algorithms

MDPI

Article

Hardness and Approximability of Dimension Reduction on the Probability Simplex

Roberto Bruno

Department of Computer Science, University of Salerno, 84084 Fisciano, Italy; rbruno@unisa.it

Abstract: Dimension reduction is a technique used to transform data from a high-dimensional space into a lower-dimensional space, aiming to retain as much of the original information as possible. This approach is crucial in many disciplines like engineering, biology, astronomy, and economics. In this paper, we consider the following dimensionality reduction instance: Given an n-dimensional probability distribution p and an integer $m < n$, we aim to find the m-dimensional probability distribution q that is the closest to p, using the Kullback–Leibler divergence as the measure of closeness. We prove that the problem is strongly NP-hard, and we present an approximation algorithm for it.

Keywords: dimension reduction; NP-completeness; approximation; bin packing; Kullback–Leibler divergence

1. Introduction

Dimension reduction [1,2] is a methodology for mapping data from a high-dimensional space to a lower-dimensional space, while approximately preserving the original information content. This process is essential in fields such as engineering, biology, astronomy, and economics, where large datasets with high-dimensional points are common.

It is often the case that the computational complexity of the algorithms employed to extract relevant information from these datasets depends on the dimension of the space where the points lie. Therefore, it is important to find a representation of the data in a lower-dimensional space that still (approximately) preserves the information content of the original data, as per given criteria.

A special case of the general issue illustrated before arises when the elements of the dataset are n-dimensional probability distributions, and the problem is to approximate them by lower-dimensional ones. This question has been extensively studied in different contexts. In [3,4], the authors address the problem of dimensionality reduction on sets of probability distributions with the aim of preserving specific properties, such as pairwise distances. In [5], Gokhale considers the problem of finding the distribution that minimizes, subject to a set of linear constraints on the probabilities, the "discrimination information" with respect to a given probability distribution. Similarly, in [6], Globerson et al. address the dimensionality reduction problem by introducing a nonlinear method aimed at minimizing the loss of mutual information from the original data. In [7], Lewis explores dimensionality reduction for reducing storage requirements and proposes an approximation method based on the maximum entropy criterion. Likewise, in [8], Adler et al. apply dimensionality reduction to storage applications, focusing on the efficient representation of large-alphabet probability distributions. More closely related to the dimensionality reduction that we deal with in this paper are the works [9–12]. In [10,11], the authors address task scheduling problems where the objective is to allocate tasks of a project in a way that maximizes the likelihood of completing the project by the deadline. They formalize the problem in terms of random variables approximation by using the Kolmogorov distance as a measure of distance and present an optimal algorithm for the problem. In contrast, in [12], Vidyasagar

Citation: Bruno, R. Hardness and Approximability of Dimension Reduction on the Probability Simplex. *Algorithms* **2024**, *17*, 296. https://doi.org/10.3390/a17070296

Academic Editor: Jesper Jansson

Received: 17 May 2024
Revised: 25 June 2024
Accepted: 4 July 2024
Published: 6 July 2024

defines a metric distance between probability distributions on two distinct finite sets of possibly different cardinalities based on the Minimum Entropy Coupling (MEC) problem. Informally, in the MEC, given two probability distributions p and q, one seeks to find a joint distribution ϕ that has p and q as marginal distributions and also has minimum entropy. Unfortunately, computing the MEC is NP-hard, as shown in [13]. However, numerous works in the literature present efficient algorithms for computing couplings with entropy within a constant number of bits from the optimal value [14–18]. We note that computing the coupling of a pair of distributions can be seen as essentially the inverse of dimension reduction. Specifically, given two distributions p and q, one constructs a third, larger distribution ϕ, such that p and q are derived from ϕ or, more formally, aggregations of ϕ. In contrast, the dimension reduction problem addressed in this paper involves starting with a distribution p and creating another, smaller distribution that is derived from p or, more formally, is an aggregation of p.

Moreover, in [12], the author demonstrates that, according to the defined metric, any optimal reduced-order approximation must be an aggregation of the original distribution. Consequently, the author provides an approximation algorithm based on the total variation distance, using an approach similar to the one we will employ in Section 4. Similarly, in [9], Cicalese et al. examine dimensionality reduction using the same distance metric introduced in [12]. They propose a general criterion for approximating p with a shorter vector q, based on concepts from Majorization theory, and provide an approximation approach to solve the problem.

We also mention that analogous problems arise in *scenario reduction* [19], where the problem is to (best) approximate a given discrete distribution with another distribution with fewer atoms in compressing probability distributions [20] and elsewhere [21–23]. Moreover, we recommend the following survey for further application examples [24].

In this paper, we study the following instantiation of the general problem described before: Given an n-dimensional probability distribution $p = (p_1, \ldots, p_n)$, and $m < n$, find the m-dimensional probability distribution $q = (q_1, \ldots, q_m)$ that is the *closest* to p, where the measure of closeness is the well-known relative entropy [25] (also known as Kullback–Leibler divergence). In Section 2, we formally state the problem. In Section 3, we prove that the problem is strongly NP-hard, and in Section 4, we provide an approximation algorithm returning a solution whose distance from p is at most 1 plus the minimum possible distance.

2. Statement of the Problem and Mathematical Preliminaries

Let

$$\mathcal{P}_n = \{p = (p_1, \ldots, p_n) \mid p_1 \geq \cdots \geq p_n > 0, \sum_{i=1}^{n} p_i = 1\} \tag{1}$$

be the $(n-1)$-dimensional probability simplex. Given two probability distributions $p \in \mathcal{P}_n$ and $q \in \mathcal{P}_m$, with $m < n$, we say that q is an *aggregation* of p if each component of q can be expressed as the sum of distinct components of p. More formally, q is an aggregation of p if there exists a *partition* $\Pi = (\Pi_1, \ldots, \Pi_m)$ of $\{1, \ldots, n\}$ such that $q_i = \sum_{j \in \Pi_i} p_j$, for each $i = 1, \ldots, m$. Notice that the aggregation operation corresponds to the following operation on random variables: Given a random variable X that takes value in a finite set $\mathcal{X} = \{x_1, \ldots, x_n\}$, such that $\Pr\{X = x_i\} = p_i$ for $i = 1, \ldots, n$, *any* function $f : \mathcal{X} \mapsto \mathcal{Y}$, with $\mathcal{Y} = \{y_1, \ldots, y_m\}$ and $m < n$, induces a random variable $f(X)$ whose probability distribution $q = (q_1, \ldots, q_m)$ is an aggregation of p. Dimension reduction in random variables through the application of deterministic functions is a common technique in the area (e.g., [10,12,26]). Additionally, the problem arises also in the area of "hard clustering" [27] where one seeks a deterministic mapping f from data, generated by an r.v. X taking values in a set $\mathcal{X}$, to "labels" in some set $\mathcal{Y}$, where typically $|\mathcal{Y}| \ll |\mathcal{X}|$.

For any probability distribution $p \in \mathcal{P}_n$ and an integer $m < n$, let us denote by $\mathcal{A}_m(p)$ the set of all $q \in \mathcal{P}_m$ that are aggregations of p. Our goal is to solve the following optimization problem:

Problem 1. *Given $p \in \mathcal{P}_n$ and $m < n$, find $q^* \in \mathcal{A}_m(p)$ such that*

$$\min_{q \in \mathcal{A}_m(p)} D(q\|p) = D(q^*\|p), \tag{2}$$

where $D(q\|p)$ is the relative entropy [25], given by

$$D(q\|p) = \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i},$$

and the logarithm is of base 2.

An additional motivation to study Problem 1 comes from the fundamental paper [28], in which the principle of minimum relative entropy (called therein *minimum cross entropy principle*) is derived in an axiomatic manner. The principle states that, of the distributions q that satisfy given constraints (in our case, that $q \in \mathcal{A}_m(p)$), one should choose the one with the least relative entropy "distance" from the prior p.

Before establishing the computational complexity of the Problem 1, we present a simple lower bound on the optimal value.

Lemma 1. *For each $p \in \mathcal{P}_n$ and $q \in \mathcal{P}_m$, $m < n$, it holds that*

$$D(q\|p) \geq D(lb(p)\|p) = -\log\left(\sum_{i=1}^{m} p_i\right). \tag{3}$$

where

$$lb(p) = \left(\frac{p_1}{\sum_{i=1}^{m} p_i}, \ldots, \frac{p_m}{\sum_{i=1}^{m} p_i}\right) \in \mathcal{P}_m. \tag{4}$$

Proof. Given an arbitrary $p \in \mathcal{P}_n$, one can see that

$$D(lb(p)\|p) = -\log\left(\sum_{i=1}^{m} p_i\right).$$

Moreover, for *any* $p \in \mathcal{P}_n$ *and* $q \in \mathcal{P}_m$, the Jensen inequality applied to the log function gives the following:

$$-D(q||p) = \sum_{i=1}^{m} q_i \log \frac{p_i}{q_i} \leq \log\left(\sum_{i=1}^{m} p_i\right).$$

□

3. Hardness

In this section, we prove that the optimization problem (2) described in Section 1 is strongly NP-hard. We accomplish this by reducing the problem from the 3-PARTITION problem, a well-known strongly NP-hard problem [29], described as follows.

3-PARTITION: Given a multiset $S = \{a_1, \ldots, a_n\}$ of $n = 3m$ positive integers for which $\sum_{i=1}^{n} a_i = mT$, for some T, the problem is to decide whether S can be partitioned into m triplets such that the sum of each triple is exactly T. More formally, the problem is to decide whether there exist $S_1, \ldots, S_m \subseteq S$ such that the following conditions hold:

$$\begin{aligned}
\sum_{a \in S_j} a &= T, \quad \forall j \in \{1, \ldots, m\}, \\
S_i \cap S_j &= \emptyset, \quad \forall i \neq j, \\
\bigcup_{i=1}^{m} S_i &= S, \\
|S_i| &= 3, \quad \forall i \in \{1, \ldots, m\}.
\end{aligned}$$

Theorem 1. *The* 3-PARTITION *problem can be reduced in polynomial time to the problem of finding the aggregation* $q^* \in \mathcal{P}_m$ *of some* $p \in \mathcal{P}_n$*, for which*

$$D(q^* \| p) = \min_{q \in \mathcal{A}_m(p)} D(q \| p).$$

Proof. The idea behind the following reduction can be summarized as follows: given an instance of 3-PARTITION, we transform it into a probability distribution p such that the lower bound $lb(p)$ is an aggregation of p if and only if the original instance of 3-PARTITION admits a solution. Let an arbitrary instance of 3-PARTITION be given, that is, let S be a multiset $\{a_1, \ldots, a_n\}$ of $n = 3m$ positive integers with $\sum_{i=1}^{n} a_i = mT$. Without loss of generality, we assume that the integers a_i are ordered in a non-increasing fashion. We construct a valid instance p of our Problem 1 as follows. We set $p \in \mathcal{P}_{n+m}$ as follows:

$$p = \left(\underbrace{\frac{1}{m+1}, \ldots, \frac{1}{m+1}}_{m \text{ times}}, \frac{a_1 + 2T}{(m+1)7mT}, \ldots, \frac{a_n + 2T}{(m+1)7mT} \right). \tag{5}$$

Note that p is a probability distribution. In fact, since $n = 3m$, we have

$$\sum_{i=1}^{n} \frac{a_i + 2T}{(m+1)7mT} = \frac{1}{(m+1)7mT} \left(\sum_{i=1}^{n} (a_i + 2T) \right) = \frac{7mT}{(m+1)7mT} = \frac{1}{m+1}.$$

Moreover, from (4) and (5), the probability distribution $lb(p) \in \mathcal{P}_m$ associated to p is as follows:

$$lb(p) = \left(\frac{p_1}{\sum_{j=1}^{m} p_j}, \ldots, \frac{p_m}{\sum_{j=1}^{m} p_j} \right) = \left(\frac{1}{m}, \ldots, \frac{1}{m} \right). \tag{6}$$

To prove the theorem, we show that the starting instance of 3-PARTITION is a YES instance *if and only if* it holds that

$$\min_{q \in \mathcal{A}_m(p)} D(q \| p) = \log \frac{m+1}{m}, \tag{7}$$

where p is given in (5).

We begin by assuming the given instance of 3-PARTITION is a YES instance, that is, there is a partition of S into triplets $S_1, \ldots, S_m$ such that

$$\sum_{a_i \in S_j} a_i = T, \quad \forall j \in \{1, \ldots, m\}, \tag{8}$$

and we show that $\min_{q \in \mathcal{A}_m(p)} D(q \| p) = \log \frac{m+1}{m}$. By Lemma 1, (5), and equality (6), we have

$$\min_{q \in \mathcal{A}_m(p)} D(q \| p) \geq D(lb(p) \| p) = \sum_{i=1}^{m} \frac{1}{m} \log \frac{1/m}{1/(m+1)} = \log \frac{m+1}{m}. \tag{9}$$

From (8), we have

$$\begin{aligned}\sum_{a_i \in S_j} \frac{a_i + 2T}{(m+1)7mT} &= \frac{T}{(m+1)7mT} + \sum_{a_i \in S_j} \frac{2T}{(m+1)7mT} \\ &= \frac{T}{(m+1)7mT} + \frac{6T}{(m+1)7mT} \\ &= \frac{1}{(m+1)m}, \qquad \forall j \in \{1, \ldots, m\}. \end{aligned} \tag{10}$$

Let us define $q' \in \mathcal{P}_m$ as follows:

$$q' = \left(\frac{1}{m+1} + \sum_{a_i \in S_1} \frac{a_i + 2T}{(m+1)7mT}, \ldots, \frac{1}{m+1} + \sum_{a_i \in S_m} \frac{a_i + 2T}{(m+1)7mT} \right), \tag{11}$$

where, by (10),

$$\sum_{a_i \in S_j} \frac{a_i + 2T}{(m+1)7mT} = \frac{1}{(m+1)m}, \quad \forall j \in \{1, \ldots, m\}. \tag{12}$$

From (12) and from the fact that $S_1, \ldots, S_m$ are a partition of $\{a_1, \ldots, a_n\}$, we obtain $q' \in \mathcal{A}_m(p)$, that is, q' is a valid aggregation of p (cfr., (5)). Moreover,

$$q' = \left(\frac{1}{m}, \ldots, \frac{1}{m} \right),$$

and $D(q' \| p) = \log \frac{m+1}{m}$. Therefore, by (9) and that $q' \in \mathcal{A}_m(p)$, we obtain

$$\min_{q \in \mathcal{A}_m(p)} D(q \| p) = \log \frac{m+1}{m},$$

as required.

To prove the opposite implication, we assume that p (as given in (5)) is a YES instance, that is,

$$\min_{q \in \mathcal{A}_m(p)} D(q \| p) = \log \frac{m+1}{m}. \tag{13}$$

We show that the original instance of 3-PARTITION is also a YES instance, that is, there is a partition of S into triplets $S_1, \ldots, S_m$ such that

$$\sum_{a_i \in S_j} a_i = T, \quad \forall j \in \{1, \ldots, m\}. \tag{14}$$

Let q^* be the element in $\mathcal{A}_m(p)$ that achieves the minimum in (13). Consequently, we have

$$\begin{aligned}\log \frac{m+1}{m} = D(q^* \| p) &= \sum_{i=1}^{m} q_i^* \log \frac{q_i^*}{p_i} = \sum_{i=1}^{m} q_i^* \log \frac{1}{p_i} - H(q^*) \\ &= \log(m+1) - H(q^*) \qquad \text{(from (5))}, \end{aligned} \tag{15}$$

where $H(q^*) = -\sum_{i=1}^{m} q_i^* \log q_i^*$ is the Shannon entropy of q^*. From (15), we obtain that $H(q^*) = \log m$; hence, $q^* = (1/m, \ldots, 1/m)$ (see [30], Thm. 2.6.4). Recalling that $q^* \in \mathcal{A}_m(p)$, we obtain that the uniform distribution

$$\left(\frac{1}{m}, \ldots, \frac{1}{m} \right) \tag{16}$$

is an *aggregation* of p. We note that the first m components of p, as defined in (5), cannot be aggregated among them to obtain (16), because $2/(m+1) > 1/m$, for $m > 2$. Therefore,

in order to obtain (16) as an aggregation of p, there must exist a partition $S_1, \ldots, S_m$ of $S = \{a_1, \ldots, a_n\}$ for which

$$\frac{1}{m+1} + \sum_{a_i \in S_j} \frac{a_i + 2T}{(m+1)7mT} = \frac{1}{m}, \quad \forall j \in \{1, \ldots, m\}. \tag{17}$$

From (17), we obtain

$$\sum_{a_i \in S_j} \frac{a_i + 2T}{(m+1)7mT} = \frac{1}{m(m+1)}, \quad \forall j \in \{1, \ldots, m\}. \tag{18}$$

From this, it follows that

$$2T|S_j| + \sum_{a_i \in S_j} a_i = 7T, \quad \forall j \in \{1, \ldots, m\}. \tag{19}$$

We note that, for (19) to be true, there cannot exist any S_j for which $|S_j| \neq 3$. Indeed, if there were a subset S_j for which $|S_j| \neq 3$, there would be at least a subset S_k for which $|S_k| > 3$. Thus, for such an S_k, we would have

$$2T|S_k| + \sum_{a_i \in S_k} a_i \geq 8T + \sum_{a_i \in S_k} a_i > 7T,$$

contradicting (19). Therefore, it holds that

$$|S_j| = 3, \quad \forall j \in \{1, \ldots, m\}. \tag{20}$$

Moreover, from (19) and (20), we obtain

$$\sum_{a_i \in S_j} a_i = 7T - 2T|S_j| = T, \quad \forall j \in \{1, \ldots, m\}. \tag{21}$$

Thus, from (21), it follows that the subsets $S_1, \ldots, S_m$ give a partition of S into triplets, such that

$$\sum_{a_i \in S_j} a_i = T, \quad \forall j \in \{1, \ldots, m\}.$$

Therefore, the starting instance of 3-PARTITION is a YES instance. □

4. Approximation

Given $p \in \mathcal{P}_n$ and $m < n$, let OPT denote the optimal value of the optimization problem (2), that is

$$OPT = \min_{q \in \mathcal{A}_m(p)} D(q\|p). \tag{22}$$

In this section, we design a greedy algorithm to compute an aggregation $\bar{q} \in \mathcal{A}_m(p)$ of p such that

$$D(\bar{q}\|p) < OPT + 1. \tag{23}$$

The idea behind our algorithm is to see the problem of computing an aggregation $q \in \mathcal{A}_m(p)$ as a bin packing problem with "overstuffing" (see [31] and references therein quoted), which is a bin packing where overfilling of bins is possible. In the classical bin packing problem, one is given a set of items, with their associated weights, and a set of bins with their associated capacities (usually, equal for all bins). The objective is to place all the items in the bins, trying to minimize a given cost function.

In our case, we have n items (corresponding to the components of p) with weights $p_1, \ldots, p_n$, respectively, and m bins, corresponding to the components of $lb(p)$ (as defined in (4)) with capacities $lb(p)_1, \ldots, lb(p)_m$. Our objective is to place all the n components of p into the m bins *without exceeding* the capacity $lb(p)_j$ of each bin j, $j = 1, \ldots, m$, by more than

$(\sum_{i=1}^{m} p_i)lb(p)_j$. For such a purpose, the idea behind Algorithm 1 is quite straightforward. It behaves like a classical First-Fit bin packing: to place the ith item, it chooses the first bin j in which the item can be inserted without exceeding its capacity by more than $(\sum_{i=1}^{m} p_i)lb(p)_j$. In the following, we will show that such a bin always exists and that fulfilling this objective is sufficient to ensure the approximation guarantee (23) we are seeking.

Algorithm 1: GreedyApprox

1. Compute $lb(p) = (p_1/\sum_{j=1}^{m} p_j, \ldots, p_m/\sum_{j=1}^{m} p_j)$;
2. Let lb_j^i be the content of bin j after the first i components of p have been placed ($lb_j^0 = 0$ for each $j \in \{1, \ldots, m\}$);
3. For $i = 0, \ldots, n-1$
   Let j be the smallest bin index for which holds that $lb_j^i + p_{i+1} < (1 + \sum_{j=1}^{m} p_j)lb(p)_j$, place p_{i+1} into the j-th bin:
   $lb_j^{i+1} = lb_j^i + p_{i+1}$,
   $lb_k^{i+1} = lb_k^i$, for each $k \neq j$;
4. Output $\overline{q} = (lb_1^n, \ldots, lb_m^n)$.

The step 3 of GreedyApprox operates as in the classical First-Fit bin packing algorithm. Therefore, it can be implemented to run in $O(n \log m)$ time, as discussed in [32]. In fact, each iteration of the loop in step 3 can be implemented in $O(\log m)$-time by using a balanced binary search tree with height $O(\log m)$ that has a leaf for each bin and in which each node keeps track of the largest remaining capacity of all the bins in its subtree.

Lemma 2. *GreedyApprox computes a valid aggregation $\overline{q} \in \mathcal{A}_m(p)$ of $p \in \mathcal{P}_n$. Moreover, it holds that*

$$D(\overline{q} \| lb(p)) < \log\left(1 + \sum_{j=1}^{m} p_j\right). \tag{24}$$

Proof. We first prove that each component p_i of p is placed in some bin. This implies that $\overline{q} \in \mathcal{A}_m(p)$.

For each step $i = 0, \ldots, m-1$, there is always a bin in which the algorithm places p_{i+1}. In fact, the capacity $lb(p)_j$ of bin j satisfies the relation:

$$lb(p)_j = \frac{p_j}{\sum_{\ell=1}^{m} p_\ell} > p_j, \quad \forall j \in \{1, \ldots, m\}.$$

Let us consider an arbitrary step $m \leq i < n$, in which the algorithm has placed the first i components of p and needs to place p_{i+1} into some bin. We show that, in this case also, there is always a bin j in which the algorithm places the item p_{i+1}, *without exceeding* the capacity $lb(p)_j$ of the bin j by more than $(\sum_{\ell=1}^{m} p_\ell)lb(p)_j$.

First, notice that in each step i, $m \leq i < n$, there is at least a bin k whose content lb_k^i *does not* exceed its capacity $lb(p)_k$; that is, for which $lb_k^i < lb(p)_k$ holds. Were this the opposite, for all bins j, we would have $lb_j^i \geq lb(p)_j$; then, we would also have

$$\sum_{j=1}^{m} lb_j^i \geq \sum_{j=1}^{m} lb(p)_j = 1. \tag{25}$$

However, this is not possible since we have placed only the first $i < n$ components of p, and therefore, it holds that

$$\sum_{j=1}^{m} lb_j^i = \sum_{j=1}^{i} p_j < \sum_{j=1}^{n} p_j = 1,$$

contradicting (25). Consequently, let k be the smallest integer for which the content of the k-th bin does not exceed its capacity, i.e., for which $lb_k^i < lb(p)_k$. For such a bin k, we obtain

$$\begin{aligned}\left(1+\sum_{j=1}^{m} p_j\right) lb(p)_k &= lb(p)_k + \left(\sum_{j=1}^{m} p_j\right) lb(p)_k \\ &= lb(p)_k + \left(\sum_{j=1}^{m} p_j\right) \frac{p_k}{\sum_{j=1}^{m} p_j} \\ &= lb(p)_k + p_k \\ &> lb_k^i + p_k \qquad (\text{since } lb(p)_k > lb_k^i) \\ &\geq lb_k^i + p_{i+1} \qquad (\text{since } p_k \geq p_{i+1}). \end{aligned} \tag{26}$$

Thus, from (26), one derives that the algorithm places p_{i+1} into the bin k without exceeding its capacity $lb(p)_k$ by more than $(\sum_{j=1}^{m} p_j) lb(p)_k$.

The reasoning applies to each $i < n$, thus proving that GreedyApprox correctly assigns each component p_i of p to a bin, effectively computing an aggregation of p. Moreover, from the instructions of step 3 of GreedyApprox, the output is an aggregation $\overline{q} = (\overline{q}_1, \ldots, \overline{q}_m) \in \mathcal{A}_m(p)$, for which the following crucial relation holds:

$$\overline{q}_i < \left(1+\sum_{j=1}^{m} p_j\right) lb(p)_i, \quad \forall i \in \{1, \ldots, m\}. \tag{27}$$

Let us now prove that $D(\overline{q} \| lb(p)) < \log\left(1+\sum_{j=1}^{m} p_j\right)$. We have

$$\begin{aligned} D(\overline{q} \| lb(p)) &= \sum_{i=1}^{m} \overline{q}_i \log \frac{\overline{q}_i}{lb(p)_i} \\ &< \sum_{i=1}^{m} \overline{q}_i \log \frac{(1+\sum_{j=1}^{m} p_j) lb(p)_i}{lb(p)_i} \qquad (\text{from (27)}) \\ &= \log\left(1+\sum_{j=1}^{m} p_j\right). \end{aligned}$$

□

We need the following technical lemma to show the approximation guarantee of GreedyApprox.

Lemma 3. *Let $q \in \mathcal{P}_m$ and $p \in \mathcal{P}_n$ be two arbitrary probability distributions with $m < n$. It holds that*

$$D(q \| p) = D(q \| lb(p)) + D(lb(p) \| p), \tag{28}$$

where $lb(p) = (lb(p)_1, \ldots, lb(b)_m) = (p_1 / \sum_{i=1}^{m} p_i, \ldots, p_m / \sum_{i=1}^{m} p_i)$.

Proof.

$$\begin{aligned} D(q\|p) &= \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i} = \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i \frac{\sum_{j=1}^{m} p_j}{\sum_{j=1}^{m} p_j}} \\ &= \sum_{i=1}^{m} q_i \log \frac{q_i}{\frac{p_i}{\sum_{j=1}^{m} p_j}} + \sum_{i=1}^{m} q_i \log \frac{1}{\sum_{j=1}^{m} p_j} \\ &= \sum_{i=1}^{m} q_i \log \frac{q_i}{lb(p)_i} + \sum_{i=1}^{m} q_i \log \frac{1}{\sum_{j=1}^{m} p_j} \qquad (\text{since } lb(p)_i = p_i / \textstyle\sum_{j=1}^{m} p_j) \\ &= D(q\|lb(p)) + \log \frac{1}{\sum_{j=1}^{m} p_j} = D(q\|lb(p)) + D(lb(p)\|p). \end{aligned}$$

□

The following theorem is the main result of this section.

Theorem 2. *For any $p \in \mathcal{P}_n$ and $m < n$, GreedyApprox produces an aggregation $\overline{q} \in \mathcal{A}_m(p)$ of p such that*

$$D(\overline{q}\|p) < OPT + 1, \tag{29}$$

where $OPT = \min_{q \in \mathcal{A}_m(p)} D(q\|p)$.

Proof. From Lemma 3, we have

$$D(\overline{q}\|p) = D(\overline{q}\|lb(p)) + D(lb(p)\|p), \tag{30}$$

and from Theorem 2, we know that the produced aggregation $\overline{q}$ of p satisfies the relation

$$D(\overline{q}\|lb(p)) < \log\left(1 + \sum_{j=1}^{m} p_j\right). \tag{31}$$

Putting it all together, we obtain:

$$\begin{aligned} D(\overline{q}\|p) &= D(\overline{q}\|lb(p)) + D(lb(p)\|p) \\ &< \log\left(1 + \sum_{j=1}^{m} p_j\right) + D(lb(p)\|p) \qquad (\text{from (31)}) \\ &= \log\left(1 + \sum_{j=1}^{m} p_j\right) - \log\left(\sum_{j=1}^{m} p_j\right) \\ &< -\log\left(\sum_{j=1}^{m} p_j\right) + 1 \qquad (\text{since } 1 + \textstyle\sum_{j=1}^{m} p_j < 2) \\ &\leq OPT + 1 \qquad (\text{from Lemma 1}). \end{aligned}$$

□

5. Concluding Remarks

In this paper, we examined the problem of approximating n-dimensional probability distributions with m-dimensional ones using the Kullback–Leibler divergence as the measure of closeness. We demonstrated that this problem is strongly NP-hard and introduced an approximation algorithm for solving the problem with guaranteed performance.

Moreover, we conclude by pointing out that the analysis of GreedyApprox presented in Theorem 2 is tight. Let $p \in \mathcal{P}_3$ be

$$p = \left(\frac{1}{2} - \epsilon, \frac{1}{2} - \epsilon, 2\epsilon\right),$$

where $\epsilon > 0$. The application of GreedyApprox on p produces the aggregation $\overline{q} \in \mathcal{P}_2$ given by

$$\overline{q} = (1 - 2\epsilon, 2\epsilon),$$

whereas one can see that the optimal aggregation $q^* \in \mathcal{P}_2$ is equal to

$$q^* = \left(\frac{1}{2} + \epsilon, \frac{1}{2} - \epsilon\right).$$

Hence, for $\epsilon \to 0$, we have

$$D(\overline{q}\|p) = (1 - 2\epsilon) \log \frac{1 - 2\epsilon}{\frac{1}{2} - \epsilon} + 2\epsilon \log \frac{2\epsilon}{\frac{1}{2} - \epsilon} \to 1,$$

while

$$OPT = D(q^*\|p) = \left(\frac{1}{2} + \epsilon\right) \log \frac{\frac{1}{2} + \epsilon}{\frac{1}{2} - \epsilon} + \left(\frac{1}{2} - \epsilon\right) \log \frac{\frac{1}{2} - \epsilon}{\frac{1}{2} - \epsilon} \to 0.$$

Therefore, to improve our approximation guarantee, one should use a bin packing heuristic different from the First-Fit as employed in GreedyApprox. Another interesting open problem is to provide an approximation algorithm with a (small) multiplicative approximation guarantee. However, both problems mentioned above would probably require a different approach, and we leave that to future investigations.

Another interesting line of research would be to extend our findings to different divergence measures (e.g., [33] and references quoted therein).

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The author wants to express his gratitude to Ugo Vaccaro for guidance throughout this research, to the anonymous referees, and to the Academic Editor for many useful suggestions that have improved the presentation of the paper.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Burges, C.J. Dimension reduction: A guided tour. *Found. Trends Mach. Learn.* **2010**, *2*, 275–365. [CrossRef]
2. Sorzano, C.O.S.; Vargas, J.; Montano, A.P. A survey of dimensionality reduction techniques. *arXiv* **2014**, arXiv:1403.2877.
3. Abdullah, A.; Kumar, R.; McGregor, A.; Vassilvitskii, S.; Venkatasubramanian, S. Sketching, Embedding, and Dimensionality Reduction for Information Spaces. *Artif. Intell. Stat. PMLR* **2016**, *51*, 948–956.
4. Carter, K.M.; Raich, R.; Finn, W.G.; Hero, A.O., III. Information-geometric dimensionality reduction. *IEEE Signal Process. Mag.* **2011**, *28*, 89–99. [CrossRef]
5. Gokhale, D.V. Approximating discrete distributions, with applications. *J. Am. Stat. Assoc.* **1973**, *68*, 1009–1012. [CrossRef]
6. Globerson, A.; Tishby, N. Sufficient dimensionality reduction. *J. Mach. Learn. Res.* **2003**, *3*, 1307–1331.
7. Lewis, P.M., II. Approximating probability distributions to reduce storage requirements. *Inf. Control.* **1959**, *2*, 214–225. [CrossRef]
8. Adler, A.; Tang, J.; Polyanskiy, Y. Efficient representation of large-alphabet probability distributions. *IEEE Sel. Areas Inf. Theory* **2022**, *3*, 651–663. [CrossRef]

9. Cicalese, F.; Gargano, L.; Vaccaro, U. Approximating probability distributions with short vectors, via information theoretic distance measures. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1138–1142.
10. Cohen, L.; Grinshpoun, T.; Weiss, G. Efficient optimal Kolmogorov approximation of random variables. *Artif. Intell.* **2024**, *329*, 104086. [CrossRef]
11. Cohen, L.; Weiss, G. Efficient optimal approximation of discrete random variables for estimation of probabilities of missing deadlines. *Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 7809–7815. [CrossRef]
12. Vidyasagar, M. A metric between probability distributions on finite sets of different cardinalities and applications to order reduction. *IEEE Trans. Autom. Control.* **2012**, *57*, 2464–2477. [CrossRef]
13. Kovačević, M.; Stanojević, I.; Šenk, V. On the entropy of couplings. *Inf. Comput.* **2015**, *242*, 369–382. [CrossRef]
14. Cicalese, F.; Gargano, L.; Vaccaro, U. Minimum-entropy couplings and their applications. *IEEE Trans. Inf. Theory* **2019**, *65*, 3436–3451. [CrossRef]
15. Compton, S. A tighter approximation guarantee for greedy minimum entropy coupling. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Espoo, Finland, 26 June–1 July 2022; pp. 168–173.
16. Compton, S.; Katz, D.; Qi, B.; Greenewald, K.; Kocaoglu, M. Minimum-entropy coupling approximation guarantees beyond the majorization barrier. *Int. Conf. Artif. Intell. Stat.* **2023**, *206*, 10445–10469.
17. Li, C. Efficient approximate minimum entropy coupling of multiple probability distributions. *IEEE Trans. Inf. Theory* **2021**, *67*, 5259–5268. [CrossRef]
18. Sokota, S.; Sam, D.; Witt, C.; Compton, S.; Foerster, J.; Kolter, J. Computing Low-Entropy Couplings for Large-Support Distributions. *arXiv* **2024**, arXiv:2405.19540.
19. Rujeerapaiboon, N.; Schindler, K.; Kuhn, D.; Wiesemann, W. Scenario reduction revisited: Fundamental limits and guarantees. *Math. Program.* **2018**, *191*, 207–242. [CrossRef]
20. Gagie, T. Compressing probability distributions. *Inf. Process. Lett.* **2006**, *97*, 133–137. [CrossRef]
21. Cohen, L.; Fried, D.; Weiss, G. An optimal approximation of discrete random variables with respect to the Kolmogorov distance. *arXiv* **2018**, arXiv:1805.07535.
22. Pavlikov, K.; Uryasev, S. CVaR distance between univariate probability distributions and approximation problems. *Ann. Oper. Res.* **2018**, *262*, 67–88. [CrossRef]
23. Pflug, G.C.; Pichler, A. Approximations for probability distributions and stochastic optimization problems. In *Stochastic Optimization Methods in Finance and Energy: New Financial Products and Energy Market Strategies*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 343–387.
24. Melucci, M. A brief survey on probability distribution approximation. *Comput. Sci. Rev.* **2019**, *33*, 91–97. [CrossRef]
25. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
26. Lamarche-Perrin, R.; Demazeau, Y.; Vincent, J.M. The best-partitions problem: How to build meaningful aggregations. In Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 17–20 November 2013; pp. 399–404.
27. Kearns, M.; Mansour, Y.; Ng, A.Y. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Learning in Graphical Models*; Springer: Dordrecht, The Netherlands, 1998; pp. 495–520.
28. Shore, J.; Johnson, R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37. [CrossRef]
29. Garey, M.; Johnson, D. Strong NP-Completeness results: Motivation, examples, and implications. *J. ACM* **1978**, *25*, 499–508. [CrossRef]
30. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
31. Dell'Olmo, P.; Kellerer, H.; Speranza, M.; Tuza, Z. A 13/12 approximation algorithm for bin packing with extendable bins. *Inf. Process. Lett.* **1998**, *65*, 229–233. [CrossRef]
32. Coffman, E.G.; Garey, M.R.; Johnson, D.S. Approximation Algorithms for Bin Packing: A Survey. In *Approximation Algorithms for NP-Hard Problems*; Hochbaum, D., Ed.; PWS Publishing Co.: Worcester, UK, 1996; pp. 46–93.
33. Sason, I. Divergence Measures: Mathematical Foundations and Applications in Information-Theoretic and Statistical Problems. *Entropy* **2022**, *24*, 712. [CrossRef]