Quantization of Random Distributions under KL Divergence

Aviv Adler
EECS (MIT)
Cambridge, MA, USA
adlera@mit.edu

Jennifer Tang
EECS (MIT)
Cambridge, MA, USA
jstang@mit.edu

Yury Polyanskiy
EECS (MIT)
Cambridge, MA, USA
yp@mit.edu

***Abstract*—Consider the problem of representing a distribution π on a large alphabet of size k up to fidelity ε in Kullback-Leibler (KL) divergence. Heuristically, arguing as for quadratic loss in high dimension, one expects that about $(k/2)\log(1/\varepsilon)$ bits would be required. We show this intuition is correct by proving explicit non-asymptotic bounds for the minimal average distortion when π is randomly sampled from a symmetric Dirichlet prior on the simplex. Our method is to reduce the single-sample problem to the traditional setting of iid samples, but for a non-standard rate distortion question with the novel distortion measure $d(x,y) = x\log(x/y)$, which we call *divergence distortion*. Practically, our results advocate using a $x \mapsto x^{2/3}$ compander (for small x) followed by a uniform scalar quantizer for storing large-alphabet distributions.**

***Index Terms*—Compression, rate distortion, quantization, Kullback-Leibler divergence, Shannon Lower Bound**

I. Introduction and Motivation

Suppose one wants to compress iid data over a large alphabet $[k] \triangleq \{1,\dots,k\}$ sampled from a distribution π. The distribution is considered to be known to the compressor (because it possesses a very large corpus of data) and unknown to the decompressor. A natural two-step compression scheme would be to first describe the distribution π and then use an optimal (Huffman, arithmetic, etc) compressor for it. Since π is to be represented with finitely many bits, only an approximation $\hat{\pi}$ can be conveyed to the decompressor. It is well known that this incurs penalty $D_{\text{KL}}(\pi\|\hat{\pi})$ in compression length [1, Ch 5]. This motivates the following definition:

Definition 1. *Given a distribution W over the simplex $\mathcal{P}([k]) \triangleq \{\pi \in \mathbb{R}^k : \pi_j \geq 0, \sum_{j=1}^k \pi_j = 1\}$,*

$$M^*(W,\varepsilon) \triangleq \inf\{|\mathcal{Q}| : \mathbb{E}_{\pi\sim W}[\min_{Q\in\mathcal{Q}} D_{\text{KL}}(\pi\|Q)] \leq \varepsilon\}\,.$$

Here $\mathcal{Q}$ is a finite set of distributions; each π is mapped to the closest $Q \in \mathcal{Q}$ (in KL divergence), with the average divergence at most ε. The quantity $M^*(W,\varepsilon)$ is intimately connected with Bayes risk or average redundancy (expressible as mutual information) in the universal compression setting.

Proposition 1. *Suppose that $\pi \sim W$ and X^n is generated iid from π. Then for any n,*

$$I(\pi; X^n) \leq \inf_{\varepsilon} \{n\varepsilon + \log M^*(W,\varepsilon)\}\,.$$

When $n = 1$, the above is equality.

This work was supported, in part, by the USAF-MIT AI Accelerator and the NSF under Grant No. CCF-17-17842.

We omit the proof which is similar to [2], [3]. Known results on $I(\pi; X^n)$ imply lower bounds on $M^*(W,\varepsilon)$ (discussed in Section I-A). Our results can also be used for lossless two-stage codes like those in the MDL literature [4]–[7].

Perhaps the most natural prior W to consider, in the absence of other knowledge, is the uniform prior which belongs to the family of symmetric Dirichlet distributions[1] $\text{Dir}_k(\alpha)$ (specifically it is the member with $\alpha = 1$). Dirichlet distributions are a popular choice for modeling unknown discrete distributions in Bayesian statistics partly because they are the conjugate priors to multinomial distributions; they are used in many fields, including Dirichlet Processes [8] and other learning and estimation problems, such as in [9]–[11].

Theorem 1. *There exists a constant $c_1 > 0$ such that for all α, k and $\varepsilon > 0$,*

$$\log M^*(\text{Dir}_k(\alpha),\varepsilon) \geq \frac{k}{2}\log\frac{c_1}{\varepsilon + \frac{1}{2\alpha k}}\,.$$

Furthermore, there exists a $c_2 > 0$ such that for all $\varepsilon > 0$ and all α, k such that $\alpha k > 1$

$$\log M^*(\text{Dir}_k(\alpha),\varepsilon) \leq \frac{k}{2}\log\frac{c_2}{\varepsilon(1-(\alpha k)^{-1/3})^2}\,.$$

A practically important consequence of our analysis is that an almost optimal quantization of large-alphabet distributions is obtained by companding $\pi_j \mapsto f(s\pi_j)$ (for some constant s), followed by the uniform (scalar) quantizer and projection back onto the simplex. For $\alpha = 1$, our method recommends

$$f(x) = \begin{cases} c(x/\tau)^{2/3} & \text{for } x \leq \tau \\ (1-c)\big(1-\exp(-(x-\tau)/3)\big) + c & \text{for } x > \tau \end{cases}$$

where $\tau = 2.954$, $c = 0.664$.

Our approach is to reduce the problem of finding $|\mathcal{Q}|$ in Definition 1, a quantization problem at heart, to a rate-distortion problem. This turns a single-sample covering question into an iid multi-sample covering question.

Definition 2. *For distribution P_X over $\mathbb{R}_{\geq 0}$ and $D > 0$,*

$$R(P_X, D) \triangleq \inf_{P_{Y|X}} \{I(X;Y) : \mathbb{E}[Y] = \mathbb{E}[X],\, \mathbb{E}[d(X,Y)] \leq D, Y > 0\}\,, \quad (1)$$

where $d(x,y) \triangleq x\log(x/y)$ (for $x \geq 0$ and $y > 0$).

[1] Symmetric Dirichlet distributions only need one scalar parameter α.

We call $d(x, y)$ the *divergence distortion*. While divergence distortion is convex in both inputs and $d(x, y) = 0$ iff $x = y$, it has the unusual property of being negative when $x < y$ (with minimum at $x = y/e$). However, from $\mathbb{E}[Y] = \mathbb{E}[X]$ and convexity we still get $\mathbb{E}[d(X, Y)] \geq 0$.[2]

It turns out (Section V) that computing (1) with P_X being the Gamma distribution yields bounds on $M^*(W, \varepsilon)$ with $W = \mathrm{Dir}_k(\alpha)$ (and exponential P_X corresponds to uniform W).

Other work on quantizing probabilities includes [12], where they use the average L_p norm, and [13]. Particularly relevant is [14] which connects rate distortion to ε-nets, inspiring subsequent works on quantizing functions [15], [16].

A. Connection to Universal Compression

Much work has been done in the setting of universal compression on the asymptotic Bayes risk or average redundancy defined in [17], [18], mostly for the case where W is the least-favorable prior, the Jeffreys' prior (or $W = \mathrm{Dir}_k(1/2)$ for discrete alphabets) [19]. The dominant term of the Bayes risk is $\frac{k-1}{2} \log n$, where $k-1$ is the dimension of the parameter space and n is the number of samples [5], [20]; the other terms are constant or order $o(1)$ in n [6], [19], [21]–[23].

Based on the above results on Jeffreys' prior and Proposition 1, we conjecture that in general $\log M^*(W, \varepsilon)$ will have a dominant term of $\frac{k-1}{2} \log \frac{1}{\varepsilon}$ as $\varepsilon \to 0$. This is close to our upper bound from Theorem 1, especially for large alphabets (our bound behaves as $\frac{k}{2} \log \frac{1}{\varepsilon}$ as $\varepsilon \to 0$). An interesting direction for future work would be to close the gap between our upper and lower bound, using the conjectured asymptotic rate as a guide. The advantage of our results, however, is in yielding an explicit non-asymptotic upper bound on the Bayes risk.

B. Rate Distortion Background

1) Shannon Lower Bound (SLB): Rate distortion functions are difficult to compute in general. For many rate distortion problems, the SLB is used for lower-bounding and even approximating $R(D)$. Asymptotically as the distortion D goes to zero, the SLB can be tight [24]–[26].

2) Reconstruction Points: For rate distortion problems which quantize X to Y, points in the support of Y are called reconstruction points. The probabilities with which each X maps to each Y are the transition probabilities. These are optimal if they minimize $I(X; Y)$. Reconstruction points can be continuous, though for squared-error distortion the optimal reconstruction is discrete if X is not supported on the whole real line [27] or if the SLB is not met [28].

3) Computing Upper Bounds: For large distortions, we can numerically compute the rate distortion function using [28] or [29]. For small distortions, while not optimal, scalar quantization is useful for computing analytic upper bounds. In particular, [30] showed that quantizing to intervals is asymptotically optimal for the squared-error distortion measure.

4) Related Work: In [31], the authors look at distortion measures which are functions of the quotients X/Y. In [32] and [33], the authors solve the rate distortion function for $d(x, y) = |x - y|$ and $d(x, y) = |\log(x) - \log(y)|$ respectively.

[2]Note that without the constraint $\mathbb{E}[Y] = \mathbb{E}[X]$, $R(P_X, D)$ would be 0 trivially by increasing Y without bound.

C. Summary

In Section II we show the following lower bound on (1) for all continuous sources (i.e. described by density functions):

Theorem 2. *For any X with density $p_X(x)$, we have*

$$R(p_X, D) \geq \frac{1}{2} \log \left(\frac{1}{D} \right) + h(X) - \frac{1}{2} \log \mathbb{E}[X] + c$$

where $h(\cdot)$ is the differential entropy and c is a constant.[3]

In Section III we describe a certain method based on the technique of [30], for upper-bounding (1), as well as lemmas allowing easy manipulation of bounds using this method. In Section IV, we use the tools from Section III to give upper bounds for the following important sources:

Theorem 3. *For a uniform source $X \sim \mathrm{Unif}_{[a,b]}$,*

$$R(p_X, D) \leq \frac{1}{2} \log \left(\frac{1}{D} \right) + \frac{1}{2} \log \left(\frac{(b-a)^2}{b} \frac{9}{32} \right)$$

Theorem 4. *For $X \sim \mathrm{Exponential}(1)$,*

$$R(p_X, D) \leq \frac{1}{2} \log \left(\frac{1}{D} \right) + \frac{1}{2} \log (9) \leq \frac{1}{2} \log \left(\frac{1}{D} \right) + 1.1$$

Theorem 5. *For source $X \sim \mathrm{Gamma}(\alpha, \beta)$ for any α, β,*

$$R(p_X, D) \leq \frac{1}{2} \log \left(\frac{1}{D} \right) + O(1)$$

The result for uniform sources is used to compute the result for exponential and Gamma sources. For all of these, our upper and lower bounds are tight up to an additive constant. This shows that $\frac{1}{2} \log \left(\frac{1}{D} \right)$ is the correct rate of growth for these sources under divergence distortion. Equating D with ε approximately gives the size of $M^*(\mathrm{Dir}_k(\alpha), \varepsilon)$. We show this precisely in Section V, where we use Theorem 2 and Theorem 5 to derive Theorem 1.

II. Lower Bound

We first state a preliminary result showing properties of the optimal reconstruction scheme for (1): a) it is discrete, and b) the optimal reconstruction points y_i are the conditional expected value of the X's mapping to them. Due to space constraints we omit the proofs; our analysis of a) follows a similar analysis in [28], and b) is a consequence of the convexity of the distortion measure.

Proposition 2. *For any source probability density p_X where $\mathbb{E}[X] < \infty$, the optimal reconstruction for* (1) *is discrete, and*

$$y_i = \frac{\int p_X(x) q_{Y|X}(y_i|x)\, x\, dx}{\int p_X(x) q_{Y|X}(y_i|x)\, dx} = \mathbb{E}[X | X \text{ maps to } y_i] \quad (2)$$

where $q_{Y|X}(\cdot|\cdot)$ are the optimal transition probabilities for (1)*.*

Our method for finding a lower bound to (1) is to use the SLB. First, we simplify by combining the distortion measure and expected value constraints into one inequality:

$$\int_y \int_x q_Y(y) q_{X|Y}(x|y) \left(x \log \frac{x}{y} - x \right) dx\, dy \leq D - \mathbb{E}[Y]$$

[3]The $-(1/2) \log \mathbb{E}[X]$ term might seem strange in light of the intuition that scaling X up will make it harder to approximate with low distortion. However, scaling X up increases $h(X)$ and the net effect is positive.

Our new distortion for finding the SLB is $d_{\text{SLB}}(x,y) = x\log(x/y) - x$. We use the following proposition, which is a slight variation on a result by Berger in [34, Ch 4]:

Proposition 3. *Let $p(x)$ be a probability density on X and $d(\cdot,\cdot)$ be a distortion measure. Let $\mathcal{A}_\lambda$ be the set of all non-negative functions $\alpha_\lambda(x,y)$ satisfying*

$$c(y) = \int_0^\infty \alpha_\lambda(x,y)p(x)e^{-\lambda d(x,y)}dx \le 1$$

for all y. For all $D > 0$, suppose $\mathcal{B}$ is the set of conditional probabilities meeting $\mathbb{E}[d(X,Y)] \le D$. Then

$$R(p,D) \ge \inf_{q_{Y|X}\in\mathcal{B}} \sup_{\lambda\ge 0,\alpha_\lambda\in\mathcal{A}_\lambda} -\lambda D + \int_y\int_x q_{Y|X}(y|x)p(x)\log\alpha_\lambda(x,y)\,dx\,dy$$

We skip the proof, which is similar to the proof in [34].

Lemma 1. *For any $y > 0$ and any $\lambda > 0$,*

$$\int_0^\infty e^{-\lambda\left(x\log\frac{x}{y}-x\right)}dx \le \frac{3}{2}e^{\lambda y}\sqrt{\frac{y}{\lambda}\frac{2\pi}{1-2c}} + \frac{1}{\lambda} \qquad (3)$$

where $c = 2\log 2 - 3/2$.

We need a bound which holds as $\lambda\to\infty$. We will use a modified version of Laplace's method.

Proof Sketch. Let $f_y(x) \triangleq e^{-\lambda\left(x\log\frac{x}{y}-x\right)}$. We will show the bound (3) by splitting up the domain into three parts.

a) $x\in[0,2y)$: We use the Taylor expansion around y to find c which bounds $x\log(x/y)-x$ by $y+\frac{1}{2}\frac{(x-y)^2}{y}(1-2c)$ for $x\le 2y$, and get $\int_0^{2y} e^{-\lambda\left(x\log\frac{x}{y}-x\right)}dx \le e^{\lambda y}\sqrt{\frac{y}{\lambda}\frac{2\pi}{1-2c}}$.

b) $x\in[2y,3y)$: Since $f_y(x)$ decreases on $(y,3y)$, the integral of $f_y(x)$ on $[2y,3y]$ is at most the integral of $f_y(x)$ on $[y,2y]$. So $\int_0^{3y} e^{-\lambda\left(x\log\frac{x}{y}-x\right)}dx \le \frac{3}{2}e^{\lambda y}\sqrt{\frac{y}{\lambda}\frac{2\pi}{1-2c}}$.

c) $x\in[3y,\infty)$: Since $f_y(ey) = 1$, for $x\in[3y,\infty)\subset[ey,\infty)$, we have $f_y(x)\le e^{-\lambda(x-ey)}$ (by checking derivatives). Hence, $\int_{ey}^\infty e^{-\lambda\left(x\log\frac{x}{y}-x\right)}dx \le \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}$. □

Proof of Theorem 2. We use Proposition 3 with $d_{\text{SLB}}(x,y)$ as the distortion measure. Let $\zeta(y,\lambda) \triangleq Ce^{\lambda y}\sqrt{\frac{y}{\lambda}}+\frac{1}{\lambda} \ge \int_0^\infty e^{-\lambda d_{\text{SLB}}(x,y)}dx$ where we pick C appropriately using Lemma 1. Fix some y and some λ. We will choose $\alpha_\lambda(x,y) = 1/(p_X(x)\zeta(y,\lambda))$. It follows from Lemma 1 that $c(y) = \int_0^\infty \alpha_\lambda(x,y)p(x)e^{-\lambda d_{\text{SLB}}(x,y)}dx \le 1$, so $\alpha_\lambda\in\mathcal{A}_\lambda$.

Let $\mathcal{B}(D)$ be the set of conditional probabilities meeting $\mathbb{E}[d_{\text{SLB}}(X,Y)] \le D - \mathbb{E}[Y]$. For any $q_{Y|X}\in\mathcal{B}(D)$, define

$$f \triangleq -\lambda D + \lambda\mathbb{E}[Y] + \int_y\int_x q_{Y|X}(y|x)p(x)\log\alpha_\lambda(x,y)\,dx\,dy$$
$$= -\lambda D + \lambda\mathbb{E}[Y] + h(X) - \int_y q_Y(y)\log\zeta(y,\lambda)\,dy$$

where $q_Y(y) = \int_x p_X(x)q_{Y|X}(y|x)\,dx$.

Next, we use $\zeta(y,\lambda)\le 2\max\{Ce^{\lambda y}\sqrt{y/\lambda}, 1/\lambda\}$. Define

$$K_{\text{small}} \triangleq \left\{y : Ce^{\lambda y}\sqrt{\frac{y}{\lambda}} < \frac{1}{\lambda}\right\},\ \rho_{\text{small}} \triangleq \int_{y\in K_{\text{small}}} q_Y(y)\,dy.$$

Let K_{large} be the complement of K_{small}. (These depend on λ and $q_{Y|X}$, though we don't explicitly write it.) Then

$$f \ge -\lambda D + \lambda\mathbb{E}[Y] + h(X) - \rho_{\text{small}}\log\frac{2}{\lambda} - \int_{y\in K_{\text{large}}} q_Y(y)\log\left(2Ce^{\lambda y}\sqrt{\frac{y}{\lambda}}\right)dy$$
$$\ge -\lambda D + h(X) + C' - \rho_{\text{small}}\log\frac{2}{\lambda} - (1-\rho_{\text{small}})\frac{1}{2}\log\frac{1}{\lambda} - \int_{y\in K_{\text{large}}}\frac{1}{2}q_Y(y)\log y\,dy.$$

We can show $-\int_{y\in K_{\text{large}}}\frac{1}{2}q_Y(y)\log y\,dy \ge -\frac{1}{2e} - \frac{1}{2}\log\mathbb{E}[Y]$. Select $\lambda = 1/D$. Then, using Proposition 3,

$$\inf_{q_{Y|X}\in\mathcal{B}(D)} R(D,p_X) \ge \inf_{q_{Y|X}\in\mathcal{B}(D)} h(X) + \rho_{\text{small}}\log\frac{1}{D} + c + (1-\rho_{\text{small}})\frac{1}{2}\log\frac{1}{D} - \frac{1}{2}\log\mathbb{E}[Y]$$
$$\ge h(X) + \frac{1}{2}\log\frac{1}{D} - \frac{1}{2}\log\mathbb{E}[X] + c$$

since $\mathbb{E}[Y] = \mathbb{E}[X]$. (We can compute that $c\approx -3.10$). □

III. Interval Method

We develop the *Interval Method* for upper-bounding (1), which is an instance of scalar quantization, similar to the technique of Gish and Pierce [30]. The Interval Method partitions the support of X into n intervals I_j. We set Y by interval, i.e. $(Y|X\in I_j) = y_j \triangleq \mathbb{E}[X|X\in I_j]$ (we refer to y_j as the *center* that $X\in I_j$ maps to). Note that this assigns Y from X in a deterministic way, i.e. $Q_{Y|X}(\cdot|\cdot)\in\{0,1\}$, which is not necessarily optimal but simplifies the analysis. The distortion of any such quantization must, of course, upper bound the minimum possible distortion. We present some lemmas; due to space constraints we omit the proofs here.

Definition 3. *For an interval $I\subseteq\mathbb{R}_{\ge 0}$, let $\mathcal{F}_I$ be the set of functions $p : I\to\mathbb{R}_{\ge 0}$ such that $\int_I p(x)\,dx < \infty$ and $\int_I p(x)x\,dx < \infty$.*[4]

Definition 4. *We define for $p\in\mathcal{F}_I$ the values*[5]

$$y^{(p,I)} \triangleq \frac{\int_I p(x)x\,dx}{\int_I p(x)\,dx} \text{ and } D^{(p,I)} \triangleq \int_I p(x)x\log\left(\frac{x}{y^{(p,I)}}\right)dx.$$

Note that when p is a probability distribution (i.e. $\int p(x)dx = 1$), this definition is equivalent to $y^{(p,I)} = \mathbb{E}_{X\sim p}[X|X\in I]$. Therefore the Interval Method automatically satisfies $y_i = \mathbb{E}[X|X \text{ maps to } y_i]$ and $\mathbb{E}[Y] = \mathbb{E}[X]$ so we will ignore this constraint moving forward.

Definition 5. *The minimum distortion on intervals of density $p\in\mathcal{F}_I$ onto n centers is*

$$D(p,n) \triangleq \inf_{\{I_1,\ldots,I_n\}}\sum_{j=1}^n D^{(p,I_j)}$$

[4] Since we will upper bound some probability density functions, it's important that this definition include functions that don't integrate to 1.

[5] We will write $y^{(p,I)}$ and $D^{(p,I)}$ even when p is defined outside of I as well, to mean $y^{(p|_I,I)}$ and $D^{(p|_I,I)}$ where $p|_I$ is the restriction of p to I.

where the infimum is taken over $I_1, \dots, I_n$ partitioning I.

Lemma 2. *For $p \in \mathcal{F}_I$ and constant $c > 0$ (where cI and $I + c$ scale and shift I by c, respectively), let $p_{\times c}$ and p_{+c} be p on cI and $I+c$ with the input scaled or shifted accordingly:*

- *$p_{\times c} : cI \to [0, \infty)$ such that $p_{\times c}(x) = p(x/c)$*
- *$p_{+c} : I + c \to [0, \infty)$ such that $p_{+c}(x) = p(x - c)$.*

Then if $p_{\times c} \in \mathcal{F}_{cI}$ and $p_{+c} \in \mathcal{F}_{I+c}$:

i. $D^{(cp,I)} = cD^{(p,I)}$;
ii. $D^{(p_{\times c},cI)} = c^2 D^{(p,I)}$;
iii. $D^{(p_{+c},I+c)} \le \frac{y^{(p,I)}}{y^{(p,I)}+c} D^{(p,I)}$

Lemma 3. *Let $p_1, p_2 \in \mathcal{F}_I$ for an interval $I \subseteq \mathbb{R}_{\ge 0}$. Then*

$$p_1 \le p_2 \implies D^{(p_1,I)} \le D^{(p_2,I)}.$$

The principal difficulty is that p_1 and p_2 can have different averages $y^{(p_1,I)}, y^{(p_2,I)}$, but this can be overcome with calculus of variations: given any $p, z \in \mathcal{F}_I$ we show that $\frac{d}{d\xi} D^{(p+\xi z, I)} \ge 0$ at $\xi = 0$. Thus, any smooth monotonic deformation of p_1 to p_2 (e.g. adding $\xi(p_2 - p_1)$) never decreases the distortion.

Lemma 4. *If $p_1 \le p_2 \in \mathcal{F}_{[0,\infty)}$, then $D(p_1, n) \le D(p_2, n)$.*

IV. Upper Bounds

In this section we derive upper bounds for (1) when p_X is a) uniform and b) exponential, and use them to show upper bounds for when p_X is Gamma. Let the support of X be $[\ell, L]$ (typically $[0, 1]$ or $[0, \infty)$); we denote the intervals as $I_j = [a_{j-1}, a_j]$ where $\ell = a_0 \le a_1 \le \cdots \le a_{n-1} \le a_n = L$. Let $y_j \triangleq y^{(p_X,I_j)} = \mathbb{E}[X | X \in I_j]$, and let $r_j \triangleq a_j - a_{j-1}$ (width of interval I_j).

A. Upper Bound for Uniform X

Since X is uniform, the computation of the centers is greatly simplified: $y_j = \frac{a_j + a_{j-1}}{2} = a_j - \frac{1}{2} r_j$.

Lemma 5. *When $X \sim \mathrm{Unif}_{[0,1]}$, the distortion on interval $I_j \subseteq [0, 1]$ is at most $\frac{1}{12} \frac{r_j^3}{y_j}$.*

Proof. Since $p_X(x) = 1$ in I_j, the distortion is

$$\int_{y_j - \frac{r_j}{2}}^{y_j + \frac{r_j}{2}} x \log\Big(\frac{x}{y_j}\Big) dx = \int_{-\frac{r_j}{2}}^{\frac{r_j}{2}} (x + y_j) \log\Big(1 + \frac{x}{y_j}\Big) dx \le \int_{-\frac{r_j}{2}}^{\frac{r_j}{2}} (x + y_j) \frac{x}{y_j} dx = \frac{1}{12} \frac{r_j^3}{y_j}$$

which uses that $x > 0$ and log is concave. □

Theorem 6. *For $X \sim \mathrm{Unif}_{[0,1]}$, $D(p_X, n) \le \frac{9}{32} \frac{1}{n^2}$.*

Proof. Using the Interval Method, we set the interval boundaries as $a_j = j^{3/2}/n^{3/2}$. Then the width of each interval is $r_j = \frac{j^{3/2} - (j-1)^{3/2}}{n^{3/2}} \le \frac{3}{2} \frac{(j - \frac{1}{2})^{1/2}}{n^{3/2}}$, and the midpoint is $y_j = \frac{j^{3/2} + (j-1)^{3/2}}{2n^{3/2}} \ge \frac{(j - \frac{1}{2})^{3/2}}{n^{3/2}}$ (since $j^{3/2}$ is convex but its derivative $(3/2) j^{1/2}$ is concave). Therefore, using Lemma 5 we can upper bound the total distortion on I_j by

$$\frac{1}{12} \frac{r_j^3}{y_j} \le \frac{1}{12} \frac{(3/2)^3 (j - \frac{1}{2})^{3/2}}{(j - \frac{1}{2})^{3/2}} \frac{1/n^{9/2}}{1/n^{3/2}} = \frac{9}{32} \frac{1}{n^3}.$$

Since there are n intervals, the total distortion is bounded above by $\frac{9}{32} \frac{1}{n^2}$, and we are done.[6] □

We then use Lemma 2 with Theorem 6 to get Theorem 3.

B. Upper Bound for Exponential X

Proposition 4. *For $X \sim \mathrm{Exponential}(1)$, $D(p_X, n) \le \frac{9}{n^2}$.*

Proof. We split the distribution into the regions $x \le \tau$ and $x > \tau$ and use the results in Section III to bound the distortion from each region separately. We will give n_1 intervals to the $[0, \tau]$ region and n_2 to the $[\tau, \infty)$ region; thus, if $n = n_1 + n_2$,

$$D(p_X, n) \le D(p_X|_{[0,\tau]}, n_1) + D(p_X|_{[\tau,\infty)}, n_2).$$

We will discuss how to set n_1, n_2 and τ later.

For the $[0, \tau]$ region, we use the upper bound $p_X(x) \le 1$. Therefore, by Lemmas 2 and 4 and Theorem 6,

$$D(p_X|_{[0,\tau]}, n_1) \le \tau^2 \frac{9}{32} \frac{1}{n_1^2}.$$

For the $[\tau, \infty)$ region, we use $a_j = 3 \log\big(n_2/(n_2 - j)\big) + \tau$ as our interval boundaries (noting that $a_{n_2} = \infty$, as it should).

First, we consider the infinite interval $I_{n_2} = [a_{n_2-1}, a_{n_2}) = [3 \log(n_2) + \tau, \infty)$. By Lemma 2 parts (i) and (iii) and the memorylessness property of the exponential distribution,

$$\begin{aligned}
&\int_{3\log(n_2)+\tau}^{\infty} e^{-x} x \log\left(\frac{x}{y_{n_2}}\right) dx \\
&\le \frac{1}{3\log(n_2) + \tau + 1} \frac{e^{-\tau}}{n_2^3} \int_0^{\infty} e^{-x} x \log(x)\, dx \\
&= \frac{(1-\gamma) e^{-\tau}}{(3\log(n_2) + \tau + 1)} \frac{1}{n_2^3} \le 18 \frac{e^{-\tau}}{\tau} \frac{1}{n_2^3}
\end{aligned}$$

where γ is the Euler-Mascheroni constant. The last inequality is meant to match it to our bound for the other intervals.

For the other intervals I_j, since $p_X(x) = e^{-x}$ is decreasing, $p_X(x) \le e^{-a_{j-1}} = e^{-\tau} \big(\frac{n_2 - j + 1}{n_2}\big)^3$ over I_j. Also,

$$\begin{aligned}
r_j &= 3 \log\left(\frac{n_2}{n_2 - j}\right) + \tau - 3 \log\left(\frac{n_2}{n_2 - j + 1}\right) - \tau \\
&= 3 \left(\log(n_2 - j + 1) - \log(n_2 - j)\right) \le 3 \frac{1}{n_2 - j}.
\end{aligned}$$

Therefore, by Lemmas 2, 3 and 5 we get, for all j,

$$\begin{aligned}
D^{(p_X,I_j)} &\le D^{(e^{-\tau}(\frac{n_2-j+1}{n_2})^3, I_j)} \\
&\le \frac{1}{12} e^{-\tau} \left(\frac{n_2 - j + 1}{n_2}\right)^3 \frac{r_j^3}{y_j} \le 18 \frac{e^{-\tau}}{\tau} \frac{1}{n_2^3}
\end{aligned}$$

since $y_j \ge \tau$ and $\big(\frac{n_2 - j + 1}{n_2 - j}\big)^3 \le 8$ (a very loose bound for most j). Summing gives $D(p_X|_{[\tau,\infty)}, n_2) \le 18 \frac{e^{-\tau}}{\tau} \frac{1}{n_2^2}$.

Therefore, we have our bound over $[0, \infty)$ for $n = n_1 + n_2$:

$$D(p_X, n) \le \tau^2 \frac{9}{32} \frac{1}{n_1^2} + 18 \frac{e^{-\tau}}{\tau} \frac{1}{n_2^2}.$$

[6] Upper bounds of $O(1/n^2)$ are achievable with interval boundaries $a_j = j^b/n^b$ for any $b > 1$, though $b = 3/2$ gives the simplest proof and the best constant with our method. When $b = 1$, the intervals are uniform and decay rate becomes $\Theta\big(\log(n)/n^2\big)$.

Let $n_1 = cn$ and $n_2 = (1-c)n$; we know $c = j/n$ for some $j \in [n-1]$, but for now let's require only $c \in (0,1)$. Thus,

$$D(p_X, n) \le \Big(\tau^2 \frac{9}{32}\frac{1}{c^2} + 18\frac{e^{-\tau}}{\tau}\frac{1}{(1-c)^2}\Big)\frac{1}{n^2}.$$

Numerical optimization gives a minimum at $\tau \approx 2.954$ and $c \approx 0.664$, giving a constant of 8.4.

Now we bring back the $c = j/n$ condition. We note that any $c \in [0.614, 0.714]$ (and $\tau = 2.954$) gives $D(p_X, n) \le \frac{9}{n^2}$; so this holds for all $n \ge 10$. For $n < 10$, we can solve each case separately. Therefore, $D(p_X, n) \le \frac{9}{n^2}$ as we wanted.[7] □

C. Upper Bound for Gamma X

Lemma 6. *There exists c such that for any $-1 < s$, if $p_X(x) = (1+s)x^s$ on $[0,1]$ (and 0 elsewhere), $D(p_X, n) \le c/n^2$.*

Proof Sketch. Let $\beta = 3/(2+s)$ and let $a_j = j^\beta/n^\beta$. Then, $D^{(p_X, I_1)} \le 1/n^3$, and Lemma 3 and Lemma 5 then give $D^{(p_X, I_j)} \le c'/n^3$, and take the sum over the n intervals. □

Proposition 5. *For $X \sim \mathrm{Gamma}(\alpha, \beta)$, $D(p_X, n) = O\big(\frac{1}{n^2}\big)$.*

Proof Sketch. We consider two cases: a) $\alpha < 1$ and b) $\alpha > 1$ ($\alpha = 1$ is the exponential distribution). For the $\alpha < 1$ case, we upper bound the portion $[0,1]$ using Lemma 6 and the tail portion $(1, \infty)$ by an exponential density and use Proposition 4 and Lemma 4. For the $\alpha > 1$ case, for some b, we upper bound the $[0, b]$ portion by a uniform density and the $[b, \infty)$ portion by an exponential density and use Lemma 4. □

V. Expected Divergence Results

Finally, we connect divergence rate distortion on Gamma sources to quantizing $\mathcal{P}([k])$ with symmetric Dirichlet prior.

Fact 1. *Let $X_i \overset{iid}{\sim} \mathrm{Gamma}(\alpha, \beta)$ for $i \in [k]$ (giving random vector X^k) and let $S \triangleq \sum_{i=1}^k X_i$. Then $P \triangleq X^k/S \sim \mathrm{Dir}_k(\alpha)$ and S and P are independent. (The division above treats P and X^k as k-length vectors.)*

Lemma 7. *For $S \sim \mathrm{Gamma}(\alpha, \alpha)$, $\mathbb{E}[S \log S] \le 1/(2\alpha)$.*

Proposition 6. *Let p_Z be the probability density of $Z \sim \mathrm{Gamma}(\alpha, \alpha)$. Then, for $k > 0$,*

$$k \cdot R\left(p_Z, D + 1/(2\alpha k)\right) \le \log M^*(\mathrm{Dir}_k(\alpha), D).$$

Proof. Let random vector $X^k = (X_1, ..., X_k)$ be such that each $X_i \sim \mathrm{Gamma}(\alpha, \alpha k)$. Let $S = \sum_{i=1}^k X_i$. Then, $S \sim \mathrm{Gamma}(\alpha k, \alpha k)$ and $\mathbb{E}[S] = 1$.

From Fact 1, we know that $X^k/S \sim \mathrm{Dir}_k(\alpha)$. Also S is independent of X^k/S and therefore it is also independent of $\min_j D_{\mathrm{KL}}((X^k/S)||Q(j))$. Fixing J as the argmin, we denote the components of $Q(J)$ as $(Y_1, \ldots, Y_k)$. Using Lemma 7,

$$\sum_{i=1}^k \mathbb{E}\left[X_i \log \frac{X_i}{Y_i}\right] = \mathbb{E}\left[\sum_{i=1}^k S\frac{X_i}{S}\log\frac{X_i}{SY_i} + X_i \log S\right]$$
$$= \mathbb{E}[S]\,\mathbb{E}[D_{\mathrm{KL}}((X^k/S)||Q(J))] + \mathbb{E}[S\log S] \le D + \frac{1}{2\alpha k}.$$

[7] Some of the bounds we used were loose, so more detailed analysis (omitted) can improve the constant, though not the decay rate.

Since we showed that a k-ary quantizer with the given average distortion exists, the standard single-letter lower bound from rate-distortion implies the bound [35, Ch 25]. If p_X is the density of each X_i, then by scaling $R(p_X, \frac{D+1/(2\alpha k)}{k}) = R(p_Z, D + 1/(2\alpha k))$ where $Z \sim \mathrm{Gamma}(\alpha, \alpha)$. □

Proposition 7. *For any α, k such that $\alpha k > 1$, there exist centers $Q(1), \ldots, Q(m)$ such that*

$$m = \Big(\frac{c}{D(1-(\alpha k)^{-1/3})^2}\Big)^{k/2}$$

and $\mathbb{E}_{P\sim \mathrm{Dir}_k(\alpha)} \min_j D_{\mathrm{KL}}(P||Q(j)) \le D$.

Proof. Let $S \sim \mathrm{Gamma}(\alpha k, \beta)$ (we will determine β later) and $P = (p_1, ..., p_k) \sim \mathrm{Dir}_k(\alpha)$. Let $X_i = Sp_i$ so $X_i \sim \mathrm{Gamma}(\alpha, \beta)$. Setting $\mathbb{E}[d(X_i, Y_i)] \le D'/k$ gives

$$D' \ge \sum_{i=1}^k \mathbb{E}[d(X_i, Y_i)] = \mathbb{E}\left[\sum_{i=1}^k X_i \log \frac{X_i}{Y_i}\right]$$
$$= \mathbb{E}\left[S\sum_{i=1}^k \frac{X_i}{S}\log\frac{X_i/S}{Y_i/\sum_{i=1}^k Y_i} + S\log\frac{S}{\sum_{i=1}^k Y_i}\right]$$
$$\ge \mathbb{E}\left[SD_{\mathrm{KL}}(P||Q)\right].$$

The log-sum inequality shows that $\mathbb{E}\left[S\log(S/\sum_{i=1}^n Y_i)\right] \ge 0$. With Proposition 5, this shows that $(c/D')^{k/2}$ centers produces $\mathbb{E}_{P,S}\big[SD_{\mathrm{KL}}(P||Q)\big] \le D'$. Note that Q depends on S and P. If $Z_1, Z_2 \ge 0$ are random variables then for any $\sigma > 0$, $\mathbb{E}[Z_2] \le \frac{\mathbb{E}[Z_1 Z_2]}{\sigma \mathbb{P}[Z_1 \ge \sigma]}$. Therefore, for any $\sigma > 0$,

$$\mathbb{E}D_{\mathrm{KL}}(P||Q) \le \frac{\mathbb{E}[SD_{\mathrm{KL}}(P||Q)]}{\sigma\mathbb{P}[S \ge \sigma]} \le \frac{D'}{\sigma\mathbb{P}[S \ge \sigma]}.$$

We can then fix s to minimize $\mathbb{E}[D_{\mathrm{KL}}(P||Q) \,|\, S = s]$ to get

$$\min_s \mathbb{E}[D_{\mathrm{KL}}(P||Q) \,|\, S = s] \le \frac{D'}{\sigma\mathbb{P}[S \ge \sigma]}.$$

We then want to set σ (and the rate parameter β) to maximize $\sigma\mathbb{P}[S \ge \sigma]$. We use $\beta = \alpha k$; then $\mathbb{E}[S] = \alpha k/\beta = 1$ and $\mathrm{Var}[S] = \alpha k/\beta^2 = \frac{1}{\alpha k}$. Thus, defining $t = 1 - \sigma$ (assuming $\sigma \le 1$), applying Chebyshev's inequality, and setting $t = (\alpha k)^{-1/3}$ gives $\mathbb{P}[S \ge 1 - t] \ge 1 - 1/(\alpha k t^2)$:

$$\mathbb{E}D_{\mathrm{KL}}(P||Q) \le \frac{D'}{(1-t)\big(1 - \frac{1}{(\alpha k t^2)}\big)} \le \frac{D'}{(1-(\alpha k)^{-1/3})^2}$$

(here the expectation is only over P, and use fixed $S = s$). We can then define our coding of $P \in \mathcal{P}([k])$ to a center $Q \in \{Q(j)\}_{j=1}^m$ to be the result of the following procedure: $P \to X^k \to Y^k \to Q$ where $X^k = sP$, Y^k is the encoding of X^k using Proposition 5, and $Q = Y^k/\sum_{i=1}^k Y_i$.

Recall that each X^k maps to one of $m = (c/D')^{k/2}$ centers. Thus, letting $D = \frac{D'}{(1-(\alpha k)^{-1/3})^2}$, we get a coding scheme on m centers where $\mathbb{E}[D_{\mathrm{KL}}(P||Q)] \le D$, giving the result. □

Finally, we can put these together to prove Theorem 1:

Proof of Theorem 1. The first part follows from Proposition 6 and Theorem 2, and the second from Proposition 7. □

Acknowledgment

The authors would like to thank Meir Feder for discussions that led to this project.

REFERENCES

[1] T.M. Cover and J.A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, USA, 2006.

[2] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *The Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.

[3] D. Haussler and A. Barron, "How well do Bayes methods work for on-line prediction of ± 1 values?," in *In Proceedings of the Third NEC Symposium on Computation and Cognition. SIAM*, 1992, pp. 74–100.

[4] J. Rissanen, "Universal coding, information, prediction, and estimation," *Information Theory, IEEE Transactions on*, vol. 30, pp. 629 – 636, 08 1984.

[5] J. Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, vol. 14, no. 3, pp. 1080–1100, 1986.

[6] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[7] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[8] T.S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[9] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, January 2003.

[10] H. Steck and T. Jaakkola, "On the Dirichlet prior and Bayesian regularization," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2002, NIPS'02, p. 713–720, MIT Press.

[11] S. Schober, "Some worst-case bounds for Bayesian estimators of discrete distributions," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2194–2198.

[12] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2007.

[13] K. Varshney and L. Varshney, "Quantization of prior probabilities for hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4553–4562, 2008.

[14] D. Donoho, "Wald lecture I: Counting bits with Kolmogorov and Shannon," 2000.

[15] Y. Zhu and J. Lafferty, "Quantized estimation of Gaussian sequence models in euclidean balls," 2014.

[16] Y. Zhu and J. Lafferty, "Quantized minimax estimation over Sobolev ellipsoids," *Information and Inference: A Journal of the IMA*, vol. 7, no. 1, pp. 31–82, 06 2017.

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

[18] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *The Annals of Statistics*, vol. 25, no. 6, pp. 2451 – 2492, 1997.

[19] B. Clarke and A. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[20] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[21] B. Clarke, "Asymptotic cumulative risk and Bayes risk under entropy loss, with applications," 1989.

[22] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, 1990.

[23] Q. Xie and A. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory*, vol. 43, no. 2, pp. 646–657, 1997.

[24] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2026–2031, 1994.

[25] T. Koch, "The Shannon lower bound is asymptotically tight," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6155–6161, 2016.

[26] T. Linder and R. Zamir, "High-resolution source coding for non-difference distortion measures: the rate-distortion function," *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 533–547, 1999.

[27] S.L. Fix, "Rate distortion functions for squared error distortion measures," *Proc. 16th Annu. Allerton Conf. Commun., Contr., Comput.*, Oct. 1978.

[28] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1939–1952, 1994.

[29] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[30] H. Gish and J. Pierce, "Asymptotically efficient quantizing," *IEEE Transactions on Information Theory*, vol. 14, no. 5, pp. 676–683, 1968.

[31] A. Buzo, F. Kuhlmann, and C. Rivera, "Rate-distortion bounds for quotient-based distortions with application to Itakura-Saito distortion measures," *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 141–147, 1986.

[32] H. Tan and K. Yao, "Evaluation of rate-distortion functions for a class of independent identically distributed sources under an absolute-magnitude criterion," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 59–64, 1975.

[33] K. Watanabe and S. Ikeda, "Rate-distortion function for gamma sources under absolute-log distortion measure," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2557–2561.

[34] T. Berger, *Rate Distortion Theory: Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

[35] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *MIT (6.441), UIUC (ECE 563)*, 2013-2016.