Lossy Compression of Permutations

Da Wang
EECS Dept., MIT
Cambridge, MA, USA
Email: dawang@mit.edu

Arya Mazumdar
ECE Dept., Univ. of Minnesota
Twin Cities, MN, USA
Email: arya@umn.edu

Gregory W. Wornell
EECS Dept., MIT
Cambridge, MA, USA
Email: gww@mit.edu

***Abstract*—We investigate the lossy compression of permutations by analyzing the trade-off between the size of a source code and the distortion with respect to Kendall tau distance, Spearman's footrule, Chebyshev distance and ℓ_1 distance of inversion vectors. We show that given two permutations, Kendall tau distance upper bounds the ℓ_1 distance of inversion vectors and a scaled version of Kendall tau distance lower bounds the ℓ_1 distance of inversion vectors with high probability, which indicates an equivalence of the source code designs under these two distortion measures. Similar equivalence is established for all the above distortion measures, every one of which has different operational significance and applications in ranking and sorting. These findings show that an optimal coding scheme for one distortion measure is effectively optimal for other distortion measures above.**

I. Introduction

In this paper we consider the lossy compression (source coding) of permutations, which is motivated by the problems of storing ranking data, and lower bounding the complexity of approximate sorting.

In a variety of applications such as college admission and recommendation systems (e.g., Yelp.com and IMDb.com), ranking, or the relative ordering of data, is the key object of interest. Noting that a ranking of n items can be represented as a permutation on n elements, 1 to n, storing a ranking is equivalent to storing a permutation. In general, to store a permutation of n elements, we need $\log_2(n!) \approx n\log_2 n - n\log_2 e$ bits. However, if we can tolerate certain error (instead of the top restaurant, say the query returns one of the top five), then how many bits are necessary for storage?

In addition to application on compression, source coding of the permutation space is also related to the analysis of comparison-based sorting algorithms. Given a group of elements of distinct values, comparison-based sorting can be viewed as the process of finding a true permutation by pairwise comparisons, and since each comparison in sorting provides at most 1 bit of information, the log-size of the permutation set $\mathcal{S}_n$ provides a lower bound to the required number of comparisons, i.e., $\log n! = n\log n - O(n)$. Similarly, the lossy source coding of permutations provides a lower bound to the problem of comparison-based approximate sorting, which can be seen as searching a true permutation subject to certain distortion. Again, the log-size of the code indicates the amount of information (in terms of bit) needed to specify the true permutation subject to certain distortion, which in turn provides a lower bound on the number of pairwise comparisons needed.

This work was supported, in part, by AFOSR under Grant No. FA9550-11-1-0183, and by NSF under Grant No. CCF-1017772. Arya Mazumdar's research was also supported in part by a startup grant from University of Minnesota.

The problem of approximate sorting has been investigated in [1], where results for the moderate distortion regime are derived with respect to the Spearman's footrule metric [2] (see below for definition).

On the other hand, every comparison-based sorting algorithm corresponds to a compression scheme of the permutation space, as we can treat the outcome of each comparison as 1 bit. This string of bits is a (lossy) representation of the permutation that is being (approximately) sorted. However, reconstructing the permutation from the compressed representation may not be straightforward.

In our earlier work [3], a rate-distortion theory for permutation space is developed, with the *worst-case distortion* as the parameter. The rate-distortion functions and source code designs for two different distortion measures, Kendall tau distance and the ℓ_1 distance of the inversion vectors, are derived. In Section III of this paper we show that under *average-case distortion*, the rate-distortion problem under Kendall tau distance and ℓ_1 distance of the inversion vectors are equivalent and hence the code design could be used interchangeably, leading to simpler coding schemes for the Kendall tau distance case (than developed in [3]), as discussed in Section IV.

Moreover, the rate-distortion problem under Chebyshev distance is also considered and its equivalence to the cases above is established. Operational meaning and importance of all these distance measures is discussed in Section II. While these distance measures usually have different intended applications, our findings show that an optimal coding scheme for one distortion measure is effectively optimal for other distortion measures.

II. Problem formulation

In this section we discuss aspects of the problem formulation. We provide a mathematical formulation of the rate-distortion problem on a permutation space in Section II-B, introduce the distortions of interest in Section II-C, and discuss their operational meaning in Section II-D.

A. Notation

Let $\mathcal{S}_n$ denote the symmetric group of n elements. We write the elements of $\mathcal{S}_n$ as arrays of natural numbers with values ranging from $1,\ldots,n$ and every value occurring only once in the array. For example, $\sigma = [3,4,1,2,5] \in \mathcal{S}_5$. This is also known as the *vector notation* for permutations. For a permutation σ, we denote its permutation inverse by σ^{-1}, where $\sigma^{-1}(x) = i$ when $\sigma(i) = x$. and $\sigma(i)$ is the i-th element in array σ. For example, the permutation inverse of $\sigma = [2,5,4,3,1]$ is $\sigma^{-1} = [5,1,4,3,2]$. Given a metric $d : \mathcal{S}_n \times \mathcal{S}_n \to \mathbb{R}^+ \cup \{0\}$, we define a *permutation space* $\mathcal{X}(\mathcal{S}_n, d)$.

Throughout the paper, we denote the set $\{1,\ldots,n\}$ as $[n]$, and let $[a:b] \triangleq \{a, a+1, \ldots, b-1, b\}$ for any two integers a and b.

B. Rate-distortion problem

Given a permutation space, we define the following rate-distortion problem.

Definition 1 (Codebook for permutations under average–case distortion)**.** *An (n, D) source code $\mathcal{C}_n \subseteq \mathcal{S}_n$ for $\mathcal{X}(\mathcal{S}_n, d)$ is a set of $M_n = |\mathcal{C}_n|$ permutations such that for a σ that is drawn uniformly at random from $\mathcal{S}_n$, there exists a permutation $\pi(\sigma) \in \mathcal{C}_n$ that*

$$\mathbb{E}\left[d(\pi(\sigma), \sigma)\right] \leq D,$$

where expected value is taken with respect to the uniform distribution over $\mathcal{S}_n$. The mapping $\pi : \mathcal{S}_n \to \mathcal{C}_n$ can be assumed to satisfy

$$\pi(\sigma) = \arg\min_{\sigma' \in \mathcal{C}_n} d(\sigma', \sigma)$$

for any $\sigma \in \mathcal{S}_n$. Given a sequence of distortions $\{D_n, n \in \mathbb{Z}^+\}$, let $A(n, D_n)$ be the minimum size of an (n, D_n) source codes in $\mathcal{X}(\mathcal{S}_n, d)$, and we define the minimal rate for distortions D_n as

$$R(D_n) \triangleq \frac{\log A(n, D_n)}{\log n!}.$$

As to the classical rate-distortion setup, we are interested in deriving the trade-off between distortion level D_n and the rate $R(D_n)$ as $n \to \infty$. In this work we show that for the distortions $d(\cdot,\cdot)$ and the sequences of distortions $\{D_n, n \in \mathbb{Z}^+\}$ of interest, $\lim_{n\to\infty} R(D_n)$ exists.

Instead of requiring $\mathbb{E}\left[d(\pi, \sigma)\right] \leq D_n$ in the above definition, we may require

$$\lim_{n\to\infty} \mathbb{P}\left[d(\pi, \sigma) > D_n\right] = 0. \tag{1}$$

This stronger requirement does not change the asymptotic rate-distortion trade-off.

C. Distortion measures

There are several distortion measures possible in permutations. Some of the most natural measures include Kendall-tau distance and ℓ_p distances, where $p \in \{1, 2, \ldots, \infty\}$. In this section we introduce the distortion measures of interest, Spearman's footrule (ℓ_1 distance between two permutation vectors), Chebyshev distance (ℓ_∞ distance between two permutation vectors), the ℓ_1 distance of inversion vectors and Kendall tau distance.

Definition 2 (Spearman's footrule)**.** *Given two permutations σ_1 and σ_2, the Spearman's footrule between σ_1 and σ_2 is*

$$d_{\ell_1}(\sigma_1, \sigma_2) \triangleq \|\sigma_1 - \sigma_2\|_1 = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|$$

Definition 3 (Chebyshev distance)**.** *Given two permutations σ_1 and σ_2, the Chebyshev distance between σ_1 and σ_2 is*

$$d_{\ell_\infty}(\sigma_1, \sigma_2) \triangleq \|\sigma_1 - \sigma_2\|_\infty = \max_{1\leq i\leq n} |\sigma_1(i) - \sigma_2(i)|$$

Inversion vector is a representation of a permutation, building upon the notion of *inversions*.

Definition 4 (inversion, inversion vector)**.** *An* inversion *in a permutation $\sigma \in \mathcal{S}_n$ is a pair $(\sigma(i), \sigma(j))$ such that $i < j$ and $\sigma(i) > \sigma(j)$. We use $I_n(\sigma)$ to denote the total number of inversions in $\sigma \in \mathcal{S}_n$, and*

$$K_n(k) \triangleq |\{\sigma \in \mathcal{S}_n : I_n(\sigma) = k\}| \tag{2}$$

to denote the number of permutations with k inversions.

A permutation $\sigma \in \mathcal{S}_n$ is associated with an inversion vector $\mathbf{x}_\sigma \in \mathcal{G}_n \triangleq [0:1] \times [0:2] \times \cdots \times [0:n-1]$, *where for $i = 1, \ldots, n-1$,*

$$\mathbf{x}_\sigma(i) = \left|\left\{j \in [n] : j < i+1, \sigma^{-1}(j) > \sigma^{-1}(i+1)\right\}\right|.$$

In words, $\mathbf{x}_\sigma(i)$ is the number of inversions in σ in which $i+1$ is the first element.

It is well known that mapping from $\mathcal{S}_n$ to $\mathcal{G}_n$ is one-to-one and straightforward [4].

Definition 5 (ℓ_1 distance of inversion vectors)**.** *Given two permutations σ_1 and σ_2, we define the ℓ_1 distance of two inversion vectors as*

$$d_{\mathbf{x},\ell_1}(\sigma_1, \sigma_2) \triangleq \sum_{i=1}^{n-1} |\mathbf{x}_{\sigma_1}(i) - \mathbf{x}_{\sigma_2}(i)|. \tag{3}$$

Example 1 (ℓ_1 distance of inversion vectors)**.** *The inversion vector for permutation $\sigma_1 = [1,5,4,2,3]$ is $\mathbf{x}_{\sigma_1} = [0,0,2,3]$, as the inversions are $(4,2),(4,3),(5,4),(5,2),(5,3)$. The inversion vector for permutation $\sigma_2 = [3,4,5,1,2]$ is $\mathbf{x}_{\sigma_2} = [0,2,2,2]$, as the inversions are $(3,1),(3,2),(4,1),(4,2),(5,1),(5,2)$. Therefore,*

$$d_{\mathbf{x},\ell_1}(\sigma_1, \sigma_2) = d_{\ell_1}([0,0,2,3],[0,2,2,2]) = 3$$

Now we introduce another distortion measure of interest, Kendall tau distance.

Definition 6 (Kendall tau distance)**.** *The* Kendall tau distance *$d_\tau(\sigma_1, \sigma_2)$ from one permutation σ_1 to another permutation σ_2 is defined as the minimum number of transpositions of pairwise adjacent elements required to change σ_1 into σ_2.*

The Kendall tau distance is upper bounded by $\binom{n}{2}$.

Example 2 (Kendall tau distance)**.** *The Kendall tau distance for $\sigma_1 = [1,5,4,2,3]$ and $\sigma_2 = [3,4,5,1,2]$ is $d_\tau(\sigma_1, \sigma_2) = 7$, as one needs at least 7 transpositions of pairwise adjacent elements to change σ_1 to σ_2. For example,*

$$\begin{aligned}\sigma_1 &= [1,5,4,2,3]\\ &\to [1,5,4,3,2] \to [1,5,3,4,2] \to [1,3,5,4,2]\\ &\to [3,1,5,4,2] \to [3,5,1,4,2] \to [3,5,4,1,2]\\ &\to [3,4,5,1,2] = \sigma_2\end{aligned}$$

Remark 1 (Bubble-sort)**.** *Let $e \triangleq [1,2,\ldots,n]$ be the identity permutation, and given a input sequence of σ, $d_\tau(\sigma, e)$ is the number of swaps needed in a bubble-sort algorithm [4].*

D. Operational meaning of the distortion measures

Spearman's footrule is a very well-known measure of disarray (see, [2]), where the sum of the deviations at different positions is measured. The *Chebyshev distance* measures maximum of the deviations at different positions. Therefore, Spearman's footrule (ℓ_1 distance) measures the total deviation, while Chebyshev distance (ℓ_∞ distance) measures the worst-case deviation. In approximate sorting, we may care about one measure over the other, or both.

In addition, *inversion vector* provides a measure of disorder in a sequence. Given a sequence $v_1, v_2, \ldots, v_n$ and permutation π such that $v_{\pi(1)} < v_{\pi(2)} < \ldots < v_{\pi(n)}$, then $\pi(n+1-k)$ is the index of the k-th largest element, and $\mathbf{x}_\pi(n-k)$ indicates the number of elements that are smaller than the k-th largest but have indices larger than that of the k-th largest element. In particular, the position of the largest element is $n - \mathbf{x}_\pi(n-1)$.

Apart from the natural sorting algorithms, *Kendall tau distance* has recently attracted substantial interest in the area of error-correcting codes for Flash memory [5], [6], Kendall tau distance is also a global measure of disarray that is very popular in statistics. It is closely related to the Spearman's footrule, as we will see next.

III. Relationships between distortion measures

In this section we show all four distortion measures defined in Section II-C are closely related to each other.

A. Spearman's footrule and Kendall tau distance

Theorem 1 (Relationship of Kendall tau distance and ℓ_1 distance of permutation vectors [2])**.** *Let σ_1 and σ_2 be any permutations in $\mathcal{S}_n$, then*

$$d_{\ell_1}(\sigma_1, \sigma_2)/2 \leq d_\tau(\sigma_1^{-1}, \sigma_2^{-1}) \leq d_{\ell_1}(\sigma_1, \sigma_2). \quad (4)$$

B. ℓ_1 distance of inverse vectors and Kendall tau distance

We show that the ℓ_1 distance of inversion vectors and the Kendall tau distance are closely related in Theorem 2, and Theorem 3, which helps to establish the equivalence of the rate-distortion problem later.

The Kendall tau distance between two permutation vectors provides upper and lower bounds to the ℓ_1 distance between the inversion vectors of the corresponding permutations, as indicated by the following theorem.

Theorem 2. *Let σ_1 and σ_2 be any permutations in $\mathcal{S}_n$, then for $n \geq 2$,*

$$\frac{1}{n-1} d_\tau(\sigma_1, \sigma_2) \leq d_{\mathbf{x},\ell_1}(\mathbf{x}_{\sigma_1}, \mathbf{x}_{\sigma_2}) \leq d_\tau(\sigma_1, \sigma_2) \quad (5)$$

The proof of this theorem is relatively straight-forward and hence omitted due to space constraint.

Remark 2. *The lower bound in Theorem 2 is tight as there exists permutations σ_1 and σ_2 that satisfy the equality. For example, when $n = 2m$, let $\sigma_1 = [1, 3, 5, \ldots, 2m-3, 2m-1, 2m, 2m-2, \ldots, 6, 4, 2]$, $\sigma_2 = [2, 4, 6, \ldots, 2m-2, 2m, 2m-1, 2m-3, \ldots, 5, 3, 1]$, then $d_\tau(\sigma_1, \sigma_2) = n(n-1)/2$ and $d_{\mathbf{x},\ell_1}(\sigma_1, \sigma_2) = n/2$. For another instance, let $\sigma_1 = [1, 2, \ldots, n-2, n-1, n]$, $\sigma_2 = [2, 3, \ldots, n-1, n, 1]$ then $d_\tau(\sigma_1, \sigma_2) = n-1$ and $d_{\mathbf{x},\ell_1}(\sigma_1, \sigma_2) = 1$.*

Theorem 2 shows that in general $d_\tau(\sigma_1, \sigma_2)$ is not a good approximation to $d_{\mathbf{x},\ell_1}(\sigma_1, \sigma_2)$ due to the $1/(n-1)$ factor. However, Theorem 3 shows that it provides a tight lower bound with high probability.

Theorem 3. *For any $\pi \in \mathcal{S}_n$, let σ be a permutation chosen uniformly from $\mathcal{S}_n$, then*

$$\mathbb{P}[c_1 \cdot d_\tau(\pi, \sigma) \leq d_{\mathbf{x},\ell_1}(\pi, \sigma)] = 1 - O(1/n) \quad (6)$$

for any positive constant $c_1 < 1/2$.

Proof: See Section V-A. ∎

C. Spearman's footrule and Chebyshev distance

Let σ_1 and σ_2 be any permutations in $\mathcal{S}_n$, then

$$d_{\ell_1}(\sigma_1, \sigma_2) \leq n \cdot d_{\ell_\infty}(\sigma_1, \sigma_2), \quad (7)$$

and additionally, the scaled Chebyshev distance lower bounds the Spearman's footrule with high probability.

Theorem 4. *For any $\pi \in \mathcal{S}_n$, let σ be a permutation chosen uniformly from $\mathcal{S}_n$, then*

$$\mathbb{P}[c_2 \cdot n \cdot d_{\ell_\infty}(\pi, \sigma) \leq d_{\ell_1}(\pi, \sigma)] = 1 - O(1/n) \quad (8)$$

for any positive constant $c_2 < 1/3$.

Proof: See Section V-B. ∎

IV. Rate distortion functions

In this section we build upon the results in Section III and prove the equivalence of lossy source codes under different distortion measures, which lead to the rate distortion functions in Theorem 6.

Theorem 5 (Equivalence of lossy source codes under different distortion measures)**.** *Below, any of the source codes of the left hand side, implies a source code of the right.*

1) *(n, D_n) source code for $\mathcal{X}(\mathcal{S}_n, d_\tau) \Rightarrow (n, D_n)$ source code for $\mathcal{X}(\mathcal{S}_n, d_{\mathbf{x},\ell_1})$,*
2) *(n, D_n) source code for $\mathcal{X}(\mathcal{S}_n, d_{\mathbf{x},\ell_1}) \Rightarrow (n, D_n/c_1 + O(n))$ source code for $\mathcal{X}(\mathcal{S}_n, d_\tau)$ for any $c_1 < 1/2$,*
3) *(n, D_n) source code for $\mathcal{X}(\mathcal{S}_n, d_{\ell_1}) \Rightarrow (n, D_n)$ source code for $\mathcal{X}(\mathcal{S}_n, d_\tau)$.*
4) *(n, D_n) source code for $\mathcal{X}(\mathcal{S}_n, d_\tau) \Rightarrow (n, 2D_n)$ source code for $\mathcal{X}(\mathcal{S}_n, d_{\ell_1})$,*
5) *$(n, D_n/n)$ source code for $\mathcal{X}(\mathcal{S}_n, d_{\ell_\infty}) \Rightarrow (n, D_n)$ source code for $\mathcal{X}(\mathcal{S}_n, d_{\ell_1})$,*
6) *(n, D_n) source code for $\mathcal{X}(\mathcal{S}_n, d_{\ell_1}) \Rightarrow (n, D_n/(nc_2) + O(1))$ source code for $\mathcal{X}(\mathcal{S}_n, d_{\ell_\infty})$ for any $c_1 < 1/3$.*

Proof: Statement 1 follow directly from (5). For statement 2, let

$$\mathcal{A}_n(\pi) \triangleq \{\sigma : c_1 \cdot d_\tau(\sigma, \pi) \leq d_{\mathbf{x},\ell_1}(\sigma, \pi)\}$$

then Theorem 3 indicates that $|\mathcal{A}_n(\pi)| = (1 - O(1/n))n!$. Let $\mathcal{C}'_n$ be the (n, D_n) source code for $\mathcal{X}(\mathcal{S}_n, d_{\mathbf{x},\ell_1})$ and σ be a permutation chosen uniformly from $\mathcal{S}_n$, then let

π_σ be the codeword for σ in $\mathcal{C}'_n$,

$$\begin{aligned}
&\mathbb{E}\left[d_\tau\left(\pi_\sigma,\sigma\right)\right]\\
&=\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n} d_\tau\left(\sigma,\pi_\sigma\right)\\
&=\frac{1}{n!}\left[\sum_{\sigma\in\mathcal{A}_n(\pi_\sigma)} d_\tau\left(\sigma,\pi_\sigma\right)+\sum_{\sigma\in\mathcal{S}_n\setminus\mathcal{A}_n(\pi_\sigma)} d_\tau\left(\sigma,\pi_\sigma\right)\right]\\
&\le\frac{1}{n!}\left[\sum_{\sigma\in\mathcal{A}_n(\pi_\sigma)} d_{\mathbf{x},\ell_1}\left(\sigma,\pi_\sigma\right)/c_1+\sum_{\sigma\in\mathcal{S}_n\setminus\mathcal{A}_n(\pi_\sigma)} n^2/2\right]\\
&\le D_n/c_1+O\left(1/n\right)n^2=D_n/c_1+O\left(n\right).
\end{aligned}$$

Statement 3 and 4 follow directly from Theorem 1. Statement 5 follows from (7). For statement 6, similar to the proof for statement 2, define

$$\mathcal{B}_n(\pi)\triangleq\left\{\sigma: c_2\cdot n\cdot d_{\ell_\infty}\left(\sigma,\pi\right)\le d_{\ell_1}\left(\sigma,\pi\right)\right\}$$

then by Theorem 4,

$$\begin{aligned}
&\mathbb{E}\left[d_{\ell_\infty}\left(\pi_\sigma,\sigma\right)\right]=\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n} d_{\ell_\infty}\left(\sigma,\pi_\sigma\right)\\
&=\frac{1}{n!}\left[\sum_{\sigma\in\mathcal{B}_n(\pi_\sigma)} d_{\ell_\infty}\left(\sigma,\pi_\sigma\right)+\sum_{\sigma\in\mathcal{S}_n\setminus\mathcal{B}_n(\pi_\sigma)} d_{\ell_\infty}\left(\sigma,\pi_\sigma\right)\right]\\
&\le\frac{1}{n!}\left[\sum_{\sigma\in\mathcal{B}_n(\pi_\sigma)} d_{\ell_1}\left(\sigma,\pi_\sigma\right)+\sum_{\sigma\in\mathcal{S}_n\setminus\mathcal{B}_n(\pi_\sigma)} n\right]\\
&\le D_n/(nc_2)+O\left(1/n\right)n=D_n/(nc_2)+O\left(1\right).
\end{aligned}$$

■

We obtain Theorem 6 as a direct consequence of Theorem 5.

Theorem 6 (Rate distortion functions for distortion measures)**.** *For permutation spaces $\mathcal{X}\left(\mathcal{S}_n,d_{\mathbf{x},\ell_1}\right)$, $\mathcal{X}\left(\mathcal{S}_n,d_\tau\right)$, and $\mathcal{X}\left(\mathcal{S}_n,d_{\ell_1}\right)$,*

$$R(D_n)=\begin{cases}1 & \text{if } D_n=O\left(n\right)\\ 1-\delta & \text{if } D_n=\Theta\left(n^{1+\delta}\right),\end{cases}\quad 0<\delta\le 1.$$

For the permutation space $\mathcal{X}\left(\mathcal{S}_n,d_{\ell_\infty}\right)$,

$$R(D_n)=\begin{cases}1 & \text{if } D_n=O\left(1\right)\\ 1-\delta & \text{if } D_n=\Theta\left(n^{\delta}\right),\end{cases}\quad 0<\delta\le 1.$$

Proof: (9) follows from [3, Theorem 5 and 6], which states the rate distortion functions for both permutation spaces $\mathcal{X}\left(\mathcal{S}_n,d_\tau\right)$ and $\mathcal{X}\left(\mathcal{S}_n,d_{\mathbf{x},\ell_1}\right)$ satisfy

$$R(D_n)=\begin{cases}1 & \text{if } D_n=O\left(n\right)\\ 1-\delta & \text{if } D_n=\Theta\left(n^{1+\delta}\right),\end{cases}\quad 0<\delta\le 1.$$

Then the rest follows from Theorem 5. ■

Theorem 5 indicates that for all the distortion measures in this paper, the lossy compression scheme for one measure preserves distortion under other measures, and hence all compression schemes can be used interchangeably under average-case distortion, after transforming the permutation representation and scaling the distortion correspondingly.

For the vector representation of permutation, compression based on Kendall tau distance is essentially optimal, which can be achieved by partitioning each permutation vector into subsequences with proper sizes and sorting them accordingly [3]. For the inversion vector representation of permutation, a simple component-wise scalar quantization achieves the optimal rate distortion trade-off, as shown in [3]. In particular, given $D=cn^{1+\delta}, 0<\delta<1$, for the $(k-1)$-th component of the inversion vector ($k=2,\cdots,n$), we quantize k points in $[0:k-1]$ uniformly with $m_k=\lceil kn/(2D)\rceil$ points, resulting component-wise average distortion $D_k=D/n$ and overall average distortion $=\sum_{k=2}^n D_k\le D$, and log of codebook size $\log M_n=\sum_{k=2}^n\log m_k=\sum_{k=2}^n\log\lceil kn/(2D)\rceil=(1-\delta)n\log n-O\left(n\right)$.

Remark 3. *This scheme is slightly different from the one in [3] as it is designed for average distortion, while the latter for worst-case distortion.*

Remark 4. *While the compression algorithm in $\mathcal{X}\left(\mathcal{S}_n,d_{\mathbf{x},\ell_1}\right)$ is conceptually simple and has time complexity $\Theta\left(n\right)$, it takes $\Theta\left(n\log n\right)$ runtime to convert a permutation from its vector representation to its inversion vector representation [4, Exercise 6 in Section 5.1.1]. Therefore, the cost of representation transformation of permutations should be taken into account when selecting the compression scheme.*

Example 3 (Rate distortion trade-off at $n=1024$)**.** *We evaluate the rate distortion functions in Theorem 6 by simulation when $n=1024$. We randomly draw a set of permutations uniformly from $\mathcal{S}_n$ and define the normalized distortion as*

$$\hat{\delta}\triangleq\log_2(c\bar{D})/\log_2 n-1, \tag{9}$$

where $\bar{D}$ is the average distortion of all generated random permutations, and c is a constant that depends on the distortion measure of interest. We also define the normalized rate as

$$\hat{R}\triangleq|\mathcal{C}_n|/\log_2 n!. \tag{10}$$

Then we plot the rate distortion trade-off for both $\mathcal{X}\left(\mathcal{S}_n,d_{\mathbf{x},\ell_1}\right)$ and $\mathcal{X}\left(\mathcal{S}_n,\tau\right)$, using the compression schemes described above. We choose $c=6$ and $c=4$ to match the expected distance of two uniformly chosen permutations respectively.

Fig. 1 shows that for $n=1024$, the trade-offs between $\hat{\delta}$ and $\hat{R}$ are very close to the theoretical limit $R(D)$.

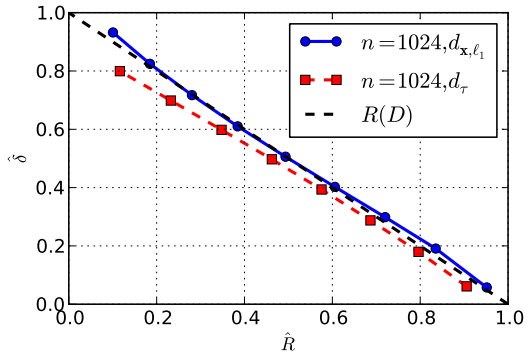

Fig. 1. Simulated rate distortion trade-offs for $n=1024$ for $\mathcal{X}\left(\mathcal{S}_n,d_{\mathbf{x},\ell_1}\right)$ and $\mathcal{X}\left(\mathcal{S}_n,\tau\right)$, where normalized rate $\hat{R}$ and normalized distortion $\hat{\delta}$ are defined in (10) and (9) respectively. For both distortion measures, the trade-offs between $\hat{\delta}$ and $\hat{R}$ are very close to the theoretical limit $R(D)$ in Theorem 6.

V. Proofs

A. Proof of Theorem 3

To prove Theorem 3, we analyze the mean and variance of the Kendall tau distance and ℓ_1 distance of inversion vectors between a permutation in $\mathcal{S}_n$ and a randomly selected permutation, in Lemma 8 and Lemma 9 respectively.

We first state the following fact without proof.

Lemma 7. *Let σ be a permutation chosen uniformly from $\mathcal{S}_n$, then $\mathbf{x}_\sigma(i)$ is uniformly distributed in $[0:i]$, $1 \le i \le n-1$.*

Lemma 8. *For any $\pi \in \mathcal{S}_n$, let σ be a permutation chosen uniformly from $\mathcal{S}_n$, and $X_\tau \triangleq d_\tau(\pi, \sigma)$, then*

$$\mathbb{E}[X_\tau] = \frac{n(n-1)}{4}, \tag{11}$$

$$\mathrm{Var}[X_\tau] = \frac{n(2n+5)(n-1)}{72}. \tag{12}$$

Proof: Let σ' be another permutation chosen independently and uniformly from $\mathcal{S}_n$, then we have both $\pi\sigma^{-1}$ and $\sigma'\sigma^{-1}$ are uniformly distributed over $\mathcal{S}_n$.

Note that Kendall tau distance is right-invariant [7], then $d_\tau(\pi, \sigma) = d_\tau\left(\pi\sigma^{-1}, e\right)$ and $d_\tau(\sigma', \sigma) = d_\tau\left(\sigma'\sigma^{-1}, e\right)$ are identically distributed, and hence the result follows [2, Table 1] and [4, Section 5.1.1]. ∎

Lemma 9. *For any $\pi \in \mathcal{S}_n$, let σ be a permutation chosen uniformly from $\mathcal{S}_n$, and $X_{\mathbf{x},\ell_1} \triangleq d_{\mathbf{x},\ell_1}(\pi, \sigma)$, then*

$$\mathbb{E}[X_{\mathbf{x},\ell_1}] > \frac{n(n-1)}{8},$$
$$\mathrm{Var}[X_{\mathbf{x},\ell_1}] < \frac{(n+1)(n+2)(2n+3)}{6}.$$

Proof: By Lemma 7, we have $X_{\mathbf{x},\ell_1} = \sum_{i=1}^{n-1} |a_i - U_i|$, where $U_i \sim \mathsf{Unif}([0:i])$ and $a_i \triangleq \mathbf{x}_\pi(i)$. Let $V_i = |a_i - U_i|$, $m_1 = \min\{i - a_i, a_i\}$ and $m_2 = \max\{i - a_i, a_i\}$, then

$$\mathbb{P}[V_i = d] = \begin{cases} 1/(i+1) & d = 0 \\ 2/(i+1) & 1 \le d \le m_1 \\ 1/(i+1) & m_1 + 1 \le d \le m_2 \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned}\mathbb{E}[V_i] &= \sum_{d=1}^{m_1} d\frac{2}{i+1} + \sum_{d=m_1+1}^{m_2} d\frac{1}{i+1} \\ &= \frac{2(1+m_1)m_1 + (m_2+m_1+1)(m_2-m_1)}{2(i+1)} \\ &= \frac{1}{2(i+1)}(m_1^2 + m_2^2 + i) \\ &\ge \frac{1}{2(i+1)}\left(\frac{(m_1+m_2)^2}{2} + i\right) = \frac{i(i+2)}{4(i+1)} > \frac{i}{4},\end{aligned}$$

$$\mathrm{Var}[V_i] \le \mathbb{E}\left[V_i^2\right] \le \frac{2}{i+1}\sum_{d=0}^{i} d^2 \le (i+1)^2.$$

Then,

$$\mathbb{E}[X_{\mathbf{x},\ell_1}] = \sum_{i=1}^{n-1}\mathbb{E}[V_i] > \frac{n(n-1)}{8},$$
$$\mathrm{Var}[X_{\mathbf{x},\ell_1}] = \sum_{i=1}^{n-1}\mathrm{Var}[V_i] < \frac{(n+1)(n+2)(2n+3)}{6}.$$

∎

With Lemma 8 and Lemma 9, now we show that the event that a scaled version of the Kendall tau distance is larger than the ℓ_1 distance of inversion vectors is unlikely.

Proof for Theorem 3: Let $c_1 = 1/3$, let $t = n^2/7$, then noting

$$\begin{aligned}t &= \mathbb{E}[c \cdot X_\tau] + \left|\Theta\left(\sqrt{n}\right)\right| \mathrm{Std}[X_\tau] \\ &= \mathbb{E}[X_{\mathbf{x},\ell_1}] - \left|\Theta\left(\sqrt{n}\right)\right| \mathrm{Std}[X_{\mathbf{x},\ell_1}],\end{aligned}$$

by Chebyshev inequality,

$$\begin{aligned}\mathbb{P}[c \cdot X_\tau > X_{\mathbf{x},\ell_1}] &\le \mathbb{P}[c \cdot X_\tau > t] + \mathbb{P}[X_{\mathbf{x},\ell_1} < t] \\ &\le O(1/n) + O(1/n) = O(1/n).\end{aligned}$$

The general case of $c_1 < 1/2$ can be proved similarly. ∎

B. Proof for Theorem 4

Lemma 10. *For any $\pi \in \mathcal{S}_n$, let σ be a permutation chosen uniformly from $\mathcal{S}_n$, and $X_{\ell_1} \triangleq d_{\ell_1}(\pi, \sigma)$, then*

$$\mathbb{E}[X_{\ell_1}] = \frac{n^2}{3} + O(n), \qquad \mathrm{Var}[X_{\ell_1}] = \frac{2n^3}{45} + O\left(n^2\right).$$

Proof: See [2, Table 1]. ∎

Proof for Theorem 4: For any $c > 0$, $cn \cdot d_{\ell_\infty}(\pi, \sigma) \le cn(n-1)$, and for any $c_2 < 1/3$, Lemma 10 and Chebyshev inequality indicate $\mathbb{P}[d_{\ell_1}(\pi, \sigma) < c_2 n(n-1)] = O(1/n)$. Therefore,

$$\begin{aligned}&\mathbb{P}[d_{\ell_1}(\pi, \sigma) \ge c_2 n \cdot d_{\ell_\infty}(\pi, \sigma)] \\ &\ge \mathbb{P}[d_{\ell_1}(\pi, \sigma) \ge c_2 n(n-1)] \\ &= 1 - \mathbb{P}[d_{\ell_1}(\pi, \sigma) < c_2 n(n-1)] \\ &= 1 - O(1/n).\end{aligned}$$

∎

References

[1] J. Giesen, E. Schuberth, and M. Stojakovi, "Approximate sorting," in *LATIN 2006: Theoretical Informatics*, J. R. Correa, A. Hevia, and M. Kiwi, Eds. Berlin, Heidelberg: Springer, 2006, vol. 3887, pp. 524–531.

[2] P. Diaconis and R. L. Graham, "Spearman's footrule as a measure of disarray," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 2, pp. 262–268, 1977.

[3] D. Wang, A. Mazumdar, and G. W. Wornell, "A rate-distortion theory for permutation spaces," in *Proc. IEEE Int. Symp. Inform. Th. (ISIT)*, 2013, pp. 2562–2566.

[4] D. E. Knuth, *Art of Computer Programming, Volume 3: Sorting and Searching*, 2nd ed. Addison-Wesley Professional, 1998.

[5] A. Jiang, M. Schwartz, and J. Bruck, "Error-correcting codes for rank modulation," in *Proc. IEEE Int. Symp. Inform. Th. (ISIT)*, 2008, pp. 1736–1740.

[6] A. Barg and A. Mazumdar, "Codes in permutations and error correction for rank modulation," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3158–3165, 2010.

[7] M. Deza and T. Huang, "Metrics on permutations, a survey," *Journal of Combinatorics, Information and System Sciences*, vol. 23, pp. 173–185, 1998.