Finding Metric Structure in Information Theoretic Clustering

Kamalika Chaudhuri
University of California, San Diego
kamalika@soe.ucsd.edu

Andrew McGregor
University of California, San Diego
andrewm@ucsd.edu

Abstract

We study the problem of clustering discrete probability distributions with respect to the Kullback-Leibler (KL) divergence. This problem arises naturally in many applications. Our goal is to pick k distributions as "representatives" such that the average or maximum KL-divergence between an input distribution and the closest representative distribution is minimized. Unfortunately, no polynomial-time algorithms with worst-case performance guarantees are known for either of these problems.

The analogous problems for l_1, l_2 and l_2^2 (i.e., k-center, k-median and k-means) have been extensively studied and efficient algorithms with good approximation guarantees are known. However, these algorithms rely crucially on the (geo-)metric properties of these metrics and do not apply to KL-divergence. In this paper, our contribution is to find a "relaxed" metric-structure for KL-divergence. In doing so, we provide the first polynomial-time algorithm for clustering using KL-divergences with provable guarantees for general inputs.

1 Introduction

In this paper, we consider the problem of clustering discrete probability distributions with respect to the Kullback-Liebler (KL) divergence where, the KL-divergence from $p = (p_1, \ldots, p_d)$ to distribution $q = (q_1, \ldots, q_d)$ is defined as

$$\mathrm{KL}(p, q) = \sum_{i \in [d]} p_i \ln \frac{p_i}{q_i} \ .$$

Specifically, we consider two problems that take n distributions $p^1, \ldots, p^n$ on $[d]$ as input. In $\mathrm{MTC}_{\mathrm{KL}}$ (minimum total cost), the goal is to find distributions $c^1, \ldots, c^k$ such that the total KL-divergence from each p^j to its closest c^i, i.e.,

$$\sum_{j \in [n]} \min_{i \in [k]} \mathrm{KL}(p^j, c^i)$$

is minimized. In $\mathrm{MMC}_{\mathrm{KL}}$ (minimum maximum cost), the goal is to find distributions $c^1, \ldots, c^k$ such that the maximum KL-divergence from each p^j to its closest c^i,

$$\max_{j \in [n]} \min_{i \in [k]} \mathrm{KL}(p^j, c^i)$$

is minimized. It turns out that polynomial time algorithms do not exist for either of these problems unless $P = NP$. Therefore, we are interested in α-approximation algorithms, i.e., algorithms that find $\tilde{c}^1, \ldots, \tilde{c}^k$ satisfying the guarantee that

$$\frac{\sum_{j \in [n]} \min_{i \in [k]} \mathrm{KL}(p^j, \tilde{c}^i)}{\min_{c^1, \ldots, c^n} \sum_{j \in [n]} \min_{i \in [k]} \mathrm{KL}(p^j, c^i)} \leq \alpha$$

for some $\alpha \geq 1$. The smaller the value of α, the better the approximation.

Both problems have been studied extensively when the input is a set of arbitrary points (not necessarily distributions), and instead of KL, the measure of distance between two points is either a metric (ℓ_1 or ℓ_2 or an arbitrary metric), with symmetry and the triangle inequality, or a measure such as ℓ_2^2. The problems are usually referred to as k-median if the measure is a metric or k-means if the measure is ℓ_2^2. However, previous algorithms for these problems typically rely crucially on the (geo-)metric properties of these distances, which do not hold for the KL-divergence. For example, KL is not symmetric and does not satisfy the triangle inequality.

In the remainder of the introduction, we motivate the need to cluster distributions and the reason why KL is a natural measure in this context. We then review the related work and summarize our contributions.

Why cluster distributions? A natural application of distributional clustering is in clustering words for document classification by topic [4]. In document classification, we are given a training set of *documents* (or collections of words) whose labels indicate the topic they represent, and the goal is to classify other similar documents according to topic. A natural approach is to look at the words in a document as *features* that are somehow correlated with the document labels; each word is viewed as a frequency distribution over labels, and given a new document containing a set of words, the distributions corresponding to the words in it are used to find

the most-likely label for the new document. However, such data is typically very sparse, because each specific word occurs a few times in the document corpora. So a common approach is to cluster together similar word distributions, for more robust inference algorithms.

Other applications of distributional clustering include clustering words according to context for language modeling [24], information bottleneck techniques [27, 24, 25], and clustering users according to their preference for movies in collaborative filtering.

Why KL-divergence? KL-divergence arises as a natural measure of the dissimilarity between two distributions in numerous ways. We direct the interested reader to Pereira et al. [24] for a wider discussion on the motivations. In what follows, we describe the motivation in terms of compressibility.

Given an alphabet Σ of size d where the i-th symbol has relative frequency p_i, an important question is to find the binary encoding of the alphabet such the average number of bits required for an encoded symbol is minimized. This classic problem in information theory was essentially solved by Huffman who presented a simple encoding scheme that achieved the optimum value of

$$H(p) = -\sum_{i\in[d]} p_i \lg p_i$$

if all p_i were negative powers of two.

We consider an issue that arises when we have two or more distributions over Σ. Consider the problem of trying to encode multiple texts with different statistics such as texts written in different languages or magazine articles covering different topics. For example, the word "perforation" may be common in articles from *Gibbons Stamp Monthly Magazine*[1] whereas "peroxide" may be more frequent in issues of *Hairdressers Journal International*[2]. Hence, if the origin of the text is known it will make sense to tailor the encoding to the statistics of the the source. However, it is likely to be unfeasible to have a different scheme for every possible periodical. Rather, we consider the problem of designing k encoding schemes and assigning each of n periodicals to one of the encoding schemes. How should this be done such that extra cost of using k encoding schemes rather than n is minimized?

More formally, let p^j be distribution over symbols in the j-th periodical. We wish to design k encoding schemes $E_1, \ldots, E_k : \Sigma \to \{0,1\}^*$ along with an assignment of distributions to encoding schemes $f : [n] \to [k]$ such that the increase in average encoding length,

$$\sum_{j\in[n]}\sum_{i\in[d]} p_i^j |E_{f(j)}(i)| + \sum_{j\in[n]}\sum_{i\in[d]} p_i^j \lg p_i^j$$

is minimized. Each encoding scheme E_j can be characterized by a distribution q^j over $[d]$ that will capture the aggregate statistics of the distributions that use E_j. Hence we may rewrite the quantity to be minimized as

$$-\sum_{j\in[n]}\sum_{i\in[d]} p_i^j \lg q_i^{f(j)} + \sum_{j\in[n]}\sum_{i\in[d]} p_i^j \lg p_i^j$$

$$= \sum_{j\in[n]}\sum_{i\in[d]} p_i^j \lg \frac{p_i^j}{q_i^{f(j)}} = (\lg e)\sum_{j\in[n]} \mathrm{KL}(p^j, q^{f(j)})$$

which is exactly the objective function to be minimized in $\mathrm{MTC}_{\mathrm{KL}}$.

1.1 Prior Work on Clustering

There has been a rich body of research on approximation algorithms for various forms of clustering. We restrict ourselves to those on hard-clustering, i.e., each input distributions is "assigned" to only the closest picked center. Even so, there is a considerable number of incomparable results in a variety of settings.

The common optimization measures when clustering points in general metrics are (a) k-median, in which the goal is to partition the input points into k sets, while minimizing the sum of the distances between each point and the center of the cluster it is assigned to, and (b) k-center, where the goal is to again partition the input points to k sets, while minimizing the maximum diameter of a cluster. When clustering in Euclidean spaces, an additional optimization measure which is commonly used is k-means, in which the goal is to partition the input points into k clusters, while minimizing the sum of the squares of the Euclidean distances between each point and the center of the cluster it is assigned to.

General "Metrics": For metric k-center, the best approximation algorithm is due to [16], which achieves an approximation factor of 2 and this is the best possible in polynomial time unless $P = NP$. For asymmetric k-center, when the directed triangle inequality holds, the best known approximation algorithm is due to [23], which achieves a factor of $O(\log^* n)$, and this is also optimal in terms of hardness [7]. For metric k-median, the best known approximation algorithm is due to [2], which achieves an approximation factor of 3, when the distances between points are symmetric, and there is a triangle inequality.

Euclidean Space: When the input points lie in Euclidean space, two versions of the clustering problems have been studied. In the *restricted version*, we require the cluster centers to be input points, while in the *unrestricted version*,we allow the cluster centers to be any point in the Euclidean space. For more details about restricted and unrestricted versions of the problems, see Section 5. Most results for clustering in Euclidean space deal with the unrestricted version of the problem.

When the input points lie in d-dimensional Euclidean spaces, Kolliopoulos and Rao [21] showed an algorithm for k-median which provides a $(1+\epsilon)$ approximation, and runs in time

$$O(2^{(O(1+\epsilon^{-1}\log\epsilon^{-1}))^{d-1}} n \log k \log n)\ .$$

[1] http://www.gibbonsstampmonthly.com/

[2] http://www.hji.co.uk/

Har-Peled and Mazumdar [19] gave a $(1+\epsilon)$ approximation algorithm which runs in time

$$O(n + 2^{O(1+\epsilon^{-1}\log\epsilon^{-1})^{d-1}} k^{O(1)} \log^{O(1)} n) .$$

A third algorithm was proposed by Badoiu et al. [3] with a running time of

$$O(d^{O(1)} n \log^{O(k)} n 2^{O(k/\epsilon)}) .$$

For Euclidean k-means, Har-Peled and Mazumdar [19] provided an $(1+\epsilon)$ approximation algorithm with running time

$$O(n + (\epsilon^{-1})^{2d+1} k^{k+2} \log^{k+1} n \log^{k} \epsilon^{-1}) .$$

A second $(1+\epsilon)$ approximation algorithm is due to Feldman et al. [15], which achieves a running time of

$$O(ndk + d(k\epsilon^{-1})^{O(1)} + 2^{O(k\epsilon^{-1})}) .$$

Kumar et al. [22] provided a simple algorithm based on sampling for Euclidean k-means which gave a $(1+\epsilon)$-approximation in

$$O(dn2^{\mathrm{poly}(k\epsilon^{-1})})$$

time. This was improved by Chen [6] to provide an algorithm which ran in

$$O(ndk + d^2 n^{\sigma} 2^{\mathrm{poly}(k\epsilon^{-1})})$$

time, for any $\sigma > 0$. Kanungo et al. [20] gives a $(9+\epsilon)$-approximation for k-means in time $O(n^3/\epsilon^d)$. For Euclidean k-center, Feder and Greene [13] show that it is NP-Hard to find an approximation-factor better than 1.822 for this problem.

KL-clustering: In this paper we are interested in KL-clustering on the probability simplex. We first note that algorithms that cluster distributions with respect to either ℓ_1 or ℓ_2^2 may give arbitrarily bad solutions for the KL-divergence. The following example shows this for $\mathrm{MTC}_{\mathrm{KL}}$.

Example 1 *Consider the following three distributions:*

$$p = (\frac{1}{2}, \frac{1-\epsilon}{2}, \frac{\epsilon}{2}),\ q = (\frac{1}{2}, \frac{1}{2}, 0),\ r = (\frac{1}{2}+\epsilon, \frac{1}{2}-\epsilon, 0) .$$

We consider the costs of all possible partitions of $\{p, q, r\}$ into two groups.

Clustering	ℓ_2^2-*cost*	ℓ_1-*cost*	KL-*cost*
$\{p,q\},\{r\}$	$\epsilon^2/4$	ϵ	$\epsilon/2 + O(\epsilon^2)$
$\{p\},\{q,r\}$	ϵ^2	2ϵ	$O(\epsilon^2)$
$\{p,r\},\{q\}$	$3\epsilon^2/4$	2ϵ	$\epsilon/2 + O(\epsilon^2)$

Note that the clustering $\{\{p,q\},\{r\}\}$ minimizes the ℓ_2^2 or ℓ_1 cost but that this clustering is a factor $\Omega(1/\epsilon)$ from optimal in terms of $\mathrm{MTC}_{\mathrm{KL}}$. *Since ϵ may be made arbitrarily small, we conclude that clustering the distributions according to either ℓ_2^2 or ℓ_1 can lead to arbitrarily bad solutions.*

There has been previous work on methods for KL-clustering [24, 4, 26, 5, 11]. However, none of these algorithms achieve guaranteed approximations in the worst case. The most directly relevant paper is a recent paper by Ackermann et al. [1]. They present a very nice algorithm that returns a good approximation for $\mathrm{MTC}_{\mathrm{KL}}$ on the assumption that all distributions to be clustered have constant mass on each coordinate, i.e., for some constant γ, $p_i^j \geq \gamma$ for all $j \in [t], i \in [d]$. This implies that $d \leq 1/\gamma$ is also constant and even for distributions with constant dimension, rules out any sparse data where some coordinates will have zero mass. Sparse data is common in many applications. In contrast, the algorithms we present are fully general and require no assumptions on the sparsity or the dimensionality of the input distributions.

1.2 Our Contributions

Our main contribution in this paper is to provide algorithms for clustering in the KL-divergence measure which achieve guaranteed approximations in the worst case. Our specific contributions are the following:

1. *Minimizing Average Distortion:* We provide the first guaranteed approximation algorithm for the problem of minimizing average distortion in the KL-divergence measure, when the input is a set of n arbitrary distributions. To show our result, we first provide constant factor approximation algorithms for the related divergences, Hellinger and Jensen-Shannon. These results exploit the fact that these divergences satisfy a relaxation of the triangle inequality and are closely related to the k-means problem on the sphere. We then show that although the KL-divergence between two distributions can be infinitely larger than the Jensen-Shannon or Hellinger divergence, one can relate the average clustering distortion in terms of the Hellinger cost to the average clustering distortion in terms of the KL-divergence. This yields an $O(\log n)$-approximation algorithm for $\mathrm{MTC}_{\mathrm{KL}}$.

   We note that while a guarantee of $O(\log n)$-factor from optimality is weaker than we would like, this does not preclude the possibility that the algorithm achieves better results in practice. Furthermore, the clustering found could be used as a preprocessing step for a improvement heuristic for which there exist no guarantees. The most important contribution of a $O(\log n)$-factor approximation is to understanding the structure of the problem.

2. *Minimizing Maximum Distortion:* We provide the first guaranteed approximation algorithm for minimizing the maximum distortion, when the input is a set of n arbitrary distributions. To show our result, we relate the maximum clustering distortion in terms of the KL-divergence to the maximum diameter of a cluster measured in terms of the JS-divergence. We then show a constant factor

approximation to the problem of minimizing the JS diameter. This yields an $O(\min(\log n, \log d))$-approximation algorithm for MMC_{KL}.

3. *Hardness Results:* Finally, we provide hardness results for the above problems. First, we show that when we restrict the cluster centers to be in the set of input distributions, no polynomial-time approximation is possible, unless $P \neq NP$. In addition, when the centers are unrestricted, we show a hardness of approximation result for k-center by demonstrating that KL behaves like ℓ_2^2 near the middle of the probability simplex.

Notation: We denote the probability simplex over $\mathbb{R}^d$ as Δ. We write $a = b \pm c$ as short hand for $a \in [b - c, b + c]$.

2 Information Geometry

In this section we review some known results about the geometry of KL and prove some new results. As we mentioned, $\text{KL}(p, q)$ is asymmetric, does not satisfy a directed triangle inequality, and can be infinite even if p and q are on the probability simplex. (It is, however, at least always positive by Gibb's inequality.) Furthermore, KL does not even satisfy a relaxed directed triangle inequality, that is

$$\frac{\text{KL}(p, r) + \text{KL}(r, q)}{\text{KL}(p, q)}$$

can be made arbitrarily small with $p, q, r \in \Delta$.[3] The following example demonstrates this.

Example 2 KL *is not a relaxed metric. Consider*

$$p = (1/2, 1/2), q = (e^{-c}, 1 - e^{-c}), r = (\epsilon, 1 - \epsilon)$$

where $1/2 \geq \epsilon > e^{-c}$. *Then*

$$\begin{aligned} \text{KL}(p, q) &\geq c/2 - \ln 2 \\ \text{KL}(p, r) &\leq (\ln \epsilon^{-1} - \ln 2)/2 \\ \text{KL}(r, q) &\leq \epsilon c - 1 \end{aligned}$$

Hence, by increasing c *and decreasing* ϵ, *the ratio*

$$(\text{KL}(p, r) + \text{KL}(r, q))/\text{KL}(p, q)$$

can be made arbitrarily small.

Two other information divergences that will play an important role in our results are the Hellinger and Jensen-Shannon divergences. These are both divergences from the family of f-divergences [10].

Definition 1 *The Hellinger and Jensen-Shannon divergence between* $p, q \in \Delta$ *are defined as*

$$\begin{aligned} \text{He}(p, q) &= \sum_{i \in [d]} (\sqrt{p_i} - \sqrt{q_i})^2 \\ \text{JS}(p, q) &= \text{KL}(p, \frac{p+q}{2}) + \text{KL}(q, \frac{p+q}{2}) \, . \end{aligned}$$

[3] We note that this ratio can be bounded below for some families of distributions in terms of the ratio of eigenvalues of a related Hessian matrix [9].

Both $\text{JS}(p, q)$ and $\text{He}(p, q)$ are symmetric and bounded: it can easily be shown that $\text{JS}(p, q) \leq 2$ and $\text{He}(p, q) \leq 2$ for all $p, q \in \Delta$. Note that since $\text{KL}(p, q)$ may be infinite this rules out any multiplicative relationship in general.

Relationships between $\text{JS}(p, q)$ and $\text{He}(p, q)$ are given in the next lemma [18, 28].

Lemma 2 *For all distributions* p *and* q,

$$\text{He}(p, q)/2 \leq \text{JS}(p, q) \leq 2 \ln(2) \, \text{He}(p, q) \, . \quad (1)$$

Unfortunately, neither JS or He are metrics but we can show that they are "almost metrics" in that they satisfy non-negativity, identity of indiscernibles, symmetry, and a relaxation of the triangle inequality. We say that a measure D satisfies an α-relaxed triangle inequality if for all $p, q, r \in \Delta$,

$$D(p, r) + D(r, q) \geq D(p, q)/\alpha.$$

(When $\alpha = 1$, this is the usual triangle inequality.)

Lemma 3 He *and* JS *obey the* 2*-relaxed triangle equality.*

Proof: We note that He and JS are both the square of metrics: this is obvious for He and the result for JS was proved in [12]. Therefore, for all $p, q, r \in \Delta$,

$$\sqrt{\text{He}(p, q)} + \sqrt{\text{He}(q, r)} \geq \sqrt{\text{He}(p, r)}$$

and hence

$$\text{He}(p, q) + \text{He}(q, r) + 2\sqrt{\text{He}(p, q)\text{He}(q, r)} \geq \text{He}(p, r) \, .$$

By an application of the AM-GM inequality we deduce:

$$2(\text{He}(p, q) + \text{He}(q, r)) \geq \text{He}(p, r)$$

as required. The result for JS follows similarly. ■

The next lemma is a well-known identity (see, e.g., [8]) that relates the KL and JS divergence.

Lemma 4 *For all* $p, q, c \in \Delta$:

$$\text{KL}(p, c) + \text{KL}(q, c) = \text{JS}(p, q) + 2\text{KL}((p + q)/2, c) \, .$$

This is referred to as the parallelogram property.

Another useful property that we will exploit is that the He-balls are convex.

Lemma 5 $B_\ell(p) = \{p' : \text{He}(p, p') \leq \ell\}$ *is convex for all* $\ell \geq 0$ *and* $p \in \Delta$. *Furthermore, for all* $p, q, r \in \Delta$ *and* $\alpha \in (0, 1)$,

$$\text{He}(p, \alpha q + (1 - \alpha) r) \leq \alpha \text{He}(p, q) + (1 - \alpha)\text{He}(p, r) \, .$$

Proof: Consider any ball $B_\ell(p) = \{p' : \text{He}(p, p') \leq \ell\}$ and let $q, s \in B_\ell(p)$ and $\alpha \in (0, 1)$. Then it suffices to show that $\alpha q + (1 - \alpha) r \in B_\ell(p)$. Let $\beta = 1 - \alpha$. Note that

$$\begin{aligned} & \frac{\alpha(\sqrt{p_i} - \sqrt{q_i})^2 + \beta(\sqrt{p_i} - \sqrt{r_i})^2}{(\sqrt{p_i} - \sqrt{\alpha q_i + \beta r_i})^2} \geq 1 \\ \Leftrightarrow \quad & \alpha\sqrt{q_i} + \beta\sqrt{r_i} \leq \sqrt{\alpha q_i + \beta r_i} \\ \Leftrightarrow \quad & \alpha^2 q_i + \beta^2 r_i + 2\alpha\beta\sqrt{q_i r_i} \leq \alpha q_i + \beta r_i \\ \Leftrightarrow \quad & 2\alpha\beta\sqrt{q_i r_i} \leq \alpha\beta q_i + \alpha\beta r_i \\ \Leftrightarrow \quad & 2\sqrt{q_i r_i} \leq q_i + r_i \end{aligned}$$

and this is true by the AM-GM inequality. ■

Properties of Cluster Centers: For the remaining result of this section we need to introduce some further notation. For any measure $D : \Delta \times \Delta \to \mathbb{R}^+$:

$$\text{SumCost}_D(p^1, \ldots, p^t; c) = \sum_{j \in [t]} D(p^j, c)$$

$$\text{SumCost}_D(p^1, \ldots, p^t) = \min_{c \in \Delta} \text{SumCost}_D(p^1, \ldots, p^t; c)$$

$$\text{MaxCost}_D(p^1, \ldots, p^t; c) = \max_{j \in [t]} D(p^j, c)$$

$$\text{MaxCost}_D(p^1, \ldots, p^t) = \min_{c \in \Delta} \text{MaxCost}_D(p^1, \ldots, p^t; c)$$

We denote the centroid of a set of distributions as

$$\text{cent}(p^1, \ldots, p^t) = t^{-1} \sum p^i \ .$$

The next lemma (a special case of more general result for all Bregman divergences [5]) shows that the center that minimizes the average ℓ_2^2 or KL distortion is the centroid of the distributions being clustered.

Lemma 6 *For any distributions $p^1, \ldots, p^t$,*

$$\text{cent}(p^1, \ldots, p^t) = \text{argmin}_{q \in \Delta} \text{SumCost}_{\ell_2^2}(p^1, \ldots, p^t; q)$$
$$= \text{argmin}_{q \in \Delta} \text{SumCost}_{\text{KL}}(p^1, \ldots, p^t; q),$$

i.e., the cluster centers for ℓ_2^2 and KL *are at centroids.*

The next lemma shows that when we are clustering distributions near the middle of the probability simplex, the centers that minimize either the maximum or average KL distortion also lie near the middle of the probability simplex. Define,

$$A(r) = \{p \in \Delta : p_j = \frac{1}{d} \pm r \text{ for all } j \in [d]\} \ . \quad (2)$$

Lemma 7 *Let $p^1, \ldots, p^t \in A(\epsilon/d)$ and $0 < \epsilon < 1/10$. Then,*

$$\text{argmin}_{c \in \Delta} \text{SumCost}_{\text{KL}}(p^1, \ldots, p^t; c) \in A(\epsilon/d) \ .$$

If

$$\frac{\text{MaxCost}_{\text{KL}}(p^1, \ldots, p^t; c)}{\text{MaxCost}_{\text{KL}}(p^1, \ldots, p^t)} \leq 10$$

then $c \in A(10\sqrt{\epsilon})$.

Proof: The first claim follows from Lemma 6 and the fact that $\text{cent}(p^1, \ldots, p^t)$ is a convex combination of $p^1, \ldots, p^t$. For the second claim note that for $i \in [t]$,

$$\text{KL}(p^i; p^1) \leq \ln \frac{1+\epsilon}{1-\epsilon} \leq 3\epsilon,$$

and hence $\text{MaxCost}_{\text{KL}}(p^1, \ldots, p^t) \leq 3\epsilon$. Consider $q \notin A(10\sqrt{\epsilon})$. Then

$$\text{KL}(p^i; q) \geq \ell_1^2(p^i; q) \geq (10\sqrt{\epsilon} - \epsilon/d)^2 > 30\epsilon \ ,$$

where the first inequality follows by Pinsker's inequality. Hence q does not satisfy,

$$\text{MaxCost}_{\text{KL}}(p^1, \ldots, p^t; q) \leq 10 \cdot \text{MaxCost}_{\text{KL}}(p^1, \ldots, p^t) \ .$$

■

3 Minimizing Average Distortion

In this section, we address the problem of computing a clustering of the input distributions which approximately minimizes the average Kullback-Liebler divergence between an input distribution and the center of the cluster it belongs to. We provide an algorithm that computes a clustering in which the average KL-divergence between an input distribution, and the center of the cluster it belongs to is at most $O(\log n)$ times the optimal cost. The main theorem in this section is the following.

Theorem 8 *There exists a polynomial time $O(\log n)$-approximation algorithm for* MTC_{KL}.

The main idea behind our algorithm is the observation that even though, in general, the KL-divergence between two distributions can be infinitely larger than the He-divergence between them, a clustering of the input distributions with low average distortion according to the He-divergence also has low average distortion by the KL-divergence. Therefore, our analysis proceeds in two steps. First, we show in Section 3.1 how to compute a clustering that approximately (within a factor of $2 + \epsilon$) minimizes the average Hellinger divergence between input distributions and the closest cluster center. Then, we show in Section 3.2 how this leads to a clustering with low average distortion in the KL-divergence.

3.1 Hellinger Clustering

In this section we present an algorithm for minimizing average He distortion. The main idea behind our algorithm is the simple observation that the Hellinger distance between two distributions p and q is the square of the Euclidean distance between the points $\sqrt{p}$ and $\sqrt{q}$ where $\sqrt{p}$ is a shorthand for the vector in the positive quadrant of the unit sphere:

$$\sqrt{p} = (\sqrt{p_1}, \ldots, \sqrt{p_d}) \ .$$

Therefore, mapping each point p^i to $\sqrt{p^i}$ and then computing a clustering that minimizes the average ℓ_2^2 measure between each transformed point and the center of the cluster it belongs to, should give us a good clustering for minimizing average Hellinger distortion. However, there will be a slight issue that arises because we insist that the cluster centers lie on the probability simplex.

Before we address the issue, we present the algorithm:

1. For each input distribution $i \in [n]$, compute $\sqrt{p^i}$
2. Compute a $(1+\epsilon)$-approximation to
$$\text{MTC}_{\ell_2^2}(\sqrt{p^1}, \ldots, \sqrt{p^n}) \ ,$$
using any $(1+\epsilon)$-approximation algorithm for k-means. Let the cluster centers be $\tilde{c}^1, \ldots, \tilde{c}^k$. Note that in general $\tilde{c}^j$ is not on the unit sphere.
3. Let $\{p^{j_1}, \ldots, p^{j_t}\}$ be the set of input distribution whose closest cluster center is $\tilde{c}^j$. Let the final center for this cluster be $\text{cent}(p^{j_1}, \ldots, p^{j_t})$.

The issue we need to address is that the cluster center c that minimizes $\text{SumCost}_{\text{He}}(p^1, \ldots, p^t; c)$ need not lie on Δ: this can be seen as a consequence of the fact that $\tilde{c}^j$ is not on the unit sphere in general. Thus the actual average Hellinger divergence for the same clustering may be much higher than the k-means cost of the transformed points. However, the following lemma establishes that setting $c = \text{cent}(p^1, \ldots, p^t)$ (which necessary lies on Δ) produces a clustering whose average Hellinger distortion is at most a factor 2 away from the k-means cost of the transformed points.

Before we state the lemma, we define some notation. For a vector $p = (p_1, \ldots, p_d)$ over d dimensions, we use p^2 to denote the vector

$$p^2 = (p_1^2, \ldots, p_d^2)$$

Lemma 9 *For $p^1, \ldots, p^t \in \Delta$, for $i \in [d]$, define*

$$a_i = \sum_{j \in [t]} p_i^j \text{ and } b_i = \sum_{j \in [t]} \sqrt{p_i^j} \ .$$

and let $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$.

$$\begin{aligned} & \text{SumCost}_{\text{He}}(p^1, \ldots, p^t; a/t) \\ \leq \ & 2\text{SumCost}_{\text{He}}(p^1, \ldots, p^t; (b/t)^2) \end{aligned}$$

Proof:

$$\begin{aligned} \sum_j (\sqrt{p_i^j} - \sqrt{a_i/t})^2 &= a_i + \sum_j a_i/t - 2\sqrt{a_i/t} b_i \\ &= 2a_i - 2t^{-1/2}\sqrt{a_i} b_i \end{aligned}$$

and

$$2\sum_j (\sqrt{p_i^j} - b_i/t)^2 \leq 2a_i - 2t^{-1} b_i^2 \ .$$

Therefore it suffices to show that $b_i \leq t^{1/2}\sqrt{a_i}$ but this follows because

$$b_i^2 = a_i + \sum_{j \neq k} \sqrt{p_i^j p_i^k} \leq a_i + (t-1)a_i \ .$$

where the inequality follows by AM-GM inequality. ■

Theorem 10 *There exists a polynomial-time $(2+\epsilon)$-approximation algorithm for MTC_{He}.*

Proof: The result for MTC_{He} is achieved as described above: first we map each distribution from the probability simplex to the positive quadrant of the unit sphere:

$$\begin{aligned} f : \Delta &\rightarrow \{x \in \mathbb{R}^d : \ell_2(x) = 1, x_i \geq 1\} \\ (p_1, \ldots, p_d) &\mapsto (\sqrt{p_1}, \ldots, \sqrt{p_d}) \ . \end{aligned}$$

We then run an algorithm for $\text{MTC}_{\ell_2^2}$. For each cluster formed, return the centroid of the original probability distributions. This is clearly a probability distribution. The cost of using this center rather than the center of mass of the probability distributions once mapped to the sphere is a factor 2 as shown in Lemma 9. ■

We conclude this section by noting that our algorithm also leads to a good clustering for minimizing average distortion according to the Jensen-Shannon measure using Eq. 1.

Lemma 11 *There exists a polynomial-time $(8 \ln 2 + \epsilon)$-approximation algorithm for MTC_{JS}.*

3.2 Kullback-Leibler Clustering

The following lemma relates $\text{SumCost}_{\text{KL}}(p^1, \ldots, p^t)$ and $\text{SumCost}_{\text{He}}(p^1, \ldots, p^t)$. We note that that a later result in Section 4 could be used (in conjunction with Lemma 2) to achieve a result with that shows the ratio scales with $\lg t$ in the worst case. However, the following proof establishes better constants and has the benefit that the proof is more geometric.

Lemma 12 *For any distributions $p^1, \ldots, p^t$,*

$$1/2 \leq \frac{\text{SumCost}_{\text{KL}}(p^1, \ldots, p^t)}{\text{SumCost}_{\text{He}}(p^1, \ldots, p^t)} \leq \lceil \lg t \rceil \, (\ln 16).$$

Proof: The first inequality follows because for $p, q \in \Delta$, $\text{JS}(p, q) = \min_{c \in \Delta}(\text{KL}(p, c) + \text{KL}(q, c)) \leq \text{KL}(p, q)$ (this follows from e.g., Lemma 6) and Eq. 1.

We now prove the second inequality. Without loss of generality assume that t is a power of 2 (otherwise consider adding $(2^{\lceil \log t \rceil} - t)$ new points at He center of $p^1, \ldots, p^t$ – this can only increases the middle term of the equation.)

Consider a balanced binary tree on the nodes of the cluster. For an internal node at height j, associate a multi-set of distributions $S(v)$ consisting of 2^j copies $p(u)$, the center of mass of the 2^j distributions at the leaves of the subtree rooted at v. Let S_j be the set of distributions at height j. Note that $S_0 = \{p^1, \ldots, p^t\}$.

The lemma follows from the next three claims.

Claim 13 *For all j, $\text{SumCost}_{\text{He}}(S_j) \leq \text{SumCost}_{\text{He}}(S_0)$.*

Proof: Let c be an arbitrary distribution. By Lemma 5,

$$2^j \text{He}(p, c) + 2^j \text{He}(q, c) \geq 2^{j+1} \text{He}((p+q)/2, c)$$

and therefore $\text{SumCost}_{\text{He}}(S_j; c)$ decreases as j increases and the result follows. ■

Claim 14 *For all j,*

$$\begin{aligned} \sum_{v : \text{height}(v) = j+1} & \text{SumCost}_{\text{KL}}(\cup_{u : u \in \text{ch}(v)} S(u)) \\ & \leq (\ln 16) \text{SumCost}_{\text{He}}(S_j) \ . \end{aligned}$$

where $\text{ch}(v)$ denotes the children of v.

Proof: Let u and w be the children of a node v at height $j+1$. Let $c = (p(u) + p(w))/2$.Then,

$$\begin{aligned} & \text{SumCost}_{\text{KL}}(S(u), S(w)) \\ = \; & 2^j \text{JS}(p(u), p(w)) \\ \leq \; & 2^{j+1}(\ln 2)\text{He}(p(u), p(w)) \\ \leq \; & 2^{j+2}(\ln 2)(\text{He}(p(u), c) + \text{He}(p(w), c)) \\ \leq \; & (\ln 16)\text{SumCost}_{\text{He}}(S(u), S(w)) \end{aligned}$$

■

Claim 15

$$\sum_j \sum_{v:\text{height}(v)=j+1} \text{SumCost}_{\text{KL}}(\cup_{u:u\in\text{ch}(v)} S(u)) = \text{SumCost}_{\text{KL}}(p^1, \ldots, p^t) \ .$$

Proof: Let v be at height $j+1$. Let v have children u and w and grandchildren u_1, u_2, w_1, w_2. Then the result follows because

$$\begin{aligned} & \text{SumCost}_{\text{KL}}(S(u_1), S(u_2)) \\ & \quad + \text{SumCost}_{\text{KL}}(S(w_1), S(w_2)) \\ & \quad + \text{SumCost}_{\text{KL}}(S(u), S(w)) \\ = \; & 2^{j-1}(\text{KL}(p(u_1), p(u)) + \text{KL}(p(u_2), p(u)) \\ & \quad + \text{KL}(p(w_1), p(w)) + \text{KL}(p(w_2), p(w)) \\ & \quad + 2\text{KL}(p(u), p(v)) + 2\text{KL}(p(w), p(v))) \\ = \; & 2^{j-1}(\text{KL}(p(u_1), p(v)) + \text{KL}(p(u_2), p(v)) \\ & \quad + \text{KL}(p(w_1), p(v)) + \text{KL}(p(w_2), p(v))) \\ = \; & \text{SumCost}_{\text{KL}}(S(u_1), S(u_2), S(w_1), S(w_2)) \end{aligned}$$

where the second inequality follows from the parallelogram property and the fact that $p(u) = (p(u_1)+p(u_2))/2$ and $p(w) = (p(w_1) + p(w_2))/2$. ■

■

We next show that the above lemma is nearly tight.

Lemma 16 *There exists $(p^i)_{i\in[t]}$ on $d \geq t$ coordinates such that,*

$$\frac{\text{SumCost}_{\text{KL}}(p^1, \ldots, p^t)}{\text{SumCost}_{\text{He}}(p^1, \ldots, p^t)} = \Omega(\log t) \ .$$

Proof: Let $(p^i)_{i\in[t]}$ be t distributions where p^i takes value i with probability 1. Then

$$\text{SumCost}_{\text{KL}}(p^1, \ldots, p^t) = t \ln t$$

whereas

$$\begin{aligned} \text{SumCost}_{\text{He}}(p^1, \ldots, p^t; c) &= t\left((1 - \frac{1}{\sqrt{t}})^2 + \frac{t-1}{t}\right) \\ &= 2t - 2\sqrt{t} \ , \end{aligned}$$

where $c = t^{-1}\sum_i p^i$. Then appeal to Lemma 9. ■

Then the proof of Theorem 8 follows immediately from Lemma 12 and Theorem 10.

4 Minimizing Maximum Distortion

In this section, we provide an algorithm for clustering the input distributions such that the maximum Kullback-Liebler divergence between an input distribution and the center of the cluster it belongs to is approximately minimized. In particular, our algorithm produces a clustering in which the maximum KL-divergence between an input distribution, and the closest center is at most a $\min(O(\log d), O(\log n))$ factor greater than optimal.

Our algorithm is pleasantly simple: we use a variant of Gonzalez's algorithm [16] to cluster the input distributions such that the Jensen-Shannon divergence between any two points in the same cluster is minimized. We then show that although the KL-divergence between two distributions can be infinitely larger than their JS-divergence, this procedure still produces a good clustering according to the KL-divergence. The main theorem in this section can be stated as follows.

Theorem 17 *There exists a polynomial-time*

$$O(\min(\log d, \log n))$$

approximation for MMC_{KL}.

Before proving the theorem, we show a lemma which establishes a general relationship between the KL-divergence and JS-divergence between two distributions, when the ratio of probability masses that the two distributions assign to any coordinate is bounded. This lemma may be of independent interest.

Lemma 18 *Let $p, q \in \Delta$ such that, for all i, $p_i/q_i \leq t$, where $t \geq e^2$. Then,*

$$\text{KL}(p, q) \leq \frac{2\ln t}{\ln(6/5)}\text{JS}(p, q)$$

Proof: For each i, let $\delta_i = (p_i - q_i)/q_i$ so that $p_i = (1+\delta_i)q_i$. Then, $\sum_i \delta_i q_i = \sum_i p_i - q_i = 0$,

$$\text{KL}(p, q) = \sum_i q_i(1+\delta_i)\ln(1+\delta_i) \ ,$$

and

$$\text{JS}(p, q) = \sum_i q_i((1+\delta_i)\ln(1+\delta_i) - (2+\delta_i)\ln(1+\frac{\delta_i}{2}))$$

Since $p_i/q_i \leq t$, and $\delta_i \leq t-1$, from Lemma 19,

$$\text{KL}(p, q) \leq \Lambda \cdot \text{JS}(p, q) + \sum_i \delta_i q_i$$

where $\Lambda = \frac{2\ln t}{\ln 6/5}$. The lemma follows from the fact that $\sum_i \delta_i q_i = 0$ and $t \geq 4$. ■

Lemma 19 *For any $x \in [-1, 2]$,*

$$\begin{aligned} (1+x)\ln(1+x) \leq 4((1+x)\ln(1+x) \\ -2(1+x/2)\ln(1+x/2)) + x \ . \end{aligned}$$

For any $x \in (2, x^*]$,

$$(1+x)\ln(1+x) \leq \frac{2\ln x^*}{\ln(6/5)}((1+x)\ln(1+x) \\ -2(1+x/2)\ln(1+x/2)) + x \ .$$

Proof: Let Λ be a parameter and let

$$Y(x) = (1+x)\ln(1+x) - \Lambda((1+x)\ln(1+x) \\ -2(1+x/2)\ln(1+x/2)) - x \ .$$

Our goal is to show that $Y(x) \leq 0$ for suitable values of the parameter Λ. The first and second order derivatives of Y can be written as follows:

$$Y'(x) = \Lambda \ln(1+x/2) - (\Lambda - 1)\ln(1+x)$$

and

$$Y''(x) = \frac{2 - \Lambda + x}{(1+x)(2+x)} \ .$$

We first consider $x \in [-1, 2)$ and $\Lambda = 4$. If $x < 2$, then $Y''(x) < 0$. Therefore, $Y'(x)$ is strictly decreasing in the range $[-1, 2)$. We note that $Y'(-1) = \infty$ and $Y'(0) = 0$; therefore Y is a strictly increasing function from $[-1, 0)$ and strictly decreasing from $(0, 2]$. As $Y(0) = 0$, $Y(x) < 0$ for $x < 0$ and $Y(x) < 0$ for $x > 0$, and the first part of the lemma follows.

To prove the second part, we write the derivative $Y'(x)$ as follows:

$$Y'(x) = \Lambda \cdot \ln \frac{1 + x/2}{1+x} + \ln(1+x)$$

If $x > 2$, then $\ln \frac{1+x/2}{1+x} < \ln(5/6)$. By plugging in $\Lambda = \frac{2\ln x^*}{\ln 6/5}$,

$$Y'(x) < -2\ln x^* + \ln(1+x) < 0$$

for x in $(2, x^*]$, which means that Y is strictly decreasing in this interval. As $t \geq e^2$, here $\Lambda \geq 4$. The previous part of the lemma implies that $Y(2) < 0$, for any $\Lambda > 4$, and hence the lemma follows. ■

Lemma 20 *Consider t distributions $p^1, \ldots, p^t$ such that $\mathrm{He}(p^i, p^j) \leq r$ for all $i, j \in [t]$. Then $\mathrm{He}(p^i, c) \leq r$ for all $i \in [t]$ where c is any convex combination of $p^1, \ldots, p^t$.*

Proof: The result follows by Lemma 5: Consider distribution p^i and the set of distributions in $B_r(p^i) = \{q : \mathrm{He}(p^i, q) \leq r\}$. By Lemma 5, $B_r(p^i)$ is convex. Since $p^j \in B_r(p^i)$ for all $j \in [t]$ and c is a convex combination of $\{p^j\}_{j\in[t]}$ we deduce that $c \in B_r(p^i)$. Hence $\mathrm{He}(p^i, c) \leq r$ as required. Since i was arbitrary the result follows. ■

Lemma 21 *Let $p^1, \ldots, p^t$ be t distributions over $[d]$ and let $c = \mathrm{cent}(p^1, \ldots, p^t)$. Then,*

$$\frac{\mathrm{MaxCost}_{\mathrm{KL}}(p^1, \ldots, p^t; c)}{\max_{i,j} \mathrm{JS}(p^i, p^j)} \leq O(\log t) \ .$$

Moreover, there exists some c^* which is a convex combination of $p^1, \ldots, p^t$ such that:

$$\frac{\mathrm{MaxCost}_{\mathrm{KL}}(p^1, \ldots, p^t; c^*)}{\max_{i,j} \mathrm{JS}(p^i, p^j)} \leq O(\log d) \ .$$

Proof: To show the first inequality, we observe that for any $i \in [d], j \in [t]$: $p_i^j / c_i \leq t$. Using this fact along with Lemma 18, we conclude that:

$$\mathrm{MaxCost}_{\mathrm{KL}}(p^1, \ldots, p^t; c) \leq O(\log t) \cdot \max_i \mathrm{JS}(p^i, c)$$

The rest of the inequality follows from the fact that JS is constant factor related to He (Lemma 2), followed by an application of Lemma 20.

To show the second inequality, let

$$q^1, \ldots, q^d \subset \{p^1, \ldots, p^t\}$$

be distributions such that for any $i \in [d]$ and any $j \in [n]$, $q_i^i \geq p_i^j$. We define

$$c^* = \mathrm{cent}(q^1 + \ldots + q^d) \ .$$

Observe that for any $i \in [d], j \in [t]$: $p_i^j / c_i^* \leq d$. Therefore, from Lemma 18, for any i,

$$\mathrm{MaxCost}_{\mathrm{KL}}(p^1, \ldots, p^t; c^*) \leq O(\log d)\mathrm{JS}(p^i, c^*)$$

From Lemma 2, $\mathrm{JS}(p^i, c^*) \leq O(\mathrm{He}(p^i, c^*))$. As c^* is a convex combination of $p^1, \ldots, p^t$, the rest of the lemma follows from an application of Lemma 20. ■

Lemma 22 *Let $p^1, \ldots, p^t$ be t distributions over $[d]$.*

$$\frac{1}{2} \leq \frac{\mathrm{MaxCost}_{\mathrm{KL}}(p^1, \ldots, p^t)}{\max_{i,j} \mathrm{JS}(p^i, p^j)}$$

Proof: Let $(i, j) = \mathrm{argmax}\, \mathrm{JS}(p^i, p^j)$. Note that since we allow unrestricted centers,

$$\mathrm{MaxCost}_{\mathrm{KL}}(p^i, p^j) \leq \mathrm{MaxCost}_{\mathrm{KL}}(p^1, \ldots, p^t; c) \ ,$$

and let q minimize $\max\{\mathrm{KL}(p^i, q), \mathrm{KL}(p^j, q)\}$. But

$$\begin{aligned} 2\max\{\mathrm{KL}(p^i, q), \mathrm{KL}(p^j, q)\} &\geq \mathrm{KL}(p^i, q) + \mathrm{KL}(p^j, q) \\ &\geq \min_q \mathrm{KL}(p^i, q) + \mathrm{KL}(p^j, q) \\ &= \mathrm{JS}(p^i, p^j) \ . \end{aligned}$$

from which the Lemma follows. ■

We now are in a position to complete the proof of Theorem 17.

Proof: From Lemmas 21 and 22, and the fact that for any c,

$$\mathrm{MaxCost}_{\mathrm{KL}}(p^1, \ldots, p^t; c) \geq \mathrm{MaxCost}_{\mathrm{KL}}(p^1, \ldots, p^t)$$

(by definition), we know that if we α-approximate the problem of minimizing the maximum JS diameter of a cluster, we get a $\min(O(\alpha \log d), O(\alpha \log n))$ approximation for k-center KL clustering. In the rest of the proof we show that we may assume $\alpha = 4$.

We use a variant of an algorithm by Gonzalez [16] that is applicable to divergences that satisfy a relaxed triangle inequality. Recall that JS satisfies,

$$\mathrm{JS}(p,q) + \mathrm{JS}(q,r) \geq \mathrm{JS}(p,r)/2 \ .$$

for all p, q, r. The algorithm assumes knowledge of the optimum JS diameter (note that there are at most n^2 possible values and thus we can check them all); let this value be D. Initially, let all p^j be "unmarked." The algorithm proceeds by picking an arbitrary unmarked distribution p^i, marking all p^j such that $\mathrm{JS}(p^j, p^i) \leq D$ and repeating until all distributions are marked. Define the each cluster as the set of distributions marked in the same iteration and call p^i the "center" of this cluster. This results in a clustering such that the maximum diameter is at most $2(D + D) = 4D$. We need to show that the process does not determine more than k centers. Suppose we pick $k + 1$ centers. Note that each of these centers are strictly greater than $(D + D)/2 = D$ apart and hence no two may be in the same cluster for the optimum clustering. This is a contradiction. ■

5 Hardness Results

In this final section of our paper, we prove hardness of approximation results for $\mathrm{MMC}_{\mathrm{KL}}$ and $\mathrm{MTC}_{\mathrm{KL}}$, i.e., we show lower bounds of the approximation factors possible in polynomial time on the assumption that $P \neq NP$. We consider two variants of these problems. For all the algorithms we presented in the previous sections, we insisted that the centers $c^1, \ldots, c^k$ lay in Δ but other than this the centers were *unrestricted*. In some of the previous work on approximation algorithms for clustering a variant is considered in which it is required that $c^1, \ldots, c^k$ are chosen from among the input distributions $\{p^1, \ldots, p^n\}$. We call this the *restricted center* version.

When a metric is used rather than KL, the restricted and unrestricted versions of the problems are closely related: it can be shown that the restriction can at most double the clustering cost. However, for KL we show that, while we have presented approximation algorithms for the unrestricted case, no approximation to any multiplicative factor is possible in the restricted case.

5.1 Unrestricted Centers

In this section, we prove an approximation hardness result for $\mathrm{MMC}_{\mathrm{KL}}$. Our result is based on demonstrating that near the center of the simplex KL behaves similarly to ℓ_2^2. We then use a result by Feder and Greene [13] that showed an approximation hardness of 1.822 for k-center in the plane where distance are measured as ℓ_2. (Hence, this gives a $1.822^2 < 3.320$ approximation hardness result for ℓ_2^2.)

Recall the definition,

$$A(r) = \{p \in \Delta : p_j = 1/d \pm r \text{ for all } j \in [d]\} \ . \quad (3)$$

Lemma 23 *For $p, q \in A(\epsilon/d)$,*

$$\mathrm{KL}(p,q) = (1 \pm 5\epsilon) d\ell_2^2(p,q) \ .$$

Proof: We apply Taylor's Theorem to the terms of the KL divergence:

$$\begin{aligned}
\mathrm{KL}(p,q) &= \sum_{i\in[d]} p_i \log \frac{p_i}{q_i} - p_i + q_i \\
&= \sum_{i\in[d]} -p_i \log\left(1 - \frac{p_i - q_i}{p_i}\right) - p_i + q_i \\
&= \sum_{i\in[d]} \frac{(p_i - q_i)^2}{p_i} + \eta_i^3 p_i
\end{aligned}$$

for some η_i with $|\eta_i| \leq |p_i - q_i|/p_i$. Note that

$$|\eta_i|^3 p_i \leq \frac{(p_i - q_i)^2}{p_i} \cdot \frac{|p_i - q_i|}{p_i} \leq 3\epsilon \frac{(p_i - q_i)^2}{p_i}$$

and therefore

$$\mathrm{KL}(p,q) = (1 \pm 3\epsilon) \sum_{i\in[d]} \frac{(p_i - q_i)^2}{p_i} \leq (1 \pm 5\epsilon) d\ell_2^2(p-q) \ .$$

■

Using the above lemma, the next theorem shows that if the distributions to be clustered are near the center of the simplex, then we can use an approximation algorithm for $\mathrm{MTC}_{\mathrm{KL}}$ or $\mathrm{MMC}_{\mathrm{KL}}$ to get a good approximation for $\mathrm{MTC}_{\ell_2^2}$ or $\mathrm{MMC}_{\ell_2^2}$ respectively.

Theorem 24 *Let $\tau \in (1, 10)$ and let*

$$p^1, \ldots, p^n \in A(\epsilon^2/(50^2 d^3)) \ .$$

Then, a τ-approximation for $\mathrm{MTC}_{\mathrm{KL}}$ yields a $(\tau + 5\epsilon)$-approximation for $\mathrm{MTC}_{\ell_2^2}$. Similarly, a τ-approximation for $\mathrm{MMC}_{\mathrm{KL}}$ yields a $(\tau + 5\epsilon)$-approximation for $\mathrm{MMC}_{\ell_2^2}$.

Proof: We first consider $\mathrm{MTC}_{\mathrm{KL}}$. Suppose we want to solve $\mathrm{MTC}_{\ell_2^2}$ on the input $p^1, \ldots, p^n \in \Delta$ and let

$$\{\tilde{c}^1, \ldots, \tilde{c}^k\}$$

be a τ-approximation for $\mathrm{MTC}_{\mathrm{KL}}$. Without loss of generality, we may assume that $\tilde{c}^1, \ldots, \tilde{c}^k$ are in the convex hull of $p^1, \ldots, p^n$ since if $\tilde{c}^j$ is the closest center to $\{p^i\}_{i\in I}$ then the objective function only decreases if we let $\tilde{c}^j = \mathrm{cent}(p^i : i \in I)$.

Denote the convex hull of $p^1, \ldots, p^n$ by H and note that $q \in H$ implies that $q \in A(\epsilon^2/(50^2 d^3)) \subset A(\epsilon/(5d))$. Hence, by appealing to Lemmas 7 and 23, we deduce,

$$\begin{aligned}
&\sum_{j\in[n]} \min_{i\in[k]} \ell_2^2(p^j, \bar{c}^i) \\
=\ & \frac{1}{d(1\pm\epsilon)^2} \sum_{j\in[n]} \min_{i\in[k]} \mathrm{KL}(p^j, \bar{c}^i) \\
\leq\ & \frac{\tau}{d(1\pm\epsilon)^2} \min_{c^1,\ldots,c^k\in H} \sum_{j\in[n]} \min_{i\in[k]} \mathrm{KL}(p^j, c^i) \\
\leq\ & \frac{\tau}{(1\pm\epsilon)^4} \min_{c^1,\ldots,c^k\in H} \sum_{j\in[n]} \min_{i\in[k]} \ell_2^2(p^j, c^i) \\
=\ & \frac{\tau}{(1\pm\epsilon)^4} \min_{c^1,\ldots,c^k\in \Delta} \sum_{j\in[n]} \min_{i\in[k]} \ell_2^2(p^j, c^i) \ .
\end{aligned}$$

where the last line follows because the optimum centers for $\text{MTC}_{\ell_2^2}$ lie in $\text{convex}(P)$.

We now consider MMC_{KL} and suppose $\{\tilde{c}^1, \ldots, \tilde{c}^k\}$ is a τ-approximation for MMC_{KL}. By appealing to Lemma 7, we may assume

$$\tilde{c}^1, \ldots, \tilde{c}^k \in A(10\sqrt{\epsilon^2/(50^2 d^2)}) = A(\epsilon/(5d)) \; .$$

Hence,

$$\begin{aligned}
& \max_{j\in[n]} \min_{i\in[k]} \ell_2^2(p^j, \tilde{c}^i) \\
= \; & \frac{1}{d(1\pm\epsilon)^2} \max_{j\in[n]} \min_{i\in[k]} \text{KL}(p^j, \tilde{c}^i) \\
\leq \; & \frac{\tau}{d(1\pm\epsilon)^2} \min_{c^1,\ldots,c^k \in A(\frac{\epsilon}{5d})} \left(\max_{j\in[n]} \min_{i\in[k]} \text{KL}(p^j, c^i) \right) \\
\leq \; & \frac{\tau}{(1\pm\epsilon)^4} \min_{c^1,\ldots,c^k \in A(\frac{\epsilon}{5d})} \left(\max_{j\in[n]} \min_{i\in[k]} \ell_2^2(p^j, c^i) \right) \\
= \; & \frac{\tau}{(1\pm\epsilon)^4} \min_{c^1,\ldots,c^k \in \Delta} \left(\max_{j\in[n]} \min_{i\in[k]} \ell_2^2(p^j, c^i) \right) \; .
\end{aligned}$$

where the last line follows because the optimum centers for $\text{MMC}_{\ell_2^2}$ lie in $A(\epsilon/(5d))$. This can be shown using ideas contained in Lemma 7:

$$\ell_2^2(p^i, p^1) \leq d \max_{j\in[d]} (p_j^i - p_j^1)^2 \leq 4\epsilon^4/(50^4 d^5)$$

while for $q \notin A(\epsilon/(5d))$,

$$\ell_2^2(p^i, q) \geq (\epsilon^2/(5d)^2 - \epsilon^2/(50^2 d^3))^2 \geq 4\epsilon^4/(50^4 d^5) \; .$$

■

To show a hardness result for unrestricted centers it is therefore sufficient to show a hardness result for $\text{MMC}_{\ell_2^2}$ when the points to be clustered lie near the middle of the probability simplex. We do this by taking Feder and Greene [13] result the showed the hardness of MMC_{ℓ_2} in the plane and demonstrating that the plane can be mapped into the middle of the probability simplex in a manner that preserves approximation factors. This will give the following theorem.

Theorem 25 *For any $\alpha < 3.320$, unless $P = NP$, no polynomial-time, α-approximation algorithm exists for* MMC_{KL}.

k-means on the middle of Δ: Given an instance I of MMC_{ℓ_2} on a bounded domain A of the $x_1 - x_2$ plane, we show how to produce an instance I' of MMC_{ℓ_2} on the three-dimensional simplex, such that, there is an approximation preserving bijection between the solutions to I and the solutions to I'.

To show this, we first assume without loss of generality that $A \subseteq [0, 1/4] \times [0, 1/4]$. We can assume this because translating and scaling scales down the distance between every pair of points by the same number. For any $x \in A$, we define a map $\phi(x)$ as follows:

$$\phi(x) = Ux + [1/3, 1/3, 1/3]^T$$

where U is the matrix:

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{pmatrix}$$

Lemma 26 *If $x \in A$, $\phi(x)$ lies on the simplex.*

Proof: We first show that if x lies on the $x_1 - x_2$ plane, then Ux lies on the plane $x_1 + x_2 + x_3 = 0$. Let

$$y_1 = (1/\sqrt{2}, -1/\sqrt{2}, 0) \text{ and } y_2 = (0, 1/\sqrt{2}, -1/\sqrt{2}) \; .$$

As $y_1 = Ux_1$ and $y_2 = Ux_2$, if $x = \alpha_1 x_1 + \alpha_2 x_2$, then, $Ux = \alpha_1 y_1 + \alpha_2 y_2$. Since

$$y_1 \cdot (1, 1, 1) = y_2 \cdot (1, 1, 1) = 0 \; ,$$

$Ux \cdot (1, 1, 1) = 0$ as well, which means that Ux lies on the plane $x_1 + x_2 + x_3 = 0$.

Since Ux lies on the plane $x_1 + x_2 + x_3 = 0$, we deduce that $\phi(x)$ lies on the plane $x_1 + x_2 + x_3 = 1$. Since $x \in [0, 1/4] \times [0, 1/4]$, for any $i \in \{1, 2, 3\}$,

$$(Ux)_i \geq -\frac{1}{4} \times \frac{1}{\sqrt{2}} \geq -\frac{1}{3} \; .$$

Therefore, for any i, $(\phi(x))_i \geq 0$. Again, as $x \in [0, 1/4] \times [0, 1/4]$, for any $i \in \{1, 2, 3\}$,

$$(Ux)_i \leq \frac{1}{4} \times \frac{1}{\sqrt{2}} \leq \frac{2}{3} \; .$$

Therefore, $(\phi(x))_i \leq 1$ for each i, from which the lemma follows. ■

To map an instance I of MMC_{ℓ_2} on the $x_1 - x_2$ plane to the probability simplex, we simply apply the map ϕ on each point of I. This produces another instance I' of the problem on the simplex, which has the following property.

Lemma 27 *There is an approximation-preserving bijection between the solutions of I and the solutions of I'.*

Proof: We observe that as U is a unitary matrix, the map ϕ is a bijection. Moreover, since ϕ consists of a translation and a rotation, it preserves the distance between every pair of points. Therefore, applying ϕ on a solution of I produces a solution of I' of the same cost. The mapping ϕ is thus approximation preserving. ■

Finally, by mapping each point $x \in I'$ to $\epsilon x + (1-\epsilon)u$ where $u = [1/3, 1/3, 1/3]^T$ we generate a set of points that lie arbitrarily close to the center of Δ (setting ϵ as small as necessary.) Again, this transformation can be seen to be approximation preserving.

5.2 Restricted Centers

In this section, we consider the restricted version of MMC_{KL} and MTC_{KL} where we insist that the cluster centers $c^1, \ldots, c^n \in \{p^1, \ldots, p^n\}$. Our result is based on relating the problem to the SET-COVER problem and appealing to a result of Feige [14].

Theorem 28 *For any $\alpha \geq 1$, unless $P = NP$, no polynomial-time, α-approximation algorithm exists for either* MTC_{KL} *or* MMC_{KL}.

Proof: Consider a reduction from the problem SET-COVER: Consider $S_1, \ldots, S_{n-d-1} \in [d-1]$ and $k \leq d$. It was shown by Feige [14] that it is NP-hard to determine if there exists $\mathcal{S} = \{S_{i_1}, \ldots, S_{i_{k-1}}\}$ such that $\bigcup S_{i_j} = [d-1]$.

We first consider MMC_{KL}. Let $c_1, c_2 > 1$ such that

$$(d-1)e^{-c_1} < 1 \text{ and } (d-1)e^{-c_2} < e^{-c_1} \ .$$

Let q^i be the probability distribution with mass e^{-c_1} on each element in S_i, and the remaining mass on $\{d\}$. Let $p^i = e_i$ (i.e., the i-th vector of the standard basis) for $i \in [d-1]$. Let r be the probability distribution with e^{-c_2} mass on each element in $[d-1]$ and the remaining mass on $\{d\}$.

Note that $\text{KL}(p^i, q^j) = c_1$, $\text{KL}(p^i, r) = c_2$, and

$$\begin{aligned} \text{KL}(q^j, r) &= (1 - |S_j|e^{-c_1}) \ln \frac{1 - |S_j|e^{-c_1}}{1 - (d-1)e^{-c_2}} \\ &\quad + |S_j|e^{-c_1} \ln(e^{-c_1}/e^{-c_2}) \\ &\leq |S_j|e^{-c_1}(c_2 - c_1) \leq de^{-c_1}c_2 \end{aligned}$$

Hence, if there exists $\mathcal{S}$, the clustering with centers $p^{i_1}, \ldots, p^{i_{k-1}}, r$ costs at most $\max\{c_1, de^{-c_1}c_2\}$ whereas otherwise the cost is

$$\max\{c_1, c_2, de^{-c_1}c_2\} \geq c_2 \ .$$

Hence the ratio difference is at least

$$\frac{c_2}{\max\{c_1, de^{-c_1}c_2\}}$$

which we can make arbitrarily large.

This also implies that no approximation is possible for MTC_{KL} because any α'-approximate solution for MTC_{KL} is also a αn-approximation solution for MMC_{KL} for k-median when centers must be original points. ■

Bi-criteria Approximation: We briefly mention an approximation algorithm for the related approximation problem of finding the minimum number k' of centers $c^1, c^2, \ldots, c^{k'} \in \{p^1, \ldots, p^n\}$ such that for all $i \in [n]$,

$$\min_{j \in [k']} \text{KL}(p^i, c^j) \leq r$$

for some given r. This can be approximated up to a factor of $O(\log n)$ using a well-known approximation algorithm for SET-COVER. Specifically, for each p^i we define a set $S_i = \{j \in [n] : \text{KL}(p^j, p^i) \leq r\}$. Then, our problem becomes picking the smallest number of sets $S_{i_1}, S_{i_2}, \ldots$ such that $\cup_{j \geq 1} S_{i_j} = [n]$. An $O(\log n)$-approximation algorithm exists for this problem.

Acknowledgements: We would like to thank Sanjoy Dasgupta for helpful discussions.

References

[1] M. R. Ackerman, J. Blomer, and C. Sohler. Clustering for metric and non-metric distance measures. In *ACM-SIAM Symposium on Discrete Algorithms*, 2008.

[2] V. Arya, N. Garg, R. Khandekar, K. Munagala, and V. Pandit. Local search heuristic for k-median and facility location problems. In *ACM Symposium on Theory of Computing*, pages 21–29, 2001.

[3] M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *ACM Symposium on Theory of Computing*, pages 250–257, 2002.

[4] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, Melbourne, AU, 1998. ACM Press, New York, US.

[5] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[6] K. Chen. On k-median clustering in high dimensions. In *Symposium on Discrete Algorithms*, 2006.

[7] J. Chuzhoy, S. Guha, E. Halperin, S. Khanna, G. Kortsarz, R. Krauthgamer, and J. Naor. Asymmetric k-center is $\log^* n$-hard to approximate. *J. ACM*, 52(4):538–551, 2005.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

[9] K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. In *NIPS*, pages 321–328, 2006.

[10] I. Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *Ann. Statist.*, 19:2032–2056, 1991.

[11] I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *JMLR*, 3:1265–1287, 2003.

[12] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

[13] T. Feder and D. Greene. Optimal algorithms for approximate clustering. In *ACM Symposium on Theory of Computing*, 1988.

[14] U. Feige. A threshold of ln for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[15] D. Feldman, M. Monemizadeh, and C. Sohler. A ptas for k-means clustering based on weak coresets. In *Symposium on Computational Geometry*, pages 11–18, 2007.

[16] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985.

[17] S. Guha, P. Indyk, and A. McGregor. Sketching

information divergences. *Submitted to Journal of Machine Learning*, 2007.

[18] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742, 2006.

[19] S. Har-Peled and S. Mazumdar. Coresets for k-means and k-median clustering and their applications. In *ACM Symposium on Theory of Computing*, 2004.

[20] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k-means clustering. *Comput. Geom.*, 28(2-3):89–112, 2004.

[21] S. G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean kappa-median problem. In *European Symposium on Algorithms*, pages 378–389, 1999.

[22] A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1 + \epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *IEEE Symposium on Foundations of Computer Science*, pages 454–462. IEEE Computer Society, 2004.

[23] R. Panigrahy and S. Vishwanathan. An $o(\log^* n)$ approximation algorithm for the asymmetric p-center problem. *J. Algorithms*, 27(2):259–268, 1998.

[24] F. C. N. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *ACL*, pages 183–190, 1993.

[25] N. Slonim, R. Somerville, N. Tishby, and O. Lahav. Objective classification of galaxy spectra using the information bottleneck method. *Monthly Notes of the Royal Astronomical Society*, 323:270–284, 2001.

[26] N. Slonim and N. Tishby. Agglomerative information bottleneck. In *NIPS*, pages 617–623, 1999.

[27] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[28] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.