Lossy joint source-channel coding in the finite blocklength regime

Victoria Kostina, *Student Member, IEEE*, Sergio Verdú, *Fellow, IEEE*

Abstract—This paper finds new tight finite-blocklength bounds for the best achievable lossy joint source-channel code rate, and demonstrates that joint source-channel code design brings considerable performance advantage over a separate one in the non-asymptotic regime. A joint source-channel code maps a block of k source symbols onto a length$-n$ channel codeword, and the fidelity of reproduction at the receiver end is measured by the probability ϵ that the distortion exceeds a given threshold d. For memoryless sources and channels, it is demonstrated that the parameters of the best joint source-channel code must satisfy $nC - kR(d) \approx \sqrt{nV + k\mathcal{V}(d)}Q^{-1}(\epsilon)$, where C and V are the channel capacity and channel dispersion, respectively; $R(d)$ and $\mathcal{V}(d)$ are the source rate-distortion and rate-dispersion functions; and Q is the standard Gaussian complementary cdf. Symbol-by-symbol (uncoded) transmission is known to achieve the Shannon limit when the source and channel satisfy a certain probabilistic matching condition. In this paper we show that even when this condition is not satisfied, symbol-by-symbol transmission is, in some cases, the best known strategy in the non-asymptotic regime.

Index Terms—Achievability, converse, finite blocklength regime, joint source-channel coding, lossy source coding, memoryless sources, rate-distortion theory, Shannon theory.

I. INTRODUCTION

In the limit of infinite blocklengths, the optimal achievable coding rates in channel coding and lossy data compression are characterized by the channel capacity C and the source rate-distortion function $R(d)$, respectively [3]. For a large class of sources and channels, in the limit of large blocklength, the maximum achievable joint source-channel coding (JSCC) rate compatible with vanishing excess distortion probability is characterized by the ratio $\frac{C}{R(d)}$ [4]. A perennial question in information theory is how relevant the asymptotic fundamental limits are when the communication system is forced to operate at a given fixed blocklength. The finite blocklength (delay) constraint is inherent to all communication scenarios. In fact, in many systems of current interest, such as real-time multi-media communication, delays are strictly constrained, while in packetized data communication, packets are frequently on the order of 1000 bits. While computable formulas for the

The authors are with the Department of Electrical Engineering, Princeton University, NJ 08544 USA (e-mail: vkostina@princeton.edu; verdu@princeton. edu).

channel capacity and the source rate-distortion function are available for a wide class of channels and sources, the luxury of being able to compute exactly (in polynomial time) the non-asymptotic fundamental limit of interest is rarely affordable. Notable exceptions where the non-asymptotic fundamental limit is indeed computable are almost lossless source coding [5], [6], and JSCC over matched source-channel pairs [7]. In general, however, one can at most hope to obtain bounds and approximations to the information-theoretic non-asymptotic fundamental limits.

Although non-asymptotic bounds can be distilled from classical proofs of coding theorems, these bounds are rarely satisfyingly tight in the non-asymptotic regime, as studied in [8], [9] in the contexts of channel coding and lossy source coding, respectively. For the JSCC problem, the classical converse is based on the mutual information data processing inequality, while the classical achievability scheme uses separate source/channel coding (SSCC), in which the channel coding block and the source coding block are optimized separately without knowledge of each other. These conventional approaches lead to disappointingly weak non-asymptotic bounds. In particular, SSCC can be rather suboptimal non-asymptotically. An accurate finite blocklength analysis therefore calls for novel upper and lower bounds that sandwich tightly the non-asymptotic fundamental limit. Such bounds were shown in [8] for the channel coding problem and in [9] for the source coding problem. In this paper, we derive new tight bounds for the JSCC problem, which hold in full generality, without any assumptions on the source alphabet, stationarity or memorylessness.

While numerical evaluation of the non-asymptotic upper and lower bounds bears great practical interest (for example, to decide how suboptimal with respect to the information-theoretic limit a given blocklength-n code is), such bounds usually involve cumbersome expressions that offer scant conceptual insight. Somewhat ironically, to get an elegant, insightful approximation of the non-asymptotic fundamental limit, one must resort to an asymptotic analysis of these non-asymptotic bounds. Such asymptotic analysis must be finer than that based on the law of large numbers, which suffices to obtain the asymptotic fundamental limit but fails to provide any estimate of the speed of convergence to that limit. There are two complementary approaches to a finer asymptotic analysis: the large deviations analysis which leads to error exponents, and the Gaussian approximation analysis which leads to dispersion. The error exponent approximation and the Gaussian approximation to the non-asymptotic fundamental limit are tight in different operational regimes. In the former, a

rate which is strictly suboptimal with respect to the asymptotic fundamental limit is fixed, and the error exponent measures the exponential decay of the error probability to 0 as the blocklength increases. The error exponent approximation is tight if the error probability a system can tolerate is extremely small. However, already for probability of error as low as 10^{-6} to 10^{-1}, which is the operational regime for many high data rate applications, the Gaussian approximation, which gives the optimal rate achievable at a given error probability as a function of blocklength, is tight [8], [9]. In the channel coding problem, the Gaussian approximation of $R^\star(n, \epsilon)$, the maximum achievable finite blocklength coding rate at blocklength n and error probability ϵ, is given by, for finite alphabet stationary memoryless channels [8],

$$ nR^\star(n, \epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n) \qquad (1) $$

where C and V are the channel capacity and dispersion, respectively. In the lossy source coding problem, the Gaussian approximation of $R^\star(k, d, \epsilon)$, the minimum achievable finite blocklength coding rate at blocklength k and probability ϵ of exceeding fidelity d, is given by, for stationary memoryless sources [9],

$$ kR^\star(k, d, \epsilon) = kR(d) + \sqrt{k\mathcal{V}(d)}Q^{-1}(\epsilon) + O(\log k) \quad (2) $$

where $R(d)$ and $\mathcal{V}(d)$ are the rate-distortion and the rate-dispersion functions, respectively.

For a given code, the excess distortion constraint, which is the figure of merit in this paper as well as in [9], is, in a way, more fundamental than the average distortion constraint, because varying d over its entire range and evaluating the probability of exceeding d gives full information about the distribution (and not just its mean) of the distortion incurred at the decoder output. Following the philosophy of [8], [9], in this paper we perform the Gaussian approximation analysis of our new bounds to show that k, the maximum number of source symbols transmissible using a given channel blocklength n, must satisfy

$$ nC - kR(d) = \sqrt{nV + k\mathcal{V}(d)}Q^{-1}(\epsilon) + O(\log n) \quad (3) $$

under the fidelity constraint of exceeding a given distortion level d with probability ϵ. In contrast, if, following the SSCC paradigm, we just concatenate the channel code in (1) and the source code in (2), we obtain

$$ nC - kR(d) \leq \min_{\eta + \zeta \leq \epsilon} \left\{ \sqrt{nV}Q^{-1}(\eta) + \sqrt{k\mathcal{V}(d)}Q^{-1}(\zeta) \right\} $$
$$ + O(\log n) \qquad (4) $$

which is usually strictly suboptimal with respect to (3).

In addition to deriving new general achievability and converse bounds for JSCC and performing their Gaussian approximation analysis, in this paper we revisit the dilemma of whether one should or should not code when operating under delay constraints. Gastpar et al. [7] gave a set of necessary and sufficient conditions on the source, its distortion measure, the channel and its cost function in order for symbol-by-symbol transmission to attain the minimum average distortion. In these curious cases, the source and the channel are probabilistically matched. In the absence of channel cost constraints, we show

that whenever the source and the channel are probabilistically matched so that symbol-by-symbol coding achieves the minimum average distortion, it also achieves the dispersion of joint source-channel coding. Moreover, even in the absence of such a match between the source and the channel, symbol-by-symbol transmission, though asymptotically suboptimal, might outperform in the non-asymptotic regime not only separate source-channel coding but also our random-coding achievability bound.

Prior research relating to finite blocklength analysis of JSCC includes the work of Csiszár [10], [11] who demonstrated that the error exponent of joint source-channel coding outperforms that of separate source-channel coding. For discrete source-channel pairs with average distortion criterion, Pilc's achievability bound [12], [13] applies. For the transmission of a Gaussian source over a discrete channel under the average mean square error constraint, Wyner's achievability bound [14], [15] applies. Non-asymptotic achievability and converse bounds for a graph-theoretic model of JSCC have been obtained by Csiszár [16]. Most recently, Tauste Campo et al. [17] showed a number of finite-blocklength random-coding bounds applicable to the almost-lossless JSCC setup, while Wang et al. [18] found the dispersion of JSCC for sources and channels with finite alphabets.

The rest of the paper is organized as follows. Section II summarizes basic definitions and notation. Sections III and IV introduce the new converse and achievability bounds to the maximum achievable coding rate, respectively. A Gaussian approximation analysis of the new bounds is presented in Section V. The evaluation of the bounds and the approximation is performed for two important special cases: the transmission of a binary memoryless source (BMS) over a binary symmetric channel (BSC) with bit error rate distortion (Section VI) and the transmission of a Gaussian memoryless source (GMS) with mean-square error distortion over an AWGN channel with a total power constraint (Section VII). Section VIII focuses on symbol-by-symbol transmission.

II. DEFINITIONS

A lossy source-channel code is a pair of (possibly randomized) mappings $\mathsf{f} \colon \mathcal{M} \mapsto \mathcal{X}$ and $\mathsf{g} \colon \mathcal{Y} \mapsto \widehat{\mathcal{M}}$. A distortion measure $\mathsf{d} \colon \mathcal{M} \times \widehat{\mathcal{M}} \mapsto [0, +\infty]$ is used to quantify the performance of the lossy code. A cost function $\mathsf{c} \colon \mathcal{X} \mapsto [0, +\infty]$ may be imposed on the channel inputs. The channel is used without feedback.

Definition 1. *The pair (f, g) is a (d, ϵ, α) lossy source-channel code for $\{\mathcal{M}, \mathcal{X}, \mathcal{Y}, \widehat{\mathcal{M}}, P_S, \mathsf{d}, P_{Y|X}, \mathsf{c}\}$ if $\mathbb{P}[\mathsf{d}(S, \mathsf{g}(Y)) > d] \leq \epsilon$ and either $\mathbb{E}[\mathsf{c}(X)] \leq \alpha$ (average cost constraint) or $\mathsf{c}(X) \leq \alpha$ a.s. (maximal cost constraint), where $\mathsf{f}(S) = X$ (see Fig. 1). In the absence of an input cost constraint we simplify the terminology and refer to the code as (d, ϵ) lossy source-channel code.*

The special case $d = 0$ and $\mathsf{d}(s, z) = 1\{s \neq z\}$ corresponds to almost-lossless compression. If, in addition, P_S is equiprobable on an alphabet of cardinality $|\mathcal{M}| = |\widehat{\mathcal{M}}| = M$, a $(0, \epsilon, \alpha)$ code in Definition 1 corresponds to an (M, ϵ, α) channel code
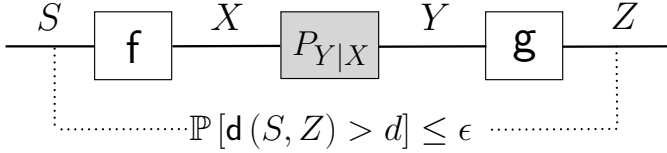
Fig. 1. A (d, ϵ) joint source-channel code.

(i.e. a code with M codewords and average error probability ϵ and cost α). On the other hand, if $P_{Y|X}$ is an identity mapping on an alphabet of cardinality M without cost constraints, a (d, ϵ) code in Definition 1 corresponds to an (M, d, ϵ) lossy compression code (as e.g. defined in [9]).

As our bounds in Sections III and IV do not foist a Cartesian structure on the underlying alphabets, we state them in the one-shot paradigm of Definition 1. When we apply those bounds to the block coding setting, transmitted objects indeed become vectors, and Definition 2 below comes into play.

Definition 2. *In the conventional fixed-to-fixed (or block) setting in which $\mathcal{X}$ and $\mathcal{Y}$ are the $n-$fold Cartesian products of alphabets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{M}$ and $\widehat{\mathcal{M}}$ are the $k-$fold Cartesian products of alphabets $\mathcal{S}$ and $\hat{\mathcal{S}}$, and $\mathsf{d}_k \colon \mathcal{S}^k \times \hat{\mathcal{S}}^k \mapsto [0, +\infty]$, $\mathsf{c}_n \colon \mathcal{A}^n \mapsto [0, +\infty]$, a (d, ϵ, α) code for $\{\mathcal{S}^k,\ \mathcal{A}^n,\ \mathcal{B}^n,\ \hat{\mathcal{S}}^k,\ P_{S^k},\ \mathsf{d}_k,\ P_{Y^n|X^n},\ \mathsf{c}_n\}$ is called a $(k, n, d, \epsilon, \alpha)$ code (or a (k, n, d, ϵ) code if there is no cost constraint).*

Definition 3. *Fix ϵ, d, α and the channel blocklength n. The maximum achievable source blocklength and coding rate (source symbols per channel use) are defined by, respectively*

$$k^\star(n, d, \epsilon, \alpha) = \sup\{k \colon \exists (k, n, d, \epsilon, \alpha)\ code\} \quad (5)$$

$$R(n, d, \epsilon, \alpha) = \frac{1}{n} k^\star(n, d, \epsilon, \alpha) \quad (6)$$

Alternatively, fix ϵ, α, source blocklength k and channel blocklength n. The minimum achievable excess distortion is defined by

$$D(k, n, \epsilon, \alpha) = \inf\{d \colon \exists (k, n, d, \epsilon, \alpha)\ code\} \quad (7)$$

Denote, for a given $P_{Y|X}$ and a cost function $\mathsf{c} \colon \mathcal{X} \mapsto [0, +\infty]$,

$$\mathbb{C}(\alpha) = \sup_{\substack{P_X \colon \\ \mathbb{E}[\mathsf{c}(X)] \leq \alpha}} I(X; Y) \quad (8)$$

and, for a given P_S and a distortion measure $\mathsf{d} \colon \mathcal{M} \times \widehat{\mathcal{M}} \mapsto [0, +\infty]$,

$$\mathbb{R}_S(d) = \inf_{\substack{P_{Z|S} \colon \\ \mathbb{E}[\mathsf{d}(S, Z)] \leq d}} I(S; Z) \quad (9)$$

We impose the following basic restrictions on $P_{Y|X}$, P_S, the input-cost function and the distortion measure:

(a) $\mathbb{R}_S(d)$ is finite for some d, i.e. $d_{\min} < \infty$, where

$$d_{\min} = \inf\{d \colon \mathbb{R}_S(d) < \infty\}; \quad (10)$$

(b) The infimum in (9) is achieved by a unique $P_{Z^\star|S}$;
(c) The supremum in (8) is achieved by a unique $P_{X^\star}$.

The dispersion, which serves to quantify the penalty on the rate of the best JSCC code induced by the finite blocklength, is defined as follows.

Definition 4. *Fix α and $d \geq d_{\min}$. The rate-dispersion function of joint source-channel coding (source samples squared per channel use) is defined as*

$$\mathcal{V}(d, \alpha) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{n\left(\frac{C(\alpha)}{R(d)} - R(n, d, \epsilon, \alpha)\right)^2}{2 \log_e \frac{1}{\epsilon}} \quad (11)$$

where $C(\alpha)$ and $R(d)$ are the channel capacity-cost and source rate-distortion functions, respectively.[1]

The distortion-dispersion function of joint source-channel coding is defined as

$$\mathcal{W}(R, \alpha) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{n\left(D\left(\frac{C(\alpha)}{R}\right) - D(nR, n, \epsilon, \alpha)\right)^2}{2 \log_e \frac{1}{\epsilon}} \quad (12)$$

where $D(\cdot)$ is the distortion-rate function of the source.

If there is no cost constraint, we will simplify notation by dropping α from (5), (6), (7), (8), (11) and (12).

Definition 5 (d$-$tilted information [9]). *For $d > d_{\min}$, the $\mathsf{d}-$tilted information in s is defined as*[2]

$$\jmath_S(s, d) = \log \frac{1}{\mathbb{E}[\exp(\lambda^\star d - \lambda^\star \mathsf{d}(s, Z^\star))]} \quad (13)$$

where the expectation is with respect to $P_{Z^\star}$, i.e. the unconditional distribution of the reproduction random variable that achieves the infimum in (9), and

$$\lambda^\star = -\mathbb{R}_S'(d) \quad (14)$$

The following properties of $\mathsf{d}-$tilted information, proven in [19], are used in the sequel.

$$\jmath_S(s, d) = \imath_{S;Z^\star}(s; z) + \lambda^\star \mathsf{d}(s, z) - \lambda^\star d \quad (15)$$

$$\mathbb{E}[\jmath_S(s, d)] = \mathbb{R}_S(d) \quad (16)$$

$$\mathbb{E}[\exp(\lambda^\star d - \lambda^\star \mathsf{d}(S, z) + \jmath_S(S, d))] \leq 1 \quad (17)$$

where (15) holds for $P_Z^\star$-almost every z, while (17) holds for all $z \in \widehat{\mathcal{M}}$, and

$$\imath_{S;Z}(s; z) = \log \frac{dP_{Z|S=s}}{dP_Z}(z) \quad (18)$$

denotes the information density of the joint distribution P_{SZ} at (s, z). We can define the right side of (18) for a given $(P_{Z|S}, P_Z)$ even if there is no P_S such that the marginal of $P_S P_{Z|S}$ is P_Z. We use the same notation $\imath_{S;Z}$ for that more general function. To extend Definition 5 to the lossless case, for discrete random variables we define 0-tilted information as

$$\jmath_S(s, 0) = \imath_S(s) \quad (19)$$

---

[1] While for memoryless sources and channels, $C(\alpha) = \mathbb{C}(\alpha)$ and $R(d) = \mathbb{R}_S(d)$ given by (8) and (9) evaluated with single-letter distributions, it is important to distinguish between the operational definitions and the extremal mutual information quantities, since the core results in this paper allow for memory.

[2] All log's and exp's are in an arbitrary common base.

where

$$\imath_S(s) = \log \frac{1}{P_S(s)} \qquad (20)$$

is the information in outcome $s \in \mathcal{M}$.

The distortion d-ball centered at $s \in \mathcal{M}$ is denoted by

$$B_d(s) = \{z \in \widehat{\mathcal{M}} : \mathsf{d}(s,z) \leq d\}. \qquad (21)$$

Given $(P_X, P_{Y|X})$, we write $P_X \to P_{Y|X} \to P_Y$ to indicate that P_Y is the marginal of $P_X P_{Y|X}$, i.e. $P_Y(y) = \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) P_X(x)$. [3]

So as not to clutter notation, in Sections III and IV we assume that there are no cost constraints. However, all results in those sections generalize to the case of a maximal cost constraint by considering X whose distribution is supported on the subset of allowable channel inputs:

$$\mathcal{F}(\alpha) = \{x \in \mathcal{X} : \mathsf{c}(x) \leq \alpha\} \qquad (22)$$

rather than the entire channel input alphabet $\mathcal{X}$.

III. CONVERSES

A. Converses via d-tilted information

Our first result is a general converse bound.

Theorem 1 (Converse). *The existence of a (d, ϵ) code for S and $P_{Y|X}$ requires that*

$$\epsilon \geq \inf_{P_{X|S}} \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma \right] \right.$$
$$\left. - \exp(-\gamma) \right\} \qquad (23)$$
$$\geq \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \mathbb{E}\left[\inf_{x \in \mathcal{X}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(x;Y) \geq \gamma \mid S \right] \right] \right.$$
$$\left. - \exp(-\gamma) \right\} \qquad (24)$$

where in (23), $S - X - Y$, and the conditional probability in (24) is with respect to Y distributed according to $P_{Y|X=x}$ (independent of S), and

$$\imath_{X;\bar{Y}}(x;y) = \log \frac{dP_{Y|X=x}}{dP_{\bar{Y}}}(y) \qquad (25)$$

Proof: Fix γ and the (d, ϵ) code $(P_{X|S}, P_{Z|Y})$. Fix an arbitrary probability measure $P_{\bar{Y}}$ on $\mathcal{Y}$. Let $P_{\bar{Y}} \to P_{Z|Y} \to$

$P_{\bar{Z}}$. We can write the probability in the right side of (23) as

$$\mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma \right]$$
$$= \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma, \mathsf{d}(S;Z) > d \right]$$
$$+ \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma, \mathsf{d}(S;Z) \leq d \right] \qquad (26)$$
$$\leq \epsilon$$
$$+ \sum_{s \in \mathcal{M}} P_S(s) \sum_{x \in \mathcal{X}} P_{X|S}(x|s) \sum_{y \in \mathcal{Y}} \sum_{z \in B_d(s)} P_{Z|Y}(z|y)$$
$$\cdot P_{Y|X}(y|x) 1\left\{ P_{Y|X}(y|x) \leq P_{\bar{Y}}(y) \exp\left(\jmath_S(s,d) - \gamma\right) \right\} \qquad (27)$$
$$\leq \epsilon + \exp(-\gamma) \sum_{s \in \mathcal{M}} P_S(s) \exp(\jmath_S(s,d)) \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y)$$
$$\cdot \sum_{z \in B_d(s)} P_{Z|Y}(z|y) \sum_{x \in \mathcal{X}} P_{X|S}(x|s) \qquad (28)$$
$$= \epsilon + \exp(-\gamma) \sum_{s \in \mathcal{M}} P_S(s) \exp(\jmath_S(s,d)) \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y)$$
$$\cdot \sum_{z \in B_d(s)} P_{Z|Y}(z|y) \qquad (29)$$
$$= \epsilon + \exp(-\gamma) \sum_{s \in \mathcal{M}} P_S(s) \exp(\jmath_S(s,d)) P_{\bar{Z}}(B_d(s)) \qquad (30)$$
$$\leq \epsilon + \exp(-\gamma) \sum_{z \in \widehat{\mathcal{M}}} P_{\bar{Z}}(z) \sum_{s \in \mathcal{M}} P_S(s)$$
$$\cdot \exp\left(\jmath_S(s,d) + \lambda^\star d - \lambda^\star \mathsf{d}(s,z)\right) \qquad (31)$$
$$\leq \epsilon + \exp(-\gamma) \qquad (32)$$

where (32) is due to (17). Optimizing over $\gamma > 0$ and $P_{\bar{Y}}$, we get the best possible bound for a given encoder $P_{X|S}$. To obtain a code-independent converse, we simply choose $P_{X|S}$ that gives the weakest bound, and (23) follows. To show (24), we weaken (23) as

$$\epsilon \geq \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma \right] \right.$$
$$\left. - \exp(-\gamma) \right\} \qquad (33)$$

and observe that for any $P_{\bar{Y}}$,

$$\inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma \right]$$
$$= \sum_{s \in \mathcal{M}} P_S(s) \inf_{P_{X|S=s}} \sum_{x \in \mathcal{X}} P_{X|S}(x|s)$$
$$\cdot \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) 1\left\{ \jmath_S(s,d) - \imath_{X;\bar{Y}}(x;y) \geq \gamma \right\} \qquad (34)$$
$$= \sum_{s \in \mathcal{M}} P_S(s)$$
$$\cdot \inf_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) 1\left\{ \jmath_S(s,d) - \imath_{X;\bar{Y}}(x;y) \geq \gamma \right\} \qquad (35)$$
$$= \mathbb{E}\left[\inf_{x \in \mathcal{X}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(x;Y) \geq \gamma \mid S \right] \right] \qquad (36)$$

$\blacksquare$

---

[3]We write summations over alphabets for simplicity. Unless stated otherwise, all our results hold for abstract probability spaces.

An immediate corollary to Theorem 1 is the following result.

Theorem 2 (Converse). *Assume that there exists a distribution $P_{\bar{Y}}$ such that the distribution of $\imath_{X;\bar{Y}}(x;Y)$ (according to $P_{Y|X=x}$) does not depend on the choice of $x \in \mathcal{X}$. If a (d,ϵ) code for S and $P_{Y|X}$ exists, then*

$$\epsilon \geq \sup_{\gamma>0} \left\{ \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(x;Y) \geq \gamma\right] - \exp\left(-\gamma\right) \right\} \tag{37}$$

for an arbitrary $x \in \mathcal{X}$. The probability measure $\mathbb{P}$ in (37) is generated by $P_S P_{Y|X=x}$.

Proof: Under the assumption, the conditional probability in the right side of (24) is the same regardless of the choice of $x \in \mathcal{X}$. ∎

The next result generalizes Theorem 1. When we apply Theorem 3 in Section V to find the dispersion of JSCC, we will let T be the number of channel input types, and we will let W be the type of the channel input block. If $T = 1$, Theorem 3 reduces to Theorem 1.

Theorem 3 (Converse). *The existence of a (d,ϵ) code for S and $P_{Y|X}$ requires that*

$$\epsilon \geq \inf_{P_{X|S}} \max_{\gamma>0,T} \left\{ -T \exp\left(-\gamma\right) \right.$$
$$+ \sup_{\substack{\bar{Y},W: \\ S-(X,W)-Y}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}|W}(X;Y|W) \geq \gamma\right] \right\} \tag{38}$$

$$\geq \max_{\gamma>0,T} \left\{ -T \exp\left(-\gamma\right) \right.$$
$$+ \sup_{\bar{Y},W} \mathbb{E}\left[\inf_{x\in\mathcal{X}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}|W}(x;Y|W) \geq \gamma \mid S\right] \right] \right\} \tag{39}$$

where T is a positive integer, the random variable W takes values on $\{1,\ldots,T\}$, and

$$\imath_{X;\bar{Y}|W}(x;y|t) = \log \frac{P_{Y|X=x,W=t}}{P_{\bar{Y}|W=t}}(y) \tag{40}$$

and in (39), the probability measure is generated by $P_S P_{W|X=x} P_{Y|X=x,W}$.

Proof: Fix a possibly randomized (d,ϵ) code $\{P_{X|S}, P_{Z|Y}\}$, a positive scalar γ, a positive integer T, an auxiliary random variable W that satisfies $S - (X,W) - Y$, and a conditional probability distribution $P_{\bar{Y}|W}: \{1,\ldots T\} \mapsto \mathcal{Y}$. Let $P_{\bar{Y}|W=t} \to P_{Z|Y} \to P_{\bar{Z}|W=t}$, i.e. $P_{\bar{Z}|W=t}(z) = \sum_{y\in\mathcal{Y}} P_{Z|Y}(z|y) P_{\bar{Y}|W=t}(y)$, for all t.

Write

$$\mathbb{P}\left[\jmath_S(S,d) - \imath_{X;Y|W}(X;Y|W) \geq \gamma\right]$$

$$\leq \epsilon + \sum_{s\in\mathcal{M}} P_S(s) \sum_{t=1}^{T} P_{W|S}(t|s) \sum_{x\in\mathcal{X}} P_{X|S,W}(x|s,t)$$
$$\cdot \sum_{y\in\mathcal{Y}} P_{Y|X,W}(y|x,t) \sum_{z\in B_d(s)} P_{Z|Y}(z|y)$$
$$\cdot \mathbb{1}\left\{P_{Y|X,W}(y|x,t) \leq P_{\bar{Y}|W=t}(y) \exp\left(\jmath_S(s,d) - \gamma\right)\right\} \tag{41}$$

$$\leq \epsilon + \exp\left(-\gamma\right) \sum_{s\in\mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) \sum_{t=1}^{T} P_{W|S}(t|s)$$
$$\cdot \sum_{y\in\mathcal{Y}} P_{\bar{Y}|W}(y|t) \sum_{z\in B_d(s)} P_{Z|Y}(z|y) \sum_{x\in\mathcal{X}} P_{X|S,W}(x|s,t) \tag{42}$$

$$\leq \epsilon + \exp\left(-\gamma\right) \sum_{t=1}^{T} \sum_{s\in\mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) \sum_{y\in\mathcal{Y}} P_{\bar{Y}|W}(y|t)$$
$$\cdot \sum_{z\in B_d(s)} P_{Z|Y}(z|y) \tag{43}$$

$$\leq \epsilon + \exp\left(-\gamma\right) \sum_{t=1}^{T} \sum_{s\in\mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) P_{\bar{Z}|W=t}(B_d(s)) \tag{44}$$

$$\leq \epsilon + \exp\left(-\gamma\right) \sum_{t=1}^{T} \sum_{s\in\mathcal{M}} P_S(s) \sum_{z\in\widehat{\mathcal{M}}} P_{\bar{Z}|W=t}(z)$$
$$\cdot \exp\left(\jmath_S(s,d) + \lambda^\star d - \lambda^\star \mathsf{d}(s,z)\right) \tag{45}$$
$$\leq \epsilon + T \exp\left(-\gamma\right) \tag{46}$$

where (46) is due to (17). Optimizing over γ, T and the distributions of the auxiliary random variables $\bar{Y}$ and W, we obtain the best possible bound for a given encoder $P_{X|S}$. To obtain a code-independent converse, we simply choose $P_{X|S}$ that gives the weakest bound, and (38) follows. To show (39), we weaken (38) by restricting the sup to W satisfying $S - X - W$ and changing the order of inf and sup as follows:

$$\max_{\gamma>0,T} \sup_{\substack{\bar{Y},W: \\ S-(X,W)-Y \\ S-X-W}} \inf_{P_{X|S}} \tag{47}$$

Observe that for any legitimate choice of $\bar{Y}$ and W,

$$\inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}|W}(X;Y|W) \geq \gamma\right] \tag{48}$$

$$= \sum_{s\in\mathcal{M}} P_S(s) \inf_{P_{X|S=s}} \sum_{x\in\mathcal{X}} P_{X|S}(x|s) \sum_{t=1}^{T} P_{W|X}(t|x)$$
$$\cdot \sum_{y\in\mathcal{Y}} P_{Y|X,W}(y|x,t) \mathbb{1}\left\{\jmath_S(s,d) - \imath_{X;\bar{Y}|W}(x;y|t) \geq \gamma\right\} \tag{49}$$

$$= \sum_{s\in\mathcal{M}} P_S(s) \inf_{x\in\mathcal{X}} \sum_{t=1}^{T} P_{W|X}(t|x) \sum_{y\in\mathcal{Y}} P_{Y|X,W}(y|x,t)$$
$$\cdot \mathbb{1}\left\{\jmath_S(s,d) - \imath_{X;\bar{Y}|W}(x;y|t) \geq \gamma\right\} \tag{50}$$

which is equal to the expectation on the right side of (39). ∎

Remark 1. Theorems 1, 2 and 3 still hold in the case $d = 0$ and $d(x, y) = 1\{x \neq y\}$, which corresponds to almost-lossless data compression. Indeed, recalling (19), it is easy to see that the proof of Theorem 1 applies, skipping the now unnecessary step (31), and, therefore, (23) reduces to

$$\epsilon \geq \inf_{P_{X|S}} \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \mathbb{P}\left[\imath_S(S) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma\right] - \exp\left(-\gamma\right) \right\} \quad (51)$$

Similar modification can be applied to the proof of Theorem 3.

Remark 2. Our converse for lossy source coding in [9, Theorem 7] can be viewed as a particular case of the result in Theorem 2. Indeed, if $\mathcal{X} = \mathcal{Y} = \{1, \ldots, M\}$ and $P_{Y|X}(m|m) = 1$, $P_Y(1) = \ldots = P_Y(M) = \frac{1}{M}$, then (37) becomes

$$\epsilon \geq \sup_{\gamma > 0} \mathbb{P}\left[\jmath_S(S, d) \geq \log M + \gamma\right] - \exp\left(-\gamma\right) \quad (52)$$

which is precisely [9, Theorem 7].

B. Converses via hypothesis testing and list decoding

To show a joint source-channel converse in [11], Csiszár used a list decoder, which outputs a list of L elements drawn from $\mathcal{M}$. While traditionally list decoding has only been considered in the context of finite alphabet sources, we generalize the setting to sources with abstract alphabets. In our setup, the encoder is the random transformation $P_{X|S}$, and the decoder is defined as follows.

Definition 6 (List decoder). *Let L be a positive real number, and let Q_S be a measure on $\mathcal{M}$. An (L, Q_S) list decoder is a random transformation $P_{\tilde{S}|Y}$, where $\tilde{S}$ takes values on Q_S-measurable sets with Q_S-measure not exceeding L:*

$$Q_S\left(\tilde{S}\right) \leq L \quad (53)$$

Even though we keep the standard "list" terminology, the decoder output need not be a finite or countably infinite set. The error probability with this type of list decoding is the probability that the source outcome S does not belong to the decoder output list for Y:

$$1 - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{\tilde{s} \in \mathcal{M}^{(L)}} \sum_{s \in \tilde{s}} P_{\tilde{S}|Y}(\tilde{s}|y) P_{Y|X}(y|x) P_{X|S}(x|s) P_S(s) \quad (54)$$

where $\mathcal{M}^{(L)}$ is the set of all Q_S-measurable subsets of $\mathcal{M}$ with Q_S-measure not exceeding L.

Definition 7 (List code). *An (ϵ, L, Q_S) list code is a pair of random transformations $(P_{X|S}, P_{\tilde{S}|Y})$ such that (53) holds and the list error probability (54) does not exceed ϵ.*

Of course, letting $Q_S = U_S$, where U_S is the counting measure on $\mathcal{M}$, we recover the conventional list decoder definition where the smallest scalar that satisfies (53) is an integer. The almost-lossless JSCC setting ($d = 0$) in Definition 1 corresponds to $L = 1$, $Q_S = U_S$. If the source is analog

(has a continuous distribution), it is reasonable to let Q_S be the Lebesgue measure.

Any converse for list decoding implies a converse for conventional decoding. To see why, observe that any (d, ϵ) lossy code can be converted to a list code with list error probability not exceeding ϵ by feeding the lossy decoder output to a function that outputs the set of all source outcomes s within distortion d from the output $z \in \widehat{\mathcal{M}}$ of the original lossy decoder. In this sense, the set of all (d, ϵ) lossy codes is included in the set of all list codes with list error probability $\leq \epsilon$ and list size

$$L = \max_{z \in \widehat{\mathcal{M}}} Q_S\left(\{s \colon \mathsf{d}(s, z) \leq d\}\right) \quad (55)$$

Denote by

$$\beta_\alpha(P, Q) = \min_{\substack{P_{W|X} : \\ \mathbb{P}[W=1] \geq \alpha}} \mathbb{Q}\left[W = 1\right] \quad (56)$$

the optimal performance achievable among all randomized tests $P_{W|X} \colon \mathcal{X} \to \{0, 1\}$ between probability distributions P and Q on $\mathcal{X}$ (1 indicates that the test chooses P).[4] In fact, Q need not be a probability measure, it just needs to be σ-finite in order for the Neyman-Pearson lemma and related results to hold.

The hypothesis testing converse for channel coding [8, Theorem 27] can be generalized to joint source-channel coding with list decoding as follows.

Theorem 4 (Converse). *Fix P_S and $P_{Y|X}$, and let Q_S be a σ-finite measure. The existence of an (ϵ, L, Q_S) list code requires that*

$$\inf_{P_{X|S}} \sup_{P_{\bar{Y}}} \beta_{1-\epsilon}(P_S P_{X|S} P_{Y|X}, Q_S P_{X|S} P_{\bar{Y}}) \leq L \quad (57)$$

where the supremum is over all probability measures $P_{\bar{Y}}$ defined on the channel output alphabet $\mathcal{Y}$.

Proof: Fix Q_S, the encoder $P_{X|S}$, and an auxiliary σ-finite conditional measure $Q_{Y|XS}$. Consider the (not necessarily optimal) test for deciding between $P_{SXY} = P_S P_{X|S} P_{Y|X}$ and $Q_{SXY} = Q_S P_{X|S} Q_{Y|XS}$ which chooses P_{SXY} if S belongs to the decoder output list. Note that this is a hypothetical test, which has access to both the source outcome and the decoder output.

According to $\mathbb{P}$, the probability measure generated by P_{SXY}, the probability that the test chooses P_{SXY} is given by

$$\mathbb{P}\left[S \in \tilde{S}\right] \geq 1 - \epsilon \quad (58)$$

Since $\mathbb{Q}\left[S \in \tilde{S}\right]$ is the measure of the event that the test chooses P_{SXY} when Q_{SXY} is true, and the optimal test cannot perform worse than the possibly suboptimal one that we selected, it follows that

$$\beta_{1-\epsilon}(P_S P_{X|S} P_{Y|X}, Q_S P_{X|S} Q_{Y|XS}) \leq \mathbb{Q}\left[S \in \tilde{S}\right] \quad (59)$$

---

[4]Throughout, P, Q denote distributions, whereas $\mathbb{P}$, $\mathbb{Q}$ are used for the corresponding probabilities of events on the underlying probability space.

Now, fix an arbitrary probability measure $P_{\bar{Y}}$ on $\mathcal{Y}$. Choosing $Q_{Y|XS} = P_{\bar{Y}}$, the inequality in (59) can be weakened as follows.

$$\mathbb{Q}\left[S \in \tilde{S}\right]$$

$$= \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y) \sum_{\tilde{s} \in \mathcal{M}^{(L)}} P_{\tilde{S}|Y}(\tilde{s}|y) \sum_{s \in \tilde{s}} Q_S(s) \sum_{x \in \mathcal{X}} P_{X|S}(x|s) \tag{60}$$

$$= \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y) \sum_{\tilde{s} \in \mathcal{M}^{(L)}} P_{\tilde{S}|Y}(\tilde{s}|y) \sum_{s \in \tilde{s}} Q_S(s) \tag{61}$$

$$\leq \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y) \sum_{\tilde{s} \in \mathcal{M}^{(L)}} P_{\tilde{S}|Y}(\tilde{s}|y) L \tag{62}$$

$$= L \tag{63}$$

Optimizing the bound over $P_{\bar{Y}}$ and choosing $P_{X|S}$ that yields the weakest bound in order to obtain a code-independent converse, (57) follows. ∎

Remark 3. Similar to how Wolfowitz's converse for channel coding can be obtained from the meta-converse for channel coding [8], the converse for almost-lossless joint source-channel coding in (51) can be obtained by appropriately weakening (57) with $L = 1$. Indeed, invoking [8]

$$\beta_\alpha(P, Q) \geq \frac{1}{\gamma}\left(\alpha - \mathbb{P}\left[\frac{dP}{dQ} > \gamma\right]\right) \tag{64}$$

and letting $Q_S = U_S$ in (57), where U_S is the counting measure on $\mathcal{M}$, we have

$$1 \geq \inf_{P_{X|S}} \sup_{P_{\bar{Y}}} \beta_{1-\epsilon}(P_S P_{X|S} P_{Y|X}, U_S P_{X|S} P_{\bar{Y}}) \tag{65}$$

$$\geq \inf_{P_{X|S}} \sup_{P_{\bar{Y}}} \sup_{\gamma > 0} \frac{1}{\gamma}\left(1 - \epsilon - \mathbb{P}\left[\imath_{X;\bar{Y}}(X;Y) - \imath_S(S) > \log\gamma\right]\right) \tag{66}$$

which upon rearranging yields (51).

In general, computing the infimum in (57) is challenging. However, if the channel is symmetric (in a sense formalized in the next result), $\beta_{1-\epsilon}(P_S P_{X|S} P_{Y|X}, U_S P_{X|S} P_{\bar{Y}})$ is independent of $P_{X|S}$.

Theorem 5 (Converse). *Fix a probability measure $P_{\bar{Y}}$. Assume that the distribution of $\imath_{X;\bar{Y}}(x; Y)$ does not depend on $x \in \mathcal{X}$ under either $P_{Y|X=x}$ or $P_{\bar{Y}}$. Then, the existence of an (ϵ, L, Q_S) list code requires that*

$$\beta_{1-\epsilon}(P_S P_{Y|X=x}, Q_S P_{\bar{Y}}) \leq L \tag{67}$$

where $x \in \mathcal{X}$ is arbitrary.

Proof: The Neyman-Pearson lemma (e.g. [20]) implies that the outcome of the optimum binary hypothesis test between P and Q only depends on the observation through $\frac{dP}{dQ}$. In particular, the optimum binary hypothesis test $W^\star$ for deciding between $P_S P_{X|S} P_{Y|X}$ and $Q_S P_{X|S} P_{\bar{Y}}$ satisfies

$$W^\star - (S, \imath_{X;\bar{Y}}(X;Y)) - (S, X, Y) \tag{68}$$

For all $s \in \mathcal{M}$, $x \in \mathcal{X}$, we have

$$\mathbb{P}\left[W^\star = 1|S = s, X = x\right]$$

$$= \mathbb{E}\left[\mathbb{P}\left[W^\star = 1|X = x, S = s, Y\right]\right] \tag{69}$$

$$= \mathbb{E}\left[\mathbb{P}\left[W^\star = 1|S = s, \imath_{X;\bar{Y}}(X;Y) = \imath_{X;\bar{Y}}(x;Y)\right]\right] \tag{70}$$

$$= \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) P_{W^\star|S, \imath_{X;\bar{Y}}(X;Y)}(1|s, \imath_{X;\bar{Y}}(x;y)) \tag{71}$$

$$= \mathbb{P}\left[W^\star = 1|S = s\right] \tag{72}$$

and

$$\mathbb{Q}\left[W^\star = 1|S = s, X = x\right] = \mathbb{Q}\left[W^\star = 1|S = s\right] \tag{73}$$

where

- (70) is due to (68),
- (71) uses the Markov property $S - X - Y$,
- (72) follows from the symmetry assumption on the distribution of $\imath_{X;\bar{Y}}(x, Y)$,
- (73) is obtained similarly to (71).

Since (72), (73) imply that the optimal test achieves the same performance (that is, the same $\mathbb{P}[W^\star = 1]$ and $\mathbb{Q}[W^\star = 1]$) regardless of $P_{X|S}$, we choose $P_{X|S} = 1_X(x)$ for some $x \in \mathcal{X}$ in the left side of (57) to obtain (67). ∎

Remark 4. In the case of finite channel input and output alphabets, the channel symmetry assumption of Theorem 5 holds, in particular, if the rows of the channel transition probability matrix are permutations of each other, and $P_{\bar{Y}^n}$ is the equiprobable distribution on the (n-dimensional) channel output alphabet, which, coincidentally, is also the capacity-achieving output distribution. For Gaussian channels with equal power constraint, which corresponds to requiring the channel inputs to lie on the power sphere, any spherically-symmetric $P_{\bar{Y}^n}$ satisfies the assumption of Theorem 5.

IV. ACHIEVABILITY

Given a source code $(\mathsf{f}_\mathsf{s}^{(M)}, \mathsf{g}_\mathsf{s}^{(M)})$ of size M, and a channel code $(\mathsf{f}_\mathsf{c}^{(M)}, \mathsf{g}_\mathsf{c}^{(M)})$ of size M, we may concatenate them to obtain the following sub-class of the source-channel codes introduced in Definition 1:

Definition 8. *An (M, d, ϵ) source-channel code is a (d, ϵ) source-channel code such that the encoder and decoder mappings satisfy*

$$\mathsf{f} = \mathsf{f}_\mathsf{c}^{(M)} \circ \mathsf{f}_\mathsf{s}^{(M)} \tag{74}$$

$$\mathsf{g} = \mathsf{g}_\mathsf{c}^{(M)} \circ \mathsf{g}_\mathsf{s}^{(M)} \tag{75}$$

where

$$\mathsf{f}_\mathsf{s}^{(M)}: \mathcal{M} \mapsto \{1, \ldots, M\} \tag{76}$$

$$\mathsf{f}_\mathsf{c}^{(M)}: \{1, \ldots, M\} \mapsto \mathcal{X} \tag{77}$$

$$\mathsf{g}_\mathsf{c}^{(M)}: \mathcal{Y} \mapsto \{1, \ldots, M\} \tag{78}$$

$$\mathsf{g}_\mathsf{s}^{(M)}: \{1, \ldots, M\} \mapsto \widehat{\mathcal{M}} \tag{79}$$

(see Fig. 2).

Note that an (M, d, ϵ) code is an $(M + 1, d, \epsilon)$ code.

The conventional separate source-channel coding paradigm corresponds to the special case of Definition 8 in which the
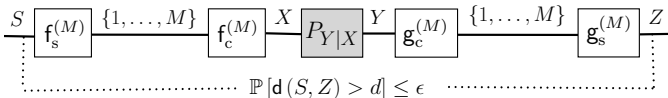
Fig. 2. An (M, d, ϵ) joint source-channel code.

source code $(\mathsf{f}_\mathsf{s}^{(M)}, \mathsf{g}_\mathsf{s}^{(M)})$ is chosen without knowledge of $P_{Y|X}$ and the channel code $(\mathsf{f}_\mathsf{c}^{(M)}, \mathsf{g}_\mathsf{c}^{(M)})$ is chosen without knowledge of P_S and the distortion measure d. A pair of source and channel codes is separation-optimal if the source code is chosen so as to minimize the distortion (average or excess) when there is no channel, whereas the channel code is chosen so as to minimize the worst-case (over source distributions) average error probability:

$$\max_{P_U} \mathbb{P}\left[U \neq \mathsf{g}_\mathsf{c}^{(M)}(Y) \right] \tag{80}$$

where $X = \mathsf{f}_\mathsf{c}^{(M)}(U)$ and U takes values on $\{1, \ldots, M\}$. If both the source and the channel code are chosen separation-optimally for their given sizes, the separation principle guarantees that under certain quite general conditions (which encompass the memoryless setting, see [21]) the asymptotic fundamental limit of joint source-channel coding is achievable. In the finite blocklength regime, however, such SSCC construction is, in general, only suboptimal. Within the SSCC paradigm, we can obtain an achievability result by further optimizing with respect to the choice of M:

Theorem 6 (Achievability, SSCC). *Fix $P_{Y|X}$, d and P_S. Denote by $\epsilon^\star(M)$ the minimum achievable worst-case average error probability among all transmission codes of size M, and the minimum achievable probability of exceeding distortion d with a source code of size M by $\epsilon^\star(M, d)$.*

Then, there exists a (d, ϵ) source-channel code with

$$\epsilon \leq \min_M \{\epsilon^\star(M) + \epsilon^\star(M, d)\} \tag{81}$$

Bounds on $\epsilon^\star(M)$ and $\epsilon^\star(M, d)$ have been obtained recently in [8] and [9], respectively.[5]

Definition 8 does not rule out choosing the source code based on the knowledge of $P_{Y|X}$ or the channel code based on the knowledge of P_S, d and d. One of the interesting conclusions in the present paper is that the optimal dispersion of JSCC is achievable within the class of (M, d, ϵ) source-channel codes introduced in Definition 8. However, the dispersion achieved by the conventional SSCC approach is in fact suboptimal.

To shed light on the reason behind the suboptimality of SSCC at finite blocklength despite its asymptotic optimality, we recall the reason SSCC achieves the asymptotic fundamental limit. The output of the optimum source encoder is, for large k, approximately equiprobable over a set of

---

[5]As the maximal (over source outputs) error probability cannot be lower than the worst-case error probability, the maximal error probability achievability bounds of [8] apply to bound $\epsilon^\star(M)$. Moreover, the random coding union (RCU) bound on average error probability of [8], although stated assuming equiprobable source, is oblivious to the distribution of the source and thus upper-bounds the worst-case average error probability $\epsilon^\star(M)$ as well.

roughly $\exp(kR(d))$ distinct messages, which allow to represent most of the source outcomes within distortion d. From the channel coding theorem we know that there exists a channel code that is capable of distinguishing, with high probability, $M = \exp(kR(d)) < \exp(nC)$ messages when equipped with the maximum likelihood decoder. Therefore, a simple concatenation of the source code and the channel code achieves vanishing probability of distortion exceeding d, for any $d > D\left(\frac{nC}{k}\right)$. However, at finite n, the output of the optimum source encoder need not be nearly equiprobable, so there is no reason to expect that a separated scheme employing a maximum-likelihood channel decoder, which does not exploit unequal message probabilities, would achieve near-optimal non-asymptotic performance. Indeed, in the non-asymptotic regime the gain afforded by taking into account the residual encoded source redundancy at the channel decoder is appreciable. The following achievability result, obtained using independent random source codes and random channel codes within the paradigm of Definition 8, capitalizes on this intuition.

Theorem 7 (Achievability). *There exists a (d, ϵ) source-channel code with*

$$\epsilon \leq \inf_{P_X, P_Z, P_{W|S}} \left\{ \mathbb{E}\left[\exp\left(-|\imath_{X;Y}(X;Y) - \log W|^+ \right) \right] \right.$$
$$\left. + \mathbb{E}\left[(1 - P_Z(B_d(S)))^W \right] \right\} \tag{82}$$

where the expectations are with respect to $P_S P_X P_{Y|X} P_Z P_{W|S}$ defined on $\mathcal{M} \times \mathcal{X} \times \mathcal{Y} \times \widehat{\mathcal{M}} \times \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers.

Proof: Fix a positive integer M. Fix a positive integer-valued random variable W that depends on other random variables only through S and that satisfies $W \leq M$. We will construct a code with separate encoders for source and channel and separate decoders for source and channel as in Definition 8. We will perform a random coding analysis by choosing random independent source and channel codes which will lead to the conclusion that there exists an (M, d, ϵ) code with error probability ϵ guaranteed in (82) with $W \leq M$. Observing that increasing M can only tighten the bound in (82) in which W is restricted to not exceed M, we will let $M \to \infty$ and conclude, by invoking the bounded convergence theorem, that the support of W in (82) need not be bounded.

Source Encoder. Given an ordered list of representation points $z^M = (z_1, \ldots, z_M) \in \widehat{\mathcal{M}}^M$, and having observed the source outcome s, the (probabilistic) source encoder generates W from $P_{W|S=s}$ and selects the lowest index $m \in \{1, \ldots, W\}$ such that s is within distance d of z_m. If no such index can be found, the source encoder outputs a pre-selected arbitrary index, e.g. M. Therefore,

$$\mathsf{f}_\mathsf{s}^{(M)}(s) = \begin{cases} \min\{m, W\} & \mathsf{d}(s, z_m) \leq d < \min_{i=1,\ldots,m-1} \mathsf{d}(s, z_i) \\ M & d < \min_{i=1,\ldots,W} \mathsf{d}(s, z_i) \end{cases} \tag{83}$$

In a good (M, d, ϵ) JSCC code, M would be chosen so large that with overwhelming probability, a source outcome would be encoded successfully within distortion d. It might seem

counterproductive to let the source encoder in (83) give up before reaching the end of the list of representation points, but in fact, such behavior helps the channel decoder by skewing the distribution of $f_s^{(M)}(S)$.

Channel Encoder. Given a codebook $(x_1, \ldots, x_M) \in \mathcal{X}^M$, the channel encoder outputs x_m if m is the output of the source encoder:

$$f_c^{(M)}(m) = x_m \qquad (84)$$

Channel Decoder. Define the random variable $U \in \{1, \ldots, M+1\}$ which is a function of S, W and z^M only:

$$U = \begin{cases} f_s^{(M)}(S) & d(S, g_s(f_s(S))) \le d \\ M+1 & \text{otherwise} \end{cases} \qquad (85)$$

Having observed $y \in \mathcal{Y}$, the channel decoder chooses arbitrarily among the members of the set:

$$g_c^{(M)}(y) = m \in \arg \max_{j \in \{1, \ldots M\}} P_{U|Z^M}(j|z^M) P_{Y|X}(y|x_j) \qquad (86)$$

A MAP decoder would multiply $P_{Y|X}(y|x_j)$ by $P_X(x_j)$. While that decoder would be too hard to analyze, the product in (86) is a good approximation because $P_{U|Z^M}(j|z^M)$ and $P_X(x_j)$ are related by

$$P_X(x_j) = \sum_{m:\, x_m = x_j} P_{U|Z^M}(m|z^M) + P_{U|Z^M}(M+1|z^M) 1\{j = M\} \qquad (87)$$

so the decoder in (86) differs from a MAP decoder only when either several x_m are identical, or there is no representation point among the first W points within distortion d of the source, both unusual events.

Source Decoder. The source decoder outputs z_m if m is the output of the channel decoder:

$$g_s^{(M)}(m) = z_m \qquad (88)$$

Error Probability Analysis. We now proceed to analyze the performance of the code described above. If there were no source encoding error, a channel decoding error can occur if and only if

$$\exists j \ne m\colon$$
$$P_{U|Z^M}(j|z^M) P_{Y|X}(Y|x_j) \ge P_{U|Z^M}(m|z^M) P_{Y|X}(Y|x_m) \qquad (89)$$

Let the channel codebook $(X_1, \ldots, X_M)$ be drawn i.i.d. from P_X, and independent of the source codebook $(Z_1, \ldots, Z_M)$, which is drawn i.i.d. from P_Z. Denote by $\epsilon(x^M, z^M)$ the excess-distortion probability attained with the source codebook z^M and the channel codebook x^M. Conditioned on the event $\{d(S, g_s(f_s(S))) \le d\} = \{U \le W\} = \{U \ne M+1\}$ (no failure at the source encoder), the probability of excess distortion is upper bounded by the probability that the channel decoder

does not choose $f_s^{(M)}(S)$, so

$$\epsilon(x^M, z^M)$$
$$\le \sum_{m=1}^M P_{U|Z^M}(m|z^m)$$
$$\cdot \mathbb{P}\left[\bigcup_{j \ne m} \left\{ \frac{P_{U|Z^M}(j|z^M) P_{Y|X}(Y|x_j)}{P_{U|Z^M}(m|z^M) P_{Y|X}(Y|x_m)} \ge 1 \right\} \mid X = x_m \right]$$
$$+ P_{U|Z^M}(U > W|z^M) \qquad (90)$$

We now average (90) over the source and channel codebooks. Averaging the m-th term of the sum in (90) with respect to the channel codebook yields

$$P_{U|Z^M}(m|z^m) \mathbb{P}\left[\bigcup_{j \ne m} \left\{ \frac{P_{U|Z^M}(j|z^M) P_{Y|X}(Y|X_j)}{P_{U|Z^M}(m|z^M) P_{Y|X}(Y|X_m)} \ge 1 \right\} \right] \qquad (91)$$

where $Y, X_1, \ldots, X_M$ are distributed according to

$$P_{YX_1 \ldots X_m}(y, x_1, \ldots, x_M) = P_{Y|X_m}(y|x_m) \prod_{j \ne m} P_X(x_j) \qquad (92)$$

Letting $\bar{X}$ be an independent copy of X and applying the union bound to the probability in (91), we have that for any given (m, z^M),

$$\mathbb{P}\left[\bigcup_{j \ne m} \left\{ \frac{P_{U|Z^M}(j|z^M) P_{Y|X}(Y|X_j)}{P_{U|Z^M}(m|z^M) P_{Y|X}(Y|X_m)} \ge 1 \right\} \right]$$
$$\le \mathbb{E}\Bigg[\min \Bigg\{ 1,$$
$$\sum_{j=1}^M \mathbb{P}\left[\frac{P_{U|Z^M}(j|z^M) P_{Y|X}(Y|\bar{X})}{P_{U|Z^M}(m|z^M) P_{Y|X}(Y|X)} \ge 1 \mid X, Y \right] \Bigg\} \Bigg] \qquad (93)$$
$$\le \mathbb{E}\left[\min \left\{ 1, \sum_{j=1}^M \frac{P_{U|Z^M}(j|z^M)}{P_{U|Z^M}(m|z^M)} \frac{\mathbb{E}\left[P_{Y|X}(Y|\bar{X})|Y\right]}{P_{Y|X}(Y|X)} \right\} \right] \qquad (94)$$
$$= \mathbb{E}\left[\min \left\{ 1, \sum_{j=1}^M \frac{P_{U|Z^M}(j|z^M)}{P_{U|Z^M}(m|z^M)} \frac{P_Y(Y)}{P_{Y|X}(Y|X)} \right\} \right] \qquad (95)$$
$$= \mathbb{E}\left[\min \left\{ 1, \frac{\mathbb{P}\left[U \le W \mid Z^M = z^M\right]}{P_{U|Z^M}(m|z^M)} \frac{P_Y(Y)}{P_{Y|X}(Y|X)} \right\} \right] \qquad (96)$$
$$= \mathbb{E}\left[\min \left\{ 1, \frac{1}{P_{U|Z^M, 1\{U \le W\}}(m|z^M, 1)} \frac{P_Y(Y)}{P_{Y|X}(Y|X)} \right\} \right] \qquad (97)$$

where (94) is due to $1\{a \ge 1\} \le a$.

Applying (97) to (90) and averaging with respect to the source codebook, we may write

$$\mathbb{E}\left[\epsilon(X^M, Z^M)\right] \le \mathbb{E}\left[\min\{1, G\}\right] + \mathbb{P}\left[U > W\right] \qquad (98)$$

where for brevity we denoted the random variable

$$G = \frac{1}{P_{U|Z^M, 1\{U \le W\}}(U|Z^M, 1)} \frac{P_Y(Y)}{P_{Y|X}(Y|X)} \qquad (99)$$

The expectation in the right side of (98) is with respect to $P_{Z^M} P_{U|Z^M} P_{W|UZ^M} P_X P_{Y|X}$. It is equal to

$$\mathbb{E}\left[\mathbb{E}\left[\min\{1,G\} \mid X,Y,Z^M,1\{U \leq W\}\right]\right]$$

$$\leq \mathbb{E}\left[\min\left\{1,\mathbb{E}\left[G \mid X,Y,Z^M,1\{U \leq W\}\right]\right\}\right] \quad (100)$$

$$= \mathbb{E}\left[\min\left\{1,W\frac{P_Y(Y)}{P_{Y|X}(Y|X)}\right\}\right] \quad (101)$$

$$= \mathbb{E}\left[\exp\left(-\left|\imath_{X;Y}(X;Y) - \log W\right|^+\right)\right] \quad (102)$$

where

- (100) applies Jensen's inequality to the concave function $\min\{1,a\}$;
- (101) uses $P_{U|X,Y,Z^M,1\{U\leq W\}} = P_{U|Z^M,1\{U\leq W\}}$;
- (102) is due to $\min\{1,a\} = \exp\left(-\left|\log\frac{1}{a}\right|^+\right)$, where a is nonnegative.

To evaluate the probability in the right side of (98), note that conditioned on $S = s$, $W = w$, U is distributed as:

$$P_{U|S,W}(m|s,w) = \begin{cases} \rho(s)(1-\rho(s))^{m-1} & m = 1,2,\ldots,w \\ (1-\rho(s))^w & m = M+1 \end{cases} \quad (103)$$

where we denoted for brevity

$$\rho(s) = P_Z(B_d(s)) \quad (104)$$

Therefore,

$$\mathbb{P}\left[U > W\right] = \mathbb{E}\left[\mathbb{P}\left[U > W|S,W\right]\right] \quad (105)$$

$$= \mathbb{E}\left[(1-\rho(S))^W\right] \quad (106)$$

Applying (102) and (106) to (98) and invoking Shannon's random coding argument, (82) follows. ∎

Remark 5. As we saw in the proof of Theorem 7, if we restrict W to take values on $\{1,\ldots,M\}$, then the bound on the error probability ϵ in (82) is achieved in the class of (M,d,ϵ) codes. The code size M that leads to tight achievability bounds following from Theorem 7 is in general much larger than the size that achieves the minimum in (81). In that case, M is chosen so that $\log M$ lies between $kR(d)$ and nC so as to minimize the sum of source and channel decoding error probabilities without the benefit of a channel decoder that exploits residual source redundancy. In contrast, Theorem 8 is obtained with an approximate MAP decoder that allows a larger choice for $\log M$, even beyond nC. Still we can achieve a good (d,ϵ) tradeoff because the channel code employs unequal error protection: those codewords with higher probabilities are more reliably decoded.

Remark 6. Had we used the ML channel decoder in lieu of (86) in the proof of Theorem 7, we would conclude that a (d,ϵ) code exists with

$$\epsilon \leq \inf_{P_X,P_Z,M}\left\{\mathbb{E}\left[\exp\left(-\left|\imath_{X;Y}(X;Y) - \log(M-1)\right|^+\right)\right] \right.$$
$$\left. + \mathbb{E}\left[(1-P_Z(B_d(S)))^M\right]\right\} \quad (107)$$

which corresponds to the SSCC bound in (81) with the worst-case average channel error probability $\epsilon^*(M)$ upper bounded using the random coding union (RCU) bound of [8] and the

source error probability $\epsilon^*(M,d)$ upper bounded using the random coding achievability bound of [9].

Remark 7. Weakening (82) by letting $W = M$, we obtain a slightly looser version of (107) in which $M-1$ in the exponent is replaced by M. To get a generally tighter bound than that afforded by SSCC, a more intelligent choice of W is needed, as detailed next in Theorem 8.

Theorem 8 (Achievability). *There exists a (d,ϵ) source-channel code with*

$$\epsilon \leq$$
$$\inf_{P_X,P_Z,\gamma>0}\left\{\mathbb{E}\left[\exp\left(-\left|\imath_{X;Y}(X;Y) - \log\frac{\gamma}{P_Z(B_d(S))}\right|^+\right)\right]\right.$$
$$\left. + e^{1-\gamma}\right\} \quad (108)$$

where the expectation is with respect to $P_S P_X P_{Y|X} P_Z$ defined on $\mathcal{M} \times \mathcal{X} \times \mathcal{Y} \times \widehat{\mathcal{M}}$.

Proof: We fix an arbitrary $\gamma > 0$ and choose

$$W = \left\lfloor\frac{\gamma}{\rho(S)}\right\rfloor \quad (109)$$

where $\rho(\cdot)$ is defined in (104). Observing that

$$(1-\rho(s))^{\lfloor\frac{\gamma}{\rho(s)}\rfloor} \leq (1-\rho(s))^{\frac{\gamma}{\rho(s)}-1} \quad (110)$$

$$\leq e^{-\rho(s)\left(\frac{\gamma}{\rho(s)}-1\right)} \quad (111)$$

$$\leq e^{1-\gamma} \quad (112)$$

we obtain (108) by weakening (82) using (109) and (112). ∎

In the case of almost-lossless JSCC, the bound in Theorem 8 can be sharpened as shown recently by Tauste Campo et al. [17].

Theorem 9 (Achievability, almost-lossless JSCC [17]). *There exists a $(0,\epsilon)$ code with*

$$\epsilon \leq \inf_{P_X}\mathbb{E}\left[\exp\left(-\left|\imath_{X;Y}(X;Y) - \imath_S(S)\right|^+\right)\right] \quad (113)$$

where the expectation is with respect to $P_S P_X P_{Y|X}$ defined on $\mathcal{M} \times \mathcal{X} \times \mathcal{Y}$.

V. GAUSSIAN APPROXIMATION

In addition to the basic conditions (a)-(c) of Section II, in this section we impose the following restrictions.

(i) The channel is stationary and memoryless, $P_{Y^n|X^n} = P_{Y|X} \times \ldots \times P_{Y|X}$. If the channel has an input cost function then it satisfies $\mathsf{c}_n(x^n) = \frac{1}{n}\sum_{i=1}^n \mathsf{c}(x_i)$.

(ii) The source is stationary and memoryless, $P_{S^k} = P_\mathsf{S} \times \ldots \times P_\mathsf{S}$, and the distortion measure is separable, $\mathsf{d}_k(s^k,z^k) = \frac{1}{k}\sum_{i=1}^k \mathsf{d}(s_i,z_i)$.

(iii) The distortion level satisfies $d_{\min} < d < d_{\max}$, where $d_{\min}$ is defined in (10), and $d_{\max} = \inf_{\mathsf{z}\in\hat{\mathcal{S}}}\mathbb{E}\left[\mathsf{d}(\mathsf{S},\mathsf{z})\right]$, where the average is with respect to the unconditional distribution of S. The excess-distortion probability satisfies $0 < \epsilon < 1$.

(iv) $\mathbb{E}\left[\mathsf{d}^9(\mathsf{S},\mathsf{Z}^\star)\right] < \infty$ where the average is with respect to $P_\mathsf{S} \times P_{\mathsf{Z}^\star}$ and $P_{\mathsf{Z}^\star}$ is the output distribution corresponding to the minimizer in (9).

The technical condition (iv) ensures applicability of the Gaussian approximation in the following result.

Theorem 10 (Gaussian approximation). *Under restrictions* (i)–(iv), *the parameters of the optimal* (k, n, d, ϵ) *code satisfy*

$$nC - kR(d) = \sqrt{nV + k\mathcal{V}(d)}Q^{-1}(\epsilon) + \theta(n) \qquad (114)$$

where

1. $\mathcal{V}(d)$ *is the source dispersion given by*

$$\mathcal{V}(d) = \mathrm{Var}\left[\jmath_{\mathsf{S}}(\mathsf{S}, d)\right] \qquad (115)$$

2. V *is the channel dispersion given by:*

   a) *If* $\mathcal{A}$ *and* $\mathcal{B}$ *are finite and the channel has no cost constraints,*

   $$V = \mathrm{Var}\left[\imath^{\star}_{\mathsf{X};\mathsf{Y}}(\mathsf{X}^{\star}; \mathsf{Y}^{\star})\right] \qquad (116)$$

   $$\imath^{\star}_{\mathsf{X};\mathsf{Y}}(x; y) = \log \frac{dP_{\mathsf{Y}|\mathsf{X}=x}}{dP_{\mathsf{Y}^{\star}}}(y) \qquad (117)$$

   where $\mathsf{X}^{\star}$, $\mathsf{Y}^{\star}$ *are the capacity-achieving input and output random variables.*

   b) *If the channel is Gaussian with either equal or maximal power constraint,*

   $$V = \frac{1}{2}\left(1 - \frac{1}{(1+P)^2}\right)\log^2 e \qquad (118)$$

   where P *is the signal-to-noise ratio.*

3. *The remainder term* $\theta(n)$ *satisfies:*

   a) *If* $\mathcal{A}$ *and* $\mathcal{B}$ *are finite, the channel has no cost constraints and* $V > 0$,

   $$-\underline{c}\log n + O(1) \leq \theta(n) \qquad (119)$$
   $$\leq \bar{c}\log n + \log\log n + O(1) \quad (120)$$

   where

   $$\underline{c} = |\mathcal{A}| - \frac{1}{2} \qquad (121)$$

   $$\bar{c} = 1 + \frac{\mathrm{Var}\left[\Lambda'_{\mathsf{Z}^{\star}}(\mathsf{S}, \lambda^{\star})\right]}{\mathbb{E}\left[|\Lambda''_{\mathsf{Z}^{\star}}(\mathsf{S}, \lambda^{\star})|\right]\log e} \qquad (122)$$

   In (122), $(\cdot)'$ *denotes differentiation with respect to* λ, $\Lambda_{\mathsf{Z}^{\star}}(s, \lambda)$ *is defined by*

   $$\Lambda_{\mathsf{Z}^{\star}}(s, \lambda) = \log \frac{1}{\mathbb{E}\left[\exp\left(\lambda d - \lambda d(s, \mathsf{Z}^{\star})\right)\right]} \qquad (123)$$

   (cf. Definition 5) and $\lambda^{\star} = -R'(d)$.

   b) *If* $\mathcal{A}$ *and* $\mathcal{B}$ *are finite, the channel has no cost constraints and* $V = 0$, (120) *still holds, while* (119) *is replaced with*

   $$\liminf_{n \to \infty} \frac{\theta(n)}{\sqrt{n}} \geq 0 \qquad (124)$$

   c) *If the channel is such that the (conditional) distribution of* $\imath^{\star}_{\mathsf{X};\mathsf{Y}}(x; \mathsf{Y})$ *does not depend on* $x \in \mathcal{A}$ *(no cost constraint), then* $\underline{c} = \frac{1}{2}$. [6]

   d) *If the channel is Gaussian with equal or maximal power constraint,* (120) *still holds, and* (119) *holds with* $\underline{c} = \frac{1}{2}$.

---

[6] Note added in proof: we have shown recently that the symmetricity condition is actually superfluous.

e) *In the almost-lossless case,* $R(d) = H(\mathsf{S})$, *and provided that the third absolute moment of* $\imath_{\mathsf{S}}(\mathsf{S})$ *is finite,* (114) *and* (119) *still hold, while* (120) *strengthens to*

$$\theta(n) \leq \frac{1}{2}\log n + O(1) \qquad (125)$$

Proof:

- Appendices C-A and C-B show the converses in (119) and (124) for cases $V > 0$ and $V = 0$, respectively, using Theorem 3.
- Appendix C-C shows the converse for the symmetric channel (3c) using Theorem 2.
- Appendix C-D shows the converse for the Gaussian channel (3d) using Theorem 2.
- Appendix D-A shows the achievability result for almost lossless coding (3e) using Theorem 9.
- Appendix D-B shows the achievability result in (120) for the DMC using Theorem 8.
- Appendix D-C shows the achievability result for the Gaussian channel (3d) using Theorem 8.

∎

Remark 8. If the channel and the data compression codes are designed separately, we can invoke channel coding [8] and lossy compression [9] results in (1) and (2) to show that (cf. (4))

$$nC - kR(d) \leq \min_{\eta + \zeta \leq \epsilon}\left\{\sqrt{nV}Q^{-1}(\eta) + \sqrt{k\mathcal{V}(d)}Q^{-1}(\zeta)\right\} + O(\log n) \qquad (126)$$

Comparing (126) to (114), observe that if either the channel or the source (or both) have zero dispersion, the joint source-channel coding dispersion can be achieved by separate coding. In that special case, either the d-tilted information or the channel information density are so close to being deterministic that there is no need to account for the true distributions of these random variables, as a good joint source-channel code would do.

The Gaussian approximations of JSCC and SSCC in (114) and (126), respectively, admit the following heuristic interpretation when n is large (and thus, so is k): since the source is stationary and memoryless, the normalized d-tilted information $J = \frac{1}{n}\jmath_{S^k}(S^k, d)$ becomes approximately Gaussian with mean $\frac{k}{n}R(d)$ and variance $\frac{k}{n}\frac{\mathcal{V}(d)}{n}$. Likewise, the conditional normalized channel information density $I = \frac{1}{n}\imath^{\star}_{X^n;Y^n}(x^n; Y^{n\star})$ is, for large k, n, approximately Gaussian with mean C and variance $\frac{V}{n}$ for all $x^n \in \mathcal{A}^n$ typical according to the capacity-achieving distribution. Since a good encoder chooses such inputs for (almost) all source realizations, and the source and the channel are independent, the random variable $I - J$ is approximately Gaussian with mean $C - \frac{k}{n}R(d)$ and variance $\frac{1}{n}\left(\frac{k}{n}\mathcal{V}(d) + V\right)$, and (114) reflects the intuition that under JSCC, the source is reconstructed successfully within distortion d if and only if the channel information density exceeds the source d-tilted information, that is, $\{I > J\}$. In contrast, in SSCC, the source is reconstructed successfully

with high probability if (I, J) falls in the intersection of half-planes $\{I > r\} \cap \{J < r\}$ for some $r = \frac{\log M}{n}$, which is the capacity of the noiseless link between the source and the channel code block that can be chosen so as to maximize the probability of that intersection, as reflected in (126). Since in JSCC the successful transmission event is strictly larger than in SSCC, i.e. $\{I > r\} \cap \{J < r\} \subset \{I > J\}$, separate source/channel code design incurs a performance loss. It is worth pointing out that $\{I > J\}$ leads to successful reconstruction even within the paradigm of the codes in Definition 8 because, as explained in Remark 5, unlike the SSCC case, it is not necessary that $\frac{\log M}{n}$ lie between I and J for successful reconstruction.

Remark 9. Using Theorem 10, it can be shown that

$$R(n, d, \epsilon) = \frac{C}{R(d)} - \sqrt{\frac{\mathcal{V}(d)}{n}} Q^{-1}(\epsilon) - \frac{1}{R(d)} \frac{\theta(n)}{n} \quad (127)$$

where the rate-dispersion function of JSCC is found as (recall Definition 4),

$$\mathcal{V}(d) = \frac{R(d)V + C\mathcal{V}(d)}{R^3(d)} \quad (128)$$

Remark 10. Under regularity conditions similar to those in [9, Theorem 14], it can be shown that

$$D(nR, n, \epsilon) = D\left(\frac{C}{R}\right) + \sqrt{\frac{\mathcal{W}(R)}{n}} Q^{-1}(\epsilon) + \frac{\partial}{\partial R} D\left(\frac{C}{R}\right) \frac{\theta(n)}{n} \quad (129)$$

where the distortion-dispersion function of JSCC is given by

$$\mathcal{W}(R) = \left(\frac{\partial}{\partial R} D\left(\frac{C}{R}\right)\right)^2 \left(V + R\mathcal{V}\left(D\left(\frac{C}{R}\right)\right)\right) \quad (130)$$

Remark 11. If the basic conditions (b) and/or (c) fail so that there are several distributions $P_{Z^\star|S}$ and/or several $P_{X^\star}$ that achieve the rate-distortion function and the capacity, respectively, then, for $\epsilon < \frac{1}{2}$,

$$\mathcal{V}(d) \leq \min \mathcal{V}_{Z^\star; X^\star}(d) \quad (131)$$
$$\mathcal{W}(R) \leq \min \mathcal{W}_{Z^\star; X^\star}(R) \quad (132)$$

where the minimum is taken over $P_{Z^\star|S}$ and $P_{X^\star}$, and $\mathcal{V}_{Z^\star; X^\star}(d)$ (resp. $\mathcal{W}_{Z^\star; X^\star}(R)$) denotes (128) (resp. (130)) computed with $P_{Z^\star|S}$ and $P_{X^\star}$. The reason for possibly lower achievable dispersion in this case is that we have the freedom to map the unlikely source realizations leading to high probability of failure to those codewords resulting in the maximum variance so as to increase the probability that the channel output escapes the decoding failure region.

Remark 12. The dispersion of the Gaussian channel is given by (118), regardless of whether an equal or a maximal power constraint is imposed. An equal power constraint corresponds to the subset of allowable channel inputs being the power sphere:

$$\mathcal{F}(P) = \left\{ x^n \in \mathbb{R}^n : \frac{|x^n|^2}{\sigma_{\mathsf{N}}^2} = nP \right\} \quad (133)$$

where σ_{N}^2 is the noise power. In a maximal power constraint, (133) is relaxed replacing '=' with '$\leq$'.

Specifying the nature of the power constraint in the subscript, we remark that the bounds for the maximal constraint can be obtained from the bounds for the equal power constraint via the following relation

$$k_{\mathrm{eq}}^\star(n, d, \epsilon) \leq k_{\max}^\star(n, d, \epsilon) \leq k_{\mathrm{eq}}^\star(n + 1, d, \epsilon) \quad (134)$$

where the right-most inequality is due to the following idea dating back to Shannon: a (k, n, d, ϵ) code with a maximal power constraint can be converted to a $(k, n + 1, d, \epsilon)$ code with an equal power constraint by appending an $(n + 1)$-th coordinate to each codeword to equalize its total power to $n\sigma_{\mathsf{N}}^2 P$. From (134) it is immediate that the channel dispersions for maximal or equal power constraints must be the same.

VI. LOSSY TRANSMISSION OF A BMS OVER A BSC

In this section we particularize the bounds in Sections III, IV and the approximation in Section V to the transmission of a BMS with bias p over a BSC with crossover probability δ. The target bit error rate satisfies $d \leq p$.

The rate-distortion function of the source and the channel capacity are given by, respectively,

$$R(d) = h(p) - h(d) \quad (135)$$
$$C = 1 - h(\delta) \quad (136)$$

The source and the channel dispersions are given by [8], [9]:

$$\mathcal{V}(d) = p(1 - p) \log^2 \frac{1 - p}{p} \quad (137)$$

$$V = \delta(1 - \delta) \log^2 \frac{1 - \delta}{\delta} \quad (138)$$

where note that (137) does not depend on d. The rate-dispersion function in (128) is plotted in Fig. 3.
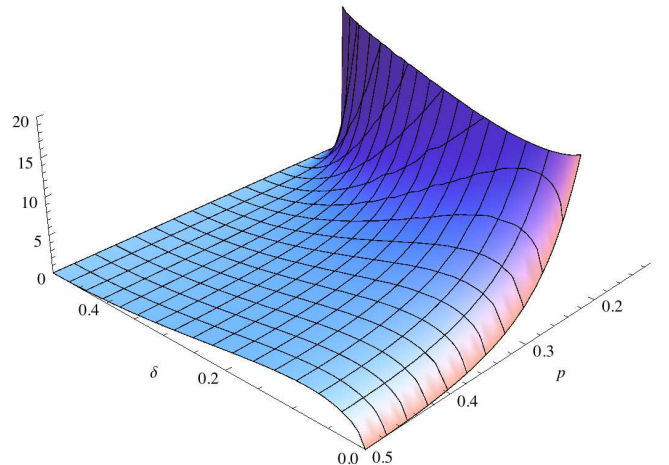

Fig. 3. The rate-dispersion function for the transmission of a BMS over a BSC with $d = 0.11$ as a function of (δ, p) in $\left(0, \frac{1}{2}\right) \times \left(d, \frac{1}{2}\right)$. It increases unboundedly as $p \to d$, and vanishes as $\delta \to \frac{1}{2}$ or $(\delta, p) \to \left(0, \frac{1}{2}\right)$.

Throughout this section, $w(a^\ell)$ denotes the Hamming weight of the binary ℓ-vector a^ℓ, and T_α^ℓ denotes a binomial random variable with parameters ℓ and α, independent of all other random variables.

For convenience, we define the discrete random variable $U_{\alpha,\beta}$ by

$$U_{\alpha,\beta} = \left(T_\alpha^k - kp\right)\log\frac{1-p}{p} + \left(T_\beta^n - n\delta\right)\log\frac{1-\delta}{\delta} \quad (139)$$

In particular, substituting $\alpha = p$ and $\beta = \delta$ in (139), we observe that the terms in the right side of (139) are zero-mean random variables whose variances are equal to $k\mathcal{V}(d)$ and nV, respectively.

Furthermore, the binomial sum is denoted by

$$\left\langle \begin{matrix} k \\ \ell \end{matrix} \right\rangle = \sum_{i=0}^{\ell} \binom{k}{i} \quad (140)$$

A straightforward particularization of the d-tilted information converse in Theorem 2 leads to the following result.

Theorem 11 (Converse, BMS-BSC). *Any (k,n,d,ϵ) code for transmission of a BMS with bias p over a BSC with bias δ must satisfy*

$$\epsilon \geq \sup_{\gamma \geq 0}\left\{\ \mathbb{P}\left[U_{p,\delta} \geq nC - kR(d) + \gamma\right] - \exp\left(-\gamma\right) \right\} \quad (141)$$

Proof: Let $P_{\bar{Y}^n} = P_{Y^{n\star}}$, which is the equiprobable distribution on $\{0,1\}^n$. An easy exercise reveals that

$$\jmath_{S^k}(s^k, d) = \imath_{S^k}(s^k) - kh(d) \quad (142)$$

$$\imath_{S^k}(s^k) = kh(p) + \left(w(s^k) - kp\right)\log\frac{1-p}{p} \quad (143)$$

$$\imath_{X^n;Y^{n\star}}(x^n; y^n) = n\left(\log 2 - h(\delta)\right)$$
$$- \left(w(y^n - x^n) - n\delta\right)\log\frac{1-\delta}{\delta} \quad (144)$$

Since $w(Y^n - x^n)$ is distributed as T_δ^n regardless of $x^n \in \{0,1\}^n$, and $w(S^k)$ is distributed as T_p^k, the condition in Theorem 2 is satisfied, and (37) becomes (141). ∎

The hypothesis-testing converse in Theorem 4 particularizes to the following result:

Theorem 12 (Converse, BMS-BSC). *Any (k,n,d,ϵ) code for transmission of a BMS with bias p over a BSC with bias δ must satisfy*

$$\mathbb{P}\left[U_{\frac{1}{2},\frac{1}{2}} < r\right] + \lambda\mathbb{P}\left[U_{\frac{1}{2},\frac{1}{2}} = r\right] \leq \left\langle \begin{matrix} k \\ \lfloor kd \rfloor \end{matrix} \right\rangle 2^{-k} \quad (145)$$

where $0 \leq \lambda < 1$ and scalar r are uniquely defined by

$$\mathbb{P}\left[U_{p,\delta} < r\right] + \lambda\mathbb{P}\left[U_{p,\delta} = r\right] = 1 - \epsilon \quad (146)$$

Proof: As in the proof of Theorem 11, we let $P_{\bar{Y}^n}$ be the equiprobable distribution on $\{0,1\}^n$, $P_{\bar{Y}^n} = P_{Y^{n\star}}$. Since under $P_{Y^n|X^n=x^n}$, $w\left(Y^n - x^n\right)$ is distributed as T_δ^n, and under $P_{Y^{n\star}}$, $w\left(Y^n - x^n\right)$ is distributed as $T_{\frac{1}{2}}^n$, irrespective of the choice of $x^n \in A^n$, the distribution of the information density in (144) does not depend on the choice of x^n under either measure, so Theorem 5 can be applied. Further, we choose Q_{S^k} to be the equiprobable distribution on $\{0,1\}^k$ and observe that under P_{S^k}, the random variable $w(S^k)$ in (143) has the same distribution as T_p^k, while under Q_{S^k} it has the

same distribution as $T_{\frac{1}{2}}^k$. Therefore, the log-likelihood ratio for testing between $P_{S^k}P_{Y^n|X^n=x^n}$ and $Q_{S^k}P_{Y^{n\star}}$ has the same distribution as ('$\sim$' denotes equality in distribution)

$$\log\frac{P_{S^k}(S^k)P_{Y^n|X^n=x^n}(Y^n)}{Q_{S^k}(S^k)P_{Y^{n\star}}(Y^n)}$$

$$= \imath_{X^n;Y^{n\star}}(x^n; Y^n) - \imath_{S^k}(S^k) + k\log 2 \quad (147)$$

$$\sim n\log 2 - nh(\delta) - kh(p)$$

$$- \begin{cases} U_{p,\delta} & \text{under } P_{S^k}P_{Y^n|X^n=x^n} \\ U_{\frac{1}{2},\frac{1}{2}} & \text{under } Q_{S^k}P_{Y^{n\star}} \end{cases} \quad (148)$$

so $\beta_{1-\epsilon}(P_{S^k}P_{Y^n|X^n=x^n}, Q_{S^k}P_{Y^{n\star}})$ is equal to the left side of (145). Finally, matching the size of the list to the fidelity of reproduction using (55), we find that L is equal to the right side of (145). ∎

If the source is equiprobable, the bound in Theorem 12 becomes particularly simple, as the following result details.

Theorem 13 (Converse, EBMS-BSC). *For $p = \frac{1}{2}$, if there exists a (k,n,d,ϵ) joint source-channel code, then*

$$\lambda\left\langle \begin{matrix} n \\ r^\star + 1 \end{matrix} \right\rangle + \left\langle \begin{matrix} n \\ r^\star \end{matrix} \right\rangle \leq \left\langle \begin{matrix} k \\ \lfloor kd \rfloor \end{matrix} \right\rangle 2^{n-k} \quad (149)$$

where

$$r^\star = \max\left\{r : \sum_{t=0}^{r}\binom{n}{t}\delta^t(1-\delta)^{n-t} \leq 1 - \epsilon\right\} \quad (150)$$

and $\lambda \in [0,1)$ is the solution to

$$\sum_{j=0}^{r^\star}\binom{n}{t}\delta^t(1-\delta)^{n-t} + \lambda\delta^{r^\star+1}(1-\delta)^{n-r^\star-1}\binom{n}{r^\star+1} = 1-\epsilon \quad (151)$$

The achievability result in Theorem 8 is particularized as follows.

Theorem 14 (Achievability, BMS-BSC). *There exists an (k,n,d,ϵ) joint source-channel code with*

$$\epsilon \leq \inf_{\gamma > 0}\left\{\mathbb{E}\left[\exp\left(-|U - \log\gamma|^+\right)\right] + e^{1-\gamma}\right\} \quad (152)$$

where

$$U = nC - (T_\delta^n - n\delta)\log\frac{1-\delta}{\delta} - \log\frac{1}{\rho(T_p^k)} \quad (153)$$

and $\rho\colon \{0,1,\ldots,k\} \mapsto [0,1]$ is defined as

$$\rho(T) = \sum_{t=0}^{k}L(T,t)q^t(1-q)^{k-t} \quad (154)$$

with

$$L(T,t) = \begin{cases} \binom{T}{t_0}\binom{k-T}{t-t_0} & t - kd \leq T \leq t + kd \\ 0 & \text{otherwise} \end{cases} \quad (155)$$

$$t_0 = \left\lceil\frac{t+T-kd}{2}\right\rceil^+ \quad (156)$$

$$q = \frac{p-d}{1-2d} \quad (157)$$

Proof: We weaken the infima over P_{X^n} and P_{Z^k} in (108) by choosing them to be the product distributions generated by the capacity-achieving channel input distribution and the rate-distortion function-achieving reproduction distribution, respectively, i.e. P_{X^n} is equiprobable on $\{0,1\}^n$, and $P_{Z^k} = P_{Z^\star} \times \ldots \times P_{Z^\star}$, where $P_{Z^\star}(1) = q$. As shown in [9, proof of Theorem 21],

$$P_{Z^k}\left(B_d(s^k)\right) \geq \rho(w(s^k)) \qquad (158)$$

On the other hand, $|Y^n - X^n|_0$ is distributed as T_δ^n, so (152) follows by substituting (144) and (158) into (108). ∎

In the special case of the BMS-BSC, Theorem 10 can be strengthened as follows.

Theorem 15 (Gaussian approximation, BMS-BSC). *The parameters of the optimal (k,n,d,ϵ) code satisfy (114) where $R(d)$, C, $\mathbb{V}(d)$, V are given by (135), (136), (137), (138), respectively, and the remainder term in (114) satisfies*

$$O(1) \leq \theta(n) \qquad (159)$$
$$\leq \log n + \log \log n + O(1) \qquad (160)$$

if $0 < d < p$, and

$$-\frac{1}{2}\log n + O(1) \leq \theta(n) \qquad (161)$$
$$\leq \frac{1}{2}\log n + O(1) \qquad (162)$$

if $d = 0$.

Proof: An asymptotic analysis of the converse bound in Theorem 12 akin to that found in [9, proof of Theorem 23] leads to (159) and (161). An asymptotic analysis of the achievability bound in Theorem 14 similar to the one found in [9, Appendix G] leads to (160). Finally, (162) is the same as (125). ∎

The bounds and the Gaussian approximation (in which we take $\theta(n) = 0$) are plotted in Fig. 4 ($d = 0$), Fig. 5 (fair binary source, $d > 0$) and Fig. 6 (biased binary source, $d > 0$). A source of fair coin flips has zero dispersion, and as anticipated in Remark 8, JSSC does not afford much gain in the finite blocklength regime (Fig. 5). Moreover, in that case the JSCC achievability bound in Theorem 8 is worse than the SSCC achievability bound. However, the more general achievability bound in Theorem 7 with the choice $W = M$, as detailed in Remark 7, nearly coincides with the SSCC curve in Fig. 5, providing an improvement over Theorem 8. The situation is different if the source is biased, with JSCC showing significant gain over SSCC (Figures 4 and 6).

VII. TRANSMISSION OF A GMS OVER AN AWGN CHANNEL

In this section we analyze the setup where the Gaussian memoryless source $S_i \sim \mathcal{N}(0, \sigma_S^2)$ is transmitted over an AWGN channel, which, upon receiving an input x^n, outputs $Y^n = x^n + N^n$, where $N^n \sim \mathcal{N}(0, \sigma_N^2 \mathbf{I})$. The encoder/decoder must satisfy two constraints, the fidelity constraint and the cost constraint:
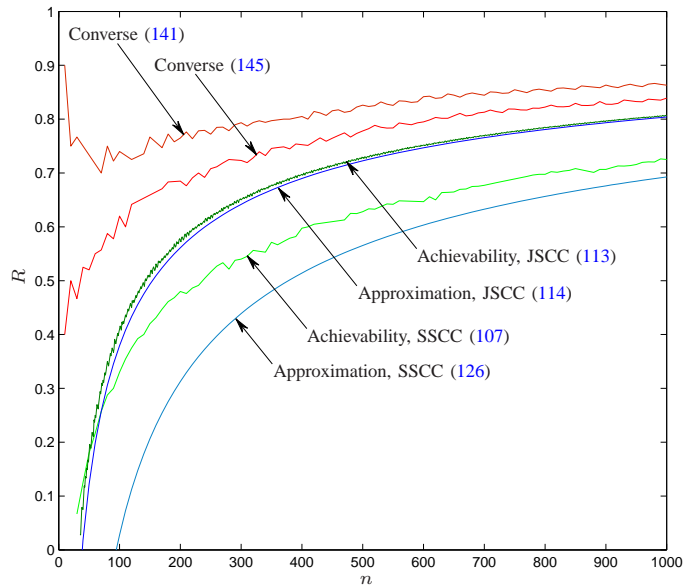

Fig. 4. Rate-blocklength tradeoff for the transmission of a BMS with bias $p = 0.11$ over a BSC with crossover probability $\delta = p = 0.11$ and $d = 0$, $\epsilon = 10^{-2}$.
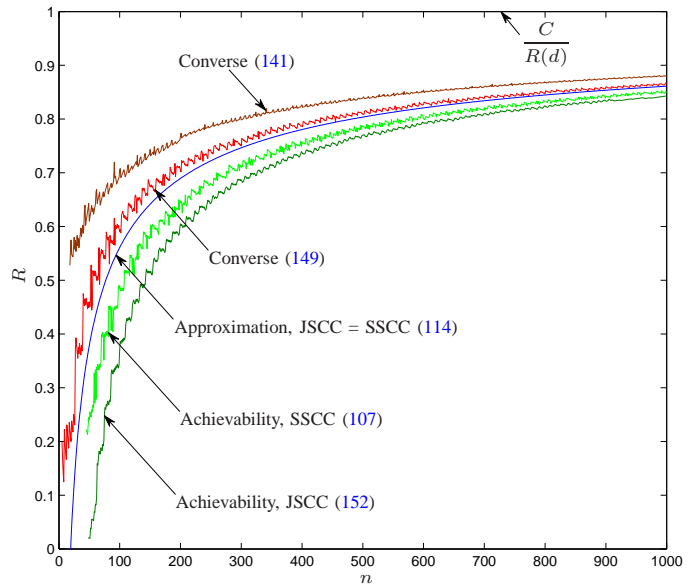

Fig. 5. Rate-blocklength tradeoff for the transmission of a fair BMS over a BSC with crossover probability $\delta = d = 0.11$ and $\epsilon = 10^{-2}$.

- the MSE distortion exceeds $0 \leq d \leq \sigma_S^2$ with probability no greater than $0 < \epsilon < 1$;
- each channel codeword satisfies the equal power constraint in (133).[7]

The capacity-cost function and the rate-distortion function are given by

$$R(d) = \frac{1}{2}\log\left(\frac{\sigma_S^2}{d}\right) \qquad (163)$$

$$C(P) = \frac{1}{2}\log(1 + P) \qquad (164)$$

---

[7]See Remark 12 in Section V for a discussion of the close relation between an equal and a maximal power constraint.
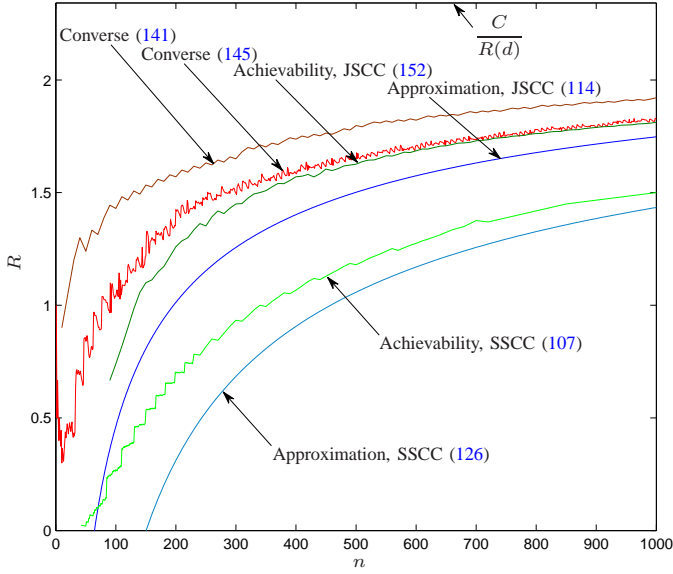
Fig. 6. Rate-blocklength tradeoff for the transmission of a BMS with bias $p = 0.11$ over a BSC with crossover probability $\delta = p = 0.11$ and $d = 0.05$, $\epsilon = 10^{-2}$.

The source dispersion is given by [9]:

$$\mathcal{V}(d) = \frac{1}{2} \log^2 e \tag{165}$$

while the channel dispersion is given by (118) [8].

In the rest of the section, W_λ^ℓ denotes a noncentral chi-square distributed random variable with ℓ degrees of freedom and non-centrality parameter λ, independent of all other random variables, and $f_{W_\lambda^\ell}$ denotes its probability density function.

A straightforward particularization of the d-tilted information converse in Theorem 2 leads to the following result.

Theorem 16 (Converse, GMS-AWGN). *If there exists a* (k, n, d, ϵ) *code, then*

$$\epsilon \geq \sup_{\gamma \geq 0} \left\{ \mathbb{P}\left[U \geq nC(P) - kR(d) + \gamma\right] - \exp\left(-\gamma\right) \right\} \tag{166}$$

where

$$U = \frac{\log e}{2}\left(W_0^k - k\right) + \frac{\log e}{2}\left(\frac{P}{1+P}W_{\frac{n}{P}}^n - n\right) \tag{167}$$

Observe that the terms to the left of the '$\geq$' sign inside the probability in (166) are zero-mean random variables whose variances are equal to $k\mathcal{V}(d)$ and nV, respectively.

Proof: The spherically-symmetric $P_{\bar{Y}^n} = P_{Y^{n\star}} = P_{Y^\star} \times \ldots \times P_{Y^\star}$, where $Y^\star \sim \mathcal{N}(0, \sigma_N^2(1 + P))$ is the capacity-achieving output distribution, satisfies the symmetry assumption of Theorem 2. More precisely, it is not hard to show (see [8, (205)]) that for all $x^n \in \mathcal{F}(\alpha)$, $\imath_{X^n;Y^{n\star}}(x^n; Y^n)$ has the same distribution under $P_{Y^{n\star}|X^n=x^n}$ as

$$\frac{n}{2} \log\left(1 + P\right) - \frac{\log e}{2}\left(\frac{P}{1+P}W_{\frac{n}{P}}^n - n\right) \tag{168}$$

The d-tilted information in s^k is given by

$$\jmath_{S^k}(s^k, d) = \frac{k}{2} \log \frac{\sigma_S^2}{d} + \left(\frac{|s^k|^2}{\sigma_S^2} - k\right) \frac{\log e}{2} \tag{169}$$

Plugging (168) and (169) into (37), (166) follows. ∎

The hypothesis testing converse in Theorem 5 is particularized as follows.

Theorem 17 (Converse, GMS-AWGN).

$$k \int_0^\infty r^{k-1} \mathbb{P}\left[PW_{n(1+\frac{1}{P})}^n + k\frac{d}{\sigma^2}r^2 \leq n\tau\right] dr \leq 1 \tag{170}$$

where τ is the solution to

$$\mathbb{P}\left[\frac{P}{1+P}W_{\frac{n}{P}}^n + W_0^k \leq n\tau\right] = 1 - \epsilon \tag{171}$$

Proof: As in the proof of Theorem 16, we let $\bar{Y}^n \sim Y^{n\star} \sim \mathcal{N}(0, \sigma_N^2(1+P)\mathbf{I})$. Under $P_{Y^n|X^n=x^n}$, the distribution of $\imath_{X^n;Y^{n\star}}(x^n; Y^{n\star})$ is that of (168), while under $P_{Y^{n\star}}$, it has the same distribution as (cf. [8, (204)])

$$\frac{n}{2} \log(1 + P) - \frac{\log e}{2}\left(PW_{n(1+\frac{1}{P})}^n - n\right) \tag{172}$$

Since the distribution of $\imath_{X^n;Y^{n\star}}(x^n; Y^{n\star})$ does not depend on the choice of $x^n \in \mathbb{R}^n$ according to either measure, Theorem 5 applies. Further, choosing Q_{S^k} to be the Lebesgue measure on $\mathbb{R}^k$, i.e. $dQ_{S^k} = ds^k$, observe that

$$\log f_{S^k}(s^k) = \log \frac{dP_{S^k}(s^k)}{ds^k} = -\frac{k}{2} \log\left(2\pi\sigma_S^2\right) - \frac{\log e}{2\sigma_S^2}|s^k|^2 \tag{173}$$

Now, (170) and (171) are obtained by integrating

$$1\Bigg\{\log f_{S^k}(s^k) + \imath_{X^n;Y^{n\star}}(x^n; y^n) > \frac{n}{2} \log(1 + P) + \frac{n}{2} \log e - \frac{k}{2} \log(2\pi\sigma_S^2) - \frac{\log e}{2} n\tau\Bigg\} \tag{174}$$

with respect to $ds^k dP_{Y^{n\star}}(y^n)$ and $dP_{S^k}(s^k)dP_{Y^n|X^n=x^n}(y^n)$, respectively. ∎

The bound in Theorem 8 can be computed as follows.

Theorem 18 (Achievability, GMS-AWGN). *There exists a* (k, n, d, ϵ) *code such that*

$$\epsilon \leq \inf_{\gamma>0} \left\{ \mathbb{E}\left[\exp\left\{-|U - \log\gamma|^+\right\}\right] + e^{1-\gamma} \right\} \tag{175}$$

where

$$U = nC(P) - \frac{\log e}{2}\left(\frac{P}{1+P}W_{\frac{n}{P}}^n - n\right) - \log \frac{F}{\rho(W_0^k)} \tag{176}$$

$$F = \max_{n\in\mathbb{N}, t\in\mathbb{R}^+} \frac{f_{W_{nP}^n}(t)}{f_{W_0^n}\left(\frac{t}{1+P}\right)} < \infty \tag{177}$$

and $\rho : \mathbb{R}^+ \mapsto [0,1]$ is defined by

$$\rho(t) = \frac{\Gamma\left(\frac{k}{2}+1\right)}{\sqrt{\pi}k\Gamma\left(\frac{k-1}{2}+1\right)}\left(1 - L\left(\sqrt{\frac{t}{k}}\right)\right)^{\frac{k-1}{2}} \tag{178}$$

where

$$L(r) = \begin{cases} 0 & r < \sqrt{\frac{d}{\sigma_S^2}} - \sqrt{1 - \frac{d}{\sigma_S^2}} \\ 1 & \left| r - \sqrt{1 - \frac{d}{\sigma_S^2}} \right| > \sqrt{\frac{d}{\sigma_S^2}} \\ \frac{\left(1 + r^2 - 2\frac{d}{\sigma_S^2}\right)^2}{4\left(1 - \frac{d}{\sigma_S^2}\right) r^2} & otherwise \end{cases} \quad (179)$$

Proof: We compute an upper bound to (108) for the specific case of the transmission of a GMS over an AWGN channel. First, we weaken the infimum over P_{Z^k} in (108) by choosing P_{Z^k} to be the uniform distribution on the surface of the k-dimensional sphere with center at $\mathbf{0}$ and radius $r_0 = \sqrt{k}\sigma\sqrt{1 - \frac{d}{\sigma_S^2}}$. We showed in [9, proof of Theorem 37] (see also [14], [22]) that

$$P_{Z^k}\left(B_d(s^k)\right) \geq \rho\left(|s^k|^2\right) \quad (180)$$

which takes care of the source random variable in (108).

We proceed to analyze the channel random variable $\imath_{X^n;Y^n}(X^n;Y^n)$. Observe that since X^n lies on the power sphere and the noise is spherically symmetric, $|Y^n|^2 = |X^n + N^n|^2$ has the same distribution as $|x_0^n + N^n|^2$, where x_0^n is an arbitrary point on the surface of the power sphere. Letting $x_0^n = \sigma_N\sqrt{P}\,(1,1,\ldots,1)$, we see that $\frac{1}{\sigma_N}|x_0^n + N^n|^2 = \sum_{i=1}^{n}\left(\frac{1}{\sigma_N^2}N_i + \sqrt{P}\right)^2$ has the non-central chi-squared distribution with n degrees of freedom and noncentrality parameter nP. To simplify calculations, we express the information density as

$$\imath_{X^n;Y^n}(x_0^n;y^n) = \imath_{X^n;Y^{n\star}}(x_0^n;y^n) - \log\frac{dP_{Y^n}}{dP_{Y^{n\star}}}(y^n) \quad (181)$$

where $Y^{n\star} \sim \mathcal{N}(0, \sigma_N^2(1 + P)\mathbf{I})$. The distribution of $\imath_{X^n;Y^{n\star}}(x_0^n;Y^n)$ is the same as (168). Further, due to the spherical symmetry of both P_{Y^n} and $P_{Y^{n\star}}$, as discussed above, we have (recall that '$\sim$' denotes equality in distribution)

$$\frac{dP_{Y^n}}{dP_{Y^{n\star}}}(Y^n) \sim \frac{f_{W_{nP}^n}(W_{nP}^n)}{f_{W_0^n}\left(\frac{W_{nP}^n}{1+P}\right)} \quad (182)$$

which is bounded uniformly in n as observed in [8, (425), (435)], thus (177) is finite, and (175) follows. ∎

The following result strengthens Theorem 10 in the special case of the GMS-AWGN.

Theorem 19 (Gaussian approximation, GMS-AWGN). *The parameters of the optimal (k, n, d, ϵ) code satisfy (114) where $R(d)$, C, $\mathcal{V}(d)$, V are given by (163), (164), (165), (118), respectively, and the remainder term in (114) satisfies*

$$O(1) \leq \theta(n) \quad (183)$$
$$\leq \log n + \log\log n + O(1) \quad (184)$$

Proof: An asymptotic analysis of the converse bound in Theorem 17 similar to that found in [9, proof of Theorem 40] leads to (183). An asymptotic analysis of the achievability bound in Theorem 18 similar to [9, Appendix K] leads to (184). ∎

Numerical evaluation of the bounds reveals that JSCC noticeably outperforms SSCC in the displayed region of blocklengths (Fig. 7).
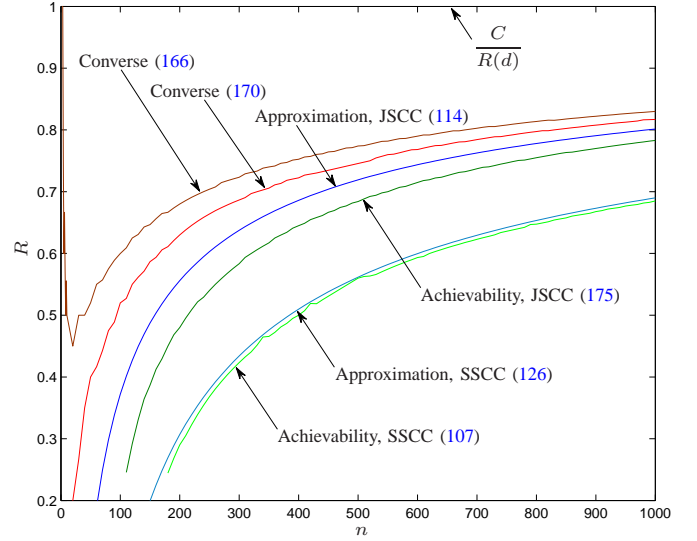

Fig. 7. Rate-blocklength tradeoff for the transmission of a GMS with $\frac{d}{\sigma_s^2} = 0.5$ over an AWGN channel with $P = 1$, $\epsilon = 10^{-2}$.

VIII. TO CODE OR NOT TO CODE

Our goal in this section is to compare the excess distortion performance of the optimal code of rate 1 at channel blocklength n with that of the optimal symbol-by-symbol code, evaluated after n channel uses, leveraging the bounds in Sections III and IV and the approximation in Section V. We show certain examples in which symbol-by-symbol coding is, in fact, either optimal or very close to being optimal. A general conclusion drawn from this section is that even when no coding is asymptotically suboptimal it can be a very attractive choice for short blocklengths [2].

A. Performance of symbol-by-symbol source-channel codes

Definition 9. *An (n, d, ϵ, α) symbol-by-symbol code is an $(n, n, d, \epsilon, \alpha)$ code (f, g) (according to Definition 2) that satisfies*

$$\mathsf{f}(s^n) = (\mathsf{f}_1(s_1), \ldots, \mathsf{f}_1(s_n)) \quad (185)$$
$$\mathsf{g}(y^n) = (\mathsf{g}_1(y_1), \ldots, \mathsf{g}_1(y_n)) \quad (186)$$

for some pair of functions $\mathsf{f}_1 : \mathcal{S} \mapsto \mathcal{A}$ and $\mathsf{g}_1 : \mathcal{B} \mapsto \hat{\mathcal{S}}$.

The minimum excess distortion achievable with symbol-by-symbol codes at channel blocklength n, excess probability ϵ and cost α is defined by

$$D_1(n, \epsilon, \alpha) = \inf\{d : \exists (n, d, \epsilon, \alpha)\ symbol\text{-}by\text{-}symbol\ code\}. \quad (187)$$

Definition 10. *The distortion-dispersion function of symbol-by-symbol joint source-channel coding is defined as*

$$\mathcal{W}_1(\alpha) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{n\left(D\left(C(\alpha)\right) - D_1(n, \epsilon, \alpha)\right)^2}{2\log_e \frac{1}{\epsilon}} \quad (188)$$

where $D(\cdot)$ is the distortion-rate function of the source.

As before, if there is no channel input-cost constraint ($\mathsf{c}^n(x^n) = 0$ for all $x^n \in \mathcal{A}^n$), we will simplify the notation and write $D_1(n, \epsilon)$ for $D_1(n, \epsilon, \alpha)$ and $\mathcal{W}_1$ for $\mathcal{W}_1(\alpha)$.

In addition to restrictions (i)–(iv) of Section V, we assume that the channel and the source are probabilistically matched in the following sense (cf. [7]).

(v) There exist α, $P_{\mathsf{X}^\star|\mathsf{S}}$ and $P_{\mathsf{Z}^\star|\mathsf{Y}}$ such that $P_{\mathsf{X}^\star}$ and $P_{\mathsf{Z}^\star|\mathsf{S}}$ generated by the joint distribution $P_\mathsf{S}P_{\mathsf{X}^\star|\mathsf{S}}P_{\mathsf{Y}|\mathsf{X}}P_{\mathsf{Z}^\star|\mathsf{Y}}$ achieve the capacity-cost function $C(\alpha)$ and the distortion-rate function $D(C(\alpha))$, respectively.

Condition (v) ensures that symbol-by-symbol transmission attains the minimum average (over source realizations) distortion achievable among all codes of any blocklength. The following results pertain to the full distribution of the distortion incurred at the receiver output and not just its mean.

Theorem 20 (Achievability, symbol-by-symbol code). *Under restrictions* (i)-(v)*, if*

$$\mathbb{P}\left[\sum_{i=1}^{n}\mathsf{d}(S_i, Z_i^\star) > nd\right] \leq \epsilon \qquad (189)$$

where $P_{Z^{n\star}|S^n} = P_{\mathsf{Z}^\star|\mathsf{S}} \times \ldots \times P_{\mathsf{Z}^\star|\mathsf{S}}$, and $P_{\mathsf{Z}^\star|\mathsf{S}}$ achieves $D(C(\alpha))$, then there exists an (n, d, ϵ, α) symbol-by-symbol code (average cost constraint).

Proof: If (v) holds, then there exist a symbol-by-symbol encoder and decoder such that the conditional distribution of the output of the decoder given the source outcome coincides with distribution $P_{\mathsf{Z}^\star|\mathsf{S}}$, so the excess-distortion probability of this symbol-by-symbol code is given by the left side of (189). ∎

Theorem 21 (Converse, symbol-by-symbol code). *Under restriction* (i) *and separable distortion measure, the parameters of any (n, d, ϵ, α) symbol-by-symbol code (average cost constraint) must satisfy*

$$\epsilon \geq \inf_{\substack{P_{\mathsf{Z}|\mathsf{S}}: \\ I(\mathsf{S};\mathsf{Z}) \leq C(\alpha)}} \mathbb{P}\left[\sum_{i=1}^{n}\mathsf{d}(S_i, Z_i) > nd\right] \qquad (190)$$

where $P_{Z^n|S^n} = P_{\mathsf{Z}|\mathsf{S}} \times \ldots \times P_{\mathsf{Z}|\mathsf{S}}$.

Proof: The excess-distortion probability at blocklength n, distortion d and cost α achievable among all single-letter codes $(P_{\mathsf{X}|\mathsf{S}}, P_{\mathsf{Z}|\mathsf{Y}})$ must satisfy

$$\epsilon \geq \inf_{\substack{P_{\mathsf{X}|\mathsf{S}}, P_{\mathsf{Z}|\mathsf{Y}}: \\ \mathsf{S}-\mathsf{X}-\mathsf{Y}-\mathsf{Z} \\ \mathbb{E}[\mathsf{c}(\mathsf{X})] \leq \alpha}} \mathbb{P}[\mathsf{d}_n(S^n, Z^n) > d] \qquad (191)$$

$$\geq \inf_{\substack{P_{\mathsf{X}|\mathsf{S}}, P_{\mathsf{Z}|\mathsf{Y}}: \\ \mathbb{E}[\mathsf{c}(\mathsf{X})] \leq \alpha \\ I(\mathsf{S};\mathsf{Z}) \leq I(\mathsf{X};\mathsf{Y})}} \mathbb{P}[\mathsf{d}_n(S^n, Z^n) > d] \qquad (192)$$

where (192) holds since $\mathsf{S} - \mathsf{X} - \mathsf{Y} - \mathsf{Z}$ implies $I(\mathsf{S};\mathsf{Z}) \leq I(\mathsf{X};\mathsf{Y})$ by the data processing inequality. The right side of (192) is lower bounded by the right side of (190) because $I(\mathsf{X};\mathsf{Y}) \leq C(\alpha)$ holds for all P_X with $\mathbb{E}[\mathsf{c}(\mathsf{X})] \leq \alpha$, and the distortion measure is separable. ∎

Theorem 22 (Gaussian approximation, optimal symbol-by-symbol code). *Assume $\mathbb{E}[\mathsf{d}^3(\mathsf{S}, \mathsf{Z}^\star)] < \infty$. Under restrictions* (i)-(v)*,*

$$D_1(n, \epsilon, \alpha) = D(C(\alpha)) + \sqrt{\frac{\mathcal{W}_1(\alpha)}{n}}Q^{-1}(\epsilon) + \frac{\theta_1(n)}{n} \qquad (193)$$

$$\mathcal{W}_1(\alpha) = \mathrm{Var}[\mathsf{d}(\mathsf{S}, \mathsf{Z}^\star)] \qquad (194)$$

where

$$\theta_1(n) \leq O(1) \qquad (195)$$

Moreover, if there is no power constraint,

$$\theta_1(n) \geq \frac{D'(R)}{R^2}\theta(n) \qquad (196)$$

$$\mathcal{W}_1 = \mathcal{W}(1) \qquad (197)$$

where $\theta(n)$ is that in Theorem 10.
If $\mathrm{Var}[\mathsf{d}(\mathsf{S}, \mathsf{Z}^\star)] > 0$ and $\mathcal{S}$, $\hat{\mathcal{S}}$ are finite, then

$$\theta_1(n) \geq O(1) \qquad (198)$$

Proof: Since the third absolute moment of $\mathsf{d}(S_i, Z_i^\star)$ is finite, the achievability part of the result, namely, (193) with the remainder satisfying (195), follows by a straightforward application of the Berry-Esseen bound to (189), provided that $\mathrm{Var}[\mathsf{d}(S_i, Z_i^\star)] > 0$. If $\mathrm{Var}[\mathsf{d}(S_i, Z_i^\star)] = 0$, it follows trivially from (189).

To show the converse in (196), observe that since the set of all (n, n, d, ϵ) codes includes all (n, d, ϵ) symbol-by-symbol codes, we have $D(n, n, \epsilon) \leq D_1(n, \epsilon)$. Since $Q^{-1}(\epsilon)$ is positive or negative depending on whether $\epsilon < \frac{1}{2}$ or $\epsilon > \frac{1}{2}$, using (130) we conclude that we must necessarily have (197), which is, in fact, a consequence of conditions (b), (c) in Section II and (v). Now, (196) is simply the converse part of (129).

The proof of the refined converse in (198) is relegated to Appendix E. ∎

In the absence of a cost constraint, Theorem 22 shows that if the source and the channel are probabilistically matched in the sense of [7], then not only does symbol-by-symbol transmission achieve the minimum average distortion, but also the dispersion of JSCC (see (197)). In other words, not only do such symbol-by-symbol codes attain the minimum average distortion but also the variance of distortions at the decoder's output is the minimum achievable among all codes operating at that average distortion. In contrast, if there is an average cost constraint, the symbol-by-symbol codes considered in Theorem 22 probably do not attain the minimum excess distortion achievable among all blocklength-n codes, not even asymptotically. Indeed, as observed in [23], for the transmission of an equiprobable source over an AWGN channel under the average power constraint and the average block error probability performance criterion, the strong converse does not hold and the second-order term is of order $n^{-\frac{1}{3}}$, not $n^{-\frac{1}{2}}$, as in (193).

Two conspicuous examples that satisfy the probabilistic matching condition (v), so that symbol-by-symbol coding is

optimal in terms of average distortion, are the transmission of a binary equiprobable source over a binary-symmetric channel provided the desired bit error rate is equal to the crossover probability of the channel [24, Sec.11.8], [25, Problem 7.16], and the transmission of a Gaussian source over an additive white Gaussian noise channel under the mean-square error distortion criterion, provided that the tolerable source signal-to-noise ratio attainable by an estimator is equal to the signal-to-noise ratio at the output of the channel [26]. We dissect these two examples next. After that, we will discuss two additional examples where uncoded transmission is optimal.

B. Uncoded transmission of a BMS over a BSC

In the setup of Section VI, if the binary source is unbiased $\left(p = \frac{1}{2}\right)$, then $C = 1 - h(\delta)$, $R(d) = 1 - h(d)$, and $D(C) = \delta$. If the encoder and the decoder are both identity mappings (uncoded transmission), the resulting joint distribution satisfies condition (v). As is well known, regardless of the blocklength, the uncoded symbol-by-symbol scheme achieves the minimum bit error rate (averaged over source and channel). Here, we are interested instead in examining the excess distortion probability criterion. For example, consider an application where, if the fraction of erroneously received bits exceeds a certain threshold, then the entire output packet is useless.

Using (130) and (194), it is easy to verify that

$$\mathcal{W}(1) = \mathcal{W}_1 = \delta(1 - \delta) \tag{199}$$

that is, uncoded transmission is optimal in terms of dispersion, as anticipated in (197). Moreover, the uncoded transmission attains the minimum bit error rate threshold $D(n, n, \epsilon)$ achievable among all codes operating at blocklength n, regardless of the allowed ϵ, as the following result demonstrates.
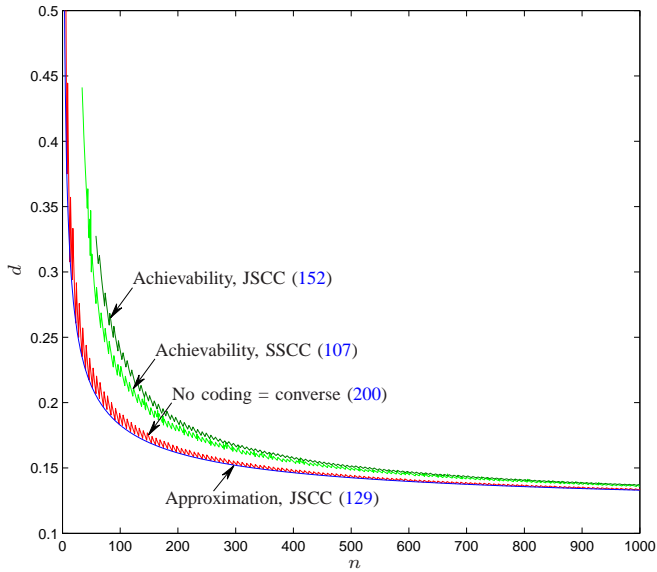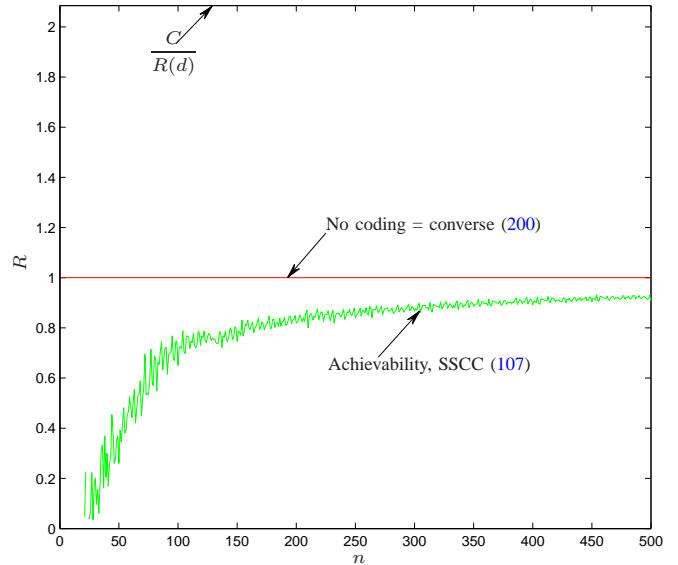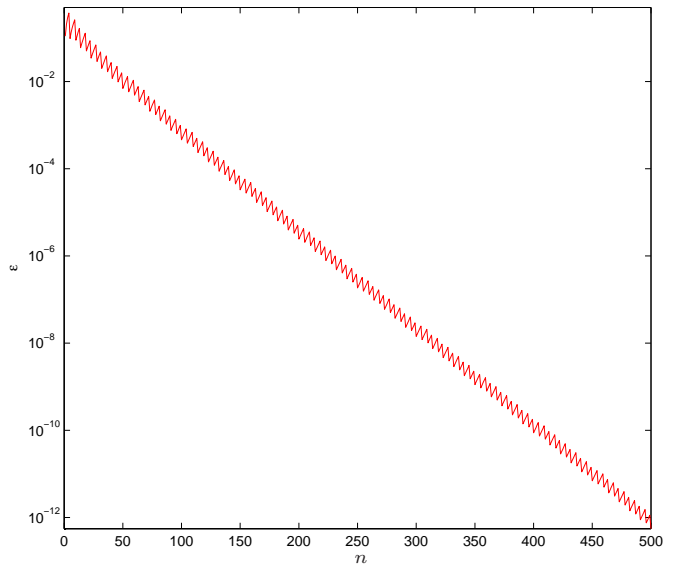
Fig. 8. Distortion-blocklength tradeoff for the transmission of a fair BMS over a BSC with crossover probability $\delta = 0.11$ and $R = 1$, $\epsilon = 10^{-2}$.

Theorem 23 (BMS-BSC, symbol-by-symbol code). *Consider the the symbol-by-symbol scheme which is uncoded if $p \geq \delta$*

Fig. 9. Rate-blocklength tradeoff (a) for the transmission of a fair BMS over a BSC with crossover probability $\delta = 0.11$ and $d = 0.22$. The excess-distortion probability ϵ is set to be the one achieved by the uncoded scheme (b).

and whose decoder always outputs the all-zero vector if $p < \delta$. It achieves, at blocklength n and excess distortion probability ϵ, regardless of $0 \leq p \leq \frac{1}{2}$, $\delta \leq \frac{1}{2}$,

$$D_1(n, \epsilon) =$$

$$\min \left\{ d : \sum_{t=0}^{\lfloor nd \rfloor} \binom{n}{t} \min\{p, \delta\}^t (1 - \min\{p, \delta\})^{n-t} \geq 1 - \epsilon \right\} \tag{200}$$

Moreover, if the source is equiprobable $\left(p = \frac{1}{2}\right)$,

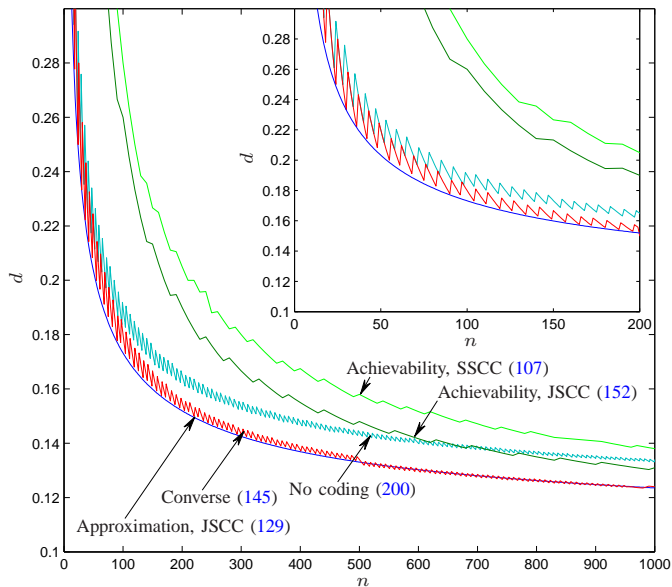$$D_1(n, \epsilon) = D(n, n, \epsilon) \tag{201}$$

Fig. 10. Distortion-blocklength tradeoff for the transmission of a BMS with $p = \frac{2}{5}$ over a BSC with crossover probability $\delta = 0.11$ and $R = 1$, $\epsilon = 10^{-2}$.

Proof: Direct calculation yields (200). To show (201), let us compare $d^\star = D_1(n, \epsilon)$ with the conditions imposed on d by Theorem 13. Comparing (200) to (150), we see that either (a) equality in (200) is achieved, $r^\star = nd^\star$, $\lambda = 0$, and (plugging $k = n$ into (149))

$$\left\langle \begin{array}{c} n \\ nd^\star \end{array} \right\rangle \leq \left\langle \begin{array}{c} n \\ \lfloor nd \rfloor \end{array} \right\rangle \tag{202}$$

thereby implying that $d \geq d^\star$, or
(b) $r^\star = nd^\star - 1$, $\lambda > 0$, and (149) becomes

$$\lambda \binom{n}{nd^\star} + \left\langle \begin{array}{c} n \\ nd^\star - 1 \end{array} \right\rangle \leq \left\langle \begin{array}{c} n \\ \lfloor nd \rfloor \end{array} \right\rangle \tag{203}$$

which also implies $d \geq d^\star$. To see this, note that $d < d^\star$ would imply $\lfloor nd \rfloor \leq nd^\star - 1$ since $nd^\star$ is an integer, which in turn would require (according to (203)) that $\lambda \leq 0$, which is impossible. ∎

For the transmission of the fair binary source over a BSC, Fig. 8 shows the distortion achieved by the uncoded scheme, the separated scheme and the JSCC scheme of Theorem 14 versus n for a fixed excess-distortion probability $\epsilon = 0.01$. The no coding / converse curve in Fig. 8 depicts one of those singular cases where the non-asymptotic fundamental limit can be computed precisely. As evidenced by this curve, the fundamental limits need not be monotonic with blocklength.

Figure 9(a) shows the rate achieved by separate coding when $d > \delta$ is fixed, and the excess-distortion probability ϵ, shown in Fig. 9(b), is set to be the one achieved by uncoded transmission, namely, (200). Figure 9(a) highlights the fact that at short blocklengths (say $n \leq 100$) separate source/channel coding is vastly suboptimal. As the blocklength increases, the performance of the separated scheme approaches that of the no-coding scheme, but according to Theorem 23 it can never outperform it. Had we allowed the excess distortion

probability to vanish sufficiently slowly, the JSCC curve would have approached the Shannon limit as $n \to \infty$. However, in Figure 9(a), the exponential decay in ϵ is such that there is indeed an asymptotic rate penalty as predicted in [11].

For the biased binary source with $p = \frac{2}{5}$ and BSC with crossover probability 0.11, Figure 10 plots the maximum distortion achieved with probability 0.99 by the uncoded scheme, which in this case is asymptotically suboptimal. Nevertheless, uncoded transmission performs remarkably well in the displayed range of blocklengths, achieving the converse almost exactly at blocklengths less than 100, and outperforming the JSCC achievability result in Theorem 14 at blocklengths as long as 700. This example substantiates that even in the absence of a probabilistic match between the source and the channel, symbol-by-symbol transmission, though asymptotically suboptimal, might outperform SSCC and even our random JSCC achievability bound in the finite blocklength regime.

C. Symbol-by-symbol coding for lossy transmission of a GMS over an AWGN channel

In the setup of Section VII, using (163) and (164), we find that

$$D(C(P)) = \frac{\sigma_S^2}{1 + P} \tag{204}$$

The next result characterizes the distribution of the distortion incurred by the symbol-by-symbol scheme that attains the minimum average distortion.

Theorem 24 (GMS-AWGN, symbol-by-symbol code). *The following symbol-by-symbol transmission scheme in which the encoder and the decoder are the amplifiers:*

$$\mathsf{f}_1(\mathsf{s}) = a\mathsf{s}, \quad a^2 = \frac{P\sigma_N^2}{\sigma_S^2} \tag{205}$$

$$\mathsf{g}_1(\mathsf{y}) = b\mathsf{y}, \quad b = \frac{a\sigma_S^2}{a^2\sigma_S^2 + \sigma_N^2} \tag{206}$$

is an (n, d, ϵ, P) symbol-by-symbol code (with average cost constraint) such that

$$\mathbb{P}\left[W_0^n D(C(P)) > nd\right] = \epsilon \tag{207}$$

where W_0^n is chi-square distributed with n degrees of freedom.

Note that (207) is a particularization of (189). Using (207), we find that

$$\mathcal{W}_1(P) = 2\frac{\sigma_S^4}{(1 + P)^2} \tag{208}$$

On the other hand, using (130), we compute

$$\mathcal{W}(1, P) = 2\frac{\sigma_S^4}{(1 + P)^2}\left(2 - \frac{1}{(1 + P)^2}\right) \tag{209}$$

$$> \mathcal{W}_1(P) \tag{210}$$

The difference between (210) and (197) is due to the fact that the optimal symbol-by-symbol code in Theorem 24 obeys an average power constraint, rather than the more stringent maximal power constraint of Theorem 10, so it is not surprising

that for the practically interesting case $\epsilon < \frac{1}{2}$ the symbol-by-symbol code can outperform the best code obeying the maximal power constraint. Indeed, in the range of blocklenghts displayed in Figure 11, the symbol-by-symbol code even outperforms the converse for codes operating under a maximal power constraint.
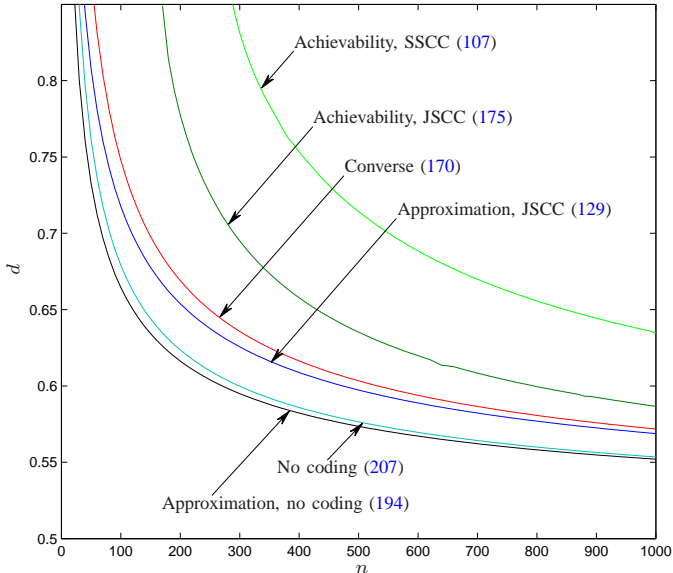

Fig. 11. Distortion-blocklength tradeoff for the transmission of a GMS over an AWGN channel with $\frac{P}{\sigma_N^2} = 1$ and $R = 1$, $\epsilon = 10^{-2}$.

D. Uncoded transmission of a discrete memoryless source (DMS) over a discrete erasure channel (DEC) under erasure distortion measure

For a discrete source, the single-letter erasure distortion measure is defined as the following mapping $\mathsf{d}\colon \mathcal{S} \times \{\mathcal{S}, \mathsf{e}\} \mapsto [0, \infty]$:[8]

$$\mathsf{d}(\mathsf{s}, \mathsf{z}) = \begin{cases} 0 & \mathsf{z} = \mathsf{s} \\ H(\mathsf{S}) & \mathsf{z} = \mathsf{e} \\ \infty & \text{otherwise} \end{cases} \tag{211}$$

For any $0 \le d \le H(\mathsf{S})$, the rate-distortion function of the equiprobable source is achieved by

$$P_{\mathsf{Z}^\star|\mathsf{S}=\mathsf{s}}(\mathsf{z}) = \begin{cases} 1 - \frac{d}{H(\mathsf{S})} & \mathsf{z} = \mathsf{s} \\ \frac{d}{H(\mathsf{S})} & \mathsf{z} = \mathsf{e} \end{cases} \tag{212}$$

The rate-distortion function and the d-tilted information for the equiprobable source with the erasure distortion measure are given by, respectively,

$$R(d) = H(\mathsf{S}) - d \tag{213}$$
$$\jmath_\mathsf{S}(\mathsf{s}, d) = \imath_\mathsf{S}(\mathsf{s}) - d \tag{214}$$

Note that, trivially, $\jmath_\mathsf{S}(\mathsf{S}, d) = R(d) = \log|\mathcal{S}| - d$ a.s. The channel that is matched to the equiprobable DMS with the

[8]The distortion measure in (211) is a scaled version of the erasure distortion measure found in literature, e.g. [27].

erasure distortion measure is the DEC, whose single-letter transition probability kernel $P_{\mathsf{Y}|\mathsf{X}}\colon \mathcal{A} \mapsto \{\mathcal{A}, \mathsf{e}\}$ is

$$P_{\mathsf{Y}|\mathsf{X}=\mathsf{x}}(\mathsf{y}) = \begin{cases} 1 - \delta & \mathsf{y} = \mathsf{x} \\ \delta & \mathsf{y} = \mathsf{e} \end{cases} \tag{215}$$

and whose capacity is given by $C = \log|\mathcal{A}| - \delta$, achieved by equiprobable $P_{\mathsf{X}^\star}$. For P_S equiprobable on $\mathcal{S} = \mathcal{A}$, we find that $D(C) = \delta \log|\mathcal{S}|$, and

$$\mathcal{W}_1 = \delta(1 - \delta)\log^2|\mathcal{S}| \tag{216}$$

E. Symbol-by-symbol transmission of a DMS over a DEC under logarithmic loss

Let the source alphabet $\mathcal{S}$ be finite, and let the reproduction alphabet $\hat{\mathcal{S}}$ be the set of all probability distributions on $\mathcal{S}$. The single-letter logarithmic loss distortion measure $\mathsf{d}\colon \mathcal{S} \times \hat{\mathcal{S}} \mapsto \mathbb{R}^+$ is defined by [28], [29]

$$\mathsf{d}(s, P_\mathsf{Z}) = \imath_\mathsf{Z}(s) \tag{217}$$

Curiously, for any $0 \le d \le H(\mathsf{S})$, the rate-distortion function and the d-tilted information are given respectively by (213) and (214), even if the source is not equiprobable. In fact, the rate-distortion function is achieved by,

$$P_{P_\mathsf{Z}^\star|\mathsf{S}=s}(P_\mathsf{Z}) = \begin{cases} \frac{d}{H(\mathsf{S})} & P_\mathsf{Z} = P_\mathsf{S} \\ 1 - \frac{d}{H(\mathsf{S})} & P_\mathsf{Z} = 1_\mathsf{S}(s) \end{cases} \tag{218}$$

and the channel that is matched to the equiprobable source under logarithmic loss is exactly the DEC in (215). Of course, unlike Section VIII-D, the decoder we need is a simple one-to-one function that outputs P_S if the channel output is e, and $1_\mathsf{S}(\mathsf{y})$ otherwise, where $\mathsf{y} \neq \mathsf{e}$ is the output of the DEC. Finally, it is easy to verify that the distortion-dispersion function of symbol-by-symbol coding under logarithmic loss is the same as that under erasure distortion and is given by (216).

IX. CONCLUSION

In this paper we gave a non-asymptotic analysis of joint source-channel coding including several achievability and converse bounds, which hold in wide generality and are tight enough to determine the dispersion of joint source-channel coding for the transmission of an abstract memoryless source over either a DMC or a Gaussian channel, under an arbitrary fidelity measure. We also investigated the penalty incurred by separate source-channel coding using both the source-channel dispersion and the particularization of our new bounds to (i) the binary source and the binary symmetric channel with bit error rate fidelity criterion and (ii) the Gaussian source and Gaussian channel under mean-square error distortion. Finally, we showed cases where symbol-by-symbol (uncoded) transmission beats any other known scheme in the finite blocklength regime even when the source-channel matching condition is not satisfied.

The approach taken in this paper to analyze the non-asymptotic fundamental limits of lossy joint source-channel coding is two-fold. Our new achievability and converse bounds apply to abstract sources and channels and allow for memory,

while the asymptotic analysis of the new bounds leading to the dispersion of JSCC is focused on the most basic scenario of transmitting a stationary memoryless source over a stationary memoryless channel.

The major results and conclusions are the following.

1) A general new converse bound (Theorem 3) leverages the concept of d-tilted information (Definition 5), a random variable which corresponds (in a sense that can be formalized [9], [30]) to the number of bits required to represent a given source outcome within distortion d and whose role in lossy compression is on a par with that of information (in (20)) in lossless compression.

2) The converse result in Theorem 4 capitalizes on two simple observations, namely, that any (d, ϵ) lossy code can be converted to a list code with list error probability ϵ, and that a binary hypothesis test between P_{SXY} and an auxiliary distribution on the same space can be constructed by choosing P_{SXY} when there is no list error. We have generalized the conventional notion of list, to allow the decoder to output a possibly uncountable set of source realizations.

3) As evidenced by our numerical results, the converse result in Theorem 5, which applies to those channels satisfying a certain symmetry condition and which is a consequence of the hypothesis testing converse in Theorem 4, can outperform the d-tilted information converse in Theorem 3. Nevertheless, it is Theorem 3 that lends itself to analysis more easily and that leads to the JSCC dispersion for the general DMC.

4) Our random-coding-based achievability bound (Theorem 7) provides insights into the degree of separation between the source and the channel codes required for optimal performance in the finite blocklength regime. More precisely, it reveals that the dispersion of JSCC can be achieved in the class of (M, d, ϵ) JSCC codes (Definition 8). As in separate source/channel coding, in (M, d, ϵ) coding the inner channel coding block is connected to the outer source coding block by a noiseless link of capacity $\log M$, but unlike SSCC, the channel (resp. source) code can be chosen based on the knowledge of the source (resp. channel). The conventional SSCC in which the source code is chosen without knowledge of the channel and the channel code is chosen without knowledge of the source, although known to achieve the asymptotic fundamental limit of joint source-channel coding under certain quite weak conditions, is in general suboptimal in the finite blocklength regime.

5) Since $\mathbb{E}[\mathsf{d}(S, Z)] = \int_0^\infty \mathbb{P}[\mathsf{d}(S, Z) > \xi] d\xi$, bounds for average distortion can be obtained by integrating our bounds on excess distortion. Note, however, that the code that minimizes $\mathbb{P}[\mathsf{d}(S, Z) > \xi]$ depends on ξ. Since the distortion cdf of any single code does not majorize the cdfs of all possible codes, the converse bound on the average distortion obtained through this approach, although asymptotically tight, may be loose at short blocklengths. Likewise, regarding achievability bounds (e.g. (108)), the optimization over channel and source random codes, P_X and P_Z, must be performed after the integration, so that the choice of code does not depend on the distortion threshold ξ.

6) For the transmission of a stationary memoryless source over a stationary memoryless channel, the Gaussian approximation in Theorem 10 (neglecting the remainder $\theta(n)$) provides a simple estimate of the maximal non-asymptotically achievable joint source-channel coding rate. Appealingly, the dispersion of joint source-channel coding decomposes into two terms, the channel dispersion and the source dispersion. Thus, only two channel attributes, the capacity and dispersion, and two source attributes, the rate-distortion and rate-dispersion functions, are required to compute the Gaussian approximation to the maximal JSCC rate.

7) In those curious cases where the source and the channel are probabilistically matched so that symbol-by-symbol coding attains the minimum possible average distortion, Theorem 22 ensures that it also attains the dispersion of joint source-channel coding, that is, symbol-by-symbol coding results in the minimum variance of distortions among all codes operating at that average distortion.

8) Even in the absence of a probabilistic match between the source and the channel, symbol-by-symbol transmission, though asymptotically suboptimal, might outperform separate source-channel coding and joint source-channel random coding in the finite blocklength regime.

APPENDIX A
THE BERRY-ESSEEN THEOREM

The following result is an important tool in the Gaussian approximation analysis.

Theorem 25 (Berry-Esseen CLT, e.g. [31, Ch. XVI.5 Theorem 2]). *Fix a positive integer n. Let W_i, $i = 1, \ldots, n$ be independent. Then, for any real t*

$$\left| \mathbb{P}\left[\sum_{i=1}^n W_i > n\left(D_n + t\sqrt{\frac{V_n}{n}} \right) \right] - Q(t) \right| \le \frac{B_n}{\sqrt{n}}, \quad (219)$$

where

$$D_n = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[W_i] \quad (220)$$

$$V_n = \frac{1}{n} \sum_{i=1}^n \mathrm{Var}[W_i] \quad (221)$$

$$T_n = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[|W_i - \mathbb{E}[W_i]|^3 \right] \quad (222)$$

$$B_n = \frac{c_0 T_n}{V_n^{3/2}} \quad (223)$$

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

APPENDIX B
AUXILIARY RESULT ON THE MINIMIZATION OF THE INFORMATION SPECTRUM

Given a finite set $\mathcal{A}$, 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 $nP_X(a)$. Let $\mathcal{P}$ be the set of all distributions on $\mathcal{A}$, which is simply the standard $|\mathcal{A}| - 1$ simplex in $\mathbb{R}^{|\mathcal{A}|}$. For an arbitrary subset $\mathcal{D} \subseteq \mathcal{P}$, denote by $\mathcal{D}_{[n]}$ the set of distributions in $\mathcal{D}$ that are also n-types, that is,

$$\mathcal{D}_{[n]} = \{P_X \in \mathcal{D} \colon \forall a \in \mathcal{A}, \ nP_X(a) \text{ is an integer}\} \quad (224)$$

Denote by $\Pi(P_X)$ the minimum Euclidean distance approximation of $P_X \in \mathcal{P}$ in the set of n-types, that is,

$$\Pi(P_X) = \arg \min_{\hat{P}_X \in \mathcal{P}_{[n]}} \left| P_X - \hat{P}_X \right| \quad (225)$$

Let $\mathcal{P}^\star$ be the set of capacity-achieving distributions:[9]

$$\mathcal{P}^\star = \{P_X \in \mathcal{P} \colon I(X;Y) = C\} \quad (226)$$

Denote the minimum (maximum) information variances achieved by the distributions in $\mathcal{P}^\star$ by

$$V_{\min} = \min_{P_X \in \mathcal{P}^\star} \text{Var}\left[\imath_{X;Y}(X;Y)\right] \quad (227)$$

$$V_{\max} = \max_{P_X \in \mathcal{P}^\star} \text{Var}\left[\imath_{X;Y}(X;Y)\right] \quad (228)$$

and let $\mathcal{P}^\star_{\min} \subseteq \mathcal{P}^\star$ be the set of capacity-achieving distributions that achieve the minimum information variance:

$$\mathcal{P}^\star_{\min} = \{P_X \in \mathcal{P}^\star \colon \text{Var}\left[\imath_{X;Y}(X;Y)\right] = V_{\min}\} \quad (229)$$

and analogously $\mathcal{P}^\star_{\max}$ for the distributions in $\mathcal{P}^\star$ with maximal variance. Lemma 1 below allows to show that in the memoryless case, the infimum inside the expectation in (39) with $W = \text{type}(X^n)$ and $P_{\bar{Y}^n | W = P_X} = P_Y \times \ldots \times P_Y$, where P_Y is the output distribution induced by the type P_X, is approximately attained by those sequences whose type is closest to the capacity-achieving distribution $P_{X^\star}$ (if it is non-unique, $P_{X^\star}$ is chosen appropriately based on the information variance it achieves). This technical result is the key to proving the converse part of Theorem 10.

Lemma 1. *There exist $\bar{\Delta} > 0$ such that for all sufficiently large n:*

1) If $V_{\min} > 0$, then there exists $K > 0$ such that for $|\Delta| \leq \bar{\Delta}$,

$$\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^{n} \imath_{X;Y}(x_i; Y_i) \leq n(C - \Delta)\right]$$
$$\geq \mathbb{P}\left[\sum_{i=1}^{n} \imath_{X;Y}(x_i^\star; Y_i) \leq n(C - \Delta)\right] - \frac{K}{\sqrt{n}} \quad (230)$$

where (230) holds for any $x^{n\star}$ with $\text{type}(x^{n\star}) = \Pi(P_X^\star)$ for $P_X^\star \in \mathcal{P}^\star_{\min}$ if $\Delta \geq 0$ and $P_X^\star \in \mathcal{P}^\star_{\max}$ if $\Delta < 0$.

2) If $V_{\max} = 0$, then for all $0 < \alpha < \frac{3}{2}$ and $\Delta \geq \frac{\bar{\Delta}}{n^{\frac{1}{2} + \alpha}}$,

$$\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^{n} \imath_{X;Y}(x_i; Y_i) \leq n(C + \Delta)\right] \geq 1 - \frac{1}{n^{\frac{1}{4} - \frac{3}{2}\alpha}} \quad (231)$$

The information densities in the left sides of (230) and (231) are computed with $\{P_{Y|X=x_i}, P_Y\}$, where P_Y is induced by the type of x^n, i.e. $\text{type}(x^n) = P_X \to P_{Y|X} \to P_Y$, and that in the right side of (230) is computed with $\{P_{Y|X=x_i^\star}, P_Y\}$, where P_Y is induced by the type of $x^{n\star}$, i.e. $\text{type}(x^{n\star}) = P_X \to P_{Y|X} \to P_Y$. The independent random variables Y_i in the left sides of (230) and (231) have distribution $P_{Y|X=x_i}$, while Y_i in the right side of (230) have distribution $P_{Y|X=x_i^\star}$.

In order to prove Lemma 1, we first show three auxiliary lemmas. The first two deal with approximate optimization of functions.

If f and g approximate each other, and the minimum of f is approximately attained at x, then g is also approximately minimized at x, as the following lemma formalizes.

Lemma 2. *Fix $\eta > 0$, $\xi > 0$. Let $\mathcal{D}$ be an arbitrary set, and let $f \colon \mathcal{D} \mapsto \mathbb{R}$ and $g \colon \mathcal{D} \mapsto \mathbb{R}$ be such that*

$$\sup_{x \in \mathcal{D}} |f(x) - g(x)| \leq \eta \quad (232)$$

Further, assume that f and g attain their minima. Then,

$$g(x) \leq \min_{y \in \mathcal{D}} g(y) + \xi + 2\eta \quad (233)$$

as long as x satisfies

$$f(x) \leq \min_{y \in \mathcal{D}} f(y) + \xi \quad (234)$$

(see Fig. 12).
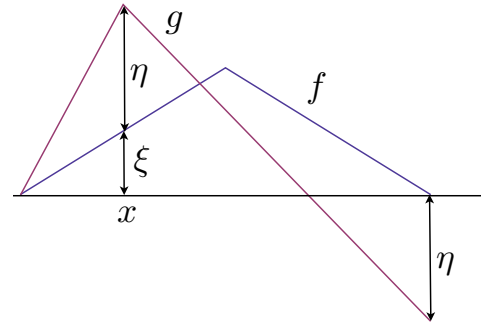
Fig. 12. An example where (233) holds with equality.

Proof of Lemma 2: Let $x^\star \in \mathcal{D}$ be such that $g(x^\star) = \min_{y \in \mathcal{D}} g(y)$. Using (232) and (234), write

$$g(x) \leq \min_{y \in \mathcal{D}} f(y) + g(x) - f(x) + \xi \quad (235)$$

$$\leq \min_{y \in \mathcal{D}} f(y) + \eta + \xi \quad (236)$$

$$\leq f(x^\star) + \eta + \xi \quad (237)$$

$$= g(x^\star) - g(x^\star) + f(x^\star) + \eta + \xi \quad (238)$$

$$\leq g(x^\star) + 2\eta + \xi \quad (239)$$

The following lemma is reminiscent of [8, Lemma 64].

Lemma 3. *Let $\mathcal{D}$ be a compact metric space, and let $d\colon \mathcal{D}^2 \to \mathbb{R}^+$ be a metric. Fix $f\colon \mathcal{D} \mapsto \mathbb{R}$ and $g\colon \mathcal{D} \mapsto \mathbb{R}$. Let*

$$\mathcal{D}^\star = \left\{ x \in \mathcal{D}\colon f(x) = \max_{y \in \mathcal{D}} f(y) \right\} \tag{240}$$

Suppose that for some constants $\ell > 0, L > 0$, we have, for all $(x, x^\star) \in \mathcal{D} \times \mathcal{D}^\star$,

$$f(x^\star) - f(x) \geq \ell d^2(x, x^\star) \tag{241}$$
$$|g(x^\star) - g(x)| \leq L d(x, x^\star) \tag{242}$$

Then, for any positive scalars φ, ψ,

$$\max_{x \in \mathcal{D}} \left\{ \varphi f(x) \pm \psi g(x) \right\} \leq \varphi f(x^\star) \pm \psi g(x^\star) + \frac{L^2 \psi^2}{4\ell\varphi} \tag{243}$$

Moreover, if, instead of (241), f satisfies

$$f(x^\star) - f(x) \geq \ell d(x, x^\star) \tag{244}$$

then, for any positive scalars ψ, φ such that

$$L\psi \leq \ell\varphi \tag{245}$$

we have

$$\max_{x \in \mathcal{D}} \left\{ \varphi f(x) \pm \psi g(x) \right\} = \varphi f(x^\star) \pm \psi g(x^\star) \tag{246}$$

Proof of Lemma 3: Let x_0 achieve the maximum on the left side of (243). Using (241) and (242), we have, for all $x^\star \in \mathcal{D}^\star$,

$$0 \leq \varphi \left(f(x_0) - f(x^\star) \right) \pm \psi \left(g(x_0) - g(x^\star) \right) \tag{247}$$
$$\leq -\ell\varphi d^2(x_0, x^\star) + L\psi d(x_0, x^\star) \tag{248}$$
$$\leq \frac{L^2 \psi^2}{4\ell\varphi} \tag{249}$$

where (249) follows because the maximum of (248) is achieved at $d(x_0, x^\star) = \frac{L\psi}{2\ell\varphi}$.

To show (246), observe using (244) and (242) that

$$0 \leq \varphi \left(f(x_0) - f(x^\star) \right) \pm \psi \left(g(x_0) - g(x^\star) \right) \tag{250}$$
$$\leq \left(-\ell\varphi + L\psi \right) d(x_0, x^\star) \tag{251}$$
$$\leq 0 \tag{252}$$

where (252) follows from (245). ∎

The following lemma deals with asymptotic behavior of the Q-function.

Lemma 4. *Fix $a \geq 0$, $b \geq 0$. Then, there exists $q \geq 0$ (explicitly computed in the proof) such that for all $z \geq -\frac{\sqrt{n}}{2b}$ and all n large enough,*

$$Q\left(z - \frac{a}{\sqrt{n}} \right) - Q\left(z + \frac{b}{\sqrt{n}} z^2 \right) \leq \frac{q}{\sqrt{n}} \tag{253}$$

Proof of Lemma 4:

$Q(x)$ is convex for $x \geq 0$, and $Q'(x) = -\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, so for $x \geq 0, \xi \geq 0$

$$Q(x + \xi) \geq Q(x) - \frac{\xi}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{254}$$

while for arbitrary x and $\xi \geq 0$,

$$Q(x + \xi) \geq Q(x) - \frac{\xi}{\sqrt{2\pi}} \tag{255}$$

If $z \geq \frac{a}{\sqrt{n}}$, we use (254) to obtain

$$Q\left(z - \frac{a}{\sqrt{n}} \right) - Q\left(z + \frac{b}{\sqrt{n}} z^2 \right) \tag{256}$$
$$\leq \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(z - \frac{a}{\sqrt{n}} \right)^2}{2}} \left(\frac{b}{\sqrt{n}} z^2 + \frac{a}{\sqrt{n}} \right) \tag{257}$$
$$\leq \frac{bz^2}{\sqrt{2\pi n}} e^{-\frac{\left(z - \frac{a}{\sqrt{n}} \right)^2}{2}} + \frac{a}{\sqrt{2\pi n}} \tag{258}$$
$$\leq \frac{3b\, e^{-1} + a}{\sqrt{2\pi n}} \tag{259}$$

where (259) holds for n large enough because the maximum of (258) is attained at $z = \sqrt{2 + \frac{a}{4n}} + \frac{a}{2\sqrt{n}}$.

If $0 \leq z \leq \frac{a}{\sqrt{n}}$, we use (255) to obtain

$$Q\left(z - \frac{a}{\sqrt{n}} \right) - Q\left(z + \frac{b}{\sqrt{n}} z^2 \right)$$
$$\leq \frac{1}{\sqrt{2\pi}} \left(\frac{b}{\sqrt{n}} z^2 + \frac{a}{\sqrt{n}} \right) \tag{260}$$
$$\leq \frac{a}{\sqrt{2\pi n}} \left(1 + \frac{ab}{n} \right) \tag{261}$$

If $-\frac{\sqrt{n}}{2b} \leq z \leq 0$, we use $Q(x) = 1 - Q(-x)$ to obtain

$$Q\left(z - \frac{a}{\sqrt{n}} \right) - Q\left(z + \frac{b}{\sqrt{n}} z^2 \right)$$
$$= Q\left(|z| - \frac{b}{\sqrt{n}} z^2 \right) - Q\left(|z| + \frac{a}{\sqrt{n}} \right) \tag{262}$$
$$\leq \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2 \left(1 - \frac{b}{\sqrt{n}} |z| \right)^2}{2}} \left(\frac{b}{\sqrt{n}} z^2 + \frac{a}{\sqrt{n}} \right) \tag{263}$$
$$\leq \frac{bz^2}{\sqrt{2\pi n}} e^{-\frac{z^2 \left(1 - \frac{b}{\sqrt{n}} |z| \right)^2}{2}} + \frac{a}{\sqrt{2\pi n}} \tag{264}$$
$$\leq \frac{bz^2}{\sqrt{2\pi n}} e^{-\frac{z^2}{8}} + \frac{a}{\sqrt{2\pi n}} \tag{265}$$
$$\leq \frac{8be^{-1} + a}{\sqrt{2\pi n}} \tag{266}$$

where (265) is due to $\left(1 - \frac{b}{\sqrt{n}} |z| \right)^2 \geq \frac{1}{4}$ in $|z| \leq \frac{\sqrt{n}}{2b}$, and (266) holds because the maximum of (265) is attained at $z^2 = 8$. ∎

We are now equipped to prove Lemma 1.

Proof of Lemma 1:

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

$$I(P_\mathsf{X}) = I(\mathsf{X}; \mathsf{Y}) = \mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X}; \mathsf{Y}) \right] \tag{267}$$
$$V(P_\mathsf{X}) = \mathbb{E}\left[\mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X}; \mathsf{Y}) \mid \mathsf{X} \right] \right] \tag{268}$$
$$T(P_\mathsf{X}) = \mathbb{E}\left[\left| \imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X}; \mathsf{Y}) - \mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X}; \mathsf{Y}) | \mathsf{X} \right] \right|^3 \mid \mathsf{X} \right] \tag{269}$$

If $P_X = \text{type}(x^n)$, then for each $a \in \mathcal{A}$, there are $nP_X(a)$ occurrences of $P_{Y|X=a}$ among the $\{P_{Y|X=x_i}, i = 1, 2, \ldots, n\}$. In the sequel we will invoke Theorem 25 with $W_i = \imath_{X;Y}(x_i; Y_i)$ where x^n is a given sequence, and (220)–(222) become

$$D_n = \frac{1}{n} \sum_{a=1}^{|\mathcal{A}|} nP_X(a) \mathbb{E}\left[\imath_{X;Y}(a; Y) \mid X = a\right] \tag{270}$$

$$= I(P_X) \tag{271}$$

$$V_n = \frac{1}{n}\left[\sum_{a=1}^{|\mathcal{A}|} nP_X(a)\text{Var}\left[\imath_{X;Y}(a; Y) \mid X = a\right]\right] \tag{272}$$

$$= V(P_X) \tag{273}$$

$$T_n = \frac{1}{n}\left[\sum_{a=1}^{|\mathcal{A}|} nP_X(a) \left|\imath_{X;Y}(a; Y) - \mathbb{E}\left[\imath_{X;Y}(a; Y)|X = a\right]\right|^3\right] \tag{}$$

$$= T(P_X) \tag{274}$$

Define the (Euclidean) δ-neighborhood of the set of capacity-achieving distributions $\mathcal{P}^\star$,

$$\mathcal{P}^\star_\delta = \left\{P_X \in \mathcal{P}: \min_{P_{X^\star} \in \mathcal{P}^\star} |P_X - P_{X^\star}| \leq \delta\right\} \tag{275}$$

We split the domain of the minimization in the left side of (230) into two sets, $\text{type}(x^n) \in \mathcal{P}^\star_{\delta,[n]}$ and $\text{type}(x^n) \in \mathcal{P}_{[n]}\backslash\mathcal{P}^\star_\delta$ (recall notation (224)), for an appropriately chosen $\delta > 0$.

We now show that (230) holds for all $\Delta \leq \frac{\Delta_I}{2}$ if the minimization is restricted to types in $\mathcal{P}_{[n]}\backslash\mathcal{P}^\star_\delta$, where $\delta > 0$ is arbitrary, and

$$\Delta_I = C - \max_{P_X \in \mathcal{P}_{[n]}\backslash\mathcal{P}^\star_\delta} I(P_X) > 0 \tag{276}$$

By Chebyshev's inequality, for all x^n whose type belongs to $\mathcal{P}_{[n]}\backslash\mathcal{P}^\star_\delta$,

$$\mathbb{P}\left[\sum_{i=1}^{n} \imath_{X;Y}(x_i; Y_i) > n(C - \Delta)\right] \tag{277}$$

$$= \mathbb{P}\left[\sum_{i=1}^{n} \imath_{X;Y}(x_i; Y_i) - nI(P_X) > n(C - I(P_X)) - n\Delta\right] \tag{278}$$

$$\leq \mathbb{P}\left[\sum_{i=1}^{n} \imath_{X;Y}(x_i; Y_i) - nI(P_X) > \frac{n\Delta_I}{2}\right] \tag{279}$$

$$\leq \mathbb{P}\left[\left(\sum_{i=1}^{n} \imath_{X;Y}(x_i; Y_i) - nI(P_X)\right)^2 > \frac{n^2\Delta_I^2}{4}\right] \tag{280}$$

$$\leq \frac{4nV(P_X)}{n^2\Delta_I^2} \tag{281}$$

$$\leq \frac{4\overline{V}}{n\Delta_I^2} \tag{282}$$

where in (279) we used

$$\Delta \leq \frac{1}{2}\Delta_I < \Delta_I \leq C - I(P_X) \tag{283}$$

and

$$\overline{V} = \max_{P_X \in \mathcal{P}} V(P_X) \tag{284}$$

Note that $\overline{V} < \infty$ by Property 1 below. Therefore,

$$\min_{\text{type}(x^n)\in\mathcal{P}_{[n]}\backslash\mathcal{P}^\star_\delta} \mathbb{P}\left[\sum_{i=1}^{n} \imath_{X;Y}(x_i; Y_i) \leq n(C - \Delta)\right] \tag{285}$$

$$> 1 - \frac{4\overline{V}}{n\Delta_I^2} \tag{286}$$

$$\geq \mathbb{P}\left[\sum_{i=1}^{n} \imath_{X;Y}(x_i^\star; Y_i) \leq n(C - \Delta)\right] - \frac{4\overline{V}}{n\Delta_I^2} \tag{287}$$

We conclude that (230) holds if the minimization is restricted to types in $\mathcal{P}_{[n]}\backslash\mathcal{P}^\star_\delta$.

Without loss of generality, we assume that all outputs in $\mathcal{B}$ are accessible (which implies that $P_{Y^\star}(y) > 0$ for all $y \in \mathcal{B}$) and choose $\delta > 0$ so that for all $P_X \in \mathcal{P}^\star_\delta$ and $y \in \mathcal{B}$,

$$P_Y(y) > 0 \tag{288}$$

where $P_X \to P_{Y|X} \to P_Y$. We recall the following properties of the functions $I(\cdot)$, $V(\cdot)$ and $T(\cdot)$ from [8, Appendices E and I].

Property 1. *The functions $I(P_X)$, $V(P_X)$ and $T(P_X)$ are continuous on the compact set $\mathcal{P}$, and therefore bounded and achieve their extrema.*

Property 2. *There exists $\ell_1 > 0$ such that for all $(P_{X^\star}, P_X) \in \mathcal{P}^\star \times \mathcal{P}^\star_\delta$,*

$$C - I(P_X) \geq \ell_1 |P_X - P_{X^\star}|^2 \tag{289}$$

Property 3. *In $\mathcal{P}^\star_\delta$, the functions $I(P_X)$, $V(P_X)$ and $T(P_X)$ are infinitely differentiable.*

Property 4. *In $\mathcal{P}^\star$, $V(P_X) = \text{Var}\left[\imath_{X;Y}(X; Y)\right]$.*

Due to Property 3, there exist nonnegative constants L_1 and L_2 such that for all $(P_X, P_{X^\star}) \in \mathcal{P}^\star_\delta \times \mathcal{P}^\star$,

$$C - I(P_X) \leq L_1 |P_X - P_{X^\star}| \tag{290}$$

$$|V(P_X) - V(P_{X^\star})| \leq L_2 |P_X - P_{X^\star}| \tag{291}$$

To treat the case $x^n \in \mathcal{P}^\star_{\delta,[n]}$, we will need to choose $\delta > 0$ carefully and to consider the cases $V_{\min} > 0$ and $V_{\max} = 0$ separately.

A. $V_{\min} > 0$.

We decrease δ until, in addition to (288),

$$V_{\min} \leq 2 \min_{P_X \in \mathcal{P}^\star_\delta} V(P_X) \tag{292}$$

is satisfied.

We now show that (230) holds if the minimization is restricted to types in $\mathcal{P}^\star_{\delta,[n]}$, for all $-\underline{\Delta} \leq \Delta \leq \frac{\Delta_I}{2}$, for an appropriately chosen $\underline{\Delta} > 0$. Using (292) and boundedness of $T(P_X)$, write

$$B = \max_{P_X \in \mathcal{P}^\star_\delta} \frac{c_0 T(P_X)}{V^{\frac{3}{2}}(P_X)} \leq \frac{2^{\frac{3}{2}} c_0 \overline{T}}{V_{\min}^{\frac{3}{2}}} < \infty \tag{293}$$

where

$$\overline{T} = \max_{P_X \in \mathcal{P}^\star_\delta} T(P_X) < \infty \tag{294}$$

Therefore, for any x^n with $\text{type}(x^n) \in \mathcal{P}^\star_{\delta,[n]}$, the Berry-Esseen bound yields:

$$\left| \mathbb{P}\left[\sum_{i=1}^n \imath_{X;Y}(x_i; Y_i) \leq n(C-\Delta) \right] - Q\left(\nu(P_X)\right) \right| \leq \frac{B}{\sqrt{n}} \tag{295}$$

where

$$\nu(P_X) = \frac{nI(P_X) - nC + n\Delta}{\sqrt{nV(P_X)}} \tag{296}$$

We now apply Lemma 2 with $\mathcal{D} = \mathcal{P}^\star_{\delta,[n]}$ and

$$f(P_X) = Q\left(\nu(P_X)\right) \tag{297}$$

$$g(P_X) = \mathbb{P}\left[\sum_{i=1}^n \imath_{X;Y}(x_i; Y_i) \leq n(C-\Delta) \right] \tag{298}$$

Condition (232) of Lemma 2 holds with $\eta = \frac{B}{\sqrt{n}}$ due to (295). As will be shown in the sequel, the following version of condition (234) holds:

$$Q(\nu(\Pi(P_{X^\star}))) \leq \min_{P_X \in \mathcal{P}^\star_{\delta,[n]}} Q(\nu(P_X)) + \frac{q}{\sqrt{n}} \tag{299}$$

where $\Pi(P_{X^\star})$, the minimum Euclidean distance approximation of $P_{X^\star}$ in the set of n-types, is formally defined in (225), and $q > 0$ will be chosen later. Applying Lemma 2, we deduce from (233) that

$$\min_{\text{type}(x^n) \in \mathcal{P}^\star_{\delta,[n]}} \mathbb{P}\left[\sum_{i=1}^n \imath_{X;Y}(x_i; Y_i) \leq n(C-\Delta) \right]$$
$$\geq \mathbb{P}\left[\sum_{i=1}^n \imath_{X;Y}(x_i^\star; Y_i) \leq n(C-\Delta) \right] - \frac{q + 2B}{\sqrt{n}} \tag{300}$$

We conclude that (230) holds if minimization is restricted to types in $\mathcal{P}^\star_{\delta,[n]}$.

We proceed to show (299). As will be proven later, for appropriately chosen $\underline{L} > 0$ and $\bar{L} > 0$ we can write

$$\frac{\sqrt{n}\Delta}{\sqrt{V(P_{X^\star})}} - \frac{\underline{L}}{\sqrt{n}} \sqrt{|\mathcal{A}|(|\mathcal{A}|-1)} \leq \nu(\Pi(P_{X^\star})) \tag{301}$$

$$\leq \max_{P_X \in \mathcal{P}^\star_{\delta,[n]}} \nu(P_X) \tag{302}$$

$$\leq \max_{P_X \in \mathcal{P}^\star_\delta} \nu(P_X) \tag{303}$$

$$\leq \frac{\sqrt{n}\Delta}{\sqrt{V(P_{X^\star})}} + \sqrt{n}\bar{L}\Delta^2 \tag{304}$$

where $P_{X^\star} \in \mathcal{P}^\star_{\min}$ if $\Delta \geq 0$, and $P_{X^\star} \in \mathcal{P}^\star_{\max}$ if $\Delta < 0$.

Denote

$$a = \underline{L}\sqrt{|\mathcal{A}|(|\mathcal{A}|-1)} \tag{305}$$

$$b = V(P_{X^\star})\bar{L} \tag{306}$$

$$z = \frac{\sqrt{n}\Delta}{\sqrt{V(P_{X^\star})}} \tag{307}$$

If

$$\Delta \geq -\frac{1}{2\bar{L}\sqrt{V_{\max}}} = -\underline{\Delta} \tag{308}$$

then $z \geq -\frac{\sqrt{n}}{2b}$, and Lemma 4 applies to z. So, using (301), (304), the fact that $Q(\cdot)$ is monotonically decreasing and Lemma 4, we conclude that there exists $q > 0$ such that

$$Q\left(\nu(\Pi(P_{X^\star}))\right) - \min_{P_X \in \mathcal{P}^\star_{\delta,[n]}} Q(\nu(P_X))$$

$$= Q\left(\nu(\Pi(P_{X^\star}))\right) - Q\left(\max_{P_X \in \mathcal{P}^\star_{\delta,[n]}} \nu(P_X) \right) \tag{309}$$

$$\leq Q\left(z - \frac{a}{\sqrt{n}} \right) - Q\left(z + \frac{b}{\sqrt{n}} z^2 \right) \tag{310}$$

$$\leq \frac{q}{\sqrt{n}} \tag{311}$$

which is equivalent to (299).

It remains to prove (301) and (304). Observing that for $a, b > 0$

$$\left| \frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}} \right| = \frac{|a-b|}{\sqrt{a}\sqrt{b}\left(\sqrt{a}+\sqrt{b}\right)} \tag{312}$$

$$\leq \frac{|a-b|}{2\min\{a,b\}^{\frac{3}{2}}} \tag{313}$$

and using (291) and (292), we have, for all $(P_X, P_{X^\star}) \in \mathcal{P}^\star_\delta \times \mathcal{P}^\star$,

$$\left| \frac{1}{\sqrt{V(P_X)}} - \frac{1}{\sqrt{V(P_{X^\star})}} \right| \leq L|P_X - P_{X^\star}| \tag{314}$$

where

$$L = L_2 \sqrt{\frac{2}{V_{\min}^3}} \tag{315}$$

Thus, recalling (290) and denoting $\zeta = |P_X - P_{X^\star}|$, we have

$$\frac{C - I(P_X) - \Delta}{\sqrt{V(P_X)}}$$

$$\leq \frac{L_1\zeta - \Delta}{\sqrt{V(P_X)}} \tag{316}$$

$$\leq \frac{L_1\zeta - \Delta}{\sqrt{V(P_{X^\star})}} + L\zeta\left(L_1\zeta + |\Delta|\right) \tag{317}$$

$$\leq -\frac{\Delta}{\sqrt{V(P_{X^\star})}} + \underline{L}\zeta \tag{318}$$

where

$$\underline{L} = \frac{L_1}{\sqrt{V_{\min}}} + L\max\left\{\underline{\Delta}, \frac{\Delta_I}{2}\right\} + LL_1\delta \tag{319}$$

So, (301) follows by observing that for any $P_X \in \mathcal{P}$,

$$|P_X - \Pi(P_X)| \leq \frac{1}{n}\sqrt{|\mathcal{A}|(|\mathcal{A}|-1)} \tag{320}$$

and letting $P_X = \Pi(P_{X^\star})$ in (316)–(318).

To show (304), we apply Lemma 3 with

$$\mathcal{D} = \mathcal{P}_\delta^\star \tag{321}$$
$$\mathcal{D}^\star = \mathcal{P}^\star \tag{322}$$
$$\varphi = \sqrt{n} \tag{323}$$
$$\psi = \sqrt{n}|\Delta| \tag{324}$$
$$f(P_{\mathsf{X}}) = \frac{I(P_{\mathsf{X}}) - C}{\sqrt{V(P_{\mathsf{X}})}} \tag{325}$$
$$g(P_{\mathsf{X}}) = \frac{1}{\sqrt{V(P_{\mathsf{X}})}} \tag{326}$$

We proceed to verify that conditions of Lemma 3 are met. Function g satisfies condition (242) with L defined in (315). Let us now show that function f satisfies condition (241) with $\ell = \frac{\ell_1}{\sqrt{\overline{V}}}$, where $\overline{V}$ and ℓ_1 are defined in (284) and (289), respectively. For any $(P_{\mathsf{X}}, P_{\mathsf{X}^\star}) \in \mathcal{P}_\delta^\star \times \mathcal{P}^\star$, write

$$f(P_{\mathsf{X}^\star}) - f(P_{\mathsf{X}}) = \frac{C - I(P_{\mathsf{X}})}{\sqrt{V(P_{\mathsf{X}})}} \tag{327}$$
$$\geq \frac{C - I(P_{\mathsf{X}})}{\sqrt{\overline{V}}} \tag{328}$$
$$\geq \frac{\ell_1}{\sqrt{\overline{V}}} |P_{\mathsf{X}} - P_{\mathsf{X}^\star}|^2 \tag{329}$$

where (328) follows from (284), and (329) applies (289). So, Lemma 3 applies to $\nu(P_{\mathsf{X}}) = \varphi f(P_{\mathsf{X}}) + \mathrm{sign}(\Delta)\psi g(P_{\mathsf{X}})$, resulting in (304) with

$$\bar{L} = \frac{L_2^2 \sqrt{\overline{V}}}{2\ell_1 V_{\min}^3} \tag{330}$$

thereby completing the proof of (300).

Combining (287) and (300), we conclude that (230) holds for all Δ in the interval

$$-\frac{\ell_1 V_{\min}^3}{L_2^2 \sqrt{\overline{V}} V_{\max}} \leq \Delta \leq \frac{\Delta_I}{2} \tag{331}$$

B. $V_{\max} = 0$.

We choose δ so that (288) is satisfied. The case $\mathrm{type}(x^n) \notin \mathcal{P}_{\delta,[n]}^\star$ was covered in (287), so we only need to consider the minimization of the left side of (231) over $\mathcal{P}_{\delta,[n]}^\star$. Fix $\alpha < \frac{3}{2}$. If

$$\Delta \geq \left(\frac{L_2^2}{2^8 \ell_1}\right)^{\frac{1}{3}} \frac{3}{n^{\frac{1}{2}+\alpha}} \tag{332}$$

we have

$$\mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) > n(C+\Delta)\right]$$
$$= \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) - nI(P_{\mathsf{X}}) > n(C - I(P_{\mathsf{X}})) + n\Delta\right] \tag{333}$$
$$\leq \frac{V(P_{\mathsf{X}})}{n\left(C - I(P_{\mathsf{X}}) + \Delta\right)^2} \tag{334}$$
$$\leq \frac{L_2|P_{\mathsf{X}} - P_{\mathsf{X}^\star}|}{n\left(\ell_1|P_{\mathsf{X}} - P_{\mathsf{X}^\star}|^2 + \Delta\right)^2} \tag{335}$$
$$\leq \frac{3^{\frac{3}{2}} L_2}{16\ell_1^{\frac{1}{2}}} \frac{1}{n\Delta^{\frac{3}{2}}} \tag{336}$$
$$\leq \frac{1}{n^{\frac{1}{4} - \frac{3}{2}\alpha}} \tag{337}$$

where

- (334) is by Chebyshev's inequality;
- (335) uses (289), (291) and $V_{\max} = 0$;
- (336) holds because the maximum of its left side is attained at $|P_{\mathsf{X}} - P_{\mathsf{X}^\star}|^2 = \frac{\Delta}{3\ell_1}$.

$\blacksquare$

APPENDIX C
PROOF OF THE CONVERSE PART OF THEOREM 10

Note that for the converse, restriction (iv) can be replaced by the following weaker one:

(iv′) The random variable $\jmath_{\mathsf{S}}(\mathsf{S}, d)$ has finite absolute third moment.

To verify that (iv) implies (iv′), observe that by the concavity of the logarithm,

$$0 \leq \jmath_{\mathsf{S}}(\mathsf{s}, d) + \lambda^\star d \leq \lambda^\star \mathbb{E}\left[\mathsf{d}(\mathsf{s}, \mathsf{Z}^\star)\right] \tag{338}$$

so

$$\mathbb{E}\left[|\jmath_{\mathsf{S}}(\mathsf{S}, d) + \lambda^\star d|^3\right] \leq \lambda^{\star 3} \mathbb{E}\left[\mathsf{d}^3(\mathsf{S}, \mathsf{Z}^\star)\right] \tag{339}$$

We now proceed to prove the converse by showing first that we can eliminate all rates exceeding

$$\frac{k}{n} \geq \frac{C}{R(d) - 3\tau} \tag{340}$$

for any $0 < \tau < \frac{R(d)}{3}$. More precisely, we show that the excess-distortion probability of any code having such rate converges to 1 as $n \to \infty$, and therefore for any $\epsilon < 1$, there is an n_0 such that for all $n \geq n_0$, no (k, n, d, ϵ) code can exist for k, n satisfying (340).

We weaken (24) by fixing $\gamma = k\tau$ and choosing a particular output distribution, namely, $P_{\bar{Y}^n} = P_{Y^{n\star}} = P_{\mathsf{Y}^\star} \times \ldots \times P_{\mathsf{Y}^\star}$. Due to restriction (ii) in Section V, $P_{Z^k}^\star = P_{\mathsf{Z}}^\star \times \ldots \times P_{\mathsf{Z}}^\star$, and the $\mathsf{d}-$tilted information single-letterizes, that is, for a.e. s^k,

$$\jmath_{S^k}(s^k, d) = \sum_{i=1}^k \jmath_{\mathsf{S}}(s_i, d) \tag{341}$$

Theorem 1 implies that error probability ϵ' of every (k, n, d, ϵ') code must be lower bounded by

$$\mathbb{E}\left[\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - \sum_{j=1}^{n} \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \geq k\tau \mid S^k\right]\right]$$
$$- \exp(-k\tau)$$

$$\geq \min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{j=1}^{n} \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \leq nC + k\tau\right]$$

$$\cdot \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) \geq nC + 2k\tau\right] - \exp(-k\tau) \qquad (342)$$

$$\geq \min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{j=1}^{n} \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \leq nC + n\tau'\right]$$

$$\cdot \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) \geq kR(d) - k\tau\right] - \exp(-k\tau) \qquad (343)$$

where in (343), we used (340) and $\tau' = \frac{C\tau}{R(d)-3\tau} > 0$. Recalling (16) and

$$\mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{x};\mathsf{Y})|\mathsf{X} = \mathsf{x}\right] \leq C \qquad (344)$$

with equality for $P_{\mathsf{X}^\star}$-a.e. x, we conclude using the law of large numbers that (343) tends to 1 as $k, n \to \infty$.

We proceed to show that for all large enough k, n, if there is a sequence of (k, n, d, ϵ') codes such that

$$-3k\tau \leq nC - kR(d) \qquad (345)$$
$$\leq \sqrt{nV + k\mathcal{V}(d)}\, Q^{-1}(\epsilon) + \theta(n) \qquad (346)$$

then $\epsilon' \geq \epsilon$.

Note that in general the bound in Theorem 1 with the choice of $P_{\bar{Y}^n}$ as above does not lead to the correct channel dispersion term. We first consider the general case, in which we apply Theorem 3, and then we show the symmetric case, in which we apply Theorem 2.

Recall 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 $nP_{\mathsf{X}}(a)$. In Theorem 3, we weaken the supremum over W by letting W map X^n to its type, $W = \text{type}(X^n)$. Note that the total number of types satisfies (e.g. [25]) $T \leq (n+1)^{|\mathcal{A}|-1}$. We weaken the supremum over $\bar{Y}^n$ in (39) by fixing $P_{\bar{Y}^n|W=P_{\mathsf{X}}} = P_{\mathsf{Y}} \times \ldots \times P_{\mathsf{Y}}$, where $P_{\mathsf{X}} \to P_{\mathsf{Y}|\mathsf{X}} \to P_{\mathsf{Y}}$, i.e. P_{Y} is the output distribution induced by the type P_{X}. In this way, Theorem 3 implies that the error probability of any (k, n, d, ϵ') code must be lower bounded by

$$\epsilon' \geq \mathbb{E}\left[\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - \sum_{i=1}^{n} \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \geq \gamma \mid S^k\right]\right]$$
$$- (n+1)^{|\mathcal{A}|-1} \exp(-\gamma) \qquad (347)$$

Choose

$$\gamma = \left(|\mathcal{A}| - \frac{1}{2}\right) \log(n+1) \qquad (348)$$

At this point we consider two cases separately, $V > 0$ and $V = 0$.

A. $V > 0$.

In order to apply Lemma 1 in Appendix B, we isolate the typical set of source sequences:

$$\mathcal{T}_{k,n} = \left\{s^k \in \mathcal{S}^k : \left|\sum_{i=1}^{k} \jmath_{\mathsf{S}}(s_i, d) - nC\right| \leq n\bar{\Delta} - \gamma\right\} \qquad (349)$$

Observe that

$$\mathbb{P}\left[S^k \notin \mathcal{T}_{k,n}\right]$$

$$= \mathbb{P}\left[\left|\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - nC\right| > n\bar{\Delta} - \gamma\right] \qquad (350)$$

$$\leq \mathbb{P}\left[\left|\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - kR(d)\right| + |nC - kR(d)| + \gamma > n\bar{\Delta}\right] \qquad (351)$$

$$\leq \mathbb{P}\left[\left|\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - kR(d)\right| > k\frac{\bar{\Delta}R(d)}{2C}\right] \qquad (352)$$

$$\leq \frac{4C^2}{R^2(d)\bar{\Delta}^2} \frac{\mathcal{V}(d)}{k} \qquad (353)$$

where

- (352) follows by lower bounding

$$n\bar{\Delta} - \gamma - |nC - kR(d)|$$
$$\geq n\bar{\Delta} - \gamma - 3k\tau \qquad (354)$$
$$\geq n\frac{3\bar{\Delta}}{4} - 3k\tau \qquad (355)$$
$$\geq k\frac{3\bar{\Delta}}{4C}(R(d) - 3\tau) - 3k\tau \qquad (356)$$
$$\geq k\frac{\bar{\Delta}R(d)}{2C} \qquad (357)$$

  where
  - (354) holds for large enough n due to (345) and (346);
  - (355) holds for large enough n by the choice of γ in (348);
  - (356) lower bounds n using (345);
  - (357) holds for a small enough $\tau > 0$.

- (353) is by Chebyshev's inequality.

Now, we let

$$\epsilon_{k,n} = \epsilon + \frac{B}{\sqrt{n+k}} + \frac{1}{\sqrt{n+1}} + \frac{4C^2}{R^2(d)\bar{\Delta}^2} \frac{\mathcal{V}(d)}{k} \qquad (358)$$

where $B > 0$ will be chosen in the sequel, and k, n are chosen so that both (345) and the following version of (346) hold:

$$nC - kR(d) \leq \sqrt{nV + k\mathcal{V}(d) - \frac{L_2|\mathcal{A}|}{n+k}}\, Q^{-1}(\epsilon_{k,n}) - \gamma \qquad (359)$$

where $L_2 < \infty$ is defined in (291). Denote for brevity

$$r(x^n, y^n, s^k) = \sum_{i=1}^{n} \imath_{\mathsf{X};\mathsf{Y}}(x_i; y_i) - \sum_{i=1}^{k} \jmath_{\mathsf{S}}(s_i, d) \qquad (360)$$

Weakening (347) using (348) and Lemma 1, we can lower bound ϵ' by

$$\mathbb{E}\left[\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[r(x^n, Y^n, S^k) \leq -\gamma \mid S^k\right] \cdot 1\left\{S^k \in \mathcal{T}_{k,n}\right\}\right]$$

$$- \frac{1}{\sqrt{n+1}}$$

$$\geq \mathbb{E}\left[\mathbb{P}\left[r(x^{n\star}, Y^n, S^k) \leq -\gamma \mid S^k\right] \cdot 1\left\{S^k \in \mathcal{T}_{k,n}\right\}\right]$$

$$- \frac{K}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \tag{361}$$

$$= \mathbb{P}\left[r(x^{n\star}, Y^n, S^k) \leq -\gamma, \ S^k \in \mathcal{T}_{k,n}\right] - \frac{K}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \tag{362}$$

$$\geq \mathbb{P}\left[r(x^{n\star}, Y^n, S^k) \leq -\gamma\right] - \mathbb{P}\left[S^k \notin \mathcal{T}_{k,n}\right] - \frac{K}{\sqrt{n}}$$

$$- \frac{1}{\sqrt{n+1}} \tag{363}$$

$$\geq \mathbb{P}\left[r(x^{n\star}, Y^n, S^k) \leq -\gamma\right] - \frac{4C^2}{R^2(d)\bar{\Delta}^2}\frac{\mathcal{V}(d)}{k} - \frac{K}{\sqrt{n}}$$

$$- \frac{1}{\sqrt{n+1}} \tag{364}$$

$$\geq \epsilon \tag{365}$$

where (361) is by Lemma 1, and (363) is by the union bound. To justify (365), observe that the quantities in Theorem 25 corresponding to the sum of independent random variables in (364) are

$$D_{n+k} = \frac{n}{n+k}I(\Pi(P_{\mathsf{X}^\star})) - \frac{k}{n+k}R(d) \tag{366}$$

$$\leq \frac{n}{n+k}C - \frac{k}{n+k}R(d) \tag{367}$$

$$V_{n+k} = \frac{n}{n+k}V(\Pi(P_{\mathsf{X}^\star})) + \frac{k}{n+k}\mathcal{V}(d) \tag{368}$$

$$\geq \frac{n}{n+k}V + \frac{k}{n+k}\mathcal{V}(d) - \frac{L_2|\mathcal{A}|}{n+k} \tag{369}$$

$$T_{n+k} = \frac{n}{n+k}T(\Pi(P_{\mathsf{X}^\star})) + \frac{k}{n+k}\mathbb{E}\left[|\jmath_\mathsf{S}(\mathsf{S}, d) - R(d)|^3\right] \tag{370}$$

where the functions $\Pi(\cdot)$, $I(\cdot)$, $V(\cdot)$, $T(\cdot)$ are defined in (225), (267)–(269) in Appendix B. To show (369), recall that $V(P_{\mathsf{X}^\star}) = V$ by Property 4 in Appendix B, and use (291) and (320). Further, T_{n+k} is bounded uniformly in P_{X}, so (223) is upper bounded by some constant $B > 0$. Finally, applying (367) and (369) to (359), we conclude that

$$-\gamma \geq (n+k)D_{n+k} - \sqrt{(n+k)V_{n+k}}Q^{-1}(\epsilon_{k,n}) \tag{371}$$

which enables us to lower bound the probability in (364) invoking the Berry-Esseen bound (Theorem 25). In view of (358), the resulting bound is equal to ϵ, and the proof of (365) is complete.

B. $V = 0$.

Fix $0 < \alpha < \frac{1}{6}$.

If $\mathcal{V}(d) > 0$, we choose γ as in (348), and

$$\epsilon_{k,n} = \epsilon + \frac{B}{\sqrt{k}} + (n+1)^{|\mathcal{A}|-1}\exp(-\gamma) + \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} \tag{372}$$

where $B > 0$ is the same as in (358), and k, n are chosen so that the following version of (346) hold:

$$nC - kR(d) \leq \sqrt{k\mathcal{V}(d)}Q^{-1}(\epsilon_{k,n}) - \gamma - \bar{\Delta}n^{\frac{1}{2}-\alpha} \tag{373}$$

where $\bar{\Delta} > 0$ was defined in Lemma 1. Weakening (347) using (231), we have

$$\epsilon' \geq \min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\imath_{\mathsf{X};\mathsf{Y}}(x_i, Y_i) \geq nC + \bar{\Delta}n^{\frac{1}{2}-\alpha}\right]$$

$$\cdot \mathbb{P}\left[\sum_{i=1}^{k}\jmath_\mathsf{S}(S_i, d) \geq nC + \bar{\Delta}n^{\frac{1}{2}-\alpha} + \gamma\right]$$

$$- (n+1)^{|\mathcal{A}|-1}\exp(-\gamma) \tag{374}$$

$$\geq \left(1 - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}}\right)$$

$$\cdot \mathbb{P}\left[\sum_{i=1}^{k}\jmath_\mathsf{S}(S_i, d) \geq kR(d) + \sqrt{k\mathcal{V}(d)}Q^{-1}(\epsilon_{k,n})\right]$$

$$- (n+1)^{|\mathcal{A}|-1}\exp(-\gamma) \tag{375}$$

$$\geq \left(1 - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}}\right)\left(\epsilon_{k,n} - \frac{B}{\sqrt{k}}\right)$$

$$- (n+1)^{|\mathcal{A}|-1}\exp(-\gamma) \tag{376}$$

$$\geq \epsilon_{k,n} - \frac{B}{\sqrt{k}} - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} - (n+1)^{|\mathcal{A}|-1}\exp(-\gamma) \tag{377}$$

$$= \epsilon \tag{378}$$

where (375) uses (231) and (373), and (376) is by the Berry-Esseen bound.

If $\mathcal{V}(d) = 0$, which implies $\jmath_S(S_i, d) = R(d)$ a.s., we let

$$\gamma = (|\mathcal{A}| - 1)\log(n+1) - \log\left(1 - \epsilon - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}}\right) \tag{379}$$

and choose k, n that satisfy

$$kR(d) - nC \geq \gamma + \bar{\Delta}n^{\frac{1}{2}-\alpha} \tag{380}$$

Then, plugging $\jmath_S(S_i, d) = R(d)$ a.s. in (347), we have

$$\epsilon' \geq \min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^{n}\imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \leq kR(d) - \gamma\right]$$

$$- (n+1)^{|\mathcal{A}|-1}\exp(-\gamma) \tag{381}$$

$$\geq \min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^{n}\imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \leq nC + \bar{\Delta}n^{\frac{1}{2}-\alpha}\right]$$

$$- (n+1)^{|\mathcal{A}|-1}\exp(-\gamma) \tag{382}$$

$$\geq 1 - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} - (n+1)^{|\mathcal{A}|-1}\exp(-\gamma) \tag{383}$$

$$= \epsilon \tag{384}$$

where (382) is by the choice of k, n in (380), (383) invokes (231), and (384) follows from the choice of γ in (379).

C. Symmetric channel.

We show that if the channel is such that the distribution of $\imath_{X;Y^\star}(x;Y)$ (according to $P_{Y|X=x}$) does not depend on the choice $x \in \mathcal{A}$, Theorem 2 leads to a tighter third-order term than (119).

If either $V > 0$ or $\mathcal{V}(d) > 0$, let

$$\gamma = \frac{1}{2}\log n \qquad (385)$$

$$\epsilon_{k,n} = \epsilon + \frac{B}{\sqrt{n+k}} + \frac{1}{\sqrt{n}} \qquad (386)$$

where $B > 0$ can be chosen as in (358), and let k, n be such that the following version of (346) (with the remainder $\theta(n)$ satisfying (119) with $\underline{c} = \frac{1}{2}$) holds:

$$nC - kR(d) \le \sqrt{nV + k\mathcal{V}(d)}Q^{-1}(\epsilon_{k,n}) - \gamma \qquad (387)$$

Theorem 2 and Theorem 25 imply that the error probability of every (k,n,d,ϵ') code must satisfy, for an arbitrary sequence $x^n \in \mathcal{A}^n$,

$$\epsilon' \ge \mathbb{P}\left[\sum_{i=1}^{k}\jmath_{S}(S_i,d) - \sum_{j=1}^{n}\imath_{X;Y^\star}(x_i;Y_i) \ge \gamma\right] - \exp(-\gamma) \qquad (388)$$

$$\ge \epsilon \qquad (389)$$

If both $V = 0$ and $\mathcal{V}(d) = 0$, choose k, n to satisfy

$$kR(d) - nC \ge \gamma \qquad (390)$$

$$= \log\frac{1}{1-\epsilon} \qquad (391)$$

Substituting (391) and $\jmath_{S}(S_i,d) = R(d)$, $\imath_{X;Y^\star}(x_i;Y_i) = C$ a.s. in (388), we conclude that the right side of (388) equals ϵ, so $\epsilon' \ge \epsilon$ whenever a (k,n,d,ϵ') code exists.

D. Gaussian channel

In view of Remark 12, it suffices to consider the equal power constraint (133). The spherically-symmetric $P_{\bar{Y}^n} = P_{Y^{n\star}} = P_{Y^\star} \times \ldots \times P_{Y^\star}$, where $Y^\star \sim \mathcal{N}(0, \sigma_N^2(1+P))$, satisfies the symmetry assumption of Theorem 2. In fact, for all $x^n \in \mathcal{F}(\alpha)$, $\imath_{X^n;Y^{n\star}}(x^n;Y^n)$ has the same distribution under $P_{Y^n|X^n=x^n}$ as (cf. (168))

$$G_n = \frac{n}{2}\log(1+P) - \frac{\log e}{2}\left(\frac{P}{1+P}\sum_{i=1}^{n}\left(W_i - \frac{1}{\sqrt{P}}\right)^2 - n\right) \qquad (392)$$

where $W_i \sim \mathcal{N}\left(\frac{1}{\sqrt{P}}, 1\right)$, independent of each other. Since G_n is a sum of i.i.d. random variables, the mean of $\frac{G_n}{n}$ is equal to $C = \frac{1}{2}\log(1+P)$ and its variance is equal to (118), the result follows analogously to (385)–(389).

APPENDIX D
PROOF OF THE ACHIEVABILITY PART OF THEOREM 10

A. Almost lossless coding ($d = 0$) over a DMC.

The proof consists of an asymptotic analysis of the bound in Theorem 9 by means of Theorem 25. Weakening (113) by

fixing $P_{X^n} = P_{X^n}^\star = P_{X^\star} \times \ldots \times P_{X^\star}$, we conclude that there exists a $(k,n,0,\epsilon')$ code with

$$\epsilon' \le \mathbb{E}\left[\exp\left(-\left|\sum_{i=1}^{n}\imath_{X;Y}^\star(X_i^\star;Y_i^\star) - \sum_{i=1}^{k}\imath_{S}(S_i)\right|^+\right)\right] \qquad (393)$$

where $(S^k, X^{n\star}, Y^{n\star})$ are distributed according to $P_{S^k}P_{X^{n\star}}P_{Y^n|X^n}$. The case of equiprobable S has been tackled in [8]. Here we assume that $\imath_S(S)$ is not a constant, that is, $\mathrm{Var}[\imath_S(S)] > 0$.

Let k and n be such that

$$nC - kH(S) \ge \sqrt{nV + k\mathcal{V}}Q^{-1}\left(\epsilon - \frac{B+1}{\sqrt{n+k}}\right) + \frac{1}{2}\log(n+k) \qquad (394)$$

where $\mathcal{V} = \mathrm{Var}[\imath_S(S)]$, and B is the Berry-Esseen ratio (223) for the sum of $n+k$ independent random variables appearing in the right side of (393). Note that B is finite due to:

- $\mathrm{Var}[\imath_S(S)] > 0$;
- the third absolute moment of $\imath_S(S)$ is finite;
- the third absolute moment of $\imath_{X;Y}^\star(X^\star;Y^\star)$ is finite, as observed in Appendix B.

Therefore, (394) can be written as (114) with the remainder therein satisfying (125). So, it suffices to prove that if k,n satisfy (394), then the right side of (393) is upper bounded by ϵ. Let

$$\mathcal{T}_{k,n} = \left\{(s^k, x^n, y^n) \in \mathcal{S}^k \times \mathcal{A}^n \times \mathcal{B}^n : \right.$$

$$\sum_{i=1}^{n}\imath_{X;Y}^\star(x_i;y_i) - \sum_{i=1}^{k}\imath_S(s_i)$$

$$\left. \ge nC - kH(S) - \sqrt{nV + k\mathcal{V}}Q^{-1}\left(\epsilon - \frac{B+1}{\sqrt{n+k}}\right)\right\} \qquad (395)$$

By the Berry-Esseen bound (Theorem 25),

$$\mathbb{P}\left[(S^k, X^{n\star}, Y^{n\star}) \notin \mathcal{T}_{k,n}\right] \le \epsilon - \frac{1}{\sqrt{n+k}} \qquad (396)$$

We now further upper bound (393) as

$$\epsilon' \le \mathbb{E}\left[\exp\left(-\left|\sum_{i=1}^{n}\imath_{X;Y}^\star(X_i^\star;Y_i^\star) - \sum_{i=1}^{k}\imath_S(S_i)\right|^+\right)\right.$$

$$\left. \cdot 1_{\mathcal{T}_{k,n}}(S^k, X^{n\star}, Y^{n\star})\right] + \mathbb{P}\left[(S^k, X^{n\star}, Y^{n\star}) \notin \mathcal{T}_{k,n}\right] \qquad (397)$$

$$\le \frac{1}{\sqrt{n+k}}\mathbb{P}\left[(S^k, X^{n\star}, Y^{n\star}) \in T_{k,n}\right] + \epsilon - \frac{1}{\sqrt{n+k}} \qquad (398)$$

$$\le \epsilon \qquad (399)$$

where we invoked (394) and (395) to upper bound the exponent in the right side of (397).

B. Lossy coding over a DMC.

The proof consists of the asymptotic analysis of the bound in Theorem 8 using Theorem 25 and Lemma 5 below, which deals with asymptotic behavior of distortion d-balls. Note that

Lemma 5 is the only step that requires finiteness of the ninth absolute moment of $d(S, Z^\star)$ as required by restriction (iv) in Section V.

Lemma 5 ([9, Lemma 2]). *Under restrictions* (ii)–(iv)*, there exist constants* $k_0, c, K > 0$ *such that for all* $k \geq k_0$,

$$\mathbb{P}\left[\log\frac{1}{P_{Z^{k\star}}(B_d(S^k))} \leq \sum_{i=1}^{k} \jmath_S(S_i, d) + \left(\bar{c} - \frac{1}{2}\right)\log k + c\right]$$
$$\geq 1 - \frac{K}{\sqrt{k}} \tag{400}$$

where $\bar{c}$ *is given by* (122).

We weaken (108) by fixing

$$P_{X^n} = P_{X^{n\star}} = P_{X^\star} \times \ldots \times P_{X^\star} \tag{401}$$

$$P_{Z^k} = P_{Z^{k\star}} = P_{Z^\star} \times \ldots \times P_{Z^\star} \tag{402}$$

$$\gamma = \frac{1}{2}\log_e k + 1 \tag{403}$$

where $\Delta > 0$, so there exists a (k, n, d, ϵ') code with error probability ϵ' upper bounded by

$$\mathbb{E}\left[\exp\left(-\left|\sum_{i=1}^{n}\imath^\star_{X;Y}(X_i^\star; Y_i^\star) - \log\frac{\gamma}{P_{Z^{k\star}}(B_d(S^k))}\right|^+\right)\right]$$
$$+ e^{1-\gamma} \tag{404}$$

where $(S^k, X^{n\star}, Y^{n\star}, Z^{k\star})$ are distributed according to $P_{S^k}P_{X^{n\star}}P_{Y^n|X^n}P_{Z^{k\star}}$. We need to show that for k, n satisfying (114), (404) is upper bounded by ϵ.

We apply Lemma 5 to upper bound (404) as follows:

$$\epsilon' \leq \mathbb{E}\left[\exp\left(-|U_{k,n}|^+\right)\right] + \frac{K+1}{\sqrt{k}} \tag{405}$$

with

$$U_{k,n} = \sum_{i=1}^{n}\imath^\star_{X;Y}(X_i^\star; Y_i^\star) - \sum_{i=1}^{k}\jmath_S(S_i, d) - \left(\bar{c} - \frac{1}{2}\right)\log k$$
$$- \log\gamma - c \tag{406}$$

We first consider the (nontrivial) case $\mathcal{V}(d) + V > 0$. Let k and n be such that

$$nC - kR(d) \geq \sqrt{nV + k\mathcal{V}(d)}Q^{-1}(\epsilon_{k,n})$$
$$+ \bar{c}\log k + \log\gamma + c \tag{407}$$

$$\epsilon_{k,n} = \epsilon - \frac{B}{\sqrt{n+k}} - \frac{K+2}{\sqrt{k}} \tag{408}$$

where constants c and $\bar{c}$ are defined in Lemma 5, and B is the Berry-Esseen ratio (223) for the sum of $n+k$ independent random variables appearing in (405). Note that B is finite because:

- either $\mathcal{V}(d) > 0$ or $V > 0$ by the assumption;
- the third absolute moment of $\jmath_S(S, d)$ is finite by restriction (iv) as spelled out in (339);
- the third absolute moment of $\imath^\star_{X;Y}(X^\star; Y^\star)$ is finite, as observed in Appendix B.

Applying a Taylor series expansion to (407) with the choice of γ in (403), we conclude that (407) can be written as (114) with the remainder term satisfying (120).

It remains to further upper bound (405) using (407). Let

$$\mathcal{T}_{k,n} = \left\{(s^k, x^n, y^n) \in \mathcal{S}^k \times \mathcal{A}^n \times \mathcal{B}^n : \right.$$
$$\sum_{i=1}^{n}\imath^\star_{X;Y}(x_i; y_i) - \sum_{i=1}^{k}\jmath_S(s_i, d)$$
$$\left. \geq nC - kR(d) - \sqrt{nV + k\mathcal{V}(d)}Q^{-1}(\epsilon_{k,n})\right\} \tag{409}$$

By the Berry-Esseen bound (Theorem 25),

$$\mathbb{P}\left[(S^k, X^{n\star}, Y^{n\star}) \notin \mathcal{T}_{k,n}\right] \leq \epsilon_{k,n} + \frac{B}{\sqrt{n+k}} \tag{410}$$

so the expectation in the right side of (405) is upper-bounded as

$$\mathbb{E}\left[\exp\left(-|U_{k,n}|^+\right)\right]$$
$$\leq \mathbb{E}\left[\exp\left(-|U_{k,n}|^+ \mathbb{1}\left\{(S^k, X^{n\star}, Y^{n\star}) \in \mathcal{T}_{k,n}\right\}\right)\right]$$
$$+ \mathbb{P}\left[(S^k, X^{n\star}, Y^{n\star}) \notin \mathcal{T}_{k,n}\right] \tag{411}$$
$$\leq \frac{1}{\sqrt{k}}\mathbb{P}\left[(S^k, X^{n\star}, Y^{n\star}) \in \mathcal{T}_{k,n}\right] + \epsilon_{k,n} + \frac{B}{\sqrt{n+k}} \tag{412}$$

where we used (407) and (409) to upper bound the exponent in the right side of (411).

Putting (405) and (412) together, we conclude that $\epsilon' \leq \epsilon$.

Finally, consider the case $V = \mathcal{V}(d) = 0$, which implies $\jmath_S(S, d) = R(d)$ and $\imath^\star_{X;Y}(X_i^\star; Y_i^\star) = C$ almost surely, and let k and n be such that

$$nC - kR(d) \geq \left(\bar{c} - \frac{1}{2}\right)\log k + \log\gamma + c + \log\frac{1}{\epsilon - \frac{K+1}{\sqrt{k}}} \tag{413}$$

where constants c and $\bar{c}$ are defined in Lemma 5. Then

$$\mathbb{E}\left[\exp\left(-|U_{k,n}|^+\right)\right] \leq \epsilon - \frac{K+1}{\sqrt{k}} \tag{414}$$

which, together with (405), implies that $\epsilon' \leq \epsilon$, as desired.

C. Lossy or almost lossless coding over a Gaussian channel

In view of Remark 12, it suffices to consider the equal power constraint (133). As shown in the proof of Theorem 18, for any distribution of X^n on the power sphere,

$$\imath_{X^n;Y^n}(X^n; Y^n) \geq G_n - F \tag{415}$$

where G_n is defined in (392) (cf. (168)) and F is a (computable) constant.

Now, the proof for almost lossless coding in Appendix D-A can be modified to work for the Gaussian channel by adding $\log F$ to the right side of (394) and replacing $\sum_{i=1}^{n}\imath^\star_{X;Y}(X_i^\star; Y_i^\star)$ in (393) and (397) with $G_n - \log F$, and in (395) with G_n.

Similarly, the proof for lossy coding in Appendix D-B is adapted for the Gaussian channel by adding $\log F$ to the right side of (407) and replacing $\sum_{i=1}^{n}\imath^\star_{X;Y}(X_i^\star; Y_i^\star)$ in (404) and (406) with $G_n - \log F$, and in (409) with G_n.

APPENDIX E
PROOF OF THEOREM 22

Applying the Berry-Esseen bound to (190), we obtain

$$D_1(n, \epsilon, \alpha)$$

$$\geq \min_{\substack{P_{Z|S}: \\ I(S;Z) \leq C(\alpha)}} \left\{ \mathbb{E}\left[d(S,Z)\right] + \sqrt{\frac{\text{Var}\left[d(S,Z)\right]}{n}} Q^{-1}\left(\epsilon + \frac{B}{\sqrt{n}}\right) \right\} \tag{416}$$

$$= D(C(\alpha)) + \sqrt{\frac{\mathcal{W}_1(\alpha)}{n}} Q^{-1}\left(\epsilon + \frac{B}{\sqrt{n}}\right) \tag{417}$$

where B is the Berry-Esseen ratio, and (417) follows by the application of Lemma 3 with

$$\mathcal{D} = \left\{ P_{SZ} = P_{Z|S}P_S : I(S;Z) \leq R(\bar{d}) \right\} \tag{418}$$

$$f(P_{SZ}) = -\mathbb{E}\left[d(S,Z)\right] \tag{419}$$

$$g(P_{SZ}) = -\sqrt{\text{Var}\left[d(S,Z)\right]} Q^{-1}\left(\epsilon + \frac{B}{\sqrt{n}}\right) \tag{420}$$

$$\varphi = 1 \tag{421}$$

$$\psi = \frac{1}{\sqrt{n}} \tag{422}$$

Note that the mean and standard deviation of $d(S,Z)$ are linear and continuously differentiable in P_{SZ}, respectively, so conditions (242) and (244) hold with the metric being the usual Euclidean distance between vectors in $\mathbb{R}^{|\mathcal{S}| \times |\hat{\mathcal{S}}|}$. So, (417) follows immediately upon observing that by the definition of the rate-distortion function, $\mathbb{E}\left[d(S,Z)\right] \geq \mathbb{E}\left[d(S,Z^\star)\right] = D(C(\alpha))$ for all $P_{Z|S}$ such that $I(S;Z) \leq C(\alpha)$.

REFERENCES

[1] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," in *2012 IEEE International Symposium on Information Theory*, Cambridge, MA, July 2012, pp. 1553–1557.

[2] ——, "To code or not to code: Revisited," in *Proceedings 2012 IEEE Information Theory Workshop*, Lausanne, Switzerland, 2012.

[3] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July and October 1948.

[4] ——, "Coding theorems for a discrete source with a fidelity criterion," *IRE Int. Conv. Rec.*, vol. 7, pp. 142–163, Mar. 1959, reprinted with changes in *Information and Decision Processes*, R. E. Machol, Ed. New York: McGraw-Hill, 1960, pp. 93-126.

[5] S. Verdú and I. Kontoyiannis, "Lossless data compression rate: Asymptotics and non-asymptotics," in *Proceedings 2012 46th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, March 2012, pp. 1–6.

[6] ——, "Lossless data compression at finite blocklengths," *submitted to IEEE Transactions on Information Theory*, 2012.

[7] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[8] 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.

[9] 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.

[10] I. Csiszár, "Joint source-channel error exponent," *Prob. Contr. & Info. Theory*, vol. 9, no. 5, pp. 315–328, Sep. 1980.

[11] ——, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 823–828, Nov. 1982.

[12] R. Pilc, "Coding theorems for discrete source-channel pairs," Ph.D. dissertation, M.I.T., 1967.

[13] ——, "The transmission distortion of a discrete source as a function of the encoding block length," *Bell Syst. Tech. J.*, vol. 47, no. 6, pp. 827–885, July/August 1968.

[14] A. D. Wyner, "Communication of analog data from a Gaussian source over a noisy channel," *Bell Syst. Tech. J.*, vol. 47, no. 5, pp. 801–812, May/June 1968.

[15] ——, "On the transmission of correlated Gaussian data over a noisy channel with finite encoding block length," *Information and Control*, vol. 20, no. 3, pp. 193–215, April 1972.

[16] I. Csiszár, "An abstract source-channel transmission problem," *Problems of Control and Information Theory*, vol. 12, no. 5, pp. 303–307, Sep. 1983.

[17] A. Tauste Campo, G. Vazquez-Vilar, A. Guillén i Fàbregas, and A. Martinez, "Random-coding joint source-channel bounds," in *Proceedings 2011 IEEE International Symposium on Information Theory*, Saint-Petersburg, Russia, Aug. 2011, pp. 899–902.

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

[19] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, no. 1, pp. 57–71, Jan. 1974.

[20] H. V. Poor, *An introduction to signal detection and estimation*. Springer, 1994.

[21] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 44 –54, Jan 1995.

[22] D. Sakrison, "A geometric treatment of the source encoding of a Gaussian random variable," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 481–486, May 1968.

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

[24] F. Jelinek, *Probabilistic information theory: discrete and memoryless models*. McGraw-Hill, 1968.

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

[26] J. Goblick, T., "Theoretical limitations on the transmission of data from analog sources," *IEEE Transactions on Information Theory*, vol. 11, no. 4, pp. 558–567, Oct. 1965.

[27] R. Gallager, *Information theory and reliable communication*. John Wiley & Sons, Inc. New York, 1968.

[28] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proceedings 2007 IEEE International Symposium on Information Theory*, Nice, France, June 2007, pp. 566–570.

[29] T. Courtade and R. Wesel, "Multiterminal source coding with an entropy-based distortion measure," in *Proceedings 2011 IEEE International Symposium on Information Theory*, Saint-Petersburg, Russia, Aug. 2011, pp. 2040–2044.

[30] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 136–152, Jan. 2000.

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

Victoria Kostina (S'12) received the Bachelors degree with honors in applied mathematics and physics from the Moscow Institute of Physics and Technology, Russia, in 2004, where she was affiliated with the Institute for Information Transmission Problems of the Russian Academy of Sciences, and the Masters degree in electrical engineering from the University of Ottawa, Canada, in 2006.

She is currently pursuing a Ph.D. degree in electrical engineering at Princeton University. Her research interests lie in information theory, theory of random processes, coding, and wireless communications.

Sergio Verdú (S'80–M'84–SM'88–F'93) received the Telecommunications Engineering degree from the Universitat Politècnica de Barcelona in 1980, and the Ph.D. degree in Electrical Engineering from the University of Illinois at Urbana-Champaign in 1984. Since 1984 he has been a member of the faculty of Princeton University, where he is the Eugene Higgins Professor of Electrical Engineering, and is a member of the Program in Applied and Computational Mathematics.

Sergio Verdú is the recipient of the 2007 Claude E. Shannon Award, and the 2008 IEEE Richard W. Hamming Medal. He is a member of the National Academy of Engineering, and was awarded a Doctorate Honoris Causa from the Universitat Politècnica de Catalunya in 2005.

He is a recipient of several paper awards from the IEEE: the 1992 Donald Fink Paper Award, the 1998 and 2012 Information Theory Paper Awards, an Information Theory Golden Jubilee Paper Award, the 2002 Leonard Abraham Prize Award, the 2006 Joint Communications/Information Theory Paper Award, and the 2009 Stephen O. Rice Prize from IEEE Communications Society. In 1998, Cambridge University Press published his book *Multiuser Detection,* for which he received the 2000 Frederick E. Terman Award from the American Society for Engineering Education.

Sergio Verdú served as President of the IEEE Information Theory Society in 1997, and on its Board of Governors (1988-1999, 2009-present). He has also served in various editorial capacities for the *IEEE Transactions on Information Theory*: Associate Editor (Shannon Theory, 1990-1993; Book Reviews, 2002-2006), Guest Editor of the Special Fiftieth Anniversary Commemorative Issue (published by IEEE Press as "Information Theory: Fifty years of discovery"), and member of the Executive Editorial Board (2010-2013). He is the founding Editor-in-Chief of *Foundations and Trends in Communications and Information Theory*.