Information-Distilling Quantizers

Bobak Nazer
BU
bobak@bu.edu

Or Ordentlich
BU/MIT
ordent@mit.edu

Yury Polyanskiy
MIT
yp@mit.edu

***Abstract*—Let X and Y be dependent random variables. We consider the problem of designing a scalar quantizer for Y to maximize the mutual information between its output and X. We study fundamental properties and bounds for this form of quantization, which is connected to the log-loss distortion criterion. Our main focus is the regime of low $I(X;Y)$, where we show that for a binary X, there always exists an M-level quantizer attaining mutual information of $\Omega(-M\cdot I(X;Y)/\log(I(X;Y))$ and that there exists pairs of X,Y for which the mutual information attained by any M-level quantizer is $\mathcal{O}(-M\cdot I(X;Y)/\log(I(X;Y)))$.**

I. Introduction

Quantization plays a central role in many information processing systems. For instance, when the data comes from a continuous alphabet, quantization is a pre-requisite for digital processing. However, even if the data comes from a discrete alphabet, reducing its cardinality often leads to more efficient processing.

Let X and Y be a pair of random variables with 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. Thus, the optimal quantizer under this setup is

$$\underset{f:\mathcal{Y}\to[M]}{\operatorname{argsup}}\ I(X;f(Y)), \tag{1}$$

where $[M] \triangleq \{1,2,\ldots,M\}$. We will use the following shorthand[1] to denote the value of the mutual information attained by the optimal M-ary quantizer.

$$I(X;[Y]_M) \triangleq \sup_{\tilde{Y}\in[Y]_M} I(X;\tilde{Y}). \tag{2}$$

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

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

When X and Y are thought of as the input and output of a channel, respectively, the problem (1) boils down to designing the M-level quantizer that maximizes the information rate, whereas (2) is the highest information rate attainable. It is therefore not surprising that this problem has received considerable attention. For example, it is well known [1, Section 2.11] that when X is a BPSK input to an AWGN channel with output Y it holds that $I(X;[Y]_2) \geq 2I(X;Y)/\pi$ and this is achieved by taking $f(\cdot)$ to be the maximum a posteriori (MAP) estimator of X from Y.[2]

A characterization of (2) is also required for practically constructing polar codes, 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 were developed to overcome this obstacle, and played a major role in the process of making polar codes implementable [3]–[5]. Specifically, it was recently shown in [5] that, for arbitrary P_{XY}, it holds that $I(X;Y) - I(X;[Y]_M) \leq \mathcal{O}(M^{-2/(|\mathcal{X}|-1)})$. The works [3]–[5], among others, also 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 [6] (drawing upon a result from [7]) that runs in time $\mathcal{O}(|\mathcal{Y}|^3)$. A supervised learning algorithm, for the scenario where P_{XY} is not known, was proposed in [8].

In this paper, we ignore the algorithmic aspects of finding the optimal M-level quantizer and instead focus on the fundamental properties of the function $I(X;[Y]_M)$. In particular, our main interest is in identifying the joint distributions P_{XY} that are the most difficult to quantize, and in the value of $I(X;[Y]_M)$ for these cases. Special attention will be given to the binary case,[3] where $X \sim \text{Bernoulli}(p)$ 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. This is also the case for the binary AWGN channel for all values of $I(X;Y)$, since the MAP quantizer always retains at least $2/\pi \approx 63.66\%$ of the mutual information.

We state our main result next, with proof deferred to Section III-C. Logarithms are generally taken w.r.t. base 2 in this paper, with the exception of the $\ln$ function that is taken w.r.t. base e.

Theorem 1: If $X \sim \text{Bernoulli}(1/2)$ and $I(X;Y) = \beta > 0$,

The work was supported in part 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 grant CCF-1618800.

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

[2] It was recently demonstrated in [2] that, if instead of BPSK, an asymmetric signaling scheme is used, the AWGN capacity can be attained at low SNR with an asymmetric 2-level quantizer.

[3] Results for general finite input alphabets will be reported in an extended version.

we have for binary quantization

$$I(X;[Y]_2) \geq \frac{1}{3e}\frac{\beta}{1+\ln\left(\frac{1}{\beta}\right)}. \quad (3)$$

Furthermore, for any $\eta \in (0,1)$ and any natural $M < \frac{12\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}{(1-\eta)^2}$

$$I(X;[Y]_M) \geq (M-1)\frac{\beta}{\max\{\log\left(\frac{1}{\beta}\right),1\}}\frac{\eta(1-\eta)^2}{12}. \quad (4)$$

Finally, for any $0 < \beta \leq 1$, there exist distributions P_{XY} with $X \sim \text{Bernoulli}(1/2)$ and $I(X;Y) = \beta$, for which

$$I(X;[Y]_M) \leq 2M\frac{\beta}{\ln\left(\frac{e\log(e)}{2\beta}\right)}, \quad (5)$$

for every natural M.

Note that this is in stark contrast to the intuition from the binary AWGN channel. While for the former, two quantization levels suffice for retaining a $2/\pi$ fraction of $I(X;Y)$, Theorem 1 shows that there exist 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”.

For a fixed distribution P_X on $\mathcal{X}$, we define and study the “information distillation” function

$$\text{ID}_M(P_X,\beta) \triangleq \inf_{P_{Y|X}\,:\,I(X;Y)\geq\beta} I(X;[Y]_M), \quad (6)$$

where the infimum is taken w.r.t. to all channels with input alphabet $\mathcal{X}$ and arbitrary (possibly continuous) output alphabet such that the mutual information is at least β. With this notation, Theorem 1 states that $\text{ID}_M(\text{Bernoulli}(1/2),\beta) = \Theta\left(M\beta/\log(1/\beta)\right)$, and in fact, as briefly argued in Section III-C, the same scaling law continues to hold for $\text{ID}_M(\text{Bernoulli}(p),\beta)$, $0<p<1$.

As discussed above, prior work [3]–[5] has focused on bounding the additive gap. In our notation, this corresponds to bounding

$$\Delta I_M^* \triangleq \sup_{\beta,P_X} \beta - \text{ID}_M(P_X,\beta).$$

In particular, the bound derived in [5] on ΔI_M^* is equivalent to the following “constant-gap” result: for every P_X, $\text{ID}_M(P_X,\beta) \geq \beta - \nu(|\mathcal{X}|)M^{-2/|(\mathcal{X}|-1)}$ for some function ν.[4] For small β, however, results of this form are less informative. Indeed, for binary-input channels and small β, this bound requires M to scale like $\beta^{-1/2}$ in order to preserve a constant fraction of the mutual information. On the other hand, our result shows that scaling M like $\mathcal{O}(\log(1/\beta))$ suffices for all binary-input channels.

A. Connection to Quantization Under Log-Loss

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., [10]. The performance of the quantizer is measured w.r.t. 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.

Assume the quantizer’s output q_Y should allow to infer information about a correlated random variable X with alphabet $\mathcal{X}$. In this case, 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 $\mathcal{P}$. 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 is the following:

$$\mathbb{E}_{XY}\log\frac{P_{X|Y}(X|Y)}{q_Y(X)} = \mathbb{E}_Y\mathbb{E}\left[\log\frac{1}{q_Y(X)}\middle| Y\right] - H(X|Y),$$

which is related to the *log-loss* criterion. Since the term $H(X|Y)$ is independent of the quantizer, our goal is to design a quantizer q_Y that minimizes $D = \mathbb{E}_Y d(Y,q_Y)$ where

$$d(y,P) \triangleq \mathbb{E}\left[\log\frac{1}{P(X)}\middle| Y=y\right],\ \forall (y,P)\in\mathcal{Y}\times\mathcal{P}. \quad (7)$$

Note that once the sets S_i, $i=1,\ldots,M$ are determined, the reconstructions that minimize D are given by $a_i = P_{X|Y\in\mathcal{S}_i}$. To see this, let $T\in[M]$ denote the cell in which Y fell, and write

$$\begin{aligned}
D &= \mathbb{E}_Y\mathbb{E}\left[\log\frac{1}{q_Y(X)}\middle| Y\right]\\
&= \mathbb{E}_T\mathbb{E}\left[\log\frac{1}{a_T(X)}\middle| Y\in\mathcal{S}_T\right]\\
&= \mathbb{E}_T\mathbb{E}\left[\log\frac{1}{P_{X|Y\in\mathcal{S}_T}(X)}\frac{P_{X|Y\in\mathcal{S}_T}(X)}{a_T(X)}\middle| Y\in\mathcal{S}_T\right]\\
&= H(X|T) + D(P_{X|T}||a_T|P_T)\\
&\geq H(X|T),
\end{aligned}$$

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

For a given distribution P_{XY}, the design of the optimal quantizer under the distortion measure (7) reduces to finding $f:\mathcal{Y}\to[M]$ which minimizes $H(X|f(Y))$, which is in turn equivalent to solving (1).

A quantity closely related to $I(X;[Y]_M)$ is the information bottleneck tradeoff [11], defined as

$$\text{IB}_R(P_{XY}) \triangleq \max_{P_{T|Y}\,:\,I(Y;T)\leq R} I(X;T). \quad (8)$$

which has been extensively studied in the machine learning literature. 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

[4] It is also demonstrated in [9] that there exist values of β, for which this bound is tight. Specifically, [9] 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 $c>0$.

be 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 (2) 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 (2) is a standard quantization/lossy compression problem (or more precisely, a remote source coding problem [12]). As such, its fundamental limit admits a single-letter solution[5] and we have that [13], [14]

$$\lim_{n\to\infty}\frac{1}{n}I(X^n;[Y^n]_{M^n}) = \mathrm{IB}_{\log M}(P_{XY}). \tag{9}$$

where $P_{X^nY^n} = P^n_{X,Y}$ and $[Y^n]_{M^n}$ refers to the set of all M^n-quantizations of Y^n. In practice, the case $n = 1$ is of major importance as inference is seldom performed in blocks of independent observations. Thus, our results indicates that when $I(X;Y)$ is small we may need at least $\Theta(\log(1/I(X;Y))$ clusters to guarantee that we use a significant fraction of the information in the observations.

II. Properties of $I(X;[Y]_M)$ and $\mathrm{ID}_M(P_X,\beta)$

Let P_{XY} be a joint distribution on $\mathcal{X}\times\mathcal{Y}$ and consider the function $I(X;[Y]_M)$, as defined in (2). The restriction to deterministic functions incurs no loss of generality, see e.g., [6]. 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,

$$I(X;f(Y,U)) \le I(X;f(Y,U),U) = I(X;f(Y,U)|U) \tag{10}$$

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 $\mathcal{I}_1,\dots,\mathcal{I}_M$, such that $f(y)=i$ iff $P_{X|Y=y}\in\mathcal{I}_i$ for $i=1,\dots,M$. A remarkable result of Burshtein et al. [7, Theorem 1] shows that the supremum in (2) can w.l.o.g. be restricted to functions for which there exists an associated partition where the regions $\mathcal{I}_1,\dots,\mathcal{I}_M$ 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 $\{1,\dots,|\mathcal{Y}|\}$ 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

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

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

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

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 (10), we can always replace $\tilde{f}$ by some deterministic function $\bar{f}:\mathcal{Y}\mapsto[M]$ such that

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

■

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

[5] 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.

the input is X. From the convexity of the mutual information w.r.t. the channel we have

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

as desired. ■

Remark 1: 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 2: 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 $I(X;[Y]_{M_1\cdot M_2})\le I(X;[Y]_{M_1})+I(X;[Y]_{M_2})$. However, as demonstrated by the following example, this is not the case. Let $X\sim\text{Uniform}(\{0,1,2,3\})$ 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,

$$I(X;[Y]_4)=I(X;Y)=2-h(\delta)-(1-\delta)\log(3),\tag{11}$$

and it can be verified that

$$I(X;[Y]_2)=\begin{cases}h\left(\frac{1}{4}\right)-\frac{1}{4}h(\delta)-\frac{3}{4}h\left(\frac{1-\delta}{3}\right) & \delta\le 1/4,\\ 1-h\left(\frac{1+2\delta}{3}\right) & \delta>1/4.\end{cases}\tag{12}$$

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

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

$$H([Y]_M)\triangleq\sup_{\tilde{Y}\in[Y]_M}H(\tilde{Y}).\tag{13}$$

Furthermore, when $M=2$ the optimization problem in (13) is equivalent to

$$\max_{\mathcal{A}\subseteq\mathcal{X}}\sum_{x\in\mathcal{A}}p_x\text{ subject to: }\sum_{x\in\mathcal{A}}p_x\le\frac{1}{2},\tag{14}$$

where $p_x\triangleq\Pr(X=x)$, $x\in\mathcal{X}$. The problem (14) is known as the *subset sum problem* and is NP-hard [15]. Nevertheless, for some special instances optimal low-complexity algorithms do exist. For example, if $\mathcal{X}$ is binary, a dynamic programming algorithm finds the optimal quantizer with complexity $\mathcal{O}(|\mathcal{Y}|^3)$, see [6].

Proposition 4: The function $\text{ID}_M(P_X,\beta)$ is convex and monotonically nondecreasing in β.

Proof. Monotonicity follows from definition. For convexity, let $\beta_1,\beta_2\ge 0$ and $\alpha\in[0,1]$, and let $P_{Y_i|X}$ be the channel that attains[6] the infimum in (6) for β_i. Consider a channel with output $Y=(Y_U,U)$, where $U=1$ with probability α and $U=2$ with probability $1-\alpha$, statistically independent of X. We have

$$I(X;Y)=I(X;Y_U,U)=I(X;Y_U|U)\ge\alpha\beta_1+\alpha\beta_2.$$

On the other hand, for any $f:\mathcal{Y}\mapsto[M]$,

$$\begin{aligned}I(X;f(Y))&=I(X;f(Y_U,U))\\&\le I(X;f(Y_U,U),U)\\&=I(X;f(Y_U,U)|U)\\&\le\alpha I(X;[Y_1]_M)+(1-\alpha)I(X;[Y_2]_M)\\&=\alpha\text{ID}_M(P_X,\beta_1)+(1-\alpha)\text{ID}_M(P_X,\beta_2),\end{aligned}$$

which shows that

$$\begin{aligned}&\text{ID}_M(P_X,\alpha\beta_1+(1-\alpha)\beta_2)\\&\le\alpha\text{ID}_M(P_X,\beta_1)+(1-\alpha)\text{ID}_M(P_X,\beta_2),\end{aligned}\tag{15}$$

as desired. ■

Remark 4 (Relations to quantization for maximizing divergence): For two distributions P,Q on $\mathcal{Y}$, $Q\ll P$, define

$$\psi_M(P,Q)\triangleq\sup_{f:\mathcal{Y}\mapsto[M]}D(P^f||Q^f),\tag{16}$$

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 $\psi_M(P,Q)\le M\epsilon$ for any $M\in\mathbb{N}$.

Proof. Consider the following two distributions:

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

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

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

$$D(P||Q)=\sum_{m=k+1}^{T}2^{-m}\log(1/g(m)),\tag{17}$$

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

$$D(P^f||Q^f)\le M\cdot\max_{A\subset\{0,1,\ldots\}}P(A)\log\frac{P(A)}{Q(A)}.\tag{18}$$

[6] More precisely, this refers to a sequence of channels approaching the infimum.

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

$$\begin{aligned} P(A) &= P(A_k) + P(A \setminus A_k) \leq P(A_k) + 2 \cdot 2^{-\ell} \\ Q(A) &= Q(A_k) + Q(A \setminus A_k) \geq P(A_k) + 2^{-\ell} g(\ell) \end{aligned} \tag{19}$$

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

$$P(A) \log \frac{P(A)}{Q(A)} \leq -2^{-\ell} t \log\left(1 - \frac{\tau}{t}\right). \tag{20}$$

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 (20) is maximized by choosing A such that $P(A_k) = 0$, for which $t = 2$, and we obtain

$$D(P^f \| Q^f) < M \cdot 2^{-(\ell - 1)} \log \frac{2}{g(\ell)}. \tag{21}$$

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 $m = 1, 2, \ldots$. We have

$$D(P\|Q) = \alpha \sum_{m=k+1}^{T} \frac{1}{m} \tag{22}$$

$$D(P^f \| Q^f) \leq 2M\left(2^{-\ell} + \frac{\alpha}{\ell}\right) \leq 2M\left(2^{-k} + \frac{\alpha}{k}\right). \tag{23}$$

The statement follows by noting that we can always choose k such that the LHS of (23) is smaller than ϵ, and then we can choose T and α such that the LHS of (22) is equal to β. ■

Proposition 5 shows that for any fixed M, and any value od $D(P\|Q)$, the ratio $\psi_M(P,Q)/D(P,Q)$ can be arbitrarily small. 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 $f(y) = \mathrm{sign}(P(y) - Q(y))$ achieves $d_{\mathrm{TV}}(P^f, Q^f) = d_{\mathrm{TV}}(P,Q)$ for any pair of distributions P, Q on $\mathcal{Y}$. An interesting question is under what φ-divergences is the ratio $\psi_M(P,Q)/D_{\varphi}(P,Q)$ always positive.

III. Bounds for $X \sim \mathrm{Bernoulli}(1/2)$

In this section we provide upper and lower bounds on $\mathrm{ID}_M(P_X, \beta)$, for the special case where $X \sim \mathrm{Bernoulli}(p)$, which we denote by $\mathrm{ID}_M(p,\beta)$. To simplify derivations, we shall further restrict attention to $p = 1/2$, though the results we obtain below remain valid for any $0 < p < 1$ with some correction terms which are qualitatively insignificant. Clearly, for any distribution P_{XY} with $\mathcal{X} = \{1,2\}$ it holds that $I(X;Y) \leq 1$. Thus, β is restricted to the interval $[0,1]$.

A. The Symmetric Quantizer for $M = 2$

We begin by analyzing the mutual information induced by the most natural binary quantizer, which is based on the maximum a posteriori (MAP) estimator

$$f_{\mathrm{MAP}}(y) = \begin{cases} 1 & \text{if } \Pr(X=1|Y=y) > 1/2 \\ 2 & \text{if } \Pr(X=1|Y=y) < 1/2 \\ \mathrm{Bernoulli}(1/2) & \text{if } \Pr(X=1|Y=y) = 1/2 \end{cases}.$$

Let $P_{e,\mathrm{MAP}}(y) \triangleq \Pr(f_{\mathrm{MAP}}(Y) \neq X | Y = y)$ and $P_{e,\mathrm{MAP}} \triangleq \mathbb{E}_Y P_{e,\mathrm{MAP}}(Y)$. By concavity of the binary entropy function $h(p) \triangleq -p\log(p) - (1-p)\log(1-p)$ we have that $h(p) \geq 2p$ for any $0 \leq p \leq 1/2$, with equality iff $p \in \{0, 1/2\}$. Consequently,

$$H(X|Y) = \mathbb{E}_Y h(P_{e,\mathrm{MAP}}(Y)) \geq 2P_{e,\mathrm{MAP}}. \tag{24}$$

We therefore have that

$$\begin{aligned} &I(X; f_{\mathrm{MAP}}(Y)) = H(X) - H(X|f_{\mathrm{MAP}}(Y)) \\ &= 1 - \mathbb{E}_Y h\left(\Pr(X \neq f_{\mathrm{MAP}}(Y))\right) \\ &\geq 1 - h(P_{e,\mathrm{MAP}}) \\ &\geq 1 - h\left(\frac{H(X|Y)}{2}\right) \\ &= 1 - h\left(\frac{1 - I(X;Y)}{2}\right). \end{aligned} \tag{25}$$

B. The High Mutual Information Regime

As a consequence of (25), we obtained

$$\mathrm{ID}_2(1/2, \beta) \geq 1 - h\left(\frac{1-\beta}{2}\right). \tag{26}$$

Furthermore, note that the bound (25) is achieved with equality when $P_{Y|X}$ is the binary erasure channel (BEC). For the BEC, however, the MAP quantizer involves randomness. Instead of flipping a coin whenever $y =?$, we can always assign a fixed value, say $f(?) = 1$, to it. This deterministic asymmetric quantizer, given by

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

induces a Z-channel from X to $f_Z(Y)$ and satisfies

$$I(X; f_Z(Y)) = \frac{\beta}{2} h\left(\frac{1-\beta}{2-\beta}\right) + 1 - h\left(\frac{1-\beta}{2-\beta}\right). \tag{27}$$

Clearly, $f_Z(y)$ is the optimal 1-bit quantizer for the BEC.

We have therefore established the following proposition.

Proposition 6: For all $0 \leq \epsilon \leq 1$ we have

$$1 - h\left(\frac{\epsilon}{2}\right) \leq \mathrm{ID}_2(1/2, 1-\epsilon) \leq 1 - \frac{1+\epsilon}{2} h\left(\frac{\epsilon}{1+\epsilon}\right). \tag{28}$$

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.

C. The Low Mutual Information Regime

In the small β regime, we arrive at qualitatively different behavior. Consider again the BEC with capacity β for $\beta \ll 1$. By (25) (which becomes an equality for the BEC) and (27), we have that

$$I(X;f_{\text{MAP}}(Y)) = 1-h\left(\frac{1-\beta}{2}\right) = \frac{\log e}{2}\beta^2 + o(\beta^2). \tag{29}$$

$$I(X;f_Z(Y)) = \frac{\beta}{2}h\left(\frac{1-\beta}{2-\beta}\right) + 1 - h\left(\frac{1-\beta}{2-\beta}\right) = \frac{\beta}{2} + o(\beta). \tag{30}$$

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.

Remark 5: One can naively attribute this effect to the randomness required by the MAP quantizer in the BEC setting. This is not the case however. To see this consider a channel with binary 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 $0 \leq \beta \leq 1$ and $0 \leq \delta \leq 1/2$. 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 (29). 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 (30) as $\delta \to 0$.

Next, we prove Theorem 1. The proof will require the following proposition, which can be verified using straightforward analysis.

Proposition 7: The function $g(t) = -t\ln(t)$ is monotone increasing in $0 < t < 1/e$ and its inverse restricted to this interval satisfies

$$\frac{1}{e}\cdot\frac{t}{-\ln(t)} < g^{-1}(t) \leq \frac{t}{-\ln(t)} \tag{31}$$

Proof of lower bounds in Theorem 1. Consider the joint distribution P_{XY}, and for any $y \in \mathcal{Y}$ define $\alpha_y \triangleq \Pr(X = 1|Y=y)$, $\bar{\alpha} \triangleq \mathbb{E}(\alpha_Y) = \frac{1}{2}$ and

$$D_y \triangleq D(P_{X|Y=y}\|P_X) = d\left(\alpha_y \,\|\, \bar{\alpha}\right), \tag{32}$$

where $d(p\|q) \triangleq p\log(p/q) + (1-p)\log((1-p)/(1-q))$ is the binary divergence function. We further define the function

$$\bar{F}(\gamma) \triangleq \Pr(D_Y \geq \gamma), \tag{33}$$

and note that it is non-increasing and satisfies

$$I(X;Y) = \mathbb{E}D_Y = \int_0^{\gamma^*} \bar{F}(\gamma)d\gamma, \tag{34}$$

where $\gamma^* = \max_{y\in\mathcal{Y}} D_y \leq 1$. Let $M = 2L+1$ for some natural number L, 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 M-level quantizer

$$f(y) = \begin{cases} 0 & d(\alpha_y\|\bar{\alpha}) \leq \gamma_1 \\ -\ell & \alpha_y < \bar{\alpha}, \gamma_\ell \leq d(\alpha_y\|\bar{\alpha}) < \gamma_{\ell+1} \\ \ell & \alpha_y > \bar{\alpha}, \gamma_\ell \leq d(\alpha_y\|\bar{\alpha}) < \gamma_{\ell+1} \end{cases}. \tag{35}$$

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,

$$\begin{aligned} I(X;f(Y)) &= \sum_{\ell=-L}^{L} \Pr(f(Y)=\ell)D(P_{X|f(Y)=\ell}\|P_X) \\ &\geq \sum_{\ell=1}^{L}\left(\bar{F}(\gamma_\ell) - \bar{F}(\gamma_{\ell+1})\right)\gamma_\ell \\ &= \bar{F}(\gamma_1)\gamma_1 + \sum_{\ell=2}^{L}\bar{F}(\gamma_\ell)(\gamma_\ell - \gamma_{\ell-1}) - \bar{F}(\gamma_{L+1})\gamma_L \\ &= \sum_{\ell=1}^{L}\bar{F}(\gamma_\ell)(\gamma_\ell-\gamma_{\ell-1}), \end{aligned} \tag{36}$$

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 (36).

For the special case of $L=1$, this reduces to $\gamma_1 = \operatorname{argmax}_\gamma \gamma\bar{F}(\gamma)$, and with this choice we have $I(X;f(Y)) = \max_\gamma \gamma\bar{F}(\gamma)$. Thus, $\bar{F}(\gamma) \leq \min\{1, I(X;f(Y))/\gamma\}$. Using the identity (34) with $\gamma^* \leq 1$, this yields

$$I(X;Y) \leq \int_0^{I(X;f(Y))} d\gamma + \int_{I(X;f(Y))}^{1} \frac{I(X;f(Y))}{\gamma}d\gamma \tag{37}$$

$$\begin{aligned} &= I(X;f(Y))\left(1+\ln\frac{1}{I(X;f(Y))}\right) \\ &= -e\frac{I(X;f(Y))}{e}\ln\left(\frac{I(X;f(Y))}{e}\right). \end{aligned} \tag{38}$$

Recalling that $L = 1$ corresponds to a quantizer with $M = 2L + 1 = 3$ levels and applying Proposition 7, we have therefore obtained

$$I(X;[Y]_3) \geq e \cdot g^{-1}\left(\frac{I(X;Y)}{e}\right) \tag{39}$$

$$\geq \frac{1}{e} \cdot \frac{I(X;Y)}{1+\ln\left(\frac{1}{I(X;Y)}\right)}. \tag{40}$$

Now, applying Corollary 1, yields (3).

For a general L, the problem of finding $\{\gamma_\ell\}$ such as to maximize (36) is more difficult. We therefore resort to a possibly suboptimal choice according to the rule

$$\gamma_1 = \epsilon I(X;Y),\ \theta = \gamma_1^{-\frac{1}{L}},\ \gamma_\ell = \gamma_1 \cdot \theta^{\ell-1}, \tag{41}$$

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

$$\gamma_{\ell+1} - \gamma_\ell = \theta\left(\gamma_\ell - \gamma_{\ell-1}\right),\ \ell = 1,\dots,L. \tag{42}$$

This implies that

$$\begin{aligned} I(X;Y) &= \sum_{\ell=0}^{L}\int_{\gamma_\ell}^{\gamma_{\ell+1}} \bar{F}(\gamma)d\gamma \\ &\leq \sum_{\ell=0}^{L}(\gamma_{\ell+1}-\gamma_\ell)\bar{F}(\gamma_\ell) \\ &= \gamma_1 + \theta\sum_{\ell=1}^{L}(\gamma_\ell-\gamma_{\ell-1})\bar{F}(\gamma_\ell) \\ &\leq \gamma_1 + \theta I(X;f(Y)). \end{aligned} \tag{43}$$

Now, setting $\epsilon = 1/(L+1)$ yields

$$\begin{aligned} I(X;f(Y)) &\geq (I(X;Y))^{\frac{L+1}{L}}\frac{L}{(1+L)^{\frac{L+1}{L}}} \\ &\geq (I(X;Y))^{\frac{L+1}{L}} \cdot \left(1-\frac{1}{\sqrt{L}}\right), \end{aligned} \tag{44}$$

where the last inequality is valid for every $L \geq 1$.

Substituting in

$$L = \left\lceil \frac{4\max\left\{\log\left(\frac{1}{I(X;Y)}\right),1\right\}}{(1-\eta)^2}\right\rceil, \tag{45}$$

it follows that

$$I(X;f(Y)) \geq 2^{-(1-\eta)^2/4}\left(\frac{1}{2}+\frac{\eta}{2}\right)I(X;Y) \geq \eta I(X;Y).$$

Since $M = 2L+1$ and $L \geq 4$, it follows that we can guarantee $I(X;f(Y)) \geq \eta I(X;Y)$ if

$$M = \left\lfloor \frac{12\max\left\{\log\left(\frac{1}{I(X;Y)}\right),1\right\}}{(1-\eta)^2}\right\rfloor \tag{46}$$

and thus $\mathrm{ID}_M(1/2,\beta) \geq \eta\beta$ for this choice of M as well. For smaller values of M, we can apply Corollary 1 to get

$$\mathrm{ID}_M(1/2,\beta) \geq \frac{M-1}{\left\lfloor\frac{12\max\left\{\log\left(\frac{1}{I(X;Y)}\right),1\right\}}{(1-\eta)^2}\right\rfloor}\eta\beta \tag{47}$$

$$\geq (M-1)\frac{\beta}{\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}\frac{\eta(1-\eta)^2}{12}. \tag{48}$$

■

Remark 6: The proof above only used the assumption that $X \sim \mathrm{Bernoulli}(1/2)$ (rather than $\mathrm{Bernoulli}(p)$ with general p) in order to bound $\gamma^* \leq 1$. The proof can be easily modified to deal with any p, in which case we have $\gamma^* \leq -\log(\min\{p,1-p\})$. This will require changing the integration limits in (37), and replacing the choice of θ in (41) with $\theta = (\gamma^*/\gamma_1)^{1/L}$.

Proof of upper bound in Theorem 1. It suffices to provide one distribution P_{XY} with $I(X;Y) \geq \beta$ for which no M-level quantizer achieves mutual information exceeding the RHS of (5). To this end, let $X \sim \mathrm{Bernoulli}(1/2)$ 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

$$f_T(t) = \begin{cases} r\delta(t) + \frac{4r}{(1-2t)^3} & 0^- < t \leq \frac{1-\sqrt{r}}{2} \\ 0 & \text{otherwise} \end{cases} \tag{49}$$

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 $\Pr(\alpha_Y = t|T = t) = \Pr(\alpha_Y = 1-t|T = t) = 1/2$.

By [7, Theorem 1], the optimal quantizer partitions the interval $[0,1]$ into M subintervals $\mathcal{I}_i = [\gamma_{i-1},\gamma_i)$ for $i = 1,\dots,M-1$ and $\mathcal{I}_M = [\gamma_{M-1},\gamma_M]$, where $0 = \gamma_0 < \gamma_1 < \dots < \gamma_M = 1$, and outputs $f(y) = i$ iff $\alpha_y \in \mathcal{I}_i$. We therefore have

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

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

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

where we have again used the symmetry of the random variable α_Y in the last equation. We have therefore obtained

$$\begin{aligned}
&I(X;f(Y))\\
&\le M \max_{0\le a\le b\le \frac{1}{2}} \Pr(a\le \alpha_Y\le b)\, d\left(\mathbb{E}[\alpha_Y|a\le \alpha_Y\le b] \,\|\, \frac{1}{2}\right)\\
&= \frac{M}{2} \max_{0\le a\le b\le \frac{1}{2}} \Pr(a\le T\le b)\, d\left(\mathbb{E}[T|a\le T\le b] \,\|\, \frac{1}{2}\right).\\
&= \frac{M}{2} \max_{0\le b\le \frac{1}{2}} \Pr(0\le T\le b)\, d\left(\mathbb{E}[T|0\le T\le b] \,\|\, \frac{1}{2}\right) \quad (50)
\end{aligned}$$

where the last equality follows since both terms are individually maximized by $a=0$. It can be verified that for any $0\le\rho\le\frac{1-\sqrt{r}}{2}$

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

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

$$\begin{aligned}
I(X;f(Y)) &\le \frac{M}{2}\cdot \max_{0\le b\le\frac{1-\sqrt{r}}{2}} r\cdot\frac{1-h(2b^2)}{(1-2b)^2}\\
&\le M\cdot\log(e) r, \qquad (51)
\end{aligned}$$

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

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

Applying Proposition 7, we have

$$r\le e g^{-1}\left(\frac{2I(X;Y)}{e\log(e)}\right)\le\frac{2I(X;Y)}{\log(e)}\frac{1}{\ln\left(\frac{e\log(e)}{2I(X;Y)}\right)} \qquad (52)$$

which gives

$$I(X;f(Y))\le 2M\frac{I(X;Y)}{\ln\left(\frac{e\log(e)}{2I(X;Y)}\right)}, \qquad (53)$$

for any M-level function f. ■

IV. Comparison with Information Bottleneck

In this section we show that the in the limit of $\beta\to 0$ 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 done in asymptotically large blocks. In particular, we prove the following theorem.

Theorem 2: If $X\sim\text{Bernoulli}(1/2)$ and $I(X;Y)=\beta>0$, then for any $\eta\in(0,1)$ there exist a quantizer $f(Y)$ such that $I(X;f(Y))\ge\eta\beta$ and

$$H(f(Y))\le\mathcal{O}\left(\log\log\log\left(\frac{1}{\beta}\right)-\log\log(1-\eta)\right). \qquad (54)$$

Contrasting this with Theorem 1 which shows that there exist distributions for which no scalar quantizer with less than $\Omega(\log\log(1/\beta)+\log\eta)$ 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 exist a quantizer $f(Y)$ with entropy $H(f(Y))=R$ for which $I(X;f(Y))=\Gamma$, then certainly $\text{IB}_R(P_{XY})\ge\Gamma$. To see this just take $T=f(Y)$ in (8).[7] It therefore follows from Theorem 1 and (2) that the information bottleneck tradeoff may be arbitrarily over-optimistic in predicting the performance of optimal scalar quantization.

Proof. in the proof of the lower bound of Theorem 1, we have proposed the M-level quantizer (35) with the parameters specified by (41). For $M=\mathcal{O}(\log(\frac{1}{\beta})/(1-\eta)^2)$, we have shown that this quantizer attains $I(X;f(Y))\ge\eta\beta$. We will now show that for the same quantizer $H(f(Y))\ll\log(M)$.

Let

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

and note that $H(f(Y))\le 1+H(\{P_\ell\})$. Our goal is therefore to derive universal upper bounds on $H(\{P_\ell\})$, that hold for all channels $P_{Y|X}$ and $X\sim\text{Bernoulli}(1/2)$.

First note that

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

where we have used (41) in the last equality. We therefore have

$$\sum_{\ell=1}^{L}\theta^\ell P_\ell\le\frac{\theta I(X;Y)}{\gamma_1}=(L+1)\theta, \qquad (55)$$

where we have used $\gamma_1=\epsilon I(X;Y)$ and $\epsilon=1/(L+1)$ in the last equality.

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

$$\begin{aligned}
f(\mathbf{a},b)\triangleq\max&\sum_{\ell=0}^{L}P_\ell\log\left(\frac{1}{P_\ell}\right)\\
&\text{subject to }\sum_{\ell=0}^{L}a_\ell P_\ell\le b,\ \sum_{\ell=1}^{L}P_\ell=1. \qquad (56)
\end{aligned}$$

[7] See [17] for an elaborate discussion on the information bottleneck tradeoff when T is restricted to be a deterministic quantizer of Y.

The problem (56) is a concave maximization problem under linear constraints, and its solution is [18]

$$f(\mathbf{a}, b) = \min_{\lambda \geq 0} \lambda b + \log\left(\sum_{\ell=0}^{L} 2^{-\lambda a_\ell}\right). \tag{57}$$

Combining (55) and (57) gives

$$H(\{P_\ell\}) \leq \min_{\lambda \geq 0} \lambda\theta(L+1) + \log\left(1 + \sum_{\ell=1}^{L} 2^{-\lambda\theta^{\ell}}\right). \tag{58}$$

Setting $\lambda = \theta^{-\kappa \log L}$, for some $\kappa > 0$ to be determined later, and recalling that $\theta > 1$, gives

$$\begin{aligned} H(\{P_\ell\}) &\leq (L+1)\theta^{1-\kappa\log L} + \log\left(1 + \sum_{\ell=1}^{L} 2^{-\theta^{\ell-\kappa\log L}}\right) \\ &\leq 2L\theta^{1-\kappa\log L} + \log\left(1 + 2\kappa\log L + \sum_{\ell=2\kappa\log L}^{L} 2^{-\theta^{\ell-\kappa\log L}}\right) \\ &\leq 2L\theta^{1-\kappa\log L} + \log\left(1 + 2\kappa\log L + L2^{-\theta^{\kappa\log L}}\right). \end{aligned}$$

Now, setting $\kappa = \frac{1}{\log\theta}$ gives

$$\begin{aligned} H(\{P_\ell\}) &\leq 2\theta + \log\left(1 + 2\frac{\log L}{\log\theta} + L2^{-L}\right) \\ &= \mathcal{O}\left(\theta + \log\left(\frac{\log L}{\log\theta}\right)\right). \end{aligned} \tag{59}$$

To complete the proof, recall that $L = (M-1)/2 = \mathcal{O}(\log(\frac{1}{\beta})/(1-\eta)^2)$, and that $\theta = (\frac{\beta}{L+1})^{-1/L} = \text{const.}$ ∎

Acknowledgment

The authors are grateful to Emre Telatar for pointing out the relation to the Knapsack problem, to Meir Feder for suggesting to consider entropy coding for $f(Y)$, and to Guy Bresler, Robert Gray, Pablo Piantanida, and Kevin Phillips for valuable discussions.

References

[1] A. J. Viterbi and J. K. Omura, *Principles of digital communication and coding*. Courier Corporation, 2013.

[2] T. Koch and A. Lapidoth, "At low snr, asymmetric quantizers are better," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5421–5445, Sept 2013.

[3] R. Pedarsani, S. H. Hassani, I. Tal, and E. Telatar, "On the construction of polar codes," in *2011 IEEE International Symposium on Information Theory Proceedings*, July 2011, pp. 11–15.

[4] I. Tal, A. Sharov, and A. Vardy, "Constructing polar codes for non-binary alphabets and macs," in *2012 IEEE International Symposium on Information Theory Proceedings*, July 2012, pp. 2132–2136.

[5] A. Kartowsky and I. Tal, "Greedy-merge degrading has optimal power-law," *arXiv preprint arXiv:1701.02119*, 2017.

[6] 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, Aug 2014.

[7] 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.

[8] 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.

[9] I. Tal, "On the construction of polar codes for channels with moderate input alphabet sizes," in *IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1297–1301.

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

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

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

[13] 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.

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

[15] H. Kellerer, U. Pferschy, and D. Pisinger, *Introduction to NP-Completeness of Knapsack Problems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 483–493.

[16] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *MIT (6.441), UIUC (ECE 563)*, 2016.

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

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