Information-Distilling Quantizers

Alankrita Bhatt, Bobak Nazer, Or Ordentlich and Yury Polyanskiy

Abstract

Let X and Y be dependent random variables. This paper considers the problem of designing a scalar quantizer for Y to maximize the mutual information between the quantizer's output and X, and develops fundamental properties and bounds for this form of quantization, which is connected to the log-loss distortion criterion. The main focus is the regime of low $I(X;Y)$, where it is shown that, if X is binary, a constant fraction of the mutual information can always be preserved using $\mathcal{O}(\log(1/I(X;Y)))$ quantization levels, and there exist distributions for which this many quantization levels are necessary. Furthermore, for larger finite alphabets $2 < |\mathcal{X}| < \infty$, it is established that an η-fraction of the mutual information can be preserved using roughly $(\log(|\mathcal{X}|/I(X;Y)))^{\eta\cdot(|\mathcal{X}|-1)}$ quantization levels.

I. Introduction

Let X and Y be a pair of random variables with alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, and a given distribution P_{XY}. This paper deals with the problem of quantizing Y into $M < |\mathcal{Y}|$ values, under the objective of maximizing the mutual information between the quantizer's output and X. With a slight abuse of notation[1], we will denote the value of the mutual information attained by the optimal M-ary quantizer by

$$I(X;[Y]_M) \triangleq \sup_{\tilde{Y}\in[Y]_M} I(X;\tilde{Y}). \tag{1}$$

where $[Y]_M$ is the set of all (deterministic) M-ary quantizations of Y,

$$[Y]_M \triangleq \{f(Y) \ : \ f:\mathcal{Y}\to[M]\}$$

and $[M] \triangleq \{1,2,\ldots,M\}$.

When X and Y are thought of as the input and output of a channel, this problem corresponds to determining the highest available information rate for M-level quantization. It is therefore not surprising that this problem has received considerable attention. For example, it is well known [2, Section 2.11] that when X is ±1 equiprobable and $Y = X + Z$ for Gaussian Z, it holds that $I(X;[Y]_2) \geq \frac{2}{\pi} I(X;Y)$, which is achieved by taking $f(\cdot)$ to be the maximum a posteriori (MAP) estimator of X from Y.[2]

A characterization of (1) is also required for the construction of good polar codes [4], since the large output cardinality of polarized channels makes it challenging to evaluate their respective capacities (and identify "frozen" bits). Efficient techniques for channel output quantization that preserve mutual information have been developed to overcome this obstacle, and played a major role in the process of making polar codes implementable [5]–[7]. One byproduct of these efforts is a sharp characterization of the *additive gap*. Specifically, it was recently shown in [7] that, for arbitrary P_{XY}, it holds that $I(X;Y) - I(X;[Y]_M) = \mathcal{O}(M^{-2/(|\mathcal{X}|-1)})$, whereas [8] demonstrates that there exist P_{XY} such that $I(X;Y) - I(X;[Y]_M) = \Omega(M^{-2/(|\mathcal{X}|-1)})$. The works [5]–[7], among others, also

A. Bhatt is with the University of California, San Diego, CA, USA (email: a2bhatt@eng.ucsd.edu). B. Nazer is with Boston Univerity, Boston, MA, USA (email: bobak@bu.edu) O. Ordentlich is with the Hebrew University of Jerusalem, Israel (email: or.ordentlich@mail.huji.ac.il). Y. Polyanskiy is with the Massachusetts Institute of Technology, MA, USA (email: yp@mit.edu)

This work was supported, in part, by ISF under Grant 1791/17, by the NSF CAREER award CCF-12-53205, the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370, and NSF grants CCF-1618800, CCF-17-17842, and ECCS-1808692. The material in this paper was presented in part at the 2017 International Symposium on Information Theory [1].

[1]This notation is meant to suggest the distance from a point to a set.

[2]In [3], it was demonstrated that if an asymmetric signaling scheme is used, instead of binary phase-shift keying (BPSK), the additive white Gaussian noise channel capacity can be attained at low signal-to-noise ratio (SNR) with an asymmetric 2-level quantizer.

provided polynomial-complexity, sub-optimal algorithms for designing such quantizers. In addition, for binary X, an algorithm for determining the optimal quantizer was proposed in [9] (drawing upon a result from [10]) that runs in time $\mathcal{O}(|\mathcal{Y}|^3)$. A supervised learning algorithm, for the scenario where P_{XY} is not known, and cannot be estimated with good accuracy, was proposed in [11].

It may at first appear surprising that the quality of quantization found in [5]–[7] depends on the alphabet size $|\mathcal{X}|$ but not on $|\mathcal{Y}|$. The reason for this is that, given $Y = y$, the relevant information about X is the the posterior distribution $P_{X|Y=y}$, which is a point on $(|\mathcal{X}|-1)$-dimensional simplex. Thus, the goal of quantizing Y is essentially a goal of quantizing the probability simplex. The goal of this paper is to understand the fundamental limits of this quantization, as a function of alphabet size. The crucial difference with [5]–[7] is that here we focus on the *multiplicative gap*, i.e., comparing the ratio of $I(X;[Y]_M)$ to $I(X;Y)$. The difference is especially profound in the case when $I(X;Y)$ is small. We ignore the algorithmic aspects of finding the optimal M-level quantizer and instead focus on the fundamental properties of the function $I(X;[Y]_M)$. To this end, we define and study the "information distillation" function

$$\mathrm{ID}_M(K,\beta) \triangleq \inf_{\substack{P_{XY}:\\ |\mathcal{X}|=K\\ I(X;Y)\geq\beta}} I(X;[Y]_M). \tag{2}$$

The infimum above is taken with respect to all joint distributions with discrete input alphabet $\mathcal{X}$ of cardinality K and *arbitrary* (possibly continuous) output alphabet $\mathcal{Y}$ such that the mutual information is at least β. One may wonder whether K has an essential role in the function $\mathrm{ID}_M(K,\beta)$. Proposition 4, stated and proved in Section II, shows that for any M and β it holds that $\inf_K \mathrm{ID}_M(K,\beta) = 0$. Thus, one must indeed restrict the cardinality of $\mathcal{X}$ in (2) in order to get a meaningful quantity.

Special attention will be given to the binary input alphabet case, where $X \sim \mathrm{Bernoulli}(p)$ for some p. In this setting, it may seem at a first glance that the optimal binary quantizer should always retain a significant fraction of $I(X;Y)$, and that the MAP quantizer should be sufficient to this end. For large $I(X;Y)$, this is indeed the case, as we show in Proposition 6. As mentioned above, this is also the case if $Y = X + Z$ with Z Gaussian for all values $I(X;Y)$, since the MAP quantizer always retains at least $2/\pi \approx 63.66\%$ of the mutual information. However, perhaps surprisingly, we show that there is no constant $c > 0$ such that $I(X;[Y]_2) > c \cdot I(X;Y)$ for all P_{XY} with $|\mathcal{X}| = 2$.

A. Main Results

Our main result is a complete characterization, up to constants, of the binary information distillation function.

Theorem 1: For any mutual information value $0 < \beta \leq 1$, the binary information distillation function is lower and upper bounded as follows:

$$\beta \cdot f_{\text{lower}}\left(\frac{M-1}{\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}\right) \leq \mathrm{ID}_M(2,\beta) \leq \beta \cdot f_{\text{upper}}\left(\frac{M-1}{\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}\right),$$

where

$$f_{\text{lower}}(t) \triangleq \begin{cases} \frac{t}{208} & t < 104 \\ 1 - \frac{52}{t} & t \geq 104 \end{cases} \qquad f_{\text{upper}}(t) \triangleq \min\{3t, 1\}.$$

The proof is deferred to Section III-D. Note that the negative aspect of this result is in stark contrast to the intuition from the binary additive white Gaussian noise (AWGN) channel. While for the former, two quantization levels suffice for retaining a $2/\pi$ fraction of $I(X;Y)$, Theorem 4 shows that there exist sequences of distributions for which at least $\Omega(\log(1/I(X;Y)))$ quantization levels are needed in order to retain a fixed fraction of $I(X;Y)$.

Furthermore, as illustrated in Section III, for small $I(X;Y)$ and $M = 2$, the MAP quantizer can be arbitrary bad w.r.t. the optimal quantizer, which is in general not "symmetric." On the positive side, $\mathcal{O}(\log(1/I(X;Y)))$ quantization levels always suffice for retaining a fixed fraction of $I(X;Y)$.

For the general case where $2 < |\mathcal{X}| < \infty$, we prove the following.

Theorem 2: Define

$$a_0(M, |\mathcal{X}|, \beta) \triangleq \frac{1}{|\mathcal{X}| - 1} \cdot \min\left\{ \frac{M-1}{208 \log\left(\frac{(|\mathcal{X}|-1)^2}{\beta}\right)}, \frac{1}{2} \right\} \tag{3}$$

$$a_{|\mathcal{X}|-1}(M, |\mathcal{X}|, \beta) \triangleq \left(1 - \left(\frac{52 \log\left(\frac{e(|\mathcal{X}|-1)}{\beta}\right)}{M^{1/(|\mathcal{X}|-1)}}\right)^{2/3}\right)^2 \tag{4}$$

and, for $k = 1, \ldots, |\mathcal{X}| - 2$, define

$$a_k(M, |\mathcal{X}|, \beta) \triangleq \frac{k}{|\mathcal{X}| - 1} \left(1 - \frac{52 \log\left(\frac{(|\mathcal{X}|-1)^2}{\beta}\right)}{M^{1/k}}\right). \tag{5}$$

Then, for any $2 < |\mathcal{X}| < \infty$ and $0 < \beta \leq \log |\mathcal{X}|$, the information distillation function is upper and lower bounded as follows

$$\begin{aligned} \beta \cdot \max_{k \in \{0,1,\ldots,|\mathcal{X}|-1\}} a_k\left(M, |\mathcal{X}|, \beta\right) &\leq \mathrm{ID}_M(|\mathcal{X}|, \beta) \\ &\leq \beta \cdot f_{\text{upper}}\left(\frac{M-1}{\max\left\{\log\left(\frac{1}{\beta}\right), 1\right\}}\right), \end{aligned} \tag{6}$$

where $f_{\text{upper}}(t)$ is as defined in Theorem 1.

The proof of the lower bound is deferred to Section IV, whereas the upper bound follows trivially by noting that $|\mathcal{X}| \mapsto \mathrm{ID}_M(|\mathcal{X}|, \beta)$ is monotone non-increasing and invoking the upper bound on $\mathrm{ID}_M(2, \beta)$ from Theorem 1.

The lower bound from Theorem 2 states that for all P_{XY} and $k \in [|\mathcal{X}|-1]$, it holds that $M = \mathcal{O}((\log(|\mathcal{X}|/I(X;Y)))^k)$ suffices to guarantee that $I(X;[Y]_M) > \frac{k}{|\mathcal{X}|-1} I(X;Y)$. In particular, choosing $k = 1$, we obtain that $M = \mathcal{O}(\log(|\mathcal{X}|/I(X;Y)))$ suffices to attain $I(X;[Y]_M) > \frac{1}{|\mathcal{X}|-1} I(X;Y)$ and, on the other hand, by the upper bound, there exist P_{XY} for which $M = \Omega(\log(1/I(X;Y)))$ is required in order to attain $I(X;[Y]_M) > \frac{1}{|\mathcal{X}|-1} I(X;Y)$. Thus, Theorem 2 gives a tight characterization (up to constants independent of $I(X;Y)$) of the number of quantization levels required in order to maintain a fraction of $0 < \eta < \frac{1}{|\mathcal{X}|-1}$ of $I(X;Y)$. However, we were not successful in establishing an upper bound that match the lower bound within the range $\eta \in \left(\frac{1}{|\mathcal{X}|-1}, 1\right)$. We nevertheless conjecture that for η close to 1 our lower bound is tight.

Conjecture 1: For any $|\mathcal{X}| > 2$, there exists some $\frac{|\mathcal{X}|-2}{|\mathcal{X}|-1} < \eta(|\mathcal{X}|) < 1$, $\beta(|\mathcal{X}|) > 0$ and a constant $c(|\mathcal{X}|) > 0$, such that for all $0 < \beta < \beta(|\mathcal{X}|)$ and $M < c(|\mathcal{X}|)(\log(1/\beta))^{|\mathcal{X}|-1}$, it holds that

$$\mathrm{ID}_M(|\mathcal{X}|, \beta) < \eta(|\mathcal{X}|) \cdot \beta. \tag{7}$$

As discussed above, prior work [5]–[7] has focused on bounding the additive gap. This corresponds to bounding the so-called "degrading cost" [7], [8], which is defined as

$$\mathrm{DC}(|\mathcal{X}|, M) \triangleq \sup_{0 < \beta \leq \log |\mathcal{X}|} \beta - \mathrm{ID}_M(|\mathcal{X}|, \beta) \tag{8}$$

in our notation. In particular, the bound derived in [7] on $\mathrm{DC}(|\mathcal{X}|, M)$ is equivalent to the following "constant-gap" result: for every $0 < \beta \leq \log |\mathcal{X}|$,

$$\mathrm{ID}_M(|\mathcal{X}|, \beta) \geq \beta - \nu(|\mathcal{X}|) M^{-2/(|\mathcal{X}|-1)}$$

for some function ν.[3] For small β, however, results of this form are less informative. Indeed, for small β, this bound requires M to scale like $\beta^{-(|\mathcal{X}|-1)/2}$ in order to preserve a constant fraction of the mutual information. On the other hand, our result shows that scaling M like $\mathcal{O}((\log(1/\beta))^{|\mathcal{X}|-1})$ suffices for joint distributions P_{XY}.

Notation: In this paper, logarithms are generally taken w.r.t. base 2, and all information measures are given in bits. When a logarithm is taken w.r.t. base e, we use the notation $\ln$ instead of $\log$. We denote the binary entropy function by $h(t) = -t\log(t) - (1-t)\log(1-t)$, and its inverse restricted to the interval $[0, 1/2]$ by $h^{-1}(t)$. The notation $\lfloor t \rfloor$ denotes the "floor" operation, i.e., the largest integer smaller than or equal to t.

II. Properties of $I(X;[Y]_M)$

Let P_{XY} be a joint distribution on $\mathcal{X}\times\mathcal{Y}$ and consider the function $I(X;[Y]_M)$, as defined in (1). The restriction to deterministic functions incurs no loss of generality, see e.g., [9]. Indeed, any random function of y, can be expressed as $f(y,U)$ where U is some random variable statistically independent of (X,Y). Thus,

$$I(X;f(Y,U)) \le I(X;f(Y,U),U) = I(X;f(Y,U)|U) \tag{9}$$

and hence there must exist some u for which $I(X;f(Y,u)) \ge I(X;f(Y;U))$. Furthermore, for any function $f:\mathcal{Y}\to[M]$, we can associate a disjoint partition of the $|\mathcal{X}|$-dimensional cube $[0,1]^{|\mathcal{X}|}$ into M regions $\mathcal{I}_1,\ldots,\mathcal{I}_M$, such that $f(y)=i$ iff $P_{X|Y=y}\in\mathcal{I}_i$ for $i=1,\ldots,M$. A remarkable result of Burshtein et al. [10, Theorem 1] shows that the maximum in (1) can without loss of generality be restricted to functions for which there exists an associated partition where the regions $\mathcal{I}_1,\ldots,\mathcal{I}_M$ are all convex.

Below, we state simple upper and lower bounds on $I(X;[Y]_M)$.

Proposition 1 (Simple bounds): For any distribution P_{XY} on $\mathcal{X}\times\mathcal{Y}$ with a finite output alphabet, and $M<|\mathcal{Y}|$,

$$\frac{M-1}{|\mathcal{Y}|}I(X;Y) \le I(X;[Y]_M) \le \min\{I(X;Y),\log(M)\}.$$

Proof. The upper bound does not require any assumptions on $\mathcal{Y}$ and follows from the data processing inequality ($X-Y-f(Y)$ forms a Markov chain in this order), and from $I(X;f(Y))\le H(f(Y))\le\log(M)$.

For the lower bound, we can identify the elements of $\mathcal{Y}$ with $\{1,\ldots,|\mathcal{Y}|\}$ such that

$$P_Y(1)D(P_{X|Y=1}||P_X) \ge \cdots \ge P_Y(|\mathcal{Y}|)D(P_{X|Y=|\mathcal{Y}|}||P_X)$$

and take the quantization function

$$f(y) = \begin{cases} y & \text{if } y < M, \\ M & \text{otherwise.} \end{cases}$$

Recalling that $I(X;Y)=\sum_y P_Y(y)D(P_{X|Y=y}||P_X)$ we see that

$$I(X;f(Y)) \ge \frac{M-1}{|\mathcal{Y}|}I(X;Y).$$

■

For $K<M$, we can construct a (possibly sub-optimal) K-level quantizer by first finding the optimal M-level quantizer and then quantizing its output to K-levels. This together with the lower bound in Proposition 1, yields the following.

Corollary 1: For natural numbers $K<M$ we have

$$I(X;[Y]_K) \ge \frac{K-1}{M}I(X;[Y]_M).$$

[3]It is also demonstrated in [8] that there exist values of β, for which this bound is tight. Specifically, [8] found a distribution P_{XY} with $X\sim\text{Bernoulli}(1/2)$ and $I(X;Y)\approx 0.2787$ for which $I(X;[Y]_M) < I(X;Y) - cM^{-2}$ for some constant $c>0$.

Remark 1: It is tempting to expect that $I(X;[Y]_M)$ will have "diminishing returns" in M for any P_{XY}, i.e., that it will satisfy the inequality $I(X;[Y]_{M_1\cdot M_2}) \leq I(X;[Y]_{M_1}) + I(X;[Y]_{M_2})$. However, as demonstrated by the following example, this is not the case. Let $X \sim \mathrm{Uniform}(\{0,1,2,3\})$ and $Y = [X+Z] \bmod 4$, where Z is additive noise statistically independent of X with $\Pr(Z=0) = \delta$ and $\Pr(Z=1) = \Pr(Z=2) = \Pr(Z=3) = (1-\delta)/3$. Clearly,

$$I(X;[Y]_4) = I(X;Y) = 2 - h(\delta) - (1-\delta)\log(3), \tag{10}$$

and it can be verified that

$$I(X;[Y]_2) = \begin{cases} h\left(\frac{1}{4}\right) - \frac{1}{4}h(\delta) - \frac{3}{4}h\left(\frac{1-\delta}{3}\right) & \delta \leq 1/4, \\ 1 - h\left(\frac{1+2\delta}{3}\right) & \delta > 1/4. \end{cases} \tag{11}$$

Thus, for this example we have that $2I(X;[Y]_2) < I(X;[Y]_4)$ for all $\delta \notin \{1/4, 1\}$.

Proposition 2 (Data processing inequality): If $X - Y - V$ form a Markov chain in this order, then

$$I(X;[V]_M) \leq I(X;[Y]_M).$$

Proof. For any function $f : \mathcal{V} \mapsto [M]$ we can generate a random function $\tilde{f} : \mathcal{Y} \mapsto [M]$ which first passes Y through the channel $P_{V|Y}$ and then applies f on its output. By (9), we can always replace $\tilde{f}$ by some deterministic function $\bar{f} : \mathcal{Y} \mapsto [M]$ such that

$$I(X;\bar{f}(Y)) \geq I(X;\tilde{f}(Y)) = I(X;f(V)).$$

■

Proposition 3: For a fixed P_X, the function $P_{Y|X} \mapsto I(X;[Y]_M)$ is convex.

Proof. For any $f : \mathcal{Y} \mapsto [M]$, let $I^f(P_X \times P_{Y|X}) \triangleq I(X;f(Y))$, and note that

$$I(X;[Y]_M) = \sup_{f:\mathcal{Y}\mapsto[M]} I^f(P_X \times P_{Y|X}).$$

Since the supremum of convex functions is also convex, it suffices to show that for a fixed P_X the function $I^f(P_X \times P_{Y|X})$ is convex in $P_{Y|X}$. To this end, consider two channels $P^1_{Y|X}$ and $P^2_{Y|X}$, and let $P^1_{f(Y)|X}$ and $P^2_{f(Y)|X}$, respectively, be the induced channels from X to $f(Y)$. Clearly, for the channel $\alpha P^1_{Y|X} + (1-\alpha)P^2_{Y|X}$, the induced channel is $\alpha P^1_{f(Y)|X} + (1-\alpha)P^2_{f(Y)|X}$. Let $Z \in [M]$ be the output of this channel, when the input is X. From the convexity of the mutual information w.r.t. the channel we have

$$\begin{aligned} I^f\left(P_X \times \left(\alpha P^1_{Y|X} + (1-\alpha)P^2_{Y|X}\right)\right) &= I(X;Z) \\ &\leq \alpha I^f(P_X \times P^1_{Y|X}) + (1-\alpha) I^f(P_X \times P^2_{Y|X}), \end{aligned}$$

as desired. ■

Remark 2: In contrast to mutual information, the functional $I(X;[Y]_M)$ is in general not concave in P_X for a fixed $P_{Y|X}$. To see this consider the following example: $\mathcal{X} = \mathcal{Y} = \{1,2,3\}$, $M = 2$, and the channel from X to Y is clean, i.e., $Y = X$. Let $P_{X_1} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ and $P_{X_2} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$. Clearly, $I(X_1;[Y]_M) = I(X_2;[Y]_M) = 1$. For any $\alpha \in (0,1)$, let $P_X = \alpha P_{X_1} + (1-\alpha)P_{X_2}$. It can be verified that

$$I(X;[Y]_M) < 1.$$

Remark 3 (Complexity of finding the optimal quantizer): For the special case where $Y = X$, the function $I(X;[Y]_M)$ reduces to[4]

$$H([Y]_M) \triangleq \sup_{\tilde{Y}\in[Y]_M} H(\tilde{Y}). \tag{12}$$

[4]Recent work by Cicalese, Gargano and Vaccaro [12] provides closed-form upper and lower bounds on $H([Y]_M)$.

Furthermore, when $M = 2$ the optimization problem in (12) is equivalent to

$$\max_{\mathcal{A}\subseteq\mathcal{X}} \sum_{x\in\mathcal{A}} p_x \text{ subject to: } \sum_{x\in\mathcal{A}} p_x \leq \frac{1}{2}, \tag{13}$$

where $p_x \triangleq \Pr(X = x)$, $x \in \mathcal{X}$. The problem (13) is known as the *subset sum problem* and is NP-hard [13]. See also [12]. Thus, when $|\mathcal{X}|$ is not constrained, the problem of finding the optimal quantizer of Y is in general NP-hard. Nevertheless, for the case where $\mathcal{X}$ is binary, a dynamic programming algorithm finds the optimal quantizer with complexity $\mathcal{O}(|\mathcal{Y}|^3)$, see [9].

Proposition 4: For any $\beta > 0$, any natural M, and n large enough, we have that

$$\mathrm{ID}_M(2^n, \beta) \leq \frac{\log(M)}{n\log(e)}\beta, \tag{14}$$

Consequently, for any $\beta > 0$ and natural M we have that $\inf_K \mathrm{ID}_M(K, \beta) = 0$, which motivates the restriction to finite input alphabets in our main theorems.

Proof. Let $Y \sim \mathrm{Bernoulli}(1/2)$, $Z \sim \mathrm{Bernoulli}(\delta)$, $Y \perp\!\!\!\perp Z$, and $X = Y \oplus Z$. Let $(X^n, Y^n) \sim P_{XY}^{\otimes n}$. For product distributions $P_{XY}^{\otimes n}$ we have that for any U satisfying the Markov chain $U - Y^n - X^n$, it holds that [14], [15]

$$\frac{I(U; X^n)}{I(U; Y^n)} \leq \sup \frac{I(U; X)}{I(U; Y)}, \tag{15}$$

where the supremum is taken w.r.t. all Markov chains $U - Y - X$ with fixed P_{XY} and $I(U; Y) > 0$. For the doubly symmetric binary source P_{XY} of interest, this supremum is $(1 - 2\delta)^2$ [15], and consequently, we obtain that for any $f : \{0, 1\}^n \mapsto [M]$, it holds that

$$\begin{aligned} I(f(Y^n); X^n) &\leq (1 - 2\delta)^2 I(f(Y^n); Y^n) \\ &\leq (1 - 2\delta)^2 H(f(Y^n)) \\ &\leq (1 - 2\delta)^2 \log(M). \end{aligned} \tag{16}$$

For any $\beta > 0$, take n large enough such that $\beta \ll n \cdot 2\log(e)$, and set

$$\delta = \frac{1}{2} - \sqrt{\frac{\beta}{n \cdot 2\log(e)}} \tag{17}$$

we obtain that

$$I(X^n; [Y^n]_M) \leq \frac{\log(M)}{n \cdot 2\log(e)}\beta. \tag{18}$$

On the other hand, we have that

$$I(X^n; Y^n) = n(1 - h(\delta)) = \beta(1 + o(1)). \tag{19}$$

Thus, for n large enough (14) indeed holds. ■

A. Relations to quantization for maximizing divergence

For two distributions P, Q on $\mathcal{Y}$, $Q \ll P$, define

$$\psi_M(P, Q) \triangleq \sup_{f:\mathcal{Y}\mapsto[M]} D(P^f || Q^f), \tag{20}$$

where P^f and Q^f are the distributions on $[M]$ induced by applying the function f on the random variables generated by P and Q, respectively. A classical characterization of Gelfand-Yaglom-Perez [16, Section 3.4], shows that $\psi_M(P, Q) \nearrow D(P\|Q)$ as $M \to \infty$. We are interested here in understanding the speed of this convergence. To this end, we prove the following result.

Proposition 5: For any $\beta, \epsilon > 0$, there exists two distributions P, Q on $\mathbb{N}$ such that $D(P\|Q) = \beta$ and $\psi_M(P, Q) \leq M\epsilon$ for any $M \in \mathbb{N}$.

Proof. Consider the following two distributions:

$$P(m) = \begin{cases} 2^{-m} & m = 1, \ldots, T \\ 2^{-(T-1)} & m = T \\ 0 & m > T \end{cases}$$

$$Q(m) = \begin{cases} P(m) & 1 \leq m \leq k \\ g(m) \cdot P(m) & k < m \leq T \\ 1 - \sum_{m=1}^{k} P(m) & \\ -\sum_{m=k+1}^{T} g(m)P(m) & m = T+1 \end{cases}$$

where $0 < g(m) \leq 1$ is some monotonically non-increasing function. We have that

$$D(P||Q) = \sum_{m=k+1}^{T} 2^{-m} \log(1/g(m)), \tag{21}$$

whereas for any $f : \{0, 1, \ldots\} \mapsto [M]$ we have that

$$D(P^f||Q^f) \leq M \cdot \max_{A \subset \{0,1,\ldots\}} P(A) \log \frac{P(A)}{Q(A)}. \tag{22}$$

Let $A_k \triangleq A \cap [k]$. Without loss of generality, we can assume that $A \setminus A_k \neq \emptyset$, as otherwise $P(A) \log(P(A)/Q(A)) = 0$. Thus, we can define $\ell \triangleq \min\{a \ : \ a \in A \setminus A_k\}$ and write

$$\begin{aligned} P(A) &= P(A_k) + P(A \setminus A_k) \leq P(A_k) + 2 \cdot 2^{-\ell} \\ Q(A) &= Q(A_k) + Q(A \setminus A_k) \geq P(A_k) + 2^{-\ell} g(\ell) \end{aligned} \tag{23}$$

Let $t = 2^{\ell} P(A_k) + 2$, and $\tau = 2 - g(\ell)$ such that the bounds above read as $P(A) \leq 2^{-\ell} t$ and $Q(A) \geq 2^{-\ell}(t - \tau)$, and

$$P(A) \log \frac{P(A)}{Q(A)} \leq -2^{-\ell} t \log\left(1 - \frac{\tau}{t}\right). \tag{24}$$

We note that the function $\varphi(t) = -t \log(1 - \frac{\tau}{t})$ is convex and monotone decreasing in the range $t > \tau$. This implies that (24) is maximized by choosing A such that $P(A_k) = 0$, for which $t = 2$, and we obtain

$$D(P^f \| Q^f) < M \cdot 2^{-(\ell-1)} \log \frac{2}{g(\ell)}. \tag{25}$$

Now, take $g(m) = 2^{-\frac{\alpha 2^m}{m}}$ for some $0 < \alpha \leq 1$, and note that it is indeed monotone non-increasing in $m = 1, 2, \ldots$, which yields

$$D(P\|Q) = \alpha \sum_{m=k+1}^{T} \frac{1}{m} \tag{26}$$

$$D(P^f \| Q^f) \leq 2M \left(2^{-\ell} + \frac{\alpha}{\ell}\right) \leq 2M \left(2^{-k} + \frac{\alpha}{k}\right). \tag{27}$$

The statement follows by noting that we can always choose k such that the left hand side of (27) is smaller than ϵ, and then we can choose $T > k$ and α such that the left hand side of (26) is equal to β. ■

Proposition 5 shows that for any fixed M, and any value of $D(P\|Q)$, the ratio $\psi_M(P, Q)/D(P\|Q)$ can be arbitrarily small.[5] Note that choosing a different φ-divergence in the definition of $\psi_M(P, Q)$ instead of the KL-divergence, could lead to very different results. In particular, under the total variation criterion, the 1-bit quantizer $f(y) = \text{sign}(P(y) - Q(y))$ achieves $d_{\text{TV}}(P^f, Q^f) = d_{\text{TV}}(P, Q)$ for any pair of distributions P, Q on $\mathcal{Y}$. An interesting question for future study is for which φ-divergences is the ratio $\psi_M(P, Q)/D_\varphi(P, Q)$ always positive.

[5]However, under some restrictions on the distributions P and Q, it is shown in [17] that a 2-level quantizer suffices to retain a constant fraction of $D(P\|Q)$.

III. Bounds for binary X

In this section, we consider the case of $|\mathcal{X}| = 2$, and provide upper and lower bounds on $\mathrm{ID}_M(2,\beta)$. We begin by studying the case where $M = |\mathcal{X}| = 2$, through which we shall demonstrate why the multiplicative decrease in mutual information is small when $I(X;Y)$ is high (close to 1). These findings illustrate that the more interesting regime for $\mathrm{ID}_M(2,\beta)$ is the one where β is small. For this regime, we derive lower and upper bound that match up to constants that do not depend on β.

A. Binary Quantization ($M = 2$)

The aim of this subsection is to analyze the performance of quantizers whose cardinality is equal to that of $\mathcal{X}$. In this case, a natural choice for the quantizer is the maximum a posteriori (MAP) estimator of X from Y. Intuitively, when $I(X;Y)$ is high (close to $H(X)$), the MAP estimator should not make many errors and the mutual information between it and X should be high as well. We make this intuition precise below. However, when $I(X;Y)$ is low, it turns out that not only does the MAP estimator fail to retain a significant fraction of $I(X;Y)$, but it can be significantly inferior to other binary quantizers.

Assume without loss of generality that $\mathcal{X} = \{1,2\}$. The maximum a posteriori (MAP) quantizer is defined by

$$f_{\text{MAP}}(y) = \begin{cases} 1 & \text{if } \Pr(X=1|Y=y) > 1/2 \\ 2 & \text{if } \Pr(X=1|Y=y) < 1/2 \\ 1\cdot U + 2(1-U) & \text{if } \Pr(X=1|Y=y) = 1/2 \end{cases}, \tag{28}$$

where $U \sim \mathrm{Bernoulli}(1/2)$ is statistically independent of (X,Y). Let $P_{e,\text{MAP}}(y) \triangleq \Pr(f_{\text{MAP}}(Y) \neq X|Y=y)$ and $P_{e,\text{MAP}} \triangleq \mathbb{E}_Y P_{e,\text{MAP}}(Y)$. By the concavity of the binary entropy function $t \mapsto h(t)$, we have that $h(t) \geq 2t$ for any $0 \leq t \leq 1/2$, with equality iff $t \in \{0, 1/2\}$. Consequently,

$$H(X|Y) = \mathbb{E}_Y h(P_{e,\text{MAP}}(Y)) \geq 2P_{e,\text{MAP}}. \tag{29}$$

Let $X \sim \mathrm{Bernoulli}(p)$ and $I(X;Y) = \beta$. We have that

$$\begin{aligned} & I(X; f_{\text{MAP}}(Y)) \qquad (30) \\ &= H(X) - H(X|f_{\text{MAP}}(Y)) \\ &= h(p) - \Pr(f_{\text{MAP}}(Y)=1)h\big(\Pr(X\neq 1|f_{\text{MAP}}(Y)=1)\big) \\ &\qquad - \Pr(f_{\text{MAP}}(Y)=2)h\big(\Pr(X\neq 2|f_{\text{MAP}}(Y)=2)\big) \\ &\geq h(p) - h(P_{e,\text{MAP}}) \\ &\geq h(p) - h\left(\frac{H(X|Y)}{2}\right) \\ &= h(p) - h\left(\frac{h(p)-\beta}{2}\right). \qquad (31) \end{aligned}$$

Since $\beta \leq h(p) \leq 1$, we have obtained that

$$\begin{aligned} I(X;f_{\text{MAP}}(Y)) &\geq \min_{\beta \leq t \leq 1} t - h\left(\frac{t-\beta}{2}\right) \\ &= \begin{cases} \beta + \frac{2}{5} - h\left(\frac{1}{5}\right) & \beta < \frac{3}{5} \\ 1 - h\left(\frac{1-\beta}{2}\right) & \beta \geq \frac{3}{5} \end{cases}. \end{aligned} \tag{32}$$

Since $I(X;[Y]_2) \geq I(X; f_{\text{MAP}}(Y))$, it follows that the right hand side of (32) is a lower bound on $\mathrm{ID}_2(2,\beta)$.

In order to obtain an upper bound on $\mathrm{ID}_2(2,\beta)$, assume $X \sim \mathrm{Bernoulli}(1/2)$ and $P_{Y|X}$ is the binary erasure channel (BEC), i.e., $\mathcal{Y} = \{0,1,?\}$ and

$$\Pr(Y=y|X=x) = \begin{cases} \beta & \text{if } y = x \\ 1-\beta & \text{if } y = ? \end{cases}, \tag{33}$$

such that $I(X;Y)=\beta$. Consider the quantizer

$$f_Z(y)=\begin{cases}1 & \text{if } y\in\{1,?\}\\ 2 & \text{if } y=0\end{cases}.$$

Since there exists an optimal deterministic quantizer, and any deterministic 1-bit quantizer for the BEC output is of the form $f_Z(y)$, this must be an optimal 1-bit quantizer. Note that the induced channel from X to $f_Z(Y)$ is a Z-channel, and it satisfies

$$I(X;f_Z(Y))=\frac{\beta}{2}h\left(\frac{1-\beta}{2-\beta}\right)+1-h\left(\frac{1-\beta}{2-\beta}\right). \tag{34}$$

By the optimality of the quantizer $f_Z(\cdot)$ for this particular distribution, it follows that the right hand side of (34) constitutes an upper bound on $\mathrm{ID}_2(2,\beta)$.

We have therefore established the following proposition.

Proposition 6: For all $0\le\epsilon\le 2/5$ we have

$$1-h\left(\frac{\epsilon}{2}\right)\le \mathrm{ID}_2(2,1-\epsilon)\le 1-\frac{1+\epsilon}{2}h\left(\frac{\epsilon}{1+\epsilon}\right). \tag{35}$$

Thus, for large β, the loss for quantizing the output to one bit is small and the fraction of the mutual information that can be retained approaches 1 as the mutual information increases. In particular, the natural MAP quantizer is never too bad, and retains a significant fraction of at least $1-h((1-\beta)/2)$ of the mutual information β.

In the small β regime, we arrive at qualitatively different behavior. We next show that the MAP quantizer can be highly sub-optimal when β is small. To that end, consider again the distribution $X\sim\mathrm{Bernoulli}(1/2)$ and $P_{Y|X}$ given by (33). i.e., a BEC. It is easy to verify that in this case both inequalities in (31) are in fact equalities for all $0\le\beta\le 1$. It follows that for a BEC with capacity $\beta\ll 1$ and uniform input, we have that

$$I(X;f_{\text{MAP}}(Y))=1-h\left(\frac{1-\beta}{2}\right)=\frac{\log e}{2}\beta^2+o(\beta^2). \tag{36}$$

$$I(X;f_Z(Y))=\frac{\beta}{2}h\left(\frac{1-\beta}{2-\beta}\right)+1-h\left(\frac{1-\beta}{2-\beta}\right)=\frac{\beta}{2}+o(\beta). \tag{37}$$

Thus, the asymmetric quantizer $f_Z(y)$ retains 50% of the mutual information, whereas the fraction of mutual information retained by the symmetric MAP quantizer vanishes as β goes to zero.

One can argue that $f_Z(y)$ is a MAP estimator just as $f_{\text{MAP}}(y)$, as the two quantizers attain the same error probability in guessing the value of X based on Y, and dismiss our findings about the sub-optimality of $f_{\text{MAP}}(y)$ by attributing it to the randomness required by the MAP quantizer, as defined in (28), in the BEC setting. This is not the case however. To see this consider a channel with binary symmetric input and output alphabet $\mathcal{Y}=\{0,1\}\times\{g,b\}$, defined by

$$\Pr(Y=y|X=x)=\begin{cases}\beta & \text{if } y=(x,g)\\ (1-\beta)\left(\frac{1}{2}+\delta\right) & \text{if } y=(x,b)\\ (1-\beta)\left(\frac{1}{2}-\delta\right) & \text{if } y=(1-x,b)\end{cases},$$

for some $0\le\beta\le 1$ and $0\le\delta\le 1/2$. Note that for $\delta=0$, this channel becomes a BEC with capacity $1-\beta$. For any $\delta>0$, the corresponding MAP quantizer is deterministic, but as $\delta\to 0$, the channel approaches a BEC, and its performance becomes closer and closer to (36). Similarly, the performance of a binary quantizer that assigns the same value to both "bad" outputs, i.e., $f(y)=2$ if $y=(0,g)$ and $f(y)=1$ otherwise, approach (37) as $\delta\to 0$.

B. Lower Bound on Quantized Mutual Information

We prove the following lower bound on $I(X;[Y]_M)$.

Theorem 3: For any P_{XY} with $|\mathcal{X}|=2$ and $I(X;Y)=\beta$, and any $\eta\in(0,1)$ we have that

$$I(X;[Y]_{\bar{M}_2(\eta,\beta)})\geq \eta\beta, \tag{38}$$

where

$$\bar{M}_2(\eta,\beta)\triangleq\left\lfloor c_1(\eta)\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}\right\rfloor, \tag{39}$$

and

$$c_1(\eta)\triangleq\frac{52}{1-\eta}. \tag{40}$$

Proof. Consider the joint distribution P_{XY}, and for any $y\in\mathcal{Y}$ define $\alpha_y\triangleq\Pr(X=1|Y=y)$, $\bar{\alpha}\triangleq\mathbb{E}(\alpha_Y)=\Pr(X=1)$ and

$$D_y\triangleq D(P_{X|Y=y}\|P_X)=d\left(\alpha_y\,\|\,\bar{\alpha}\right), \tag{41}$$

where $d(p_1\|p_2)\triangleq p_1\log(p_1/p_2)+(1-p_1)\log((1-p_1)/(1-p_2))$ is the binary KL divergence function. Let $\kappa\triangleq\max\{\log(\frac{1}{\bar{\alpha}}),\log(\frac{1}{1-\bar{\alpha}})\}$. We further define the function

$$\bar{F}(\gamma)\triangleq\Pr(D_Y\geq\gamma), \tag{42}$$

and note that it is non-increasing and satisfies

$$I(X;Y)=\mathbb{E}D_Y=\int_0^{\gamma^*}\bar{F}(\gamma)d\gamma, \tag{43}$$

where $\gamma^*=\sup_{y\in\mathcal{Y}}D_y\leq\kappa$. Let L be some natural number, let $0=\gamma_0\leq\gamma_1\leq\cdots\leq\gamma_L\leq\gamma_{L+1}=\gamma^*+\delta$, for some arbitrary small $\delta>0$, and define the following $(2L+1)$-level quantizer

$$f(y)=\begin{cases}0 & d(\alpha_y\|\bar{\alpha})\leq\gamma_1\\ -\ell & \alpha_y<\bar{\alpha},\gamma_\ell\leq d(\alpha_y\|\bar{\alpha})<\gamma_{\ell+1}\\ \ell & \alpha_y>\bar{\alpha},\gamma_\ell\leq d(\alpha_y\|\bar{\alpha})<\gamma_{\ell+1}\end{cases}. \tag{44}$$

We have that for $\ell=1,\ldots,L$

$$d\left(\mathbb{E}[\alpha_Y|f(Y)=-\ell]\|\bar{\alpha}\right)\geq\gamma_\ell,\ d\left(\mathbb{E}[\alpha_Y|f(Y)=\ell]\|\bar{\alpha}\right)\geq\gamma_\ell$$

and by the definition of $\bar{F}(\gamma)$ we also have

$$\Pr\left(\{f(Y)=-\ell\}\cup\{f(Y)=\ell\}\right)=\bar{F}(\gamma_\ell)-\bar{F}(\gamma_{\ell+1}).$$

Thus,

$$\begin{aligned}
&I(X;f(Y))\\
&=\sum_{\ell=-L}^{L}\Pr(f(Y)=\ell)D(P_{X|f(Y)=\ell}\|P_X)\\
&\geq\sum_{\ell=1}^{L}\left(\bar{F}(\gamma_\ell)-\bar{F}(\gamma_{\ell+1})\right)\gamma_\ell\\
&=\bar{F}(\gamma_1)\gamma_1+\sum_{\ell=2}^{L}\bar{F}(\gamma_\ell)(\gamma_\ell-\gamma_{\ell-1})-\bar{F}(\gamma_{L+1})\gamma_L\\
&=\sum_{\ell=1}^{L}\bar{F}(\gamma_\ell)(\gamma_\ell-\gamma_{\ell-1}),
\end{aligned} \tag{45}$$

where in the last equality we used $\gamma_0 = 0$ and $\bar{F}(\gamma_{L+1}) = \bar{F}(\gamma^* + \delta) = 0$. Our goal is therefore to choose the numbers $\{\gamma_\ell\}_{\ell=1}^L$ such as to maximize (45). For a general L, this problem is difficult and we therefore resort to a possibly suboptimal choice according to the rule

$$\gamma_1 = \epsilon I(X;Y),\ \theta = \left(\frac{\gamma_1}{\kappa}\right)^{-\frac{1}{L}},\ \gamma_\ell = \gamma_1 \cdot \theta^{\ell-1}, \tag{46}$$

for $\ell = 2, \ldots, L, L+1$ and some $0 < \epsilon < 1$ to be specified. Note that this choice guarantees that

$$\gamma_{\ell+1} - \gamma_\ell \leq \theta\left(\gamma_\ell - \gamma_{\ell-1}\right),\ \ell = 1, \ldots, L. \tag{47}$$

This implies that

$$\begin{aligned} I(X;Y) &= \int_0^\kappa \bar{F}(\gamma) d\gamma \\ &= \sum_{\ell=0}^{L} \int_{\gamma_\ell}^{\gamma_{\ell+1}} \bar{F}(\gamma) d\gamma \\ &\leq \sum_{\ell=0}^{L} (\gamma_{\ell+1} - \gamma_\ell) \bar{F}(\gamma_\ell) \\ &\leq \gamma_1 + \theta \sum_{\ell=1}^{L} (\gamma_\ell - \gamma_{\ell-1}) \bar{F}(\gamma_\ell) \\ &\leq \gamma_1 + \theta I(X; f(Y)). \end{aligned} \tag{48}$$

Therefore,

$$I(X; f(Y)) \geq \frac{(1-\epsilon)}{\theta} I(X;Y). \tag{49}$$

Substituting in

$$\epsilon = \frac{1-\eta}{2}, L = \left\lceil \frac{2\log\left(\frac{2\kappa}{(1-\eta)I(X;Y)}\right)}{(1-\eta)} \right\rceil, \tag{50}$$

it follows that $\theta \leq 2^{\frac{(1-\eta)}{2}}$. Using this and the fact that $\frac{1-x}{2^x} \geq 1 - 2x$ for $0 \leq x \leq \frac{1}{2}$, from (49)

$$I(X; f(Y)) \geq \frac{(1 - \frac{(1-\eta)}{2})}{2^{\frac{(1-\eta)}{2}}} I(X;Y) \geq \eta I(X;Y).$$

Thus, as $|f(\cdot)| = 2L + 1$, and $L \geq 4$, we have shown that $I(X[Y]_M) \geq \eta I(X;Y)$ for

$$M = \left\lfloor \frac{8\log\left(\frac{2\kappa}{(1-\eta)I(X;Y)}\right)}{(1-\eta)} \right\rfloor. \tag{51}$$

Now consider the case where $I(X;Y) \leq \frac{1-\eta}{2}$. Note that

$$\begin{aligned} &\frac{8\log\left(\frac{2\kappa}{(1-\eta)I(X;Y)}\right)}{(1-\eta)} \\ &= \frac{8}{1-\eta}\left(\log(\kappa) + \log\left(\frac{2}{1-\eta}\right) + \log\left(\frac{1}{I(X;Y)}\right)\right) \\ &\leq \frac{8}{1-\eta}\left(\log(\kappa) + 2\log\left(\frac{1}{I(X;Y)}\right)\right) \end{aligned} \tag{52}$$

where (52) follows since $I(X;Y) \leq \frac{1-\eta}{2}$. Next, we show that $\log(\kappa) = \mathcal{O}\left(\log\left(\frac{1}{I(X;Y)}\right)\right)$. Without loss of generality, assume $\bar{\alpha} \leq (1-\bar{\alpha})$, and so $\kappa = \log\left(\frac{1}{\bar{\alpha}}\right)$. Now note that $I(X;Y) \leq H(X) = h(\bar{\alpha})$. Thus, $\bar{\alpha} \geq h^{-1}(I(X;Y))$. We then have the following bound on $h^{-1}(I(X;Y))$ [18, Theorem 2.2]

$$h^{-1}(I(X;Y)) \geq \frac{I(X;Y)}{2\log(\frac{6}{I(X;Y)})}. \tag{53}$$

Therefore, $\frac{1}{\bar{\alpha}} \leq \frac{1}{h^{-1}(I(X;Y))} \leq \frac{2}{I(X;Y)}\log\left(\frac{6}{I(X;Y)}\right)$ which in turn implies that $\log(\kappa) = \log\log(\frac{1}{\bar{\alpha}}) \leq \log\log\left(\frac{2}{I(X;Y)}\right) + \log\log\log\left(\frac{6}{I(X;Y)}\right)$. Thus, for any $I(X;Y) \leq \frac{1}{2}$,

$$\log(\kappa) \leq 2\log\left(\frac{1}{I(X;Y)}\right). \tag{54}$$

Using this in (52) we get

$$\frac{8\log\left(\frac{2\kappa}{(1-\eta)I(X;Y)}\right)}{(1-\eta)} \leq \frac{32}{1-\eta}\log\left(\frac{1}{I(X;Y)}\right) \tag{55}$$

for any $I(X;Y) = \beta \leq \frac{1-\eta}{2}$. Therefore, $I(X;[Y]_M) \geq \eta\beta$ for $M \geq \left\lfloor \frac{32}{1-\eta}\log\left(\frac{1}{\beta}\right)\right\rfloor$ whenever $\beta \leq \frac{1-\eta}{2}$.

When $\beta \geq \frac{1-\eta}{2}$, we use the bound

$$\beta - I(X;[Y]_M) \leq 1268M^{-2} \tag{56}$$

which is established in [7, Theorem 1]. This implies that for $\beta \geq \frac{1-\eta}{2}$,

$$I(X;[Y]_M) \geq \beta\left(1 - \frac{2536M^{-2}}{1-\eta}\right) \geq \eta\beta \tag{57}$$

whenever $M \geq \left\lfloor \frac{52}{1-\eta}\right\rfloor$.

Since

$$\max\left\{\frac{52}{1-\eta}, \frac{32\log\left(\frac{1}{\beta}\right)}{1-\eta}\right\} \leq \frac{52\max\left\{\log\left(\frac{1}{\beta}\right), 1\right\}}{1-\eta}, \tag{58}$$

combining the results obtained for each case $\left(\beta \leq \frac{1-\eta}{2}$ and $\beta \geq \frac{1-\eta}{2}\right)$ establishes

$$I(X;[Y]_M) \geq \eta\beta \quad \text{for} \quad M \geq \left\lfloor \frac{52\max\left\{\log\left(\frac{1}{\beta}\right), 1\right\}}{1-\eta}\right\rfloor,$$

as desired. ■

C. Upper Bound on Quantized Mutual Information

Theorem 4: For any $0 < \beta \leq 1$, there exist a distribution P_{XY} with $I(X;Y) \geq \beta$, for which

$$I(X;[Y]_M) \leq 2M\frac{\beta}{\ln\left(\frac{e\log(e)}{2\beta}\right)}, \tag{59}$$

for every natural M.

Proof. We provide a distribution P_{XY} with $I(X;Y) \geq \beta$ for which no M-level quantizer achieves mutual information exceeding the right hand side of (59). Let $X \sim \text{Bernoulli}(1/2)$ and $Y = (X \oplus Z_T, T)$ be the output

of a binary-input memoryless output-symmetric (BMS) whose input is X, where T is a mixed random variable in $[0, 1/2)$ whose probability density function is given by

$$f_T(t) = \begin{cases} r\delta(t) + \frac{4r}{(1-2t)^3} & 0^- < t \leq \frac{1-\sqrt{r}}{2} \\ 0 & \text{otherwise} \end{cases} \tag{60}$$

for some $0 < r \leq 1$, Z_T is a binary random variable with $\Pr(Z_T = 1|T = t) = t$, and (Z_T, T) is statistically independent of X. It can be easily verified that

$$\Pr(\alpha_Y = t|T = t) = \Pr(\alpha_Y = 1 - t|T = t) = 1/2 \ . \tag{61}$$

By [10, Theorem 1], the optimal quantizer partitions the interval $[0, 1]$ into M subintervals $\mathcal{I}_i = [\gamma_{i-1}, \gamma_i)$ for $i = 1, \ldots, M-1$ and $\mathcal{I}_M = [\gamma_{M-1}, \gamma_M]$, where $0 = \gamma_0 < \gamma_1 < \cdots < \gamma_M = 1$, and outputs $f(y) = i$ iff $\alpha_y \in \mathcal{I}_i$. We therefore have

$$\begin{aligned} I(X; f(Y)) &= \sum_{i=1}^{M} \Pr(\alpha_Y \in \mathcal{I}_i) d\left(\mathbb{E}[\alpha_Y|\alpha_Y \in \mathcal{I}_i] \,\middle\|\, \frac{1}{2}\right) \\ &\leq M \max_{0 \leq a < b \leq 1} \Pr(a \leq \alpha_Y \leq b) d\left(\mathbb{E}[\alpha_Y|a \leq \alpha_Y \leq b] \,\middle\|\, \frac{1}{2}\right). \end{aligned}$$

By the symmetry of the random variable α_Y around $1/2$, we can restrict the optimization to $a < 1/2$ and $a < b \leq 1$. Let $\underline{b} = \min\{b, 1-b\}$ and $\bar{b} = \max\{b, 1-b\}$ and define the two intervals $\mathcal{T}_0 = [a, \underline{b})$, $\mathcal{T}_1 = [\underline{b}, \bar{b}]$. By the convexity of KL divergence we have that

$$\begin{aligned} &d\left(\mathbb{E}[\alpha_Y|a \leq \alpha_Y \leq b] \,\middle\|\, \frac{1}{2}\right) \\ &= d\left(\sum_{i=0}^{1} \Pr(\alpha_Y \in \mathcal{T}_i|a \leq \alpha_Y \leq b)\mathbb{E}[\alpha_Y|\alpha_Y \in \mathcal{T}_i] \,\middle\|\, \frac{1}{2}\right) \\ &\leq \sum_{i=0}^{1} \Pr(\alpha_Y \in \mathcal{T}_i|a \leq \alpha_Y \leq b) d\left(\mathbb{E}[\alpha_Y|\alpha_Y \in \mathcal{T}_i] \,\middle\|\, \frac{1}{2}\right) \\ &= \Pr(\alpha_Y \in \mathcal{T}_0|a \leq \alpha_Y \leq b) d\left(\mathbb{E}[\alpha_Y|a \leq \alpha_Y \leq \underline{b}] \,\middle\|\, \frac{1}{2}\right), \end{aligned}$$

where in the last equation we have used the fact that $\mathbb{E}[\alpha_Y|\alpha_Y \in \mathcal{T}_1] = 1/2$, due to the symmetry of the random variable α_Y. We have therefore obtained

$$\begin{aligned} &I(X; f(Y)) \\ &\leq M \max_{0 \leq a \leq b \leq \frac{1}{2}} \Pr(a \leq \alpha_Y \leq b) d\left(\mathbb{E}[\alpha_Y|a \leq \alpha_Y \leq b] \,\middle\|\, \frac{1}{2}\right) \\ &\overset{(i)}{=} \frac{M}{2} \max_{0 \leq a \leq b \leq \frac{1}{2}} \Pr(a \leq T \leq b) d\left(\mathbb{E}[T|a \leq T \leq b] \,\middle\|\, \frac{1}{2}\right). \\ &\overset{(ii)}{=} \frac{M}{2} \max_{0 \leq b \leq \frac{1}{2}} \Pr(0 \leq T \leq b) d\left(\mathbb{E}[T|0 \leq T \leq b] \,\middle\|\, \frac{1}{2}\right) \end{aligned} \tag{62}$$

where (i) follows since for any interval $\mathcal{A} \subset [0, 1/2)$ we have that $\Pr(\alpha_Y \in \mathcal{A}) = \frac{1}{2}\Pr(T \in \mathcal{A})$ and $\mathbb{E}[\alpha_Y|\alpha_Y \in \mathcal{A}] = \mathbb{E}[T|T \in \mathcal{A}]$ by (61) and (ii) follows since for any choice of $0 < b \leq 1/2$, both terms are individually maximized by $a = 0$. It can be verified that for any $0 \leq \rho \leq \frac{1-\sqrt{r}}{2}$

$$\int_0^\rho t f_T(t) dt = \frac{2r\rho^2}{(1-2\rho)^2}; \ \Pr(0 \leq T \leq \rho) = \frac{r}{(1-2\rho)^2},$$

and therefore $\mathbb{E}[T|0 \le T \le b] = 2b^2$, and we have that for any M-level quantizer

$$\begin{aligned} I(X; f(Y)) &\le \frac{M}{2} \cdot \max_{0 \le b \le \frac{1-\sqrt{r}}{2}} r \cdot \frac{1 - h(2b^2)}{(1-2b)^2} \\ &\le M \cdot \log(e) r, \end{aligned} \tag{63}$$

where the last inequality follows by noting that the function $\frac{1-h(2b^2)}{(1-2b)^2}$ is monotone increasing in $0 < b < 1/2$, and taking the limit as $b \to 1/2$. It remains to relate r and $I(X;Y)$. Recalling that $h(\frac{1}{2} - p) \le 1 - 2\log(e)p^2$, we have

$$\begin{aligned} I(X;Y) &= 1 - \mathbb{E}h(T) \\ &\ge 2\log(e)\mathbb{E}\left(\frac{1}{2} - T\right)^2 \\ &= 2\log(e)\frac{r}{4}\ln\left(\frac{e}{r}\right) \\ &= \frac{e\log(e)}{2}\frac{r}{e}\ln\left(\frac{e}{r}\right). \end{aligned}$$

It can be verified that the function $g(t) \triangleq -t\ln(t)$ is monotone increasing in $0 < t < 1/e$ and its inverse restricted to this interval satisfies

$$\frac{1}{e} \cdot \frac{t}{-\ln(t)} < g^{-1}(t) \le \frac{t}{-\ln(t)}. \tag{64}$$

It therefore follows that

$$r \le e g^{-1}\left(\frac{2I(X;Y)}{e\log(e)}\right) \le \frac{2I(X;Y)}{\log(e)} \frac{1}{\ln\left(\frac{e\log(e)}{2I(X;Y)}\right)} \tag{65}$$

which gives

$$I(X; f(Y)) \le 2M \frac{I(X;Y)}{\ln\left(\frac{e\log(e)}{2I(X;Y)}\right)}, \tag{66}$$

for any M-level function f. ■

D. Proof of Theorem 1

We begin by proving the lower bound. Using Theorem 3 and solving for η, we obtain that

$$I(X; [Y]_M) \ge \left(1 - \frac{52\max\left\{\log\left(\frac{1}{\beta}\right), 1\right\}}{M}\right) \cdot \beta. \tag{67}$$

As a consequence of Theorem 3, we also have that

$$I\left(X; [Y]_{\left\lfloor 104\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}\right\rfloor}\right) \ge \frac{1}{2}\beta. \tag{68}$$

Now, applying Corollary 1, we obtain that for any $M < 104\max\left\{\log\left(\frac{1}{\beta}\right), 1\right\}$ it holds that

$$I(X; [Y]_M) \ge \frac{M-1}{104\max\left\{\log\left(\frac{1}{\beta}\right), 1\right\}} \frac{1}{2}\beta. \tag{69}$$

Combining (67) and (69) establishes that $\mathrm{ID}_M(2, \beta) \ge f_{\text{lower}}\left(\frac{M-1}{\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}\right)$.

To establish the upper bound, we use Theorem 4. Note first that for any $M > 1$ and $\beta > 1/2$ we have that $f_{\text{upper}}\left(\frac{M-1}{\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}\right) = 1$. Thus, it suffices to prove that $\mathrm{ID}_M(2,\beta) \leq f_{\text{upper}}\left(\frac{M-1}{\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}\right)$ for $\beta < 1/2$. In this case Theorem 4 shows that

$$\begin{aligned}
\mathrm{ID}_M(2,\beta) &\leq 2M \frac{\beta}{\ln\left(\frac{e\log(e)}{2\beta}\right)} \\
&= \frac{2}{\ln 2} \frac{M\beta}{\log\left(\frac{1}{\beta}\right)\left(1+\frac{\log(e\log(e)/2)}{\log\left(\frac{1}{\beta}\right)}\right)} \\
&\leq \frac{2}{\ln 2\,(1+\log(e\log(e)/2))} \cdot \frac{M}{\log\left(\frac{1}{\beta}\right)} \cdot \beta \\
&< \frac{3}{2} \cdot \frac{M}{\log\left(\frac{1}{\beta}\right)} \cdot \beta. \qquad (70)
\end{aligned}$$

Combining this with the trivial bound $\mathrm{ID}_M(2,\beta) \leq \beta$, we obtain

$$\begin{aligned}
\mathrm{ID}_M(2,\beta) &\leq \min\left\{\frac{3}{2} \cdot \frac{M}{\log\left(\frac{1}{\beta}\right)}, 1\right\} \cdot \beta \\
&\leq \min\left\{\frac{3}{2} \cdot \frac{M}{\max\left\{\log\left(\frac{1}{\beta}\right), 1\right\}}, 1\right\} \cdot \beta. \qquad (71)
\end{aligned}$$

Noting that for $M = 1$ we trivially have $\mathrm{ID}_M(2,\beta) = 0$ and that $3M/2 \leq 3(M-1)$ for $M > 1$, we obtain the desired result.

IV. Bounds for $|\mathcal{X}| > 2$

In the previous section we have shown that if X is binary, then $\mathcal{O}(\log(1/I(X;Y)))$ quantization levels always suffice in order to retain any constant fraction $0 < \eta < 1$ of $I(X;Y)$. In this section, we leverage this result in order to show that in general $\mathcal{O}(\log(1/I(X;Y))^k)$ quantization levels suffice in order to retain a constant fraction $0 < \eta < \frac{k}{|\mathcal{X}|-1}$, for $k \in [|\mathcal{X}|-1]$.

For a random variable $X \in \{1,\ldots,|\mathcal{X}|\}$, we can define the $|\mathcal{X}|-1$ binary random variables $A_i \triangleq \mathbb{1}_{\{X=i\}}$, $i = 1,\ldots,|\mathcal{X}|-1$. Clearly, X fully determines $\{A_1,\ldots,A_{|\mathcal{X}|-1}\}$ and vice versa. In particular, the encoding of X by $\{A_1,\cdots,A_{|\mathcal{X}|-1}\}$ can be thought of as the "one-hot" encoding of X, with the last bit, whose value is deterministically dictated by the preceding $|\mathcal{X}|-1$ bits, omitted. Representing X in this manner, nevertheless, allows us to reduce the problem of quantizing Y in order to retain information on X, into $|\mathcal{X}|-1$ separate problems of quantizing Y in order to retain information on A_i. Since the random variables $\{A_1,\ldots,A_{|\mathcal{X}|-1}\}$ are binary, the results from Theorem 3 can be applied.

The main result of this section is Theorem 5, that lower bounds the worst-case multiplicative loss due to quantization. Before stating this result, and giving its proof, we demonstrate the technique of reducing to the binary case via "one-hot" encoding for the setup considered in [7], [8]. Recall that for all $M \geq 2|\mathcal{X}|$ and $|\mathcal{Y}| > 2|\mathcal{X}|$, [7, Theorem 1] bounds the worst-case additive gap due to quantization as

$$\sup_{P_{XY}} I(X;Y) - I(X;[Y]_M) \leq \nu(|\mathcal{X}|) \cdot M^{-\frac{2}{|\mathcal{X}|-1}}, \qquad (72)$$

where the function $\nu(|\mathcal{X}|)$ is explicitly defined in [7] and the supremum is with respect to all P_{XY} with input alphabet of cardinality $|\mathcal{X}|$, and output alphabet of cardinality $|\mathcal{Y}|$. We further note that $\nu(|\mathcal{X}|)$ satisfies $\nu(2) \leq 1268$ and $\nu(|\mathcal{X}|) \approx 16\pi e|\mathcal{X}|^3$ for large $|\mathcal{X}|$.

Below, we use a "one-hot" encoding technique combined with (72) for $|\mathcal{X}| = 2$ only to obtain a slight refinement of the constant in the additive gap for $|\mathcal{X}| > 2$ (for large enough values of M).

Proposition 7: For $|\mathcal{X}| \geq 2$ and any M such that $M^{\frac{1}{|\mathcal{X}|-1}} \geq 4$ is an integer, we have

$$\mathrm{DC}(|\mathcal{X}|, M) \leq 1268(|\mathcal{X}| - 1) \cdot M^{-\frac{2}{|\mathcal{X}|-1}} \ . \tag{73}$$

Proof. Without loss of generality we may assume $|\mathcal{Y}| > 4$, as otherwise the assumption $M^{\frac{1}{|\mathcal{X}|-1}} \geq 4$ implies that $M \geq |\mathcal{Y}|$, in which case $I(X;[Y]_M) = I(X;Y)$.

The case $|\mathcal{X}| = 2$ is therefore obtained from (72), which reads

$$I(X;Y) - I(X;[Y]_M) \leq \nu(2) \cdot M^{-2} \tag{74}$$

for all P_{XY} with binary X.

Now let $|\mathcal{X}| > 2$, and without loss of generality assume $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$. Define $A_i \triangleq \mathbb{1}_{\{X=i\}}$, for $i = 1, 2, \ldots, |\mathcal{X}| - 1$. Then,

$$\begin{aligned} I(X;Y) &= I(A_1, \ldots, A_{|\mathcal{X}|-1}; Y) \\ &= \sum_{i=1}^{|\mathcal{X}|-1} I(A_i; Y | A_1^{i-1} = 0) \Pr(A_1^{i-1} = 0) \end{aligned} \tag{75}$$

where $A_1^{i-1} = 0$ denotes the event $A_1 = \cdots = A_{i-1} = 0$.

Let $f(y)$ be an M-level quantizer of the form $f(y) = (f_1(y), \ldots, f_{|\mathcal{X}|-1}(y))$. Then,

$$\begin{aligned} I(X; f(Y)) &= \sum_{i=1}^{|\mathcal{X}|-1} I(A_i; f(Y) | A_1^{i-1} = 0) \Pr(A_1^{i-1} = 0) \\ &\geq \sum_{i=1}^{|\mathcal{X}|-1} I(A_i; f_i(Y) | A_1^{i-1} = 0) \Pr(A_1^{i-1} = 0). \end{aligned} \tag{76}$$

Thus, combining (75) and (76), gives

$$\begin{aligned} I(X;Y) - I(X; f(Y)) \leq \sum_{i=1}^{|\mathcal{X}|-1} (&I(A_i; Y | A_1^{i-1} = 0) - \\ &I(A_i; f_i(Y) | A_1^{i-1} = 0)) \Pr(A_1^{i-1} = 0). \end{aligned} \tag{77}$$

From (74), it holds that by choosing $|f_i(y)| = M^{\frac{1}{|\mathcal{X}|-1}} \geq 4$, for all $1 \leq i \leq |\mathcal{X}| - 1$, we can find quantizers $f_1(y), \ldots, f_{|\mathcal{X}|-1}(y)$ for which

$$I(A_i; Y | A_1^{i-1} = 0) - I(A_i; f_i(Y) | A_1^{i-1} = 0) \leq \nu(2) \cdot M^{\frac{-2}{|\mathcal{X}|-1}}. \tag{78}$$

Consequently, with this choice, we obtain

$$\begin{aligned} I(X;Y) - I(X; f(Y)) &\leq \nu(2) \cdot M^{\frac{-2}{|\mathcal{X}|-1}} \sum_{i=1}^{|\mathcal{X}|-1} \Pr(A_1^{i-1} = 0) \\ &\leq (|\mathcal{X}| - 1)\nu(2) \cdot M^{\frac{-2}{|\mathcal{X}|-1}}, \end{aligned} \tag{79}$$

as desired. ■

Next, we focus on the regime of $I(X;Y) \ll 1$ and prove an upper bound on the number of quantization levels M, required to attain a fraction $0 < \eta < 1$ of $I(X;Y)$. Roughly, we show that it suffices to take M that scales like $(\log(1/I(X;Y)))^{\eta \cdot (|\mathcal{X}|-1)}$. More precisely, for any $k \in [|\mathcal{X}| - 1]$, if $\eta < \frac{k}{|\mathcal{X}|-1}$, then $\mathcal{O}\left((\log(1/I(X;Y)))^k\right)$ levels suffice.

Theorem 5: For any P_{XY} with $I(X;Y)=\beta$, and any $\eta\in(0,1)$, we have that

$$I(X;[Y]_{\bar{M}_{|\mathcal{X}|}(\eta,\beta)})\geq\eta\beta, \tag{80}$$

where

$$\bar{M}_{|\mathcal{X}|}(\eta,\beta)\triangleq\left\lfloor\left[c_1\left(\sqrt{\eta}\right)\log\left(\frac{|\mathcal{X}|-1}{(1-\sqrt{\eta})\beta}\right)\right]^{|\mathcal{X}|-1}\right\rfloor, \tag{81}$$

with $c_1(\eta)$ as defined in (40). Furthermore, if $\eta<\frac{k}{|\mathcal{X}|-1}$ for some natural $k\leq|\mathcal{X}|-2$, we have that

$$I(X;[Y]_{\tilde{M}_{|\mathcal{X}|}(\eta,\beta,k)})\geq\eta\beta, \tag{82}$$

where

$$\tilde{M}_{|\mathcal{X}|}(\eta,\beta,k)\triangleq\left\lfloor\left[c_1\left(\frac{\eta}{k/(|\mathcal{X}|-1)}\right)\log\left(\frac{(|\mathcal{X}|-1)^2}{\beta}\right)\right]^{k}\right\rfloor. \tag{83}$$

Proof. As above, define $A_i\triangleq\mathbb{1}_{\{X=i\}}$, such that

$$I(X;Y)=I(A_1,\ldots,A_{|\mathcal{X}|-1};Y)=\sum_{i=1}^{|\mathcal{X}|-1}I_i\cdot p_i \tag{84}$$

where

$$I_i\triangleq I(A_i;Y|A_1^{i-1}=0),\ \ i=1,\ldots,|\mathcal{X}|-1, \tag{85}$$

$$p_i\triangleq\Pr(A_1^{i-1}=0),\qquad i=1,\ldots,|\mathcal{X}|-1, \tag{86}$$

and $A_1^{i-1}=0$ denotes the event $A_1=\cdots=A_{i-1}=0$. Furthermore, set

$$v_i\triangleq\frac{I_i\cdot p_i}{\beta},\ \ i=1,\ldots,|\mathcal{X}|-1, \tag{87}$$

and let the permutation $\pi:[|\mathcal{X}|-1]\mapsto[|\mathcal{X}|-1]$ be such that $v_{\pi(1)}\geq\cdots\geq v_{\pi(|\mathcal{X}|-1)}$. For $0\leq k\leq|\mathcal{X}|-1$, define the function

$$F(k)\triangleq\sum_{i=1}^{k}v_{\pi(i)}, \tag{88}$$

with the convention that $F(0)=0$, and note that

1) $F(t)\geq\frac{t}{|\mathcal{X}|-1}$ for any natural $t\leq|\mathcal{X}|-1$, and in particular, $F(|\mathcal{X}|-1)=1$;
2) $F(|\mathcal{X}|-1)-F(t-1)=\sum_{i=t}^{|\mathcal{X}|-1}v_{\pi(i)}\leq(|\mathcal{X}|-t)v_{\pi(t)}$ and therefore, for any natural $t\leq|\mathcal{X}|-1$ we have that $v_{\pi(t)}\geq\frac{1-F(t-1)}{|\mathcal{X}|-t}$;

Let $\eta\in(0,1)$ and let $\bar{\eta}$ be some number satisfying $0<\eta<\bar{\eta}<1$. Let

$$k_{\bar{\eta}}=\min\{k\ :\ F(k)\geq\bar{\eta}\}. \tag{89}$$

By the definition of $k_{\bar{\eta}}$, we have that $F(k_{\bar{\eta}}-1)<\bar{\eta}$. Thus, by the second property, we have that

$$v_{\pi(k_{\bar{\eta}})}\geq\frac{1-\bar{\eta}}{|\mathcal{X}|-k_{\bar{\eta}}}\geq\frac{1-\bar{\eta}}{|\mathcal{X}|-1}. \tag{90}$$

Let $\eta'=\frac{\eta}{\bar{\eta}}<1$. Consider the conditional joint distribution $P_{A_iY|A_1^{i-1}=0}$. Since A_i is a binary random variable, by Theorem 3 we can design a quantizer $f_i:\mathcal{Y}\mapsto[M_i]$ with $M_i\leq\left\lfloor c_1(\eta')\log\max\left\{\left(\frac{1}{I_i}\right),1\right\}\right\rfloor$ quantization levels, such that $I(A_i;f_i(Y)|A_1^{i-1}=0)\geq\eta'\cdot I_i$. Let $f(y)=(f_{\pi(1)}(y),\ldots,f_{\pi(k_{\bar{\eta}})}(y)):\mathcal{Y}\mapsto[M_{\pi(1)}]\times\cdots\times[M_{\pi(k_{\bar{\eta}})}]$ be

the Cartesian product of the quantizers $f_{\pi(1)}(y),\ldots,f_{\pi(k_{\bar{\eta}})}(y)$ attaining this tradeoff between η' and the number of quantization levels. We have that

$$\begin{aligned}
I(X;f(Y)) &= I(A_1,\ldots,A_{|\mathcal{X}|-1};f(Y)) \\
&= \sum_{i=1}^{|\mathcal{X}|-1} I(A_i;f(Y)|A_1^{i-1}=0)\Pr(A_1^{i-1}=0) \\
&\geq \sum_{i=1}^{k_{\bar{\eta}}} I(A_{\pi(i)};f(Y)|A_1^{\pi(i)-1}=0)\Pr(A_1^{\pi(i)-1}=0) \\
&\geq \sum_{i=1}^{k_{\bar{\eta}}} I(A_{\pi(i)};f_{\pi(i)}(Y)|A_1^{\pi(i)-1}=0)p_{\pi(i)} \\
&\geq \sum_{i=1}^{k_{\bar{\eta}}} \eta' I_{\pi(i)}p_{\pi(i)} \\
&= \eta'\beta\sum_{i=1}^{k_{\bar{\eta}}} v_{\pi(i)} \\
&= \frac{\eta}{\bar{\eta}}\beta F(k_{\bar{\eta}}) \\
&\geq \eta\beta. \qquad (91)
\end{aligned}$$

Since $I_{\pi(i)} = \beta v_{\pi(i)}/p_{\pi(i)} \geq \beta v_{\pi(i)} > \frac{\beta(1-\bar{\eta})}{|\mathcal{X}|-1}$, $\forall i \leq k_{\bar{\eta}}$, by (90), we have that $\forall i \leq k_{\bar{\eta}}$

$$\begin{aligned}
M_{\pi(i)} &\leq \left\lfloor c_1(\eta')\max\left\{\log\left(\frac{|\mathcal{X}|-1}{(1-\bar{\eta})\beta}\right),1\right\}\right\rfloor \\
&= \left\lfloor c_1(\eta')\log\left(\frac{|\mathcal{X}|-1}{(1-\bar{\eta})\beta}\right)\right\rfloor, \qquad (92)
\end{aligned}$$

where the last inequality follows since $\frac{|\mathcal{X}|-1}{\beta} \geq \frac{|\mathcal{X}|-1}{\log|\mathcal{X}|} \geq 2$ for all $|\mathcal{X}|>2$. Consequently, we obtained

$$|f(y)| \leq \left\lfloor\left[c_1\left(\frac{\eta}{\bar{\eta}}\right)\log\left(\frac{|\mathcal{X}|-1}{(1-\bar{\eta})\beta}\right)\right]^{k_{\bar{\eta}}}\right\rfloor. \qquad (93)$$

To establish the first part of the statement, take $\bar{\eta}=\sqrt{\eta}$ and recall that $k_{\bar{\eta}} \leq |\mathcal{X}|-1$ by definition.

For the second part, note that if $\eta < \frac{k}{|\mathcal{X}|-1}$, for some $k<|\mathcal{X}|-1$ we may take $\bar{\eta}=\frac{k}{|\mathcal{X}|-1}$, and that $k_{\bar{\eta}} \leq k$, as $F(t) \geq \frac{t}{|\mathcal{X}|-1}$. Substituting into (93), and noting that $1-\bar{\eta} \geq \frac{1}{|\mathcal{X}|-1}$, establishes the second part of the statement. ■

Theorem 2 now follows as a rather simple corollary.

Proof of Theorem 2. We first show that for $1 \leq k \leq |\mathcal{X}|-2$, it holds that $I(X;[Y]_M) \geq a_k(M,|\mathcal{X}|,\beta)\cdot\beta$. To that end, for any $0<\eta'<1$, and $1\leq k\leq|\mathcal{X}|-2$ define

$$M(|\mathcal{X}|,\beta,\eta',k) = \left\lfloor\left[\frac{52}{1-\eta'}\log\left(\frac{(|\mathcal{X}|-1)^2}{\beta}\right)\right]^k\right\rfloor. \qquad (94)$$

By the second part of Theorem 5, we have that

$$I(X;[Y]_{M(|\mathcal{X}|,\beta,\eta',k)}) \geq \frac{k}{|\mathcal{X}|-1}\eta'\beta. \qquad (95)$$

Solving for η' shows that

$$I(X;[Y]_M) \geq \frac{k}{|\mathcal{X}|-1}\left(1-\frac{52\log\left(\frac{(|\mathcal{X}|-1)^2}{\beta}\right)}{M^{\frac{1}{k}}}\right)\beta, \qquad (96)$$

and maximizing with respect to k yields the bound

$$I(X;[Y]_M) \geq \max_{1\leq k\leq |\mathcal{X}|-2} a_k(M,|\mathcal{X}|,\beta)\cdot\beta. \tag{97}$$

Moreover, (95) applied with $\eta' = 1/2$ and $k = 1$ shows that

$$I\left(X;[Y]_{\left\lfloor 104\log\left(\frac{(|\mathcal{X}|-1)^2}{\beta}\right)\right\rfloor}\right) \geq \frac{1}{2(|\mathcal{X}|-1)}\beta. \tag{98}$$

Now, applying Corollary 1, we obtain that for any $M < 104\log\left(\frac{(|\mathcal{X}|-1)^2}{\beta}\right)$ it holds that

$$I(X;[Y]_M) \geq \frac{M-1}{104\log\left(\frac{(|\mathcal{X}|-1)^2}{\beta}\right)}\frac{1}{2(|\mathcal{X}|-1)}\beta. \tag{99}$$

which is equivalent to

$$I(X;[Y]_M) \geq a_0(M,|\mathcal{X}|,\beta)\cdot\beta. \tag{100}$$

Finally, we use the first part of Theorem 5 to show that $I(X;[Y]_M) \geq a_{|\mathcal{X}|-1}(M,|\mathcal{X}|,\beta)\cdot\beta$. Recalling the definition of $\bar{M}_{|\mathcal{X}|}(\eta,\beta)$ in (81), we have that for any $0<\eta<1$

$$\begin{aligned}
\bar{M}_{|\mathcal{X}|}(\eta,\beta) &= \left\lfloor\left[c_1\left(\sqrt{\eta}\right)\log\left(\frac{|\mathcal{X}|-1}{(1-\sqrt{\eta})\beta}\right)\right]^{|\mathcal{X}|-1}\right\rfloor \\
&\leq \left\lfloor\left[c_1(\sqrt{\eta})\left(\log\left(\frac{|\mathcal{X}|-1}{\beta}\right)+\frac{\log(e)}{(1-\sqrt{\eta})^{1/2}}\right)\right]^{|\mathcal{X}|-1}\right\rfloor \\
&\leq \left\lfloor\left[\frac{52}{(1-\sqrt{\eta})^{3/2}}\left(\log\left(\frac{e(|\mathcal{X}|-1)}{\beta}\right)\right)\right]^{|\mathcal{X}|-1}\right\rfloor \\
&\triangleq M'_{|\mathcal{X}|}(\eta,\beta).
\end{aligned} \tag{101}$$

Thus, by the first part of Theorem 5, we have that

$$I(X;[Y]_{M'_{|\mathcal{X}|}(\eta,\beta)}) \geq \eta\beta. \tag{102}$$

Solving for η yields

$$I(X;[Y]_M) \geq a_{|\mathcal{X}|-1}(M,|\mathcal{X}|,\beta)\cdot\beta. \tag{103}$$

The theorem now follows by combining (97), (100), and (103). ■

V. Connections to Quantization Under Log-Loss and the Information Bottleneck Problem

In general, an M-level quantizer q for a random variable Y consists of a disjoint partition of its alphabet $\mathcal{Y} = \bigcup_{i=1}^{M}\mathcal{S}_i$, and a set of corresponding reproduction values $a_i \in \mathcal{A}$, such that $q_y = \sum_{i=1}^{M} a_i \mathbb{1}_{\{y\in\mathcal{S}_i\}}$, see, e.g., [19]. The performance of the quantizer is measured with respect to some predefined distortion function $d : \mathcal{Y}\times\mathcal{A}\mapsto\mathbb{R}$, which quantifies the "important features" of Y that the quantizer should aim to retain. The expected distortion $\mathbb{E}d(Y,q_Y)$ is then typically taken as the quantizer's main figure of merit.

In our considerations, we observe and quantize the random variable Y, but the distortion measure is evaluated with respect to X, where X and Y are jointly distributed according to P_{XY}. This setup is sometimes referred to as remote source coding (or quantization). If the distortion measure of interest between X and the reconstruction q_y is $\tilde{d} : \mathcal{X}\times\mathcal{A}\mapsto\mathbb{R}$, one can define the induced distortion measure

$$d(y,q_y) = \mathbb{E}\left[\tilde{d}(X,q_y)|Y=y\right], \tag{104}$$

such that $\mathbb{E}[\tilde{d}(X,q_Y)] = \mathbb{E}[d(Y,q_Y)]$. Consequently, the remote quantization problem is reduced to a direct quantization problem, with an induced distortion measure [20].

Under various tasks of inferring information about X from Y, it is natural to take the reconstruction alphabet $\mathcal{A}$ to be the set of all distributions on $\mathcal{X}$, i.e., the $|\mathcal{X}|-1$ dimensional simplex $\mathcal{P}^{|\mathcal{X}|-1}$ [21]–[23] . Ideally, we would like the reconstructed distribution q_y to be as close as possible to the conditional distribution $P_{X|Y=y}$, for all $y \in \mathcal{Y}$. Various loss functions can be used to measure the distance between two distributions, depending on the ultimate performance criterion for the inference of X. One such loss function, that has enjoyed a special status in the information theory and machine learning literature [23]–[26] is the logarithmic-loss:

$$d(x,P) = \log\left(\frac{1}{P(x)}\right),\ \forall (x,P) \in \mathcal{X} \times \mathcal{P}^{|\mathcal{X}|-1}. \tag{105}$$

For the remote quantization setup, the induced distortion measure is

$$d(y,P) \triangleq \mathbb{E}\left[\log\frac{1}{P(X)}\middle| Y=y\right],\ \forall (y,P) \in \mathcal{Y} \times \mathcal{P}^{|\mathcal{X}|-1}. \tag{106}$$

Thus, the design of a quantizer for Y under $d(y,P)$ reduces to determining a disjoint partition $\mathcal{Y} = \bigcup_{i=1}^{M} \mathcal{S}_i$ of the alphabet $\mathcal{Y}$, and assigning a representative distribution $a_i \in \mathcal{P}^{|\mathcal{X}|-1}$ for each quantization cell S_i, such that $q_y = a_i$ iff $y \in S_i$. Note that once the sets S_i, $i = 1, \ldots, M$ are determined, the reconstructions that minimize $D = \mathbb{E}d(Y, q_Y)$ are given by $a_i = P_{X|Y\in\mathcal{S}_i}$. To see this, let $f : \mathcal{Y} \to [M]$ be such that $f(y) = i$ if $i \in \mathcal{S}_i$, set $T = f(Y)$, and write

$$\begin{aligned}
D &= \mathbb{E}_{XY}\left[\log\left(\frac{1}{q_Y(X)}\right)\right] \\
&= \mathbb{E}_{XT}\left[\log\left(\frac{1}{a_T(X)}\right)\right] \\
&= \mathbb{E}_T\left[\mathbb{E}\left[\log\left(\frac{1}{P_{X|T}(X|T)}\frac{P_{X|T}(X|T)}{a_T(X)}\right)\middle|T\right]\right] \\
&= H(X|T) + D\big(P_{X|T}\big\|a_T\big|P_T\big) \\
&\geq H(X|T) \\
&= H(X|f(Y)),
\end{aligned} \tag{107}$$

with equality if and only if $a_t = P_{X|T=t} = P_{X|Y\in\mathcal{S}_t}$ for all $t \in [M]$.

It follows that, for a given distribution P_{XY}, the design of the optimal quantizer under the distortion measure (106) reduces to finding $f : \mathcal{Y} \to [M]$ which minimizes $H(X|f(Y))$. Clearly, determining the minimum value of $H(X|f(Y))$ is equivalent to our maximization problem (1).

A quantity closely related to $I(X;[Y]_M)$ is the information bottleneck tradeoff [27], defined as

$$\mathrm{IB}_R(P_{XY}) \triangleq \max_{P_{T|Y}\,:\, I(Y;T)\leq R} I(X;T), \tag{108}$$

which has been extensively studied in the machine learning literature, see e.g. [28]–[30]. There, Y is thought of as an high-dimensional observation containing information about X, that must be first "compressed" to a simpler representation before inference can be efficiently performed. The random variable $T = f(Y)$ represents a clustering operation, where for the task of inferring X, all members in the cluster are treated as indistinguishable. A major difference, however, between the information bottleneck formulation and that of (1) is that the latter restricts $|f(\cdot)|$ to M, whereas the former allows for random quantizers and restricts the compression rate $I(T;Y)$. The discussion above indicates that the problem (1) is a standard quantization/lossy compression problem (or more precisely, a remote source coding problem). As such, its fundamental limit admits a single-letter solution[6] and we have that [24], [31]

$$\lim_{n\to\infty}\frac{1}{n} I(X^n;[Y^n]_{M^n}) = \mathrm{IB}_{\log M}(P_{XY}). \tag{109}$$

[6]One subtle point to be noted is that the relevant distortion measure for $I(X^n;[Y]^n_{M^n})$ is not separable. Nevertheless, it is not difficult to show that restricting the reconstruction distribution to the form $q_{y^n}(x^n) = \prod_{i=1}^n q^i_{y^n}(x_i)$ entails no loss asymptotically.

where $P_{X^nY^n} = P_{XY}^{\otimes n}$ and $[Y^n]_{M^n}$ refers to the set of all M^n-quantizations of Y^n. That is, in the asymptotic limit, our problem (1) corresponds to an information bottleneck problem. However, the scalar setting $n = 1$ is of major importance as inference is seldom performed in blocks consisting of multiple independent samples from P_{XY}. Overall, our results in the previous sections indicate that when $I(X;Y)$ is small we may need at least $\Theta(\log(1/I(X;Y))$ clusters to guarantee that we retain a significant fraction of the original information.

A. On the Gap Between Scalar Quantization and Information Bottleneck

In this subsection, we show that in the limit $I(X;Y) \to 0$, the restriction to using a scalar quantizer results in a significantly worse performance than the one predicted by the information bottleneck, which implicitly assumes quantization is performed in asymptotically large blocks. In particular, we prove the following theorem.

Theorem 6: For any P_{XY} with $|\mathcal{X}| = 2$ and $I(X;Y) = \beta$, and any $\eta \in (0,1)$ there exists a quantizer $f(Y)$ such that $I(X;f(Y)) \geq \eta\beta$ and

$$H(f(Y)) \leq \log\log\log\left(\frac{1}{\beta}\right) - 2\log(1-\eta) + 11. \tag{110}$$

Contrasting this with Theorem 4, and its simplification in (70), which show that there exist distributions P_{XY} with $|\mathcal{X}| = 2$, for which no scalar quantizer with less than $\log\log(1/\beta) + \log(\eta) + 1$ bits can attain $I(X;f(Y)) > \eta\beta$, we see that the restriction to quantization in blocklength $n = 1$ entails a significant cost w.r.t. quantization in long blocks. In particular, if for a distribution P_{XY} there exists a quantizer $f(Y)$ with entropy $H(f(Y)) = R$ for which $I(X;f(Y)) = \Gamma$, then certainly $\mathrm{IB}_R(P_{XY}) \geq \Gamma$. To see this just take $T = f(Y)$ in (108).[7] It therefore follows from Theorem 4 and Theorem 6 that the information bottleneck tradeoff may be over-optimistic in predicting the performance of optimal scalar quantization.

Proof. In the proof of the lower bound of Theorem 4, we have proposed the M-level quantizer (44) with the parameters specified by (46). For $M = \lfloor \frac{52}{1-\eta}\log\left(\frac{1}{\beta}\right)\rfloor$, and $\beta \leq \frac{1-\eta}{2}$, we have shown that this quantizer attains $I(X;f(Y)) \geq \eta\beta$. We will now show that for the same quantizer $H(f(Y)) = \mathcal{O}\left(\log\log(M)\right)$.

Let

$$P_\ell \triangleq \Pr\left(\{f(Y) = -\ell\} \cup \{f(Y) = \ell\}\right),\ \ell = 0, \ldots, L$$

and note that

$$H(f(Y)) \leq 1 + H(\{P_\ell\}). \tag{111}$$

Our goal is therefore to derive universal upper bounds on $H(\{P_\ell\})$ that hold for all joint distributions P_{XY} with $|\mathcal{X}| = 2$.

First, recall from the proof of Theorem 3 that

$$I(X;Y) = \mathbb{E}D_Y \geq \sum_{\ell=0}^{L} \gamma_\ell P_\ell = \gamma_1 \sum_{\ell=1}^{L} \theta^{\ell-1} P_\ell,$$

where we have used (46) in the last equality. We therefore have

$$\begin{aligned}
\sum_{\ell=0}^{L} \theta^\ell P_\ell &= P_0 + \sum_{\ell=1}^{L} \theta^\ell P_\ell \\
&\leq 1 + \frac{\theta I(X;Y)}{\gamma_1} \\
&= 1 + \frac{\theta}{(1-\eta)/2} \\
&\leq \frac{4\theta}{1-\eta}
\end{aligned} \tag{112}$$

[7] See [32] for an elaborate discussion on the information bottleneck tradeoff when T is restricted to be a deterministic quantizer of Y.

where in (112) we have used $\gamma_1 = \epsilon I(X;Y)$, due to (46), and $\epsilon = (1-\eta)/2$, due to (50).

For a vector $\mathbf{a} = \{a_0, a_1, \ldots, a_L\} \in \mathbb{R}_+^{L+1}$ and a scalar $\min_\ell\{a_\ell\} \le b \le \max_\ell\{a_\ell\}$, define the function

$$f(\mathbf{a}, b) \triangleq \max \sum_{\ell=0}^{L} P_\ell \log\left(\frac{1}{P_\ell}\right) \quad \text{subject to } \sum_{\ell=0}^{L} a_\ell P_\ell \le b,\ \sum_{\ell=0}^{L} P_\ell = 1. \tag{113}$$

The problem (113) is a concave maximization problem under linear constraints, and its solution is [33, p.228]

$$f(\mathbf{a}, b) = \min_{\lambda \ge 0} \lambda b + \log\left(\sum_{\ell=0}^{L} 2^{-\lambda a_\ell}\right). \tag{114}$$

Combining (112) and (114) with $a_\ell = \theta^\ell$ and $b = \frac{4\theta}{1-\eta}$, gives

$$H(\{P_\ell\}) \le \min_{\lambda \ge 0} \lambda \frac{4\theta}{1-\eta} + \log\left(\sum_{\ell=0}^{L} 2^{-\lambda\theta^\ell}\right). \tag{115}$$

Setting $\lambda = \frac{1}{L}$, gives

$$\begin{aligned} H(\{P_\ell\}) &\le \frac{4}{1-\eta}\cdot\frac{\theta}{L} + \log\left(\sum_{\ell=0}^{L} 2^{-\frac{\theta^\ell}{L}}\right) \\ &\le \frac{4}{1-\eta}\cdot\frac{\theta}{L} + \log\left(\sum_{\ell=0}^{\lfloor 2\frac{\log L}{\log\theta}\rfloor} 2^{-\frac{\theta^\ell}{L}} + \sum_{\ell=\lfloor 2\frac{\log L}{\log\theta}\rfloor+1}^{L} 2^{-\frac{\theta^\ell}{L}}\right) \\ &\le \frac{4}{1-\eta}\cdot\frac{\theta}{L} + \log\left(2\frac{\log L}{\log\theta} + 1 + L2^{-L}\right). \end{aligned}$$

Recalling that $1 < \theta < 2^{(1-\eta)/2}$, and noting that $1 + L2^{-L} < 2\log(L)$ for $L > 1$, we obtain

$$H(\{P_\ell\}) \le \frac{1}{L}\frac{2^{(1-\eta)/2}}{(1-\eta)/4} + \log\left(4\frac{\log L}{\log\theta}\right).$$

For the first term, we can use the definition of L in (50) to obtain

$$\begin{aligned} \frac{1}{L}\frac{2^{(1-\eta)/2}}{(1-\eta)/4} &\le \frac{(1-\eta)/2}{\log\left(\frac{2\kappa}{(1-\eta)I(X;Y)}\right)}\cdot\frac{2^{(1-\eta)/2}}{(1-\eta)/4} \\ &\le 2\cdot 2^{(1-\eta)/2} \\ &\le 3, \end{aligned} \tag{116}$$

where we have used the fact that $\kappa \ge 1$, and consequently $\log\left(\frac{2\kappa}{(1-\eta)I(X;Y)}\right) \ge 1$. For the second term, we have that

$$\begin{aligned} \log(\theta) &\overset{(i)}{\ge} \frac{1}{L}\log\left(\frac{\kappa}{\epsilon\beta}\right) \\ &\overset{(ii)}{\ge} \frac{(1-\eta)/4}{\log\left(\frac{\kappa}{\beta\cdot(1-\eta)/2}\right)}\log\left(\frac{\kappa}{\beta\cdot(1-\eta)/2}\right) \\ &= \frac{1-\eta}{4}. \end{aligned} \tag{117}$$

where (i) follows from (46) and (ii) follows from (50), where we have used

$$L = \left\lceil \frac{2}{1-\eta} \log\left(\frac{\kappa}{\beta \cdot (1-\eta)/2}\right) \right\rceil \leq \frac{4}{1-\eta} \log\left(\frac{\kappa}{\beta \cdot (1-\eta)/2}\right).$$

We have therefore obtained that

$$H(\{P_\ell\}) \leq 7 + \log\log L - \log(1-\eta).$$

Now, recalling that $L < M < \frac{52\log(1/\beta)}{1-\eta}$, we have that

$$\begin{aligned} H(\{P_\ell\}) &\leq 10 - \log\log(1-\eta) - \log(1-\eta) + \log\log\log\left(\frac{1}{\beta}\right) \\ &\leq 10 - 2\log(1-\eta) + \log\log\log\left(\frac{1}{\beta}\right). \end{aligned} \tag{118}$$

Now, applying (111) with (118) establishes the result. ■

ACKNOWLEDGMENT

The authors are grateful to Emre Telatar for pointing out the relation to the Knapsack problem, to Meir Feder for suggesting to compare the performance of scalar quantizers with that of scalar quantization followed by entropy coding for $f(Y)$, and to Guy Bresler, Robert Gray, Pablo Piantanida, and Shlomo Shamai (Shitz) for valuable discussions.

REFERENCES

[1] B. Nazer, O. Ordentlich, and Y. Polyanskiy, "Information-distilling quantizers," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 96–100.
[2] A. J. Viterbi and J. K. Omura, *Principles of digital communication and coding*. Courier Corporation, 2013.
[3] T. Koch and A. Lapidoth, "At low SNR, asymmetric quantizers are better," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5421–5445, September 2013.
[4] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.
[5] R. Pedarsani, S. H. Hassani, I. Tal, and E. Telatar, "On the construction of polar codes," in *Proceedings of the IEEE International Symposium on Information Theory*, July 2011, pp. 11–15.
[6] I. Tal, A. Sharov, and A. Vardy, "Constructing polar codes for non-binary alphabets and MACs," in *Proceedings of the IEEE International Symposium on Information Theory*, July 2012, pp. 2132–2136.
[7] A. Kartowsky and I. Tal, "Greedy-merge degrading has optimal power-law," *IEEE Transactions on Information Theory*, to appear 2019.
[8] I. Tal, "On the construction of polar codes for channels with moderate input alphabet sizes," in *Proceedings of the IEEE International Symposium on Information Theory*, June 2015, pp. 1297–1301.
[9] B. M. Kurkoski and H. Yagi, "Quantization of binary-input discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4544–4552, August 2014.
[10] D. Burshtein, V. D. Pietra, D. Kanevsky, and A. Nadas, "Minimum impurity partitions," *The Annals of Statistics*, vol. 20, no. 3, pp. 1637–1646, 1992.
[11] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1294–1309, July 2009.
[12] F. Cicalese, L. Gargano, and U. Vaccaro, "Bounds on the entropy of a function of a random variable and their applications," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2220–2230, April 2018.
[13] H. Kellerer, U. Pferschy, and D. Pisinger, *Introduction to NP-Completeness of Knapsack Problems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 483–493.
[14] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *CoRR*, vol. abs/1304.6133, 2013. [Online]. Available: http://arxiv.org/abs/1304.6133
[15] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*, E. Carlen, M. Madiman, and E. M. Werner, Eds. Springer, 2017, pp. 211–249.
[16] ——, "Lecture notes on information theory," *MIT (6.441), UIUC (ECE 563)*, 2016.
[17] W. Huleihel, M. Brennan, and G. Bresler, "Quantizations preserving Kullback-Leibler divergence," in *Proceedings of the International Zurich Seminar on Information and Communication (IZS 2018) Proceedings*. ETH Zurich, February 2018, p. 61.
[18] C. Calabro, "The exponential complexity of satisfiability problems," Ph.D. dissertation, UC San Diego, 2009.

[19] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, October 1998.

[20] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, September 1962.

[21] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[22] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *studia scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.

[23] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, October 1998.

[24] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[25] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman, "Justification of logarithmic loss via the benefit of side information," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5357–5365, October 2015.

[26] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[27] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *37th Annual Allerton Conference on Communications, Control, and Computing*, Monticello, IL, 1999, pp. 368–377.

[28] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000, pp. 208–215.

[29] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, 2015, pp. 1–5.

[30] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[31] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 595–609.

[32] D. Strouse and D. J. Schwab, "The deterministic information bottleneck," *arXiv preprint arXiv:1604.00268*, 2016.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.