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 10^3-sized alphabets, the quality of representation improves from a loss (under KL divergence) of 0.5(1.0) bits/entry to 10^{-3}(10^{-4}) bits/entry. Compared to floating point representations, our compander method improves the loss from 10^{-1}(10^{-6}) to 10^{-3}(10^{-4}) 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) achieves near-optimal performance, attaining a KL-quantization loss of 10^{-4} effective quantization method. For example, for a K-letter alphabet and b = 16, 20, or 32, the compander computes f(x) and applies a uniform quantizer with N levels, i.e., encoding x to n = n_N(x) ∈ [N] if f(x) ∈ (n_K^{-1}, n_K), this is equivalent to n_N(x) = f(x)N.

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

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

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

1) Encoding: Companders require two things: a monotonically increasing function f : [0, 1] → [0, 1] (we denote the set of such functions as 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 x = (x_1, ..., x_K) ∈ Δ_{K-1}. To quantize x ∈ [0, 1], the compander computes f(x) and applies a uniform quantizer with N levels, i.e., encoding x to n = n_N(x) ∈ [N] if f(x) ∈ (n_K^{-1}, n_K). This is equivalent to n_N(x) = f(x)N.

2) Decoding: To decode n ∈ [N], we pick some y^{(n)} ∈ I^{(n)} to represent all x ∈ I^{(n)}; for a given x (at granularity N), its representation is denoted \( \hat{y}(x) = y^{(n_N(x))} \). 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 exactly using the inverse of f.

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

\[ y_i(x) = \frac{\hat{y}(x_i)}{\sum_{j=1}^{K} \hat{y}(x_j)} \]  

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

Thus, any x ∈ Δ_{K-1} requires K[log_2 N] bits to store; to encode and decode, only f and N need to be stored (as well as

1Priors on Δ_{K-1} induce priors over [0, 1] for each letter.)
the prior if using centroid decoding). Another major advantage of
comparers 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
$x$ as $y(x)$ is the KL divergence

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

4) Distributions from a prior: Much of our work concerns
the case where $x \in \Delta_{K-1}$ is drawn from some prior $P_x$ (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
$\mathcal{P}_K$.

We let $\mathcal{P}$ denote the class of continuous probability distribu-
tions 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 monoton-
ically, its derivative exists almost everywhere). We include 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 $\mathcal{P}_K \subseteq \mathcal{P}$ be the set of $p$ where $\mathbb{E}_{X \sim p}[X] = 1/K$. Note
that $P \in \mathcal{P}_K$ implies its marginals are in $\mathcal{P}_K$.

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

$$L_K(P, f, N) = \mathbb{E}_{X \sim P}[D_{KL}(X||y(X))] .$$

This is the value we want to minimize.

Note that $L_K(P, f, N)$ 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):

$$\hat{L}_K(P, f, N) = \mathbb{E}_{X \sim P} \left[ \sum_{i=1}^{K} X_i \log \frac{X_i}{\hat{y}(X_i)} \right]$$

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

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

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

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

$$L_K(P, f, N) \leq \hat{L}_K(P, f, N) = K \hat{L}(p, f, N)$$

To derive our results about worst-case priors (for instance, 
Theorem 3), we will also be interested in $\hat{L}(p, f, N)$ even
when $p$ is not known to be a marginal of some $P \in \mathcal{P}_K$.

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 $\mathbb{E}[\hat{y}(X)]$,
potentially even making it negative; specifically, the Taylor
expansion of $X \log \frac{X}{\hat{y}(X)}$ has $X - \hat{y}(X)$ 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:

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

We similarly define $\hat{L}_K(P, f)$ and $L_K(P, f)$. 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 ef-
ficacy 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’ comparers:

Definition 1. Let $\mathcal{F} \subset \mathcal{F}$ 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 $f'(x) \geq c \alpha x^{\alpha - 1}$ 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

$$L_1(p, f) = \frac{1}{2} \int_0^1 p(x) f'(x)^{-2} x^{-1} \, dx$$

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:

$$\liminf_{N \to \infty} N^2 \hat{L}(p, f, N) \geq L_1(p, f) .$$

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

$$\hat{L}(p, f) = L_1(p, f) < \infty .$$

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

$$\hat{L}(p, f, N) \approx N^{-2} L_1(p, f) .$$

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

$$\hat{L}(p, f, N) \geq N^{-2} L_1(p, f)$$
i.e. the quantizer cannot do better than $N^{-2} L_1(p, f)$ loss (as
$N \to \infty$) by choosing $f \notin \mathcal{F}$. 

187
Theorem 2. The best loss against source $p \in \mathcal{P}$ is
\[
\inf_{f \in \mathcal{F}} \tilde{L}(p, f) = \min_{f \in \mathcal{F}} L^1(p, f) = \frac{1}{24} \left( \int_0^1 (p(x)x^{-1})^{1/3} dx \right)^3
\]  
(6)
where the optimal compander against $p$ is
\[
f_p(x) = \arg \min_{f \in \mathcal{F}} L^1(p, f) = \frac{\int_0^1 (p(t)t^{-1})^{1/3} dt}{\int_0^1 (p(t)(t^{-1})^{1/3} dt}
\]  
(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 $L^1(p, f)$) it has the smallest asymptotic loss against $p$. If $f_p \not\in \mathcal{F}^\dagger$, we use the following:

Proposition 2. For any $f \in \mathcal{F}$ and $\delta \in (0, 1)$, the functions
\[
f_{p,\delta}(x) = (1 - \delta)f_p(x) + \delta x^{1/2}
\]  
(8)
satisfy $f_{p,\delta} \in \mathcal{F}^\dagger$ and
\[
\lim_{\delta \to 0} \tilde{L}(p, f_{p,\delta}) = \lim_{\delta \to 0} L^1(p, f_{p,\delta}) = L^1(p, f_p)
\]
Thus, you can imitate $f_p$ arbitrarily closely by mixing it with $x^{1/2}$ (or any $x^\alpha$ 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 $f \not\in \mathcal{F}^\dagger$, so we restrict our analysis to $f \in \mathcal{F}^\dagger$, for which (3) holds.

Since the prior $P$ generating $x$ 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 $c_K \in [\frac{1}{3}, 2]$ such that if $a_K = (4/\log K + 1)^{1/3}$ and $b_K = 4/a_K - a_K$, then the following density is in $\mathcal{P}_{1/K}$:
\[
p_K^* = (a_K x^{1/3} + b_K x^{4/3})^{-3/2}
\]  
(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:
\[
f_K^*(x) = \frac{\text{ArcSinh}(\sqrt{c_K} K \log K) x}{\text{ArcSinh}(\sqrt{c_K} K \log K)}
\]  
(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
\[
\sup_{p \in \mathcal{P}_{1/K}} \tilde{L}(p, f_K^*) = \inf_{f \in \mathcal{F}^\dagger} \sup_{p \in \mathcal{P}_{1/K}} \tilde{L}(p, f) = \sup_{p \in \mathcal{P}_{1/K}} \inf_{f \in \mathcal{F}^\dagger} \tilde{L}(p, f)
\]  
(11)
which is equal to $\tilde{L}(p_K^*, f_K^*)$ and satisfies
\[
\tilde{L}(p_K^*, f_K^*) = \Theta(K^{-1} \log^2 K)
\]  
(13)
This theorem importantly implies the following:

Corollary 1. For any prior $P \in \mathcal{P}_K^\Delta$,
\[
\mathcal{L}_K(P, f_k^*) = \mathcal{L}_K(P, f_k^*) = \Theta(\log^2 K)
\]
There also exists $P^* \in \mathcal{P}_K^\Delta$ such that for any $P \in \mathcal{P}_K^\Delta$
\[
\inf_{f \in \mathcal{F}} \mathcal{L}_K(P^*, f) \geq \frac{K - 1}{2K} \mathcal{L}_K(P, f_k^*) = \Theta(\log^2 K)
\]  
(14)
The $\frac{K - 1}{2K}$-factor gap in (14) is because $P^* \in \mathcal{P}_K^\Delta$ 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:
\[
f_K^{**}(x) = \frac{\text{ArcSinh}(\sqrt{(1/2)(K \log K)} x)}{\text{ArcSinh}(\sqrt{(1/2) K \log K})}
\]  
(15)
This is close to optimal without needing to compute $c_K$.

Theorem 4. If $c_k \in [\frac{1}{3}, \frac{1}{2}]$, then for any $p \in \mathcal{P}$,
\[
\tilde{L}(p, f_k^*) \leq (1 + \varepsilon) \tilde{L}(p, f_k^*)
\]
Remark 2. While $f_k^*$ and $f_k^{**}$ might appear complicated, $\text{ArcSinh}(\sqrt{z}) = \log(\sqrt{z} + \sqrt{z + 1})$ 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 $P \in \mathcal{P}_{1/K}$ with granularity $N$, we have a bound on the average KL divergence:
\[
\mathbb{E}_x \mathcal{L}_KL(X|Y) = O(N^{-2} \log^2 K)
\]  
(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 $X$ is drawn from a prior $P$ (which is fixed as $N \to \infty$). In the worst-case problem, $x$ is chosen to maximize loss and can depend on $N$.

Theorem 5. The minimax compander with midpoint decoding achieves worst-case loss of
\[
\max_{x \in \Delta^K_{K-1}} D_{KL}(x|y) = O(N^{-2} \log^2 K)
\]  
(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 $2^{-2b \log^{b/2} K}$ and $2^{-2b \log^{b/2} \log K}$. Thus, our result using comparandizers is within a factor $2^{2b/(K-1)} \log^2 K$ of the optimal loss. (The bound $2^{-2b \log^{b/2} K}$ is not associated with an explicit quantization scheme. One is only shown to exist.)
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 Compa...r performance with 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 \( x \) and its quantized version \( y \) 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.

### III. Background

Compressors (also spelled “componders”) 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 compressors, 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 simplifying assumptions. Generalizations of Bennett’s formula are also studied for the case of expected \( r \)th moment loss \( E [\cdot | \cdot ]^r \). This is computed for length-\( K \) vectors in [7] and [8]. The typical examples of compressors used in engineering are \( \mu \)-law and \( A \)-law compressors [9]. For the \( \mu \)-law, [6] and [10] argue that for sufficiently large \( \mu \) 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 comander \(f\), we construct the following: the local loss function at granularity \(N\), defined as

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

and the asymptotic local loss function, defined as

\[
g(x) = \frac{1}{2^N} f'(x)^{-2} x^{-1}.
\]

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:

\[
\lim_{N \to \infty} \frac{1}{N} \int g_N dp \geq \int g dp \quad \text{for all } f \in \mathcal{F}, p \in \mathcal{P}
\]

\[
\lim_{N \to \infty} \frac{1}{N} \int g_N dp = \int g dp < \infty \quad \text{for all } f \in \mathcal{F}, p \in \mathcal{P}
\]

To do this, we show the following:

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

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

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 \(\approx \frac{1}{2^N} r^{-2} x^{-1}\) when \(r\) is very small. Combining these yields the result (see [1] for details).

Proposition 5. For all \(p \in \mathcal{P}, f \in \mathcal{F}\), we have \(\int g dp < \infty\) and there exists \(h\) s.t. \(h \geq g_N\) for all \(N\) and \(\int h dp < \infty\).

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

V. MINIMAX COMPANADER

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

\[
\tilde{L}(p, f_p) = L^1(p, f_p) \leq L^1(p, f) \leq \lim_{N \to \infty} \inf \{N^2 \tilde{L}(p, f, N)\}
\]

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

\[
\tilde{L}(p, f_{p, \delta}) = L^1(p, f_{p, \delta}) \leq L^1(p, f_p)(1 - \delta)^{-2}.
\]

Taking \(\delta \to 0\) thus shows that \(L^1(p, f_p) = \inf_{f \in \mathcal{F}} \tilde{L}(p, f)\). 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

\[
\int_0^1 (p(x)x^{-1})^{1/3} \, dx
\]

(20)

(which of course maximizes (6)) subject to \(p(x) \geq 0\) and \(\int_0^1 p(x) dx = 1\). Furthermore, since \(p\) must be the marginal of some symmetric prior over \(\Delta_{K-1}\), we know \(p \in \mathcal{P}_{1/K}\), which adds an additional constraint \(\int_0^1 p(x) dx = 1/K\). 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 comander for (9) is proportional to

\[
\int_0^1 z^{-1/3} (a_K z^{1/3} + b_K z^{2/3})^{-1/2} \, dz = \frac{2 \text{ArcSin}(\frac{\sqrt{b_K}}{a_K})}{\sqrt{b_K}}
\]

Using the constants \(a_K\) and \(b_K\) which meet the constraints, and normalizing so \(f(1) = 1\), gives \(f^*_K\) (10). The function \(L^1(p, f)\) 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, \(f^*_K \in \mathcal{F}\) (it behaves as a multiple of \(x^{1/2}\) near 0), so \(\tilde{L}(p, f^*_K) = L^1(p, f^*_K)\) 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 \(\Delta_{K-1}\) 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 \(\sum_{i=1}^{K} X_i \leq 2\).

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 \(\Delta_{K-1}\), as needed. Then:

\[
\inf_{f \in \mathcal{F}} \tilde{L}_K(P^*, f) \geq (K - 1) \inf_{f \in \mathcal{F}} \tilde{L}(2p^*_K(2x), f) = (K - 1) \frac{1}{2} L^1(p^*_K, f^*_K) \geq \frac{1}{2} K - \frac{1}{2} \sup_{P \in \mathcal{P}_{1/K}} \tilde{L}_K(P, f^*_K)
\]

where the last inequality holds because \(p^*_K\) is the worst-case density (under expectation constraints). To make it symmetric, we perturb 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


