Electronic Colloquium on Computational Complexity, Report No. 35 (1999)

# Clustering for Edge-Cost Minimization

*preliminary version*

Leonard J. Schulman
College of Computing
Georgia Inst. Technology
Atlanta GA 30332-0280

**Abstract**

We address the problem of partitioning a set of $n$ points into clusters, so as to minimize the sum, over all intracluster pairs of points, of the cost associated with each pair. We obtain a randomized approximation algorithm for this problem, for the cost functions $\ell_2^2, \ell_1$ and $\ell_2$, as well as any cost function isometrically embeddable in $\ell_2^2$.

The maximization problem (maximize the costs of intercluster edges) is approximated with high probability to within multiplicative factor $(1 - \varepsilon)$; while the minimization problem either receives a multiplicative $1 + \varepsilon$ approximation, or else the optimal clustering is correctly identified except for a mislabelling of an $\varepsilon$ fraction of the point set. Given a fixed approximation parameter $\varepsilon$, the runtime is linear in $n$ for $\ell_2^2$ problems of dimension $o(\log n / \log \log n)$; and $n^{O(\log \log n)}$ in the general case.

The case $\ell_2^2$ is addressed by combining three elements: (a) Variable-probability sampling of the given points, to reduce the size of the data set. (b) Near-isometric dimension reduction. (c) A deterministic exact algorithm which runs in time exponential in the dimension (rather than the number of points). The remaining cases are addressed by reduction to $\ell_2^2$.

## Contents:

ISSN 1433-8092

# 1 Introduction

We consider the problem of clustering, or classifying, a set of data points $T$. We adopt an edge-cost minimization approach to this problem: a cost $\phi_{u,v}$ is charged for every pair of points $u, v \in T$ assigned to the same cluster. (Other terms for such a cost are "weight", "penalty", "energy", "dissimilarity" and "distance".) $\phi$ increases with dissimilarity of points, although it is not necessarily a metric. The clustering task is to partition $T$ into clusters $S$ and $\bar{S} = T - S$ so that the total cost is minimized. In other words, if we define $\phi(S) = \sum_{u,v \in S} \phi_{u,v}$, the task is to find a partition $(S, \bar{S})$ minimizing $\phi(S) + \phi(\bar{S})$ (which we will also write $\phi(S, \bar{S})$). More generally the task is to find a "$k$-partition" into clusters $\{S_1, ..., S_k\}$ whose sum of costs is minimal among $k$-partitions.

There is a great variety of problems calling for clustering, and it is unlikely that any one framework can suit all of these. The edge-cost approach offers two benefits. First, like any approach in which an objective function is to be optimized, it allows clustering algorithms to be compared both on the basis of their runtimes and on the basis of the quality of their output (namely whether they find an optimum or only approximately optimum partition; and whether this is accomplished with certainty or only with high probability). Second, we find out more than just what is the best partition with respect to the stated criterion. The ratio $\phi(S, \bar{S})/\phi(T)$, or $\phi(S_1, ..., S_k)/\phi(T)$, gives an indication of the quality of the partition: the achievable ratio may guide the choice of the most appropriate $k$, including the possibility that no partitioning hypothesis is supported by the data ($k = 1$).

At this level of generality, however, there is no way to find an optimum partition without an exhaustive examination of all $2^{|T|-1}$ (or generally Stirling number, second kind, $S_{|T|,k} \approx k^{|T|}$) partitions of the input set.

In this paper we provide an efficient randomized approximation algorithm for clustering with respect to a range of cost functions $\phi$. This is achieved by a combination of sampling; near-isometric dimension reduction; and reduction to the interesting special case in which $\phi$ is the square of a Euclidean metric, for which case we describe an exact deterministic algorithm running in time polynomial in the number of points and exponential in the dimension (as opposed to exponential in the number of points), hence polynomial for bounded dimension.

The cost functions we address are $\ell_2^2$ and any cost function isometrically (or nearly isometrically) embeddable in $\ell_2^2$, including $\ell_1$ and $\ell_2$. The main focus of this paper is an algorithm which, for any such function and any fixed $k, \varepsilon, \delta$, given a clustering problem on $n$ points in any dimension, computes in time $n^{O(\log \log n)}$, with probability at least $1 - \delta$, a $k$-clustering that is "$\varepsilon$-close" to optimum. In the special case that $\phi = \ell_2^2$ and that the dimension is $o(\log n / \log \log n)$, the runtime is linear in $n$. By "$\varepsilon$-close" we mean the following, where $\phi_{\text{opt}}$ is the cost of an optimal clustering: (a) The *cut* cost, $(\phi(T) - \phi_{\text{opt}})/\phi(T)$ is multiplicatively approximated to within factor $(1 - \varepsilon)$. (b) If $\phi_{\text{opt}} \geq \varepsilon\phi(T)$ then $\phi_{\text{opt}}$ is multiplicatively approximated to within factor $(1 + \varepsilon)$; and otherwise, in the case $k = 2$, a clustering $S$ is identified such that the fraction of points whose membership must be switched between $S$ and $\bar{S}$ in order to convert $S$ into an optimal clustering, is less than $\varepsilon$.

For the case of metric spaces, an approximation algorithm for max cut (i.e. $k = 2$) has been provided in recent independent work by Fernandez de la Vega and Kenyon [14] (building upon [4, 12, 13]). Max cut (maximization of $\phi(T) - \phi(S, \bar{S})$) is NP-complete [42, 25] (even for metric spaces); min cluster (minimization of $\phi(S, \bar{S})$) is equivalent, but multiplicative approximation of these quantities is not equivalent, with min cluster being harder since there is always a clustering for which $\phi(S, \bar{S}) \leq \phi(T)/2$.

The clustering problem is fully interesting already for the case in which the points of $T$ are of equal "significance" or "weight". However, all our results go through, and are in fact more naturally stated, for the case in which the points are weighted by a nonnegative real valued function $w$; this possibility will also be very useful in the algorithm. So, the more general formulation of the cost function is

$$\phi(S) = \sum_{\{u,v\} \subseteq S} w_u w_v \phi_{u,v}. \tag{1}$$

Observe that without loss of generality the points may be assumed distinct. We will assume throughout that $\phi$ is symmetric and that $\phi_{u,u} = 0 \; \forall u \in T$; hence equivalently $\phi(S) = \frac{1}{2} \sum_{u,v \in S} w_u w_v \phi_{u,v}$. We will also assume throughout that $\phi$ is nonnegative.

For general references in the field of clustering see [10, 38, 33, 69, 27, 36, 53, 54, 2, 7]; for discussions of a variety of interesting methods and application areas see [24, 68, 65, 58, 60, 67, 48, 26, 46].

A key role in our method is played by a random sampling process which, given $T$, picks a very small weighted collection of points. We can show that for a range of cost functions, the cost of this collection is with high probability close to that of the original collection $T$. In the case $\phi = \ell_2^2$, a clustering computation for such small samples can also be used to induce a good clustering of $T$. We therefore begin by describing the sampling process.

## 2 Sampling process

An optimistic idea for the clustering problem is simply to select each point with some small probability $p$; hopefully then good partitions of the sample will "lift" to good partitions of the original data. Indeed there is some promise in this approach, for if we let $T'$ denote the selected set, then $E(\phi(T')) = p^2 E(\phi(T))$. However, this is futile as a method of identifying good clusterings of $T$. For, there are sets $T$ with the following property: for almost all small subsets $S$ of $T$, the partition of $S$ inherited from the optimum partition of $T$, is much more expensive than the optimal partition of $S$. So examining the partitions of random subsets of $T$ does not contribute substantially toward partitioning $T$.

We propose instead a more interesting sampling method. In this method the selection probabilities are determined by the original weights and by the location of points within $T$; a variable-weighting method is used to balance the effects of the uneven sampling probabilities.

We analyze the sampling process for any nonnegative cost function $\phi$ on the edges (pairs of points) satisfying the following "$c$-metric" condition: there is a positive constant $c$ such that $\phi^{1/c}$ is a metric. Thus $\phi_{x,y}^{1/c} \leq \phi_{x,z}^{1/c} + \phi_{z,y}^{1/c}$ and consequently also $\phi_{x,y} \leq 2^c \max\{\phi_{x,z}, \phi_{z,y}\}$. With a little more care note that

$$\phi_{x,y} \leq 2^{c-1}(\phi_{x,z} + \phi_{z,y}). \tag{2}$$

Given a collection of points $T$ with weight function $w$, form a new variable-weights collection $T'$ in the following random process. For each point $u \in T$ let

$$\alpha_u = \frac{s w_u \sum_{v \in T} w_v \phi_{u,v}}{2\phi(T)}. \tag{3}$$

(A satisfactory choice for $s$ is, for example, 4. In practice it may be desirable to adjust this value to optimize the performance of the algorithm.) Observe that $s = \sum_u \alpha_u$. Let $\beta_u = \frac{\alpha_u}{1+\alpha_u}$. To each point $u$ assign an independently chosen random variable $K_u$ with the integral exponential distribution with expectation $\alpha_u$; namely, $P(K_u = i) = (1-\beta_u)\beta_u^i$. We will also denote this quantity $p_{u,i}$.

Observe for future reference that $\alpha_u = \frac{\beta_u}{1-\beta_u} = \sum_0^\infty i(1-\beta_u)\beta_u^i = \sum_0^\infty i p_{u,i} = \sum_1^\infty \beta^i$. Moreover, the variance of an integer exponential r.v. with expectation $\alpha$ is $\alpha(1+\alpha)$.

Now form the collection $T'$ by assigning weight $w'_u = w_u K_u / \alpha_u$ to each point $u$ of $T$.

We examine the r.v. $\phi(T')$. We begin with its first moment, establishing that it is an estimator of the desired quantity. This relies only on the independence of the random variables $\{K_u\}$ and the weight selection $w'_u = w_u K_u / E(K_u)$.

$$\begin{aligned}
E(\phi(T')) &= \frac{1}{2} \sum_{x,y \in T} \phi_{x,y} E(w'_x w'_y) && (4)\\
&= \frac{1}{2} \sum_{x,y \in T} \phi_{x,y} E(w'_x) E(w'_y) && (5)\\
&= \frac{1}{2} \sum_{x,y \in T} \phi_{x,y} w_x w_y && (6)\\
&= \phi(T) && (7)
\end{aligned}$$

Next we turn to the second moment.

$$E(\phi(T')^2) \quad = \quad \frac{1}{4}\sum_{x,y,z,t\in T}\phi_{x,y}\phi_{z,t}\frac{w_x w_y w_z w_t}{\alpha_x\alpha_y\alpha_z\alpha_t}\sum_{i,j,k,\ell\geq 0} ijk\ell P((K_x=i)\wedge(K_y=j)\wedge(K_z=k)\wedge(K_t=\ell))$$

We calculate this by beginning (first line below) as if the variables $K_x, K_y, K_z, K_t$ are independent even when some of $x, y, z, t$ collide; and then correcting for the effects of collisions. The second line corrects the first line for single collisions, and the third corrects the first for double collisions. The fourth line accounts properly for single collisions, and the fifth for double collisions. $E(\phi(T')^2) =$

$$\begin{aligned}
= \quad & \frac{1}{4}([\sum_{x,y,z,t\in T}\phi_{x,y}\phi_{z,t}\frac{w_x w_y w_z w_t}{\alpha_x\alpha_y\alpha_z\alpha_t}\sum_{i,j,k,\ell\geq 0} ijk\ell p_{x,i}p_{y,j}p_{z,k}p_{t,\ell} \\
& -4(\sum_{x,y,z\in T}\phi_{x,y}\phi_{z,y}\frac{w_x w_y^2 w_z}{\alpha_x\alpha_y^2\alpha_z}\sum_{i,j,k,\ell\geq 0} ijk\ell p_{x,i}p_{y,j}p_{z,k}p_{y,\ell} - \sum_{x,y\in T}\phi_{x,y}^2\frac{w_x^2 w_y^2}{\alpha_x^2\alpha_y^2}\sum_{i,j,k,\ell\geq 0} ijk\ell p_{x,i}p_{y,j}p_{x,k}p_{y,\ell}) \\
& -2\sum_{x,y\in T}\phi_{x,y}^2\frac{w_x^2 w_y^2}{\alpha_x^2\alpha_y^2}\sum_{i,j,k,\ell\geq 0} ijk\ell p_{x,i}p_{y,j}p_{x,k}p_{y,\ell}] \\
& +4[\sum_{x,y,z\in T}\phi_{x,y}\phi_{z,y}\frac{w_x w_y^2 w_z}{\alpha_x\alpha_y^2\alpha_z}\sum_{i,j,k\geq 0} ij^2 k p_{x,i}p_{y,j}p_{z,k} - \sum_{x,y\in T}\phi_{x,y}^2\frac{w_x^2 w_y^2}{\alpha_x^2\alpha_y^2}\sum_{i,j,k\geq 0} ij^2 k p_{x,i}p_{y,j}p_{x,k}] \\
& +2[\sum_{x,y\in T}\phi_{x,y}^2\frac{w_x^2 w_y^2}{\alpha_x^2\alpha_y^2}\sum_{i,j\geq 0} i^2 j^2 p_{x,i}p_{y,j}])
\end{aligned}$$

$$\begin{aligned}
= \quad & (\frac{1}{2}\sum_{x,y\in T}\phi_{x,y}\frac{w_x w_y}{\alpha_x\alpha_y}\sum_{i,j\geq 0} ijp_{x,i}p_{y,j})^2 \\
& +\sum_{x,y,z\in T}\phi_{x,y}\phi_{z,y}\frac{w_x w_y^2 w_z}{\alpha_x\alpha_y^2\alpha_z}\sum_{i,j,k\geq 0} ijkp_{x,i}p_{y,j}p_{z,k}(j-\sum_{\ell\geq 0}\ell p_{y,\ell}) \\
& +\sum_{x,y\in T}\phi_{x,y}^2\frac{w_x^2 w_y^2}{\alpha_x^2\alpha_y^2}[\sum_{i,j,k,\ell\geq 0} ijk\ell p_{x,i}p_{y,j}p_{x,k}p_{y,\ell} - \frac{1}{2}\sum_{i,j,k,\ell\geq 0} ijk\ell p_{x,i}p_{y,j}p_{x,k}p_{y,\ell} \\
& -\sum_{i,j,k\geq 0} ij^2 kp_{x,i}p_{y,j}p_{x,k} + \frac{1}{2}\sum_{i,j\geq 0} i^2 j^2 p_{x,i}p_{y,j}] \\
= \quad & (\frac{1}{2}\sum_{x,y\in T}\phi_{x,y}[\frac{w_x}{\alpha_x}\sum_{i\geq 0} ip_{x,i}][\frac{w_y}{\alpha_y}\sum_{j\geq 0} jp_{y,j}])^2 \\
& +\sum_{x,y,z\in T}\phi_{x,y}\phi_{z,y}\frac{w_x w_y^2 w_z}{\alpha_x\alpha_y^2\alpha_z}\sum_{i,j,k\geq 0} ijkp_{x,i}p_{y,j}p_{z,k}(j-\alpha_y) \\
& +\sum_{x,y\in T}\phi_{x,y}^2\frac{w_x^2 w_y^2}{\alpha_x^2\alpha_y^2}[\frac{1}{2}\sum_{i,j,k,\ell\geq 0} ijk\ell p_{x,i}p_{y,j}p_{x,k}p_{y,\ell} - \sum_{i,j,k\geq 0} ij^2 kp_{x,i}p_{y,j}p_{x,k} + \frac{1}{2}\sum_{i,j\geq 0} i^2 j^2 p_{x,i}p_{y,j}] \\
= \quad & \phi(T)^2 \\
& +\sum_{y\in T}\frac{w_y^2}{\alpha_y^2}\sum_{j\geq 0} jp_{y,j}(j-\alpha_y)[\sum_{x\in T}\phi_{x,y}\frac{w_x}{\alpha_x}\sum_{i\geq 0} ip_{x,i}]^2 \\
& +\sum_{x,y\in T}\phi_{x,y}^2\frac{w_x^2 w_y^2}{\alpha_x^2\alpha_y^2}\sum_{i,j\geq 0} ijp_{x,i}p_{y,j}[\frac{1}{2}\sum_{k,\ell\geq 0} k\ell p_{x,k}p_{y,\ell} - \sum_{k\geq 0} jkp_{x,k} + \frac{1}{2}ij]
\end{aligned}$$

$$
\begin{aligned}
&= \phi(T)^2 + \sum_{y \in T} \frac{w_y^2}{\alpha_y^2} \sum_{j \geq 0} j p_{y,j}(j - \alpha_y)[\sum_{x \in T} \phi_{x,y} w_x]^2 \\
&\quad + \frac{1}{2} \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\alpha_x^2 \alpha_y^2} \sum_{i,j \geq 0} ij p_{x,i} p_{y,j} [\alpha_x \alpha_y - 2j\alpha_x + ij] \\
&= \phi(T)^2 + \sum_{y \in T} \frac{w_y^2}{\alpha_y^2} \sum_{j \geq 0} j p_{y,j}(j - \alpha_y)[\frac{2\phi(T)\alpha_y}{s w_y}]^2 \\
&\quad + \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\alpha_x^2 \alpha_y^2} \sum_{i,j \geq 0} ij p_{x,i} p_{y,j} [\alpha_x \alpha_y - j\alpha_x - i\alpha_y + ij] \\
&= \phi(T)^2 + [\frac{2\phi(T)}{s}]^2 \sum_{y \in T} \sum_{j \geq 0} j p_{y,j}(j - \alpha_y) + \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\alpha_x^2 \alpha_y^2} \sum_{i,j \geq 0} ij p_{x,i} p_{y,j} (\alpha_x - i)(\alpha_y - j) \\
&= \phi(T)^2 + [\frac{2\phi(T)}{s}]^2 \sum_{y \in T} \alpha_y(1 + \alpha_y) + \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\alpha_x^2 \alpha_y^2} (\sum_{i \geq 0} i p_{x,i}(\alpha_x - i))(\sum_{j \geq 0} j p_{y,j}(\alpha_y - j)) \\
&= \phi(T)^2 + [\frac{2\phi(T)}{s}]^2 \sum_{y \in T} \alpha_y(1 + \alpha_y) + \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\alpha_x^2 \alpha_y^2} \alpha_x(1 + \alpha_x)\alpha_y(1 + \alpha_y) \\
&= \phi(T)^2 + [\frac{2\phi(T)}{s}]^2 \sum_{y \in T} \alpha_y(1 + \alpha_y) + \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\beta_x \beta_y}
\end{aligned}
$$

To summarize,

$$
E(\phi(T')^2) = \phi(T)^2 + [\frac{2\phi(T)}{s}]^2 \sum_{y \in T} \alpha_y(1 + \alpha_y) + \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\beta_x \beta_y} \tag{8}
$$

Now we analyze this expression. Beginning with the second term, recall that $\sum \alpha_u = s$, therefore $\sum \alpha_u^2 \leq s^2$, and so

$$
[\frac{2\phi(T)}{s}]^2 \sum_{y \in T} \alpha_y(1 + \alpha_y) \leq 4\phi(T)^2(1 + 1/s). \tag{9}
$$

Next we examine the third term. Recalling equation (3) and that $\beta_u = \alpha_u/(1 + \alpha_u)$, we write

$$
\sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\beta_x \beta_y} = \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2 (1 + \alpha_x)(1 + \alpha_y)}{\alpha_x \alpha_y} = \frac{4\phi(T)^2}{s^2} \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x w_y (1 + \alpha_x)(1 + \alpha_y)}{(\sum_z w_z \phi_{x,z})(\sum_z w_z \phi_{y,z})}
$$

and recalling $\sum \alpha_u = s$ we have

$$
\sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\beta_x \beta_y} \leq 4 \frac{(1+s)^2}{s^2} \phi(T)^2 \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x w_y}{(\sum_z w_z \phi_{x,z})(\sum_z w_z \phi_{y,z})}
$$

$$
= 4 \frac{(1+s)^2}{s^2} \phi(T)^2 \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x w_y}{\min\{\sum_z w_z \phi_{x,z}, \sum_z w_z \phi_{y,z}\} \max\{\sum_z w_z \phi_{x,z}, \sum_z w_z \phi_{y,z}\}} \tag{10}
$$

Let us lower bound each of the terms in the last denominator. For the min term, recalling inequality 2, consider that for any $x \in T$

$$
\phi(T) = \frac{1}{2} \sum_{u,v \in T} w_u w_v \phi_{u,v} \leq \frac{1}{2} \sum_{u,v \in T} w_u w_v 2^{c-1} (\phi_{u,x} + \phi_{x,v})
$$

$$= \frac{1}{2} \sum_{u,v \in T} w_u w_v 2^c \phi_{u,x} = 2^{c-1} w_T \sum_{u \in T} w_u \phi_{u,x}$$

Hence

$$\min\{\sum_z w_z \phi_{x,z}, \sum_z w_z \phi_{y,z}\} \geq 2^{1-c} \phi(T)/w_T.$$

For the max term, write (again using inequality 2)

$$\max\{\sum_z w_z \phi_{x,z}, \sum_z w_z \phi_{y,z}\} \geq \frac{1}{2} \sum_z w_z(\phi_{x,z} + \phi_{y,z}) \geq \frac{1}{2} \sum_z w_z 2^{1-c} \phi_{x,y} = 2^{-c} w_T \phi_{x,y}$$

Now combine the min and max analyses to continue from equation 10:

$$\sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x^2 w_y^2}{\beta_x \beta_y} \leq 4 \frac{(1+s)^2}{s^2} \phi(T)^2 \sum_{x,y \in T} \phi_{x,y}^2 \frac{w_x w_y}{(2^{1-c}\phi(T)/w_T)(2^{-c} w_T \phi_{x,y})}$$

$$= 2^{1+2c} \frac{(1+s)^2}{s^2} \phi(T) \sum_{x,y \in T} w_x w_y \phi_{x,y}$$

$$= 2^{2+2c} (1+s)^2 \phi(T)^2 / s^2 \qquad (11)$$

Unifying equations 8, 9 and 11 we find that

**Theorem 1** *If $\phi$ is a c-metric then for the random collection $T'$ selected as described above,*

$$E(\phi(T')^2) \leq \phi(T)^2 (5 + 4/s + 2^{2+2c}(1+s)^2/s^2)$$

□

For the clustering procedure we will need a slight extension of this statement. Let $G$ be an undirected graph whose vertices are the points of a collection $S$, and let

$$\phi_G(S) = \sum_{\{u,v\} \in G} w_u w_v \phi_{u,v}. \qquad (12)$$

If $S' \subseteq S$ then we let $\phi_G(S')$ equal $\phi_{G'}(S')$ where $G'$ is the induced graph on vertex set $S'$. By an analysis identical to equation 4, the prescribed sampling procedure on a set $T$ yields

$$E(\phi_G(T')) = \phi_G(T). \qquad (13)$$

Moreover, since $0 \leq \phi_G(T') \leq \phi(T')$,

$$E(\phi_G(T')^2) \leq E(\phi(T')^2). \qquad (14)$$

We therefore have:

**Theorem 2** *If $\phi$ is a c-metric then for any graph $G$ and for the random collection $T'$ selected as described above,*

$$E(\phi_G(T')^2) \leq \phi(T)^2 (5 + 4/s + 2^{2+2c}(1+s)^2/s^2)$$

□

**Note 3** The analysis depends only on 4-wise independence of the random variables $\{K_u\}$, so it can in fact be implemented with a small sample space. Actually strictly speaking in order to use exactly the 4-wise distribution called for, an infinite sample space would be needed because each random variable is supported on infinitely many values; however, large values occur rarely and so can be omitted without substantially increasing the probability of error of the algorithm. To be specific: our algorithm repeats the above sampling process, using $s = 4$, a total of at most $C\varepsilon^{-2} \log(n) \log(n/\delta)$ times (for some constant $C$). Since $\alpha_u \leq s = 4$ $\forall u$, it follows that $\beta_u = \frac{\alpha_u}{1+\alpha_u} \leq 4/5$ $\forall u$. Hence the probability that any of the vertices, in any of the samples, receives a random variable $K_u \geq j$, is at most $Cn\varepsilon^{-2}\log(n)\log(n/\delta)\frac{1}{5}\sum_{j'\geq j}(4/5)^{j'} = Cn\varepsilon^{-2}\log(n)\log(n/\delta)(4/5)^j$. If we therefore pick $j = \log_{5/4}[Cn\varepsilon^{-2}\delta^{-1}\log(n)\log(n/\delta)]$, and change the probability space so that $P(K_u = j) = \beta_u^j$, while $P(K_u = j') = 0$ for $j' > j$, then we are guaranteed of introducing an additional error probability of no more than $\delta$ into the analysis of the algorithm. Koller and Megiddo [47] provide a deterministic algorithm, running in time polynomial in $n$, which produces a 4-wise independent sample space of size $O((nj)^4) \subseteq \tilde{O}((n\log(n\varepsilon^{-1}\delta^{-1}))^4)$, with the assumed marginals (distributions of each variable $K_u$). (By slight modification of the marginals, earlier and somewhat simpler methods based on linear error correcting codes can likely be used as well, see [39, 50, 1, 9]. This would require modification of the analysis of the sampling process.)

# 3 The case $\phi = \ell_2^2$

## 3.1 Preliminaries and related literature

We focus now on the very interesting special case in which the cost function is the square of Euclidean distance. We denote the Euclidean distance between $u$ and $v$ by $\rho_{u,v}$, thus in this section $\phi_{u,v} = \rho_{u,v}^2$.

We first mention a physical analogy which may motivate consideration of this case (and which perhaps goes a little way toward explaining its tractability). Imagine that each pair of points is connected by an identical spring (thinking just of the uniform-weights case), whose resting length is 0. In this case $\phi(T)$ is the potential energy of the original system (Hooke's law), and we seek a partition which, when we sever the inter-cluster springs, minimizes the remaining energy.

We will speak of the "interior" and "exterior" of a sphere $C$ in $\mathbb{R}^d$: these are respectively the bounded and unbounded components of $\mathbb{R}^d - C$. We will say that two sets $S_1, S_2 \subset \mathbb{R}^d$ are *separated* (or *strictly separated*) by sphere $C$ if one of these sets is contained in the interior and the other in the exterior of $C$. We will say that $S_1, S_2$ are *weakly separated* by sphere $C$ if one of these sets is disjoint from the interior and the other from the exterior of $C$.

The key to the deterministic algorithm computing the optimal 2-partition or $k$-partition of a point set, is the following corollary of proposition 8:

**Corollary 4** *If $(S, T - S)$ is an optimal partition with respect to $\phi = \ell_2^2$ of a (possibly variable-weight) point set $T$, then there exists a sphere separating $S$ and $T - S$.*

Let $c(S)$ be the center of gravity of a set of points $S$, $c(S) = w_S^{-1}\sum_{u\in S} w_u u$. (Here $w_S = \sum_{u\in S} w_u$, or simply $|S|$ in the uniform-weights case.) Let $\mathrm{Var}(S) = w_S^{-1}\sum_{v\in S} w_v \rho_{v,c(S)}^2$. Calculation shows that an equivalent formulation of the $\ell_2^2$ cost function is $\phi(S) = w_S^2\mathrm{Var}(S)$.

Weak separation was shown previously by Boros and Hammer [8]. The distinction between the kinds of separation was overlooked. Later Inaba, Katoh and Imai [35] proposed examining all sphere partitions to find an optimal partition. That proposal is justified only on the basis of the present work, because weak separation does not imply a sub-exponential time algorithm.

The proper handling of data that is "singular" in the sense that it contains more than $d + 1$ co-spherical points (not necessarily a rarity in integer-coordinate data) has turned out to be the aspect requiring the most care both in the description of the deterministic exact algorithm, and in its implementation (by the author and by students).

On the other hand, since perturbations of the point locations affect $\phi$ continuously, this issue can, in the case of approximation algorithms, be circumvented (though there is no need to) by first perturbing the points into general position with respect to spheres.

Based upon corollary 4 we provide a deterministic algorithm computing an optimal 2-partition in section 3.4.

The first mention of $\ell_2^2$ as a clustering criterion may be by Kiseleva, Muchnik and Novikov [45], for point sets in one dimension.

## 3.2 Necessary condition for local optimality

Since points are allowed varying weights, it is natural to allow clusterings in which the weight of a point is allocated among several clusters. However as can easily be verified, to any such clustering there corresponds another of lesser or equal cost which splits no points. (This is true of any cost function, requiring only $\phi_{u,u} = 0 \; \forall u$; and also for criteria such as $\psi$ discussed in section 5.) Hence in the sequel only partitions which assign points to unique clusters will be considered.

Versions of the results in this section hold for any measure, but for simplicity and because of the algorithmic focus, we restrict ourselves to the case of measures supported on finite point sets.

**Definition 5** *The distance between two $k$-partitions $\mathcal{S} = \{S_1, ..., S_k\}$ and $\mathcal{R} = \{R_1, ..., R_k\}$ of a set $T$, is the least number of elements whose memberships must be changed so that $\forall i \; \exists j \; S_i = R_j$. (Equivalently, the least Hamming distance between the vectors in $\{1, ..., k\}^n$ specifying the partitions $\mathcal{S}$ and $\mathcal{R}$, that can be obtained by permutation of the alphabet $\{1, ..., k\}$.)*

**Definition 6** *A $k$-partition $\mathcal{S} = \{S_1, ..., S_k\}$ is $j$-stable if its cost $\phi(\mathcal{S}) = \sum \phi(S_i)$ is minimal among all $k$-partitions within distance $j$.*

If $\mathcal{S}$ is optimal then it is $j$-stable for any $j$. If $\mathcal{S}$ is $(n-1)$-stable then it is optimal.

Consider a set $R$ and a point $v$. Then $\phi(R \cup \{v\}) - \phi(R) = w_v \sum_{u \in R} w_u \rho^2_{u,v}$. This can be rewritten

$$\phi(R \cup \{v\}) - \phi(R) = w_v[w_R^{-1}\phi(R) + w_R \rho^2_{v,c(R)}]. \tag{15}$$

For, place $c(R)$ at the origin of the coordinate system, and $v$ at the position $\rho_{v,c(R)}$ on the first axis. Then $w_v \sum_{u \in R} w_u \rho^2_{u,v} = w_v \sum_{u \in R} w_u [\sum_2^d u_i^2 + (u_1 - \rho_{v,c(R)})^2] = w_v[\sum_{u \in R} w_u \sum_1^d u_i^2 + w_R \rho^2_{v,c(R)} - 2\rho_{v,c(R)} \sum_u w_u u_1] = w_v[\sum_{u \in R} w_u \rho^2_{u,c(R)} + w_R \rho^2_{v,c(R)}] = w_v[w_R^{-1}\phi(R) + w_R \rho^2_{v,c(R)}]$.

Note that equation 15 is a special case of the more general

$$\phi(R) = \sum_{i,j} [\frac{w_{R_i}\phi(R_j)}{w_{R_j}} + \frac{1}{2} w_{R_i} w_{R_j} \rho^2_{R_i,R_j}]. \tag{16}$$

expressing the cost of $R$ in terms of constituents $\{R_i\}$ with centers $\{c(R_i)\}$, weights $\{w_{R_i}\}$ and costs $\{\phi(R_i)\}$; we write $\rho_{R_i,R_j}$ for $\rho_{c(R_i),c(R_j)}$. This expression in turn can be written in the following way, which will be useful in the sequel:

$$\phi(R) = \sum \phi(R_i) + \sum_{i<j} (\phi(R_i \cup R_j) - \phi(R_i) - \phi(R_j)) \tag{17}$$

Now consider an existing partition $\{S, \bar{S}\}$ and a new point $v$. To which cluster is it preferable to adjoin $v$?

Define three regions partitioning space as follows: the region in which it is preferable to adjoin the new point to $S$, $\eta(S) = \{v \in \mathbb{R}^d : \phi(S \cup \{v\}) - \phi(S) < \phi(\bar{S} \cup \{v\}) - \phi(\bar{S})\}$; the region in which it is preferable to adjoin to $\bar{S}$, $\eta(\bar{S}) = \{v \in \mathbb{R}^d : \phi(S \cup \{v\}) - \phi(S) > \phi(\bar{S} \cup \{v\}) - \phi(\bar{S})\}$; and the boundary between these two regions, where there is a tie, $\nu = \eta(S)^c \cap \eta(\bar{S})^c$.

The surface $\nu$ is defined by the equation

$$\phi(S \cup \{v\}) - \phi(S) = \phi(\bar{S} \cup \{v\}) - \phi(\bar{S}),$$

equivalently

$$\sum_{u \in S} w_u \rho^2_{u,v} = \sum_{u \in \bar{S}} w_u \rho^2_{u,v},$$

equivalently

$$w_S \rho^2_{v,c(S)} + w_S^{-1}\phi(S) = w_{\bar{S}} \rho^2_{v,\bar{S}} + w_{\bar{S}}^{-1}\phi(\bar{S}). \tag{18}$$

Examination of the last condition shows that $\nu$, if not empty, is a sphere (a hyperplane if $w_S = w_{\bar{S}}$). Details in section 3.3.

**Proposition 7** *$S$ is 1-stable if and only if $S \subseteq \eta(S)^c$ and $\bar{S} \subseteq \eta(\bar{S})^c$.*

**Proof:** If $S$ is 1-stable then, for any point $v \in S$, $\sum_{u \in S-\{v\}} w_u \rho^2_{u,v} \leq \sum_{u \in \bar{S}} w_u \rho^2_{u,v}$. Equivalently, $\sum_{u \in S} w_u \rho^2_{u,v} \leq \sum_{u \in \bar{S}} w_u \rho^2_{u,v}$, implying that $v \in \eta(S)^c$. The argument for $\bar{S}$ is identical. The converse is immediate. □

**Proposition 8** *If $S$ is 2-stable then either $S \cap \nu = \emptyset$ or $\bar{S} \cap \nu = \emptyset$.*

**Proof:** Suppose that $u \in S \cap \nu$ and $v \in \bar{S} \cap \nu$. Now exchange the memberships of these points. The change in $\phi$ is

$$\Delta\phi = w_u( \sum_{r \in \bar{S}-\{v\}} w_r \rho^2_{u,r} - \sum_{s \in S} w_s \rho^2_{u,s}) + w_v( \sum_{s \in S-\{u\}} w_s \rho^2_{v,s} - \sum_{r \in \bar{S}} w_r \rho^2_{v,r})$$

Since $u$ and $v$ are on $\nu$, each of these terms equals $-w_u w_v \rho^2_{u,v}$. Hence $\Delta\phi = -2 w_u w_v \rho^2_{u,v} < 0$, contradicting 2-stability. □

Let $\Phi_d(n) = \sum_0^d \binom{n}{i}$.

**Proposition 9** *A set of $n$ points in $\mathbb{R}^d$ has at most $\Phi_{d+1}(n-1)$ 2-stable 2-partitions. The clusters of every such partition are separated by a sphere.*

Although an $O(n^{d+1})$ bound is immediate (the sphere $\nu$ needs only to be perturbed slightly in case it contains points of one of the clusters), it is worth examining the combinatorics of sphere partitions a little more closely.

First, a detour into the literature concerning partitions of space by hyperplanes. Schläfli showed in the last century that the number of cells in a partition of $\mathbb{R}^d$ by $n$ hyperplanes is at most $\Phi_d(n)$ ([62] p. 209), and that this bound is achieved for hyperplanes in general position. The number of partitions of $n$ points in general position in $\mathbb{R}^d$ by hyperplanes is deduced to be $\Phi_d(n-1)$ by an inversion in a sphere centered at one of the points. (It is not hard to see that the number is only reduced for points not in general position). This bound has been rediscovered a few times, including by Harding [32] (who was aware of Schläfli's work but not of the connection), in the context of clustering.

Now to prove the theorem lift the points of the set by stereographic projection (see e.g. [59]) to the surface of a sphere of dimension $d$ embedded in $\mathbb{R}^{d+1}$ (the reverse process of terrestrial mapmaking). Partitions of the original set by spheres of dimension $d-1$ are in bijective correspondence with partitions of the lifted set by hyperplanes of dimension $d$.

Almost the same result can be obtained from some broader considerations (which, however, rest upon the same proof technique). Vapnik and Chervonenkis [71], Sauer [61] and Perles and Shelah [64] showed that the number of subsets of a set of size $n$ that can be defined by intersections with elements of a range space of VC dimension $d$, is at most $\Phi_d(n)$. (For a discussion of applications in learning theory, statistics, and combinatorial and computational geometry see [43, 70, 52].) Dudley has shown that the range space of balls in dimension $d$ has VC dimension $d+1$ [17], and the bound $\Phi_{d+1}(n)$ follows. □

The 2-stable partitions are even further restricted:

**Proposition 10** *If $S \subset R$ then $\eta(R) \subset \eta(S)$. Furthermore $\nu(R) \cap \nu(S)$ can contain at most one point.*

**Proof:** A consequence of the equation $\phi(S \cup \{v\}) - \phi(S) = w_v \sum_{u \in S} w_u \rho^2_{u,v}$. □

A collection of elements of a poset is termed a $j$-family if it contains no chains of length $j+1$ [29]. (A 1-family is an antichain.)

**Corollary 11** *(a) The collection of sets which occur as clusters in 1-stable partitions of a set $T$ are a 2-family in the poset of subsets of $T$. (b) The collection of sets $S \cap \eta(S)$ for 1-stable partitions $(S, \bar{S})$ of a set $T$ are an antichain in the poset of subsets of $T$.*

**Proof:** (a) Follows from propositions (7,10): to begin with, if $S \subset R$ are both clusters in 1-stable partitions, then so is every $Q$ satisfying $S \subseteq Q \subseteq R$, since then $Q \subseteq R \subseteq \eta(R)^c \subseteq \eta(Q)^c$ and $\bar{Q} \subseteq \bar{S} \subseteq \eta(\bar{S})^c \subseteq \eta(\bar{Q})^c$. It suffices therefore to rule out chains of length 3. Only a point $v$ on $\nu(S)$ can be adjoined to a 1-stable cluster $S$ to yield a new 1-stable cluster $Q = S \cup \{v\}$. Having adjoined $v$, every other point $u$ of $T$ either lies in $\eta(\bar{Q})$, in which case it is strictly preferable to leave $u$ in $\bar{Q}$; or else $u$ lies in $(\eta(Q) \cup \nu(Q)) - \{v\} \subset \eta(S)$, in which case by 1-stability of $S$, $u$ already belongs to $S$. (b) As indicated in (a), a 2-chain can only be a pair of sets $S$, $S \cup \{v\}$ for $v \in \nu(S)$. □

## 3.3 The spherical boundary

We show here the details of the solution of equation (18). Let $r$ be the center of gravity of $S$, let $s$ be the center of gravity of $\bar{S}$, and let $a_r = w_S, b_r = \phi(S)/w_S, a_s = w_{\bar{S}}$, and $b_s = \phi(\bar{S})/w_{\bar{S}}$. Each region "belongs to" the center of gravity of one of the clusters (although the centers might not lie within their own regions). The surface $\nu$ is specified by

$$a_r \rho_{r,v}^2 + b_r = a_s \rho_{s,v}^2 + b_s. \tag{19}$$

Let $\rho = \rho_{r,s}$. Parameterize any point $v$ by letting $h$ be the vector that is perpendicular to the line $m$ passing through $r$ and $s$, and such that $v - \rho h$ is on $m$. Parameterize $m$ as $(\frac{1}{2} - z)r + (\frac{1}{2} + z)s$, for a real parameter $z$. (E.g. $z = -1/2$ indicates the point $r$.) Let $z$ be the value specifying the point $v - \rho h$. Hence $v = (\frac{1}{2} - z)r + (\frac{1}{2} + z)s + \rho h$.

The boundary between the regions occurs on the surface

$$a_r \rho^2[(z + \frac{1}{2})^2 + |h|^2] + b_r = a_s \rho^2[(z - \frac{1}{2})^2 + |h|^2] + b_s \tag{20}$$

where $|h|$ denotes the length of $h$. If $a_r \neq a_s$ then this is equivalent, writing $y = z + \frac{1}{2}\frac{a_r + a_s}{a_r - a_s}$, to the sphere

$$\rho^2(y^2 + |h|^2) = \rho^2 \frac{a_r a_s}{(a_r - a_s)^2} - \frac{b_r - b_s}{a_r - a_s}.$$

In other words, the sphere of radius $[\rho^2 \frac{a_r a_s}{(a_r - a_s)^2} - \frac{b_r - b_s}{a_r - a_s}]^{1/2}$, centered at the point corresponding to $y = 0$ (equivalently $z = -\frac{1}{2}\frac{a_r + a_s}{a_r - a_s}$) on the line $m$. Note that one of the regions may be empty, in case $\rho^2 \frac{a_r a_s}{(a_r - a_s)^2} - \frac{b_r - b_s}{a_r - a_s} < 0$.

If $a_r = a_s = a$ then the boundary is the hyperplane $z = \frac{b_s - b_r}{2a\rho^2}$.

## 3.4 Exact deterministic algorithm for 2-partitions

**Theorem 12** *For fixed $d$ an optimal 2-partition in $\mathbb{R}^d$ may be found in $O(n^{d+1})$ time.*

**Proof:** A time bound of $O(n^{d+2})$ is easily obtained by expending $O(n)$ time computing $\phi$ for each sphere partition. Time $O(n^{d+1})$ is achieved as follows.

Begin by applying a stereographic projection, mapping the points in $V \cong \mathbb{R}^d$ into a sphere $S^d \subset W \cong \mathbb{R}^{d+1}$, thereby transforming spherical partitions of the points to hyperplane partitions of their images in $W$. Next use geometric duality to transform the $n$ points to $n$ hyperplanes in $W$. Every cell in the arrangement of $n$ hyperplanes, corresponds to a hyperplane partition of the $n$ points. (The different points in the cell represent different hyperplanes yielding the same partition.) Standard methods exist for constructing the incidence graph of this arrangement; the size of the arrangement (total number of faces of all dimensions) is $O(n^{d+1})$, and the arrangement can be constructed in the same amount of time ([18] §7). From this arrangement, a graph (represented by its adjacency list) can be constructed, whose nodes are the cells of the arrangement, connected by an edge if the corresponding cells share a hyperplane boundary; to each edge is also attached a label indicating the identity of the hyperplane boundary (hence of the corresponding point in $W$, hence of the corresponding point in $V$). Now begin by picking any vertex of the graph (cell of the arrangement) and computing the cost of the partition it defines. Then execute a breadth-first-search of the graph. Each time an edge is crossed, update the cost of the partition in unit time by toggling the membership of the corresponding point in $V$, and using equation 15. Breadth first search runs in time linear in the size of the graph, $O(n^{d+1})$. □

The same runtime was claimed for this problem in [35], without argument; support for this claim could not be found.

It is an open question whether corollary 11 can be used to improve the asymptotic runtime of an algorithm examining all potential 2-stable 2-partitions. To begin with, a $o(n^{d+1})$ bound on the size of an antichain in the poset of intersections with spheres would be required; the existence of such a bound seems to be an open question. (The related "$k$-sets" question for intersections with halfspaces is a long-standing challenge; see [52, 49, 22, 19, 5, 3, 57, 21, 15].) It would also be necessary to implement an efficient search of sphere partitions which did not expend much effort on spheres not in the relevant 2-family.

## 3.5 Exact deterministic algorithm for $k$-partitions

Consider a 1-stable $k$-partition. Just as for 2-partitions, space is partitioned into $k$ open cluster regions (and their boundaries) dictating to which cluster a new point would be adjoined. The same derivation which led to equation (18) indicates that cluster region $i$ contains all points $v$ for which

$$w_{S_i}\rho^2_{v,c(S_i)} + w^{-1}_{S_i}\phi(S_i) < w_{S_j}\rho^2_{v,c(S_j)} + w^{-1}_{S_j}\phi(S_j) \;\; \forall j \neq i.$$

Hence:

**Proposition 13** *To a 1-stable $k$-partition there corresponds a set of $\binom{k}{2}$ spheres, such that each cluster region is the union of regions defined by intersections of interiors or exteriors of the spheres. If the $k$-partition is 2-stable then (just as for the 2-partition case) the boundary region between two of these clusters can only contain points belonging to one of them; the same number of spheres therefore also suffice in order to separate the clusters.* □

Inaba, Katoh and Imai [35] observed that 1-stability (but actually this requires 2-stability) implies that a $k$-partition is further constrained. Form the space $\mathbb{R}^{(d+2)k}$ whose dimensions are the weights, costs, and cartesian coordinates of the centers of gravity of the clusters. Now each triple $(a, b, c)$, in which $1 \leq a \leq n$ represents a point of the data, and $1 \leq b < c \leq k$ represent labels of clusters, specifies a polynomial of degree 3 over these variables (see equation 19). Let $E$ denote the union of the zero-sets of these polynomials. Each 2-stable partition is associated with a distinct open region of space in the complement of $E$. (A "sign pattern" of the polynomials; possibly a union of several topological components.) As suggested in [35], the partitions defined by these regions can be examined exhaustively, and that partition which gives the least-cost clustering can be identified as optimal. A theorem of Warren [73] shows that the number of components in $\mathbb{R}^N - E$, where $E$ is the union of the zero-sets of $M$ polynomials in $N$ variables each of degree at most $D$, is less than $((4eDM)/N)^N$. Therefore the number of regions to be examined in our case is at most $(\frac{12en\binom{k}{2}}{(d+2)k})^{(d+2)k}$, which, for fixed $d$ and $k$, is $O(n^{(d+2)k})$. In the recent [34], theorem 5.6, the improvement $O(n^{(d+2)k-1})$ is claimed. The best method (as a function of $n$) known to the present author, for the task of examining each of these regions, uses Gröbner basis computations (and, eventually, numerical root-finding for high-degree univariate polynomials), and has a worst case time analysis of $O(n^{(d+2)k+1}\exp(\exp((d+2)k)))$. The dependence in this bound on $d$ and $k$ is evidently severe.

In [35] a runtime of $n^{(d+2)k+1}$ is claimed, without substantiation, for the process of examining the regions and choosing the best partition; the claim is repeated in [34], but it has not been possible to glean an intended algorithm from that source or from queries to the authors. No suggestion of a resort to Gröbner basis computations is made in these references.

Let $F(n, d, k) = \min\{O(n^{(d+1)\binom{k}{2}}k^{O(\binom{k}{2}^{d+1})}), O(n^{(d+2)k+1}\exp(\exp((d+2)k)))\}$.

**Proposition 14** *There is an algorithm finding the optimal $k$-partition of a set of $n$ points which runs in time $F(n, d, k)$.*

For the $O(n^{(d+1)\binom{k}{2}}k^{O(\binom{k}{2}^{d+1})})$ bound, the 2-partition algorithm is mimicked: range over all $n^{(d+1)\binom{k}{2}}$ collections of $\binom{k}{2}$ spheres defined by points of the set. Each such collection defines a partition of the points into $O(\binom{k}{2}^{d+1})$ regions, and this partition can be coarsened in at most $k^{O(\binom{k}{2}^{d+1})}$ ways to form a partition into $k$ clusters; for each such possibility, compute the center of gravity of each cluster, obtain the coefficients "$a_{s_i}$" and "$b_{s_i}$" as in section 3.3, and solve for the spherical boundaries which these centers define. The spheres (or at the least the last sphere) should be changed incrementally (using breadth-first search) so that with each change, the partition cost can be updated in unit time. Then either check whether the partition defined by these boundaries agrees with that under consideration (in which case it is indeed 2-stable), or simply compute the value of $\phi$ and store the partition if it is the best seen so far. □

In practice, Gröbner basis computations are often substantially faster than is guaranteed by the worst case analysis. Perhaps the relative merits of the two methods described above are best clarified by experience. The runtime, in either method, is polynomial in $n$, but impractical for large $n$ for any but modest $d$ and $k$. However, the approximation method described below will address the dependence on $n$ and $d$. The dependence on $k$ remains steep. Further improvement is highly desirable, and an interesting problem. One may, of course, adopt a "divisive" or "top-down" strategy, as is commonly done in clustering work: first compute a good 2-partition of the original point set, then further 2-partition each of the clusters thus obtained, and so forth. There are no guarantees for the quality of results obtained in this way, nor does it seem clear whether optimism is warranted. Still, the approach merits exploration; though it is unlikely that a divisive approach can guarantee optimal results, it may be realistic to seek an approximation algorithm in this way. It is possible that the method of choice for large $k$ is to use the divisive strategy to partition into more parts than is desired; and then choose the least costly combination of these parts into $k$ clusters.

### 3.6 Simplification of $\ell_2^2$ clustering by dimension reduction

The set $T$ of $n$ points to be clustered may lie in a high dimensional Euclidean space. We need never consider a space of dimension greater than $n-1$: input given in a higher dimensional space should be reduced to this case by projection onto the affine subspace containing $T$.

**Proposition 15** *Fix any $\varepsilon > 0$. Given a set $T$ of $n$ points in $\mathbb{R}^{n-1}$, a $k$-partition of $T$ can be computed whose cost $\phi$ is within a factor of $1+\epsilon$ of optimal, in time $n^{O(\varepsilon^{-2}\log n)}$ (for $k=2$) and $F(n, \varepsilon^{-2}\log n, k)$ (for general $k$).*

The same method can be applied to reduce dimension before applying any clustering algorithm for the objective function $\psi$.

**Proof:** Johnson and Lindenstrauss showed that if a set $T$ of $n$ points in Euclidean space is mapped under a random orthogonal projection $M$ to an $O(\frac{\log n}{\epsilon^2})$-dimensional subspace, then with high probability the distortion of the metric on these points is no more than $1+\epsilon$ [40] (a constant of 4 is achievable in this theorem). (The distortion is $\max_{a,b,c,d\in T}(\frac{\rho_{Ma,Mb}\rho_{c,d}}{\rho_{a,b}\rho_{Mc,Md}})$.) Such a mapping may be found efficiently (in time $\tilde{O}(n^2)$) by trial and error.

Once a suitable mapping has been found, 2-partition or $k$-partition algorithms (deterministic exact, for a guaranteed approximation, or randomized approximate, for a high probability result) can be applied. □

For the objective function $\psi$ to be defined in section (5) one need range in the deterministic algorithm over hyperplane rather than sphere partitions. As shown in [31], improving on the immediate $O(n^{d+1})$, this can be done in time $O(n^d \log n)$ in dimension $d$. This can be improved:

**Note 16** The optimal clustering for $\psi$ can be found in time $O(n^d)$ for $d \geq 2$, by using geometric duality and computing the incidence graph of the arrangement of hyperplanes, in the same manner as in section 3.4.

The cluster regions in the original space are the preimages of cluster regions in the lower dimensional space, hence cylinders. In the case of 2-partitions for $\phi$ the cross-sections of the cylinders are spheres or their complements. (Halfspaces for $\psi$).

### 3.7 Simplification of $\ell_2^2$ clustering by sampling

We now show what is the consequence for $\phi = \ell_2^2$ of using the sampling process described in section 2. The cost function $\ell_2^2$ is of course a 2-metric in the sense discussed in section 2. Hence theorem 1, with $s \geq 4$ and $c = 2$, implies that

$$E(\phi(T')^2) \leq 106\phi(T)^2 = 106E(\phi(T'))^2. \tag{21}$$

Fix $s = 4$. Let $D(p;q)$ denote the information divergence or Kullback-Liebler divergence, $D(p;q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$.

Given $\varepsilon$, $\delta$, set $a = 636\varepsilon^{-2}$, $b = (\log(n^{d+1}\delta^{-1}))/D(1/2;2/3)$, and repeat the sampling process described in section (2) $t = ab$ times. Let $T_i$ (for $1 \leq i \leq t$) be the collection of points (with appropriate weights) obtained in trial $i$.

For a collection of points $S$ and a sphere $\gamma$ containing no point of $S$ (with $\gamma_1$ and $\gamma_2$ denoting the two closed regions of space bounded by $\gamma$), let $\phi_\gamma(S)$ be the cost of a partition of $S$ by $\gamma$, namely $\phi_\gamma(S) = \phi(S \cap \gamma_1) + \phi(S \cap \gamma_2)$; this corresponds to the notation of equation 12 with the understanding that the graph consists of all pairs of points not separated by $\gamma$.

Let $U$ be the set of points sampled with nonzero weight in any of the sampling processes. For every spherical partition of the set $U$ (represented by a sphere $\gamma$ passing through no point of $U$), consider the following quantity:

$$h(\gamma) = \text{median}_{j=1}^{b}\{\frac{1}{a}\sum_{i=1}^{a}\phi_\gamma(T_{a(j-1)+i})\}. \tag{22}$$

**Lemma 17** *For any given sphere $\gamma$, the inequality*

$$|\frac{h(\gamma) - \phi_\gamma(T)}{\phi(T)}| < \varepsilon/2 \tag{23}$$

*holds with probability at least $1 - \delta n^{-d-1}$. For the optimal sphere partition $\gamma$ of $U$, the inequality*

$$|\frac{h(\gamma) - \phi_\gamma(T)}{\phi_\gamma(T)}| < \varepsilon \tag{24}$$

*holds with probability at least $1 - \delta n^{-d-1}$.*

**Proof:** For a fixed sphere $\gamma$, and any $i$, the random variable $\phi_\gamma(T_i)$ is an unbiased estimator of $\phi_\gamma(T)$; and, using theorem 2, its variance is at most $106\phi(T)^2$.

Correspondingly, for a fixed sphere $\gamma$, and any $j$, the random variable $M = \frac{1}{a}\sum_{i=1}^{a}\phi_\gamma(T_{a(j-1)+i})$ is an unbiased estimator of $\phi_\gamma(T)$, with variance at most $106\phi(T)^2/a$. Hence using the Chebychev inequality,

$$P(|\frac{M - \phi_\gamma(T)}{\phi(T)}| > \varepsilon/2) < \frac{4\text{Var}(M)}{\varepsilon^2\phi(T)^2} = \frac{424}{a\varepsilon^2} = 2/3 \tag{25}$$

By an application of the Chernoff bound this implies the first statement of the lemma. The best sphere partition $\gamma$ satisfies $\phi_\gamma(T) \leq \phi(T)/2$, which implies the second statement of the lemma. □

As noted earlier, the number of distinct sphere partitions of $T$ is bounded by $\Phi_{d+1}(n-1) < n^{d+1}$. We can now conclude that with high probability, the sphere partitions of $U$ are an "$\varepsilon$-approximate" set of representatives for the sphere partitions of $T$, in the following sense:

**Theorem 18** *With probability at least $1-\delta$: $|\frac{h(\gamma)-\phi_\gamma(T)}{\phi(T)}| < \varepsilon/2$ for all spheres $\gamma$, and, for the optimal sphere cut, $|\frac{h(\gamma)-\phi_\gamma(T)}{\phi_\gamma(T)}| < \varepsilon$.* □

**Note 19** This is quite different from saying that $|\frac{\phi_\gamma(U)-\phi_\gamma(T)}{\phi(T)}| < \varepsilon$.

**Note 20** Theorem (18) offers the possibility that instead of carrying out the analysis of the algorithm with a union bound over all $O(n^{d+1})$ sphere partitions of $T$, it is enough to analyze a union bound over all sphere partitions of a sample of size $O(\varepsilon^{-2}\log(n^{d+1}\delta^{-1}))$, as described above (the sample need not actually be selected). This allows the actual algorithm to use fewer points in its sample, and to save time in analyzing all sphere partitions of that sample. However, this gain is offset by the need to use smaller values of $\varepsilon$ in the intermediate steps in order to obtain the same value desired in the specification of the problem.

**Note 21** Theorem (18) implies the existence of a subset $U$ of $T$ of size $O(\varepsilon^{-2}\log(n^{d+1}))$, such that any two spheres which define the same partition of $U$, define partitions of $T$ whose cut costs are within $\varepsilon\phi(T)$ of each other.

## 3.8 Randomized approximation algorithm for 2-partitions

1. Depending on the dimension $d$ execute either 1A or 1B:

1A. If the dimension is low, $d \in o(\varepsilon^{-2}\log n)$:
Carry out the above sampling procedure $t = 636\varepsilon^{-2}\log(n^{d+1}\delta^{-1})/D(1/2;2/3)$ times, and proceed to either option 2A or 2B below.

1B. If the dimension is high, $d \in \Omega(\varepsilon^{-2}\log n)$:
First carry out the dimension-reduction procedure described in section (3.6), reducing the dimension to $d' = O((\varepsilon/3)^{-2}\log n)$ while distorting all distances by at most $1+\varepsilon/3$. Then we carry out the sampling procedure, again using the parameter $\varepsilon/3$, i.e. setting $t = 636(\varepsilon/3)^{-2}\log(n^{d'+1}\delta^{-1})/D(1/2;2/3)$. (These choices guarantee that the combined allowed error $(1+\varepsilon/3)^2$ is less than $1+\varepsilon$, provided $\varepsilon \leq 1$; the case $\varepsilon > 1$ is less interesting.) Then we proceed to either option 2A or 2B below.

2. In the second step carry out either of the following options:

2A. For each sphere partition $\gamma$ of $U$, calculate $h(\gamma)$. Select a sphere $\gamma$ minimizing $h$ and use it to partition $T$. This partition (which will generally include some arbitrary choices for points near $\gamma$) is the output of the algorithm.

2B. For each sphere partition $\gamma$ of $U$, evaluate $\phi_\gamma(T)$ (again there will generally be some arbitrary choices for points near $\gamma$), and output the partition minimizing this quantity.

**Corollary 22** *With probability at least $1-\delta$ the value of the cut output by the algorithm (using either option 2A or 2B) is within a multiplicative factor of $1-\varepsilon$ of the optimal value.* □

The output of option 2B is always of course at least as good as that of option 2A, but it necessitates a slightly higher runtime, about $n|U|^{d+1}$; both the improvement in output quality, and the increase in runtime, are fairly slight, so either option seems reasonable. Since option 2B comes with no better guarantees than option 2A, we evaluate the runtime in terms of option 2A.

Run time of the algorithm: linear time suffices to compute the quantities $\{\alpha_u\}_{u\in T}$ required for the sampling procedures. Generation in the simplest way of the r.v.s $\{K_u\}$ used for each of the trials, requires time $O(n\log n)$. However since almost all of these coefficients are likely to equal 0, they can be generated in sublinear expected time (without explicitly listing the zero-valued r.v.s).

Finally, time $|T_i||U|^{d+1}$ suffices to evaluate all spherical partitions $\gamma$ of $U$ with respect to each of the samples $T_i$ $(1 \leq i \leq t)$, and so time $|U|^{d+2}$ suffices to compute $h(\gamma)$. $|U|$ is bounded by the sum of the variables $K_i$, and the expectation of this sum is $st$. Since the $K_i$ are exponentially distributed and independent, the distribution of their sum has exponential tails, hence the expected runtime of the computation is $O((\varepsilon^{-2}\log(n^{d+1}\delta^{-1}))^{d+2})$. (Alternatively we can allow another probability $\delta$ of the algorithm failing, and simply restart it whenever the $|U|$ is too large.) Recall that due to section 3.6, $d$ may be assumed here to be the minimum of $O(\varepsilon^{-2}\log n)$ and the original dimension.

In conclusion, we have shown (neglecting the cost of the linear algebra that may be required at initialization to isometrically reduce the dimension of the problem to $n-1$):

**Theorem 23** *Given a clustering problem for cost function $\phi = \ell_2^2$ on $n$ points in dimension $d$, the above algorithm runs in time*

$$O((\varepsilon^{-2}\log(n^{d'+1}\delta^{-1}))^{d'+2})$$

*(where $d' = \min\{d, O(\varepsilon^{-2}\log n)\}$), and with probability at least $1-\delta$ outputs a clustering with a cut cost within a factor of $1-\varepsilon$ of optimum.* □

If we simplify somewhat by assuming $\varepsilon$ and $\delta$ constant, this gives a runtime of $O(((d+1)\log n)^{d+2})$. In the worst case this is $n^{O(\log\log n)}$. If $d \in o(\frac{\log n}{\log\log n})$ then the runtime of the algorithm is linear, and is dominated by the time to compute the sampling probabilities $\alpha_u$.

## 3.9 Few points of a good clustering are mislabelled

Multiplicative approximation of the maximum cut value, obtained above, does not imply multiplicative approximation of the min cluster value; however, it does imply, as we now show, that the minimum cluster has been determined correctly except for a small fraction of misidentified points. This is the main goal of an automated clustering method, since a realistic classification problem will generally be only roughly modeled by a simple criterion such as $\phi$, so that there is little reason to think that the decree of the optimum $\phi$-clustering, concerning points at the fringes of the clusters, carries much meaning. Nonetheless, multiplicative approximation of the min cluster would be a stronger result and remains, at the least, an outstanding theoretical problem.

Consider a cut $(S, \bar{S})$ of $T$ (we will have in mind that this is the optimal clustering, although this plays no role in the following arguments, only in their application in note 27). Let $G$ denote the graph containing all edges that do not cross this cut. Consider any other clustering $(S', \bar{S}')$ of $T$; and let $G'$ be the graph containing all edges that do not cross the second cut. Hence $\phi_G(T)$ and $\phi_{G'}(T)$ are the costs of the two clusterings. Define the distance $\Delta(S, S')$ between the two cuts to be $\frac{1}{w_T}\min\{w_{S\cap S'} + w_{\bar{S}\cap\bar{S}'}, w_{S\cap\bar{S}'} + w_{\bar{S}\cap S'}\}$. In case the points have unit weights this is the same (after scaling) as the Hamming-type distance of definition 5.

**Lemma 24** $\phi(T) \leq (9 + \frac{4}{\Delta(S,S')})(\phi_G(T) + \phi_{G'}(T))$.

This bound is not far from the truth; it is easy to construct an example with $\phi_G(T) = 0$ and $\phi(T) = \frac{1}{\Delta(S,S')}\phi_{G'}(T)$.

The lemma immediately implies:

**Theorem 25** *Let a clustering $(S, \bar{S})$ be given (with corresponding graph $G$). Let $c$ be any positive number. Let $0 < \varepsilon \leq \frac{1}{17(1+c)}$ and suppose that $\phi_G(T) \leq \varepsilon\phi(T)$. Let $(S', \bar{S}')$ be another clustering (with corresponding graph $G'$), such that $\Delta(S, S') \geq \frac{4(1+c)\varepsilon}{1-9(1+c)\varepsilon}$, i.e. the two clusterings differ in the labelling of a substantial part of the data. Then $\phi_{G'}(T) \geq c\varepsilon\phi(T)$.*

So $(S', \bar{S}')$ is a worse clustering than $(S, \bar{S})$, by a factor of at least $c$.

To simplify the above theorem, in the particular case $c = 1$ we can for example say:

**Corollary 26** *If $\phi_G(T) \leq \varepsilon\phi(T) \leq \frac{1}{34}\phi(T)$ and $\Delta(S, S') \geq 17\varepsilon$, then $\phi_{G'}(T) \geq \varepsilon\phi(T)$.*

Theorem 25 and this corollary are a precise formulation, for $\phi$, of the intuitive statement that should hold for any useful clustering criterion: that if the data set can be clustered very well, then that clustering must be "meaningful" or "nearly unique" — there cannot be an entirely different way of achieving a clustering of similar quality.

**Note 27** The algorithmic implication is as follows: the algorithm of the preceding section can be run to identify a $(1-\varepsilon)$ multiplicative approximation of max cut. If the value of that cut is less than $(1-\varepsilon)\phi(T)$, then we can also obtain a multiplicative approximation of the min cluster by using $\varepsilon^2$ in place of $\varepsilon$ in the algorithm. On the other hand if it is found that the max cut value is at least $(1-\varepsilon)\phi(T)$, then we can conclude that the clustering $S'$ so identified is very close to the optimal clustering $S$, specifically $\Delta(S, S') < 17\varepsilon$.

**Proof of lemma 24:** We begin with a triangle inequality concerning $\phi$. Given two weighted collections of points $A$ and $B$, let

$$r_{AB} = [\frac{\phi(A \cup B) - \phi(A) - \phi(B)}{w_A w_B}]^{1/2}.$$

From equation 16 we read:

$$r_{AB}^2 = \rho_{A,B}^2 + \frac{\phi(A)}{w_A^2} + \frac{\phi(B)}{w_B^2}$$

Now, we show the triangle inequality:

$$\begin{aligned}
r_{AB} + r_{BC} &= [\rho_{A,B}^2 + \frac{\phi(A)}{w_A^2} + \frac{\phi(B)}{w_B^2}]^{1/2} + [\rho_{B,C}^2 + \frac{\phi(B)}{w_B^2} + \frac{\phi(C)}{w_C^2}]^{1/2} \\
&\geq [\rho_{A,B}^2 + \frac{\phi(A)}{w_A^2}]^{1/2} + [\rho_{B,C}^2 + \frac{\phi(C)}{w_C^2}]^{1/2} \\
&= [\rho_{A,B}^2 + \frac{\phi(A)}{w_A^2} + \rho_{B,C}^2 + \frac{\phi(C)}{w_C^2} + 2[(\rho_{A,B}^2 + \frac{\phi(A)}{w_A^2})(\rho_{B,C}^2 + \frac{\phi(C)}{w_C^2})]^{1/2}]^{1/2} \\
&\geq [\rho_{A,B}^2 + \rho_{B,C}^2 + 2\rho_{A,B}\rho_{B,C} + \frac{\phi(A)}{w_A^2} + \frac{\phi(C)}{w_C^2}]^{1/2} \\
&= [(\rho_{A,B} + \rho_{B,C})^2 + \frac{\phi(A)}{w_A^2} + \frac{\phi(C)}{w_C^2}]^{1/2} \\
&\geq [\rho_{A,C}^2 + \frac{\phi(A)}{w_A^2} + \frac{\phi(C)}{w_C^2}]^{1/2} \\
&= r_{AC}
\end{aligned}$$

This does not make $r$ a metric on clusters, because $r_{AA} > 0$ (unless $A$ is a point).

We conclude also that

$$r_{AB}^2 + r_{BC}^2 \geq r_{AC}^2/2. \tag{26}$$

Now, consider the four subcollections defined by the cuts $S$ and $S'$: $A = S \cap S'$, $B = S \cap \bar{S}'$, $C = \bar{S} \cap S'$, and $D = \bar{S} \cap \bar{S}'$. In these terms,

$$\phi_G(T) = \phi(A \cup B) + \phi(C \cup D)$$

$$\phi_{G'}(T) = \phi(A \cup C) + \phi(B \cup D)$$

$$\Delta(S, S') = \frac{1}{w_T} \min\{w_A + w_D, w_B + w_C\}$$

We assume without loss of generality that

$$\Delta(S, S') = \frac{w_B + w_C}{w_T}.$$

Let $E \in \{B, C\}$ be the heavier of the two, so that

$$w_E = \max\{w_B, w_C\} \geq \Delta(S, S')w_T/2$$

Similarly let $F \in \{A, D\}$ be the heavier of the two, so that

$$w_F = \max\{w_A, w_D\} \geq w_T/4$$

Employ equation 17 to write:

$$\begin{aligned}
\phi(T) &= \phi(A) + \phi(B) + \phi(C) + \phi(D) \\
&\quad + (\phi(A \cup B) - \phi(A) - \phi(B)) + (\phi(C \cup D) - \phi(C) - \phi(D)) \\
&\quad + (\phi(A \cup C) - \phi(A) - \phi(C)) + (\phi(B \cup D) - \phi(B) - \phi(D))
\end{aligned}$$

$$
\begin{aligned}
& +(\phi(A\cup D)-\phi(A)-\phi(D))+(\phi(B\cup C)-\phi(B)-\phi(C))\\
\leq\ & \phi(A\cup B)+\phi(C\cup D)+\phi(A\cup C)+\phi(B\cup D)\\
& +(\phi(A\cup D)-\phi(A)-\phi(D))+(\phi(B\cup C)-\phi(B)-\phi(C))\\
=\ & \phi(A\cup B)+\phi(C\cup D)+\phi(A\cup C)+\phi(B\cup D)\\
& +w_Aw_Dr_{AD}^2+w_Bw_Cr_{BC}^2
\end{aligned}
$$

which, by inequality 26, is

$$
\begin{aligned}
\leq\ & \phi(A\cup B)+\phi(C\cup D)+\phi(A\cup C)+\phi(B\cup D)\\
& +2w_Aw_D(r_{AE}^2+r_{ED}^2)+2w_Bw_C(r_{BF}^2+r_{FC}^2)\\
=\ & \phi(A\cup B)+\phi(C\cup D)+\phi(A\cup C)+\phi(B\cup D)\\
& +2w_Aw_D(\frac{\phi(A\cup E)-\phi(A)-\phi(E)}{w_Aw_E}+\frac{\phi(E\cup D)-\phi(E)-\phi(D)}{w_Ew_D})\\
& +2w_Bw_C(\frac{\phi(B\cup F)-\phi(B)-\phi(F)}{w_Bw_F}+\frac{\phi(F\cup C)-\phi(F)-\phi(C)}{w_Fw_C}))\\
\leq\ & \phi(A\cup B)+\phi(C\cup D)+\phi(A\cup C)+\phi(B\cup D)\\
& +\frac{2w_D}{w_E}\phi(A\cup E)+\frac{2w_A}{w_E}\phi(E\cup D)+\frac{2w_C}{w_F}\phi(B\cup F)+\frac{2w_B}{w_F}\phi(F\cup C)
\end{aligned}
$$

Every argument of $\phi$ in the last line is the same as one of those in the preceding line. Both $\frac{2w_D}{w_E}$ and $\frac{2w_A}{w_E}$ are bounded above by $\frac{2w_T}{w_E}\leq\frac{4}{\Delta(S,S')}$, while both $\frac{2w_C}{w_F}$ and $\frac{2w_B}{w_F}$ are bounded above by $\frac{2w_T}{w_F}\leq 8$. Hence

$$\phi(T)\leq(9+\frac{4}{\Delta(S,S')})(\phi(A\cup B)+\phi(C\cup D)+\phi(A\cup C)+\phi(B\cup D))\leq(9+\frac{4}{\Delta(S,S')})(\phi_G(T)+\phi_{G'}(T)).$$

□

### 3.10 Randomized approximation algorithm for $k$-partitions

The argument of section 3.8 goes through with little change, with $(\frac{12en\binom{k}{2}}{(d+2)k})^{(d+2)k}$ (or $O(n^{(d+2)k})$ for fixed $d$ and $k$) replacing $n^{d+1}$ in the union bound over partitions. Specifically, take $a=636\varepsilon^{-2}$ as before; $s=4$ as before; now $b=(\log((\frac{12en\binom{k}{2}}{(d+2)k})^{(d+2)k}\delta^{-1}))/D(1/2;2/3)$; and use the sampling procedure to reduce the problem to one on fewer points, which is then solved using the deterministic procedure of section 3.5.

**Proposition 28** *Given a clustering problem for cost function $\phi=\ell_2^2$ on $n$ points in dimension $d$, this algorithm runs in time*

$$F(O(\varepsilon^{-2}\log((\frac{n\binom{k}{2}}{(d'+2)k})^{(d'+2)k}\delta^{-1})),d',k).$$

*(where $F$ is as defined for proposition 14, and $d'=\min\{d,O(\varepsilon^{-2}\log n)\}$), and with probability at least $1-\delta$ outputs a clustering into $k$ clusters, with a cut cost within a factor of $1-\varepsilon$ of optimum.* □

If $\varepsilon$, $\delta$ and $k$ are fixed then, just as for $k=2$, this runtime is $n^{O(\log\log n)}$ in the worst case, and linear if $d\in o(\frac{\log n}{\log\log n})$.

### 3.11 Examples

We show two examples of optimally 2-partitioned point sets in the plane. In each case the boundary circle $\nu$ between the clusters has been marked. These examples were produced using a program written by the author in the "Maple" programming language, implementing a version of the deterministic algorithm.

In figure 1, two mildly cohesive clusters of points are evident to the eye; $\phi$ yields the sensible partition. In the figure 2 $\phi$ again identifies the basic structure, in this case a cohesive central collection and several outliers.

Figure 1: in hardcopy only

Figure 2: in hardcopy only

# 4 The cases $\phi = \ell_1$, $\ell_2$, and other cost functions

Having obtained a clustering algorithm for $\ell_2^2$, we are positioned to take advantage of the generosity of $\ell_2^2$ as a host space.

By a cost function on a set of points $T$ we mean a function $\lambda : T^2 \to \mathbb{R}$ which is symmetric, nonnegative, and 0 on the diagonal. An embedding of $(T, \lambda)$ in $(T', \lambda')$ with distortion $C$ is a map $\iota : T \to T'$ such that $\sup_{a,b,c,d \in T}(\frac{\lambda'(\iota(a),\iota(b))\lambda(c,d)}{\lambda(a,b)\lambda'(\iota(c),\iota(d))}) = C$. If $C = 1$ we say $\iota$ is isometric (regardless of whether the domain and range are metric spaces). We will abbreviate by writing simply $\ell_1, \ell_2$ or $\ell_2^2$ when all that matters is that the dimension of the space is finite. Note that for $\ell_2$ and $\ell_2^2$ no dimension beyond $n - 1$ need be considered. A space is finite if $T$ is finite.

**Theorem 29 [Linial, London and Rabinovich] (a)** *There is an algorithm which, given a cost function $\lambda$ on a set of $n$ points $T$, identifies a minimum-distortion embedding of $(T, \lambda)$ in $\ell_2^2$, in time polyomial in $n$.* **(b)** *Every finite $\ell_1$ or $\ell_2$ space is isometrically embeddable in $\ell_2^2$.*

For completeness we provide a proof of this theorem in the appendix.

Hence our approximation scheme solves also the cases $\phi = \ell_1$ and $\phi = \ell_2$ in the same asymptotic runtime (i.e. $n^{O(\log \log n)}$ for fixed $\varepsilon$, $\delta$ and $k$) guaranteed for the case $\ell_2^2$. Note that this does not supply a way of taking advantage of an initially low dimension to obtain an improved runtime.

Finally note that whether or not a given cost function is known to be $\ell_2^2$-embeddable, one may solve the PSD program, and provide a good clustering opportunistically if a low distortion embedding exists.

# 5 Discussion and other objective functions for clustering

Some interesting objective functions for clustering do not fall within the framework discussed in this paper, of the sum of a cost function over all intra-cluster pairs of points.

One such criterion which has attracted considerable attention is $\psi(S) = |S|\mathrm{Var}(S) = \sum_{v \in S} \rho_{v,S}^2 = \phi(S)/|S|$ (in this section we let $\phi = \ell_2^2$). Clustering so as to minimize $\sum \psi(S_i)$ (sometimes known as "sum of squares minimization") appears to have been discussed first for one dimension by Fisher in 1958 [23] and for higher dimensions by Ward in 1963 [72] and Shlezinger in 1965 [66]; a partial list of subsequent literature is [20, 37, 28, 51, 63, 6, 41, 11, 35, 31, 16], and some surveys touching on the subject are [27, 55]. (Point weights appear not to have been discussed in this literature, but can be accomodated without harm to any existing result.) The regions containing optimal clusters relative to the $\psi$ criterion are Voronoi cells centered on the centers of gravity of the clusters; the $\phi$ and $\psi$ criteria generally lead to quite different kinds of optimum partitions. Which criterion, if either, is preferable will depend on the application domain. The restriction that optimal regions for the $\psi$ criterion must be convex can be regarded as either an advantage or a limitation of the criterion. An examination of the examples in section 3.11 may help clarify the relative advantages of the two criteria. The first example would be partitioned in the same manner as by $\phi$; but the second example would be bisected by a line, which would not reflect the structure of the data set.

On the other hand, it is not hard to provide an example in which $\psi$ performs more "reasonably" than $\phi$. If in the first example, one of the populations was much more numerous than the other (without changing their locations), $\phi$ would "misbehave" by including in the small cluster some of the nearer points of the larger population.

The earliest related discussion we are aware of is by Neyman; it appears that his proposed clustering criterion corresponds to the function $\sum_{v \in S} \rho_{v,S}$ [56]. We are not aware of any existing algorithmic work specifically concerning this criterion. However, just as for any criterion of the form $\sum_{v \in S} g(\rho_{v,S})$ ($g$ monotone increasing), optimal cluster regions must be Voronoi cells, so an exhaustive examination of such partitions will find an optimal partition in time $O(n^{d+1})$.

It should be noted that the Johnson-Lindenstrauss dimension-reduction step is useful for both of the above objective functions, since, for exactly the same reasons described earlier in the paper, it reduces the effective dimension of any $1 \pm \varepsilon$ approximation problem, to $O(\varepsilon^{-2} \log n)$.

Humans have a complex lexicon for describing patterns. It is easy to "fool" any simple mathematical criterion, in the sense of providing a data set for which the optimal clustering does not correspond to the way a person would "typically" describe the data set. However it is difficult of course for a human to analyze a large data set in a high dimensional space (where "high" is anything more than 2 or perhaps 3). A similar problem arises for any formal clustering criterion which, though perhaps highly suited to a particular situation, lacks an efficient clustering algorithm. In such situations, good algorithms for clustering according to simple cost functions can perform a useful service by digesting large volumes of data, possibly in high dimensional spaces, and providing reasonable candidate partitions. Given a complex data set, one may wish to use a fast algorithm to obtain a partition into a number of clusters that is small, but larger than needed (or anticipated) for the final solution; these clusters can then be examined and joined into fewer clusters either by more complex automated methods or by hand.

In many applications there is no natural sense in which lengths along the different axes are comparable. In addition, some of the measurements may be correlated. In such cases use of the Euclidean distance is questionable. However, given the availability of efficient algorithms and keeping in mind that an automatically computed clustering may be only a starting point for further optimization, rather than a finished product, it is worth exploring heuristics for transforming inputs into a form in which clustering by $\phi$ can serve as a useful automated first step. One simple transformation of the data is to normalize the scale of each dimension so that all variances are equal. Another approach is to search for a scaling of the axes, or more generally a linear transformation of the space, for which the quality of the minimum partition ratio $((\sum \phi(S_i))/\phi(T))$ is maximized. This may capture (in a limited sense) the "unavoidable" clusterings of the data.

Another goal for clustering algorithms is to perform accurate estimation for a mixture model (i.e. a collection of multi-peak probability densities). It remains to be seen whether the methods of this paper are useful for this purpose.

## Acknowledgments

Thanks to V. Vazirani, N. Linial, A. Wigderson, A. Magen, I. Muchnik, C. Kenyon, P. Indyk, U. Vazirani and D. Klain for very helpful discussions.

## References

[1] N. Alon, L. Babai, and A. Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *J. Algorithms*, 7:567–583, 1986.

[2] P. Arabie, L. J. Hubert, and G. De Soete, editors. *Clustering and Classification.* World Scientific, 1996.

[3] B. Aronov, B. Chazelle, H. Edelsbrunner, L. J. Guibas, M. Sharir, and R. Wenger. Points and triangles in the plane and halving planes in space. *Discrete Comput. Geom.*, 6:435–442, 1991.

[4] S. Arora, D. Karger, and M. Karpinski. Polynomial time approximation schemes for dense instances of NP-hard problems. In *27th Annual ACM Symposium on the Theory of Computing*, pages 284–293, Las Vegas, 1995.

[5] I. Bárány, Z. Füredi, and L. Lovász. On the number of halving planes. In *Proc. 5'th Ann. Symp. Computational Geometry*, pages 140–144, 1989.

[6] J. P. Benzécri. Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Dannées*, VII(2):209–218, 1982.

[7] M. Bern and D. Eppstein. Approximation algorithms for geometric problems. In D. Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, pages 296–345. PWS Publishing, 1996.

[8] E. Boros and P. L. Hammer. On clustering problems with connected optima in Euclidean spaces. *Discrete Mathematics*, 75:81–88, 1989.

[9] B. Chor, O. Goldreich, J. Hastad, J. Friedman, S. Rudich, and R. Smolenski. The bit extraction problem or t-resilient functions. In *Proceedings of the 26th Annual Symposium on Foundations of Computer Science*, pages 396–407, 1985.

[10] R. M. Cormack. A review of classification. *J. Roy. Stat. Soc. A*, 134:321–367, 1971.

[11] H. E. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1:7–24, 1984.

[12] W. Fernandez de la Vega. Max-cut has a randomized approximation scheme in dense graphs. *Random Structures and Algorithms*, 8(3):187–198, 1996.

[13] W. Fernandez de la Vega and M. Karpinski. Polynomial time approximation of dense weighted instances of max-cut. manuscript.

[14] W. Fernandez de la Vega and C. Kenyon. A randomized approximation scheme for metric max-cut. In *39th Annual Symposium on Foundations of Computer Science*, pages 468–471, Palo Alto, 1998. IEEE.

[15] T. K. Dey and H. Edelsbrunner. Counting triangle crossings and halving planes. *Discrete Comput. Geom.*, 12:281–289, 1994.

[16] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[17] R. M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6:899–929, 1978.

[18] H. Edelsbrunner. *Algorithms in Combinatorial Geometry.* Springer-Verlag, 1987.

[19] H. Edelsbrunner and E. Welzl. On the number of line separations of a finite set in the plane. *J. Combin. Theory Ser. A*, 38:15–29, 1985.

[20] A. W. F. Edwards and L. L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21:362–375, 1965.

[21] D. Eppstein. Improved bounds for intersecting triangles and halving planes. *J. Combin. Theory Ser. A*, 62:176–182, 1993.

[22] P. Erdös, L. Lovász, A. Simmons, and E. Strauss. Dissection graphs of planar point sets. In J. Srivastava, editor, *A Survey of Combinatorial Theory*, pages 139–149. North-Holland, 1973.

[23] W. D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53:789–798, 1958.

[24] W. D. Fisher. *Clustering and Aggregation in Economics.* Johns Hopkins Press, 1969.

[25] R. M. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete graph problems. *Theor. Comput. Sci.*, 1:237–267, 1976.

[26] C. Glymour, D. Madigan, D. Pregibon, and P. Smyth. Statistical inference and data mining. *Communications of the ACM*, 39(11), November 1996.

[27] A. D. Gordon. *Classification.* Chapman and Hall, 1981.

[28] J. C. Gower. A comparison of some methods of cluster analysis. *Biometrics*, 23:623–637, 1967.

[29] C. Greene and D. J. Kleitman. Proof techniques in the theory of finite sets. In G.-C. Rota, editor, *Studies in Combinatorics.* The Mathematical Association of America, 1978.

[30] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization.* Springer-Verlag, 1988.

[31] P. Hansen, B. Jaumard, and N. Mladenović. Minimum sum of squares clustering in a low dimensional space. *Journal of Classification*, 1998. to appear.

[32] E. F. Harding. The number of partitions of a set of $n$ points in $k$ dimensions induced by hyperplanes. *Proc. Edinburgh Mathematical Society, series II*, 15:285–289, 1967.

[33] J. A. Hartigan. *Clustering Algorithms.* Wiley, 1975.

[34] M. Inaba. *Geometric Clustering on Feature Manifold.* PhD thesis, University of Tokyo, 1999.

[35] M. Inaba, N. Katoh, and H. Imai. Applications of weighted Voronoi diagrams and randomization to variance-based $k$-clustering. In *Proc. 10'th ACM Symp. Comp. Geom.*, pages 332–339, 1994.

[36] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice-Hall, 1988.

[37] R. C. Jancey. Multidimensional group analysis. *Australian Journal of Botany*, 14:127–130, 1966.

[38] N. Jardine and R. Sibson. *Mathematical Taxonomy.* Wiley, 1971.

[39] A. Joffe. On a set of almost deterministic $k$-independent random variables. *Ann. Probability*, 2:161–162, 1974.

[40] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.*, 26:189–206, 1984.

[41] J. Juan. Programme de classification hiérarchique par l'algorithme de la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Dannées*, VII(2):229–225, 1982.

[42] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[43] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, 1994.

[44] L. G. Khachiyan. A polynomial algorithm for linear programming. *Doklady Academiia Nauk USSR*, 244:1093–1096, 1979. In Russian. Translation in *Soviet Mathematics Doklady* 20:191-194, 1979.

[45] N. E. Kiseleva, I. B. Muchnik, and S. G. Novikov. Stratified samples in the problem of representative sampling. *Automation and Remote Control*, 47(5):684–693, 1986.

[46] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. 9'th ACM-SIAM Symp. on Discr. Alg.*, 1998.

[47] Daphne Koller and Nimrod Megiddo. Constructing small sample spaces satisfying given constraints. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on the Theory of Computing*, pages 268–277, San Diego, California, 16–18 May 1993.

[48] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. In *Proc. 35'th Annual Symposium on Foundations of Computer Science*, pages 577–591. IEEE Press, 1994.

[49] L. Lovász. On the number of halving lines. *Ann. Univ. Sci. Budapest Eötvös Sect. Math.*, 14:107–108, 1971.

[50] M. Luby. A simple parallel algorithm for the maximal independent set problem. *SIAM J. Comput.*, 15(4):1036–1053, Nov. 1986.

[51] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the 5'th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. U. California Press, 1967.

[52] J. Matousek. Geometric set systems. to appear in Proc. 2'nd European Math. Congress.

[53] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, 1988.

[54] B. Mirkin. *Mathematical Classification and Clustering.* Kluwer, 1996.

[55] B. G. Mirkin and I. Muchnik. Clustering and multidimensional scaling in Russia (1960-1990): a review. In P. Arabie, L. J. Hubert, and G. De Soete, editors, *Clustering and Classification*, pages 295–339. World Scientific, 1996.

[56] J. Neyman. On the two different aspects of the representative model: the method of stratified sampling and the method of purposive selection. *J. R. Statis. Soc.*, 97:558–606, 1934.

[57] J. Pach, W. Steiger, and E. Szemerédi. An upper bound on the number of planar k-sets. *Discrete Comput. Geom.*, 7:109–123, 1992.

[58] D. Pollard. Quantization and the method of $k$-means. *IEEE Trans. Inform. Theory*, IT-28:199–205, March 1982.

[59] E. G. Rees. *Notes on Geometry.* Springer Verlag, 1983.

[60] M. J. Sabin and R.M. Gray. Global convergence and empirical consistency of the generalized Lloyd algorithm. *IEEE Trans. Inform. Theory*, IT-32(2):148–155, 1986.

[61] N. Sauer. On the density of families of sets. *J. Combin. Theory Ser. A*, 13:145–147, 1972.

[62] L. Schläfli. Theorie der vielfachen Kontinuität (1901). In *Gesammelte Mathematische Abhandlungen*, volume I, pages 167–387. Verlag Birkhäuser, 1950.

[63] A. J. Scott and M. J. Symons. On the Edwards and Cavalli-Sforza method of cluster analysis. *Biometrics*, 27:217–219, 1971.

[64] S. Shelah. A combinatorial problem: stability and order for models and theories in infinitary languages. *Pacific J. Mathematics*, 41:247–261, 1972.

[65] J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Conf. Computer Vision and Pattern Recognition*, 1997.

[66] M. I. Shlezinger. On unsupervised pattern recognition. In V. M. Glushkov, editor, *Reading Automata*, pages 62–70. Naukova Dumka, 1965.

[67] R. R. Sokal and P. H. A. Sneath. *Principles of Numerical Taxonomy.* Freeman, 1963.

[68] R. C. Tryon and D. E. Bailey. *Cluster Analysis.* McGraw-Hill, 1970.

[69] J. van Ryzin, editor. *Classification and Clustering.* Academic Press, 1977.

[70] V. N. Vapnik. *Estimation of dependencies based on empirical data.* Springer Verlag, 1982.

[71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[72] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[73] H. E. Warren. Lower bounds for approximation by nonlinear manifolds. *Trans. Amer. Math. Soc.*, 133:167–178, 1968.

# 6 Appendix: proof of the [LLR] $\ell_2^2$-embedding theorem

Theorem 29(a): With any embedding $\iota$ of $T$ in $\mathbb{R}$ associate the matrix $B$ whose rows represent the images of the points. Let $A$ be the positive semidefinite matrix $A = BB^t$. The square of the distance between the images of points $i, j \in T$ is $A_{ii} + A_{jj} - 2A_{ij}$. The problem of obtaining an optimal embedding is, then, that of optimizing the semidefinite program

$\min C$  
subject to:  
$A$ is positive semidefinite  
$\lambda_{ij} \leq A_{ii} + A_{jj} - 2A_{ij} \leq C\lambda_{ij} \quad \forall i, j \in T$

This program can be solved in polynomial time using the ellipsoid method [44, 30]. Once an optimal solution $A$ is found an embedding is identified by factoring $A$ into $BB^t$.

Theorem 29(b): $\ell_1 \rightarrow \ell_2^2$: It is sufficient to consider the case $(\mathbb{R}, \ell_1) \rightarrow \ell_2^2$. Let $n$ be the number of points. An embedding $(\mathbb{R}, \ell_1) \rightarrow (\mathbb{R}^{n-1}, \ell_2^2)$ is achieved by sorting the points and then letting the $i$'th coordinate $(1 \leq i \leq n-1)$ be 0 for the first $i$ points, and $\lambda^{1/2}(i, i+1)$ for the remaining points.

$(\mathbb{R}^m, \ell_2) \rightarrow \ell_1$: Let $\mathbb{R}_M^{S^{m-1}}$ denote the space of measurable functions on the sphere $S^{m-1}$. The metric space $(\mathbb{R}_M^{S^{m-1}}, \ell_1)$ is defined by letting $\int |f-g| d\mu$ be the distance between $f, g \in \mathbb{R}_M^{S^{m-1}}$, where $\mu$ is invariant under rotations. It is easily verified that the map $\iota_1 : (\mathbb{R}^m, \ell_2) \rightarrow (\mathbb{R}_M^{S^{m-1}}, \ell_1)$, carrying $v \in \mathbb{R}^m$ to the function $\iota_1^v$ defined by $\iota_1^v(x) = v \cdot x$, is isometric. For a finite dimensional version $\iota_2$ of this construction, partition $S^{m-1}$ into open regions such that for every region $U$, every $x, y \in U$ and every $u, v \in T$, $(\iota_1^u(x) - \iota_1^v(x))(\iota_1^u(y) - \iota_1^v(y)) > 0$. At most $\min\{\Phi_m(\binom{n}{2}), n!\}$ regions are required. To each such region $U$ associate one coordinate of the embedding, and set $\iota_2^v(U) = \int_{x \in U} \iota_1^v(x) d\mu$. $\square$