities was recently given in [6]. For a class of discrete channels with finite memory, the strong converse was shown by Wolfowitz [7] and independently by Feinstein [8], a result soon generalized to a more general class of stationary discrete channels with finite memory [9]. In a more general setting not requiring the assumption of stationarity or finite memory, Verdú and Han [10] showed a necessary and sufficient condition for a channel without cost constraints to satisfy the strong converse, while Han [11, Theorem 3.7.1] generalized that condition to the setting with cost constraints. In the special case of finite-input channels, that necessary and sufficient condition boils down to the capacity being equal to the limit of maximal normalized mutual informations. In turn, that condition is implied by the information stability of the channel [12], a condition which in general is not easy to verify. Using a novel notion of strong information stability, a general strong converse result was recently shown in [13, Theorem 3]. The strong converse for DMC with separable cost was shown by Csiszár and Körner [14, Theorem 6.11] and by Han [11, Theorem 3.7.2]. Regarding continuous channels, in the most basic case of the memoryless additive white Gaussian noise (AWGN) channel with the cost function being the power of the channel input block, $\mathsf{b}_n(x^n) = \frac{1}{n}|x^n|^2$, the strong converse was shown by Shannon [15] (contemporaneously with Wolfowitz's finite-alphabet strong converse). Yoshihara [16] proved the strong converse for the time-continuous channel with additive Gaussian noise having an arbitrary spectrum and also gave a simple proof of Shannon's strong converse result. Under the requirement that the power of each message converges stochastically to a given constant β, the strong converse for the AWGN channel with feedback was shown by Wolfowitz [17]. Note that in all those analyses of the power-constrained AWGN channel the cost constraint is meant on a per-codeword basis. In fact, the strong converse ceases to hold if the cost constraint is averaged over the codebook [18, Section 4.3.3].

Channel dispersion quantifies the backoff from capacity, unescapable at finite blocklengths due to the random nature of the channel coming into play, as opposed to the asymptotic representation of the channel as a deterministic bit pipe of a given capacity. More specifically, for coding over the DMC, the maximum achievable code rate at blocklength n compatible with error probability ϵ is approximated by $C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon)$ [19], [20] where C is the channel capacity, V is the channel dispersion, and $Q^{-1}(\cdot)$ is the inverse of the Gaussian complementary cdf. Polyanskiy et al. [20] found the dispersion of the DMC without cost constraints as well as that of the AWGN channel with a power constraint. In parallel,

This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1016625 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370.

Hayashi [21, Theorem 3] gave the dispersion of the DMC with and without cost constraints (with the loose estimate of $o\left(\sqrt{n}\right)$ for the third order term). For constant composition codes over the DMC, Polyanskiy [18, Sec. 3.4.6] showed the dispersion of constant composition codes over the DMC, while Moulin [22] refined the third-order term in the expansion of the maximum achievable code rate, under regularity conditions. Wang et al. [23] gave a second-order analysis of joint source-channel coding over finite alphabets based on constant composition codebooks.

In this paper, we demonstrate that the nonasymptotic fundamental limit for coding over channels with cost constraints is closely approximated in terms of the cdf of a random variable we refer to as the b-tilted information density, which parallels the notion of d-tilted information for lossy compression [24]. We show a simple non-asymptotic converse bound for general channels with input cost constraints in terms of b-tilted information density. Not only does this bound lead to a general strong converse result, but it is also tight enough to find the channel dispersion-cost function and the third order term equal to $\frac{1}{2}\log n$ when coupled with the corresponding achievability bound. More specifically, we show that for the DMC, $\log M^{\star}(n,\epsilon,\beta)$, the logarithm of the maximum achievable code size at blocklength n, error probability ϵ and cost β, is given by, under mild regularity assumptions

$$\log M^{\star}(n,\epsilon,\beta) = nC(\beta) - \sqrt{nV(\beta)}Q^{-1}\left(\epsilon\right) + \frac{1}{2}\log n + O\left(1\right) \tag{1}$$

where $V(\beta)$ is the dispersion-cost function, thereby refining Hayashi's result [21] and providing a matching converse to the result of Moulin [22]. We observe that the capacity-cost and the dispersion-cost functions are given by the mean and the variance of the b-tilted information density. This novel interpretation juxtaposes nicely with the corresponding results in [24] (d-tilted information in rate-distortion theory). Furthermore, we generalize (1) to lossy joint source-channel coding of general memoryless sources over channels with cost.

Section II introduces the b-tilted information density. Section III states the new non-asymptotic converse bound which holds for a general channel with cost constraints, without making any assumptions on the channel (e.g. alphabets, stationarity, memorylessness). An asymptotic analysis of the converse and achievability bounds, including the proof of the strong converse and the expression for the channel dispersion-cost function, is presented in Section IV in the context of memoryless channels. Section V generalizes the results in Sections III and IV to the lossy joint source-channel coding setup.

II. b-TILTED INFORMATION DENSITY

In this section, we introduce the concept of b-tilted information density and several relevant properties in a general single-shot approach.

Fix the transition probability kernel $P_{Y|X}\colon \mathcal{X}\to\mathcal{Y}$ and the cost function $\mathsf{b}\colon \mathcal{X}\mapsto[0,\infty]$. In the application of this single-shot approach in Section IV, $\mathcal{X}$, $\mathcal{Y}$, $P_{Y|X}$ and b will become $\mathcal{A}^n$, $\mathcal{B}^n$, $P_{Y^n|X^n}$ and b_n, respectively. Denote

$$\mathbb{C}(\beta) = \sup_{\substack{P_X:\\ \mathbb{E}[\mathsf{b}(X)]\le\beta}} I(X;Y), \tag{2}$$

$$\lambda^{\star} = \mathbb{C}'(\beta). \tag{3}$$

Since $\mathbb{C}(\beta)$ is non-decreasing concave function of β [14, Theorem 6.11], $\lambda^{\star}\ge 0$. For random variables Y and $\bar{Y}$ defined on the same space, denote

$$\imath_{Y\|\bar{Y}}(y) = \log\frac{dP_Y}{dP_{\bar{Y}}}(y). \tag{4}$$

If Y is distributed according to $P_{Y|X=x}$, we abbreviate the notation as

$$\imath_{X;\bar{Y}}(x;y) = \log\frac{dP_{Y|X=x}}{dP_{\bar{Y}}}(y). \tag{5}$$

in lieu of $\imath_{Y|X=x\|\bar{Y}}(y)$. The information density $\imath_{X;Y}(x;y)$ between realizations of two random variables with joint distribution $P_XP_{Y|X}$ follows by particularizing (5) to $\{P_{Y|X},P_Y\}$, where $P_X\to P_{Y|X}\to P_Y$[1]. In general, however, the function in (5) does not require $P_{\bar{Y}}$ to be induced by any input distribution.

Further, define the function

$$\jmath_{X;\bar{Y}}(x;y,\beta) = \imath_{X;\bar{Y}}(x;y) - \lambda^{\star}\left(\mathsf{b}(x)-\beta\right). \tag{6}$$

The special case of (6) with $P_{\bar{Y}}=P_{Y^\star}$, where $P_{Y^\star}$ is the unique output distribution that achieves the supremum in (2) [25], defines b-tilted information density:

Definition 1 (b-tilted information density)**.** *The* b*-tilted information density between $x\in\mathcal{X}$ and $y\in\mathcal{Y}$ is $\jmath_{X;Y^\star}(x;y,\beta)$.*

Since $P_{Y^\star}$ is unique even if there are several (or none) input distributions $P_{X^\star}$ that achieve the supremum in (2), there is no ambiguity in Definition 1. If there are no cost constraints (i.e. $\mathsf{b}(x)=0\ \forall x\in\mathcal{X}$), then $\mathbb{C}'(\beta)=0$ regardless of β, and

$$\jmath_{X;\bar{Y}}(x;y,\beta) = \imath_{X;\bar{Y}}(x;y). \tag{7}$$

The counterpart of the b-tilted information density in rate-distortion theory is the d-tilted information [24].

Example 1. For n uses of a memoryless AWGN channel with unit noise power and maximal power not exceeding nP, $\mathbb{C}(P)=\frac{n}{2}\log(1+P)$, and the output distribution that achieves (2) is $Y^{n\star}\sim\mathcal{N}\left(0,(1+P)\,\mathbf{I}\right)$. Therefore

$$\jmath_{X^n;Y^{n\star}}(x^n;y^n,P) = \frac{n}{2}\log\left(1+P\right) - \frac{\log e}{2}\left|y^n-x^n\right|^2 + \frac{\log e}{2(1+P)}\left(\left|y^n\right|^2-\left|x^n\right|^2+nP\right), \tag{8}$$

where the Euclidean norm is denoted by $\left|x^n\right|^2=\sum_{i=1}^n x_i^2$. It is easy to check that under $P_{Y^n|X^n=x^n}$, the distribution of

[1]We write $P_X\to P_{Y|X}\to P_Y$ to indicate that P_Y is the marginal of $P_XP_{Y|X}$, i.e. $P_Y(y)=\int_{\mathcal{X}}dP_{Y|X}(y|x)dP_X(x)$.

$\jmath_{X^n;Y^{n\star}}(x^n;Y^n,P)$ is the same as that of (by '$\sim$' we mean equality in distribution)

$$\jmath_{X^n;Y^{n\star}}(x^n;Y^n,P) \sim \frac{n}{2}\log(1+P) - \frac{P\log e}{2(1+P)}\left[W^n_{\frac{|x^n|^2}{P^2}} - n - \frac{|x^n|^2}{P^2}\right], \quad (9)$$

where W^ℓ_λ denotes a non central chi-square distributed random variable with ℓ degrees of freedom and non-centrality parameter λ. The mean of (9) is $\frac{n}{2}\log(1+P)$, in accordance with (16), while its variance is $\frac{1}{2}\frac{\left(nP^2+2|x^n|^2\right)}{(1+P)^2}\log^2 e$ which becomes $nV(P)$ (found in [20] and displayed in (46)) after averaging with respect to $X^{n\star}$ distributed according to $P_{X^{n\star}} \sim \mathcal{N}(0, P\mathbf{I})$.

Denote [2]

$$\beta_{\min} = \inf_{x\in\mathcal{X}} \mathsf{b}(x), \quad (10)$$

$$\beta_{\max} = \sup\{\beta \geq 0 \colon \mathbb{C}(\beta) < \mathbb{C}(\infty)\}. \quad (11)$$

Theorem 1 below highlights the importance of b-tilted information density in the optimization problem (2). Of key significance in the asymptotic analysis in Section IV, Theorem 1 gives a nontrivial generalization of the well-known properties of information density to the setting with cost constraints.

Theorem 1. *Fix $\beta_{\min} < \beta < \beta_{\max}$. Assume that $P_{X^\star}$ achieving (2) is such that the constraint is achieved with equality:*

$$\mathbb{E}[\mathsf{b}(X^\star)] = \beta. \quad (12)$$

Then, the following equalities hold.

$$\mathbb{C}(\beta) = \sup_{P_X} \mathbb{E}[\jmath_{X;Y}(X;Y,\beta)] \quad (13)$$

$$= \sup_{P_X} \mathbb{E}[\jmath_{X;Y^\star}(X;Y,\beta)] \quad (14)$$

$$= \mathbb{E}[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)] \quad (15)$$

$$= \mathbb{E}[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)|X^\star], \quad (16)$$

where (16) holds $P_{X^\star}$-a.s., and $P_X \to P_{Y|X} \to P_Y$, $P_{X^\star} \to P_{Y|X} \to P_{Y^\star}$.

Proof. Appendix A. □

Throughout the paper, we assume that the assumptions of Theorem 1 hold.

For channels without cost, the inequality

$$D(P_{Y|X=x}\|P_{Y^\star}) \leq C \ \forall x \in \mathcal{X} \quad (17)$$

is key to proving strong converses. Theorem 1 generalizes this result to channels with cost, showing that

$$\mathbb{E}[\jmath_{X;Y^\star}(x;Y,\beta)|X=x] \leq C(\beta) \ \forall x \in \mathcal{X}. \quad (18)$$

Note that (18) is crucial for showing both the strong converse and the refined asymptotic analysis.

[2] We allow $\beta_{\max} = +\infty$.

Remark 1. The general strong converse result in [13, Theorem 3] includes channels with cost using the concept of 'quasi-caod', which is defined as any output distribution P_{Y^n} such that

$$D(P_{Y^n|X^n=x^n}\|P_{Y^n}) \leq I^\star_n + o(I^\star_n) \ \forall x^n \in \mathcal{A}^n \colon \mathsf{b}_n(x^n) \leq \beta, \quad (19)$$

where $\mathcal{A}$ is the single-letter channel input alphabet, and $I^\star_n = \max_{P_{X^n}\colon \mathsf{b}(X^n)\leq b \text{ a.s.}} I(X^n;Y^n)$. Since $C(\beta) = \lim_{n\to\infty}\frac{1}{n}I^\star_n$, (18) implies that $P_{Y^\star}\times\ldots\times P_{Y^\star}$ is always a quasi-caod.

Corollary 2. *For all $P_X \ll P_{X^\star}$*

$$\mathrm{Var}[\jmath_{X;Y^\star}(X;Y,\beta)] = \mathbb{E}[\mathrm{Var}[\jmath_{X;Y^\star}(X;Y,\beta)|X]] \quad (20)$$

$$= \mathbb{E}[\mathrm{Var}[\imath_{X;Y^\star}(X;Y)|X]]. \quad (21)$$

Proof. Appendix B. □

III. Nonasymptotic bounds

Converse and achievability bounds give necessary and sufficient conditions, respectively, on (M,ϵ,β) in order for a code to exist with M codewords and average error probability not exceeding ϵ and cost not exceeding β. Such codes (allowing stochastic encoders and decoders) are rigorously defined next.

Definition 2 ((M,ϵ,β) code)**.** *An (M,ϵ,β) code for $\{P_{Y|X},\mathsf{b}\}$ is a pair of random transformations $P_{X|S}$ (encoder) and $P_{Z|Y}$ (decoder) such that $\mathbb{P}[S\neq Z]\leq\epsilon$, where $S-X-Y-Z$, the probability is evaluated with S equiprobable on an alphabet of cardinality M, and the codewords satisfy the maximal cost constraint (a.s.)*

$$\mathsf{b}(X) \leq \beta. \quad (22)$$

The non-asymptotic quantity of principal interest is $M^\star(\epsilon,\beta)$, the maximum code size achievable at error probability ϵ and cost β.

Theorem 3 (Converse)**.** *The existence of an (M,ϵ,β) code for $\{P_{Y|X},\mathsf{b}\}$ requires that*

$$\epsilon \geq \max_{\gamma>0}\Bigg\{\sup_{\bar{Y}}\inf_{x\colon \mathsf{b}(x)\leq\beta}\mathbb{P}\left[\imath_{X;\bar{Y}}(x;Y)\leq\log M-\gamma|X=x\right] - \exp(-\gamma)\Bigg\} \quad (23)$$

$$\geq \max_{\gamma>0}\Bigg\{\sup_{\bar{Y}}\inf_{x\in\mathcal{X}}\mathbb{P}\left[\jmath_{X;\bar{Y}}(x;Y,\beta)\leq\log M-\gamma|X=x\right] - \exp(-\gamma)\Bigg\}. \quad (24)$$

Proof. The bound in (23) is due to Wolfowitz [26]. The bound in (24) simply weakens (23) using $\mathsf{b}(x)\leq\beta$. □

By restricting the channel input space appropriately, converse bounds for channels with cost constraints can be obtained from the converse bounds in [20], [27]. Their analysis

becomes tractable by the introduction of b-tilted information density in (24) and an application of (18).

Achievability bounds for channels with cost constraints can be obtained from the random coding bounds in [20], [27] by restricting the distribution from which the codewords are drawn to satisfy $\mathsf{b}(X) \le \beta$ a.s. In particular, for the DMC, we may choose P_{X^n} to be equiprobable on the set of codewords of the type closest (among types satisfying the cost constraint) to the input distribution $P_{\mathsf{X}^\star}$ that achieves the capacity-cost function. As shown in [21], such constant composition codes achieve the dispersion of channel coding under input cost constraints. Unfortunately, the computation of such bounds may become challenging in high dimension, particularly with continuous alphabets.

IV. Asymptotic analysis

To introduce the blocklength into the non-asymptotic converse of Section III, we consider (M, ϵ, β) codes for $\{P_{Y^n|X^n}, \mathsf{b}_n\}$, where $P_{Y^n|X^n}\colon \mathcal{A}^n \mapsto \mathcal{B}^n$ and $\mathsf{b}_n\colon \mathcal{A}^n \mapsto [0,\infty]$. We call such codes (n, M, ϵ, β) codes, and denote the corresponding non-asymptotically achievable maximum code size by $M^\star(n, \epsilon, \beta)$.

A. Assumptions

The following basic assumptions hold throughout Section IV.

(i) The channel is stationary and memoryless, $P_{Y^n|X^n} = P_{\mathsf{Y}|\mathsf{X}} \times \ldots \times P_{\mathsf{Y}|\mathsf{X}}$.
(ii) The cost function is separable, $\mathsf{b}_n(x^n) = \frac{1}{n}\sum_{i=1}^n \mathsf{b}(x_i)$, where $\mathsf{b}\colon \mathcal{A} \mapsto [0,\infty]$.
(iii) Each codeword is constrained to satisfy the maximal cost constraint, $\mathsf{b}_n(x^n) \le \beta$.
(iv) $\sup_{\mathsf{x}\in\mathcal{A}} \operatorname{Var}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{x};\mathsf{Y},\beta)|\mathsf{X}=x\right] = V_{\max} < \infty$.

Under these assumptions, the capacity-cost function is given by

$$C(\beta) = \sup_{P_{\mathsf{X}}\colon \mathbb{E}[\mathsf{b}(\mathsf{X})]\le\beta} I(\mathsf{X};\mathsf{Y}). \tag{25}$$

Observe that in view of assumptions (i) and (ii), as long as $P_{\bar{Y}^n}$ is a product distribution, $P_{\bar{Y}^n} = P_{\bar{\mathsf{Y}}} \times \ldots \times P_{\bar{\mathsf{Y}}}$,

$$\jmath_{X^n;\bar{Y}^n}(x^n;y^n,\beta) = \sum_{i=1}^n \jmath_{\mathsf{X};\bar{\mathsf{Y}}}(x_i;y_i,\beta). \tag{26}$$

B. Strong converse

Although the tools developed in Sections II and III are able to result in a strong converse for channels that exhibit ergodic behavior (see also Remark 1), for the sake of concreteness and length, we only deal here with the memoryless setup described in Section IV-A.

We show that if transmission occurs at a rate greater than the capacity-cost function, the error probability must converge to 1, regardless of the specifics of the code. Towards this end, we fix some $\alpha > 0$, we choose $\log M \ge nC(\beta) + 2n\alpha$, and we weaken the bound (24) in Theorem 3 by fixing $\gamma = n\alpha$ and $P_{\bar{Y}^n} = P_{\mathsf{Y}^\star} \times \ldots \times P_{\mathsf{Y}^\star}$, where $\mathsf{Y}^\star$ is the output distribution that achieves $C(\beta)$, to obtain

$$\epsilon \ge \inf_{x^n\in\mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^n \jmath_{\mathsf{X};\mathsf{Y}^\star}(x_i;Y_i,\beta) \le nC(\beta) + n\alpha\right] - \exp(-n\alpha) \tag{27}$$

$$\ge \inf_{x^n\in\mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^n \jmath_{\mathsf{X};\mathsf{Y}^\star}(x_i;Y_i,\beta) \le \sum_{i=1}^n c(x_i) + n\alpha\right] - \exp(-n\alpha), \tag{28}$$

where for notational convenience we have abbreviated

$$c(\mathsf{x}) = \mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{x};\mathsf{Y},\beta)|\mathsf{X}=\mathsf{x}\right], \tag{29}$$

and (28) employs (14).

To show that the right side of (28) converges to 1, we invoke the following law of large numbers for non-identically distributed random variables.

Lemma 1 (e.g. [28])**.** *Suppose that W_i are uncorrelated and $\sum_{i=1}^\infty \operatorname{Var}\left[\frac{W_i}{c_i}\right] < \infty$ for some strictly positive sequence (c_n) increasing to $+\infty$. Then,*

$$\frac{1}{c_n}\left(\sum_{i=1}^n W_i - \mathbb{E}\left[\sum_{i=1}^n W_i\right]\right) \to 0 \text{ in } L^2. \tag{30}$$

Let $W_i = \jmath_{\mathsf{X};\mathsf{Y}^\star}(x_i;Y_i,\beta)$ and $c_i = i$. Since (recall (iv))

$$\sum_{i=1}^\infty \operatorname{Var}\left[\frac{1}{i}\jmath_{\mathsf{X};\mathsf{Y}^\star}(x_i;Y_i,\beta)|X_i=x_i\right] \le V_{\max}\sum_{i=1}^\infty \frac{1}{i^2} \tag{31}$$

$$< \infty, \tag{32}$$

by virtue of Lemma 1, the right side of (28) converges to 1, so any channel satisfying (i)–(iv) also satisfies the strong converse.

As noted in [18, Theorem 77] in the context of the AWGN channel, the strong converse does not hold if the cost constraint is averaged over the codebook, i.e. if, in lieu of (22), the cost requirement is

$$\frac{1}{M}\sum_{m=1}^M \mathbb{E}\left[\mathsf{b}(X)|S=m\right] \le \beta. \tag{33}$$

To see why the strong converse does not hold in general, fix a code of rate $C(\beta) < R < C(2\beta)$ none of whose codewords cost more than 2β and whose error probability satisfies $\epsilon_n \to 0$. Since $R < C(2\beta)$, such a code exists. Now, replace half of the codewords with the all-zero codeword (assuming $\mathsf{b}(\mathbf{0}) = 0$) while leaving the decision regions of the remaining codewords untouched. The average cost of the new code satisfies (33), its rate is greater than the capacity-cost function, $R > C(\beta)$, yet its average error probability does not exceed $\epsilon_n + \frac{1}{2} \to \frac{1}{2}$.

C. Dispersion

First, we give the operational definition of the dispersion-cost function of any channel.

Definition 3 (Dispersion-cost function)**.** *The channel dispersion-cost function, measured in squared information units per channel use, is defined by*

$$V(\beta)=\lim_{\epsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\frac{\left(nC(\beta)-\log M^{\star}(n,\epsilon,\beta)\right)^2}{2\log_e\frac{1}{\epsilon}}. \quad (34)$$

An explicit expression for the dispersion-cost function of a discrete memoryless channel is given in the next result.

Theorem 4. *In addition to assumptions (i)–(iv), assume that the capacity-achieving input distribution $P_{\mathsf{X}^\star}$ is unique and that the channel has finite input and output alphabets.*

$$\log M^{\star}(n,\epsilon,\beta)=nC(\beta)-\sqrt{nV(\beta)}Q^{-1}(\epsilon)+\theta(n), \quad (35)$$

$$C(\beta)=\mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X}^\star;\mathsf{Y}^\star,\beta)\right], \quad (36)$$

$$V(\beta)=\mathrm{Var}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X}^\star;\mathsf{Y}^\star,\beta)\right], \quad (37)$$

where the remainder term $\theta(n)$ satisfies:

a) If $V(\beta)>0$,

$$-\frac{1}{2}\left(|\mathrm{supp}\left(P_{\mathsf{X}^\star}\right)|-1\right)\log n+O(1)\le\theta(n) \quad (38)$$

$$\le\frac{1}{2}\log n+O(1). \quad (39)$$

b) If $V(\beta)=0$, (38) holds, and (39) is replaced by

$$\theta(n)\le O\left(n^{\frac{1}{3}}\right). \quad (40)$$

Proof. Converse. Full details are given in Appendix D. The main steps of the refined asymptotic analysis of the bound in Theorem 3 are as follows. First, building on the ideas of [29], [30], we weaken the bound in (24) by a careful choice of a non-product auxiliary distribution $P_{\bar{Y}^n}$. Second, using Theorem 1 and the technical tools developed in Appendix C, we show that the infimum in the right side of (24) is lower bounded by ϵ for the choice of M in (35).

Achievability. Full details are given in Appendix E, which provides an asymptotic analysis of the Dependence Testing bound of [20] in which the random codewords are of type closest to $P_{\mathsf{X}^\star}$, rather than drawn from the product distribution $P_{\mathsf{X}}\times\ldots\times P_{\mathsf{X}}$, as in achievability proofs for channel coding without cost constraints. We use Corollary 2 to establish that such constant composition codes achieve the dispersion-cost function. □

Remark 2. According to a recent result of Moulin [22], the achievability bound on the remainder term in (38) can be tightened to match the converse bound in (39), thereby establishing that

$$\theta(n)=\frac{1}{2}\log n+O(1), \quad (41)$$

provided that the following regularity assumptions hold:

- The random variable $\imath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X}^\star;\mathsf{Y}^\star)$ is of nonlattice type;
- $\mathrm{supp}(P_{\mathsf{X}^\star})=\mathcal{A}$;
- $\mathrm{Cov}\left[\imath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X}^\star;\mathsf{Y}^\star),\imath_{\mathsf{X};\mathsf{Y}^\star}(\bar{\mathsf{X}}^\star;\mathsf{Y}^\star)\right]<\mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X}^\star;\mathsf{Y}^\star)\right]$ where $P_{\bar{\mathsf{X}}^\star\mathsf{X}^\star\mathsf{Y}^\star}(\bar{\mathsf{x}},\mathsf{x},\mathsf{y})=\frac{1}{P_{\mathsf{Y}^\star}(\mathsf{y})}P_{\mathsf{X}^\star}(\bar{\mathsf{x}})P_{\mathsf{Y}|\mathsf{X}}(\mathsf{y}|\bar{\mathsf{x}})P_{\mathsf{Y}|\mathsf{X}}(\mathsf{y}|\mathsf{x})P_{\mathsf{X}^\star}(\mathsf{x})$.

Remark 3. As we show in Appendix F, Theorem 4 applies to channels with abstract alphabets provided that in addition to (i)–(ii), they meet the following criteria:

(a) The cost function $\mathsf{b}\colon\mathcal{A}\to[0,\infty]$ is such that for all $\gamma\in[\beta,\infty)$, $\mathsf{b}^{-1}(\gamma)$ is nonempty. In particular, this condition is satisfied if the channel input alphabet $\mathcal{A}$ is a metric space, and b is continuous and unbounded with $\mathsf{b}(0)=0$.

(b) The distribution of $\imath_{X^n;Y^{n\star}}(x^n;Y^n)$, where $P_{Y^{n\star}}=P_{\mathsf{Y}^\star}\times\ldots\times P_{\mathsf{Y}^\star}$ does not depend on the choice of $x^n\in\mathcal{F}_n$, where $\mathcal{F}_n=\{x^n\in\mathcal{A}^n\colon\mathsf{b}_n(x^n)=\beta\}$.

(c) For all x in the projection of $\mathcal{F}_n$ onto $\mathcal{A}$, i.e. for all x such that $(\mathsf{x},x_2,\ldots,x_n)\in\mathcal{F}_n$ for some $x_2,\ldots,x_n$,

$$\mathbb{E}\left[\left|\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X};\mathsf{Y},\beta)-C(\beta)\right|^3|\mathsf{X}=\mathsf{x}\right]<\infty. \quad (42)$$

(d)[3] There exists a distribution P_{X^n} supported on $\mathcal{F}_n$ such that $\imath_{Y^n\|Y^{n\star}}(Y^n)$, where $P_{X^n}\to P_{Y^n|X^n}\to P_{Y^n}$, is almost surely bounded by $f_n=o(\sqrt{n})$ from above.

Then, (35) holds identifying (43)–(45) or all $\mathsf{x}\in\mathcal{A}$ s.t. $\mathsf{b}(\mathsf{x})=\beta$:

$$C(\beta)=D(P_{\mathsf{Y}|\mathsf{X}=\mathsf{x}}\|P_{\mathsf{Y}^\star}), \quad (43)$$

$$V(\beta)=\mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{x};\mathsf{Y})|\mathsf{X}=\mathsf{x}\right], \quad (44)$$

$$-f_n+O(1)\le\theta(n)\le\frac{1}{2}\log n+O(1), \quad (45)$$

where $f_n=o(\sqrt{n})$ is specified in (d).

Remark 4. Theorem 4 with the remainder in (41) [31] also holds for the AWGN channel with maximal signal-to-noise ratio P, offering a novel interpretation of the dispersion of the Gaussian channel [20]

$$V(P)=\frac{1}{2}\left(1-\frac{1}{(1+P)^2}\right)\log^2 e \quad (46)$$

as the variance of the b-tilted information density. We note that the AWGN channel satisfies the conditions of Remark 3 with P_{X^n} uniform on the power sphere and $f_n=O(1)$ [20].

Remark 5. As we show in Appendix G, a stationary memoryless channel with $\mathsf{b}(\mathsf{x})=\mathsf{x}$ which takes a nonnegative input and adds an exponential noise of unit mean to it [32], satisfies the conditions of Remark 3 with $f_n=O(1)$, and

$$\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{x};\mathsf{y},\beta)=\log(1+\beta)+\frac{\beta}{1+\beta}(\mathsf{x}-\mathsf{y}+1)\log e, \quad (47)$$

$$C(\beta)=\log(1+\beta), \quad (48)$$

$$V(\beta)=\frac{\beta^2}{(1+\beta)^2}\log^2 e. \quad (49)$$

Remark 6. As should be clear from the proof of Theorem 4, if the capacity-achieving distribution is not unique, then

$$V(\beta)=\begin{cases}\min\mathrm{Var}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X}^\star;\mathsf{Y}^\star,\beta)\right] & 0<\epsilon\le\frac{1}{2}\\ \max\mathrm{Var}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X}^\star;\mathsf{Y}^\star,\beta)\right] & \frac{1}{2}<\epsilon<1\end{cases} \quad (50)$$

where the optimization is performed over all $P_{\mathsf{X}^\star}$ that achieve $C(\beta)$. This parallels the dispersion result for channels without cost [20].

[3] For the converse result, assumptions (a)–(c) suffice.

V. Joint source-channel coding

In this section we state the counterparts of Theorems 3 and 4 in the lossy joint source-channel coding setting. Proofs of the results in this section are obtained by fusing the proofs in Sections III and IV and those in [27].

In the joint source-channel coding setup the source is no longer equiprobable on an alphabet of cardinality M, as in Definition 2, rather it is arbitrarily distributed on an abstract alphabet $\mathcal{M}$. Further, instead of reproducing the transmitted S under a probability of error criterion, we might be interested in approximating S within a certain distortion, so that a decoding failure occurs if the distortion between the source and its reproduction exceeds a given distortion level d, i.e. if $\mathsf{d}(S,Z) > d$, where $Z \in \widehat{\mathcal{M}}$ is the representation of S, $\widehat{\mathcal{M}}$ is a reproduction alphabet, and $\mathsf{d}\colon \mathcal{M} \times \widehat{\mathcal{M}} \mapsto \mathbb{R}_+$ is the distortion measure. A (d,ϵ,β) code is a code for a fixed source-channel pair such that the probability of exceeding distortion d is no larger than ϵ and no channel codeword costs more than β. A (d,ϵ,β) code in a block coding setting, when a source block of length k is mapped to a channel block of length n, is called a (k,n,d,ϵ,β) code. The counterpart of the b-tilted information density in lossy compression is the d-tilted information, $\jmath_S(s,d)$, which can be computed using the equality

$$\jmath_S(s,d) = \imath_{Z^\star;S}(z;s) + \lambda_S \mathsf{d}(s,z) - \lambda_S d, \tag{51}$$

where $Z^\star$ is the random variable that achieves the infimum on the right side of

$$\mathbb{R}_S(d) \triangleq \min_{\substack{P_{Z|S}:\\ \mathbb{E}[\mathsf{d}(S,Z)]\le d}} I(S;Z), \tag{52}$$

$\lambda_S = -\mathbb{R}_S'(d) > 0$, and equality in (51) holds for $P_{Z^\star}$-a.e. z [24]. In a certain sense, the d-tilted information quantifies the number of bits required to reproduce the source outcome $s \in \mathcal{M}$ within distortion d. For rigorous definitions and further details we refer the reader to [27].

Theorem 5 (Converse)**.** *The existence of a (d,ϵ,β) code for S and $P_{Y|X}$ requires that*

$$\epsilon \ge \inf_{P_{X|S}} \max_{\gamma>0} \left\{ \sup_{\bar{Y}} \mathbb{P}\left[\jmath_S(S,d) - \jmath_{X;\bar{Y}}(X;Y,\beta) \ge \gamma\right] - \exp(-\gamma) \right\} \tag{53}$$

$$\ge \max_{\gamma>0} \left\{ \sup_{\bar{Y}} \mathbb{E}\left[\inf_{x\in\mathcal{X}} \mathbb{P}\left[\jmath_S(S,d) - \jmath_{X;\bar{Y}}(x;Y,\beta) \ge \gamma \mid S\right]\right] - \exp(-\gamma) \right\}, \tag{54}$$

where the probabilities in (53) and (54) are with respect to $P_S P_{X|S} P_{Y|X}$ and $P_{Y|X=x}$, respectively.

Proof. The bound is obtained by weakening [27, Theorem 1] (23) using $\mathsf{b}(x) \le \beta$. □

Under the usual memorylessness assumptions, applying Theorem 1 to the bound in (54), it is easy to show that the strong converse holds for lossy joint source-channel coding over channels with input cost constraints. A more refined analysis leads to the following result.

Theorem 6 (Gaussian approximation)**.** *Assume the channel has finite input and output alphabets. For stationary memoryless sources satisfying the regularity assumptions (i)–(iv) of [27] and channels satisfying assumptions (ii)–(iv) of Section IV-A, the parameters of the optimal (k,n,d,ϵ) code satisfy*

$$nC(\beta) - kR(d) = \sqrt{nV(\beta) + k\mathcal{V}(d)}\, Q^{-1}(\epsilon) + \theta(n), \tag{55}$$

where $\mathcal{V}(d) = \mathrm{Var}[\jmath_{\mathsf{S}}(\mathsf{S},d)]$, $V(\beta)$ is given in (37), and the remainder $\theta(n)$ satisfies, if $V(\beta) > 0$,

$$-\frac{1}{2}\log n + O\left(\sqrt{\log n}\right) \le \theta(n) \tag{56}$$

$$\le \bar{\theta}(n) + \left(\frac{1}{2}|\mathrm{supp}(P_{\mathsf{X}^\star})| - 1\right)\log n, \tag{57}$$

where $\bar{\theta}(n) = O(\log n)$ denotes the upper bound on the remainder term given in [27, Theorem 10]. If $V(\beta) = \mathcal{V}(d) = 0$, the upper bound on $\theta(n)$ stays the same, and the lower one becomes $O\left(n^{\frac{1}{3}}\right)$.

Proof outline. The achievability part is proven joining the asymptotic analyses of [27, Theorem 8] and of Theorem 9, shown in Appendix E. For the converse part, $P_{\bar{Y}}$ is chosen as in (146), and similar to the proof of the converse part of [27, Theorem 10], a typical set of source outcomes is identified, and it is shown using Theorem 7.2 that for every source outcome in that set, the inner infimum in (54) is approximately achieved by the capacity-achieving channel input type. □

VI. Conclusion

We introduced the concept of b-tilted information density (Definition 1), a random variable whose distribution governs the analysis of optimal channel coding under input cost constraints. The properties of b-tilted information density listed in Theorem 1 play a key role in the asymptotic analysis of the converse bound in Theorem 3 in Section IV, which does not only lead to the strong converse and the dispersion-cost function when coupled with the corresponding achievability bound, but it also proves that the third order term in the asymptotic expansion (1) is upper bounded (in the most common case of $V(\beta) > 0$) by $\frac{1}{2}\log n + O(1)$. In addition, we showed in Section V that the results of [27] generalize to coding over channels with cost constraints and also tightened the estimate of the third order term in [27]. As propounded in [29], [30], the gateway to the refined analysis of the third order term is an apt choice of a non-product distribution $P_{\bar{Y}^n}$ in the bounds in Theorems 3 and 5.

VII. Acknowledgement

We thank the referees for their unusually thorough reviews, which are reflected in the final version.

APPENDIX A
PROOF OF THEOREM 1

We note first two auxiliary results.

Lemma 2 ([33]). *Let $0 \leq \alpha \leq 1$, and let $P \ll Q$ be distributions on the same probability space. Then,*

$$\lim_{\alpha \to 0} \frac{1}{\alpha} D(\alpha P + (1-\alpha)Q \| Q) = 0. \tag{58}$$

Lemma 3 (Donsker-Varadhan [34]). *Let $g\colon \mathcal{X} \mapsto [-\infty, +\infty]$ and let $\bar{X}$ be a random variable on $\mathcal{X}$ such that $\mathbb{E}\left[\exp\left(g(\bar{X})\right)\right] < \infty$. Then,*

$$\mathbb{E}\left[g(X)\right] - D(X\|\bar{X}) \leq \log \mathbb{E}\left[\exp\left(g(\bar{X})\right)\right] \tag{59}$$

with equality if and only if X has distribution $P_{X^\star}$ such that

$$\imath_{X^\star \| \bar{X}}(x) = g(x) - \log \mathbb{E}\left[\exp\left(g(\bar{X})\right)\right]. \tag{60}$$

Proof. If the left side of (59) is not $-\infty$, we can write

$$\mathbb{E}\left[g(X)\right] - D(X\|\bar{X}) = \mathbb{E}\left[g(X) - \imath_{X\|X^\star}(X) - \imath_{X^\star\|\bar{X}}(X)\right] \tag{61}$$

$$= \log \mathbb{E}\left[\exp\left(g(\bar{X})\right)\right] - D(X\|X^\star), \tag{62}$$

which is maximized by letting $P_X = P_{X^\star}$. □

We proceed to prove Theorem 1 by generalizing [35, Theorem 6.1]. Equality (13) is a standard result in convex optimization. By the assumption, the supremum in the right side of (13) is attained by $P_{X^\star}$, therefore $\mathbb{C}(\alpha)$ is equal to the right side of (15).

To show (14), fix $0 \leq \alpha \leq 1$. Denote

$$P_{\bar{X}} \to P_{Y|X} \to P_{\bar{Y}}, \tag{63}$$

$$P_{\hat{X}} = \alpha P_{\bar{X}} + (1-\alpha) P_{X^\star}, \tag{64}$$

$$P_{\hat{X}} \to P_{Y|X} \to P_{\hat{Y}} = \alpha P_{\bar{Y}} + (1-\alpha) P_{Y^\star}, \tag{65}$$

and write

$$\begin{aligned}
&\alpha\left(\mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)\right] - \mathbb{E}\left[\jmath_{X;Y^\star}(\bar{X};\bar{Y},\beta)\right]\right) \\
&\quad + D(\hat{Y}\|Y^\star) \\
&= \alpha D(P_{Y|X}\|P_{Y^\star}|P_{X^\star}) - \alpha D(P_{Y|X}\|P_{Y^\star}|P_{\bar{X}}) + D(\hat{Y}\|Y^\star) \\
&\quad + \lambda^\star \alpha \mathbb{E}\left[\mathsf{b}(\bar{X})\right] - \lambda^\star \alpha \mathbb{E}\left[\mathsf{b}(X^\star)\right] \qquad (66) \\
&= D(P_{Y|X}\|P_{Y^\star}|P_{X^\star}) - D(P_{Y|X}\|P_{Y^\star}|P_{\hat{X}}) + D(\hat{Y}\|Y^\star) \\
&\quad - \lambda^\star \mathbb{E}\left[\mathsf{b}(X^\star)\right] + \lambda^\star \mathbb{E}\left[\mathsf{b}(\hat{X})\right] \qquad (67) \\
&= D(P_{Y|X}\|P_{Y^\star}|P_{X^\star}) - D(P_{Y|X}\|P_{\hat{Y}}|P_{\hat{X}}) - \lambda^\star \mathbb{E}\left[\mathsf{b}(X^\star)\right] \\
&\quad + \lambda^\star \mathbb{E}\left[\mathsf{b}(\hat{X})\right] \qquad (68) \\
&= \mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)\right] - \mathbb{E}\left[\jmath_{X;\hat{Y}}(\hat{X};\hat{Y},\beta)\right] \qquad (69) \\
&\geq 0, \qquad (70)
\end{aligned}$$

where (70) holds because $X^\star$ achieves the supremum in the right side of (13). Assume for the moment that $P_{\bar{Y}} \ll P_{Y^\star}$. Lemma 2 implies that $D(\hat{Y}\|Y^\star) = o(\alpha)$. Thus, supposing that $\mathbb{E}\left[\jmath_{X;Y^\star}(\bar{X};\bar{Y},\beta)\right] > \mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)\right]$ would lead to a contradiction, since then the left side of (66) would be negative for a sufficiently small α.

To complete the proof of (14), it remains to show $P_{Y^\star}$ dominates all $P_{\bar{Y}}$ such that $P_{\bar{X}} \to P_{Y|X} \to P_{\bar{Y}}$. By contradiction, assume that $P_{\bar{X}}$ and $\mathcal{F} \subseteq \mathcal{Y}$ are such that $P_{\bar{Y}}(\mathcal{F}) > P_{Y^\star}(\mathcal{F}) = 0$, and define the mixture $P_{\hat{X}}$ as in (64). Note that

$$\begin{aligned}
D(P_{Y|X}\|P_{\hat{Y}}|P_{\bar{X}}) &\geq D(\bar{Y}\|\hat{Y}) && (71)\\
&\geq D(1\{\bar{Y}\in\mathcal{F}\}\|1\{\hat{Y}\in\mathcal{F}\}) && (72)\\
&\geq P_{\bar{Y}}(\mathcal{F}) \log \frac{P_{\bar{Y}}(\mathcal{F})}{P_{\hat{Y}}(\mathcal{F})} && (73)\\
&= P_{\bar{Y}}(\mathcal{F}) \log \frac{1}{\alpha}. && (74)
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}\left[\jmath_{X;\hat{Y}}(\hat{X};\hat{Y},\beta)\right] - \mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)\right] \\
&= \alpha\mathbb{E}\left[\jmath_{X;\hat{Y}}(\bar{X};\bar{Y},\beta)\right] + (1-\alpha)\mathbb{E}\left[\jmath_{X;\hat{Y}}(X^\star;Y^\star,\beta)\right] \\
&\quad - \mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)\right] \qquad (75)\\
&\geq \alpha\mathbb{E}\left[\jmath_{X;\hat{Y}}(\bar{X};\bar{Y},\beta)\right] - \alpha\mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)\right] \qquad (76)\\
&\geq \alpha\Big(P_{\bar{Y}}(\mathcal{F})\log\frac{1}{\alpha} - \lambda^\star\mathbb{E}\left[\mathsf{b}(\bar{X})\right] + \lambda^\star\beta \\
&\quad - \mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)\right]\Big) \qquad (77)\\
&> 0, \qquad (78)
\end{aligned}$$

where (76) is due to $D(Y^\star\|\hat{Y}) \geq 0$, (77) invokes (74), and (78) holds for sufficiently small α, thereby contradicting (13). We conclude that indeed $P_{\bar{Y}} \ll P_{Y^\star}$.

To show (16), define the following function of a pair of probability distributions on $\mathcal{X}$:

$$\begin{aligned}
F(P_X, P_{\bar{X}}) &= \mathbb{E}\left[\jmath_{X;\bar{Y}}(X;Y,\beta)\right] - D(X\|\bar{X}) && (79)\\
&= \mathbb{E}\left[\jmath_{X;Y}(X;Y,\beta)\right] - D(X\|\bar{X}) + D(Y\|\bar{Y}) && (80)\\
&\leq \mathbb{E}\left[\jmath_{X;Y}(X;Y,\beta)\right], && (81)
\end{aligned}$$

where (81) holds by the data processing inequality for relative entropy. Since equality in (81) is achieved by $P_X = P_{\bar{X}}$, $\mathbb{C}(\beta)$ can be expressed as the double maximization

$$\mathbb{C}(\beta) = \max_{P_{\bar{X}}} \max_{P_X} F(P_X, P_{\bar{X}}). \tag{82}$$

To solve the inner maximization in (82), we invoke Lemma 3 with

$$g(x) = \mathbb{E}\left[\jmath_{X;\bar{Y}}(x;Y,\beta)|X=x\right] \tag{83}$$

to conclude that

$$\max_{P_X} F(P_X, P_{\bar{X}}) = \log \mathbb{E}\left[\exp\left(\mathbb{E}\left[\jmath_{X;\bar{Y}}(\bar{X};\bar{Y},\beta)|\bar{X}\right]\right)\right], \tag{84}$$

which in the special case $P_{\bar{X}} = P_{X^\star}$ yields, using representation (82),

$$\begin{aligned}
\mathbb{C}(\beta) &\geq \log \mathbb{E}\left[\exp\left(\mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y,\beta)|X^\star\right]\right)\right] && (85)\\
&\geq \mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y^\star,\beta)\right] && (86)\\
&= \mathbb{C}(\beta) && (87)
\end{aligned}$$

where (86) applies Jensen's inequality to the strictly convex function $\exp(\cdot)$, and (87) holds by the assumption. We conclude that, in fact, (86) holds with equality, which implies that $\mathbb{E}\left[\jmath_{X;Y^\star}(X^\star;Y,\beta)|X^\star\right]$ is almost surely constant, thereby showing (16).

Appendix B
Proof of Corollary 2

To show (21), we invoke (6) to write, for any $x \in \mathcal{X}$,

$$\begin{aligned}
&\mathrm{Var}\left[\jmath_{X;Y^\star}(X;Y,\beta)|X=x\right] \\
&= \mathrm{Var}\left[\imath_{X;Y^\star}(X;Y) - \lambda^\star\left(\mathsf{b}(X)-\beta\right)|X=x\right] &(88)\\
&= \mathrm{Var}\left[\imath_{X;Y^\star}(X;Y)|X=x\right]. &(89)
\end{aligned}$$

To show (20), we invoke (16) to write

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Var}\left[\jmath_{X;Y^\star}(X;Y,\beta)|X\right]\right] \\
&= \mathbb{E}\left[\left(\jmath_{X;Y^\star}(X;Y,\beta)\right)^2\right] \\
&- \mathbb{E}\left[\left(\mathbb{E}\left[\jmath_{X;Y^\star}(X;Y,\beta)|X\right]\right)^2\right] &(90)\\
&= \mathbb{E}\left[\left(\jmath_{X;Y^\star}(X;Y,\beta)\right)^2\right] - \mathbb{C}^2(\beta) &(91)\\
&= \mathrm{Var}\left[\jmath_{X;Y^\star}(X;Y,\beta)\right]. &(92)
\end{aligned}$$

Appendix C
Auxiliary result on the minimization of the cdf of a sum of independent random variables

Let $\mathcal{D}$ is a metric space with metric $d : \mathcal{D}^2 \mapsto \mathbb{R}^+$. Let $W_i(z), i = 1,\ldots,n$ be independent random variables parameterized by $z \in \mathcal{D}$. Denote

$$D_n(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[W_i(z)\right], \tag{93}$$

$$V_n(z) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left[W_i(z)\right], \tag{94}$$

$$T_n(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left|W_i(z)-\mathbb{E}\left[W_i(z)\right]\right|^3\right]. \tag{95}$$

Let ℓ_1, ℓ_2, ℓ_3, L_1, L_2, F_1, F_2, $V_{\min}$ and $T_{\max}$ be positive constants. We assume that there exist $z^\star \in \mathcal{D}$ and sequences $D_n^\star$, $V_n^\star$ such that for all $z \in \mathcal{D}$,

$$D_n^\star - D_n(z) \geq \ell_1 d^2(z,z^\star) - \frac{\ell_2}{\sqrt{n}} d(z,z^\star) - \frac{\ell_3}{n}, \tag{96}$$

$$D_n^\star - D_n(z^\star) \leq \frac{L_1}{n}, \tag{97}$$

$$|V_n(z) - V_n^\star| \leq F_1 d(z,z^\star) + \frac{F_2}{\sqrt{n}}, \tag{98}$$

$$V_{\min} \leq V_n(z), \tag{99}$$

$$T_n(z) \leq T_{\max}. \tag{100}$$

Theorem 7. *In the setup described above, under assumptions (96)–(100), for any $A > 0$, there exists a $K \geq 0$ such that, for all $|\Delta| \leq \delta_n$ (where δ_n is specified below) and all sufficiently large n:*

1. *If $\delta_n = \frac{A}{\sqrt{n}}$,*

$$\min_{z\in\mathcal{D}} \mathbb{P}\left[\sum_{i=1}^n W_i(z) \leq n\left(D_n^\star - \Delta\right)\right] \geq Q\left(\Delta\sqrt{\frac{n}{V_n^\star}}\right) - \frac{K}{\sqrt{n}}. \tag{101}$$

2. *For $\delta_n = A\sqrt{\frac{\log n}{n}}$,*

$$\min_{z\in\mathcal{D}} \mathbb{P}\left[\sum_{i=1}^n W_i(z) \leq n\left(D_n^\star - \Delta\right)\right] \geq Q\left(\Delta\sqrt{\frac{n}{V_n^\star}}\right) - K\sqrt{\frac{\log n}{n}}. \tag{102}$$

3. *Fix $0 \leq \beta \leq \frac{1}{6}$. If in (98), $V_n^\star = 0$ (which implies that $V_{\min} = 0$ in (99), i.e. we drop the requirement in Theorems 7.1 and 7.2 that $V_{\min}$ be positive), then there exists $K \geq 0$ such that for all $\Delta > \frac{A}{n^{\frac{1}{2}+\beta}}$, where $A > 0$ is arbitrary*

$$\min_{z\in\mathcal{D}} \mathbb{P}\left[\sum_{i=1}^n W_i(z) \leq n\left(D_n^\star + \Delta\right)\right] \geq 1 - \frac{K}{A^{\frac{3}{2}}}\frac{1}{n^{\frac{1}{4}-\frac{3}{2}\beta}}. \tag{103}$$

Theorem 7 gives a general result on the minimization of a cdf of a sum of independent random variables parameterized by elements of a metric space: it says that the minimum is approximately achieved by the sum with the largest mean, under regularity conditions. The metric nature of the parameter space is essential in making sure the means and the variances of $W_i(\cdot)$ behave like continuous functions: assumptions (98) and (97) essentially ensure that functions $D_n(\cdot)$ and $D_n(z)$ are well-behaved in the neighborhood of the optimum, while assumption (96) guarantees that $D_n(\cdot)$ decays fast enough near its maximum.

Before we proceed to prove Theorem 7, we recall the Berry-Esseen refinement of the central limit theorem.

Theorem 8 (Berry-Esseen CLT, e.g. [36, Ch. XVI.5 Theorem 2])**.** *Fix a positive integer n. Let W_i, $i = 1,\ldots,n$ be independent. Then, for any real t*

$$\left|\mathbb{P}\left[\sum_{i=1}^n W_i > n\left(D_n + t\sqrt{\frac{V_n}{n}}\right)\right] - Q(t)\right| \leq \frac{B_n}{\sqrt{n}}, \tag{104}$$

where

$$D_n = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[W_i\right], \tag{105}$$

$$V_n = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left[W_i\right], \tag{106}$$

$$T_n = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left|W_i - \mathbb{E}\left[W_i\right]\right|^3\right], \tag{107}$$

$$B_n = \frac{c_0 T_n}{V_n^{3/2}}, \tag{108}$$

and $0.4097 \leq c_0 \leq 0.5600$ ($c_0 \leq 0.4784$ for identically distributed W_i).

We also make note of the following lemma, which deals with the behavior of the Q-function.

Lemma 4 ([27, Lemma 4]). *Fix $b \geq 0$. Then, there exists $q \geq 0$ such that for all $z \geq -\frac{1}{2b}$ and all $n \geq 1$,*

$$Q\left(\sqrt{n}z\right) - Q\left(\sqrt{n}z\left(1+bz\right)\right) \leq \frac{q}{\sqrt{n}}. \tag{109}$$

We are now equipped to prove Theorem 7.

Proof of Theorem 7. To show (103), denote for brevity $\zeta = d\left(z, z^\star\right)$ and write

$$\begin{aligned}
&\mathbb{P}\left[\sum_{i=1}^{n} W_i(z) > n\left(D_n^\star + \Delta\right)\right] \\
&\leq \mathbb{P}\left[\sum_{i=1}^{n} W_i(z) > n\left(D_n(z) + \ell_1\zeta^2 - \frac{\ell_2}{\sqrt{n}}\zeta - \frac{\ell_3}{n} + \frac{A}{n^{\frac{1}{2}+\beta}}\right)\right] && (110)\\
&\leq \frac{1}{n}\frac{F_1\zeta + \frac{F_2}{\sqrt{n}}}{\left(\ell_1\zeta^2 - \frac{\ell_2}{\sqrt{n}}\zeta - \frac{\ell_3}{n} + \frac{A}{n^{\frac{1}{2}+\beta}}\right)^2} && (111)\\
&\leq \frac{K}{A^{\frac{3}{2}}}\frac{1}{n^{\frac{1}{4}-\frac{3}{2}\beta}}, && (112)
\end{aligned}$$

where

- (110) uses (96) and the assumption on the range of Δ;
- (111) is due to Chebyshev's inequality and $V_n^\star = 0$;
- (112) is by a straightforward algebraic exercise revealing that ζ that maximizes the left side of (112) is proportional to $\frac{A^{\frac{1}{2}}}{n^{\frac{1}{4}+\frac{1}{2}\beta}}$.

We proceed to show (101) and (102).

Denote

$$g_n(z) = \mathbb{P}\left[\sum_{i=1}^{n} W_i(z) \leq n(D_n^\star - \Delta)\right]. \tag{113}$$

Using (99) and (100), observe

$$\frac{c_0 T_n(z)}{V_n^{\frac{3}{2}}(z)} \leq B = \frac{c_0 T_{\max}}{V_{\min}^{\frac{3}{2}}} < \infty. \tag{114}$$

Therefore the Berry-Esseen bound yields:

$$\left|g_n(z) - Q\left(\sqrt{n}\nu_n(z)\right)\right| \leq \frac{B}{\sqrt{n}}, \tag{115}$$

where

$$\nu_n(z) \triangleq \frac{D_n(z) - D_n^\star + \Delta}{\sqrt{V_n(z)}}. \tag{116}$$

Denote

$$\nu_n^\star \triangleq \frac{\Delta}{\sqrt{V_n^\star}} \tag{117}$$

Since

$$\begin{aligned}
g_n(z) &= Q(\sqrt{n}\nu_n^\star) + \left[g_n(z) - Q\left(\sqrt{n}\nu_n(z)\right)\right] \\
&\quad + \left[Q\left(\sqrt{n}\nu_n(z)\right) - Q(\sqrt{n}\nu_n^\star)\right] && (118)\\
&\geq Q(\sqrt{n}\nu_n^\star) - \frac{B}{\sqrt{n}} + \left[Q\left(\sqrt{n}\nu_n(z)\right) - Q(\sqrt{n}\nu_n^\star)\right], && (119)
\end{aligned}$$

to show (101), it suffices to show that

$$Q(\sqrt{n}\nu_n^\star) - \min_{z\in\mathcal{D}} Q\left(\sqrt{n}\nu_n(z)\right) \leq \frac{q}{\sqrt{n}} \tag{120}$$

for some $q \geq 0$, and to show (102), replacing q with $q\sqrt{\log n}$ in the right side of (120) would suffice.

Since Q is monotonically decreasing, to achieve the minimum in (120) we need to maximize $\sqrt{n}\nu_n(z)$. As will be proven shortly, for appropriately chosen $a, b, c > 0$ we can write

$$\max_{z\in\mathcal{D}} \nu_n(z) \leq \nu_n^\star + b\nu_n^{\star 2} + \frac{c\delta_n}{\sqrt{n}} \tag{121}$$

for n large enough.

If

$$\Delta \geq -\frac{\sqrt{V_{\min}}}{2b} = -A, \tag{122}$$

then $\nu_n^\star \geq -\frac{1}{2b}$, and Lemma 4 applies to $\nu_n^\star$. So, using (121), the fact that $Q(\cdot)$ is monotonically decreasing and Lemma 4, we conclude that there exists $q > 0$ such that

$$\begin{aligned}
&Q\left(\sqrt{n}\nu_n^\star\right) - \min_{z\in\mathcal{D}} Q\left(\sqrt{n}\nu_n(z)\right) \\
&\leq Q\left(\sqrt{n}\nu_n^\star\right) - Q\left(\sqrt{n}\nu_n^\star + \sqrt{n}b\nu_n^{\star 2} + c\delta_n\right) && (123)\\
&\leq Q\left(\sqrt{n}\nu_n^\star\right) - Q\left(\sqrt{n}\nu_n^\star + \sqrt{n}b\nu_n^{\star 2}\right) + \frac{c}{\sqrt{2\pi}}\delta_n && (124)\\
&\leq \frac{q}{\sqrt{n}} + \frac{c}{\sqrt{2\pi}}\delta_n, && (125)
\end{aligned}$$

where

- (124) is due to

$$Q(z+\xi) \geq Q(z) - \frac{\xi}{\sqrt{2\pi}}, \tag{126}$$

  which holds for arbitrary z and $\xi \geq 0$,
- (125) holds by Lemma 4 as long as $\nu_n^\star \geq -\frac{1}{2b}$.

Thus, (125) establishes (101) and (102). It remains to prove (121). To upper-bound $\max_{z\in\mathcal{D}} \nu_n(z)$, denote for convenience

$$f_n(z) = \frac{D_n(z) - D_n^\star}{\sqrt{V_n(z)}}, \tag{127}$$

$$g_n(z) = \frac{1}{\sqrt{V_n(z)}}, \tag{128}$$

and note, using (96), (97), (99), (100) and (by Hölder's inequality)

$$V_n(z) \leq T_{\max}^{\frac{2}{3}}, \tag{129}$$

that

$$\begin{aligned}
f_n(z^\star) - f_n(z) &= \frac{D_n(z^\star) - D_n^\star}{\sqrt{V_n(z^\star)}} - \frac{D_n(z) - D_n^\star}{\sqrt{V_n(z)}} && (130)\\
&\geq \ell_1' d^2(z, z^\star) - \frac{\ell_2'}{\sqrt{n}} d(z, z^\star) - \frac{\ell_3'}{n}, && (131)
\end{aligned}$$

where

$$\ell_1' = T_{\max}^{-\frac{1}{3}}\ell_1, \tag{132}$$

$$\ell_2' = V_{\min}^{-\frac{1}{2}}\ell_2, \tag{133}$$

$$\ell_3' = V_{\min}^{-\frac{1}{2}}(L_1 + \ell_3). \tag{134}$$

Observe that for $a, b > 0$

$$\left|\frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}}\right| \le \frac{|a-b|}{2\min\{a,b\}^{\frac{3}{2}}}, \tag{135}$$

so, using (98) and (99), we conclude

$$\left|\frac{1}{\sqrt{V_n(z)}} - \frac{1}{\sqrt{V_n^\star}}\right| \le F_1' d(z, z^\star) + \frac{F_2'}{\sqrt{n}}, \tag{136}$$

where

$$F_1' = \frac{1}{2} V_{\min}^{-\frac{3}{2}} F_1, \tag{137}$$

$$F_2' = \frac{1}{2} V_{\min}^{-\frac{3}{2}} F_2. \tag{138}$$

Let z_0 achieve the maximum $\max_{z\in\mathcal{D}} \nu_n(z)$, i.e.

$$\max_{z\in\mathcal{D}} \nu_n(z) = f_n(z_0) + \Delta g_n(z_0). \tag{139}$$

Using (136) and (131), we have,

$$\begin{aligned}
&\nu_n(z_0) - \nu_n(z^\star) \\
&= (f_n(z_0) - f_n(z^\star)) + \Delta\left(g_n(z_0) - g_n(z^\star)\right) && (140)\\
&\le -\ell_1' d^2(z_0, z^\star) + \left(\frac{\ell_2'}{\sqrt{n}} + |\Delta| F_1'\right) d(z_0, z^\star) + \frac{2F_2'|\Delta|}{\sqrt{n}} \\
&\quad + \frac{\ell_3'}{n} && (141)\\
&\le \frac{1}{4\ell_1'}\left(\frac{\ell_2'}{\sqrt{n}} + |\Delta| F_1'\right)^2 + \frac{2F_2'|\Delta|}{\sqrt{n}} + \frac{\ell_3'}{n}, && (142)
\end{aligned}$$

where (142) follows because the maximum of its left side is achieved at $d(z_0, z^\star) = \frac{1}{2\ell_1'}\left(\frac{\ell_2'}{\sqrt{n}} + |\Delta| F_1'\right)$. Using (96), (99), (136), we upper-bound

$$\nu_n(z^\star) \le \nu_n^\star + \frac{F_2'|\Delta|}{\sqrt{n}} + \frac{\ell_3}{n V_{\min}} + \frac{\ell_3 F_2'}{n^{\frac{3}{2}}}. \tag{143}$$

Applying (142) and (143) to upper-bound $\max_{z\in\mathcal{D}} \nu_n(z)$, we have established (121) in which

$$b = \frac{F_1'^2 T_{\max}^{\frac{2}{3}}}{4\ell_1'}, \tag{144}$$

where we used (98) and (129) to upper-bound $\Delta^2 = \nu_n^{\star 2} V_n^\star$, thereby completing the proof.

□

Appendix D
Proof of the converse part of Theorem 4

Given a finite set $\mathcal{A}$, let $\mathcal{P}$ be the set of all distributions on $\mathcal{A}$ that satisfy the cost constraint,

$$\mathbb{E}\left[\mathsf{b}(\mathsf{X})\right] \le \beta, \tag{145}$$

which is a convex set in $\mathbb{R}^{|\mathcal{A}|}$.

Leveraging an idea of Tomamichel and Tan [30], we will weaken (24) by choosing $P_{\bar{Y}^n}$ to be a convex combination of non-product distributions with weights chosen to favor those distributions that are close to $P_{Y^{\star n}}$. Specifically (cf. [30]),

$$P_{\bar{Y}^n}(y^n) = \frac{1}{A}\sum_{\mathbf{k}\in\mathcal{K}} \exp\left(-|\mathbf{k}|^2\right) \prod_{i=1}^n P_{\mathsf{Y}|\mathbf{K}=\mathbf{k}}(y_i), \tag{146}$$

where $\{P_{\mathsf{Y}|\mathbf{K}=\mathbf{k}},\ \mathbf{k}\in\mathcal{K}\}$ are defined as follows, for some $c > 0$,

$$P_{\mathsf{Y}|\mathbf{K}=\mathbf{k}}(\mathsf{y}) = P_{\mathsf{Y}^\star}(\mathsf{y}) + \frac{k_{\mathsf{y}}}{\sqrt{nc}}, \tag{147}$$

$$\begin{aligned}
\mathcal{K} = \Bigg\{ &\mathbf{k}\in\mathbb{Z}^{|\mathcal{B}|}\colon \sum_{\mathsf{y}\in\mathcal{B}} k_{\mathsf{y}} = 0, \\
&-P_{\mathsf{Y}^\star}(\mathsf{y}) + \frac{1}{\sqrt{nc}} \le \frac{k_{\mathsf{y}}}{\sqrt{nc}} \le 1 - P_{\mathsf{Y}^\star}(\mathsf{y}) \Bigg\}, && (148)
\end{aligned}$$

$$A = \sum_{\mathbf{k}\in\mathcal{K}} \exp\left(-|\mathbf{k}|^2\right) < \infty. \tag{149}$$

Denote by $P_{\Pi(\mathsf{Y})}$ the minimum Euclidean distance approximation of an arbitrary $P_{\mathsf{Y}} \in \mathcal{Q}$, where $\mathcal{Q}$ is the set of distributions on the channel output alphabet $\mathcal{B}$, in the set $\left\{P_{\mathsf{Y}|\mathbf{K}=\mathbf{k}}\colon\ \mathbf{k}\in\mathcal{K}\right\}$:

$$P_{\Pi(\mathsf{Y})} = P_{\mathsf{Y}|\mathbf{K}=\mathbf{k}^\star} \text{ where } \mathbf{k}^\star = \arg\min_{\mathbf{k}\in\mathcal{K}} \left|P_{\mathsf{Y}} - P_{\mathsf{Y}|\mathbf{K}=\mathbf{k}}\right|. \tag{150}$$

The quality of approximation (150) is governed by [30]

$$\left|P_{\Pi(\mathsf{Y})} - P_{\mathsf{Y}}\right| \le \sqrt{\frac{|\mathcal{B}|(|\mathcal{B}|-1)}{nc}}. \tag{151}$$

We say that $x^n \in \mathcal{A}^n$ has type P_{X} if the number of times each letter $a\in\mathcal{A}$ is encountered in x^n is $nP_{\mathsf{X}}(a)$. An n-type is a distribution whose masses are multiples of $\frac{1}{n}$. Denote by $P_{\hat{\mathsf{X}}}$ the minimum Euclidean distance approximation of P_{X} in the set of n-types, that is,

$$P_{\hat{\mathsf{X}}} = \arg \min_{\substack{P\in\mathcal{P}:\\ P \text{ is an } n\text{-type}}} \left|P_{\mathsf{X}} - P\right|. \tag{152}$$

The accuracy of approximation in (152) is controlled by the following inequality:

$$\left|P_{\mathsf{X}} - P_{\hat{\mathsf{X}}}\right| \le \frac{\sqrt{|\mathcal{A}|\left(|\mathcal{A}|-1\right)}}{n}. \tag{153}$$

For each $P_{\mathsf{X}} \in \mathcal{P}$, let $x^n \in \mathcal{A}^n$ be an arbitrary sequence of type $P_{\hat{\mathsf{X}}}$, and lower-bound the sum in (146) by the term containing $P_{\Pi(\mathsf{Y})}$ to obtain:

$$\begin{aligned}
\jmath_{X^n;\bar{Y}^n}(x^n; y^n, \beta) &\le \sum_{i=1}^n \jmath_{\mathsf{X};\Pi(\mathsf{Y})}(x_i, y_i, \beta) \\
&\quad + nc\left|P_{\Pi(\mathsf{Y})} - P_{\mathsf{Y}^\star}\right|^2 + A. && (154)
\end{aligned}$$

Applying (146) and (154) to loosen (24), we conclude by Theorem 3 that, as long as an (n, M, ϵ') code exists, for an arbitrary $\gamma > 0$,

$$\epsilon' \ge \min_{P_{\mathsf{X}}\in\mathcal{P}} \mathbb{P}\left[\sum_{i=1}^n W_i(P_{\mathsf{X}}) \le \log M - \gamma - A\right] - \exp\left(-\gamma\right), \tag{155}$$

where

$$W_i(P_{\mathsf{X}}) = \jmath_{\mathsf{X};\Pi(\mathsf{Y})}(x_i, Y_i, \beta) + c\left|P_{\Pi(\mathsf{Y})} - P_{\mathsf{Y}^\star}\right|^2, \tag{156}$$

and Y_i is distributed according to $P_{\mathsf{Y}|\mathsf{X}=x_i}$.[4] To evaluate the minimization on the right side of (155), we will apply Theorem

[4] Strictly speaking, the order of $W_i(P_{\mathsf{X}})$, $i = 1, \ldots, n$ depends on the particular choice of sequence x^n of type $P_{\hat{\mathsf{X}}}$. However, since the distribution of the sum $\sum_{i=1}^n W_i(P_{\mathsf{X}})$ does not depend on their relative order, we may choose this sequence arbitrarily.

7 with $\mathcal{D}=\mathcal{P}$, $z=P_{\mathsf{X}}$, $z^\star=P_{\mathsf{X}^\star}$, $W_i(\cdot)$ in (156), and the metric being the usual Euclidean distance in $\mathbb{R}^n$.

Define the following functions $\mathcal{P}\times\mathcal{Q}\mapsto\mathbb{R}_+$:

$$D(P_{\mathsf{X}},P_{\bar{\mathsf{Y}}})=\mathbb{E}\left[\jmath_{\mathsf{X};\bar{\mathsf{Y}}}(\mathsf{X};\mathsf{Y},\beta)\right]+c\left|P_{\bar{\mathsf{Y}}}-P_{\mathsf{Y}^\star}\right|^2, \tag{157}$$

$$V(P_{\mathsf{X}},P_{\bar{\mathsf{Y}}})=\mathbb{E}\left[\mathrm{Var}\left[\jmath_{\mathsf{X};\bar{\mathsf{Y}}}(\mathsf{X};\mathsf{Y},\beta)\mid\mathsf{X}\right]\right], \tag{158}$$

$$T(P_{\mathsf{X}},P_{\bar{\mathsf{Y}}})=\mathbb{E}\left[\left|\jmath_{\mathsf{X};\bar{\mathsf{Y}}}(\mathsf{X};\mathsf{Y},\beta)-\mathbb{E}\left[\jmath_{\mathsf{X};\bar{\mathsf{Y}}}(\mathsf{X};\mathsf{Y},\beta)|\mathsf{X}\right]\right|^3\right], \tag{159}$$

where the expectations are with respect to $P_{\mathsf{Y}|\mathsf{X}}P_{\mathsf{X}}$.

With the choice in (156) the functions (93)–(95) are particularized to the following mappings $\mathcal{P}\mapsto\mathbb{R}_+$:

$$D_n(P_{\mathsf{X}})=D\left(P_{\hat{\mathsf{X}}},P_{\Pi(\mathsf{Y})}\right), \tag{160}$$

$$V_n(P_{\mathsf{X}})=V\left(P_{\hat{\mathsf{X}}},P_{\Pi(\mathsf{Y})}\right), \tag{161}$$

$$T_n(P_{\mathsf{X}})=T\left(P_{\hat{\mathsf{X}}},P_{\Pi(\mathsf{Y})}\right). \tag{162}$$

and $D_n^\star$, $V_n^\star$ are

$$D_n^\star=C(\beta), \tag{163}$$

$$V_n^\star=V(\beta). \tag{164}$$

We perform the minimization on the right side of (155) separately for $P_{\mathsf{X}}\in\mathcal{P}_\delta^\star$ and $P_{\mathsf{X}}\in\mathcal{P}\backslash\mathcal{P}_\delta^\star$, where

$$\mathcal{P}_\delta^\star=\left\{P_{\mathsf{X}}\in\mathcal{P}\colon\ |P_{\mathsf{X}}-P_{\mathsf{X}^\star}|\le\delta\right\}. \tag{165}$$

Assuming without loss of generality that all outputs in $\mathcal{B}$ are accessible (meaning that for each $\mathsf{y}\in\mathcal{B}$, there exists $\mathsf{x}\in\mathcal{A}$ with $P_{\mathsf{Y}|\mathsf{X}}(\mathsf{y}|\mathsf{x})>0$; this implies in particular that $P_{\mathsf{Y}^\star}(\mathsf{y})>0$ for all $\mathsf{y}\in\mathcal{B}$), we choose $\delta>0$ so that

$$\min_{P_{\mathsf{X}}\in\mathcal{P}_\delta^\star}\min_{\mathsf{y}\in\mathcal{B}}P_{\mathsf{Y}}(\mathsf{y})=p_{\min}>0, \tag{166}$$

$$2\min_{P_{\mathsf{X}}\in\mathcal{P}_\delta^\star}V\left(P_{\mathsf{X}}\right)\ge V(\beta). \tag{167}$$

To perform the minimization on the right side of (155) over $\mathcal{P}_\delta^\star$, we will invoke Theorem 7 with $\mathcal{D}=\mathcal{P}_\delta^\star$, the metric being the usual Euclidean distance between $|\mathcal{A}|$-vectors. Let us check that the assumptions of Theorem 7 are satisfied. It is easy to verify directly that the functions $P_{\mathsf{X}}\mapsto D(P_{\mathsf{X}},P_{\mathsf{Y}})$, $P_{\mathsf{X}}\mapsto V(P_{\mathsf{X}},P_{\mathsf{Y}})$, $P_{\mathsf{X}}\mapsto T(P_{\mathsf{X}},P_{\mathsf{Y}})$ are continuous (and therefore bounded) on $\mathcal{P}$ and infinitely differentiable on $\mathcal{P}_\delta^\star$. Therefore, assumptions (99) and (100) of Theorem 7 are met. To verify that (96) holds, write, for $\zeta=|P_{\mathsf{X}}-P_{\mathsf{X}^\star}|$,

$$C(\beta)-D\left(P_{\hat{\mathsf{X}}},P_{\Pi(\mathsf{Y})}\right)=C(\beta)-D\left(P_{\mathsf{X}},P_{\mathsf{Y}}\right)-\frac{\ell_2}{\sqrt{n}}\zeta-\frac{\ell_3}{n} \tag{168}$$

$$\ge\ell_1\zeta^2-\frac{\ell_2}{\sqrt{n}}\zeta-\frac{\ell_3}{n}, \tag{169}$$

where all constants ℓ_1, ℓ_2, ℓ_3 are positive, and:

- to show (168), observe that for a fixed $P_{\bar{\mathsf{Y}}}$, $D\left(\cdot,P_{\bar{\mathsf{Y}}}\right)$ is a linear function of P_{X}, so in view of (153)

$$\left|D\left(P_{\hat{\mathsf{X}}},P_{\Pi(\mathsf{Y})}\right)-D\left(P_{\mathsf{X}},P_{\Pi(\mathsf{Y})}\right)\right|\le\frac{L_1}{n}. \tag{170}$$

Furthermore,

$$\begin{aligned}
&D\left(P_{\mathsf{X}},P_{\Pi(\mathsf{Y})}\right)\\
&=D\left(P_{\mathsf{X}},P_{\mathsf{Y}}\right)+c|P_{\Pi(\mathsf{Y})}-P_{\mathsf{Y}^\star}|^2-c|P_{\mathsf{Y}}-P_{\mathsf{Y}^\star}|^2\\
&+D(P_{\mathsf{Y}}\|P_{\Pi(\mathsf{Y})}) &(171)\\
&\le D\left(P_{\mathsf{X}},P_{\mathsf{Y}}\right)+c|P_{\Pi(\mathsf{Y})}-P_{\mathsf{Y}}|^2\\
&+2c|P_{\mathsf{Y}}-P_{\mathsf{Y}^\star}||P_{\Pi(\mathsf{Y})}-P_{\mathsf{Y}}|+D(P_{\mathsf{Y}}\|P_{\Pi(\mathsf{Y})}) &(172)\\
&\le D\left(P_{\mathsf{X}},P_{\mathsf{Y}}\right)+\frac{\ell_2}{\sqrt{n}}\zeta+\frac{\ell_3'}{n}, &(173)
\end{aligned}$$

where we used the triangle inequality, (151), a "reverse Pinsker inequality" [37, Lemma 6.3]:

$$D(\mathsf{Y}\|\bar{\mathsf{Y}})\le\frac{\log e}{\min_{b\in\mathcal{B}}P_{\bar{\mathsf{Y}}}(b)}\left|P_{\mathsf{Y}}-P_{\bar{\mathsf{Y}}}\right|^2 \tag{174}$$

and

$$|P_{\mathsf{Y}}-P_{\bar{\mathsf{Y}}}|\le|P_{\mathsf{Y}|\mathsf{X}}||P_{\mathsf{X}}-P_{\bar{\mathsf{X}}}|, \tag{175}$$

where $P_{\bar{\mathsf{X}}}\to P_{\mathsf{Y}|\mathsf{X}}\to P_{\bar{\mathsf{Y}}}$, and the spectral norm of $P_{\mathsf{Y}|\mathsf{X}}$ satisfies $|P_{\mathsf{Y}|\mathsf{X}}|\le\sqrt{|\mathcal{A}|}$.

- (169) uses

$$\mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y},\beta)\right]\le C(\beta)-\ell_1'\zeta^2, \tag{176}$$

where $\ell_1'>0$, and

$$\ell_1=\ell_1'-c|\mathcal{A}| \tag{177}$$

can be made positive for a small enough c. Inequality (176) can be shown following the reasoning in [20, (497)–(505)] invoking (16) in lieu of the corresponding property for the conventional information density. Here we provide a simpler proof using Pinsker's inequality. Viewing P_{X} as a vector and $P_{\mathsf{Y}|\mathsf{X}}$ as a matrix, write

$$P_{\mathsf{X}}=P_{\mathsf{X}^\star}+v_0+v_\perp, \tag{178}$$

where v_0 and $v_\perp$ are projections of $P_{\mathsf{X}}-P_{\mathsf{X}^\star}$ onto $\mathrm{Ker}P_{\mathsf{Y}|\mathsf{X}}$ and $(\mathrm{Ker}P_{\mathsf{Y}|\mathsf{X}})^\perp$ respectively, where

$$\mathrm{Ker}P_{\mathsf{Y}|\mathsf{X}}=\left\{v\in\mathbb{R}^{|\mathcal{A}|}\colon\ v^TP_{\mathsf{Y}|\mathsf{X}}=0\right\}. \tag{179}$$

We consider two cases $v_\perp=0$ and $v_\perp\neq0$ separately. Condition $v_\perp=0$ implies $P_{\mathsf{X}}\to P_{\mathsf{Y}|\mathsf{X}}\to P_{\mathsf{Y}^\star}$, which combined with $P_{\mathsf{X}}\neq P_{\mathsf{X}^\star}$ and (16) means that the complement of $F=\mathrm{supp}(P_{\mathsf{X}^\star})$ is nonempty and

$$a\triangleq C(\beta)-\max_{x\notin F}\mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{x};\mathsf{Y},\beta)|\mathsf{X}=\mathsf{x}\right] \tag{180}$$

is positive. Therefore

$$\begin{aligned}
&\mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y},\beta)\right]\\
&=\mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X};\mathsf{Y},\beta)\right] &(181)\\
&=\mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X};\mathsf{Y},\beta),\mathsf{X}\in F\right]+\mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X};\mathsf{Y},\beta),\mathsf{X}\notin F\right] &(182)\\
&\le C(\beta)P_{\mathsf{X}}\left(F\right)+P_{\mathsf{X}}\left(F^c\right)\left(C(\beta)-a\right) &(183)\\
&\le C(\beta)-(\lambda_{\min}^+(P_F^2))^{1/2}a|v| &(184)\\
&\le C(\beta)-\frac{1}{4}(\lambda_{\min}^+(P_F^2))^{1/2}a|v|^2, &(185)
\end{aligned}$$

where (183) uses (16), P_F is the orthogonal projection matrix onto F^c and $\lambda_{\min}^+(\cdot)$ is the minimum nonzero eigenvalue of the indicated positive semidefinite matrix.

If $v_\perp \neq 0$, write

$$\begin{aligned}
&\mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y},\beta)\right] \\
&= \mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X};\mathsf{Y},\beta)\right] - D(P_{\mathsf{Y}}\|P_{\mathsf{Y}^\star}) && (186)\\
&\le \mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{X};\mathsf{Y},\beta)\right] - \frac{1}{2}\left|P_{\mathsf{Y}} - P_{\mathsf{Y}^\star}\right|^2 \log e && (187)\\
&\le C(\beta) - \frac{1}{2}\left|P_{\mathsf{Y}} - P_{\mathsf{Y}^\star}\right|^2 \log e, && (188)
\end{aligned}$$

where (187) is by Pinsker's inequality, and (188) is by (14). To conclude the proof of (176), we lower bound the second term in (188) as follows.

$$\begin{aligned}
|P_{\mathsf{Y}} - P_{\mathsf{Y}^\star}|^2 &= \left|(P_{\mathsf{X}} - P_{\mathsf{X}^\star})^T P_{\mathsf{Y}|\mathsf{X}}\right|^2 && (189)\\
&= \left|v_\perp^T P_{\mathsf{Y}|\mathsf{X}}\right|^2 && (190)\\
&\ge \lambda_{\min}(P_{\mathsf{Y}|\mathsf{X}})|v_\perp|^2 && (191)\\
&\ge \lambda^+_{\min}(P_{\mathsf{Y}|\mathsf{X}}P^T_{\mathsf{Y}|\mathsf{X}})\lambda^+_{\min}(P_\perp^2)|v|^2, && (192)
\end{aligned}$$

where $P_\perp$ is the orthogonal projection matrix onto $(\mathrm{Ker}P_{\mathsf{Y}|\mathsf{X}})^\perp$.

To establish (97), write

$$\begin{aligned}
C(\beta) - D(P_{\hat{\mathsf{X}}}, P_{\Pi(\mathsf{Y})}) &\le C(\beta) - D(P_{\mathsf{X}}, P_{\Pi(\mathsf{Y})}) + \frac{L_1}{n} && (193)\\
&\le C(\beta) - \mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y},\beta)\right] + \frac{L_1}{n}, && (194)
\end{aligned}$$

where (193) is due to (170). Substituting $\mathsf{X} = \mathsf{X}^\star$ into (194), we obtain (97).

Finally, to verify (98), write

$$\begin{aligned}
&\left|V\left(P_{\hat{\mathsf{X}}}, P_{\Pi(\mathsf{Y})}\right) - V(\beta)\right| \\
&\le \left|V(P_{\mathsf{X}}, P_{\mathsf{Y}}) - V(\beta)\right| + \left|V(P_{\mathsf{X}}, P_{\mathsf{Y}}) - V\left(P_{\hat{\mathsf{X}}}, P_{\mathsf{Y}}\right)\right| \\
&+ \left|V\left(P_{\hat{\mathsf{X}}}, P_{\Pi(\mathsf{Y})}\right) - V\left(P_{\hat{\mathsf{X}}}, P_{\mathsf{Y}}\right)\right| && (195)\\
&\le F_1|P_{\mathsf{X}} - P_{\mathsf{X}^\star}| + F_2'|P_{\mathsf{X}} - P_{\hat{\mathsf{X}}}| + F_2''\left|P_{\Pi(\mathsf{Y})} - P_{\mathsf{Y}}\right| && (196)\\
&\le F_1\zeta + \frac{F_2}{\sqrt{n}} && (197)
\end{aligned}$$

where all constants F are positive, and

- (196) uses continuous differentiability of $P_{\mathsf{X}} \mapsto V(P_{\mathsf{X}}, P_{\mathsf{Y}})$ (in $\mathcal{P}^\star_\delta$) and $P_{\bar{\mathsf{Y}}} \mapsto V(P_{\mathsf{X}}, P_{\bar{\mathsf{Y}}})$ (at any $P_{\bar{\mathsf{Y}}}$ with $P_{\bar{\mathsf{Y}}}(\mathsf{Y}) > 0$ a.s.).
- (197) applies (153) and (151).

Theorem 7 is thereby applicable.

If $V(\beta) > 0$, letting

$$\begin{aligned}
\gamma &= \frac{1}{2}\log n && (198)\\
\log M &= nC(\beta) - \sqrt{nV(\beta)}\, Q^{-1}\left(\epsilon + \frac{K+1}{\sqrt{n}}\right) + \frac{1}{2}\log n \\
&\quad + A, && (199)
\end{aligned}$$

where constant K is the same as in (101), we apply Theorem 7.1 to conclude that the right side of (155) with minimization constrained to types in $\mathcal{P}^\star_\delta$ s lower bounded by ϵ:

$$\min_{P_{\mathsf{X}} \in \mathcal{P}^\star_\delta} \mathbb{P}\left[\sum_{i=1}^n W_i(P_{\mathsf{X}}) \le \log M - \gamma - A\right] - \exp\left(-\gamma\right) \ge \epsilon. \tag{200}$$

If $V(\beta) = 0$, we fix $0 < \eta < 1 - \epsilon$ and let

$$\begin{aligned}
\gamma &= \log\frac{1}{\eta}, && (201)\\
\log M &= nC(\beta) + \left(\frac{K}{1-\epsilon-\eta}\right)^{\frac{2}{3}} n^{\frac{1}{3}} + \log\frac{1}{\eta}, && (202)
\end{aligned}$$

where A is that in (103). Applying Theorem 7.3 with $\beta = \frac{1}{6}$, we conclude that (200) holds for the choice of M in (202) if $V(\beta) = 0$.

To evaluate the minimum over $\mathcal{P}\backslash\mathcal{P}^\star_\delta$ on the right side of (155), define

$$C(\beta) - \max_{P_{\mathsf{X}} \in \mathcal{P}\backslash\mathcal{P}^\star_\delta} \mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y},\beta)\right] = 2\Delta > 0 \tag{203}$$

and observe

$$\begin{aligned}
&D(P_{\mathsf{X}}, P_{\Pi(\mathsf{Y})}) \\
&= \mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y},\beta)\right] + D(\mathsf{Y}\|\Pi(\mathsf{Y})) + c|P_{\Pi(\mathsf{Y})} - P_{\mathsf{Y}^\star}|^2 && (204)\\
&\le \mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y},\beta)\right] + D(\mathsf{Y}\|\Pi(\mathsf{Y})) + 4c && (205)\\
&\le \mathbb{E}\left[\jmath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y},\beta)\right] + \frac{|\mathcal{B}|(|\mathcal{B}|-1)\log e}{\sqrt{nc}} + 4c, && (206)
\end{aligned}$$

where

- (205) holds because the Euclidean distance between two distributions satisfies

$$|P_{\mathsf{Y}} - P_{\bar{\mathsf{Y}}}| \le 2, \tag{207}$$

- (206) is due to (151), (174), and

$$\min_{\mathsf{Y}} \min_{\mathsf{y} \in \mathcal{B}} P_{\Pi(\mathsf{Y})}(\mathsf{y}) \ge \frac{1}{\sqrt{nc}}, \tag{208}$$

  which is a consequence of (148).

Therefore, choosing $c < \frac{\Delta}{4}$, we can ensure that for all n large enough,

$$C(\beta) - \max_{P_{\mathsf{X}} \in \mathcal{P}\backslash\mathcal{P}^\star_\delta} D(P_{\mathsf{X}}, P_{\Pi(\mathsf{Y})}) \ge \Delta > 0. \tag{209}$$

Also, it is easy to show using (208) that there exists $a > 0$ such that

$$V(P_{\mathsf{X}}, P_{\Pi(\mathsf{Y})}) \le a\log^2 n. \tag{210}$$

By Chebyshev's inequality, we have, for the choice of γ in (198) and M in (199),

$$\begin{aligned}
&\max_{P_{\mathsf{X}} \in \mathcal{P}\backslash\mathcal{P}^\star_\delta} \mathbb{P}\left[\sum_{i=1}^n W_i(P_{\mathsf{X}}) > \log M - \gamma - A\right] \\
&\le \mathbb{P}\left[\sum_{i=1}^n W_i(P_{\mathsf{X}}) - \mathbb{E}\left[W_i(P_{\mathsf{X}})\right] > \frac{n\Delta}{2}\right] && (211)\\
&\le \frac{4a}{\Delta^2}\frac{\log^2 n}{n}. && (212)
\end{aligned}$$

Combining (200) and (212) concludes the proof.

APPENDIX E
PROOF OF THE ACHIEVABILITY PART OF THEOREM 4

The proof consists of the asymptotic analysis of the following bound.

Theorem 9 (Dependence Testing bound [20])**.** *There exists an (M,ϵ,β) code with*

$$\epsilon \le \inf_{P_X} \mathbb{E}\left[\exp\left(-\left|\imath_{X;Y}(X;Y) - \log\frac{M-1}{2}\right|^+\right)\right], \quad (213)$$

where the infimum is over all distributions supported on $\{x \in \mathcal{X} : \mathsf{b}(x) \le \beta\}$.

The following lemma will be instrumental.

Lemma 5 ([20, Lemma 47])**.** *Let $W_1,\dots,W_n$ be independent, with $V_n > 0$ and $T_n < \infty$ where V_n and T_n are defined in (106) and (107), respectively. Then for any $\gamma > 0$,*

$$\mathbb{E}\left[\exp\left\{-\sum_{i=1}^n W_i\right\} 1\left\{\sum_{i=1}^n W_i > \log\gamma\right\}\right] \le 2\left(\frac{\log 2}{\sqrt{2\pi}} + \frac{2T_n}{\sqrt{nV_n}}\right)\frac{1}{\gamma\sqrt{nV_n}}. \quad (214)$$

Let P_{X^n} be equiprobable on the set of sequences of type $P_{\hat{\mathsf{X}}^\star}$, where $P_{\hat{\mathsf{X}}^\star}$ is the minimum Euclidean distance approximation of $P_{\mathsf{X}^\star}$ formally defined in (152). Let $P_{X^n} \to P_{Y^n|X^n} \to P_{Y^n}$, $P_{\hat{\mathsf{X}}^\star} \to P_{\mathsf{Y}|\mathsf{X}} \to P_{\hat{\mathsf{Y}}^\star}$, and $P_{\hat{Y}^{n\star}} = P_{\hat{\mathsf{Y}}^\star} \times \ldots \times P_{\hat{\mathsf{Y}}^\star}$.

The following lemma demonstrates that P_{Y^n} is close to $P_{\hat{Y}^{n\star}}$.

Lemma 6. *Almost surely, for n large enough and some constant c,*

$$\imath_{Y^n\|\hat{Y}^{n\star}}(Y^n) \le \frac{1}{2}\left(|\mathrm{supp}\,(P_{\mathsf{X}^\star})| - 1\right)\log n + c \quad (215)$$

Proof. For a vector $\mathbf{k} = (k_1,\dots,k_{|\mathcal{B}|})$, denote the multinomial coefficient

$$\binom{n}{\mathbf{k}} = \frac{n!}{k_1!k_2!\dots k_{|\mathcal{B}|}!} \quad (216)$$

By Stirling's approximation, the number of sequences of type $P_{\hat{\mathsf{X}}^\star}$ satisfies, for n large enough and some constant $c_1 > 0$

$$\binom{n}{nP_{\hat{\mathsf{X}}^\star}} \ge c_1 n^{-\frac{1}{2}(|\mathrm{supp}(P_{\mathsf{X}^\star})|-1)} \exp\left(nH(\hat{\mathsf{X}}^\star)\right) \quad (217)$$

On the other hand, for all x^n of type $P_{\hat{X}^{\star n}}$,

$$P_{\hat{X}^{\star n}}(x^n) = \exp\left(-nH(\hat{\mathsf{X}}^\star)\right) \quad (218)$$

Assume without loss of generality that all outputs in $\mathcal{B}$ are accessible, which implies that $P_{\mathsf{Y}^\star}(\mathsf{y}) > 0$ for all $\mathsf{y} \in \mathcal{B}$. Hence, the left side of (215) is almost surely finite, and for all $y^n \in \mathcal{Y}^n$ with nonzero probability according to P_{Y^n},

$$\frac{P_{Y^n}(y^n)}{P_{\hat{Y}^{n\star}}(y^n)} = \frac{\binom{n}{nP_{\hat{\mathsf{X}}^\star}}^{-1}\sum^\star 1}{\sum_{x^n\in\mathcal{A}^n} P_{Y^n|X^n=x^n}(y^n)P_{\hat{X}^{n\star}}(x^n)} \quad (219)$$

$$\le \frac{\binom{n}{nP_{\hat{\mathsf{X}}^\star}}^{-1}\sum^\star 1}{\sum^\star P_{Y^n|X^n=x^n}(y^n)P_{\hat{X}^{n\star}}(x^n)} \quad (220)$$

$$= \frac{\binom{n}{nP_{\hat{\mathsf{X}}^\star}}^{-1}\sum^\star 1}{\exp\left(-nH(\hat{\mathsf{X}}^\star)\right)\sum^\star 1} \quad (221)$$

$$= \binom{n}{nP_{\hat{\mathsf{X}}^\star}}^{-1}\exp\left(nH(\hat{\mathsf{X}}^\star)\right) \quad (222)$$

$$\le c_1 n^{\frac{1}{2}(|\mathrm{supp}(P_{\mathsf{X}^\star})|-1)}, \quad (223)$$

where we abbreviated $\sum^\star = \sum_{x^n:\,\mathrm{type}(x^n)=P_{\hat{\mathsf{X}}^\star}}$. □

We first consider the case $V(\beta) > 0$. For c in (215) and some $\gamma > 0$, let

$$\log\frac{M-1}{2} \triangleq S_n - \frac{1}{2}\left(|\mathrm{supp}\,(P_{\mathsf{X}^\star})| - 1\right)\log n - c, \quad (224)$$

$$S_n \triangleq nD_n - \sqrt{nV_n}Q^{-1}(\epsilon_n), \quad (225)$$

$$\epsilon_n \triangleq \epsilon - 2\left(\frac{\log 2}{\sqrt{2\pi}} + \frac{2T_n}{\sqrt{nV_n}}\right)\frac{1}{\gamma\sqrt{nV_n}} - \frac{B_n}{\sqrt{n}}, \quad (226)$$

where D_n and V_n are those in (105) and (106), computed with $W_i = \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(x_i,Y_i)$, namely

$$D_n = \mathbb{E}\left[\imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(\hat{\mathsf{X}}^\star,\hat{\mathsf{Y}}^\star)\right] \quad (227)$$

$$V_n = \mathrm{Var}\left[\imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(\hat{\mathsf{X}}^\star,\hat{\mathsf{Y}}^\star)|\hat{\mathsf{X}}^\star\right] \quad (228)$$

Since the functions $P_{\mathsf{X}} \mapsto \mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X},\mathsf{Y})\right]$ and $P_{\mathsf{X}} \mapsto \mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X},\mathsf{Y})|\mathsf{X}\right]$ are continuously differentiable in a neighborhood of $P_{\mathsf{X}^\star}$ in which $P_{\mathsf{Y}}(\mathsf{Y}) > 0$ a.s., there exist constants $L_1 \ge 0$, $F_1 \ge 0$ such that

$$|D_n - C(\beta)| \le L_1|P_{\hat{\mathsf{X}}^\star} - P_{\mathsf{X}^\star}|, \quad (229)$$

$$|V_n - V(\beta)| \le F_1|P_{\hat{\mathsf{X}}^\star} - P_{\mathsf{X}^\star}|, \quad (230)$$

where we used (21). Applying (153), we observe that the choice of $\log M$ in (224) satisfies (35), (38). Therefore, to prove the claim we need to show that the right side of (213) with the choice of M in (224) is upper bounded by ϵ.

Weakening (213) by choosing P_{X^n} equiprobable on the set of sequences of type $P_{\hat{\mathsf{X}}^\star}$, as above, we infer that an (M,ϵ',β) code exists with

$$\epsilon' \le \mathbb{E}\left[\exp\left(-\left|\imath_{X^n;Y^n}(X^n;Y^n) - \log\frac{M-1}{2}\right|^+\right)\right] \tag{231}$$

$$= \mathbb{E}\Big[\exp\Big(-\Big|\sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(X_i;Y_i) - \imath_{Y^n\|\hat{Y}^{n\star}}(Y^n) - \log\frac{M-1}{2}\Big|^+\Big)\Big] \tag{232}$$

$$\le \mathbb{E}\left[\exp\left(-\left|\sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(X_i;Y_i) - S_n\right|^+\right)\right] \tag{233}$$

$$= \mathbb{E}\left[\exp\left(-\left|\sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(x_i;Y_i) - S_n\right|^+\right)\right] \tag{234}$$

$$\le \exp(S_n)\cdot \mathbb{E}\left[\exp\left(-\sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(x_i;Y_i)\right) 1\left\{\sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(x_i;Y_i) > S_n\right\}\right] + \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(x_i;Y_i) \le S_n\right] \tag{235}$$

$$\le \epsilon, \tag{236}$$

where

- (233) applies Lemma 6 and substitutes (224);
- (234) holds for any choice of x^n of type $P_{\hat{\mathsf{X}}^\star}$ because the (conditional on $X^n = x^n$) distribution of $\imath_{X^n;\hat{Y}^{n\star}}(x^n;Y^n) = \sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(x_i;Y_i)$ depends the choice of x^n only through its type;
- (236) upper-bounds the first term using Lemma 5, and the second term using Theorem 8.

If $V(\beta) = 0$, let S_n in (224) be

$$S_n = nD_n - 2\gamma, \tag{237}$$

and let $\gamma > 0$ be the solution to

$$\exp(-\gamma) + \frac{F_1\sqrt{|\mathcal{A}|(|\mathcal{A}|-1)}}{\gamma^2} = \epsilon, \tag{238}$$

where F_1 is that in (230). Note that such solution exists because the function in the left side of (238) is continuous on $(0,\infty)$, unbounded as $\gamma \to 0$ and vanishing as $\gamma \to \infty$. The reasoning up to (234) still applies, at which point we upper-bound the right-side of (234) in the following way:

$$\epsilon' \le \exp(-\gamma)\,\mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(x_i;Y_i) > S_n + \gamma\right] + \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\hat{\mathsf{Y}}^\star}(x_i;Y_i) \le S_n + \gamma\right] \tag{239}$$

$$\le \exp(-\gamma) + \frac{nV_n}{\gamma^2} \tag{240}$$

$$\le \epsilon, \tag{241}$$

where

- (240) upper-bounds the second probability using Chebyshev's inequality;
- (241) uses $V(\alpha) = 0$, (153) and (230).

Appendix F
Proof of Theorem 4 under the assumptions of Remark 3

Under assumption (a), every (n, M, ϵ, β) code with a maximal cost constraint can be converted to an $(n+1, M, \epsilon, \beta)$ code with an equal cost constraint (i.e. equality in (22) is requested) by appending to each codeword a coordinate x_{n+1} with

$$\mathsf{b}(x_{n+1}) = \beta - \sum_{i=1}^n \mathsf{b}(x_i). \tag{242}$$

Since $\sum_{i=1}^n \mathsf{b}(x_i) \le \beta n$, the right side of (242) is no smaller than β, and so by assumption (a) a coordinate x_{n+1} satisfying (242) can be found. It follows that

$$M^\star_{\mathrm{eq}}(n,\epsilon,\beta) \le M^\star_{\max}(n,\epsilon,\beta) \le M^\star_{\mathrm{eq}}(n+1,\epsilon,\beta), \tag{243}$$

where the subscript specifies the nature of the cost constraint. We thus may focus only on the codes with equal cost constraint. The capacity-cost function can be expressed as (43) due to (16). The converse part now follows by invoking (24) with $P_{\bar{Y}^n} = P_{\mathsf{Y}^\star} \times \ldots \times P_{\mathsf{Y}^\star}$ and $\gamma = \frac{1}{2}\log n$. A simple application of the Berry-Esseen bound (Theorem 8) using assumption (c) leads to the desired result.

To show the achievability part, we follow the proof in Appendix E, drawing the codewords from P_{X^n} appearing in assumption (d), replacing all minimum distance approximations by the true distributions, and replacing the right side of (215) by f_n.

Appendix G
Dispersion-cost function of an additive exponential channel

As shown in [32], the capacity-cost function is given by (48), and $\mathsf{Y}^\star$ is exponential with mean $1+\beta$, i.e.

$$dP_{\mathsf{Y}^\star}(y) = \frac{1}{1+\beta} e^{-\frac{y}{1+\beta}}\,dy, \tag{244}$$

which leads to the expression for b-tilted information density in (47). Conditions (a)–(c) in Remark 3 are clearly satisfied. To verify condition (d), let P_{X^n} be uniform on the $(n-1)$-simplex $\{x^n \in \mathbb{R}_+^n \colon \sum_{i=1}^n x_i = n\beta\}$. Then, the distribution of $Y^n = X^n + N^n$, where N^n is a vector of i.i.d. exponential components with means 1, is a function of $\sum_{i=1}^n N_i$ only. Since the same holds for $Y^{n\star}$, the log-likelihood ratio $\imath_{Y^n\|Y^{n\star}}(y^n)$ is also a function of $\sum_{i=1}^n y_i$ only. Now, the sum of n exponentially distributed random variables with mean a has Erlang distribution, whose pdf is $\frac{t^{n-1}e^{-t/a}}{a^n(n-1)!}dt$, so (assuming natural logarithms for ease of computation)

$$\imath_{Y^n\|Y^{n\star}}(y^n) = L\left(\sum_{i=1}^n y_i, n\right), \tag{245}$$

$$L(t,n) \triangleq n\beta - \frac{\beta}{1+\beta}t + n\log_e(1+\beta) + (n-1)\log_e\left(1 - \frac{n\beta}{t}\right). \tag{246}$$

A direct algebraic computation shows that for each n, the maximum of $L(\cdot, n)$ is achieved at

$$t^\star(n) \triangleq \frac{1}{2}\left(n\beta + \sqrt{n}\sqrt{n\beta^2 + 4n(1+\beta) - 4(1+\beta)}\right). \tag{247}$$

Another computation verifies that $L(t^\star(n), n)$ is monotonically decreasing in n, so

$$\max_{n,t} L(t,n) = L(t^\star(1), 1) \tag{248}$$

$$= \frac{\beta}{1+\beta} + \log_e(1+\beta), \tag{249}$$

i.e. $\imath_{Y^n \| Y^{n\star}}(y^n)$ is bounded by a constant, and condition (d) is satisfied.

REFERENCES

[1] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois Journal of Mathematics*, vol. 1, no. 4, pp. 591–606, 1957.

[2] ——, "Strong converse of the coding theorem for semicontinuous channels," *Illinois Journal of Mathematics*, vol. 3, no. 4, pp. 477–489, 1959.

[3] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 19, no. 3, pp. 357–359, 1973.

[4] G. Dueck and J. Körner, "Reliability function of a discrete memoryless channel at rates above capacity," *IEEE Transactions on Information Theory*, vol. 25, no. 1, pp. 82–85, Jan 1979.

[5] J. H. B. Kemperman, "Strong converses for a general memoryless channel with feedback," in *Proceedings 6th Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes*, 1971, pp. 375–409.

[6] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proceedings 48th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, 2010, pp. 1327–1333.

[7] J. Wolfowitz, "The maximum achievable length of an error correcting code," *Illinois Journal of Mathematics*, vol. 2, no. 3, pp. 454–458, 1958.

[8] A. Feinstein, "On the coding theorem and its converse for finite-memory channels," *Information and Control*, vol. 2, no. 1, pp. 25–44, 1959.

[9] J. Wolfowitz, "A note on the strong converse of the coding theorem for the general discrete finite-memory channel," *Information and Control*, vol. 3, no. 1, pp. 89 – 93, 1960.

[10] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[11] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer, Berlin, 2003.

[12] M. Pinsker, *Information and information stability of random variables and processes*. San Francisco: Holden-Day, 1964.

[13] Y. Polyanskiy and S. Verdú, "Relative entropy at the channel output of a capacity-achieving code," in *Proceedings 49th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, Sep. 2011, pp. 52–59.

[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge Univ Press, 2011.

[15] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, no. 3, pp. 611–656, 1959.

[16] K. Yoshihara, "Simple proofs for the strong converse theorems in some channels," *Kodai Mathematical Journal*, vol. 16, no. 4, pp. 213–222, 1964.

[17] J. Wolfowitz, "Note on the Gaussian channel with feedback and a power constraint," *Information and Control*, vol. 12, no. 1, pp. 71–78, 1968.

[18] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Dept. Electrical Engineering, Princeton University, 2010.

[19] V. Strassen, "Asymptotische abschätzungen in Shannon's informationstheorie," in *Proceedings 3rd Prague Conference on Information Theory*, Prague, 1962, pp. 689–723.

[20] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

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

[22] P. Moulin, "The log-volume of optimal constant-composition codes for memoryless channels, within $O(1)$ bits," in *Proceedings 2012 IEEE International Symposium on Information Theory*, Cambridge, MA, July 2012, pp. 826–830.

[23] D. Wang, A. Ingber, and Y. Kochman, "The dispersion of joint source-channel coding," in *Proceedings 49th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, Sep. 2011.

[24] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.

[25] J. H. B. Kemperman, "On the Shannon capacity of an arbitrary channel," *Indagationes Mathematicae (Proceedings)*, vol. 77, no. 2, pp. 101–115, 1974.

[26] J. Wolfowitz, "Notes on a general strong converse," *Information and Control*, vol. 12, no. 1, pp. 1–4, 1968.

[27] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2545–2575, May 2013.

[28] E. Çinlar, *Probability and Stochastics*. Springer, 2011.

[29] Y. Polyanskiy, "Saddle point in the minimax converse for channel coding," *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2576–2595, 2013.

[30] M. Tomamichel and V. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7041–7051, 2013.

[31] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the AWGN channel," in *2014 IEEE International Symposium on Information Theory*, Honolulu, HI, June 2014, pp. 2077–2081.

[32] S. Verdú, "The exponential distribution in information theory," *Problemy Peredachi Informatsii*, vol. 32, no. 1, pp. 100–111, 1996.

[33] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, pp. 146–158, 1975.

[34] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, I," *Communications on Pure and Applied Mathematics*, vol. 28, no. 1, pp. 1–47, 1975.

[35] S. Verdú, *Information Theory*, in preparation.

[36] W. Feller, *An Introduction to Probability Theory and its Applications*, 2nd ed. John Wiley & Sons, 1971, vol. II.

[37] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1007–1016, March 2006.