Efficient Representation of Large-Alphabet Probability Distributions

Aviv Adler, *Member, IEEE*, Jennifer Tang, *Member, IEEE*, and Yury Polyanskiy, *Senior Member, IEEE*

***Abstract*—A number of engineering and scientific problems require representing and manipulating probability distributions over large alphabets, which we may think of as long vectors of reals summing to 1. In some cases it is required to represent such a vector with only b bits per entry. A natural choice is to partition the interval $[0, 1]$ into 2^b uniform bins and quantize entries to each bin independently. We show that a minor modification of this procedure – applying an entrywise non-linear function (compander) $f(x)$ prior to quantization – yields an extremely effective quantization method. For example, for $b = 8(16)$ and 10^5-sized alphabets, the quality of representation improves from a loss (under KL divergence) of $0.5(0.1)$ bits/entry to $10^{-4}(10^{-9})$ bits/entry. Compared to floating point representations, our compander method improves the loss from $10^{-1}(10^{-6})$ to $10^{-4}(10^{-9})$ bits/entry. These numbers hold for both real-world data (word frequencies in books and DNA k-mer counts) and for synthetic randomly generated distributions. Theoretically, we analyze a minimax optimality criterion and show that the closed-form compander $f(x) \propto \text{ArcSinh}(\sqrt{c_K(K \log K)}x)$ is (asymptotically as $b \to \infty$) optimal for quantizing probability distributions over a K-letter alphabet. Non-asymptotically, such a compander (substituting $1/2$ for c_K for simplicity) has KL-quantization loss bounded by $\leq 8 \cdot 2^{-2b} \log^2 K$. Interestingly, a similar minimax criterion for the quadratic loss on the hypercube shows optimality of the standard uniform quantizer. This suggests that the ArcSinh quantizer is as fundamental for KL-distortion as the uniform quantizer for quadratic distortion.**

***Index Terms*—Quantization (signal), data compression.**

I. Compander Basics and Definitions

CONSIDER the problem of *quantizing* the probability simplex $\triangle_{K-1} = \{\boldsymbol{x} \in \mathbb{R}^K : \boldsymbol{x} \geq \boldsymbol{0}, \sum_i x_i = 1\}$ of alphabet size K,[1] i.e., of finding a finite subset $\mathcal{Z} \subseteq \triangle_{K-1}$ to represent the entire simplex. Each $\boldsymbol{x} \in \triangle_{K-1}$ is associated with some $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{x}) \in \mathcal{Z}$, and the objective is to find a set $\mathcal{Z}$ and an assignment such that the difference between the values $\boldsymbol{x} \in \triangle_{K-1}$ and their representations $\boldsymbol{z} \in \mathcal{Z}$ are minimized; while this can be made arbitrarily small by making $\mathcal{Z}$ arbitrarily large, the goal is to do this efficiently for any given fixed size $|\mathcal{Z}| = M$. Since $\boldsymbol{x}, \boldsymbol{z} \in \triangle_{K-1}$, they both represent probability distributions over a size-K alphabet. Hence, a natural way to measure the quality of the quantization is to use the KL (Kullback-Leibler) divergence $D_{\text{KL}}(\boldsymbol{x}\|\boldsymbol{z})$, which corresponds to the excess code length for lossless compression and is commonly used as a way to compare probability distributions. (Note that we want to minimize the KL divergence.)

While one can consider how to best represent the vector $\boldsymbol{x}$ as a whole, in this paper we consider only *scalar quantization* methods in which each element x_j of $\boldsymbol{x}$ is handled separately, since we showed in [1] that for Dirichlet priors on the simplex, methods using scalar quantization perform nearly as well as optimal vector quantization. Scalar quantization is also typically simpler and faster to use, and can be parallelized easily. Our scalar quantizer is based on *companders* (portmanteau of 'compressor' and 'expander'), a simple, powerful and flexible technique first explored by Bennett in 1948 [2] in which the value x_j is passed through a nonlinear function f before being uniformly quantized. We discuss the background in greater depth in Section III.

In what follows, log is always base-e unless otherwise specified. We denote $[N] \coloneqq \{1, \ldots, N\}$.

A. Encoding

Companders require two things: a monotonically increasing[2] function $f : [0, 1] \to [0, 1]$ (we denote the set of such functions as $\mathcal{F}$) and an integer N representing the number of quantization levels, or *granularity*. To simplify the problem and algorithm, we use the same f for each element of the vector $\boldsymbol{x} = (x_1, \ldots, x_K) \in \triangle_{K-1}$ (see Remark 1). To quantize $x \in [0, 1]$, the compander computes $f(x)$ and applies a uniform quantizer with N levels, i.e., encoding x to $n_N(x) \in [N]$ if $f(x) \in (\frac{n-1}{N}, \frac{n}{N}]$; this is equivalent to $n_N(x) = \lceil f(x)N \rceil$.

This encoding partitions $[0, 1]$ into *bins* $I^{(n)}$:

$$x \in I^{(n)} = f^{-1}\left(\left(\frac{n-1}{N}, \frac{n}{N}\right]\right) \iff n_N(x) = n$$

where f^{-1} denotes the preimage under f.

Manuscript received 15 April 2022; accepted 23 August 2022. Date of publication 6 January 2023; date of current version 13 June 2023. This work was supported in part by NSF under Grant CCF-2131115; in part by the United States Air Force Research Laboratory; and in part by the United States Air Force Artificial Intelligence Accelerator through the Cooperative Agreement under Grant FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. *(Corresponding author: Jennifer Tang.)*

The authors are with the EECS Department, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: adlera@mit.edu; jstang@mit.edu; yp@mit.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSAIT.2023.3234502, provided by the authors.

Digital Object Identifier 10.1109/JSAIT.2023.3234502

[1] While the alphabet has K letters, $\triangle_{K-1}$ is $(K-1)$-dimensional due to the constraint that the entries sum to 1.

[2] We require increasing functions as a convention, so larger x_i map to larger values in $[N]$. Note that f does *not* need to be *strictly* increasing; if f is flat over interval $I \subseteq [0, 1]$ then all $x_i \in I$ will always be encoded by the same value. This is useful if no x_i in I ever occurs, i.e., I has zero probability mass under the prior.

As an example, consider the function $f(x) = x^s$. Varying s gives a natural class of functions from [0, 1] to [0, 1], which we call the class of *power companders*. If we select $s = 1/2$ and $N = 4$, then the 4 bins created by this encoding are

$$I^{(1)} = (0, 1/16], I^{(2)} = (1/16, 1/4],$$
$$I^{(3)} = (1/4, 9/16], I^{(4)} = (9/16, 1].$$

B. Decoding

To decode $n \in [N]$, we pick some $y_{(n)} \in I^{(n)}$ to represent all $x \in I^{(n)}$; for a given x (at granularity N), its representation is denoted $y(x) = y_{(n_N(x))}$. This is generally either the *midpoint* of the bin or, if x is drawn randomly from a known prior[3] p, the *centroid* (the mean within bin $I^{(n)}$). The midpoint and centroid of $I^{(n)}$ are defined, respectively, as

$$\bar{y}_{(n)} = \frac{1}{2}\left(f^{-1}\left(\frac{n-1}{N}\right) + f^{-1}\left(\frac{n}{N}\right)\right)$$
$$\widetilde{y}_{(n)} = \mathbb{E}_{X\sim p}\left[X \mid X \in I^{(n)}\right].$$

We will discuss this in greater detail in Section I-D.

Handling each element of $\boldsymbol{x}$ separately means the decoded values may not sum to 1, so we normalize the vector after decoding. Thus, if $\boldsymbol{x}$ is the input,

$$z_i(\boldsymbol{x}) = \frac{y(x_i)}{\sum_{j=1}^{K} y(x_j)} \tag{1}$$

and the vector $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{x}) = (z_1(\boldsymbol{x}), \ldots, z_K(\boldsymbol{x})) \in \triangle_{K-1}$ is the output of the compander. This notation reflects the fact that each entry of the normalized reconstruction depends on all of $\boldsymbol{x}$ due to the normalization step. We refer to $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) = (y(x_1), \ldots, y(x_K))$ as the *raw* reconstruction of $\boldsymbol{x}$, and $\boldsymbol{z}$ as the *normalized* reconstruction. If the raw reconstruction uses centroid decoding, we likewise denote it using $\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{y}}(\boldsymbol{x}) = (\widetilde{y}(x_1), \ldots, \widetilde{y}(x_K))$. For brevity we may sometimes drop the $\boldsymbol{x}$ input in the notation, e.g., $\boldsymbol{z} \coloneqq \boldsymbol{z}(\boldsymbol{x})$; if $\boldsymbol{X}$ is random we will sometimes denote its quantization as $\boldsymbol{Z} \coloneqq \boldsymbol{z}(\boldsymbol{X})$.

Thus, any $\boldsymbol{x} \in \triangle_{K-1}$ requires $K\lceil \log_2 N \rceil$ bits to store; to encode and decode, only f and N need to be stored (as well as the prior if using centroid decoding). Another major advantage is that a single f can work well over many or all choices of N, making the design more flexible.

C. KL Divergence Loss

The loss incurred by representing $\boldsymbol{x}$ as $\boldsymbol{z} \coloneqq \boldsymbol{z}(\boldsymbol{x})$ is the KL divergence

$$D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{z}) = \sum_{i=1}^{K} x_i \log \frac{x_i}{z_i}.$$

Although this loss function has some unusual properties (for instance $D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{z}) \neq D_{\mathrm{KL}}(\boldsymbol{z}\|\boldsymbol{x})$ and it does not obey the triangle inequality) it measures the amount of 'mis-representation' created by representing the probability vector $\boldsymbol{x}$ by another probability vector $\boldsymbol{z}$, and is hence is a natural quantity to minimize. In particular, it represents the excess code length created by trying to encode the output of $\boldsymbol{x}$ using a code built for $\boldsymbol{z}$, as well as having connections to hypothesis testing (a natural setting in which the 'difference' between probability distributions is studied).

D. Distributions From a Prior

Much of our work concerns the case where $\boldsymbol{x} \in \triangle_{K-1}$ is drawn from some prior $P_{\boldsymbol{x}}$ (to be commonly denoted as simply P). Using a single f for each entry means we can WLOG assume that P is symmetric over the alphabet, i.e., for any permutation σ, if $\boldsymbol{X} \sim P$ then $\sigma(\boldsymbol{X}) \sim P$ as well. This is because for any prior P over $\triangle_{K-1}$, there is a symmetric prior P' such that

$$\mathbb{E}_{\boldsymbol{X}\sim P}[D_{\mathrm{KL}}(\boldsymbol{X}\|\boldsymbol{z}(\boldsymbol{X}))] = \mathbb{E}_{\boldsymbol{X}'\sim P'}\left[D_{\mathrm{KL}}\left(\boldsymbol{X}'\|\boldsymbol{z}\left(\boldsymbol{X}'\right)\right)\right]$$

for all f, where $\boldsymbol{z}(\boldsymbol{X})$ is the result of quantizing (to any number of levels) with f as the compander. To get $\boldsymbol{X}' \sim P'$, generate $\boldsymbol{X} \sim P$ and a uniformly random permutation σ, and let $\boldsymbol{X}' = \sigma(\boldsymbol{X})$.

We denote the set of symmetric priors as $\mathcal{P}_K^{\triangle}$. Note that a key property of symmetric priors is that their marginal distributions are the same across all entries, and hence we can speak of $P \in \mathcal{P}_K^{\triangle}$ having a single marginal p.

Remark 1: In principle, given a nonsymmetric prior $P_{\boldsymbol{x}}$ over $\triangle_{K-1}$ with marginals $p_1, \ldots, p_K$, we could quantize each letter's value with a different compander $f_1, \ldots, f_K$, giving more accuracy than using a single f (at the cost of higher complexity). However, the symmetrization of $P_{\boldsymbol{x}}$ over the letters (by permuting the indices randomly after generating $\boldsymbol{X} \sim P_{\boldsymbol{x}}$) yields a prior in $\mathcal{P}_K^{\triangle}$ on which any single f will have the same (overall) performance and cannot be improved on by using varying f_i. Thus, considering symmetric $P_{\boldsymbol{x}}$ suffices to derive our minimax compander.

While the random probability vector comes from a prior $P \in \mathcal{P}_K^{\triangle}$, our analysis will rely on decomposing the loss so we can deal with one letter at a time. Hence, we work with the marginals p of P (which are identical since P is symmetric), which we refer to as *single-letter distributions* and are probability distributions over [0, 1].

We let $\mathcal{P}$ denote the class of probability distributions over [0, 1] that are absolutely continuous with respect to the Lebesgue measure. We denote elements of $\mathcal{P}$ by their probability density functions (PDF), e.g., $p \in \mathcal{P}$; the cumulative distribution function (CDF) associated with p is denoted F_p and satisfies $F_p'(x) = p(x)$ and $F_p(x) = \int_0^x p(t)\,dt$ (since F_p is monotonic, its derivative exists almost everywhere). Note that while $p \in \mathcal{P}$ does not have to be continuous, its CDF F_p must be absolutely continuous. Following common terminology [3], we refer to such probability distributions as *continuous*.

Let $\mathcal{P}_{1/K} = \{p \in \mathcal{P} : \mathbb{E}_{X\sim p}[X] = 1/K\}$. Note that $P \in \mathcal{P}_K^{\triangle}$ implies its marginals p are in $\mathcal{P}_{1/K}$.

E. Expected Loss and Preliminary Results

For $P \in \mathcal{P}_K^{\triangle}$, $f \in \mathcal{F}$ and granularity N, we define the *expected loss*:

$$\mathcal{L}_K(P, f, N) = \mathbb{E}_{\boldsymbol{X}\sim P}[D_{\mathrm{KL}}(\boldsymbol{X}\|\boldsymbol{z}(\boldsymbol{X}))]. \tag{2}$$

This is the value we want to minimize over f.

Remark 2: While $\boldsymbol{X}$ and $\boldsymbol{z}(\boldsymbol{X})$ are random, they are also probability vectors. The KL divergence $D_{\mathrm{KL}}(\boldsymbol{X}\|\boldsymbol{z}(\boldsymbol{X}))$ is the

[3]Priors on $\triangle_{K-1}$ induce priors over [0, 1] for each entry.

divergence between $\boldsymbol{X}$ and $\boldsymbol{z}(\boldsymbol{X})$ themselves, not the prior distributions over $\triangle_{K-1}$ they are drawn from.

Note that $\mathcal{L}_K(P,f,N)$ can almost be decomposed into a sum of K separate expected values, except the normalization step (1) depends on the random vector $\boldsymbol{X}$ as a whole. Hence, we define the *raw loss*:

$$\widetilde{\mathcal{L}}_K(P,f,N) = \mathbb{E}_{\boldsymbol{X}\sim P}\Big[\sum_{i=1}^{K} X_i \log(X_i/\widetilde{y}(X_i))\Big]. \tag{3}$$

We also define for $p \in \mathcal{P}$, the *single-letter loss* as

$$\widetilde{L}(p,f,N) = \mathbb{E}_{X\sim p}\big[X\log(X/\widetilde{y}(X))\big]. \tag{4}$$

The raw loss is useful because it bounds the (normalized) expected loss and is decomposable into single-letter losses. Note that both raw and single-letter loss are defined with centroid decoding.

Proposition 1: For $P \in \mathcal{P}_K^{\triangle}$ with marginals p,

$$\mathcal{L}_K(P,f,N) \le \widetilde{\mathcal{L}}_K(P,f,N) = K\widetilde{L}(p,f,N).$$

Proof: Separating out the normalization term gives

$$\begin{aligned}\mathcal{L}(P,f,N) &= \mathbb{E}_{\boldsymbol{X}\sim P}[D_{\mathrm{KL}}(\boldsymbol{X}||\boldsymbol{z}(\boldsymbol{X}))]\\ &= \widetilde{\mathcal{L}}_K(P,f,N) + \mathbb{E}_{X\sim P}\left[\log\left(\sum_{i=1}^{K}\widetilde{y}(X_i)\right)\right].\end{aligned}$$

Since $\mathbb{E}[\widetilde{y}(X_i)] = \mathbb{E}[X_i]$ for all i, $\sum_{i=1}^{K}\mathbb{E}[\widetilde{y}(X_i)] = \sum_{i=1}^{K}\mathbb{E}[X_i] = 1$. Because log is concave, by Jensen's Inequality

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}\sim P}\left[\log\Big(\sum_{i=1}^{K}\widetilde{y}(X_i)\Big)\right] &\le \log\left(\mathbb{E}\Big[\sum_{i=1}^{K}\widetilde{y}(X_i)\Big]\right)\\ &= \log(1) = 0\end{aligned}$$

and we are done.[4] ∎

To derive our results about worst-case priors (for instance, Theorem 1), we will also be interested in $\widetilde{L}(p,f,N)$ even when p is not known to be a marginal of some $P \in \mathcal{P}_K^{\triangle}$.

Remark 3: Though one can define raw and single-letter loss without centroid decoding (replacing $\widetilde{y}$ in (3) or (4) with another decoding method $\widehat{y}$), this removes much of their usefulness. This is because the resulting expected loss can be dominated by the difference between $\mathbb{E}[X]$ and $\mathbb{E}[\widehat{y}(X)]$, potentially even making it negative; specifically, the Taylor expansion of $X\log(X/\widehat{y}(X))$ has $X-\widehat{y}(X)$ in its first term, which can have negative expectation.

While this can make the expected 'raw loss' negative under general decoding, it cannot be exploited to make the (normalized) expected loss negative because the normalization step $z_i(\boldsymbol{X}) = \widehat{y}(X_i)/\sum_j \widehat{y}(X_j)$ cancels out the problematic term. Centroid decoding avoids this problem by ensuring $\mathbb{E}[X] = \mathbb{E}[\widetilde{y}(X)]$, removing the issue.

As we will show, when N is large these values are roughly proportional to N^{-2} (for well-chosen f) and so we define the *asymptotic single-letter loss*:

$$\widetilde{L}(p,f) = \lim_{N\to\infty} N^2\widetilde{L}(p,f,N). \tag{5}$$

[4] An upper bound similar to Proposition 1 can be found in [4, Lemma 1].

We similarly define $\widetilde{\mathcal{L}}_K(P,f)$ and $\mathcal{L}_K(P,f)$. While the limit in (5) does not necessarily exist for every p,f, we will show that one can ensure it exists by choosing an appropriate f (which works against any $p \in \mathcal{P}$), and cannot gain much by not doing so.

II. Results

We demonstrate, theoretically and experimentally, the efficacy of companding for quantizing probability distributions with KL divergence loss.

A. Theoretical Results

While we will occasionally give intuition for how the results here are derived, our primary concern in this section is to fully state the results and to build a clear framework for discussing them.

Our main results concern the formulation and evaluation of a *minimax compander* f_K^* for alphabet size K, which satisfies

$$f_K^* = \arg\min_{f\in\mathcal{F}} \sup_{p\in\mathcal{P}_{1/K}} \widetilde{L}(p,f). \tag{6}$$

We require $p \in \mathcal{P}_{1/K}$ because if $P \in \mathcal{P}_K^{\triangle}$ and is symmetric, its marginals are in $\mathcal{P}_{1/K}$.

The natural counterpart of the minimax compander f_K^* is the *maximin density* $p_K^* \in \mathcal{P}_{1/K}$, satisfying

$$p_K^* = \arg\max_{p\in\mathcal{P}_{1/K}} \inf_{f\in\mathcal{F}} \widetilde{L}(p,f). \tag{7}$$

We call (6) and (7), respectively, the *minimax condition* and the *maximin condition*.

In the same way that the minimax compander gives the best performance guarantee against an unknown single-letter prior $p \in \mathcal{P}_{1/K}$ (asymptotic as $N\to\infty$), the maximin density is the most difficult prior to quantize effectively as $N\to\infty$. Since they are highly related, we will define them together:

Proposition 2: For alphabet size $K > 4$, there is a unique $c_K \in [\frac{1}{4},\frac{3}{4}]$ such that if $a_K = (4/(c_K K\log K + 1))^{1/3}$ and $b_K = 4/a_K^2 - a_K$, then the following density is in $\mathcal{P}_{1/K}$:

$$p_K^*(x) = \left(a_K x^{1/3} + b_K x^{4/3}\right)^{-3/2}. \tag{8}$$

Furthermore, $\lim_{K\to\infty} c_K = 1/2$.

Note that this is both a result and a definition: we show that a_K, b_K, c_K exist which make the definition of p_K^* possible. With the constant c_K, we define the minimax compander:

Definition 1: Given the constant c_K as shown to exist in Proposition 2, the *minimax compander* is the function $f_K^* : [0,1] \to [0,1]$ where

$$f_K^*(x) = \frac{\mathrm{ArcSinh}\left(\sqrt{c_K(K\log K)}\,x\right)}{\mathrm{ArcSinh}\left(\sqrt{c_K K\log K}\right)}.$$

The *approximate minimax compander* f_K^{**} is

$$f_K^{**}(x) = \frac{\mathrm{ArcSinh}\left(\sqrt{(1/2)(K\log K)}\,x\right)}{\mathrm{ArcSinh}\left(\sqrt{(1/2)K\log K}\right)}. \tag{9}$$

Remark 4: While f_K^* and f_K^{**} might seem complex, $\mathrm{ArcSinh}(\sqrt{w}) = \log(\sqrt{w}+\sqrt{w+1})$ so they are relatively simple functions to work with.

We will show that f_K^*, p_K^* as defined above satisfy their respective conditions (6) and (7).

Theorem 1: The minimax compander f_K^* and maximin single-letter density p_K^* satisfy

$$\sup_{p\in\mathcal{P}_{1/K}} \widetilde{L}(p,f_K^*) = \inf_{f\in\mathcal{F}} \sup_{p\in\mathcal{P}_{1/K}} \widetilde{L}(p,f) \tag{10}$$

$$= \sup_{p\in\mathcal{P}_{1/K}} \inf_{f\in\mathcal{F}} \widetilde{L}(p,f) = \inf_{f\in\mathcal{F}} \widetilde{L}(p_K^*,f) \tag{11}$$

which is equal to $\widetilde{L}(p_K^*,f_K^*)$ and satisfies

$$\widetilde{L}(p_K^*,f_K^*) = \frac{1}{24}(1+o(1))K^{-1}\log^2 K. \tag{12}$$

Since any symmetric $P \in \mathcal{P}_K^{\triangle}$ has marginals $p \in \mathcal{P}_{1/K}$, this (with Proposition 1) implies an important corollary for the normalized KL-divergence loss incurred by using the minimax compander:

Corollary 1: For any prior $P \in \mathcal{P}_K^{\triangle}$,

$$\mathcal{L}_K(P,f_K^*) \le \widetilde{\mathcal{L}}_K(P,f_K^*) = \frac{1}{24}(1+o(1))\log^2 K.$$

However, the set of symmetric $P \in \mathcal{P}_K^{\triangle}$ does not correspond exactly with $p \in \mathcal{P}_{1/K}$: while any symmetric $P \in \mathcal{P}_K^{\triangle}$ has marginals $p \in \mathcal{P}_{1/K}$, it is not true that any given $p \in \mathcal{P}_{1/K}$ has a corresponding symmetric prior $P \in \mathcal{P}_K^{\triangle}$. Thus, it is natural to ask: can the minimax compander's performance be improved by somehow taking these 'shape' constraints into account? The answer is 'not by more than a factor of ≈ 2':

Proposition 3: There is a prior $P^* \in \mathcal{P}_K^{\triangle}$ such that for any $P \in \mathcal{P}_K^{\triangle}$

$$\inf_{f\in\mathcal{F}} \widetilde{\mathcal{L}}_K(P^*,f) \ge \frac{K-1}{2K}\widetilde{\mathcal{L}}_K(P,f_K^*).$$

While the minimax compander satisfies the minimax condition (6), it requires working with the constant c_K, which, while bounded, is tricky to compute or use exactly. Hence, in practice we advocate using the *approximate minimax compander* (9), which yields very similar asymptotic performance without needing to know c_K:

Proposition 4: Suppose that K is sufficiently large so that $c_K \in [\frac{1}{2(1+\varepsilon)}, \frac{1+\varepsilon}{2}]$. Then for any $p \in \mathcal{P}$,

$$\widetilde{L}(p,f_K^{**}) \le (1+\varepsilon)\widetilde{L}(p,f_K^*).$$

Before we show how we get Theorem 1, we make the following points:

Remark 5: If we use the uniform quantizer instead of minimax there exists a $P \in \mathcal{P}_K^{\triangle}$ where

$$\mathbb{E}_{X\sim P}[D_{\mathrm{KL}}(X\|Z)] = \Theta\Big(K^2N^{-2}\log N\Big).$$

This is done by using marginal density p uniform on $[0, 2/K]$. To get a prior $P \in \mathcal{P}_K^{\triangle}$ with these marginals, if K is even, we can pair up indices so that $x_{2j-1} = 2/K - x_{2j}$ for all $j = 1,\dots,K/2$ (for odd K, set $x_K = 1/K$) and then symmetrize by permuting the indices. See Appendix F in the supplementary material for more details.

The dependence on N is worse than N^{-2} resulting in $\widetilde{L}(p,f) = \infty$. This shows theoretical suboptimality of the uniform quantizer. Note also that the quadratic dependence on K is significantly worse than the $\log^2 K$ dependence achieved by the minimax compander.

Incidentally, other single-letter priors such as $p(x) = (1-\alpha)x^{-\alpha}$ where $\alpha = \frac{K-2}{K-1}$ can achieve worse dependence on N (specifically, $N^{-(2-\alpha)}$ for this prior). However, the example above achieves a bad dependence on both N and K simultaneously, showing that in all regimes of K, N the uniform quantizer is vulnerable to bad priors.

Remark 6: Instead of the KL divergence loss on the simplex, we can do a similar analysis to find the minimax compander for L_2^2 loss on the unit hypercube. The solution is given by the identity function $f(x) = x$ corresponding to the standard (non-companded) uniform quantization. (See Section VI.)

To show Theorem 1 we formulate and show a number of intermediate results which are also of significant interest for a theoretical understanding of companding under KL divergence, in particular studying the asymptotic behavior of $\widetilde{L}(p,f,N)$ as $N\to\infty$. We define:

Definition 2: For $p \in \mathcal{P}$ and $f \in \mathcal{F}$, let

$$L^\dagger(p,f) = \frac{1}{24}\int_0^1 p(x)f'(x)^{-2}x^{-1}\,dx$$

$$= \mathbb{E}_{X\sim p}\Big[\frac{1}{24}f'(X)^{-2}X^{-1}\Big]. \tag{13}$$

For full rigor, we also need to define a set of 'well-behaved' companders:

Definition 3: Let $\mathcal{F}^\dagger \subseteq \mathcal{F}$ be the set of f such that for each f there exist constants $c > 0$ and $\alpha \in (0, 1/2]$ for which $f(x) - cx^\alpha$ is still monotonically increasing.

Then the following describes the asymptotic single-letter loss of compander f on prior p (with centroid decoding):

Theorem 2: For any $p \in \mathcal{P}$ and $f \in \mathcal{F}$,

$$\liminf_{N\to\infty} N^2\widetilde{L}(p,f,N) \ge L^\dagger(p,f). \tag{14}$$

Furthermore, if $f \in \mathcal{F}^\dagger$ then an exact result holds:

$$\widetilde{L}(p,f) = L^\dagger(p,f) < \infty. \tag{15}$$

The intuition behind the formula for $L^\dagger(p,f)$ is that as $N\to\infty$, the density p becomes roughly uniform within each bin $I^{(n)}$. Additionally, the bin containing a given $x \in [0,1]$ will have width $r_{(n)} \approx N^{-1}f'(x)^{-1}$. Then, letting $\mathrm{unif}_{I^{(n)}}$ be the uniform distribution over $I^{(n)}$ and $\bar{y}_{(n)} \approx x$ be the midpoint of $I^{(n)}$ (which is also the centroid under the uniform distribution), we apply the approximation

$$\mathbb{E}_{X\sim\mathrm{unif}_{I^{(n)}}}\big[X\log(X/\bar{y}_{(n)})\big] \approx \frac{1}{24}r_{(n)}^2\bar{y}_{(n)}^{-1}$$

$$\approx \frac{1}{24}N^{-2}f'(x)^{-2}x^{-1}.$$

Averaging over $X \sim p$ and multiplying by N^2 then gives (13). One wrinkle is that we need to use the Dominated Convergence Theorem to get the exact result (15), but we cannot necessarily apply it for all $f \in \mathcal{F}$; instead, we can apply it for all $f \in \mathcal{F}^\dagger$, and outside of $\mathcal{F}^\dagger$ we get (14) using Fatou's Lemma.

While limiting ourselves to $f \in \mathcal{F}^\dagger$ might seem like a serious restriction, it does not lose anything essential because $\mathcal{F}^\dagger$ is 'dense' within $\mathcal{F}$ in the following way:

Proposition 5: For any $f \in \mathcal{F}$ and $\delta \in (0, 1]$,

$$f_\delta(x) \ = (1-\delta)f(x) + \delta x^{1/2} \tag{16}$$

satisfies $f_\delta \in \mathcal{F}^\dagger$ and

$$\lim_{\delta\to 0} \widetilde{L}(p, f_\delta) = \lim_{\delta\to 0} L^\dagger(p, f_\delta) = L^\dagger(p, f).$$

Remark 7: It is important to note that strictly speaking the limit represented by $\widetilde{L}(p, f)$ may not always exist if $f \notin \mathcal{F}^\dagger$. However: (i) one can always guarantee that it exists by selecting $f \in \mathcal{F}^\dagger$; (ii) by (14), it is impossible to use f outside $\mathcal{F}^\dagger$ to get asymptotic performance better than $L^\dagger(p, f)$; and (iii) by Proposition 5, given f outside $\mathcal{F}^\dagger$, one can get a compander in $\mathcal{F}^\dagger$ with arbitrarily close (or better) performance to f by using $f_\delta(x) = (1-\delta)f(x) + \delta x^{1/2}$ for δ close to 0. This suggests that considering only $f \in \mathcal{F}^\dagger$ is sufficient since there is no real way to benefit by using $f \notin \mathcal{F}^\dagger$.

Additionally, both f_K^* and f_K^{**} are in $\mathcal{F}^\dagger$. Thus, in Theorem 1, although the limit might not exist for certain $f \in \mathcal{F}, p \in \mathcal{P}_{1/K}$, the minimax compander still performs better since it has less loss than even the lim inf of the loss of other companders.

Given Theorem 2, it's natural to ask: for a given $p \in \mathcal{P}$, what compander f minimizes $L^\dagger(p, f)$? This yields the following by calculus of variations:

Theorem 3: The best loss against source $p \in \mathcal{P}$ is

$$\begin{aligned}\inf_{f\in\mathcal{F}} \widetilde{L}(p, f) &= \min_{f\in\mathcal{F}} L^\dagger(p, f) \\ &= \frac{1}{24}\Big(\int_0^1 \Big(p(x)x^{-1}\Big)^{1/3} dx\Big)^3 \end{aligned} \tag{17}$$

where the *optimal compander against* p is

$$f_p(x) = \arg\min_{f\in\mathcal{F}} L^\dagger(p, f) = \frac{\int_0^x \left(p(t)t^{-1}\right)^{1/3} dt}{\int_0^1 \left(p(t)t^{-1}\right)^{1/3} dt} \tag{18}$$

(satisfying $f_p'(x) \propto (p(x)x^{-1})^{1/3}$).

Note that f_p may not be in $\mathcal{F}^\dagger$ (for instance, if p assigns zero probability mass to an interval $I \subseteq [0, 1]$, then f_p will be constant over I). However, this can be corrected by taking a convex combination with $x^{1/2}$ as described in Proposition 5.

The expression (17) represents in a sense how hard $p \in \mathcal{P}$ is to quantize with a compander, and the maximin density p_K^* is the density in $\mathcal{P}_{1/K}$ which maximizes it;[5] in turn, the minimax compander f_K^* is the optimal compander against p_K^*, i.e.,

$$f_K^* = f_{p_K^*}.$$

So far we considered quantization of a random probability vector with a known prior. We next consider the case where the quantization guarantee is given pointwise, i.e., we cover $\triangle_{K-1}$ with a finite number of KL divergence balls of fixed radius. Note that since the prior is unknown, only the midpoint decoder can be used.

[5]The maximizing density over all $p \in \mathcal{P}$ happens to be $p(x) = \frac{1}{2}x^{-1/2}$; however, $\mathbb{E}_{X\sim p}[X] = 1/3$ so it cannot be the marginal of any symmetric $P \in \mathcal{P}_K^\triangle$ when $K > 3$.

Theorem 4 (Divergence Covering): On alphabet size $K > 4$ and $N \geq 8\log(2\sqrt{K\log K} + 1)$ intervals, the minimax and approximate minimax companders with midpoint decoding achieve *worst-case loss* over $\triangle_{K-1}$ of

$$\max_{\boldsymbol{x}\in\triangle_{K-1}} D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{z}) \leq (1 + \mathrm{err}(K))N^{-2}\log^2 K$$

where $\mathrm{err}(K)$ is an error term satisfying

$$\mathrm{err}(K) \leq 18\frac{\log\log K}{\log K} \leq 7 \text{ when } K > 4.$$

Note that the non-asymptotic worst-case bound matches (up to a constant factor) the known-prior asymptotic result (12). We remark that condition on N is mild: for example, if $N = 256$ (i.e., we are representing the probability vector with 8 bits per entry), then $N > 8\log(2\sqrt{K\log K} + 1)$ for all $K \leq 2.6 \times 10^{25}$.

Remark 8: When b is the number of bits used to quantize each value in the probability vector, using the approximate minimax compander yields a worst-case loss on the order of $2^{-2b}\log^2 K$. In [5] we prove bounds on the optimal loss under arbitrary (vector) quantization of probability vectors and show that this loss is sandwiched between $2^{-2b\frac{K}{K-1}}$ ([5, Proposition 2]) and $2^{-2b\frac{K}{K-1}}\log K$ ([5, Th. 2]). Thus, the entrywise companders in this work are quite competitive.

We also consider the natural family of *power companders* $f(x) = x^s$, both in terms of average asymptotic raw loss and worst-case non-asymptotic normalized loss. By definition, $f(x) \in \mathcal{F}^\dagger$ and hence $\widetilde{L}(p, f)$ is well-defined and Theorem 2 applies.

Theorem 5: The power compander $f(x) = x^s$ with exponent $s \in (0, 1/2]$ has asymptotic loss

$$\sup_{p\in\mathcal{P}_{1/K}} \widetilde{L}(p, f) = \frac{1}{24}s^{-2}K^{2s-1}. \tag{19}$$

For $K > 7$, (19) is minimized by setting $s = \frac{1}{\log K}$ (when $K \leq 7$, $\frac{1}{\log K} > 1/2$) and $f(x) = x^s$ achieves

$$\sup_{p\in\mathcal{P}_{1/K}} \widetilde{L}(p, f) = \frac{e^2}{24}\frac{1}{K}\log^2 K$$

$$\text{and} \quad \sup_{P\in\mathcal{P}_K^\triangle} \widetilde{\mathcal{L}}(P, f) = \frac{e^2}{24}\log^2 K.$$

Additionally, when $s = \frac{1}{\log K}$, it achieves the following worst-case bound with midpoint decoding for $K > 7$ and $N > \frac{e}{2}\log K$:

$$\max_{\boldsymbol{x}\in\triangle_{K-1}} D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{z}) \leq (1 + \mathrm{err}(K, N))\frac{e^2}{2}N^{-2}\log^2 K$$

where $\mathrm{err}(K, N) = \frac{e}{2}\frac{\log K}{N - \frac{e}{2}\log K}$.

Note in particular that when $N \geq e\log K$, we have $\mathrm{err}(K, N) \leq 1$, giving a bound of $\max_{\boldsymbol{x}\in\triangle_{K-1}} D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{z}) \leq e^2N^{-2}\log^2 K$.

We can think of $s = \frac{1}{\log K}$ as a 'minimax' among the class of power companders. This result shows $f(x) = x^{\frac{1}{\log K}}$ has performance within a constant factor of the minimax compander, and hence might be a good alternative.

B. Experimental Results

We compare the performance of five quantizers, with granularities $N = 2^8$ and $N = 2^{16}$, on three types of datasets of various alphabet sizes:

- Random synthetic distributions drawn from the uniform prior over the simplex: We draw and take the average over 1000 random samples for our results.
- Frequency of words in books: These frequencies are computed from text available on the Natural Language Toolkit (NLTK) libraries for Python. For each text, we get tokens (single words or punctuation) from each text and simply count the occurrence of each token
- Frequency of k-mers in DNA: For a given sequence of DNA, the set of k-mers are the set of length k substrings which appear in the sequence. We use the human genome as the source for our DNA sequences. Parts of the sequence marked as repeats are removed.

Our quantizers are:

- *Approximate Minimax Compander:* As given by equation (9). Using the approximate minimax compander is much simpler than the minimax compander since the constant c_K does not need to be computed.
- *Truncation:* Uniform quantization (equivalent to $f(x) = x$), which truncates the least significant bits. This is the natural way of quantizing values in [0, 1].
- *Float and bfloat16:* For 8-bit encodings ($N = 2^8$), we use a floating point implementation which allocates 4 bits to the exponent and 4 bits to the mantissa. For 16-bit encodings ($N = 2^{16}$), we use bfloat16, a standard which is commonly used in machine learning [6].
- *Exponential Density Interval (EDI):* This is the quantization method we used in an achievability proof in [1]. It is designed for the uniform prior over the simplex.
- *Power Compander:* Recall that the compander is $f(x) = x^s$. We optimize s and find that $s = \frac{1}{\log_e K}$ asymptotically minimizes KL divergence, and also gives close to the best performance among power companders empirically. To see the effects of different powers s on the performance of the power compander, see Figure 1.

Because a well-defined prior does not always exist for these datasets (and for simplicity) we use midpoint decoding for all the companders. When a probability value of exactly 0 appears, we do not use companding and instead quantize the value to 0, i.e., the value 0 has its own bin.

Our main experimental results are given in Figure 2, showing the KL divergence between the empirical distribution $\boldsymbol{x}$ and its quantized version $\boldsymbol{z}$ versus alphabet size K. The approximate minimax compander performs well against all sources. For truncation, the KL divergence increases with K and is generally fairly large. The EDI quantizer works well for the synthetic uniform prior (as it should), but for real-world datasets like word frequency in books, it performs badly (sometimes even worse than truncation). The loss of the power compander is similar to the minimax compander (only worse by a constant factor), as predicted by Theorem 5.

The experiments show that the approximate minimax compander achieves low loss on the entire ensemble of data (even for relatively small granularity, such as $N = 256$) and

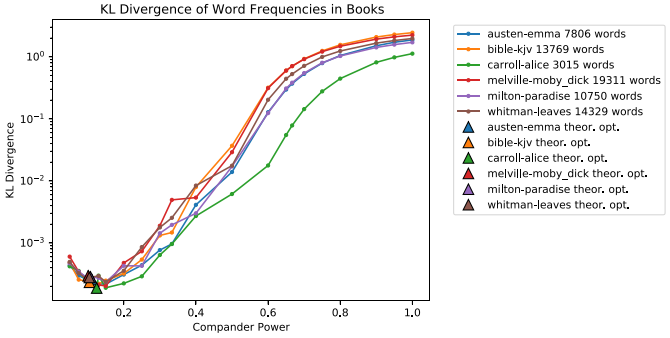

Fig. 1. Power compander $f(x) = x^s$ performance with different powers s used to quantize frequency of words in books. The number K of distinct words in each book is shown in the legend. The theoretical optimal power $s = \frac{1}{\log K}$ is plotted.

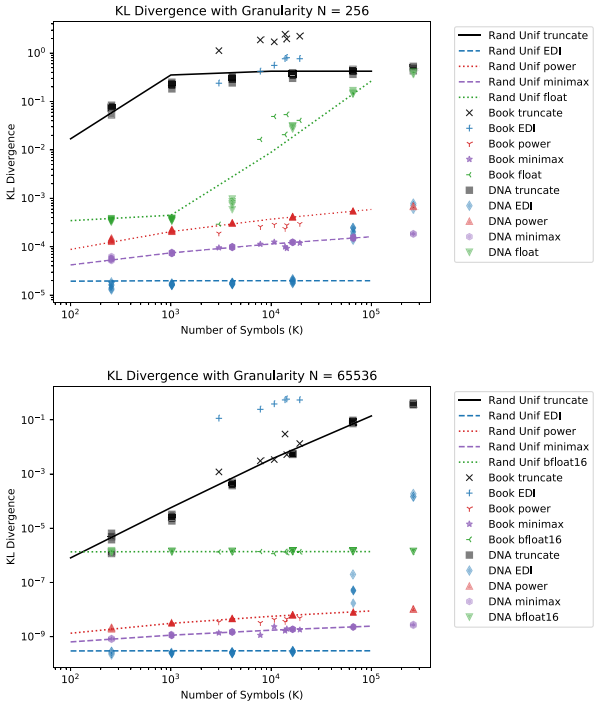

Fig. 2. Plot comparing the performance of the truncation compander, the EDI compander, floating points, the power compander, and the approximate minimax compander (9) on probability distributions of various sizes.

outperforms both truncation and floating-point implementations on the same number of bits. Additionally, its closed-form expression (and entrywise application) makes it simple to implement and computationally inexpensive, so it can be easily added to existing systems to lower storage requirements at little or no cost to fidelity.

C. Paper Organization

We provide background and discuss previous work on companders in Section III. We prove Theorem 2 in Section IV (though proofs of some lemmas and propositions leading up to it are given in Appendix A in the supplementary material). Proposition 5 is proved in Appendix B in the supplementary material. In Section V, we optimize over (13) to get the maximin single-letter distribution (showing part of Proposition 2 with other parts left to Appendix D-A in the supplementary material) and the minimax compander, thus

showing Theorem 1, 3, Corollary 1 and Proposition 3 (leaving Theorem 4 for Appendix D-B in the supplementary material). We prove Theorem 4 and the worst-case part of Theorem 5 in Appendix E in the supplementary material. Other parts of Theorem 5 are discussed in Appendix C-B in the supplementary material. In Section VI we discuss companders for losses other than KL divergence. Finally, in Section VII we discuss a connection of our problem to the problem of information distillation with proofs given in Appendix G in the supplementary material.

III. Background

Companders (also spelled "compandors") were introduced by Bennett in 1948 [2] as a way to quantize speech signals, where it is advantageous to give finer quantization levels to weaker signals and coarser levels to larger signals. Bennett gives a first order approximation that the mean-square error in this system is given by

$$\frac{1}{12N^2}\int_a^b \frac{p(x)}{(f'(x))^2}dx \tag{20}$$

where N is the number quantization levels, a and b are the minimum and maximum values of the input signal, p is the probability density of the input signal, and f' is the slope of the compressor function placed before the uniform quantization. This formula is similar to our (13) except that we have an extra x^{-1} since we are working with KL divergence. Others have expanded on this line of work. In [7], the authors studied the same problem and determined the optimal compressor under mean-square error, a result which parallels our result (17). However, results like those in [2], [7] are stated either as first order approximations or make simplifying assumptions. For example, in [7], the authors state that they assume the values $\widehat{y}_{(n)}$ are close together enough that probability density within any given bin can be treated as a constant. In contrast, we rigorously show that this fundamental logic holds under very general conditions ($f \in \mathcal{F}^\dagger$).

Generalizations of Bennett's formula are also studied when instead of mean-square error, the loss is the expected rth moment loss $\mathbb{E}\|\cdot\|^r$. This is computed for vectors of length K in [8] and [9].

The typical examples of companders used in engineering and signals processing are the μ-law and A-law companders [10]. For the μ-law compander, [7] and [11] argue that for mean-squared error, for a large enough constant μ the distortion becomes independent of the signal.

Quantizing probability distributions is a well-studied topic, though typically the loss function is a norm and not KL divergence [12]. Quantizing for KL divergence is considered in our earlier work [1], focusing on average KL loss for Dirichlet priors.

A similar problem to quantizing under KL divergence is *information k-means*. This is the problem of clustering n points a_i to k centers $\hat{a}_j$ to minimize the KL divergences between the points and their associated centers. Theoretical aspects of this are explored in [13] and [14]. Information k-means has been implemented for several different applications [15], [16], [17]. There are also other works that study clustering with a slightly different but related metric [18], [19], [20]; however, the focus of these works is to analyze data rather than reduce storage.

Remark 9: A variant of the classic problem of prediction with log-loss is an equivalent formulation to quantizing the simplex with KL loss: let $\boldsymbol{x} \in \triangle_{K-1}$ and $A \sim \boldsymbol{x}$ (in the alphabet $[K]$); we want to predict A by positing a distribution $\boldsymbol{z} \in \triangle_{K-1}$, and our loss is $-\log z_A$. In the standard version, the problem is to pick the best $\boldsymbol{z}$ given limited information about $\boldsymbol{x}$; however, if we *know* $\boldsymbol{x}$ but are required to express $\boldsymbol{z}$ using only $\log_2 M$ bits, it is equivalent to quantizing the simplex with KL divergence loss.

IV. Asymptotic Single-Letter Loss

In this section we give the proof of Theorem 2 (though the proofs of some lemmas must be sketched). We use the following notation:

Given an interval I we define $\bar{y}_I$ to be its midpoint and r_I to be its width, so that by definition

$$I = \left[\bar{y}_I - r_I/2, \bar{y}_I + r_I/2\right].$$

Note that if $I \subseteq [0,1]$ then $r_I \le 2\bar{y}_I$.

Given probability distribution p and interval I, we denote the following: $p|_I$ is p restricted to I; $\pi_{p,I} \coloneqq \mathbb{P}_{X\sim p}[X \in I]$ is the probability mass of I; and the *centroid of I under p* is

$$\widetilde{y}_{p,I} \coloneqq \mathbb{E}_{X\sim p|_I}[X] = \mathbb{E}_{X\sim p}[X \,|\, X \in I].$$

If they are undefined because $\mathbb{P}_{X\sim p}[X \in I] = 0$ then by convention $p|_I$ is uniform on I and $\widetilde{y}_{p,I} = \bar{y}_I$.

When $I = I^{(n)}$ is a bin of the compander, we can replace it with (n) in the notation, i.e., $\bar{y}_{(n)} = \bar{y}_{I^{(n)}}$ (so the midpoint of the bin containing x at granularity N is denoted $\bar{y}_{(n_N(x))}$ and the width of the bin is $r_{(n_N(x))}$). When I and/or p are fixed, we sometimes drop them from the notation, i.e., $\widetilde{y}_I$ or even just $\widetilde{y}$ to denote the centroid of I under p.

A. The Local Loss Function

One key to the proof is the following perspective: instead of considering $X \sim p$ directly, we (equivalently) first select bin $I^{(n)}$ with probability $\pi_{p,(n)}$, and then select $X \sim p|_{(n)}$. The expected loss can then be considered within bin $I^{(n)}$. This makes it useful to define:

Definition 4: Given probability measure p and interval I, the *single-interval loss of I under p* is

$$\ell_{p,I} = \mathbb{E}_{X\sim p|_I}\left[X\log\left(X/\widetilde{y}_{p,I}\right)\right].$$

As before, if p and/or I is fixed and clear, we can drop it from the notation (and if $I = I^{(n)}$ is a bin, we can denote the local loss as $\ell_{p,(n)}$). This can be interpreted as follows: if we quantize all $x \in I$ to the centroid $\widetilde{y}_I$, then $\ell_{p,I}$ is the expected loss of $X \sim p$ conditioned on $X \in I$. Thus the values of $\ell_{p,(n)}$ can be used as an alternate means of computing the single-letter loss:

$$\begin{aligned}\widetilde{L}(p,f,N) &= \mathbb{E}_{X\sim p}\left[X\log(X/\widetilde{y}(X))\right]\\ &= \sum_{n=1}^{N}\pi_{p,(n)}\mathbb{E}_{X\sim p|_{(n)}}\left[X\log\left(X/\widetilde{y}_{p,(n)}\right)\right]\\ &= \sum_{n=1}^{N}\pi_{p,(n)}\ell_{p,(n)} = \int_{[0,1]}\ell_{p,(n_N(x))}\,dp.\end{aligned}$$

Thus the normalized single-letter loss (whose limit is the asymptotic single-letter loss (5)) is

$$N^2 \widetilde{L}(p,f,N) = \int_{[0,1]} N^2 \, \ell_{p,(n_N(x))} \, dp.$$

For single-letter density p and compander f, we define the *local loss function at granularity N*:

$$g_N(x) = N^2 \, \ell_{p,(n_N(x))}. \tag{21}$$

We also define the *asymptotic local loss function*:

$$g(x) = \frac{1}{24} f'(x)^{-2} x^{-1}.$$

Theorem 2 is therefore equivalent to:

$$\liminf_{N\to\infty} \int g_N \, dp \geq \int g \, dp \;\; \forall \; p \in \mathcal{P}, f \in \mathcal{F} \tag{22}$$

$$\text{and } \lim_{N\to\infty} \int g_N \, dp = \int g \, dp \;\; \forall \; p \in \mathcal{P}, f \in \mathcal{F}^\dagger. \tag{23}$$

To prove (22) and (23), we show:

Proposition 6: For all $p \in \mathcal{P}, f \in \mathcal{F}$, if $X \sim p$ then

$$\lim_{N\to\infty} g_N(X) = g(X) \quad \text{almost surely.}$$

Proposition 7: Let $f \in \mathcal{F}^\dagger$ be a compander and $c > 0$ and $\alpha \in (0,1]$ such that $f(x) - cx^\alpha$ is monotonically increasing. Letting g_N be the local loss functions as in (21) and

$$h(x) = \left(2^{2/\alpha} + \alpha^2 2^{1/\alpha - 2}\right)(c\alpha)^{-2} x^{1-2\alpha} + c^{-1/\alpha} 2^{1/\alpha - 2}$$

then $g_N(x) \leq h(x)$ for all x, N. Additionally, if $\alpha \leq 1/2$ then $\int_{[0,1]} h \, dp < \infty$.

The lower bound (22) then follows immediately from Proposition 6 and Fatou's Lemma; and when $f \in \mathcal{F}^\dagger$, by Proposition 7 there is some h which is integrable over p and dominates all g_N, thus showing (23) by the Dominated Convergence Theorem.

To prove Proposition 6, we use the following:

- For any x at which f is differentiable, when N is large, the width of the interval x falls in is
$$r_{(n_N(x))} \approx N^{-1} f'(x)^{-1}.$$
- For any x at which F_p is differentiable, $p|_I$ will be approximately uniform over any sufficiently small I containing x.
- For a sufficiently small interval I containing x and such that $p|_I$ is approximately uniform,
$$\ell_{p,I} \approx \frac{1}{24} r_I^2 x^{-1}.$$

Putting these together, we get that if F_p and f are both differentiable at x then when N is large,

$$\begin{aligned} g_N(x) &= N^2 \, \ell_{p,(n_N(x))} \\ &\approx N^2 \frac{1}{24} r_{(n_N(x))}^2 x^{-1} \approx \frac{1}{24} f'(x)^{-2} x^{-1} = g(x) \end{aligned}$$

as we wanted. We formally state each of these steps in Appendix A-B in the supplementary material and combine them to prove Proposition 6 in Appendix A-C in the supplementary material.

The proof of Proposition 7 is given in Appendix A-D in the supplementary material, along with its own set of definitions and lemmas needed to show it.

V. Minimax Compander

Theorem 2 showed that for $f \in \mathcal{F}^\dagger$, the asymptotic single-letter loss is equivalent to

$$\widetilde{L}(p,f) = \frac{1}{24} \int_0^1 p(x) f'(x)^{-2} x^{-1} dx.$$

Using this, we can analyze what is the 'best' compander f we can choose and what is the 'worst' single-letter density p in order to show Theorem 1 and 3 and their related results.

A. Optimizing the Compander

We show Theorem 3, which follows from Theorem 2 by finding $f \in \mathcal{F}$ which minimizes $L^\dagger(p,f)$. This is achieved by optimizing over f'; we will also use some concepts from Proposition 5 to connect it back to $\inf_{f\in\mathcal{F}} \widetilde{L}(p,f)$ when the resulting f is not in $\mathcal{F}^\dagger$. Since $f : [0,1] \to [0,1]$ is monotonic, we use constraints $f'(x) \geq 0$ and $\int_0^1 f'(x)\,dx = 1$. We solve the following:

$$\begin{aligned} \text{minimize} \quad & L^\dagger(p,f) = \frac{1}{24} \int_0^1 p(x) f'(x)^{-2} x^{-1} \, dx \\ \text{subject to} \quad & \int_0^1 f'(x)\,dx = 1 \\ & \text{and } f'(x) \geq 0 \text{ for all } x \in [0,1]. \end{aligned}$$

The function $L^\dagger(p,f)$ is convex in f', and thus first order conditions show optimality. Let $\lambda(x)$ satisfy $\int_0^1 \lambda(x) dx = 0$. If $f'(x) \propto (p(x)x^{-1})^{1/3}$, we derive:

$$\begin{aligned} &\frac{d}{dt} \frac{1}{24} \int_0^1 p(x) \big(f'(x) + t\,\lambda(x)\big)^{-2} x^{-1} \, dx \\ &\quad = \frac{1}{24} \int_0^1 p(x) x^{-1} \frac{d}{dt} \big(f'(x) + t\,\lambda(x)\big)^{-2} dx \\ &\quad = -\frac{1}{12} \int_0^1 p(x) x^{-1} \big(f'(x) + t\,\lambda(x)\big)^{-3} \lambda(x) \, dx \\ &\quad = -\frac{1}{12} \int_0^1 p(x) x^{-1} f'(x)^{-3} \lambda(x) \, dx \quad (\text{at } t = 0) \\ &\quad \propto -\frac{1}{12} \int_0^1 \lambda(x) \, dx = 0. \end{aligned} \tag{24}$$

Thus, such f satisfies the first-order optimality condition under the constraint $\int f'(x)\,dx = 1$. This gives $f_p'(x) \propto (p(x)x^{-1})^{1/3}$ and $f(0) = 0$ and $f(1) = 1$, from which (17) and (18) follow. If $f_p \in \mathcal{F}^\dagger$, then $f_p = \arg\min_f \widetilde{L}(p,f)$, and for any other $f \in \mathcal{F}$,

$$\begin{aligned} \widetilde{L}(p,f_p) = L^\dagger(p,f_p) &\leq L^\dagger(p,f) \\ &\leq \liminf_{N\to\infty} N^2 \widetilde{L}(p,f,N). \end{aligned}$$

If $f_p \notin \mathcal{F}^\dagger$, for any $\delta > 0$ define $f_{p,\delta} = (1-\delta)f_p + \delta x^{1/2}$ (as in (16)). Then $f_{p,\delta} - \delta x^{1/2} = (1-\delta)f_p$ is monotonically increasing so $f_{p,\delta} \in \mathcal{F}^\dagger$, so Theorem 2 applies to $f_{p,\delta}$; additionally, $f_{p,\delta} - (1-\delta)f_p = \delta x^{1/2}$ is monotonically increasing as well so $f'_{p,\delta} \geq (1-\delta)f'_p$. Hence, plugging into the $L^\dagger$ formula gives:

$$\widetilde{L}\big(p, f_{p,\delta}\big) = L^\dagger\big(p, f_{p,\delta}\big) \leq L^\dagger\big(p, f_p\big)(1-\delta)^{-2}.$$

Taking $\delta \to 0$ (and since $\mathcal{F}^\dagger \subseteq \mathcal{F}$) shows that

$$L^\dagger\big(p, f_p\big) = \inf_{f \in \mathcal{F}^\dagger} \widetilde{L}(p, f),$$

finishing the proof of Theorem 3.

Remark 10: Since we know the corresponding single-letter source p for a Dirichlet prior, using this p with Theorem 3 gives us the optimal compander for Dirichlet priors on any alphabet size. This gives us a better quantization method than EDI which was discussed in Section II-B. This optimal compander for Dirichlet priors is called the *beta compander* and its details are given in Appendix C-A in the supplementary material.

B. The Minimax Companders and Approximations

To prove Theorem 1 and Corollary 1, we first consider what density p maximizes equation (17):

$$\frac{1}{24}\left(\int_0^1 \Big(p(x)x^{-1}\Big)^{1/3} dx\right)^3$$

i.e., is most difficult to quantize with a compander. Using calculus of variations to maximize

$$\int_0^1 \Big(p(x)x^{-1}\Big)^{1/3} dx \tag{25}$$

(which of course maximizes (17)) subject to $p(x) \geq 0$ and $\int_0^1 p(x)\,dx = 1$, we find that maximizer is $p(x) = \frac{1}{2}x^{-1/2}$. However, while interesting, this is only for a single letter; and because $\mathbb{E}[X] = 1/3$ under this distribution, it is clearly impossible to construct a prior over $\triangle_{K-1}$ (whose output vector *must* sum to 1) with this marginal (unless $K = 3$).

Hence, we add an expected value constraint to the problem of maximizing (25), giving:

$$\begin{aligned} \text{maximize} \quad & \int_0^1 \Big(p(x)x^{-1}\Big)^{1/3} dx \\ \text{subject to} \quad & \int_0^1 p(x)\,dx = 1; \qquad (26) \\ & \int_0^1 p(x)x\,dx = \frac{1}{K}; \qquad (27) \\ & \text{and } p(x) \geq 0 \text{ for all } x. \end{aligned}$$

We can solve this again using variational methods (we are maximizing a concave function so we only need to satisfy first-order optimality conditions). A function $p(x) > 0$ is optimal if, for any $\lambda(x)$ where

$$\int_0^1 \lambda(x)\,dx = 0 \text{ and } \int_0^1 \lambda(x)x\,dx = 0$$

the following holds:

$$\frac{d}{dt}\int_0^1 x^{-1/3}\big(p(x) + t\,\lambda(x)\big)^{1/3}\,dx = 0.$$

We have by the same logic as before:

$$\begin{aligned} &\frac{d}{dt}\int_0^1 x^{-1/3}\big(p(x) + t\,\lambda(x)\big)^{1/3}\,dx \\ &\quad = \frac{1}{3}\int_0^1 x^{-1/3}\big(p(x) + t\,\lambda(x)\big)^{-2/3}\lambda(x)\,dx \\ &\quad = \frac{1}{3}\int_0^1 x^{-1/3}p(x)^{-2/3}\lambda(x)\,dx \text{ (at } t = 0). \end{aligned} \tag{28}$$

Thus, if we can arrange things so that there are constants a_K, b_K such that

$$x^{-1/3}p(x)^{-2/3} = a_K + b_K x$$

this ensures (28) equals zero. In that case,

$$\begin{aligned} & x^{-1/3}p(x)^{-2/3} = a_K + b_K x \\ \iff\ & p(x)^{-2/3} = a_K x^{1/3} + b_K x^{4/3} \\ \iff\ & p(x) = \big(a_K x^{1/3} + b_K x^{4/3}\big)^{-3/2}. \end{aligned} \tag{29}$$

This is the maximin density p_K^* from Proposition 2 (8), where a_K, b_K are set to meet the constraints (26) and (27). Exact formulas for a_K, b_K are difficult to find; we give more details on after the next step.

We want to determine the optimal compander for the maximin density (29). We know from (24) that we need to first compute

$$\begin{aligned} \phi(x) &= \int_0^x w^{-1/3}\Big(a_K w^{1/3} + b_K w^{4/3}\Big)^{-1/2} dw \\ &= \frac{2\text{ArcSinh}\Big(\sqrt{\frac{b_K x}{a_K}}\Big)}{\sqrt{b_K}}. \end{aligned} \tag{30}$$

The best compander $f(x)$ is proportional to (30) and is exactly given by $f(x) = \phi(x)/\phi(1)$. The resulting compander, which we call the *minimax compander*, is

$$f(x) = \frac{\text{ArcSinh}\Big(\sqrt{\frac{b_K x}{a_K}}\Big)}{\text{ArcSinh}\Big(\sqrt{\frac{b_K}{a_K}}\Big)}. \tag{31}$$

Given the form of $f(x)$, it is natural to determine an expression for the ratio b_K/a_K. We can parameterize both a_K and b_K by b_K/a_K and then examine how b_K/a_K behaves as a function of K. The constraints on a_K and b_K give that

$$\begin{aligned} a_K &= 4^{1/3}(b_K/a_K + 1)^{-1/3} \\ b_K &= 4a_K^{-2} - a_K. \end{aligned}$$

The ratio b_K/a_K grows approximately as $K \log K$. Hence, we choose to parameterize

$$b_K/a_K = c_K K \log K.$$

To satisfy the constraints, we get $0.25 < c_K < 0.75$ so long as $K > 24$ (see Appendix D-A in the supplementary material for details), and Lemma 11 in Appendix D-A2 in the supplementary material shows that $c_K \to 1/2$ as $K \to \infty$. Combining these gives Proposition 2.

We can then express a_K, b_K in terms of c_K:

$$\begin{aligned} a_K &= 4^{1/3}(c_K K\log K+1)^{-1/3} \\ b_K &= 4a_K^{-2} - a_K \\ &= 4^{1/3}(c_K K\log K+1)^{2/3} - 4^{1/3}(c_K K\log K+1)^{-1/3} \\ &= 4^{1/3}(c_K K\log K)^{2/3}(1+o(1)). \end{aligned} \tag{32}$$

When K is large, the second term in (32) is negligible compared to the first. Thus, plugging into (31) we get the minimax compander and approximate minimax compander, respectively:

$$\begin{aligned} f_K^*(x) &= \frac{\mathrm{ArcSinh}\left(\sqrt{(c_K K\log K)x}\right)}{\mathrm{ArcSinh}\left(\sqrt{c_K K\log K}\right)} \\ \approx f_K^{**}(x) &= \frac{\mathrm{ArcSinh}\left(\sqrt{((1/2)K\log K)x}\right)}{\mathrm{ArcSinh}\left(\sqrt{(1/2)K\log K}\right)}. \end{aligned}$$

The minimax compander minimizes the maximum (raw) loss against all densities in $\mathcal{P}_{1/K}$, while the approximate minimax compander performs very similarly but is more applicable since it can be used without computing c_K.

To compute the loss of the minimax compander, we can use (17) to get

$$L^\dagger(p_K^*, f_K^*) = \frac{1}{24}\left(\frac{2\mathrm{ArcSinh}\left(\sqrt{c_K K\log K}\right)}{\sqrt{b_K}}\right)^3.$$

Substituting we get

$$\begin{aligned} &L^\dagger(p_K^*, f_K^*) \\ &= \frac{1}{24}\frac{8\left(\log\left(\sqrt{c_K K\log K}+\sqrt{c_K K\log K+1}\right)\right)^3}{2c_K K\log K(1+o(1))} \\ &= \frac{1}{24}\frac{(\log 4(c_K K\log K))^3}{2c_K K\log K}(1+o(1)) \\ &= \frac{1}{24}\frac{\log^2 K}{K}(1+o(1)). \end{aligned} \tag{33}$$

In fact, not only is f_K^* optimal against the maximin density p_K^*, but (as alluded to in the name 'minimax compander') it minimizes the maximum asymptotic loss over all $p \in \mathcal{P}_{1/K}$. More formally we show that (f_K^*, p_K^*) is a saddle point of $L^\dagger$.

The function $L^\dagger(p,f)$ is concave (actually linear) in p and convex in f', and we can show that the pair (f_K^*, p_K^*) form a saddle point, thus proving (10)-(11) from Theorem 1.

We can compute that

$$\begin{aligned} \left(f_K^*\right)'(x) &\propto \left(p_K^*(x)x^{-1}\right)^{1/3} \\ &= x^{-1/3}\left(a_K x^{1/3} + b_K x^{4/3}\right)^{-1/2} \\ &= \frac{1}{\sqrt{a_K x + b_K x^2}}. \end{aligned}$$

Assume we set a_K and b_K to the appropriate values for K. For any $p \in \mathcal{P}_{1/K}$,

$$\begin{aligned} L^\dagger(p, f_K^*) &= \int_0^1 p(x)x^{-1}\left(\left(f_K^*\right)'(x)\right)^{-2}dx \\ &= \int_0^1 p(x)x^{-1}\left(a_K x + b_K x^2\right)dx \\ &= a_K + b_K\frac{1}{K} \end{aligned}$$

i.e., $L^\dagger(p, f_K^*)$ does not depend on p. Since f_K^* is the optimal compander against the maximin compander p_K^* we can therefore conclude:

$$\begin{aligned} &\sup_{p\in\mathcal{P}_{1/K}} L^\dagger(p, f_K^*) = L^\dagger(p_K^*, f_K^*) \\ &= \inf_{f\in\mathcal{F}} L^\dagger(p_K^*, f) = \sup_{p\in\mathcal{P}_{1/K}}\inf_{f\in\mathcal{F}} L^\dagger(p,f). \end{aligned}$$

Since it is always true that

$$\sup_{p\in\mathcal{P}_{1/K}}\inf_{f\in\mathcal{F}} L^\dagger(p,f) \le \inf_{f\in\mathcal{F}}\sup_{p\in\mathcal{P}_{1/K}} L^\dagger(p,f),$$

this shows that (f_K^*, p_K^*) is a saddle point.

Furthermore, $f_K^* \in \mathcal{F}^\dagger$ (specifically it behaves as a multiple of $x^{1/2}$ near 0), so $\widetilde{L}(p, f_K^*) = L^\dagger(p, f_K^*)$ for all p, thus showing that f_K^* performs well against any $p \in \mathcal{P}_{1/K}$. Using (13) with the expressions for p_K^* and f_K^* and (33) gives (12). This completes the proof of Theorem 1.

Remark 11: While the power compander $f(x) = x^{1/\log K}$ is not minimax optimal, it has similar properties to the minimax compander and differs in loss by at most a constant factor. We analyze the power compander in Appendix C-B in the supplementary material.

C. Existence of Priors With Given Marginals

While p_K^* is the most difficult density in $\mathcal{P}_{1/K}$ to quantize, it is unclear whether a prior P^* on $\triangle_{K-1}$ exists with marginals p_K^* – even though K copies of p_K^* will correctly sum to 1 in expectation, it may not be possible to correlate them to guarantee they sum to 1. However, it is possible to construct a prior P^* whose marginals are as hard to quantize, up to a constant factor, as p_K^*, by use of clever correlation between the letters. We start with a lemma:

Lemma 1: Let $p \in \mathcal{P}_{1/K}$. Then there exists a joint distribution of $(X_1, \ldots, X_K)$ such that (i) $X_i \sim p$ for all $i \in [K]$ and (ii) $\sum_{i\in[K]} X_i \le 2$, guaranteed.

Proof: Let F be the cumulative distribution function of p. Define the quantile function F^{-1} as

$$F^{-1}(u) = \inf\{x : F(x) \ge u\}.$$

We break $[0,1]$ into K uniform sub-intervals $I_i = ((i-1)/K, i/K]$ (let $I_1 = [0, 1/K]$). We then generate $X_1, X_2, \ldots, X_K$ jointly by the following procedure:

1) Choose a permutation $\sigma : [K] \to [K]$ uniformly at random (from $K!$ possibilities).
2) Let $U_k \sim \mathrm{unif}_{I_{\sigma(k)}}$ independently for all k.
3) Let $X_k = F^{-1}(U_k)$.

Now we consider $\sum_k X_k$. Let $b_i = F^{-1}(i/k)$ for $i = 0, 1, \ldots, K$. Note that if $\sigma(k) = i$ then $U_k \in ((i-1)/K, i/K]$ and hence $X_k = F^{-1}(U_k) \in [b_{i-1}, b_i]$. Therefore $X_{\sigma^{-1}(i)} \in$

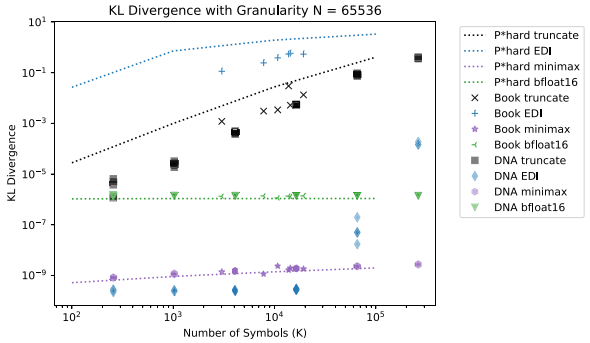

Fig. 3. Each compander (or quantization method) is used on random distributions drawn from the prior P^*_{hard}. Comparison is given to when each compander is used on the books and DNA datasets.

$[b_{i-1}, b_i]$ and thus for any permutation σ,

$$\sum_{i=1}^{K} b_{i-1} \le \sum_{i=1}^{K} X_{\sigma^{-1}(i)} \le \sum_{i=1}^{K} b_i$$
$$= \Big(\sum_{i=1}^{K} b_{i-1}\Big) + b_K - b_0$$
$$\le \Big(\sum_{i=1}^{K} b_{i-1}\Big) + 1 \le 2$$

as $\sum_i b_{i-1} \le \sum_i \mathbb{E}[X_{\sigma^{-1}(i)}] = K\mathbb{E}_{X\sim p}[X] = 1$. ■

Lemma 1 shows a joint distribution of $W_1, \ldots, W_{K-1}$ such that $W_i \sim p^*_K$ for all i and $\sum_{i=1}^{K-1} W_i \le 2$ (guaranteed) exists. Then, if $X_i = W_i/2$ for all $i \in [K-1]$, we have $\sum_{i=1}^{K-1} X_i \le 1$. Then setting $X_K = 1 - \sum_{i=1}^{K-1} X_i \ge 0$ ensures that $(X_1, \ldots, X_K)$ is a probability vector. Denoting this prior P^*_{hard} and letting $p^{**}_K(x) = 2p^*_K(2x)$ (so $W_i \sim p^*_K \implies X_i \sim p^{**}_K$) we get that

$$\inf_{f\in\mathcal{F}} \widetilde{\mathcal{L}}_K\big(P^*_{\text{hard}}, f\big) \ge (K-1) \inf_{f\in\mathcal{F}} \widetilde{L}\big(p^{**}_K, f\big) \tag{34}$$
$$= (K-1)\frac{1}{2}L^{\dagger}\big(p^*_K, f^*_K\big) \ge \frac{1}{2}\frac{K-1}{K} \sup_{P\in\mathcal{P}^{\triangle}_K} \widetilde{\mathcal{L}}_K\big(P, f^*_K\big). \tag{35}$$

The last inequality holds because p^*_K is the maximin density (under expectation constraints). To make P^*_{hard} symmetric, we permute the letter indices randomly without affecting the raw loss; thus we get Corollary 1. To get (35) from (34), we have

$$\inf_{f\in\mathcal{F}} \widetilde{L}\big(2p^*_K(2x), f\big) = \frac{1}{24}\left(\int_0^1 \Big(2p^*_K(2x)x^{-1}\Big)^{1/3} dx\right)^3$$
$$= \frac{1}{24}\left(\int_0^1 \Big(2p^*_K(u)2u^{-1}\Big)^{1/3} \frac{1}{2} du\right)^3$$
$$= \frac{1}{2}L^{\dagger}\big(p^*_K, f_*\big).$$

This shows Proposition 3. In Figure 3, we validate the distribution P^*_{hard} by showing the performance of each compander when quantizing random distributions drawn from P^*_{hard}. For the minimax compander, the KL divergence loss on the worst-case prior looks to be within a constant of that for the other datasets.

VI. Companding Other Metrics and Spaces

While our primary focus has been KL divergence over the simplex, for context we compare our results to what the same compander analysis would give for other loss functions like squared Euclidean distance (L_2^2) and absolute distance (L_1 or TV distance). For a vector $\boldsymbol{x}$ and its representation $\boldsymbol{z}$ let

$$L_2^2(\boldsymbol{x}, \boldsymbol{z}) = \sum_i (x_i - z_i)^2$$
$$L_1(\boldsymbol{x}, \boldsymbol{z}) = \sum_i |x_i - z_i|.$$

For squared Euclidean distance, asymptotic loss was already given by (20) in [2], and scales as N^{-2}. It turns out that the maximin single-letter distribution over a bounded interval is the uniform distribution. Thus, the minimax compander for L_2^2 is simply the identity function, i.e., uniform quantization is the minimax for quantizing a hypercube in high-dimensional space under L_2^2 loss. (For unbounded spaces, L_2^2 loss does not scale with N^{-2}.)

If we add the expected value constraint to the L_2^2 compander optimization problem, we can derive the best square distance compander for the probability simplex. For alphabet size K, we get that the minimax compander for L_2^2 is given by

$$f_{L_2^2,K}(x) = \frac{\sqrt{1 + K(K-2)x} - 1}{K-2}$$

and the total L_2^2 loss for probability vector $\boldsymbol{x}$ and its quantization $\boldsymbol{z}$ has the relation

$$\lim_{N\to\infty} N^2 L_2^2(\boldsymbol{x}, \boldsymbol{z}) \le \frac{1}{3}.$$

For L_1, unlike KL divergence and L_2^2, the loss scales as $1/N$. Like L_2^2, the minimax single-letter compander for L_1 loss in the hypercube $[0, 1]^K$ is the identity function, i.e., uniform quantization. In general, the derivative of the optimal compander for single-letter density $p(x)$ has the form

$$f'_{L_1,K}(x) \propto \sqrt{p(x)}.$$

On the probability simplex for alphabet size K, the worst case prior $p(x)$ has the form

$$p(x) = (\alpha_K x + \beta_K)^{-2}$$

where α_K, β_K are constants scaling to allow $\int_{[0,1]} dp = 1$ (i.e., p is a valid probability density) and $\int_{[0,1]} x\, dp = 1/K$ (i.e., $\mathbb{E}_{X\sim p}[X] = 1/K$ so K copies of it are expected to sum to 1).

Thus, the minimax compander on the simplex for L_1 loss (and letting $\gamma_K = \alpha_K/\beta_K$) satisfies

$$f'_{L_1,K}(x) \propto (\alpha_K x + \beta_K)^{-1}$$
$$\implies f_{L_1,K}(x) \propto \log((\alpha_K/\beta_K)x + 1)$$
$$\implies f_{L_1,K}(x) = \frac{\log(\gamma_K x + 1)}{\log(\gamma_K + 1)}$$

since $f_{L_1,K}(x)$ has to be scaled to go from 0 to 1.

The asymptotic L_1 loss for probability vector $\boldsymbol{x}$ and its quantization $\boldsymbol{z}$ is bounded by

$$\lim_{N\to\infty} N L_1(\boldsymbol{x}, \boldsymbol{z}) = O(\log K).$$

Loss	Space	Optimal Compander	Asymptotic Upper Bound
KL	Simplex	$f_K^*(x) = \frac{\mathrm{ArcSinh}(\sqrt{c_K(K\log K)}\,x)}{\mathrm{ArcSinh}(\sqrt{c_K K\log K})}$	$N^{-2}\log^2 K$
L_2^2	Simplex	$f_{L_2^2,K}(x) = \frac{\sqrt{1+K(K-2)x}-1}{K-2}$	N^{-2}
L_2^2	Hypercube	$f_{L_2^2}(x) = x$ (uniform quantizer)	$N^{-2}K$
L_1 (TV)	Simplex	$f_{L_1,K}(x) = \frac{\log(\gamma_K x+1)}{\log(\gamma_K+1)}$	$N^{-1}\log K$
L_1 (TV)	Hypercube	$f_{L_1}(x) = x$ (uniform quantizer)	$N^{-1}K$

Fig. 4. Summary of results for various losses and spaces. Asymptotic Upper Bound is an upper bound on how we expect the loss of the optimal compander to scale with N and K (constant terms are neglected).

VII. Connection to Information Distillation

It turns out that the general problem of quantizing the simplex under the *average* KL divergence loss, as defined in (2), is equivalent to recently introduced problem of *information distillation*. Information distillation has a number of applications, including in constructing polar codes [21], [22]. In this section we establish this equivalence and also demonstrate how the compander-based solutions to the KL-quantization can lead to rather simple and efficient information distillers.

A. Information Distillation

In the information distillation problem we have two random variables $A \in \mathcal{A}$ and $B \in \mathcal{B}$, where $|\mathcal{A}| = K$ (and $\mathcal{B}$ can be finite or infinite) under joint distribution $P_{A,B}$ with marginals P_A, P_B. The goal is, given some finite $M < |\mathcal{B}|$, to find an *information distiller* (which we will also refer to as a *distiller*), which is a (deterministic) function $h{:}\mathcal{B} \to [M]$, which minimizes the information loss

$$I(A;B) - I(A;h(B))$$

associated with quantizing $B \to h(B)$. The interpretation here is that B is a (high-dimensional) noisy observation of some important random variable A and we want to record observation B, but only have $\log_2 M$ bits to do so. Optimal h minimizes the additive loss entailed by this quantization of B.

To quantify the amount of loss incurred by this quantization, we use the *degrading cost* [21], [22]

$$\mathrm{DC}(K,M) = \sup_{P_{A,B}} \inf_h I(A;B) - I(A;h(B)).$$

Note that in supremizing over $P_{A,B}$ there is no restriction on $\mathcal{B}$, only on $|\mathcal{A}|$ and the size of the range of h. It has been shown in [22] that there is a $P_{A,B}$ such that

$$\inf_h I(A;B) - I(A;h(B)) = \Omega\left(M^{-2/(K-1)}\right)$$

giving a lower bound to $\mathrm{DC}(K,M)$. For an upper bound, [23] showed that if $2K < M < |\mathcal{B}|$, then

$$\mathrm{DC}(K,M) = O\left(M^{-2/(K-1)}\right).$$

Specifically, $\mathrm{DC}(K,M) \le \nu(K)M^{-2/(K-1)}$ where $\nu(K) \approx 16\pi eK^2$ for large K. While [21] focused on multiplicative loss, their work also implied an improved bound on the additive loss as well; namely, for all $K \ge 2$ and $M^{1/(K-1)} \ge 4$, we have

$$\mathrm{DC}(K,M) \le 1268(K-1)M^{-2/(K-1)}. \tag{36}$$

B. Info Distillation Upper Bounds Via Companders

Using our KL divergence quantization bounds, we will show an upper bound to $\mathrm{DC}(K,M)$ which improves on (36) for K which are not too small and for M which are not exceptionally large. First, we establish the relation between the two problems:

Proposition 8: For every $P_{A,B}$ define a random variable $\boldsymbol{X} \in \triangle_{K-1}$ by setting $X_a = P[A = a \,|\, B]$. Then, for every information distiller $h : \mathcal{B} \to [M]$ there is a vector quantizer $z : \triangle_{K-1} \to \triangle_{K-1}$ with range of cardinality M such that

$$I(A;B) - I(A;h(B)) \ge \mathbb{E}[D_{\mathrm{KL}}(\boldsymbol{X}\|z(\boldsymbol{X}))]. \tag{37}$$

Conversely, for any vector quantizer z there exists a distiller h such that

$$I(A;B) - I(A;h(B)) \le \mathbb{E}[D_{\mathrm{KL}}(\boldsymbol{X}\|z(\boldsymbol{X}))].$$

The inequalities in Proposition 8 can be replaced by equalities if the distiller h and the quantizer z avoid certain trivial inefficiencies. If they do so, there is a clean 'equivalent' quantizer z for any distiller h, and vice versa, which preserves the expected loss. This equivalence and Proposition 8 are shown in Appendix G in the supplementary material.

Thus, we can use KL quantizers to bound the degrading cost above (see Appendix G in the supplementary material for details):

$$\begin{aligned}\mathrm{DC}(K,M) &= \sup_{P_{A,B}} \inf_h I(A;B) - I(A;h(B)) \\ &= \sup_P \inf_z \mathbb{E}_{\boldsymbol{X}\sim P}[D_{\mathrm{KL}}(\boldsymbol{X}\|\boldsymbol{Z})] \\ &\le \inf_z \sup_P \mathbb{E}_{\boldsymbol{X}\sim P}[D_{\mathrm{KL}}(\boldsymbol{X}\|\boldsymbol{Z})].\end{aligned} \tag{38}$$

We then use the approximate minimax compander results to give an upper bound to (38). This yields:

Proposition 9: For any $K \ge 5$ and $M^{1/K} > \lceil 8\log(2\sqrt{K\log K}+1)\rceil$

$$\mathrm{DC}(K,M) \le \left(1 + 18\frac{\log\log K}{\log K}\right) M^{-\frac{2}{K}} \log^2 K.$$

Proof: Consider the right-hand side of (37). The compander-based quantizer from Theorem 4 gives a guaranteed bound on $D(\boldsymbol{X}\|z(\boldsymbol{X}))$ (and $M = N^K$ substituted), which also holds in expectation. ∎

Remark 12: Similarly, an upper bound on the divergence covering problem [5, Th. 2] implies

$$\mathrm{DC}(K,M) \le 800(\log K)M^{-2/(K-1)}.$$

(This appears to be the best known upper bound on DC.) The lower bound on the divergence covering, though, does not imply lower bounds on DC, since divergence covering seeks one collection of M points that are good for quantizing any P, whereas DC permits the collection to depend on P. For distortion measures that satisfy the triangle inequality, though, we have a provable relationship between the metric entropy and rate-distortion for the least-favorable prior, see [24, Sec. 27.7].

ACKNOWLEDGMENT

The authors would like to thank Anthony Philippakis for his guidance on the DNA k-mer experiments.

REFERENCES

[1] A. Adler, J. Tang, and Y. Polyanskiy, "Quantization of random distributions under KL divergence," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2021, pp. 2762–2767.
[2] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 446–472, 1948.
[3] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford, U.K.: Oxford Univ. Press, 2001.
[4] A. Ben-Yishai and O. Ordentlich, "Constructing multiclass classifiers using binary classifiers under log-loss," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2021, pp. 2435–2440.
[5] J. Tang, "Divergence covering," Ph.D. dissertation, Dept. Electr. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2022.
[6] D. Kalamkar et al., "A study of BFLOAT16 for deep learning training," 2019, *arXiv:1905.12322*.
[7] P. F. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, no. 1, pp. 44–48, Jan. 1951.
[8] P. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 139–149, Mar. 1982.
[9] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inf. Theory*, vol. 25, no. 4, pp. 373–380, Jul. 1979.
[10] M. Lewis and C. W. Brokish, "A-law and mu-law companding implementations using the TMS320c54x," Application Note SPRA163A, Texas Instrum., Dallas, TX, USA, 1997.
[11] B. Smith, "Instantaneous companding of quantized signals," *Bell Syst. Tech. J.*, vol. 36, no. 3, pp. 653–710, May 1957.
[12] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*. Berlin, Germany: Springer-Verlag, 2000.
[13] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Proc. 12th Int. Conf. Neural Inf. Process. Syst.*, Cambridge, MA, USA, 1999, pp. 617–623.
[14] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *arXiv:physics/0004057*.
[15] F. Pereira, N. Tishby, and L. Lee, "Distributional clustering of english words," in *Proc. ACL*, 1993, pp. 183–190.
[16] B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 751–763, Apr. 2013.
[17] J. Cao, Z. Wu, J. Wu, and W. Liu, "Towards information-theoretic k-means clustering for image indexing," *Signal Process.*, vol. 93, no. 7, pp. 2026–2037, 2013.
[18] I. Dhillon and S. Mallela, "A divisive information-theoretic feature clustering algorithm for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1265–1287, Mar. 2003.
[19] F. Nielsen, "Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 657–660, Jul. 2013.
[20] R. Veldhuis, "The centroid of the symmetrical Kullback–Leibler distance," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 96–99, May 2002.
[21] A. Bhatt, B. Nazer, O. Ordentlich, and Y. Polyanskiy, "Information-distilling quantizers," *IEEE Trans. Inf. Theory*, vol. 67, no. 4, pp. 2472–2487, Apr. 2021.
[22] I. Tal, "On the construction of polar codes for channels with moderate input alphabet sizes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2015, pp. 1297–1301.
[23] A. Kartowsky and I. Tal, "Greedy-merge degrading has optimal power-law," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2017, pp. 1618–1622.
[24] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2022. [Online]. Available: https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf

Aviv Adler (Member, IEEE) received the A.B. degree in mathematics from Princeton University and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, as a member of LIDS. He is joining AUTOLab with the University of California at Berkeley as a Postdoctoral Scholar. His research interests include robotics, motion planning, optimization, complexity theory, and information theory.

Jennifer Tang (Member, IEEE) received the B.S.E. degree in electrical engineering from Princeton University and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT) as a member of LIDS. She is a Postdoctoral Associate with IDSS, MIT. Her research interests include information theory, prediction and learning theory, quantization and data compression, high-dimensional statistics, data analytics, defect tolerance, and models for social dynamics and inference.

Yury Polyanskiy (Senior Member, IEEE) received the M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia, in 2005, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2010. He is a Professor of Electrical Engineering and Computer Science and a member of LIDS, IDSS, and the Center of Statistics, Massachusetts Institute of Technology. His research interests span information theory, statistical machine learning, error-correcting codes, wireless communication, and fault tolerance. He won the 2020 IEEE Information Theory Society James Massey Award, the 2013 NSF CAREER Award, and the 2011 IEEE Information Theory Society Paper Award.