Efficient Representation of Large-Alphabet Probability Distributions via Arcsinh-Compander

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*—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 set up a minimax optimality criterion and show that the compander $f(x) \propto \mathrm{ArcSinh}(\sqrt{(1/2)(K \log K)x})$ achieves near-optimal performance, attaining a KL-quantization loss of $\asymp 2^{-2b} \log^2 K$ for a K-letter alphabet and $b \to \infty$. Interestingly, a similar minimax criterion for the quadratic loss on the hypercube shows optimality of the standard uniform quantizer. This suggests that the $\mathrm{ArcSinh}$ quantizer is as fundamental for KL-distortion as the uniform quantizer for quadratic distortion.**

I. Compander Basics and Definitions

Consider the problem of finding a *quantization scheme* on $\triangle_{K-1}$ (probability simplex of alphabet size K) minimizing the KL (Kullback-Leibler) divergence between probability vectors and their representations, which corresponds to the excess code length for lossless compression and is commonly used as a way to measure the difference between probability distributions. Specifically, a probability vector $\boldsymbol{x} \in \triangle_{K-1}$ is represented by some $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ from some finite subset of $\triangle_{K-1}$ (so of course many $\boldsymbol{x}$ must map to the same $\boldsymbol{y}$); the goal is to minimize the KL divergence between the vectors $\boldsymbol{x}$ and their representations $\boldsymbol{y}(\boldsymbol{x})$.

(Full proofs and additional discussion are in [1].)

This work was supported in part by the NSF grant CCF-2131115 and sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number 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.

In this paper we consider only scalar quantization methods, which handle each element x_i of $\boldsymbol{x}$ separately since we showed in [2] that for Dirichlet priors on the simplex, scalar quantization performs nearly as well as optimal vector quantization; scalar quantization is typically simpler and faster to use, and can be parallelized easily. Our technique is based on *companders* (portmanteau of 'compressor' and 'expander').

1) Encoding: Companders require two things: a monotonically increasing 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 $\boldsymbol{x} = (x_1, \ldots, x_K) \in \triangle_{K-1}$. To quantize $x \in [0,1]$, the compander computes $f(x)$ and applies a uniform quantizer with N levels, i.e. encoding x to $n = n_N(x) \in [N]$ 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 system partitions $[0,1]$ into *bins* $I^{(n)}$:

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

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

2) Decoding: To decode $n \in [N]$, we pick some $\widehat{y}^{(n)} \in I^{(n)}$ to represent all $x \in I^{(n)}$; for a given x (at granularity N), its representation is denoted $\widehat{y}(x) = \widehat{y}^{(n_N(x))}$. This is usually the *midpoint* of the bin or, if x is drawn randomly from a prior,[1] the *centroid* (the mean within bin $I^{(n)}$). The midpoint of $I^{(n)}$ can be computed quickly using the inverse of f.

Using scalar quantization means the decoded values may not sum to 1, so we normalize. Thus, if $\boldsymbol{x}$ is the input, let

$$y_i(\boldsymbol{x}) = \frac{\widehat{y}(x_i)}{\sum_{j=1}^K \widehat{y}(x_j)}\,; \tag{1}$$

then the vector $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) = (y_1(\boldsymbol{x}), \ldots, y_K(\boldsymbol{x})) \in \triangle_{K-1}$ is the output of the compander. We refer to $\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{y}}(\boldsymbol{x}) = (\widehat{y}(x_1), \ldots, \widehat{y}(x_K))$ as the *raw* reconstruction of $\boldsymbol{x}$, and $\boldsymbol{y}$ as the *normalized* reconstruction. If the raw reconstruction uses centroid decoding, we likewise denote it using $\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{y}}(\boldsymbol{x}) = (\widetilde{y}(x_1), \ldots, \widetilde{y}(x_K))$; in general, we use $\widetilde{\cdot}$ to denote values dependent on centroid decoding.

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

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

978-1-6654-2159-1/22/$31.00 ©2022 IEEE

the prior if using centroid decoding). Another major advantage of companders is that a single f can work well over many or all choices of N, making the design more flexible.

3) KL divergence loss: The loss incurred by representing $\boldsymbol{x}$ as $\boldsymbol{y}(\boldsymbol{x})$ is the KL divergence

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

4) 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, as permuting the letter indices does not affect the KL divergence. We denote the set of such priors as $\mathcal{P}_K^{\triangle}$.

We let $\mathcal{P}$ denote the class of continuous probability distributions on $[0,1]$; these have a probability density function (PDF) p and a cumulative distribution function (CDF) F_p satisfying $p(x) = F_p'(x)$ and $F_p(x) = \int_0^x p(t)\,dt$ (since F_p is monotonic, its derivative exists almost everywhere). We denote elements of $\mathcal{P}$ by their PDFs, i.e. as $p \in \mathcal{P}$ (the PDF p does not have to be continuous, but the CDF F_p has to be absolutely continuous).

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

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

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

This is the value we want to minimize.

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

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

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

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

The raw loss is useful because it bounds the (normalized) expected loss and is decomposable into single-letter losses:

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

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

To derive our results about worst-case priors (for instance, Theorem 3), 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 1. *Though one can define raw loss and single-letter loss without centroid decoding, doing so 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. However, this cannot be exploited to make the (normalized) expected loss negative as the normalization step removes this term.*

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

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

We similarly define $\widetilde{\mathcal{L}}_K(P,f)$ and $\mathcal{L}_K(P,f)$. While the limit in (2) does not 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. Main Results

We demonstrate, theoretically and experimentally, the efficacy of companding for quantizing probability distributions with KL divergence loss. Though our theoretical results are asymptotic as $N \to \infty$ and focus on raw loss, the experimental (normalized) loss of the various companders closely tracks the (raw) loss predicted theoretically, even for quantization levels as low as $N = 256$ (8 bits per value).

1) Theory: We define a set of 'well-behaved' companders:

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

This is equivalent to $f'(x) \geqslant c\,\alpha\,x^{\alpha-1}$ for all x where f' is defined (which is almost everywhere since f is monotonic). We also define the following function on p and f:

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

$$L^\dagger(p,f) = \frac{1}{24}\int_0^1 p(x) f'(x)^{-2} x^{-1}\,dx \tag{3}$$
$$= \int_{[0,1]} \frac{1}{24} f'(x)^{-2} x^{-1}\,dp$$

Then the asymptotic loss of f against p satisfies:

Theorem 1. *For any $p \in \mathcal{P}$ and $f \in \mathcal{F}$, the bound holds:*

$$\liminf_{N\to\infty} N^2 \widetilde{L}(p,f,N) \geqslant L^\dagger(p,f)\,. \tag{4}$$

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

$$\widetilde{L}(p,f) = L^\dagger(p,f) < \infty\,. \tag{5}$$

Essentially, as long as you select a compander f from the 'well-behaved' set $\mathcal{F}^\dagger$, for large granularities N the single-letter loss will be approximated by

$$\widetilde{L}(p,f,N) \approx N^{-2} L^\dagger(p,f)\,.$$

The lower bound (4) shows that even for $f \notin \mathcal{F}^\dagger$,

$$\widetilde{L}(p,f,N) \gtrsim N^{-2} L^\dagger(p,f)$$

i.e. the quantizer cannot do better than $N^{-2}L^\dagger(p,f)$ loss (as $N \to \infty$) by choosing $f \notin \mathcal{F}^\dagger$.

Theorem 2. *The best loss against source $p \in \mathcal{P}$ is*

$$\inf_{f \in \mathcal{F}} \widetilde{L}(p,f) = \min_{f \in \mathcal{F}} L^\dagger(p,f) = \frac{1}{24}\Big(\int_0^1 (p(x)x^{-1})^{1/3} dx\Big)^3 \quad (6)$$

where the optimal compander against p *is*

$$f_p(x) = \arg\min_{f \in \mathcal{F}} L^\dagger(p,f) = \frac{\int_0^x (p(t)t^{-1})^{1/3}\, dt}{\int_0^1 (p(t)t^{-1})^{1/3}\, dt} \quad (7)$$

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

If $f_p \in \mathcal{F}^\dagger$, it achieves the value from (6) and (as the minimizer of $L^\dagger(p,f)$) it has the smallest asymptotic loss against p. If $f_p \notin \mathcal{F}^\dagger$, we use the following:

Proposition 2. *For any $f \in \mathcal{F}$ and $\delta \in (0,1]$, the functions*

$$f_{p,\delta}(x) = (1-\delta) f_p(x) + \delta x^{1/2} \quad (8)$$

satisfy $f_{p,\delta} \in \mathcal{F}^\dagger$ and

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

Thus, you can imitate f_p arbitrarily closely by mixing it with $x^{1/2}$ (or any x^α for $\alpha \in (0, 1/2]$ will also work); the mixture is by definition in $\mathcal{F}^\dagger$. This (with Theorem 1) shows there is no real advantage to using $f \notin \mathcal{F}^\dagger$, so we restrict our analysis to $f \in \mathcal{F}^\dagger$, for which (3) holds.

Since the prior P generating $\boldsymbol{x}$ is usually unknown, we give a compander which performs well against *any* prior. This is closely linked to the following probability density on $[0,1]$:

Proposition 3. *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 $\mathcal{P}_{1/K}$:*

$$p_K^*(x) = (a_K x^{1/3} + b_K x^{4/3})^{-3/2} \quad (9)$$

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

We call p_K^* the *maximin single-letter density.*
The optimal compander against p_K^* is the *minimax compander*:

$$f_K^*(x) = \frac{\mathrm{ArcSinh}(\sqrt{c_K (K \log K)\, x})}{\mathrm{ArcSinh}(\sqrt{c_K K \log K})} \quad (10)$$

Note that $f_K^* \in \mathcal{F}^\dagger$ (see Remark 2). The source p_K^* and compander f_K^* then form an 'equilibrium':

Theorem 3. *The minimax compander f_K^* and maximin single-letter density p_K^* satisfy*

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

$$= \sup_{p \in \mathcal{P}_{1/K}} \inf_{f \in \mathcal{F}^\dagger} \widetilde{L}(p,f) = \inf_{f \in \mathcal{F}^\dagger} \widetilde{L}(p_K^*, f) \quad (12)$$

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

$$\widetilde{L}(p_K^*, f_K^*) = \Theta(K^{-1} \log^2 K) \quad (13)$$

This theorem importantly implies the following:

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

$$\mathcal{L}_K(P, f_K^*) \leqslant \widetilde{\mathcal{L}}_K(P, f_K^*) = \Theta(\log^2 K)$$

There also exists $P^* \in \mathcal{P}_K^\triangle$ such that for any $P \in \mathcal{P}_K^\triangle$

$$\inf_{f \in \mathcal{F}} \widetilde{\mathcal{L}}_K(P^*, f) \geqslant \frac{K-1}{2K} \widetilde{\mathcal{L}}_K(P, f_K^*) = \Theta(\log^2 K) \quad (14)$$

The $\frac{K-1}{2K}$-factor gap in (14) is because $P^* \in \mathcal{P}_K^\triangle$ is a stronger constraint than $p_K^* \in \mathcal{P}_{1/K}$; however, whether the gap can be improved further remains open.

For any K, c_K can be approximated numerically. We can also simplify the quantizer by noting that $c_K \approx \frac{1}{2}$ for large K to get the *approximate minimax compander*:

$$f_K^{**}(x) = \frac{\mathrm{ArcSinh}(\sqrt{(1/2)(K \log K)\, x})}{\mathrm{ArcSinh}(\sqrt{(1/2) K \log K})} \quad (15)$$

This is close to optimal without needing to compute c_K:

Theorem 4. *If $c_K \in [\frac{1}{2(1+\varepsilon)}, \frac{1+\varepsilon}{2}]$, then for any $p \in \mathcal{P}$,*

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

Remark 2. *While f_K^* and f_K^{**} might appear complicated, $\mathrm{ArcSinh}(\sqrt{z}) = \log(\sqrt{z} + \sqrt{z+1})$ is fairly simple. Taking the Taylor expansion also confirms that they are in $\mathcal{F}^\dagger$.*

Note that (7) *(Theorem 2) suggests that the natural form of an optimal compander against p is a normalized incomplete integral, which is hard to use. Thus, the closed-form expressions of f_K^* and f_K^{**} is a welcome surprise.*

Using the minimax compander f_K^* or approximate minimax compander f_K^{**} on $P \in \mathcal{P}_K^\triangle$ with granularity N, we have a bound on the average KL divergence:

$$\mathbb{E}_{\boldsymbol{X} \sim P}[D_{\mathrm{KL}}(\boldsymbol{X} \| \boldsymbol{Y})] = O\left(N^{-2} \log^2 K\right) . \quad (16)$$

Remark 3. *Instead of the KL divergence loss on the simplex, we can do a similar analysis to find the minimax compander for mean-square error on the unit hypercube. The solution is given by the identity function $f(x) = x$ corresponding to the standard (non-companded) uniform quantization.*

The above are all 'average case' results, where $\boldsymbol{X}$ is drawn from a prior P (which is fixed as $N \to \infty$). In the worst-case problem, $\boldsymbol{x}$ is chosen to maximize loss and can depend on N:

Theorem 5. *The minimax compander with midpoint decoding achieves* worst-case loss *of*

$$\max_{\boldsymbol{x} \in \triangle_{K-1}} D_{\mathrm{KL}}(\boldsymbol{x} \| \boldsymbol{y}) = O\left(N^{-2} \log^2 K\right) . \quad (17)$$

Due to space constraints, we omit the proofs of Theorems 4 and 5 (see [1]). We sketch the rest in Sections IV and V.

Remark 4. *When b is the number of bits used to quantize each value in the probability vector, we get a loss on the order of $2^{-2b} \log^2 K$. If we use optimal vector quantization (for worst-case loss instead of average; explored in [3]), the loss is an order between $2^{-2b\frac{K}{K-1}}$ and $2^{-2b\frac{K}{K-1}} \log K$. Thus, our result using companders is within a factor $2^{2b/(K-1)} \log^2 K$ of the optimal loss. (The bound $2^{-2b\frac{K}{K-1}} \log K$ is not associated with an explicit quantization scheme. One is only shown to exist.)*

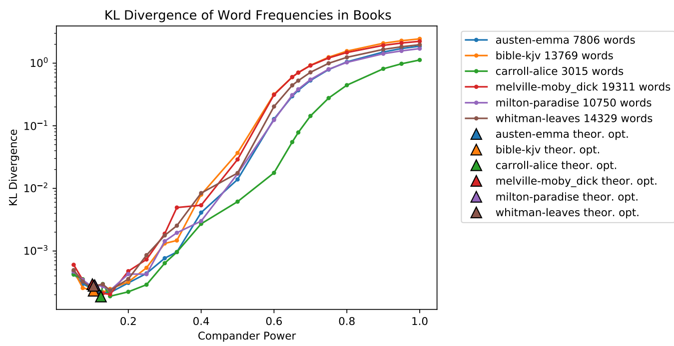

Fig. 1. Power compander $f(x) = x^s$ performance with different powers s used to quantize frequency of words in books. Number of distinct words in each book is shown in the legend. The theoretical optimal power $s = \frac{1}{\log K}$ is plotted where K is the number of distinct words.

2) Experiments: We test the performance of the approximate minimax compander (15) on three types of datasets: (i) random synthetic distributions drawn from the uniform prior over the simplex; (ii) frequency of words in books; and (iii) frequency of k-mers in DNA. We compare it against four alternatives, for granularities $N = 2^8$ and $N = 2^{16}$:

- **Truncation:** Values are quantized uniformly (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 [4].
- **Exponential Density Interval (EDI):** This is the quantization method we used in an achievability proof in [2]. It is designed for the uniform prior over the simplex.
- **Power Compander:** The compander where $f(x) = x^s$, a natural class of functions from $[0, 1]$ to $[0, 1]$. We optimize s and find that $s = \frac{1}{\log K}$ minimizes KL divergence. To see the effects of different powers s on the performance of the power compander, see Figure 1.

Our main experimental results are given in Figure 2, showing the KL divergence between the original distribution $\boldsymbol{x}$ and its quantized version $\boldsymbol{y}$ 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 power compander performs similarly to the minimax compander and is worse only by a constant.[2]

The experiments demonstrate that the approximate minimax compander achieves low loss on the entire ensemble of data (even for relatively small granularity, such as $N = 256$) and 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. Thus it can be easily added to existing systems to lower storage requirements at little or no cost to fidelity.

[2]Theorem 5 also holds for the power compander with different constants.

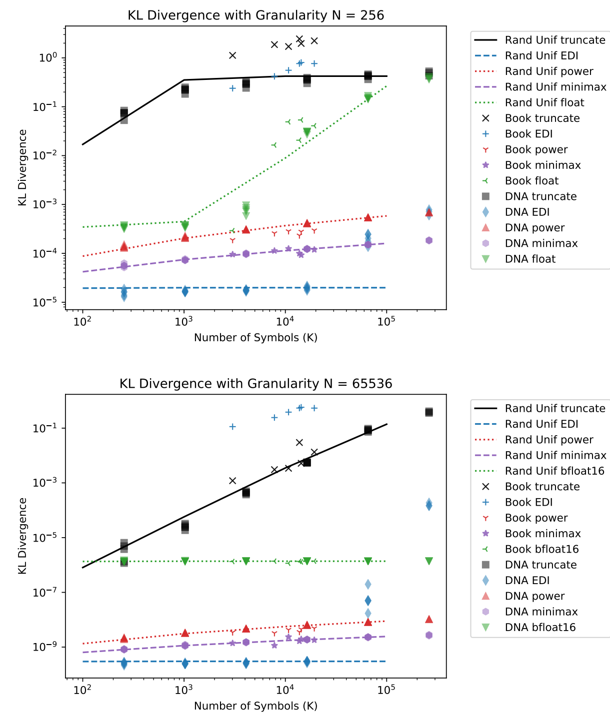

Fig. 2. Plot comparing the performance of the truncation compander, the EDI compander, floating points, the power compander, and the approximate minimax compander (15) on probability distributions of various sizes.

III. Background

Companders (also spelled "compandors") were introduced by Bennett in 1948 [5] as a way to quantize speech signals. Bennett gives a first order approximation of the mean-square error given by companders, which is similar to our (3) (though we measure expected KL divergence loss instead). Others have expanded on this line of work. In [6], the authors studied the same problem and determined the optimal compressor under mean-square error, a result which parallels our result (6). However, the results from [5], [6] are stated either as first order approximations or make simplfying assumptions. Generalizations of Bennett's formula are also studied for the case of expected rth moment loss $\mathbb{E}\| \cdot \|^r$. This is computed for length-K vectors in [7] and [8]. The typical examples of companders used in engineering are μ-law and A-law companders [9]. For the μ-law, [6] and [10] argue that for sufficiently large μ and mean-squared error, the distortion becomes independent of the signal.

Quantizing probability distributions is a common topic, though typically the loss function is a norm and not KL divergence [11]. We studied average KL divergence loss in our earlier work [2], where we focus on 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 [12] and [13]. Information k-means has been implemented for several different applications [14], [15], [16]. There are also other works that study clustering with a slightly different but related metric [17], [18], [19]; the focus of these works is to analyze data rather than reduce storage.

IV. Asymptotic Single-Letter Loss

In this section we give the outline of the proof of Theorem 1. Given density p and compander f, we construct the following: the *local loss function at granularity N*, defined as

$$g_N(x) = N^2\, \mathbb{E}_{X\sim p}[X \log(X/\widetilde{y}(X)) \mid X \in I^{(n_N(x))}]$$

and the *asymptotic local loss function*, defined as

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

The function g_N basically takes each x and returns the expected loss for $X \sim p$ which fall in the same bin as x, thus averaging the losses in each bin. The expressions (4) and (5) we need to show in Theorem 1 are thus equivalent to:

$$\liminf_{N\to\infty} \int g_N\, dp \geqslant \int g\, dp \qquad \text{for all } f \in \mathcal{F}, p \in \mathcal{P} \tag{18}$$

$$\lim_{N\to\infty} \int g_N\, dp = \int g\, dp < \infty \text{ for all } f \in \mathcal{F}^\dagger, p \in \mathcal{P} \tag{19}$$

To do this, we show the following:

Proposition 4. *For all $p \in \mathcal{P}$, $f \in \mathcal{F}$, if $X \sim p$ then*

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

The basic intuition for Proposition 4 follows from three facts: (i) as $N \to \infty$, the width of the bin containing x becomes $\approx N^{-1} f'(x)^{-1}$; (ii) as the width of an interval approaches 0, $p \in \mathcal{P}$ becomes approximately uniform; (iii) the divergence produced by the uniform distribution on an interval I of width r containing x (where all values in I are represented by the same value y) is $\approx \frac{1}{24} r^{-2} x^{-1}$ when r is very small. Combining these yields the result (see [1] for details).

Proposition 5. *For all $p \in \mathcal{P}$, $f \in \mathcal{F}^\dagger$, we have $\int g\, dp < \infty$ and there exists h s.t. $h \geqslant g_N$ for all N and $\int h\, dp < \infty$.*

Proposition 4 then implies (18) by Fatou's Lemma. If $f \in \mathcal{F}^\dagger$ then Proposition 5 (with Proposition 4) gives (19) via the Dominated Convergence Theorem, thus showing Theorem 1.

V. Minimax Compander

We show Theorem 2 and Proposition 2 together. They follow from Theorem 1 by finding $f \in \mathcal{F}$ which minimizes $L^\dagger(p, f)$, by optimizing over f'. Since $f : [0, 1] \to [0, 1]$ is monotonic, we use constraints $f'(x) \geqslant 0$ and $\int_0^1 f'(x)\, dx = 1$. Using calculus of variations, we get $f_p'(x) \propto (p(x)x^{-1})^{1/3}$ and $f(0) = 0$ and $f(1) = 1$, from which (6) and (7) follow. If $f_p \in \mathcal{F}^\dagger$, then $f_p = \arg\min_f \widetilde{L}(p, f)$, as for any other $f \in \mathcal{F}$,

$$\widetilde{L}(p, f_p) = L^\dagger(p, f_p) \leqslant L^\dagger(p, f) \leqslant \liminf_{N\to\infty} N^2 \widetilde{L}(p, f, N)$$

If $f_p \notin \mathcal{F}^\dagger$, for any $\delta > 0$ define $f_{p,\delta} \in \mathcal{F}^\dagger$ as in (8). Then

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

Taking $\delta \to 0$ thus shows that $L^\dagger(p, f_p) = \inf_{f\in\mathcal{F}^\dagger} \widetilde{L}(p, f)$. This finishes the proofs of Theorem 2 and Proposition 2.

To prove Theorem 3 and Corollary 1, we ask: what density p maximizes (6)? To do this, we instead maximize

$$\int_0^1 (p(x)x^{-1})^{1/3}\, dx \tag{20}$$

(which of course maximizes (6)) subject to $p(x) \geqslant 0$ and $\int_0^1 p(x)\, dx = 1$. Furthermore, since p must be the marginal of some symmetric prior over $\triangle_{K-1}$, we know $p \in \mathcal{P}_{1/K}$, which adds an additional constraint $\int_0^1 p(x)x\, dx = 1/K$. Solving this problem with calculus of variations yields the *maximin density* p_K^* (9) from Theorem 3. We then know from (7) that the best compander for (9) is proportional to

$$\int_0^x z^{-\frac{1}{3}} \left(a_K z^{\frac{1}{3}} + b_K z^{\frac{4}{3}}\right)^{-\frac{1}{2}} dz = \frac{2\mathrm{ArcSinh}\left(\sqrt{\frac{b_K x}{a_K}}\right)}{\sqrt{b_K}}$$

Using the constants a_K and b_K which meet the constraints, and normalizing so $f(1) = 1$, gives f_K^* (10). The function $L^\dagger(p, f)$ is linear in p and convex in f', and we can show that the pair (f_K^*, p_K^*) form a saddle point, thus proving (11)-(12) from Theorem 3. Furthermore, $f_K^* \in \mathcal{F}^\dagger$ (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 (3) with the expressions for p_K^* and f_K^* gives (13).

While p_K^* is the hardest density in $\mathcal{P}_{1/K}$ to quantize, it is unclear whether a prior P^* on $\triangle_{K-1}$ exists with marginals p_K^*. However, it is possible to construct a prior P^* whose marginals are as hard to quantize, up to a constant, as p_K^*.

Lemma 1. *For $p \in \mathcal{P}_{1/K}$, there is a joint distribution of $(X_1, \ldots, X_K)$ such that $X_i \sim p$ for all i, and $\sum_{i\in[K]} X_i \leqslant 2$.*

Lemma 1 yields a joint distribution of $K-1$ values, with marginals p_K^*, that sums to at most 2; scaling by $1/2$ and adding a (nonnegative) residual random variable gives a prior P^* on $\triangle_{K-1}$, as needed. Then:

$$\inf_{f\in\mathcal{F}} \widetilde{\mathcal{L}}_K(P^*, f) \geqslant (K-1) \inf_{f\in\mathcal{F}} \widetilde{L}(2p_K^*(2x), f)$$
$$= (K-1)\frac{1}{2} L^\dagger(p_K^*, f_K^*) \geqslant \frac{1}{2}\frac{K-1}{K} \sup_{P\in\mathcal{P}_K^\triangle} \widetilde{\mathcal{L}}_K(P, f_K^*)$$

where the last inequality holds because p_K^* is the worst-case density (under expectation constraints). To make it symmetric, we permute the letter indices randomly without affecting the raw loss, thus getting the prior P^* which shows Corollary 1.

VI. Acknowledgements

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

REFERENCES

[1] Jennifer Tang Aviv Adler and Yury Polyanskiy, "Efficient representation of large-alphabet probability distributions," *arXiv preprint arXiv:2205.03752*, 2022.

[2] Aviv Adler, Jennifer Tang, and Yury Polyanskiy, "Quantization of random distributions under KL divergence," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2762–2767.

[3] Jennifer Tang, *Divergence Covering*, Ph.D. thesis, Massachusetts Institute of Technology, 2022.

[4] Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al., "A study of bfloat16 for deep learning training," *arXiv preprint arXiv:1905.12322*, 2019.

[5] W. R. Bennett, "Spectra of quantized signals," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 446–472, 1948.

[6] P.F. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proceedings of the IRE*, vol. 39, no. 1, pp. 44–48, 1951.

[7] P. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 139–149, 1982.

[8] A. Gersho, "Asymptotically optimal block quantization," *IEEE Transactions on Information Theory*, vol. 25, no. 4, pp. 373–380, 1979.

[9] Michele Lewis and SC MTSA, "A-law and mu-law companding implementations using the tms320c54x," 1997.

[10] Bernard Smith, "Instantaneous companding of quantized signals," *The Bell System Technical Journal*, vol. 36, no. 3, pp. 653–710, 1957.

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

[12] Noam Slonim and Naftali Tishby, "Agglomerative information bottleneck," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 1999, NIPS'99, p. 617–623, MIT Press.

[13] Naftali Tishby, Fernando C Pereira, and William Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[14] Fernando Pereira, Naftali Tishby, and Lillian Lee, "Distributional clustering of English words," in *Proceedings of the ACL*, 1993, pp. 183–190.

[15] Bin Jiang, Jian Pei, Yufei Tao, and Xuemin Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 751–763, 2013.

[16] Jie Cao, Zhiang Wu, Junjie Wu, and Wenjie Liu, "Towards information-theoretic k-means clustering for image indexing," *Signal Processing*, vol. 93, no. 7, pp. 2026–2037, 2013.

[17] Inderjit Dhillon and Subramanyam Mallela, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of machine learning research*, vol. 3, pp. 1265–1287, 04 2003.

[18] Frank Nielsen, "Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 657–660, 2013.

[19] R. Veldhuis, "The centroid of the symmetrical Kullback-Leibler distance," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 96–99, 2002.