Strong Functional Representation Lemma and Applications to Coding Theorems

Cheuk Ting Li, *Member, IEEE*, and Abbas El Gamal, *Life Fellow, IEEE*

***Abstract*— This paper shows that for any random variables X and Y, it is possible to represent Y as a function of (X, Z) such that Z is independent of X and $I(X; Z|Y) \leq \log(I(X; Y)+1)+4$ bits. We use this strong functional representation lemma (SFRL) to establish a bound on the rate needed for one-shot exact channel simulation for general (discrete or continuous) random variables, strengthening the results by Harsha *et al.* and Braverman and Garg, and to establish new and simple achievability results for one-shot variable-length lossy source coding, multiple description coding, and Gray–Wyner system. We also show that the SFRL can be used to reduce the channel with state noncausally known at the encoder to a point-to-point channel, which provides a simple achievability proof of the Gelfand–Pinsker theorem.**

***Index Terms*— Functional representation lemma, channel simulation, one-shot achievability, lossy source coding, channel with state.**

I. Introduction

THE functional representation lemma [1, p. 626] states that for any random variables X and Y, there exists a random variable Z independent of X such that Y can be represented as a function of X and Z. This result has been used to establish several results in network information theory beginning with the early work of Hajek and Pursley on the broadcast channel [2] and Willems and van der Meulen on the multiple access channel with cribbing encoders [3].

The random variable Z in the functional representation lemma can be intuitively viewed as the part of Y which is not contained in X. However, Z is not necessarily unique. For example, let B_1, B_2, B_3, B_4 be i.i.d. Bern$(1/2)$ random variables and define $X = (B_1, B_2, B_3)$ and $Y = (B_2, B_3, B_4)$. Then both $Z_1 = B_4$ and $Z_2 = B_1 \oplus B_4$ satisfy the functional representation lemma. However, $H(Y|Z_1) = 2$ while $H(Y|Z_2) = 3$, that is, Z_1 provides more information about Y than Z_2. In general, $H(Y|Z) = I(X; Y|Z) + H(Y|X, Z) = I(X; Y, Z) \geq I(X; Y)$. For our example $H(Y|Z_1) = I(X; Y) = 2$, that is, Z_1 is the most informative Z about Y. What is the most informative Z about Y in general? Does it always achieve the lower bound $H(Y|Z) \geq I(X; Y)$?

Manuscript received April 19, 2017; revised January 19, 2018 and June 16, 2018; accepted August 3, 2018. Date of publication August 16, 2018; date of current version October 18, 2018. This work was supported by a gift from Huawei Technologies. This paper was presented in part at the 2017 IEEE International Symposium on Information Theory.

C. T. Li was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is now with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: ctli@berkeley.edu).

A. El Gamal is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: abbas@ee.stanford.edu).

Communicated by V. Vaishampayan, Associate Editor for Source Coding.
Digital Object Identifier 10.1109/TIT.2018.2865570

In this paper, we show that for general (X, Y), their exists a Z such that $H(Y|Z)$ is close to $I(X; Y)$. Specifically, we strengthen the functional representation lemma to show that for any X and Y, there exists a Z independent of X such that Y is a function of X and Z, *and*

$$I(X; Z|Y) \leq \log(I(X; Y) + 1) + 4.$$

Alternatively this can be expressed as

$$H(Y|Z) \leq I(X; Y) + \log(I(X; Y) + 1) + 4. \tag{1}$$

We use the above *strong functional representation lemma* (SFRL) together with an optimal prefix code such as a Huffman code to establish one-shot, *variable-length* achievability results for channel simulation [4], Shannon's lossy source coding [5], multiple description coding [6], [7] and lossy Gray–Wyner system [8]. These one-shot achievability results can be stated in terms of mutual information, without the need of information density or other quantities. We then show how the SFRL can be used to reduce the channel with state known at the encoder to a point-to-point channel, providing a simple proof to the Gelfand-Pinsker theorem [9]. The asymptotic block coding counterparts of these one-shot results can be readily obtained by converting the variable-length code into a block code and incurring an error probability that vanishes as the block length approaches infinity.

A weaker form of the SFRL for discrete random variables follows from the result by Harsha *et al.* [4] on the one-shot exact channel simulation with unlimited common randomness. Their result implies that $I(X; Z|Y) \leq (1 + \epsilon) \log(I(X; Y) + 1) + c_\epsilon$ is achievable, where $\epsilon > 0$ and c_ϵ is a function of ϵ. This result was later strengthened by Braverman and Garg [10] to $I(X; Z|Y) \leq \log(I(X; Y) + 1) + c$ (note that replacing the universal code in [4] by a code for a suitable power law distribution can also yield the same improvement). It is also shown in [10] that there exist examples for which the log term is necessary. SFRL strengthens these results in two ways; first it generalizes the bound to random variables with arbitrary distributions (whereas the results in [4] and [10] only applies to discrete distributions), and second it provides a bound with a small additive constant of 4 (whereas the constants in [4] and [10] are unspecified). Our stronger result is established using a new construction of Z and g that we refer to as the *Poisson functional representation*, instead of the rejection sampling approach in [4] and [10]. Perhaps more importantly, we are the first to show that the result in [4] can be considered as a strengthened functional representation lemma, which led us to explore applications in source and channel coding.

One-shot achievability results using fixed length (random) coding have been recently established for lossy source coding and several settings in network information theory. Liu *et al.* [11] established a one-shot achievability result for lossy source coding using channel resolvability. One-shot quantum lossy source coding settings were investigated by Datta *et al.* [12]. Verdú [13] introduced non-asymptotic packing and covering lemmas and used them to establish one-shot achievability results for several settings including Gelfand-Pinsker. Liu *et al.* [14] proved a one-shot mutual covering lemma and used it to establish a one-shot achievability result for the broadcast channel. Watanabe *et al.* [15] established several one-shot achievability results for coding with side-information (including Gelfand-Pinsker). Yassaee *et al.* [16] established several one-shot achievability results, including Gelfand-Pinsker and multiple description coding. Most of these results are stated in terms of information density and various other quantities. In contrast, our one-shot achievability results using variable-length codes are all stated in terms of only mutual information. Moreover, given the SFRL, our proofs are generally simpler.

Variable-length (one-shot, finite blocklength or asymptotic) lossy source coding settings have been studied, see [17]–[21]. Some of these works concern the universal setting in which the distribution of the source is unknown, hence the use of variable-length codes is justified. In contrast, the reason we consider variable-length codes in this paper is that it allows us to give one-shot results that subsume their asymptotic fixed-length counterparts.

In the following section, we state the SFRL, introduce the Poisson functional representation construction and provide a sketch of the proof of the lemma. The complete proof is given in Appendix A. In Sections III and IV we use SFRL to establish one-shot achievability results for channel simulation and three source coding settings, respectively. In Section V, we use SFRL together with Shannon's channel coding theorem to provide a simple achievability proof of the Gelfand–Pinsker theorem. Finally in Section VI we prove a lower bound on $I(X;Z|Y)$ in SFRL (whereas SFRL is an upper bound) and discuss several other properties.

Notation

Throughout this paper, we assume that log is base 2 and the entropy H is in bits. We use the notation: $X_a^b = (X_a, \ldots, X_b)$, $X^n = X_1^n$, $[a:b] = [a,b] \cap \mathbb{Z}$ and $[a] = [1:a]$.

For discrete X, we write the probability mass function as p_X. For continuous X, we write the probability density function as f_X. For general random variable X, we write the probability measure (push-forward measure by X) as P_X.

II. Strong Functional Representation Lemma

The main result in this paper is given in the following.

Theorem 1 (Strong Functional Representation Lemma): For any pair of random variables $(X,Y) \sim \mathsf{P}_{XY}$ (over a Polish space with Borel probability measure) with $I(X;Y) < \infty$, there exists a random variable Z independent of X such that Y can be expressed as a function $g(X,Z)$ of X and Z, and

$$I(X;Z|Y) \le \log(I(X;Y)+1) + 4.$$

Moreover, if X and Y are discrete with cardinalities $|\mathcal{X}|$ and $|\mathcal{Y}|$, respectively, then $|\mathcal{Z}| \le |\mathcal{X}|(|\mathcal{Y}|-1)+2$.

Note that SFRL can be applied conditionally; given $\mathsf{P}_{XY|U}$, we can represent Y as a function $g(X,Z,U)$ such that Z is independent of (X,U) and

$$I(X;Z|Y,U) \le \log\left(I(X;Y|U)+1\right) + 4. \tag{2}$$

We can have $Z \perp\!\!\!\perp (X,U)$, not only $Z \perp\!\!\!\perp X \mid U$ which follows from directly applying SFRL for each value of U. The reason is that by the functional representation lemma, we can represent Z as a function of U and $\tilde{Z}$ such that $\tilde{Z} \perp\!\!\!\perp U$ (which, together with $\tilde{Z} \perp\!\!\!\perp X \mid U$, gives $\tilde{Z} \perp\!\!\!\perp (X,U)$), and use $\tilde{Z}$ instead of Z.

Note that SFRL applies to general distributions P_{XY}. Although $H(Y)$ may be infinite, as long as $I(X;Y)$ is finite, the cardinality of Y conditioned on Z is countable and $H(Y|Z)$ is finite by SFRL. Since $Z \perp\!\!\!\perp X$ and $H(Y|X,Z)=0$ imply that $I(X;Z|Y) = H(Y|Z) - I(X;Y)$, the SFRL implies the existence of a $Z \perp\!\!\!\perp X$ such that $H(Y|Z)$ is close to $I(X;Y)$.

To prove the SFRL, we use the following random variable Z and function g construction.

Definition 1 (Poisson Functional Representation): Fix any joint distribution P_{XY}. Let $0 \le T_1 \le T_2 \le \cdots$ be a Poisson point process with rate 1 (i.e., the increments $T_i - T_{i-1}$ are i.i.d. Exp(1) for $i = 1,2,\ldots$ with $T_0 = 0$), and $\tilde{Y}_1, \tilde{Y}_2, \ldots$ be i.i.d. with $\tilde{Y}_1 \sim \mathsf{P}_Y$. Take $Z = \{(\tilde{Y}_i, T_i)\}_{i=1,2,\ldots}$, i.e., a marked Poisson point process. Then we can let $Y = g_{X\to Y}(X,Z)$, where

$$g_{X\to Y}(x,\ \{(\tilde{y}_i,t_i)\}) = \tilde{y}_{k_{X\to Y}(x,\ \{(\tilde{y}_i,t_i)\})},$$

and

$$k_{X\to Y}(x,\ \{(\tilde{y}_i,t_i)\}) = \arg\min_i\ t_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{y}_i),$$

where we write $(d\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x))(y) = ((d\,\mathsf{P}_{Y|X}(\cdot|x)/d\mathsf{P}_Y)(y))^{-1}$ for the Radon-Nikodym derivative (i is not chosen if $(d\,\mathsf{P}_{Y|X}(\cdot|x)/d\mathsf{P}_Y)(\tilde{y}_i) = 0$).

To illustrate this Poisson functional representation, consider the following.

Example 1: Let $Y \sim \mathrm{Unif}[0,1]$ and $Y|\{X=x\} \sim f_{Y|X}(y|x)$. Then $g_{X\to Y}(x,z) = \tilde{y}_k$ where $k = \arg\min_i t_i / f_{Y|X}(y|x)$. Figure 1 shows an example of $z = \{(\tilde{y}_i,t_i)\}$. The index k is selected by scaling up the graph of $f_{Y|X}(y|x)$ until it hits the first point, then we output $\tilde{y}_k$ of that point ($\tilde{y}_3$ in the figure). It is straightforward to check that this procedure gives the correct conditional distribution $Y|\{X=x\} \sim f_{Y|X}(y|x)$. Roughly speaking, if $I(X;Y)$ is small, then $Y|\{X=x\}$ will be close to the uniform distribution for most x's, and the $\tilde{y}_k$'s with smaller indices k's will be more likely to be output, and therefore $H(Y|Z)$ will be smaller. (If $I(X;Y)=0$, then $Y|\{X=x\} \sim \mathrm{Unif}[0,1]$ and $\tilde{y}_1$ is output for almost all x, and hence $H(Y|Z)=0$.)

Remark 1: Exponential representation for discrete Y Let $Y \in \{1,\ldots,l\}$ be discrete, then we can simplify the construction of Z in the definition of Poisson functional representation

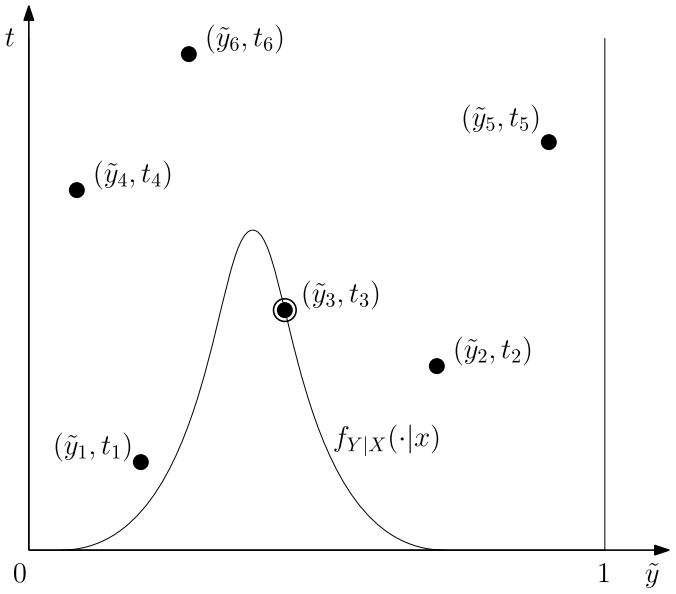

Fig. 1. Illustration of the Poisson functional representation construction for Example 1.

to first generating a sequence of i.i.d. exponential random variables $Z_1, \ldots, Z_l$ with $Z_1 \sim \text{Exp}(1)$ and setting

$$g_{X\to Y}(x, z^l) = \arg\min_{y\in\mathcal{Y}} \frac{z_y}{p_{Y|X}(y|x)}.$$

Since the arg min of independent exponential random variables with different rates has a pmf proportional to the rates, we have $g_{X\to Y}(x, Z^l) \sim p_{Y|X}(\cdot|x)$. Moreover, if $|\mathcal{Y}|$ is finite, then normalizing $\{Z_y\}_{y\in\mathcal{Y}}$ by scaling each Z_y by the same factor such that $\sum_y Z_y = 1$ (making $\{Z_y\}$ uniformly distributed over the probability simplex on $\mathcal{Y}$) would not affect $g_{X\to Y}(X, Z^l)$. Hence we can assume $\{Z_y\}$ is uniformly distributed over the simplex.

To see the connection to the Poisson construction, note that we can equivalently let $Z_y = p_Y(y) \cdot \min_{i:\,\tilde{Y}_i = y} T_i$, where T_i is as defined before, and $Y = g_{X\to Y}(X, Z^l)$.

We now proceed to give a sketch of the proof of Theorem 1 by showing that the Poisson functional representation satisfies the constraints. The complete proof is given in Appendix A.

Proof: [Sketch of the Proof of Theorem 1] Consider the Poisson functional representation. Let $Y = \tilde{Y}_K$,

$$K = k_{X\to Y}(X, \{(\tilde{Y}_i, T_i)\}) = \arg\min_i T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|X)}(\tilde{Y}_i).$$

Since Y is a function of Z and K, we have $H(Y|Z) \le H(K)$. We now proceed to bound $H(K)$.

Condition on $X = x$. Since $T_1 \le T_2 \le \cdots$, K is small when $d\,\mathsf{P}_Y(y)/d\,\mathsf{P}_{Y|X}(y|x)$ for different y's are close to 1, i.e., P_Y is close to $\mathsf{P}_{Y|X}(\cdot|x)$ (if $\mathsf{P}_Y = \mathsf{P}_{Y|X}(\cdot|x)$ for all y, then $d\,\mathsf{P}_Y(y)/d\,\mathsf{P}_{Y|X}(y|x) = 1$, and $K = 1$). In fact we can prove that

$$\mathsf{E}\left[\log K|X = x\right] \le D(\mathsf{P}_{Y|X}(\cdot|x)\|\,\mathsf{P}_Y) + e^{-1}\log e + 1.$$

The proof is given in Appendix A. Therefore $\mathsf{E}\left[\log K\right] \le I(X;Y) + e^{-1}\log e + 1$. By the maximum entropy distribution subject to a given $\mathsf{E}\left[\log K\right]$, we have

$$H(K) \le \mathsf{E}\left[\log K\right] + \log\left(\mathsf{E}\left[\log K\right] + 1\right) + 1.$$

The proof of this bound is given in Appendix B for the sake of completeness. Hence

$$\begin{aligned} &H(K)\\ &\le I(X;Y) + e^{-1}\log e + 2 + \log\left(I(X;Y) + e^{-1}\log e + 2\right)\\ &\le I(X;Y) + \log(I(X;Y)+1) + e^{-1}\log e + 2 + \log(e^{-1}\log e + 2)\\ &< I(X;Y) + \log\left(I(X;Y)+1\right) + 4. \end{aligned}$$

Operationally, K can be encoded using the optimal prefix-free code for the Zipf distribution $q(k) \propto k^{-\lambda}$, where

$$\lambda = 1 + 1/(I(X;Y) + e^{-1}\log e + 1). \tag{3}$$

It can be checked that the expected length of the codeword is upper bounded by $I(X;Y) + \log(I(X;Y)+1) + 5$. ■

Remark 2: The Poisson functional representation is a *non-causal* scheme, meaning that in order to determine whether to output $\tilde{y}_i$, one has to look at future $\tilde{y}_j$'s, $j > i$. While a future $\tilde{y}_j$ has larger t_j, it can be chosen if it has a much smaller $(d\,\mathsf{P}_Y/d\,\mathsf{P}_{Y|X}(\cdot|x))(\tilde{y}_i)$. In comparison, the schemes in [4] and [10], which are based on rejection sampling, are causal. Causality is irrelevant in the applications in this paper, however, and will not be discussed further.

III. One-Shot Channel Simulation

Channel simulation aims to find the minimum amount of communication over a noiseless channel needed to simulate a memoryless channel $\mathsf{P}_{Y|X}$. Several settings of this problem have been studied, see [22]–[24]. Consider the one-shot channel simulation with unlimited common randomness setup [4] in which Alice and Bob share unlimited common randomness W. Alice observes $X \sim \mathsf{P}_X$ and sends a prefix-free description M to Bob via a noiseless channel such that Bob can generate Y (from M and W) according to a prescribed conditional distribution $\mathsf{P}_{Y|X}$. The problem is to find the minimum expected description length of M, $\mathsf{E}[L(M)]$, needed. Since we have the Markov chain $X - M - Y$ conditional on W,

$$\begin{aligned} \mathsf{E}[L(M)] &\ge H(M|W)\\ &\ge I(X;Y|W)\\ &= I(X;Y,W) - I(X;W)\\ &= I(X;Y,W)\\ &\ge I(X;Y). \end{aligned}$$

In [10], which strengthens the result in [4], it is shown that for X and Y discrete,

$$\mathsf{E}[L(M)] \le I(X;Y) + \log(I(X;Y)+1) + c$$

is achievable, where c is an unspecified constant.

We now show that the SFRL provides an upper bound on $\mathsf{E}[L(M)]$ that applies to arbitrary (not only discrete) channels. By the SFRL (1), there exists a Z independent of X such that $Y = g_{X\to Y}(X, Z)$ and

$$H(Y|Z) \le I(X;Y) + \log(I(X;Y)+1) + 4.$$

We use $W = Z$ as the common randomness. Upon observing $X = x$, Alice computes $y = g_{X\to Y}(x, z)$ and encodes y using a Huffman code for the pmf $p_{Y|Z}(\cdot|z)$ into the description m (note that Y can be arbitrary but by the SFRL $Y|\{Z = z\}$ is discrete). Bob then recovers y from m and z. The expected length is

$$\mathsf{E}[L(M)] \le I(X; Y) + \log(I(X; Y) + 1) + 5.$$

In practice, instead of using a Huffman code (which may be impractical since $p_{Y|Z}(\cdot|z)$ is not easy to compute), we can compress $k = k_{X\to Y}(x, z)$ in the Poisson functional representation into m using the optimal prefix-free code for the Zipf distribution (3).

Moreover, for discrete X, Y, the amount of the common randomness can be bounded by $\log|\mathcal{W}| \le \log(|\mathcal{X}|(|\mathcal{Y}| - 1) + 2)$. In comparison, the amount of the common randomness in [4] can be bounded by $O(\log(|\mathcal{X}||\mathcal{Y}|))$ only if the expected description length is increased by $O(\log\log(|\mathcal{X}| + |\mathcal{Y}|))$.

Remark 3: In [4], the setting in which $X = x$ is an arbitrary input (instead of $X \sim p_X$) is studied. It is shown that

$$\mathsf{E}[L(M)] \le C + (1 + \epsilon)\log(C + 1) + c_\epsilon$$

for all $x \in \mathcal{X}$ is achievable, where C is the capacity of the channel $p_{Y|X}$ and c_ϵ is a function of ϵ.

The Poisson functional representation can still be applied to this setting. If we encode $k = k_{X\to Y}(x, z)$ into M using the optimal prefix-free code for the Zipf distribution $q(k) \propto k^{-\lambda}$, where $\lambda = 1+1/(C+e^{-1}\log e+1)$, then by the same argument in the proof of the SFRL, and [4, Claim 3.1],

$$\mathsf{E}[L(M)] \le C + \log(C + 1) + 5$$

is achievable.

We can also prove a cardinality bound of the common randomness Z in this setting. Applying Carathéodory's theorem on the $(|\mathcal{X}||\mathcal{Y}|)$-dimensional vectors with entries $\mathsf{E}[\log K|X = x, Z = z]$ and $p(x, y|z)$ for $x \in \{1, \ldots, |\mathcal{X}|\}$, $y \in \{1, \ldots, |\mathcal{Y}| - 1\}$, we have the cardinality bound $|\mathcal{Z}| \le |\mathcal{X}||\mathcal{Y}| + 1$.

Remark 4: A generalization to this problem, referred as message compression in the study of communication complexity, and related to Slepian-Wolf coding [25], concerns the case in which Bob also observes the side information U correlated with X (see [26], [27]). SFRL cannot be directly applied in this case since the conditional version of SFRL requires Y to be a function of (X, Z, U), though Alice does not observe U. Such generalization is beyond the scope of this paper.

IV. Lossy Source Coding

We use the SFRL to establish one-shot achievability results for three lossy source coding settings.

A. Lossy Source Coding

Consider the following one-shot variable-length lossy source coding problem. We are given a random variable (source) $X \in \mathcal{X}$ with $X \sim \mathsf{P}_X$, a reproduction alphabet $\mathcal{Y}$, and a distortion function $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ (note that X, Y can be arbitrary, and $d(x, y)$ can be infinite). Given X, the encoder selects $\tilde{Y} \in \mathcal{Y}$ and encodes it using a prefix-free code into $M \in \{0, 1\}^*$. The decoder recovers $\tilde{Y}$ from M. Let $\bar{R} = \mathsf{E}[L(M)]$ be the expected value of the length of the description M and $\mathsf{E}[d(X, \tilde{Y})]$ be the average distortion of representing X by $\tilde{Y}$. An expected length-distortion pair $(\bar{R}, D)$ is said to be achievable if there exists a variable-length code with expected description length $\bar{R}$ such that $\mathsf{E}[d(X, \tilde{Y})] \le D$.

In the following we use the SFRL to establish a set of achievable $(\bar{R}, D)$ pairs.

Theorem 2: A pair $(\bar{R}, D)$ is achievable for the one-shot variable-length lossy source coding problem with source $X \sim \mathsf{P}_X$, reproduction alphabet $\mathcal{Y}$, and distortion measure $d(x, y)$ if

$$\bar{R} > R(D) + \log(R(D) + 1) + 6,$$

where

$$R(D) = \inf_{\mathsf{P}_{Y|X}:\, \mathsf{E}[d(X,Y)]\le D} I(X; Y)$$

is the (asymptotic) rate-distortion function [5].

Proof: Let Y be the random variable that attains $\mathsf{E}[d(X, Y)] \le D$ and $I(X; Y) \le R(D) + \epsilon$. By the SFRL (1), there exists Z independent of X such that $Y = g_{X\to Y}(X, Z)$ and

$$H(g_{X\to Y}(X, Z)|Z) \le I(X; Y) + \eta,$$

where $\eta = \log(I(X; Y) + 1) + 4$. Consider the set

$$A = \{(H(g_{X\to Y}(X, z)),\ \mathsf{E}_X[d(X,\ g_{X\to Y}(X, z))]) : z \in \mathcal{Z}\}.$$

The point $(H(g_{X\to Y}(X, Z)|Z),\ \mathsf{E}[d(X, Y)])$ is a weighted average of the points in A (and thus is in the convex hull of A). Hence there exists z satisfying the rate constraint $H(g_{X\to Y}(X, z)) \le H(g_{X\to Y}(X, Z)|Z)$, and there exists z' satisfying the distortion constraint $\mathsf{E}_X\left[d(X,\ g_{X\to Y}(X, z'))\right] \le \mathsf{E}[d(X, Y)]$. However, there may not exist a single z simultaneously satisfying both constraints. Hence we invoke Carathéodory's theorem to find a mixture between two points z_0, z_1 and $\lambda \in [0, 1]$ such that both constraints are satisfied:

$$\begin{aligned}(1-\lambda)&H(g_{X\to Y}(X, z_0)) + \lambda H(g_{X\to Y}(X, z_1))\\ &\le H(g_{X\to Y}(X, Z)|Z) \le I(X; Y) + \eta,\\ &\quad \times(1-\lambda)\mathsf{E}_X[d(X,\ g_{X\to Y}(X, z_0))]\\ &\quad + \lambda\mathsf{E}_X[d(X,\ g_{X\to Y}(X, z_1))]\\ &\le \mathsf{E}[d(X, Y)].\end{aligned}$$

Note that to satisfy the above inequalities, we need one point less than stated in Carathéodory's theorem. Take $Q \sim \mathrm{Bern}(\lambda)$, $\tilde{Y} = g_{X\to Y}(X, z_Q)$. Then

$$H(\tilde{Y}) \le H(\tilde{Y}|Q) + H(Q) \le H(\tilde{Y}|Q) + 1 \le I(X; Y) + \eta + 1.$$

We use a Huffman code to encode $\tilde{Y}$ and obtain an expected length $\bar{R} \le H(\tilde{Y}) + 1$. The result follows by letting $\epsilon \to 0$. ∎

Note that a stochastic encoder is used in the proof. Nevertheless, the encoder only needs to randomize between two deterministic encoding functions in order to achieve Theorem 2.

An interesting implication of Theorem 2 is that for any source P_X, distortion measure $d(x, y)$, and distortion level D, the optimal asymptotic rate $R(D)$ cannot be too far from the optimal one-shot expected description length $\bar{R}^*(D) = \inf\{\bar{R} : (\bar{R}, D)$ achievable$\} \le R(D)+\log(R(D)+1)+6$. For example, there does not exist $(\mathsf{P}_X, d(x, y), D)$, where $R(D) = 100$ but $\bar{R}^*(D) \ge 113$. This is a benefit of considering variable-length codes. Such conclusion does not hold if we consider fixed-length codes instead (e.g., if $X \sim \text{Geom}(1/2)$, $d(x, y) = \mathbf{1}\{x \ne y\}$, then $R(D) \le 2$ for any $D \ge 0$, but the optimal length of the one-shot fixed-length code tends to infinity as $D \to 0$).

Although the above achievability proof does not use random coding, it can be interpreted as using the following *soft random coding* scheme.

Soft codebook generation. The random variable $Z = \{(\tilde{Y}_i, T_i)\}_{i=1,2,\ldots}$ produced by the Poisson functional representation represents the choice of the codebook. We select a "soft codebook" by conditioning on $Z = \{(\tilde{y}_i, t_i)\}_{i=1,2,\ldots}$. Unlike conventional codebook $\mathcal{C} \subseteq Y$ which contains a fixed number of y's, a soft codebook $\{(\tilde{y}_i, t_i)\}$ contains an infinite sequence of $\tilde{y}_i$'s, each with a weight t_i (the smaller t_i is, the more likely $\tilde{y}_i$ is chosen).
Encoding. The encoder observes x and finds the reconstruction $\tilde{y}_k$ where

$$k = \arg\min_i \; t_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{y}_i).$$

It then encodes the index k using an optimal prefix-free code for the Zipf distribution (3). This is analogous to a conventional codebook generation in which we find the closest $y \in \mathcal{C}$ to x and encodes it into its index in $\mathcal{C}$. Here we use a prefix-free code over the positive integers to encode the index into the description m because the index k can be unbounded, but the smaller k's (with smaller t_k's) are more likely to be used so they are assigned shorter descriptions.
Decoding. The decoder receives m, recovers k, then outputs $\tilde{y}_k$.

Note that the soft random coding scheme shares some similarity with the likelihood encoder in [28], which uses a conventional i.i.d. random codebook generation $y(m) \sim \mathsf{P}_Y$, $m = 1, \ldots, 2^R$, but uses a stochastic encoder which chooses m with probability proportional to the likelihood function

$$\mathcal{L}(m|x) = p_{X|Y}(x|y(m)) \propto \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(y(m)).$$

The soft random coding scheme can be viewed as fixing the randomness in the likelihood encoder as part of the codebook.

A related one-shot variable-length lossy source coding setting with a constraint on the probability that the distortion exceed certain level (instead of average distortion) was studied in [21]. In [29], a result similar to Theorem 2 is given in the context of epsilon entropy.

The finite blocklength variable-length lossy source coding problem [17] concerns the case in which the source is memoryless and average per symbol distortion $d(x^n, y^n) = (1/n)\sum_i d(x_i, y_i)$. In [30] it is shown that the expected per symbol description length $\bar{R}/n = R(D)+(1+o(1))(1/n)\log n$ is achievable via d-semifaithful codes [31] with $d(X^n, \tilde{Y}^n) \le D$ surely. Applying Theorem 2 to X^n, we have

$$\begin{aligned}\bar{R}/n &= R(D) + (1/n)(\log(nR(D) + 1) + 6)\\ &= R(D) + (1 + o(1))(1/n)\log n.\end{aligned}$$

Hence we achieve the same redundancy as [30] albeit under the expected distortion constraint instead of the stronger sure distortion constraint using the d-semifaithful codes.

We can use Theorem 2 to establish the achievability of Shannon's (asymptotic) lossy source coding theorem [5], assuming there exists a symbol $y_0 \in \mathcal{Y}$ with finite $d(x, y_0)$ for all x. First note that the redundancy $(1 + o(1))(1/n)\log n$ in the finite block length extension can be made arbitrarily small, hence $\bar{R}/n$ can be made arbitrarily close to $R(D)$. Now we use the finite block length scheme over l blocks of n source symbols each of length n (for a total block length of nl). By the law of large numbers, the probability that the total description length is greater than $nl(R(D) + \epsilon)$ tends to 0 as the block length approaches infinity. Hence, we can construct a fixed length code out of the variable-length code by simply discarding descriptions longer than $nl(R(D) + \epsilon)$ and assigning the reconstruction sequence $(y_0, \ldots, y_0)$ to the discarded descriptions.

B. Multiple Description Coding

In this section, we use the SFRL to establish a one-shot inner bound for the variable-length multiple description coding problem, which yields an alternative proof of the El Gamal-Cover inner bound [6] and the Zhang-Berger inner bound [7], [32], [33] in the asymptotic regime. The encoder observes $X \sim \mathsf{P}_X$ and produces two prefix-free descriptions $M_1, M_2 \in \{0, 1\}^*$. Decoder 1 observes M_1 and generates $\tilde{Y}_1$ with distortion $d_1(X, \tilde{Y}_1)$. Similarly, Decoder 2 observes M_2 and produces $\tilde{Y}_2$ with distortion $d_2(X, \tilde{Y}_2)$. Decoder 0 observes M_1 and M_2 and produces $\tilde{Y}_0$ with distortion $d_0(X, \tilde{Y}_0)$. An expected description length-distortion tuple $(\bar{R}_1, \bar{R}_2, D_0, D_1, D_2)$ is said to be achievable if there exists a scheme with expected description length $\mathsf{E}[L(M_i)] \le \bar{R}_i$ and expected distortion $\mathsf{E}[d_i(X, \tilde{Y}_i)] \le D_i$.

Theorem 3: The tuple $(\bar{R}_1, \bar{R}_2, D_0, D_1, D_2)$ is achievable if

$$\begin{aligned}\bar{R}_1 &\ge I(X; Y_1, U) + 2\eta,\\ \bar{R}_2 &\ge I(X; Y_2, U) + 2\eta,\\ \bar{R}_1+\bar{R}_2 &\ge I(X; Y_0, Y_1, Y_2|U)+2I(X; U)+I(Y_1; Y_2|U)+5\eta,\\ D_i &\ge \mathsf{E}[d_i(X, Y_i)] \;\text{ for } i = 0, 1, 2\end{aligned}$$

for some $\mathsf{P}_{U,Y_0,Y_1,Y_2|X}$, where

$$\eta = \log\big(I(X; Y_0, Y_1, Y_2, U) + I(Y_1; Y_2|U) + 1\big) + 7.$$

Note that the only difference between the above region and Zhang-Berger inner bound is the addition of η, which grows like $\log n$ if we consider X^n and does not affect the asymptotic rate.

Proof: It suffices to prove the achievability of the corner point:

$$\bar{R}_1 = I(X; Y_1|U) + I(X; U) + 2\eta - 1, \tag{4}$$

$$\bar{R}_2 = I(X,Y_1; Y_2|U) + I(X; Y_0|Y_1, Y_2, U) + I(X; U) + 3\eta - 1, \tag{5}$$

$$D_i = \mathsf{E}[d_i(X, Y_i)] \quad \text{for } i = 0, 1, 2. \tag{6}$$

The desired rate region can be achieved by time sharing between this corner point and the other corner point where Y_1, Y_2 are flipped, resulting in a penalty of at most 1 bit (we can use the first bits of M_1 and M_2 to represent which corner point it is).

Applying the SFRL (1) to X, U, we have $U = g_{X\to U}(X, Z_3)$, where $Z_3 \perp\!\!\!\perp X$ such that

$$\begin{aligned} H(U|Z_3) &\le I(X; U) + \log(I(X; U) + 1) + 4 \\ &\le I(X; U) + \eta - 3. \end{aligned}$$

Applying the SFRL to X, Y_1 conditioned on U (2), we have $Y_1 = g_{X\to Y_1|U}(X, Z_1, U)$, where $Z_1 \perp\!\!\!\perp (X, U)$ such that

$$\begin{aligned} H(Y_1|U, Z_1) &\le I(X; Y_1|U) + \log(I(X; Y_1|U) + 1) + 4 \\ &\le I(X; Y_1|U) + \eta - 3. \end{aligned}$$

Applying the SFRL to $(X, Y_1), Y_2$ conditioned on U, we have $Y_2 = g_{XY_1\to Y_2|U}(X, Y_1, Z_2, U)$, $Z_2 \perp\!\!\!\perp (X, Y_1, U)$ such that

$$\begin{aligned} H(Y_2|U, Z_2) &\le I(X, Y_1; Y_2|U) + \log(I(X, Y_1; Y_2|U) + 1) + 4 \\ &\le I(X, Y_1; Y_2|U) + \eta - 3. \end{aligned}$$

Applying the SFRL to X, Y_0 conditioned on (Y_1, Y_2, U), we have $Y_0 = g_{X\to Y_0|Y_1Y_2U}(X, Z_0, Y_1, Y_2, U)$, $Z_0 \perp\!\!\!\perp (X, Y_1, Y_2, U)$ such that

$$\begin{aligned} &H(Y_0|Y_1, Y_2, U, Z_0) \\ &\quad \le I(X; Y_0|Y_1, Y_2, U) + \log(I(X; Y_0|Y_1, Y_2, U) + 1) + 4 \\ &\quad \le I(X; Y_0|Y_1, Y_2, U) + \eta - 3. \end{aligned}$$

Note that $Z_0^3 \perp\!\!\!\perp X$. Consider the convex hull of the 7-dimensional vectors

$$\begin{bmatrix} H(U|Z_0^3 = z_0^3) \\ H(Y_1|U, Z_0^3 = z_0^3) \\ H(Y_2|U, Z_0^3 = z_0^3) \\ H(Y_0|Y_1, Y_2, U, Z_0^3 = z_0^3) \\ \mathsf{E}[d_0(X, Y_0) \mid Z_0^3 = z_0^3] \\ \mathsf{E}[d_1(X, Y_1) \mid Z_0^3 = z_0^3] \\ \mathsf{E}[d_2(X, Y_2) \mid Z_0^3 = z_0^3] \end{bmatrix}$$

for different $z_0^3 \in \mathcal{Z}_0 \times \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathcal{Z}_3$. By Carathéodory's theorem, there exists a pmf p_Q with cardinality $|\mathcal{Q}| \le 7$ and $\tilde{z}_0^3(q)$ such that

$$H(U|Q,\ Z_0^3 = \tilde{z}_0^3(Q)) \le I(X; U) + \eta - 3,$$

and similarly for the other 6 inequalities. Take $\tilde{U} = g_{X\to U}(X, \tilde{z}_3(Q))$, $\tilde{Y}_1 = g_{X\to Y_1|U}(X, \tilde{z}_1(Q), \tilde{U})$, $\tilde{Y}_2 = g_{XY_1\to Y_2|U}(X, \tilde{Y}_1, \tilde{z}_2(Q), \tilde{U})$ and $\tilde{Y}_0 = g_{X\to Y_0|Y_1Y_2U}(X, \tilde{z}_0(Q), \tilde{Y}_1, \tilde{Y}_2, \tilde{U})$. Write $C_{p_Y}(y) \in \{0, 1\}^*$ for the Huffman codeword of y for the distribution p_Y. We set M_1 to be the concatenation of Q (3 bits), $C_{p_{\tilde{U}|Q}(\cdot|Q)}(\tilde{U})$ and $C_{p_{\tilde{Y}_1|\tilde{U}Q}(\cdot|\tilde{U},Q)}(\tilde{Y}_1)$, and M_2 to be the concatenation of Q, $C_{p_{\tilde{U}|Q}(\cdot|Q)}(\tilde{U})$, $C_{p_{\tilde{Y}_2|\tilde{U}Q}(\cdot|\tilde{U},Q)}(\tilde{Y}_2)$ and $C_{p_{\tilde{Y}_0|\tilde{Y}_1\tilde{Y}_2\tilde{U}Q}(\cdot|\tilde{Y}_1,\tilde{Y}_2,\tilde{U},Q)}(\tilde{Y}_0)$. The expected length of M_1 is upper bounded by

$$\begin{aligned} &3 + (I(X; U) + \eta - 3 + 1) + (I(X; Y_1|U) + \eta - 3 + 1) \\ &\qquad = I(X; Y_1|U) + I(X; U) + 2\eta - 1. \end{aligned}$$

Hence (4) is satisfied. By similar arguments, (5) and (6) hold.

Decoder 1 receives M_1 and recovers Q, and then recovers $\tilde{U}$ by decoding the Huffman code for the distribution $p_{\tilde{U}|Q}(\cdot|Q)$, and then recovers $\tilde{Y}_1$ similarly. Decoder 2 receives M_2 and recovers $Q, \tilde{U}$ and $\tilde{Y}_2$. Decoder 0 receives M_1, M_2 and recovers $Q, \tilde{U}, \tilde{Y}_1, \tilde{Y}_2$ and $\tilde{Y}_0$. ■

C. Lossy Gray–Wyner System

In this section, we use the SFRL to establish a one-shot inner bound for the lossy Gray–Wyner system [8], which yields an alternative proof of the achievability of the rate region in the asymptotic regime. The encoder observes $(X_1, X_2) \sim \mathsf{P}_{X_1,X_2}$ and produces three prefix-free descriptions $M_0, M_1, M_2 \in \{0, 1\}^*$. Decoder 1 observes M_0, M_1 and generates $\tilde{Y}_1$ with distortion $d_1(X_1, \tilde{Y}_1)$. Similarly, Decoder 2 observes M_0, M_2 and produces $\tilde{Y}_2$ with distortion $d_2(X_2, \tilde{Y}_2)$. An expected description length-distortion tuple $(\bar{R}_0, \bar{R}_1, \bar{R}_2, D_1, D_2)$ is said to be achievable if there exists a scheme with expected description length $\mathsf{E}[L(M_i)] \le \bar{R}_i$ and expected distortion $\mathsf{E}[d_i(X_i, \tilde{Y}_i)] \le D_i$.

Theorem 4: The tuple $(\bar{R}_0, \bar{R}_1, \bar{R}_2, D_1, D_2)$ is achievable if

$$\bar{R}_0 \ge I(X_1, X_2; U) + \log(I(X_1, X_2; U) + 1) + 8, \tag{7}$$

$$\bar{R}_1 \ge I(X_1; Y_1|U) + \log(I(X_1; Y_1|U) + 1) + 5, \tag{8}$$

$$\bar{R}_2 \ge I(X_2; Y_2|U) + \log(I(X_2; Y_2|U) + 1) + 5, \tag{9}$$

$$D_i \ge \mathsf{E}[d_i(X_i, Y_i)] \quad \text{for } i = 1, 2 \tag{10}$$

for some $\mathsf{P}_{U|X_1,X_2}$, $\mathsf{P}_{Y_1|X_1,U}$, $\mathsf{P}_{Y_2|X_2,U}$.

Note that the only difference between the above region and the lossy Gray–Wyner rate region [1, p. 357] is the addition of the logarithm terms, which grows like $\log n$ if we consider X_1^n, X_2^n and does not affect the asymptotic rate.

Proof: Applying the SFRL to $(X_1, X_2), U$, we have $U = g_{X_1X_2\to U}(X_1, X_2, Z_0)$, where $Z_0 \perp\!\!\!\perp (X_1, X_2)$ such that

$$H(U|Z_0) \le I(X_1, X_2; U) + \log(I(X_1, X_2; U) + 1) + 4.$$

Applying the SFRL to X_1, Y_1 conditioned on U (2), we have $Y_1 = g_{X_1\to Y_1|U}(X_1, Z_1, U)$, where $Z_1 \perp\!\!\!\perp (X_1, U)$ such that

$$H(Y_1|U, Z_1) \le I(X_1; Y_1|U) + \log(I(X_1; Y_1|U) + 1) + 4.$$

Applying the SFRL to X_2, Y_2 conditioned on U, we have $Y_2 = g_{X_2\to Y_2|U}(X_2, Z_2, U)$, where $Z_2 \perp\!\!\!\perp (X_2, U)$ such that

$$H(Y_2|U, Z_2) \le I(X_2; Y_2|U) + \log(I(X_2; Y_2|U) + 1) + 4.$$

Note that $Z_0^2 \perp\!\!\!\perp (X_1, X_2)$. Consider the convex hull of the 5-dimensional vectors

$$\begin{bmatrix} H(U|Z_0^2 = z_0^2) \\ H(Y_1|U, Z_0^2 = z_0^2) \\ H(Y_2|U, Z_0^2 = z_0^2) \\ \mathsf{E}[d_1(X_1, Y_1) \mid Z_0^2 = z_0^2] \\ \mathsf{E}[d_2(X_2, Y_2) \mid Z_0^2 = z_0^2] \end{bmatrix}$$

for different $z_0^2 \in \mathcal{Z}_0 \times \mathcal{Z}_1 \times \mathcal{Z}_2$. By Carathéodory's theorem, there exists a pmf p_Q with cardinality $|\mathcal{Q}| \le 5$ and $\tilde{z}_0^2(q)$ such that

$$H(U|Q,\, Z_0^2 = \tilde{z}_0^2(Q)) \le I(X_1, X_2; U) + \log(I(X_1, X_2; U) + 1) + 4,$$

and similarly for the other 4 inequalities. Take $\tilde{U} = g_{X_1X_2 \to U}(X_1, X_2, \tilde{z}_0(Q))$, $\tilde{Y}_1 = g_{X_1 \to Y_1|U}(X_1, \tilde{z}_1(Q), \tilde{U})$ and $\tilde{Y}_2 = g_{X_2 \to Y_2|U}(X_2, \tilde{z}_2(Q), \tilde{U})$. Write $C_{p_Y}(y) \in \{0,1\}^*$ for the Huffman codeword of y for the distribution p_Y. We set M_0 to be the concatenation of Q (3 bits) and $C_{p_{\tilde{U}|Q}(\cdot|Q)}(\tilde{U})$, $M_1 = C_{p_{\tilde{Y}_1|\tilde{U}Q}(\cdot|\tilde{U},Q)}(\tilde{Y}_1)$ and $M_2 = C_{p_{\tilde{Y}_2|\tilde{U}Q}(\cdot|\tilde{U},Q)}(\tilde{Y}_2)$. The expected length of M_0 is upper bounded by

$$\begin{aligned} &3 + (H(U|Z_0) + 1) \\ &\le 3 + (I(X_1, X_2; U) + \log(I(X_1, X_2; U) + 1) + 4 + 1) \\ &= I(X_1, X_2; U) + \log(I(X_1, X_2; U) + 1) + 8. \end{aligned}$$

Hence (7) is satisfied. By similar arguments, (8), (9) and (10) hold.

Decoder 1 receives M_0, M_1 and recovers Q, and then recovers $\tilde{U}$ by decoding the Huffman code for the distribution $p_{\tilde{U}|Q}(\cdot|Q)$, and then recovers $\tilde{Y}_1$ by decoding the Huffman code for the distribution $p_{\tilde{Y}_1|\tilde{U}Q}(\cdot|\tilde{U}, Q)$. Similar for Decoder 2. ∎

V. Achievability of Gelfand–Pinsker

In this section, we use the SFRL to prove the achievability part of the Gelfand-Pinsker theorem [9] for discrete memoryless channels with discrete memoryless state $p_S p_{Y|X,S}$, where the state is noncausally available at the encoder. The asymptotic capacity of this setting is

$$C_{\mathrm{GP}} = \max_{p_{U|S},\, x(u,s)} \left(I(U; Y) - I(U; S)\right).$$

We show the achievability of any rate below C_{GP} directly by using the SFRL to reduce the channel to a point-to-point memoryless channel. Fix $p_{U|S}$ and $x(u,s)$ that attain the capacity. Applying the SFRL to S, U, there exists a random variable $V \perp\!\!\!\perp S$ such that

$$H(U|V) \le I(U; S) + \log(I(U; S) + 1) + 4.$$

Note that

$$\begin{aligned} I(V; Y) &= I(U; Y) - I(U; Y|V) + I(V; Y|U) \\ &\ge I(U; Y) - H(U|V) \\ &\ge I(U; Y) - I(U; S) - \log(I(U; S) + 1) - 4. \end{aligned}$$

Hence we have constructed a memoryless point-to-point channel $p_{Y|V}$ with achievable rate close to $I(U; Y) - I(U; S)$.

For n channel uses, let $U^n|\{S^n = s^n\} \sim \prod_i p_{U|S}(u_i|s_i)$. The SFRL applied to S^n, U^n gives

$$I(V; Y^n) \ge nI(U; Y) - nI(U; S) - \log(nI(U; S) + 1) - 4.$$

Now we use the channel $p_{Y^n|V}$ l times (for a total block length of nl). By the channel coding theorem, we can communicate $l(nI(U; Y) - nI(U; S) - \log(nI(U; S) + 1) - 4) - o(l)$ bits with error probability that tends to 0 as $l \to \infty$. Letting $n \to \infty$ completes the proof.

In the above proof, we see that the SFRL can be used to convert a channel with state into a point-to-point channel by "orthogonalizing" the auxiliary input U and the state S. The point-to-point channel can be constructed explicitly via Poisson functional representation. This construction can be useful for designing codes for channels with state based on codes for point-to-point channels. It is interesting to note that this reduction makes the achievability proof for the Gelfand–Pinsker quite similar to that for the causal case in which the channel is reduced to a point-to-point channel using the "Shannon strategy" (see [1, p. 176]).

Note that Marton's inner bound for the broadcast channels with private messages [34] can also be proved using the SFRL in a similar manner. The idea is to "orthogonalize" the dependent auxiliary random variables U_1, U_2 by applying the SFRL on U_1, U_2 to produce two independent input random variables, and treat them with Y_1, Y_2 as an interference channel, and finally to treat interference as noise.

VI. Lower Bound and Properties of $I(X; Z|Y)$

Define the *excess functional information* as

$$\Psi(X \to Y) = \inf_{Z:\, Z \perp\!\!\!\perp X,\, H(Y|X,Z)=0} I(X; Z|Y).$$

An equivalent way to state SFRL is $\Psi(X \to Y) \le \log(I(X; Y)+1)+4$. In this section, we explore the properties of $\Psi(X \to Y)$. We first establish a lower bound.

Proposition 1: For discrete Y,

$$\Psi(X \to Y) \ge -\sum_{y \in \mathcal{Y}} \int_0^1 \Big(\mathsf{P}_X\left\{p_{Y|X}(y|X) \ge t\right\} \cdot \log\left(\mathsf{P}_X\left\{p_{Y|X}(y|X) \ge t\right\}\right)\Big) dt - I(X; Y).$$

Moreover for $|\mathcal{Y}| = 2$, equality holds in the above inequality, and the infimum in $\Psi(X \to Y)$ is attained via the Poisson functional representation.

Proof: Fix $Z \perp\!\!\!\perp X$ such that $Y = g(X, Z)$. For any y, let $V_y = p_{Y|Z}(y|Z)$, $U \sim \mathrm{Unif}[0,1]$, $\tilde{X}_y = p_{Y|X}(y|X)$, $\tilde{V}_y = \mathsf{P}\left\{\tilde{X}_y \ge U \mid U\right\}$, then $\mathsf{E}[V_y] = \mathsf{E}[\tilde{V}_y] = p_Y(y)$. We have

$$\begin{aligned} &\int_v^1 \mathsf{P}\{V_y \ge t\} dt \\ &= \mathsf{E}\left[\max\left\{V_y - v,\, 0\right\}\right] \\ &= \mathsf{E}_Z\left[\max\left\{p_{Y|Z}(y|Z) - v,\, 0\right\}\right] \\ &= \mathsf{E}_Z\left[\max\left\{\mathsf{P}_X\left\{g(X, Z) = y \mid Z\right\} - v,\, 0\right\}\right] \\ &= \mathsf{E}_Z\left[\max\left\{\mathsf{E}_{\tilde{X}_y}\left[\mathsf{P}_X\left\{g(X, Z) = y \mid Z, \tilde{X}_y\right\} \,\middle|\, Z\right] - v,\, 0\right\}\right] \\ &= \mathsf{E}_Z\Big[\max\left\{\mathsf{E}_{\tilde{X}_y}\left[\mathsf{P}_X\left\{g(X, Z) = y \mid Z, \tilde{X}_y\right\} \,\middle|\, Z\right]\right. \end{aligned}$$

$$
\begin{aligned}
&- \mathsf{E}_{\tilde{X}_y}\left[\mathbf{1}\left\{\tilde{X}_y > F^{-1}_{\tilde{X}_y}(1-\upsilon)\right\}\right], 0\Big\}\Big] \\
&\le \mathsf{E}_Z\Big[\mathsf{E}_{\tilde{X}_y}\Big[\max\Big\{\mathsf{P}_X\left\{g(X,Z)=y \mid Z, \tilde{X}_y\right\} \\
&\quad - \mathbf{1}\left\{\tilde{X}_y > F^{-1}_{\tilde{X}_y}(1-\upsilon)\right\}, 0\Big\} \,\Big|\, Z\Big]\Big] \\
&= \mathsf{E}_Z\Big[\mathsf{E}_{\tilde{X}_y}\Big[\mathsf{P}_X\left\{g(X,Z)=y \mid Z, \tilde{X}_y\right\} \\
&\quad \cdot \mathbf{1}\left\{\tilde{X}_y \le F^{-1}_{\tilde{X}_y}(1-\upsilon)\right\} \,\Big|\, Z\Big]\Big] \\
&= \mathsf{E}_{\tilde{X}_y}\Big[\mathsf{E}_Z\Big[\mathsf{P}_X\left\{g(X,Z)=y \mid Z, \tilde{X}_y\right\} \,\Big|\, \tilde{X}_y\Big] \\
&\quad \cdot \mathbf{1}\left\{\tilde{X}_y \le F^{-1}_{\tilde{X}_y}(1-\upsilon)\right\}\Big] \\
&= \mathsf{E}_{\tilde{X}_y}\Big[\mathsf{E}_X\Big[\mathsf{P}_Z\left\{g(X,Z)=y \mid X\right\} \,\Big|\, \tilde{X}_y\Big] \\
&\quad \cdot \mathbf{1}\left\{\tilde{X}_y \le F^{-1}_{\tilde{X}_y}(1-\upsilon)\right\}\Big] \\
&= \mathsf{E}_{\tilde{X}_y}\left[\mathsf{E}_X\left[p_{Y|X}(y|X) \mid \tilde{X}_y\right]\mathbf{1}\left\{\tilde{X}_y \le F^{-1}_{\tilde{X}_y}(1-\upsilon)\right\}\right] \\
&= \mathsf{E}_{\tilde{X}_y}\left[\tilde{X}_y \cdot \mathbf{1}\left\{\tilde{X}_y \le F^{-1}_{\tilde{X}_y}(1-\upsilon)\right\}\right] \\
&= \mathsf{E}_U\left[\max\left\{\mathsf{P}\left\{\tilde{X}_y \ge U \mid U\right\} - \upsilon,\, 0\right\}\right] \\
&= \mathsf{E}\left[\max\left\{\tilde{V}_y - \upsilon,\, 0\right\}\right] \\
&= \int_\upsilon^1 \mathsf{P}\{\tilde{V}_y \ge t\}dt.
\end{aligned}
$$

Hence V_y dominates $\tilde{V}_y$ stochastically in the second order. By the concavity of $-t\log t$, we have

$$
\begin{aligned}
H(Y|Z) &= -\sum_y \mathsf{E}_Z\left[p_{Y|Z}(y|Z)\log p_{Y|Z}(y|Z)\right] \\
&= -\sum_y \mathsf{E}\left[V_y \log V_y\right] \\
&\ge -\sum_y \mathsf{E}\left[\tilde{V}_y \log \tilde{V}_y\right] \\
&= -\sum_y \int_0^1 \Big(\mathsf{P}_X\left\{p_{Y|X}(y|X) \ge u\right\} \\
&\qquad \cdot \log\left(\mathsf{P}_X\left\{p_{Y|X}(y|X) \ge u\right\}\right)\Big)du. \qquad (11)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
I(X;Z|Y) \ge -\sum_y \int_0^1 \Big(&\mathsf{P}_X\left\{p_{Y|X}(y|X) \ge t\right\} \\
&\cdot \log\left(\mathsf{P}_X\left\{p_{Y|X}(y|X) \ge t\right\}\right)\Big)dt - I(X;Y).
\end{aligned}
$$

One can verify that for $|\mathcal{Y}| = 2$, equality in (11) holds by the definition of Poisson functional representation. ∎

The following proposition shows that there exists a sequence of (X, Y) for which the bound $\Psi(X,Y) \le \log(I(X;Y)+1)+4$ given in the SFRL is tight within 5 bits. An example where the log term is tight is also given in [10], though the additive constant is not specified there.

Proposition 2: For every $\alpha \ge 0$, there exists discrete X, Y such that $I(X;Y) \ge \alpha$ and

$$\Psi(X \to Y) \ge \log(I(X;Y)+1) - 1.$$

The proof is given in Appendix C. Besides the upper bound given by the SFRL and its tightness, in the following we establish other properties of $\Psi(X \to Y)$. We write the conditional excess functional information as

$$\Psi(X \to Y \mid Q) = \mathsf{E}_Q\left[\Psi(X \to Y \mid Q = q)\right].$$

Proposition 3: The excess functional information $\Psi(X \to Y)$ satisfies the following properties.

1) Alternative characterization.
$$\Psi(X \to Y) = \inf_{Z:\, Z \perp\!\!\!\perp X} H(Y|Z) - I(X;Y).$$
2) Monotonicity. If $X_1 \perp\!\!\!\perp X_2$ and $X_1 \perp\!\!\!\perp (X_2, Y_2) \mid Y_1$, then
$$\Psi((X_1,X_2) \to (Y_1,Y_2)) \ge \Psi(X_1 \to Y_1).$$
3) Subadditivity. If $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2)$, then
$$\Psi((X_1,X_2) \to (Y_1,Y_2)) \le \Psi(X_1 \to Y_1) + \Psi(X_2 \to Y_2).$$
As a result, if we further have $X_2 \perp\!\!\!\perp Y_2$, then $\Psi((X_1,X_2) \to (Y_1,Y_2)) = \Psi(X_1 \to Y_1)$ by monotonicity.
4) Data processing of $\Psi + I$. If $X_2 - X_1 - Y_1 - Y_2$ forms a Markov chain,
$$\Psi(X_1 \to Y_1) + I(X_1;Y_1) \ge \Psi(X_2 \to Y_2) + I(X_2;Y_2).$$
5) Upper bound by common entropy.
$$\begin{aligned}\Psi(X \to Y) &\le G(X;Y) - I(X;Y) \\ &\le \min\{H(X|Y),\, H(Y|X)\},\end{aligned}$$
where $G(X;Y) = \min_{X \perp\!\!\!\perp Y|W} H(W)$ is the common entropy [35], [36].
6) Conditioning. If Q satisfies $H(Q|X) = 0$, then
$$\Psi(X \to Y) \ge \Psi(X \to Y \mid Q).$$
If we further have $H(Q|Y) = 0$, then equality holds in the above inequality.
7) Successive minimization.
$$\Psi(X \to Y) = \inf_{V:\, V \perp\!\!\!\perp X}\left\{I(X;V|Y) + \Psi(X \to Y \mid V)\right\}.$$

Proof:

1) *Alternative characterization*. Note that if $Z \perp\!\!\!\perp X$ and $H(Y|X,Z) = 0$, then $H(Y|Z) - I(X;Y) = I(X;Z|Y)$, hence
$$\begin{aligned}\inf_{Z:\, Z \perp\!\!\!\perp X,\, H(Y|X,Z)=0} I(X;Z|Y) \\ \ge \inf_{Z:\, Z \perp\!\!\!\perp X} H(Y|Z) - I(X;Y).\end{aligned}$$

For the other direction, assume $Z \perp\!\!\!\perp X$. By the functional representation lemma, let $Y = g(X, Z, \tilde{Z})$, $\tilde{Z} \perp\!\!\!\perp (X, Z)$. We have

$$\begin{aligned} H(Y|Z) - I(X;Y) &\geq H(Y|Z,\tilde{Z}) - I(X;Y) \\ &= I(X; Z, \tilde{Z}|Y) \\ &\geq \inf_{Z' : Z' \perp\!\!\!\perp X,\, H(Y|X,Z')=0} I(X; Z'|Y). \end{aligned}$$

2) *Monotonicity*. Let Z satisfies $Z \perp\!\!\!\perp (X_1, X_2)$ and $H(Y_1, Y_2|X_1, X_2, Z) = 0$. Note that $(Z, X_2) \perp\!\!\!\perp X_1$ and $H(Y_1|X_1, Z, X_2) = 0$. Hence

$$\begin{aligned} &I(X_1, X_2; Z|Y_1, Y_2) \\ &\quad \geq I(X_1; Z|X_2, Y_1, Y_2) \\ &\quad = I(X_1; Z|X_2, Y_1, Y_2) + I(X_1; Y_2|X_2, Y_1) \\ &\quad = I(X_1; Z|X_2, Y_1) + I(X_1; Y_2|X_2, Y_1, Z) \\ &\quad \geq I(X_1; Z|X_2, Y_1) \\ &\quad = I(X_1; Z|X_2, Y_1) + I(X_1; X_2|Y_1) \\ &\quad = I(X_1; Z, X_2|Y_1) \\ &\quad \geq \Psi(X_1 \to Y_1). \end{aligned}$$

3) *Subadditivity*. let Z_1, Z_2 satisfies $Z_i \perp\!\!\!\perp X_i$ and $H(Y_i|X_i, Z_i) = 0$, then

$$\begin{aligned} &\Psi((X_1, X_2) \to (Y_1, Y_2)) \\ &\quad \leq I(X_1, X_2;\, Z_1, Z_2 \,|\, Y_1, Y_2) \\ &\quad = I(X_1; Z_1|Y_1) + I(X_2; Z_2|Y_2). \end{aligned}$$

4) *Data processing of* $\Psi + I$. let $Z \perp\!\!\!\perp X_1$, and let $Y_2 = g(Y_1, W)$ be the functional representation of Y_2. Then $(Z, W) \perp\!\!\!\perp X_2$, and by the the alternative characterization,

$$\begin{aligned} &\Psi(X_2 \to Y_2) + I(X_2; Y_2) \\ &\quad \leq H(Y_2|Z, W) \\ &\quad = H(Y_2|Z, W, Y_1) + I(Y_1; Y_2|Z, W) \\ &\quad \leq H(Y_1|Z, W) \\ &\quad = H(Y_1|Z). \end{aligned}$$

5) The upper bound by common entropy is a direct consequence of the data processing inequality in the previous part.

6) *Conditioning*. Assume that $H(Q|X) = 0$, $Z \perp\!\!\!\perp X$ and $H(Y|X, Z) = 0$, then $Z \perp\!\!\!\perp X|\{Q = q\}$and $H(Y|X, Z, Q = q) = 0$ for all q, hence

$$\begin{aligned} I(X; Z|Y) &\geq I(X; Z|Y, Q) \\ &= \mathsf{E}_{q\sim \mathsf{P}_Q}\left[I(X; Z|Y, Q = q)\right] \\ &\geq \mathsf{E}_{q\sim \mathsf{P}_Q}\left[\Psi(X \to Y \mid Q = q)\right]. \end{aligned}$$

To show the equality case, assume $H(Q|Y) = 0$. Let $\tilde{Z}$ satisfies $\tilde{Z} \perp\!\!\!\perp X|\{Q = q\}$ and $H(Y|X, \tilde{Z}, Q = q) = 0$ for all q. By functional representation lemma, let $\tilde{Z} = g(Q, Z)$, $Z \perp\!\!\!\perp Q$, and since we are invoking functional representation lemma over the marginal distribution of $(Q, \tilde{Z})$, we can assume $Z \perp\!\!\!\perp (X, Y)|(Q, \tilde{Z})$. Hence $Z \perp\!\!\!\perp X$. We have

$$\begin{aligned} \mathsf{E}_{q\sim \mathsf{P}_Q}\left[I(X; \tilde{Z}|Y, Q = q)\right] &= I(X; \tilde{Z}|Y, Q) \\ &= I(X; Z|Y, Q) \\ &= I(X; Z|Y) \\ &\geq \Psi(X \to Y). \end{aligned}$$

7) *Successive minimization*. Assume that $V \perp\!\!\!\perp X$, and let $\tilde{Z}$ satisfy $\tilde{Z} \perp\!\!\!\perp X|\{V = v\}$ and $H(Y|X, \tilde{Z}, V = v) = 0$ for all v, then $X \perp\!\!\!\perp (\tilde{Z}, V)$. We have

$$\begin{aligned} &\mathsf{E}_{q\sim \mathsf{P}_Q}\left[I(X; \tilde{Z}|Y, V = v)\right] \\ &\quad = I(X; \tilde{Z}|Y, V) \\ &\quad = I(X; \tilde{Z}, V|Y) - I(X; V|Y) \\ &\quad = I(X; \tilde{Z}, V|Y) - I(X; V|Y) \\ &\quad \geq \Psi(X \to Y) - I(X; V|Y). \end{aligned}$$

Note that $I(X; V|Y) + \Psi(X \to Y \mid V) = \Psi(X \to Y)$ if $V = \emptyset$. Also note that

$$\begin{aligned} &\inf_{V:\, V \perp\!\!\!\perp X} \{I(X; V|Y) + \Psi(X \to Y \mid V)\} \\ &\quad \leq \inf_{V:\, V \perp\!\!\!\perp X,\, H(Y|X,Z)=0} \{I(X; V|Y) + \Psi(X \to Y \mid V)\} \\ &\quad = \inf_{V:\, V \perp\!\!\!\perp X,\, H(Y|X,Z)=0} I(X; V|Y) \\ &\quad = \Psi(X \to Y). \end{aligned}$$

■

Remark 5: If $\Psi(X, Y) = 0$, then it means that there exists Z such that $Z \perp\!\!\!\perp X$, $Z \perp\!\!\!\perp X|Y$, $H(Y|Z) = I(X; Y)$ and $H(Y|X, Z) = 0$. This implies there exists z such that $H(Y|Z = z) \geq I(X; Y)$ and $H(Y|X, Z = z) = 0$. Hence it is possible to perform one-shot zero error channel coding on the channel $\mathsf{P}_{X|Y}$ with input distribution $\mathsf{P}_{Y|Z=z}$ to communicate a message with entropy $\geq I(X; Y)$.

APPENDIX

A. Proof of Theorem 1

Condition on the event $\{X = x\}$ where $\mathsf{P}_{Y|X}(\cdot|x) \ll \mathsf{P}_Y$ (which is true for P_X-almost all x's since $I(X; Y) < \infty$). First we show that $g_{X\to Y}(x, \{(\tilde{Y}_i, T_i)\})$ follows the distribution $\mathsf{P}_{Y|X}(\cdot|x)$. By the marking theorem of the Poisson point process [37], [38], $\{(\tilde{Y}_i, T_i)\}$ is a non-homogeneous Poisson point process with intensity measure $\mathsf{P}_Y \times \mu$ (where μ is the Lebesgue measure on $[0, \infty)$). Applying the mapping theorem [37], [38] for the mapping $(y, t) \mapsto (y,\, t \cdot (d\mathsf{P}_Y/d\,\mathsf{P}_{Y|X}(\cdot|x))(y))$ over the set $\{(y, t) : (d\,\mathsf{P}_{Y|X}(\cdot|x)/d\mathsf{P}_Y)(y) > 0\}$ (note that the mapping is measurable since $d\,\mathsf{P}_{Y|X}(\cdot|x)/d\mathsf{P}_Y$ is measurable),

$$\left\{\left(\tilde{Y}_i\,,\; T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{Y}_i)\right)\right\}$$

is a Poisson point process with intensity measure $\mathsf{P}_{Y|X}(\cdot|x) \times \mu$, since the number of occurrences of this process in the set

A follows the Poisson distribution with rate

$$\begin{aligned}
&(\mathsf{P}_Y \times \mu)(\{(y,t) : (y,\, t \cdot (d\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x))(y)) \in A\}) \\
&\quad = \int \mu(\{t : (y,\, t \cdot (d\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x))(y)) \in A\}) d\mathsf{P}_Y \\
&\quad = \int \Big(\mu(\{t : (y,\, t \cdot (d\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x))(y)) \in A\}) \\
&\qquad \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}\Big) d\mathsf{P}_{Y|X}(\cdot|x) \\
&\quad = \int \mu(\{t : (y,t) \in A\}) d\mathsf{P}_{Y|X}(\cdot|x) \\
&\quad = (\mathsf{P}_{Y|X}(\cdot|x) \times \mu)(A).
\end{aligned}$$

Hence if we consider $\{T_i \cdot (d\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x))(\tilde{Y}_i)\}$ alone, it is a Poisson point process with intensity measure μ. Therefore the first occurrence has a distribution

$$\min_i \Big(T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{Y}_i)\Big) \sim \mathrm{Exp}(1).$$

Now, let

$$\begin{aligned}
\Theta &= \min_i \Big(T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{Y}_i)\Big), \\
K &= \arg\min_i \Big(T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|X)}(\tilde{Y}_i)\Big).
\end{aligned}$$

Letting $Y = \tilde{Y}_K$, then as desired we have

$$Y|\{X = x\} \sim \mathsf{P}_{Y|X}(\cdot|x).$$

Since Y is a function of $\{(\tilde{Y}_i, T_i)\})$ and K, $H(Y \mid \{(\tilde{Y}_i, T_i)\}) \le H(K)$. Conditioned on $\Theta = \theta$, we have $\tilde{Y}_K \sim \mathsf{P}_{Y|X}(\cdot|x)$ since we can consider $\{\tilde{Y}_i,\, T_i \cdot (d\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x))(\tilde{Y}_i)\}$ as a marked Poisson point process with i.i.d. marks distributed as $\mathsf{P}_{Y|X}(\cdot|x)$, and $\{\tilde{Y}_i,\, T_i \cdot (d\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x))(\tilde{Y}_i)\}_{i \neq K}$ is a Poisson point process with intensity measure $\mathsf{P}_{Y|X}(\cdot|x) \times \mu|_{[\theta,\infty)}$ (where $\mu|_{[\theta,\infty)}$ is the restriction of μ to $[\theta,\infty)$). By mapping theorem, $\{(\tilde{Y}_i, T_i)\}_{i \neq K}$ is a Poisson point process with intensity measure

$$\nu(A \times B) = \int_A \mu\left(B \cap \left[\theta \cdot \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(y)\,,\, \infty\right)\right) d\mathsf{P}_Y(y).$$

Note that $K - 1 = |\{i : T_i < T_K\}|$. Hence $K - 1$ conditioned on $\Theta = \theta$ and $\tilde{Y}_K = \tilde{y}$ follows the Poisson distribution with rate

$$\begin{aligned}
&\nu\,(\mathcal{Y} \times [0,\, T_K)) \\
&= \nu\left(\mathcal{Y} \times \left[0,\, \theta \cdot \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(\tilde{y})\right)\right) \\
&= \int_{\mathcal{Y}} \mu\Bigg(\left[0,\, \theta \cdot \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(\tilde{y})\right) \\
&\quad \cap \left[\theta \cdot \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(y)\,,\, \infty\right)\Bigg) d\mathsf{P}_Y(y) \\
&= \theta \int_{\mathcal{Y}} \max\left\{0,\, \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(\tilde{y}) - \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(y)\right\} d\mathsf{P}_Y(y) \\
&\le \theta \int_{\mathcal{Y}} \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(\tilde{y}) \cdot d\mathsf{P}_Y(y) \\
&= \theta \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(\tilde{y}).
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathsf{E}[\log K | X = x] \\
&= \mathsf{E}_{Y \sim \mathsf{P}_{Y|X}(\cdot|x)}\left[\int_0^\infty e^{-\theta} \mathsf{E}\left[\log K \mid \Theta = \theta,\, \tilde{Y}_K = Y\right] d\theta\right] \\
&\le \mathsf{E}_{Y \sim \mathsf{P}_{Y|X}(\cdot|x)}\left[\int_0^\infty e^{-\theta} \log\left(\theta \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(Y) + 1\right) d\theta\right] \\
&\le \mathsf{E}_{Y \sim \mathsf{P}_{Y|X}(\cdot|x)}\left[\log\left(\int_0^\infty e^{-\theta} \theta \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(Y) d\theta + 1\right)\right] \\
&= \mathsf{E}_{Y \sim \mathsf{P}_{Y|X}(\cdot|x)}\left[\log\left(\frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(Y) + 1\right)\right] \\
&\le \mathsf{E}_{Y \sim \mathsf{P}_{Y|X}(\cdot|x)}\left[\max\left\{\log \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(Y)\,, 0\right\} + 1\right] \\
&= D(\mathsf{P}_{Y|X}(\cdot|x) \,\|\, \mathsf{P}_Y) \\
&\quad - \mathsf{E}_{Y \sim \mathsf{P}_{Y|X}(\cdot|x)}\left[\min\left\{\log \frac{d\,\mathsf{P}_{Y|X}(\cdot|x)}{d\,\mathsf{P}_Y}(Y)\,, 0\right\}\right] + 1 \\
&\le D(\mathsf{P}_{Y|X}(\cdot|x) \,\|\, \mathsf{P}_Y) + e^{-1} \log e + 1,
\end{aligned}$$

where the last line follows by the same arguments as in [4, Appendix A]. For $X \sim \mathsf{P}_X$,

$$\mathsf{E}\left[\log K\right] \le I(X;Y) + e^{-1} \log e + 1.$$

By the maximum entropy distribution subject to a given $\mathsf{E}\left[\log K\right]$ (see Appendix B), we have

$$\begin{aligned}
&H(K) \\
&\le I(X;Y) + e^{-1} \log e + 2 + \log\left(I(X;Y) + e^{-1} \log e + 2\right) \\
&\le I(X;Y) + \log\left(I(X;Y) + 1\right) + e^{-1} \log e + 2 \\
&\quad + \log\left(e^{-1} \log e + 2\right) \\
&< I(X;Y) + \log\left(I(X;Y) + 1\right) + 4.
\end{aligned}$$

To prove the cardinality bound, first note that if $|\mathcal{X}|$, $|\mathcal{Y}|$ are finite, then $|\mathcal{Z}| \le |\mathcal{Y}|^{|\mathcal{X}|}$ can be assumed to be finite since it is the number of different functions $x \mapsto g_{X \to Y}(x,z)$ for different z. To further reduce the cardinality, we apply Carathéodory's theorem on the $(|\mathcal{X}|(|\mathcal{Y}| - 1) + 1)$-dimensional vectors with entries $H(Y|Z = z)$ and $p(x, y|z)$ for $x \in \{1, \ldots, |\mathcal{X}|\}$, $y \in \{1, \ldots, |\mathcal{Y}| - 1\}$; see [39], [40]. The cardinality bound can be proved using Fenchel-Eggleston-Carathéodory theorem [41], [42].

B. Proof of the Bound on Entropy in Theorem 1

The proof of the following proposition follows from the standard argument in maximum entropy distribution. It is well-known that Zipf distribution maximizes the entropy for a fixed $\mathsf{E}\left[\log \Theta\right]$, see [43]. A similar lemma (with an unspecified constant) is also used in [10]. It is included here for the sake of completeness.

Proposition 4: Let $\Theta \in \{1, 2, \ldots\}$ be a random variable, then

$$H(\Theta) \le \mathsf{E}\left[\log \Theta\right] + \log\left(\mathsf{E}\left[\log \Theta\right] + 1\right) + 1.$$

Proof: Let $q(\theta) = c\theta^{-\lambda}$ where $\lambda = 1 + 1/\mathsf{E}\left[\log \Theta\right]$, and $c > 0$ such that $\sum_{\theta=1}^\infty q(\theta) = 1$. Note that

$$\sum_{\theta=1}^\infty \theta^{-\lambda} \le 1 + \int_1^\infty \theta^{-\lambda} d\theta = 1 + \frac{1}{\lambda - 1}.$$

Therefore

$$\begin{aligned} H(\Theta) &\le \sum_{\theta=1}^{\infty} p_\Theta(\theta) \log \frac{1}{q(\theta)} \\ &= \sum_{\theta=1}^{\infty} p_\Theta(\theta) \left(\lambda \log \theta - \log c\right) \\ &= \lambda \mathsf{E}\left[\log \Theta\right] + \log\left(\sum_{\theta=1}^{\infty} \theta^{-\lambda}\right) \\ &\le \lambda \mathsf{E}\left[\log \Theta\right] + \log\left(1 + \frac{1}{\lambda - 1}\right) \\ &= \mathsf{E}\left[\log \Theta\right] + \log\left(\mathsf{E}\left[\log \Theta\right] + 1\right) + 1. \end{aligned}$$

Operationally, we would use the optimal prefix-free code for the Zipf distribution $q(\theta)$ to encode Θ.

■

C. Proof of Proposition 2

Let $k \in \{0, 1, \ldots\}$, $V \in [0 : 2^k - 1]$,

$$p_V(v) = \gamma^{-1} 2^{k - \lceil \log(v+1) \rceil},$$

where $\gamma = 2^{k-1}(k+2)$, and let $X \sim \text{Unif}[0 : 2^k - 1]$ independent of V, and $Y = (X + V) \bmod 2^k$. Note that $|\{v : \gamma\, p_V(v) > t\}| = \gamma\, p_V(\lfloor t \rfloor)$ for $t \ge 0$. We have

$$\begin{aligned} &- \sum_{y \in \mathcal{Y}} \int_0^1 \mathsf{P}_X\left\{p_{Y|X}(y|X) \ge t\right\} \log\left(\mathsf{P}_X\left\{p_{Y|X}(y|X) \ge t\right\}\right) dt \\ &= - \sum_{y \in \mathcal{Y}} \int_0^1 2^{-k} \left|\{v : p_V(v) \ge t\}\right| \log\left(2^{-k} \left|\{v : p_V(v) \ge t\}\right|\right) dt \\ &= k - \int_0^1 \left|\{v : p_V(v) \ge t\}\right| \log \left|\{v : p_V(v) \ge t\}\right| dt \\ &= k - \int_0^1 \gamma\, p_V\left(\lfloor \gamma t \rfloor\right) \log\left(\gamma\, p_V\left(\lfloor \gamma t \rfloor\right)\right) dt \\ &= k - \sum_{v=0}^{2^k - 1} p_V(v) \log\left(\gamma\, p_V(v)\right) dt \\ &= k - \log \gamma + H(V). \end{aligned}$$

And

$$I(X; Y) = H(Y) - H(Y|X) = k - H(V).$$

By Proposition 1,

$$\begin{aligned} \Psi(X \to Y) &\ge k - \log \gamma + H(V) - (k - H(V)) \\ &= 2H(V) - \log \gamma. \end{aligned}$$

One can check that

$$H(V) = \frac{1}{2}k + \log(k+2) - \frac{3}{2} + \frac{1}{k+2}.$$

Hence

$$I(X; Y) = \frac{1}{2}k - \log(k+2) + \frac{3}{2} - \frac{1}{k+2} \le \frac{1}{2}k,$$

and

$$\begin{aligned} &\Psi(X \to Y) \\ &\quad \ge k + 2\log(k+2) - 3 + \frac{2}{k+2} - \log\left(2^{k-1}(k+2)\right) \\ &\quad = \log(k+2) - 2 + \frac{2}{k+2} \\ &\quad \ge \log(I(X; Y) + 1) - 1. \end{aligned}$$

ACKNOWLEDGMENTS

This work was partially supported by a gift from Huawei Technologies. The authors would like to thank the anonymous reviewers for their insightful remarks and for pointing out to us the Braverman-Garg paper. Their comments have helped improve the presentation of the results and their connections to previous work.

REFERENCES

[1] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[2] B. Hajek and M. B. Pursley, "Evaluation of an achievable rate region for the broadcast channel," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 1, pp. 36–46, Jan. 1979.
[3] F. M. J. Willems and E. van der Meulen, "The discrete memoryless multiple-access channel with cribbing encoders," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 313–327, May 1985.
[4] P. Harsha, R. Jain, D. McAllester, and J. Radhakrishnan, "The communication complexity of correlation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 438–449, Jan. 2010.
[5] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Int. Conv. Rec.*, vol. 7, no. 1, pp. 142–163, Mar. 1959.
[6] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 6, pp. 851–857, Nov. 1982.
[7] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 4, pp. 502–521, Jul. 1987.
[8] R. M. Gray and A. D. Wyner, "Source coding for a simple network," *Bell Syst. Tech. J.*, vol. 53, no. 9, pp. 1681–1721, Nov. 1974.
[9] S. Gel'fand and M. Pinsker, "Coding for channel with random parameters," *Problems Control Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.
[10] M. Braverman and A. Garg, "Public vs private coin in bounded-round information," in *Automata, Languages, and Programming—ICALP* (Lecture Notes in Computer Science), vol. 8572, J. Esparza, P. Fraigniaud, T. Husfeldt, and E. Koutsoupias, Eds. Berlin, Germany: Springer, 2014.
[11] J. Liu, P. Cuff, and S. Verdú, "Resolvability in E_γ with applications to lossy compression and wiretap channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 755–759.
[12] N. Datta, J. M. Renes, R. Renner, and M. M. Wilde, "One-shot lossy quantum data compression," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8057–8076, Dec. 2013.
[13] S. Verdú, "Non-asymptotic achievability bounds in multiuser information theory," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2012, pp. 1–8.
[14] J. Liu, P. Cuff, and S. Verdú, "One-shot mutual covering lemma and Marton's inner bound with a common message," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 1457–1461.
[15] S. Watanabe, S. Kuzuoka, and V. Y. F. Tan, "Non-asymptotic and second-order achievability bounds for source coding with side-information," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 3055–3059.
[16] M. H. Yassaee, M. R. Aref, and A. Gohari, "A technique for deriving one-shot achievability results in network information theory," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 1287–1291.
[17] J. T. Pinkston, "Encoding independent sample information sources," Res. Lab. Electron., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. 462, 1967.
[18] M. Pursley and L. Davisson, "Variable rate coding for nonergodic sources and classes of ergodic sources subject to a fidelity constraint," *IEEE Trans. Inf. Theory*, vol. 22, no. 3, pp. 324–337, May 1976.
[19] K. Mackenthun and M. Pursley, "Variable-rate universal block source coding subject to a fidelity constraint," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 3, pp. 349–360, May 1978.

[20] O. Kosut and L. Sankar, "Universal fixed-to-variable source coding in the finite blocklength regime," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 649–653.
[21] V. Kostina, Y. Polyanskiy, and S. Verdú, "Variable-length compression allowing errors," *IEEE Trans. Inf. Theory*, vol. 61, no. 8, pp. 4316–4330, Aug. 2015.
[22] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem," *IEEE Trans. Inf. Theory*, vol. 48, no. 10, pp. 2637–2655, Oct. 2002.
[23] P. Cuff, "Distributed channel synthesis," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7071–7096, Nov. 2013.
[24] C. H. Bennett, I. Devetak, A. W. Harrow, P. W. Shor, and A. Winter, "The quantum reverse Shannon theorem and resource tradeoffs for simulating quantum channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 2926–2959, May 2014.
[25] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.
[26] R. Jain, J. Radhakrishnan, and P. Sen, "A direct sum theorem in communication complexity via message compression," in *Automata, Languages and Programming—ICALP* (Lecture Notes in Computer Science), vol. 2719, J. C. M. Baeten, J. K. Lenstra, J. Parrow, and G. J. Woeginger, Eds. Berlin, Germany: Springer, 2003.
[27] M. Braverman and A. Rao, "Information equals amortized communication," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6058–6069, Oct. 2014.
[28] E. C. Song, P. Cuff, and H. V. Poor, "The likelihood encoder for lossy compression," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1836–1849, Apr. 2016.
[29] E. C. Posner and E. R. Rodemich, "Epsilon entropy and data compression," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 2079–2125, 1971.
[30] Z. Zhang, E.-H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion. 1. Known statistics," *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 71–91, Jan. 1997.
[31] D. S. Ornstein and P. C. Shields, "Universal almost sure data compression," *Ann. Probability*, vol. 18, no. 2, pp. 441–452, 1990.
[32] R. Venkataramani, G. Kramer, and V. K. Goyal, "Multiple description coding with many channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 9, pp. 2106–2114, Sep. 2003.
[33] J. Wang, J. Chen, L. Zhao, P. Cuff, and H. Permuter, "On the role of the refinement layer in multiple description coding and scalable coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1443–1456, Mar. 2011.
[34] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 3, pp. 306–311, May 1979.
[35] G. R. Kumar, C. T. Li, and A. A. El Gamal, "Exact common information," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 161–165.
[36] C. T. Li and A. A. El Gamal, "Distributed simulation of continuous random variables," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2016, pp. 565–569.
[37] J. F. C. Kingman, *Poisson Processes*. London, U.K.: Oxford Univ. Press, 1993.
[38] G. Last and M. Penrose, *Lectures on the Poisson Process*, vol. 7. Cambridge, U.K.: Cambridge Univ. Press, 2017.
[39] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 6, pp. 629–637, Nov. 1975.
[40] A. D. Wyner and J. Ziv, "The rate–distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.
[41] H. G. Eggleston, *Convexity*. Cambridge, U.K.: Cambridge Univ. Press, 1958.
[42] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.
[43] M. Visser, "Zipf's law, power laws and maximum entropy," *New J. Phys.*, vol. 15, no. 4, p. 043021, 2013.

Cheuk Ting Li (S'12) received the B.Sc. degree in mathematics and B.Eng. degree in information engineering from The Chinese University of Hong Kong in 2012, and the M.S. and Ph.D. degree in electrical engineering from Stanford University in 2014 and 2018 respectively. He is currently a postdoctoral scholar at the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. His research interests include generation of random variables, one-shot schemes in information theory, wireless communications and information-theoretic secrecy.

Abbas El Gamal (S'71–M'73–SM'83–F'00–LF'16) is the Hitachi America Professor in the School of Engineering at Stanford University. He received his B.Sc. Honors degree from Cairo University in 1972, and his M.S. in Statistics and Ph.D. in Electrical Engineering both from Stanford University in 1977 and 1978, respectively. From 1978 to 1980, he was an Assistant Professor of Electrical Engineering at USC. From 2003 to 2012, he was the Director of the Information Systems Laboratory at Stanford University. From 2012-2017, he was the Fortinet Founders Chair of the Department of Electrical Engineering. His research contributions have been in network information theory, FPGAs, and digital imaging devices and systems. He has authored or coauthored over 230 papers and holds 35 patents in these areas. He is coauthor of the book *Network Information Theory* (Cambridge Press 2011). He is a member of the US National Academy of Engineering and a Life Fellow of the IEEE. He received several honors and awards for his research contributions, including the 2016 IEEE Richard Hamming Medal, the 2014 Viterbi Lecture, the 2013 Shannon Memorial Lecture, the 2012 Claude E. Shannon Award, the inaugural Padovani Lecture, and the 2004 INFOCOM Paper Award. He served on the Board of Governors of the Information Theory Society from 2009 to 2016 and was President in 2014.