Data reduction for weighted and outlier-resistant clustering

Dan Feldman* Leonard J. Schulman†

September 21, 2011

Abstract

Statistical data frequently includes outliers; these can distort the results of estimation and optimization problems. For this reason, loss functions which deemphasize the effect of outliers are widely used by statisticians. However, there are relatively few algorithmic results about clustering with outliers.

For instance, the *k-median with outliers* problem uses a loss function $f_{c_1,\ldots,c_k}(x)$ which is equal to the minimum of a penalties h, and the least distance between the data point x and a center c_i. The loss-minimizing choice of $\{c_1,\ldots,c_k\}$ is an outlier-resistant clustering of the data. This problem is also a natural special case of the *k-median with penalties* problem considered by [Charikar, Khuller, Mount and Narasimhan SODA'01].

The essential challenge encountered here is data reduction for the *weighted k-median* problem. We solve this problem, which was previously solved only in one dimension ([Har-Peled FSTTCS'06], [Feldman, Fiat and Sharir FOCS'06]). As a corollary, we also achieve improved data reduction for the k-line-median problem.

*Caltech, Pasadena CA 91125. Email: `dannyf@caltech.edu`.

†Caltech, Pasadena CA 91125. Email: `schulman@caltech.edu`. Work supported in part by the NSF.

1 Introduction

1.1 Weighted optimization problems

We show how to perform data reduction for a variety of problems in optimization and statistical estimation. The problems are of the following form: a metric space M and a family of functions F are specified. Then, given a set P of n points in M, the optimization problem is to find an f which is a $(1+\varepsilon)$-approximate minimizer of $f(P)$ among all $f \in F$, where $f(P) = \sum_{p \in P} f(p)$. We focus particularly on families F which are appropriate for minimization of a loss function (e.g., max likelihood estimation) in statistical inference. A key case we treat pertains to the problem of clustering with outliers:

k-median with outliers in ℓ_2^d: (ℓ_2^d is $\mathbb{R}^d$ with the Euclidean metric.) Here one is interested in modeling data as being distributed about k centers, with points that are beyond a threshold distance h being considered as outliers.

$$\begin{aligned} F_{d,k}^{\text{out}} &: \mathbb{R}^d \to \mathbb{R}_+ \qquad \text{where } \mathbb{R}_+ = \text{nonnegative reals} \\ F_{d,k}^{\text{out}} &= \{f_{C,h}\}_{C \subseteq \mathbb{R}^d, |C|=k, h>0} \\ f_{C,h}(p) &= \min\{h, \min_{c \in C} \|p - c\|\} \end{aligned}$$

Note that, conditional on a point being treated as an outlier (assigned value h), it has no further effect on the cost-minimizing choice of the centers C. This way of formalizing the treatment of outliers is a slightly simplified form of the *Tukey biweight* loss function used by statisticians to perform outlier-resistant estimation. It is also a special case of the *k-median with penalties* problem considered by Charikar, Khuller, Mount and Narasimhan [14] (the distinction being that they allow each point p a custom penalty $h(p)$ for being an outlier).

Data reduction means the replacement of the input P (implicitly, the uniform probability distribution on P) by a probability distribution ν on a much smaller set A, such that for all $f \in F$,

$$f(\nu) = (1 \pm \varepsilon) f(P)/n, \quad \text{where} \quad f(\nu) = \sum_{p \in A} \nu(p) f(p).$$

(A, ν) is often called an ε-approximation or core-set for the data P w.r.t. the family of functions F. In recent years a substantial body of work has gone into providing (or showing non-existence of) data reduction for various problems, because one can then run a relatively inefficient optimization algorithm (possibly even exhaustive search) on the core-set.

Results, in brief: We show for any constant k how to efficiently construct core-sets of cardinality $O(\Delta \log^2 n)$ for certain types of k-clustering problems; here Δ is a Vapnik-Chervonenkis-type measure of the combinatorial complexity of the clustering problem. (If P is in ℓ_2^d then $\Delta \in O(d)$; if P is in a finite metric space then $\Delta \in O(\log n)$.) Our algorithm is randomized and runs in time linear in the input size. These clustering problems include the well-known k-median, k-means, etc., but go beyond these to include treatment of outliers and also variations such as k-line-medians. The key obstacle we overcome, which has not been overcome previously except in one dimension, is that of handling "weights" on the clustering centers.

Using the map-reduce technique [29], our core-sets imply polylogarithmic space and polylogarithmic update-time algorithms for clustering streaming data with outliers; this is apparently the first result of this type. Similarly, the techniques can be adapted to parallelization[24, 5].

The ability to handle weights is what allows us to provide core-sets for the outliers family $F_{d,k}^{\text{out}}$ defined above; the following is another example which explains what we mean by weights and is also, from the mathematical point of view, perhaps the central example to keep in mind:

Weighted k-median in ℓ_2^d: Here one is interested in modeling possible heterogeneity among cluster centers. This is natural, for example, in the context of *mixture models* in which the components of the mixture have varying standard deviations. Heterogeneity also arises naturally from geometric considerations, in the reduction of the (unweighted) k-line-median problem to weighted k-median [28, 22].

$$\begin{aligned} F_{d,k}^{\text{wght}} &: \mathbb{R}^d \to \mathbb{R}_+ \\ F_{d,k}^{\text{wght}} &= \{f_C\}_{C \subseteq \mathbb{R}^d \times \mathbb{R}_+, |C|=k} \\ f_C(p) &= \min_{(c,w)\in C} w \cdot \|p-c\| \end{aligned}$$

Our approach is more general than is implied by these two examples, but is somewhat technical and is deferred to Section 2; our main theorem is Theorem 4.3 and in Section 5 we discuss additional examples covered by the approach. In brief, these include k-means and other finite exponents for clustering; standard "M-estimators" in robust statistics such as the Huber and Tukey loss functions; and (by the aforementioned reduction) the k-line-median problem.

Robust Statistics. As described, our approach provides core-sets for (at least) two of the most important outlier-resistant statistical estimator. Huber's estimator is used very widely [26, 33]; Zhang [50] writes that "this estimator is so satisfactory that it has been recommended for almost all situations". Hardin et al. [30] write that "Tukey's biweight has been well established as a resistant measure of location and scale for multivariate data [44, 33, 32]". Both of these estimators are types of *M-estimators*; very little is known about the computational complexity of optimizing M-estimators [42, 41, 44, 30, 26, 33] and our paper shows how to make considerable improvement in this direction.

1.2 Literature

Sampling. Data reduction by uniform sampling goes back to the foundations of statistics; the most relevant line of work for our purpose is that initiated by Vapnik and Chervonenkis [49, 43, 31, 13, 48, 40, 10, 34, 8, 9, 3, 7] (and see [45, 47]). However, for estimation of general nonnegative (esp., unbounded) loss functions, and for the design of approximation algorithms for (related) optimization problems, it is essential to design *weighted* sampling methods. This is a more recent line of work, beginning at least with [11, 16, 35, 4, 46]. There are also methods for deterministic data reduction [39, 19, 22] but the results are generally weaker and we shall not emphasize this aspect of the problem in the paper.

Clustering. The k-median problem was shown to be NP-hard by a reduction from dominating set [37]. This problem is a special case of k-clustering problems with various exponents $r > 0$, with loss function $f_C(p) = \min_{c\in C} \operatorname{dist}(p,c)^r$ for centers $C = \{c_1, \ldots, c_k\}$. The k-means problem (exponent $r = 2$) is NP-hard even for $k = 2$ [18] or in the Euclidean plane [38]. The case $r = \infty$ refers to the k-center problem $f_C(P) = \max_{p\in P} \min_{c\in C} \operatorname{dist}(p,c)$; it is NP-hard to approximate this to within a factor of 1.822 even in the Euclidean plane [20].

The current best approximation guarantee for k-median in general metrics is $(3+\varepsilon)$ [6]. When k is fixed, [23] provided a weak core-set of size independent of d for k-means that yields an algorithm that takes time $O(nd) + (k/\varepsilon^2)^{O(k/\varepsilon)}$. (A weak core-set is sufficient for optimization but not for evaluation of general queries.) Recently, this result was generalized and improved for any constant $r \geq 1$, with weak core-sets of size only linear in k [22]. Strong core-sets of size $(dk)^{O(1)}$ for the k-median problem for any constant $r > 0$ were provided in [36].

In the *k-median with penalties problem* [14], for each input point we may decide to either provide service and pay the service cost to its nearest center, or to pay the penalty. Setting all penalties to 1 gives the standard notion, which has has also been studied earlier in the context of TSP and Steiner trees, see [25, 12] and references therein. As mentioned above, this is precisely our $F^{\text{out}}_{d,k}$ problem; it is also very close to clustering with Tukey loss, see Sec. 5.

An alternative approach to handling outliers is the *robust k-median with m outliers problem* due to [14]. Here there is, besides the usual k-median formulation, an additional parameter m which is the number of points we are allowed to "discard". The problem is to place the k centers so as to minimize the sum, over the best set of $n - m$ data points, of the distance to the closest center. This is a less "continuous" way of treating outliers and, correspondingly, m enters significantly into the time complexities of algorithms. Our weighted-k-median algorithm can be used to address this problem, see Sec. 5. [14] also considered relaxing the number of discarded points, and provided a polynomial time algorithm that outputs a k-clustering serving $(1-\varepsilon)(n-m)$ points with cost within $4(1 + 1/\varepsilon)$ times the optimum cost (for $n - m$ points). Recently, [22] improved the running time for this problem to linear in n by showing that an ε-approximation of P for k balls (in particular, a small uniform random sample) is a core-set for this problem.

The *weighted k-median* problem was introduced in Har-Peled [28]; that paper provided an $O((\log n)^k)$-size core-set for this problem in one dimension, and posed the construction of core-sets in higher dimension as an open problem. The same paper proved a lower bound of $\Omega(\max\left\{(k/\varepsilon)\log(n/k), 2^k\right\})$ for the size of a core-set for weighted k-medians, even in one dimension. We do not know a stronger lower bound in higher dimension. Thus our results are optimal, as a function of n, up to a log factor.

In the k-line-median problem, the "centers" are actually lines in $\mathbb{R}^d$. This problem can be reduced to the *weighted* k-median problem in one dimension [28, 22]. Our core-set for this problem, of size $O((\varepsilon^{-1}\log n)^2)$, improves on the best previous $O((\varepsilon^{-1}\log n)^{O(k)})$. Our method also considerably simplifies, even for the one-dimensional version of the problem, the construction in [28] (which both these papers depend on).

2 Preliminaries

Let (M, dist) be the metric space in which our points (or data items) lie. Our framework depends upon a distortion (or "loss") function

$$D : M \times M \to \mathbb{R}_+.$$

We require that D satisfy the following conditions:

1. *Symmetry* condition
$$D(p, q) = D(q, p) \tag{1}$$

2. D is a function of dist, and, as a univariate function, is monotone non-decreasing.

3. *Log-Log Lipschitz Condition,* parameterized by $r \in (0, \infty)$: For all $x, \delta > 0$,
$$D(xe^{\delta}) \leq e^{r\delta} D(x). \tag{2}$$

Optimization Problem	**Metric**	**ℓ_r loss**	**Approx.**	**Time**	**Ref.**
k-Median with penalties	Arbitrary	$r = \infty$	3	$O(n^3)$	[14]
k-Median with penalties	Arbitrary	$r = 1$	4	$n^{3+O(1)}$	[14]
k-Median with penalties	Arbitrary	$r \in O(1)$	$O(1)$	$k^{O(k)} n \log(n) + nk^{O(k)} \log^2 n$	⋆⋆
k-Median with penalties	$\mathbb{R}^d$	$r \in O(1)$	$1+\varepsilon$	$ndk^{O(k)} + \left(k\varepsilon^{-2}\log(n)\right)^k$	⋆⋆
M-Estimators	Arbitrary	$r \in O(1)$	Heuristics	?	e.g.[30]
M-Estimators	$\mathbb{R}^d$	$r \in O(1)$	$1+\varepsilon$	$ndk^{O(k)} + \left(k\varepsilon^{-2}\log(n)\right)^k$	⋆⋆
M-Estimators	Arbitrary	$r \in O(1)$	$O(1)$	$k^{O(k)} n \log(n) + nk^{O(k)} \log^2 n$	⋆⋆
Robust k-Median with m outliers	Arbitrary	$r = 1$	$O(1)$	$O(k^2(k+m)^2 n^3 \log n)$	[15]
Robust 2-Median with m outliers	$\mathbb{R}^2$	$r = \infty$	1	$O(nm^7 \log^3 n)$	[1]
Robust 4-Median with m outliers	$\mathbb{R}^2$	$r = \infty$	1	$nm^{O(1)} \log n$	[1]
Robust 5-Median with m outliers	$\mathbb{R}^2$	$r = \infty$	1	$nm^{O(1)} \log^5 n)$	[1]
Robust k-Median with m outliers	Arbitrary	$r = \infty$	3	$O(n^3)$	[14]
Robust k-Median with m outliers	$\mathbb{R}^d$	$r \in O(1)$	$1+\varepsilon$	$nd(m+k)^{O(m+k)}$ $+(\varepsilon^{-1} k \log n)^{O(1)}$	⋆⋆
Robust k-Median with m outliers	Arbitrary	$r \in O(1)$	$O(1)$	$n\log(n)(m+k)^{O(m+k)}$ $+(k \log n)^{O(1)}$	⋆⋆
k-Line Median	$\mathbb{R}^d$	$r = 1, 2$	$1+\varepsilon$	$nd(k/\varepsilon)^{O(1)} + d(\log d)^{(k/\varepsilon)^{O(1)}}$	[17]
k-Line Median	$\mathbb{R}^d$	$r = 1$	$1+\varepsilon$	$ndk^{O(1)} + (\varepsilon^{-d} \log n)^{O(dk)}$	[28]
k-Line Median	$\mathbb{R}^d$	$r = 2$	$1+\varepsilon$	$ndk^{O(1)} + (\varepsilon^{-d} \log n)^{O(dk^2)}$	[21]
k-Line Median	$\mathbb{R}^d$	$r = 1, 2$	$1+\varepsilon$	$ndk^{O(1)} + (\varepsilon^{-1} \log n)^{O(k)}$	[22]
k-Line Median	$\mathbb{R}^d$	$r \in O(1)$	$1+\varepsilon$	$ndk^{O(1)} + \varepsilon^{-2}\log(n) \cdot k^{O(k)}$	⋆⋆

Table 1: Approximation Algorithms. The input is a set P of n points in $\mathbb{R}^d$ or in an arbitrary metric space. The results of this paper are marked with ⋆⋆.

Core-set	**Metric**	**ℓ_r loss**	**Size**	**Ref.**
Weighted k-median	$\mathbb{R}^1$	$r = 1$	$(\varepsilon^{-1}\log n)^{O(k)}$	[28]
Weighted k-median	$\mathbb{R}^1$	$r = 2$	$(\varepsilon^{-1}\log n)^{O(k^2)}$	[21]
Weighted k-median	$\mathbb{R}^1$	$r = \infty$	$(k/\varepsilon)^{O(k)}$	[2]
Weighted k-median	$\mathbb{R}^d$	$r = \infty$	$O(k!/\varepsilon^{dk})$	[27]
Weighted k-median	$\mathbb{R}^d$/Arbitrary	$r \in O(1)$	$k^{O(k)}(\varepsilon^{-1} d \log n)^2$	⋆⋆
k-Line median	$\mathbb{R}^d$	$r = 1, 2$	$dk\varepsilon^{-2} + (\varepsilon^{-1}\log n)^{O(k^2)}$	[22]
k-Line median	$\mathbb{R}^d$	$r \in O(1)$	$dk\varepsilon^{-2} + k^{O(k)}(\varepsilon^{-1}\log n)^2$	⋆⋆
k-Median with penalties	$\mathbb{R}^d$/Arbitrary	$r \in O(1)$	$k^{O(k)}(\varepsilon^{-1} d \log n)^2$	⋆⋆
Robust k-median with m outliers	$\mathbb{R}^d$/Arbitrary	$r \in O(1)$	$(k+m)^{O(k+m)}(\varepsilon^{-1} d \log n)^2$	⋆⋆
M-Estimators	$\mathbb{R}^d$/Arbitrary	$r \in O(1)$	$k^{O(k)}(\varepsilon^{-1} d \log n)^2$	⋆⋆
k-Mean+median	$\mathbb{R}^d$/Arbitrary	$r \in O(1)$	$k^{O(k)}(\varepsilon^{-1} d \log n)^2$	⋆⋆

Table 2: Core-sets. The input is a set P of $n$points in $\mathbb{R}^d$ or in an arbitrary metric space. New results of this paper are marked with ⋆⋆. We denote $d = O(\log n)$ for the case of arbitrary metric space.

Lemma 2.1. *The conditions above imply*

(i) For $\phi = (4r)^r$,

$$D(p,c) - D(q,c) \leq \phi D(p,q) + \frac{D(p,c)}{4} \tag{3}$$

(ii) (Weak triangle inequality) For $\rho = \max\{2^{r-1}, 1\}$,

$$D(p,q) \leq \rho(D(p,c) + D(c,q)). \tag{4}$$

Proof. [1] (i) Let $x = \mathrm{dist}(p,c), y = \mathrm{dist}(q,c), z = \mathrm{dist}(p,q)$. So we are to show $D(x) - D(y) \leq \phi D(z) + D(x)/4$. We suppose that $x > y$ and $D(x) > \phi D(z)$, otherwise the lemma is immediate. So by Eqn 2, $x > z\phi^{1/r}$.

An equivalent form of Eqn 2 is that for $\delta > 0$, $D(xe^{-\delta}) \geq e^{-r\delta} D(x)$. So $D(x) - D(y) \leq D(x) \cdot (1 - (y/x)^r)$.

Note that for $u \geq 0$, $1 - u^r \leq r(1-u)$; this follows because, viewing each side as a function of u, the two functions are tangent at $u = 1$, and the LHS is convex-cap while the RHS is linear. Applying this we have $D(x) - D(y) \leq D(x) \cdot r \cdot (x-y)/x$. Applying the triangle inequality $x - y \leq z$ we have that $D(x) - D(y) \leq D(x) \cdot r \cdot z/x$. By our earlier bound this is $< D(x) \cdot r \cdot \phi^{-1/r}$. Plugging in $\phi = (4r)^r$ implies Eqn 3.

(ii) By the triangle inequality and Eqn 2, for any $0 < p < 1$, $D(p,q) \leq pD(p,c)\left(\frac{\mathrm{dist}(p,c)+\mathrm{dist}(c,q)}{\mathrm{dist}(p,c)}\right)^r + (1-p)D(c,q)\left(\frac{\mathrm{dist}(p,c)+\mathrm{dist}(c,q)}{\mathrm{dist}(c,q)}\right)^r = (\mathrm{dist}(p,c) + \mathrm{dist}(c,q))^r \left(\frac{pD(p,c)}{\mathrm{dist}(p,c)^r} + \frac{(1-p)D(c,q)}{\mathrm{dist}(c,q)^r}\right)$. Substituting $p = \mathrm{dist}(p,c)^r/(\mathrm{dist}(p,c)^r + \mathrm{dist}(c,q)^r)$ we have $D(p,q) \leq (D(p,c) + D(c,q))\frac{(\mathrm{dist}(p,c)+\mathrm{dist}(c,q))^r}{\mathrm{dist}(p,c)^r+\mathrm{dist}(c,q)^r}$. By convexity considerations, for $r \geq 1$ the factor is maximized with $\mathrm{dist}(p,c) = \mathrm{dist}(c,q)$ and for $r \leq 1$ it is maximized with $\mathrm{dist}(c,q) = 0$, yielding Eqn 4. □

Definition 2.2 (Tractable (M, D) Problems)**.** *Let (M, dist) be a metric space. Let D be a function from $M \times M$ to $[0, \infty)$. We call the problem (M, D)* tractable *if* (1),(4)*, and* (3) *hold for some constants $\phi, \rho \in (0, \infty)$.*

In Theorem 4.3 we show how to perform data reduction for tractable (M, D) problems, conditional on a shatter function (essentially, VC dimension) bound.

Let $P \subseteq M$ be a finite set of points. For $B \subseteq M$, we denote by **closest**(P, B, γ) the set that consists of the $\lceil \gamma |P| \rceil$ points $p \in P$ with the smallest values of $\min_{q \in B} D(p,q)$. For $p \in M$ and a set $C \subseteq M \times \mathbb{R}_+$ define $D_W(p, C) = \min_{(c,w) \in C} w \cdot D(p,c)$. Each $(c, w) \in C$ is called a *weighted center*. For integer $k \geq 1$ write $[k] := \{1, \ldots, k\}$.

We show how to perform data reduction for a variety of statistical problems by considering appropriate choices of M, and D and showing that the above properties are satisfied. The families of functions we consider have the following description:

$$\begin{aligned} F_{M,D} &: M \to \mathbb{R}_+ \\ F_{M,D} &= \{f_C\}_{C \subseteq M \times \mathbb{R}_+, |C| = k} \\ f(p) &= D_W(p, C) \end{aligned}$$

In this notation, the weighted k-median problem is ($M = \mathbb{R}^d$, dist = Euclidean metric, and D = dist); the weighted k-mean problem is ($M = \mathbb{R}^d$, dist = Euclidean metric, and $D = \mathrm{dist}^2$);

[1] Notation has been switched here to univariate D; fix the notation up for final version

and the k-median with outliers problem is (a special case of) the problem ($M = \mathbb{R}^d$, dist $=$ Euclidean metric, and $D = \min\{\text{dist}, 1\}$).

As established in [36, 22], a sufficient condition for data reduction is that the *total sensitivity* $\mathcal{T} = \mathcal{T}(F_{M,D})$ be small, and that we be able to effectively compute good upper bounds $s(p)$ for the sensitivities of the points of P[2]; the cardinality of the resulting set A is then approximately $\mathcal{T}^2\, d/\varepsilon^2$, where d is a Vapnik-Chervonenkis measure of the combinatorial complexity of the family $F_{M,D}$.

Before showing how to compute bounds on the sensitivities of points we need two more definitions.

Definition 2.3. *For a finite set $Q \subseteq M$ and $\gamma \in [0,1]$, define*

$$D^*(Q,\gamma) := \min_{c \in M} \sum_{p \in \mathbf{closest}(Q,c,\gamma)} D(p,c).$$

A point c which achieves the above min is, in a sense, a median of a densest region of the data. (One may also think of it as a good "median with outliers" for the data.) In what follows it would be very useful to have a subroutine to compute such a point, but this is a nearly circular request (though not quite as hard as the full goal of the paper). Instead we will be able to achieve our results using a subroutine which produces a point with the following weaker property.

Definition 2.4 (Robust Median)**.** *For $\gamma \in [0,1]$, $\tau \in (0,1)$ and $\alpha > 0$, the point $q \in M$ is a* (γ, τ, α)-median *of the finite set $Q \subseteq M$ if*

$$\sum_{p \in \mathbf{closest}(Q,\{q\},(1-\tau)\gamma)} D(p,q) \le \alpha \cdot D^*(Q,\gamma). \tag{5}$$

3 Bounding point sensitivities

3.1 Sensitivity bound for weighted medians

The key technical advance in this paper lies in the following lemma, which shows how to translate the new definitions of the previous section into good upper bounds on the sensitivities of data points. This lemma is what enables us to handle weighted clustering problems.

In each application one needs only to ensure that the problem is "tractable" as in definition 2.2, and that the appropriate shatter function ($\sim$ VC dimension) is bounded.

Lemma 3.1. *Let (M,D) be tractable and let $P \subseteq M$ be a finite set. Suppose that (q_k, Q_k) is the output of the algorithm* `Recursive-Robust-Median`(P,k)*. Then for every set $C = \{(c_1,w_1),\cdots,(c_k,w_k)\} \subseteq M \times [0,\infty)$ and $p \in Q_k$ such that $D_W(p,C) > 0$, we have*

$$\frac{D_W(p,C)}{\sum_{q \in P} D_W(q,C)} \le \frac{O(k)}{|Q_k|}.$$

Proof of Lemma 3.1: Consider the variables $Q_0,\ldots,Q_k$ and $q_1,\ldots,q_k$ that are computed during the execution of `Recursive-Robust-Median`(P,k). A point $p \in P$ is *served* by the weighted center $(c,w) \in C$ if $D_W(p,C) = w \cdot D(p,c)$. For every $i \in [k+1]$, let $(c_i, w_i) \in C$ denote a center

[2] For a family F and n data points P, the sensitivity of $p \in P$ is $s(p) = \sup_{f \in F} f(p)/((1/n)\sum_{q \in P} f(q))$; the total sensitivity $\mathcal{T}(F)$ is $\sup_P \sum_{p \in P} s(p)$.

Algorithm 1: `Recursive-Robust-Median`(P, k)

Input: A set $P \subseteq M$, an integer $k \geq 1$.
Output: A pair (q_k, Q_k) that satisfies Lemma 3.1.

$Q_0 \leftarrow P$
for $i = 1$ ***to*** k **do**
    Compute a $(1/k, \tau, \alpha)$-median $q_i \in M$ of Q_{i-1} for some constants $\tau \in (0, 1)$ and $\alpha \in (0, \infty)$
    $Q_i \leftarrow$ **closest**$(Q_{i-1}, \{q_i\}, (1-\tau)/(2k))$
return (q_k, Q_k)

that serves at least $|Q_{i-1}|/k$ points from Q_{i-1}. Let P_i denote the points of P that are served by (c_i, w_i). For every $i \in [k]$, let $Q'_i :=$ **closest**$(Q_{i-1}, \{q_i\}, (1-\tau)/k)$, and $D^*_i = \sum_{q \in Q'_i} D(q, q_i)$. Since $|P_i \cap Q_{i-1}| \geq |Q_{i-1}|/k \geq |Q'_i|$, we have by Definition 2.3,

$$\sum_{q \in P_i \cap Q_{i-1}} D(q, c_i) \geq D^*(Q_{i-1}, 1/k). \tag{6}$$

We prove the lemma using the following case analysis. **Case (i):** There is an $i \in [k]$ such that

$$D(p, c_i) \leq \frac{16\phi\rho\alpha \cdot D^*_i}{|Q'_k|}. \tag{7}$$

Case (ii): Otherwise.
Proof of Case (i): By (7) we have

$$\begin{aligned} \frac{D_W(p, C)}{\sum_{q \in P} D_W(q, C)} &\leq \frac{w_i \cdot D(p, c_i)}{w_i \sum_{q \in P_i} D_W(q, C)} \\ &\leq \frac{D(p, c_i)}{\sum_{q \in P_i} D(q, c_i)} \leq \frac{16\phi\rho\alpha \cdot D^*_i / |Q'_k|}{\sum_{q \in P_i \cap Q_{i-1}} D(q, c_i)}. \end{aligned} \tag{8}$$

By Definition 2.4, we have $D^*(Q_{i-1}, 1/k) \geq D^*_i/\alpha$. Using this with (6) yields $\sum_{q \in P_i \cap Q_{i-1}} D(q, c_i) \geq D^*_i/\alpha$. By the last inequality and (8) we obtain

$$\frac{D_W(p, C)}{\sum_{q \in P} D_W(C)} \leq \frac{16\phi\rho\alpha \cdot D^*_i / |Q'_k|}{D^*_i/\alpha} \leq \frac{16\phi\rho\alpha^2}{|Q_k|}.$$

Proof of Case (ii): By the pigeonhole principle, $c_i = c_j$ for some $i, j \in [k+1]$, $i < j$. Put $q \in P_j \cap Q_{j-1}$ and note that $p \in Q_k \subseteq Q_{j-1}$. Using the Markov inequality,

$$D(q, q_{j-1}), D(p, q_{j-1}) \leq \frac{2D^*_{j-1}}{|Q'_{j-1}|}.$$

By this, the symmetry of $D(\cdot, \cdot)$ and (4),

$$D(p, q) \leq \rho\big(D(p, q_{j-1}) + D(q_{j-1}, q)\big) \leq \frac{4\rho \cdot D^*_{j-1}}{|Q'_{j-1}|}.$$

Using the last inequality with (3) yields

$$D(p, c_j) - D(q, c_j) \leq \phi D(p, q) + \frac{D(p, c_j)}{4}$$
$$\leq \frac{4\phi\rho \cdot D^*_{j-1}}{|Q'_{j-1}|} + \frac{D(p, c_j)}{4} \leq \frac{4\phi\rho\alpha \cdot D^*_i}{|Q'_k|} + \frac{D(p, c_j)}{4}.$$

Since Case (i) does not hold, we have $16\phi\rho\alpha \cdot D^*_i/|Q'_k| < D(p, c_i) = D(p, c_j)$. Combining the last two inequalities yields

$$D(p, c_j) - D(q, c_j) < \frac{D(p, c_j)}{4} + \frac{D(p, c_j)}{4} = \frac{D(p, c_j)}{2}.$$

That is, $D(q, c_j) > D(p, c_j)/2$. Hence,

$$\frac{D_W(p, C)}{\sum_{q \in P} D_W(C)} \leq \frac{D(p, c_j)}{\sum_{q \in P_j \cap Q_{j-1}} D(q, c_j)}$$
$$< \frac{2D(p, c_j)}{D(p, c_j) \cdot |P_j \cap Q_{j-1}|} \leq \frac{2k}{|Q_{j-1}|} \leq \frac{2k}{|Q_k|}.$$

□

3.2 Data reduction for robust medians

Theorem 3.2. *Let (M, D) be tractable. Let $Q \subseteq M$ be a finite set of points, $k \geq 1$ an integer, and $\delta \in (0, 1)$. Let $q \in M$ be the output of a call to* `Median`(Q, k, δ)*; See Algorithm 2. Then, with probability at least $1 - \delta$, the point q is a $\big(1/k, 1/4, 2\big)$-median for Q.*

Algorithm 2: `Median`(Q, k, δ)

Input: A finite set $Q \subseteq M$, an integer $k \geq 1$, and $\delta \in (0, 1/10)$.

Output: A point $q \in M$ that satisfies Theorem 3.2.

1. $b \leftarrow$ a universal constant that can be determined from the proof of Theorem 3.2
2. Pick a uniform i.i.d. sample S of $bk^2 \log(1/\delta)$ points from Q
3. $q \leftarrow$ a point that minimizes $\sum_{p \in \mathbf{closest}(S, \{q\}, 15/(16k))} D(p, q)$ over $q \in S$
4. **return** q

Proof. We consider the variables b, Q' and S as defined in Algorithm 2. Put $\tau = 1/16$ and $\gamma = 1/k$. Let $q^* \in M$ be a $((1 - \tau)\gamma, 0, 1)$-median of S. Let q be the closest point to q^* in S. By (4), for every $p \in M$ we have

$$D(p, q) \leq \rho(D(p, q^*) + D(q^*, q)) \leq 2\rho \cdot D(p, q^*).$$

Summing this over every $p \in \mathbf{closest}(S, \{q^*\}, (1 - \tau)\gamma)$ yields

$$\sum_{p \in \mathbf{closest}(S, \{q\}, (1-\tau)\gamma)} D(p, q) \leq 2\rho D^*(S, (1 - \tau)\gamma).$$

Hence, q is a $((1 - \tau)\gamma, 0, 2)$-median of S, which is also a $((1 - \tau)\gamma, \tau, 2)$-median of S. The theorem now follows from Theorem 3.3. □

The following is a special case of Lemma 9.6 that is proven in [22]:

Theorem 3.3 ([22])**.** *Let (M, D) be tractable. Let $Q \subseteq M$ be a finite set of points. Let $\gamma \in (0, 1)$, and $\tau, \delta \in (0, 1/10)$. Pick uniformly, i.i.d., a (multi)-set S of*

$$s = \frac{b}{\tau^4 \gamma^2} \cdot \log\left(\frac{1}{\delta}\right)$$

points from Q, where b is a sufficiently large universal constant. With probability at least $1-\delta$, any $((1-\tau)\gamma, \tau, 2)$-median for S is a $(\gamma, 4\tau, 2)$-median for Q.

Proof. For every $p \in P$ and $c \in M$ let $f(c) = D(p, c)$. Let $\mathcal{D}(S) = S$ for every $S \subset P$. Using the (weak) triangle inequality, we have that one of the points of S is a constant factor approximation for the median of S. The theorem now follows from [22, Lemma 9.6]. □

4 Data reduction for tractable (M, D) problems

Definition 4.1 ($\dim(M, D, k)$ [49])**.** *Let (M, D) be tractable. For every $r \geq 0$ and $C \subseteq M \times [0, \infty)$ of size $|C| = k$, let $\mathrm{ball}(C, r) = \{p \in P \mid D_W(p, C) \leq r\}$. Let*

$$\mathrm{balls} = \{\mathrm{ball}(C, r) \mid C \subseteq M \times [0, \infty), |C| = k, r \geq 0\}.$$

The dimension $\dim(M, D, k)$ is the smallest integer d such that for every finite $S \subseteq M$ we have

$$\Big|\{S \cap \mathrm{ball} \mid \mathrm{ball} \in \mathrm{balls}\}\Big| \leq |S|^d \ .$$

The following is a corollary of [22, Theorem 13.1].

Corollary 4.2. *Let (M, D) be tractable, and $P \subseteq M$ be a finite set of points. Let $\varepsilon \in (0, 1/4)$. Let $s : P \to [0, \infty)$ be a function on P such that*

$$s(p) \geq \max_{C \in M \times [0,\infty), |C|=k} \frac{D_W(p, C)}{\sum_{q \in P} D_W(q, C)}.$$

Let $\mathcal{T} = \sum_{p \in P} s(p)$, and b be a sufficiently large constant. Pick a (multi)-set A of $b\mathcal{T}^2(\dim(M, D, k) + \log(1/\delta))/\varepsilon^2$ points from P by repeatedly, i.i.d., selecting $p \in P$ with probability $s(p)/\mathcal{T}$. For $p \in A$ let $\nu(p) = \mathcal{T}/(|A| \cdot s(p))$. Then, with probability at least $1-\delta$:

$$\text{For all } C \in M \times [0, \infty) \text{ and } |C| = k : \quad \left|\sum_{p \in P} D_W(p, c) - \sum_{p \in A} \nu(p) D_W(p, c)\right| \leq \varepsilon \sum_{p \in P} D_W(p, c).$$

Proof. Let $X = (M \times [0, \infty))^k$. For every $p \in P$ and $C \in X$, let $f_p(C) = D_W(p, C)$, $s_{f_p} = f'_p = f_p$, $m(f_p) = n \cdot s(p)/\mathcal{T}$, and $g_{f_p}(C) = f_p(C)/m(f_p)$. Let G_{f_p} consists of $m(f_p)$ copies of g_f and let $G = \bigcup_{p \in P} G_{f_p}$. Hence, $S = \left\{g_{f_p} \mid p \in A\right\}$ is a uniform random sampling from G. By [22, Theorem 6.9], for a sufficiently large b, S is an $(\varepsilon/(2\mathcal{T}))$-approximation of G, with probability at least $1-\delta$. Assume that this event indeed occurs. Let $U = \{g \cdot |G|/|S| \mid g \in S\}$. By Theorem 13.1 of [22], we obtain that

$$\forall C \in X : |\sum_{p \in P} f_p(C) - \sum_{f \in U} f(C)| \leq \frac{\varepsilon}{\mathcal{T}} \max_{p \in P} \frac{f_p(C)}{m(f_p)} \sum_{p \in P} m(f_p). \tag{9}$$

We have

$$\sum_{p\in P} m(f_p) = \sum_{p\in P} \frac{ns(p)}{\mathcal{T}} = n,$$

and, for every $C \in X$

$$\max_{p\in P} \frac{f_p(C)}{m(f_p)} = \frac{\mathcal{T}}{n} \max_{p\in P} \frac{D_W(p,C)}{s(p)} \leq \frac{\mathcal{T}}{n} \sum_{p\in P} D_W(p,C).$$

For every $f = g_{f_p} \cdot |G|/|S| \in U$ we have

$$f(C) = \frac{g_{f_p}(C)\cdot |G|}{|S|} = \frac{f_p(C)\cdot n}{m(f_p)|A|} = \frac{D_W(p,C)\cdot n}{m(f_p)|A|} = \frac{D_W(p,C)\cdot \mathcal{T}}{s(p)\cdot |A|} = \nu(p) D_W(p,C).$$

Substituting the last three inequalities in (9) yields

$$\begin{aligned}
\forall C \in X : &|\sum_{p\in P} f_p(C) - \sum_{p\in A} w(p) D_W(p,C)| \\
&\leq \frac{\varepsilon}{\mathcal{T}} \cdot \frac{\mathcal{T}}{n} \sum_{q\in P} D_W(q,C)\cdot n \\
&= \varepsilon \sum_{p\in P} D_W(p,C).
\end{aligned}$$

□

Algorithm 3: `CoreSet`$(P, k, \varepsilon, \delta)$

Input: A set $P \subseteq M$, an integer $k \geq 1$, and $\tau, \delta \in (0, 1/10)$ where (M, D) is tractable.
Output: A set A and a probability measure ν on A that satisfy Theorem 4.3.

$b \leftarrow$ a constant that can be determined from the proof of Theorem 4.3
$Q_0 \leftarrow P$
while $|Q_0| > b$ **do**
  for $i \leftarrow 1$ **to** k **do**
    $q_i \leftarrow$ `Median`$(Q_{i-1}, k, \delta/(k^{bk}\log n))$
    $Q_i \leftarrow$ **closest**$(Q_{i-1}, \{q_i\}, 1/(bk))$
  for each $p \in Q_k$ **do** $s(p) \leftarrow \frac{bk}{|Q_k|}$
  $Q_0 \leftarrow Q_0 \setminus Q_k$
for each $p \in Q_0$ **do** $s(p) \leftarrow 1$
$\mathcal{T} \leftarrow \sum_{p\in P} s(p)$
Pick a (multi)-set A of $b\mathcal{T}^2(\dim(M, D, k) + \log(1/\delta))/\varepsilon^2$ points from P by repeatedly, i.i.d., selecting $p \in P$ with probability $s(p)/\mathcal{T}$
for each $p \in A$ **do** $\nu(p) \leftarrow \dfrac{\mathcal{T}}{|A|\cdot s(p)}$
return (A, ν)

Theorem 4.3. *Let (M, D) be tractable. Let $P \subseteq M$ be a set of n points, and $(\varepsilon, \delta) \in (0, 1/10)$. Let A be the output of the procedure* `CoreSet`$(P, k, \varepsilon, \delta)$*; see Algorithm 3. Then the following hold:*

(i)

$$|A| = \frac{k^{O(k)}(\log n)^2}{\varepsilon^2} \cdot \big(\dim(M, D, k) + \log(1/\delta)\big).$$

(ii) *With probability at least* $1 - \delta$,

$$\forall C \subseteq M \times [0, \infty), |C| = k :$$

$$\left| \sum_{p \in P} D_W(p, C) - \sum_{p \in A} \nu(p) \cdot D_W(p, C) \right| \leq \varepsilon \sum_{p \in P} D_W(p, C).$$

(iii) *The construction of* A *takes time*

$$ntk^{O(k)} + tk^{O(k)} \log(1/\delta) \log(n) + \frac{k^{O(k)}(\log n)^3}{\varepsilon^2} \cdot \big(\dim(M, D, k) + \log(1/\delta)\big),$$

where t *is the time it takes to compute* $D(p, q)$ *for some* $p, q \in M$.

The structure of the algorithm is this: The equivalent of Algorithm 1 is run repeatedly to identify a "dense" cluster in the data. (Lines 4-7.) Due to Lemma 3.1 we the sensitivity of each point in this cluster is bounded by some constant divided by the current number of points. The cluster is then removed, and we repeat.

Proof. ***(i)*** For every $i \in [k]$, let $Q_i^{(j)}$ denote the value of Q_i at Line 7 of Algorithm 3 during the jth "while" iteration. Let J denote the total number of "while" iterations. By Line 6 of Algorithm 3, we have that $|Q_i^{(j)}| \geq |Q_{i-1}^{(j)}|/(bk)$. Hence,

$$|Q_k^{(j)}| \geq \frac{|Q_0^{(j)}|}{(bk)^k} = \frac{|Q_0^{(j)}|}{k^{O(k)}}.$$

By the last equation and Line 8, for every $j \in [J - 1]$ we have

$$\begin{aligned} \left|Q_0^{(j+1)}\right| &= \left|Q_0^{(j)}\right| - \left|Q_k^{(j)}\right| \leq \left|Q_0^{(j)}\right| - \frac{|Q_0^{(j)}|}{k^{O(k)}} \\ &= \left|Q_0^{(1)}\right| \left(1 - \frac{1}{k^{O(k)}}\right)^j = n \left(1 - \frac{1}{k^{O(k)}}\right)^j. \end{aligned} \tag{10}$$

Since $\left|Q_0^{(J)}\right| \geq 1$, substituting $j = J - 1$ in the previous inequality we conclude that

$$J \leq k^{O(k)} \log n. \tag{11}$$

By Line 3, the size of Q_0 during the execution of Line 9 is $O(1)$. By the definition of $s(p)$ in Lines 7 and 9 we have

$$\begin{aligned} \mathcal{T} = \sum_{p \in P} s(p) &= \sum_{j \in [J]} \sum_{p \in Q_k} s(p) + O(1) \\ &= \sum_{j \in [J]} \sum_{p \in Q_k} \frac{bk}{|Q_k|} + O(1) = J \cdot bk + O(1). \end{aligned}$$

Together with (11) we obtain $\mathcal{T} \leq k^{O(k)} \log n$. By this and Line 11 we thus have

$$|A| = \frac{b\mathcal{T}^2\left(\dim(M,D,k) + \log(1/\delta)\right)}{\varepsilon^2} = \frac{k^{O(k)}(\log n)^2}{\varepsilon^2} \cdot \left(\dim(M,D,k) + \log(1/\delta)\right).$$

(ii) Put $i \in [k]$, $j \in [J]$, $C = \{(c_1, w_1), \ldots, (c_k, w_k)\} \subseteq M \times [0, \infty)$ and $p \in Q_k^{(j)}$ such that $D_W(p, C) > 0$. Let $q_i^{(j)}$ denote the value of q_i after the execution of Line 5 of the jth "while" iteration. By Theorem 3.2, with probability at least $1 - \delta/(k^{bk} \log n)$ we have that $q_i^{(j)}$ is a $(1/k, 1/4, 2)$-median for $Q_{i-1}^{(j)}$ Assume that this event occurs for every $i \in [k]$, and $j \in [J]$. This assumption holds with probability at least $1 - Jk\delta/k^{bk} \geq 1 - \delta$ for a sufficiently large b.

By substituting $P = Q_0^{(j)}$, $\tau = 1/4$ and $\alpha = 2$ in Algorithm 1, we thus have that the pair $(q_k^{(j)}, Q_k^{(j)})$ satisfies Lemma 3.1 for every $j \in [J]$. That is,

$$\frac{D_W(p,C)}{\sum_{q \in Q_0^{(j)}} D_W(q,C)} \leq \frac{O(k)}{|Q_k^{(j)}|}.$$

Hence, for the value $s(p)$ that is defined in Line 7 of Algorithm 3, and an appropriate b,

$$s(p) = \frac{bk}{|Q_k^{(j)}|} \geq \frac{D_W(p,C)}{\sum_{q \in Q_0^{(j)}} D_W(q,C)} \geq \frac{D_W(p,C)}{\sum_{q \in P} D_W(q,C)}.$$

By Corollary (4.2), with probability at least $1 - \delta$ we have

$$\forall C \in M \times [0,\infty), |C| = k : \left|\sum_{p \in P} D_W(p,c) - \sum_{p \in A} \nu(p) D_W(p,c)\right| \leq \varepsilon \sum_{p \in P} D_W(p,c).$$

(iii) The running time of Algorithm 3 is dominated by Lines 5, 6 and 11 as follows. For a set $Q \subseteq M$, the running time of `Median`(Q, k, δ) (see Algorithm 2) is dominated by Line 3 which can be implemented in $d|S|^2 = dbk^2 \log(1/\delta)$ time by computing the distance between every pair of points in S. Using (11), the overall time of Line 5 of Algorithm 3 is

$$J \cdot k \cdot (d|S|^2) = k^{O(k)} \log n \cdot dbk^2 \log(1/\delta) = dk^{O(k)} \log(1/\delta) \log(n).$$

Line 6 of Algorithm 3 takes $d \cdot |Q_{i-1}^{(j)}| \leq d \cdot |Q_0^{(j)}|$ time for the jth "while" iteration using order statistics. Using (10), the overall execution time of Line 6 is

$$d \sum_{j \in [J], i \in [k]} |Q_{i-1}^{(j)}| \leq dk \sum_{j \in [J]} |Q_0^{(j)}| \leq ndk \sum_{j \in [J]} \left(1 - \frac{1}{k^{O(k)}}\right)^j \leq ndk^{O(k)}.$$

Line 11 can be implemented using a binary tree in time

$$\log(n) \cdot |A| = \frac{k^{O(k)}(\log n)^3}{\varepsilon^2} \cdot \left(\dim(M,D,k) + \log(1/\delta)\right).$$

□

5 Applications

5.1 $F_{d,k}^{\mathbf{wght}}$ and $F_{d,k}^{\mathbf{out}}$

The following theorem includes both $F_{d,k}^{\text{wght}}$ and $F_{d,k}^{\text{out}}$ as special cases by taking, respectively, $h = \infty$ or all weights equal.

Theorem 5.1. *Let P be a set of points in a metric space (M, dist), $r > 0$ be a constant and $h \in (0, \infty)$. Let $k \geq 1$ be an integer, and $\varepsilon \in (0, 1/10)$. A set A of size*

$$|A| = \frac{k^{O(k)}(\log n)^2}{\varepsilon^2} \cdot \big(\dim(M, D, k) + \log(1/\delta)\big)$$

and a weight function $\nu : A \to \mathbb{R}_+$ can be computed in time

$$ntk^{O(k)} + tk^{O(k)} \log(1/\delta) \log(n) + \frac{k^{O(k)}(\log n)^3}{\varepsilon^2} \cdot \big(\dim(M, D, k) + \log(1/\delta)\big),$$

such that, with probability at least $1 - \delta$,

$$\forall C \subseteq M \times [0, \infty), |C| = k : \left| \sum_{p \in P} f_C(p) - \sum_{p \in A} \nu(p) f_C(p) \right| \leq \varepsilon \sum_{p \in P} f_C(p),$$

where $f_C(p) := \min_{(c,w) \in C} \big(w \min\{h, \text{dist}^r(p, c)\}\big)$, and t is the time it takes to compute $D(p, q)$ for some $p, q \in M$.

Proof. Define $D : M \times M \to \mathbb{R}_+$ as $D(p, q) = \min\{h, \text{dist}^r(p, q)\}$. Using Theorem 4.3 it suffices to prove that (M, D) is tractable; see Definition 2.2. Condition (1) is merely that D is symmetric. It remains to show (4) and (3).
Proof of (4): If $\max\{D(p, c), D(c, q)\} = h$ then

$$D(p, q) \leq h = \max\{D(p, c), D(c, q)\} \leq \rho\big(D(p, c) + D(c, q)\big).$$

Otherwise, $D(p, c) = \text{dist}^r(p, c)$ and $D(c, q) = \text{dist}^r(c, q)$. It is straightforward to show (and well-known) that for $\rho = \max\{1, 2^{r-1}\}$, $\text{dist}^r(p, q) \leq \rho\big(\text{dist}^r(q, c) + \text{dist}^r(p, c)\big)$.
Proof of (3): If $D(q, c) \geq D(p, c)$ then (3) holds trivially.

Otherwise, $D(p, c) > D(q, c) = \text{dist}^r(q, c)$. So

$$\begin{aligned}
D(p, c) - D(q, c) &= \min\{h, \text{dist}^r(p, c)\} - \text{dist}^r(q, c) \\
&= \min\{h - \text{dist}^r(q, c), \text{dist}^r(p, c) - \text{dist}^r(q, c)\} \\
&\leq \min\left\{h, \phi \text{dist}^r(p, q) + \frac{\text{dist}^r(q, c)}{4}\right\} \quad \text{using Lemma 5.2} \\
&\leq \min\{h, \phi \text{dist}^r(p, q)\} + \frac{D(p, c)}{4} \\
&\leq \phi D(p, q) + \frac{D(p, c)}{4}.
\end{aligned}$$

$\square$

Lemma 5.2. *Let* (M, dist) *be a metric space and* $r \geq 1$*. Then for every* $p, q, c \in M$*,*

$$\mathrm{dist}^r(p,c) - \mathrm{dist}^r(q,c) \leq \phi \mathrm{dist}^r(p,q) + \frac{\mathrm{dist}^r(p,c)}{4} \tag{12}$$

with $\phi = (4r)^r$*.*

Proof. We suppose $\mathrm{dist}(p,c) > \mathrm{dist}(q,c)$, otherwise the lemma is trivial. If $\mathrm{dist}(p,c)^r < \phi \mathrm{dist}^r(p,q)$ then the lemma is immediate, so we suppose that

$$\mathrm{dist}(p,c) \geq \phi^{1/r} \mathrm{dist}(p,q). \tag{13}$$

Note that for $x \geq 0$, $1 - x^r \leq r(1-x)$; this follows because, viewing each side as a function of x, the two functions are tangent at $x = 1$, and the LHS is convex-cap while the RHS is linear. Applying this we have

$$\mathrm{dist}^r(p,c) - \mathrm{dist}^r(q,c) = \mathrm{dist}^r(p,c) \left(1 - \left(\frac{\mathrm{dist}(q,c)}{\mathrm{dist}(p,c)}\right)^r\right) \leq \mathrm{dist}^r(p,c) \cdot r \cdot \frac{\mathrm{dist}(p,c) - \mathrm{dist}(q,c)}{\mathrm{dist}(p,c)}$$

The proof is now completed by applying (13) and the triangle inequality $\mathrm{dist}(p,c) - \mathrm{dist}(q,c) \leq \mathrm{dist}(p,q)$ to obtain:

$$r \cdot \frac{\mathrm{dist}(p,c) - \mathrm{dist}(q,c)}{\mathrm{dist}(p,c)} \leq r \cdot \frac{\mathrm{dist}(p,q)}{\mathrm{dist}(p,c)} \leq r/\phi^{1/r} = 1/4$$

□

5.2 Classic M-estimators

The following loss (or distortion) functions, known as *M-estimators*, are popular with statisticians performing robust (i.e., outlier-resistant) estimation [33, 50]. (In these expressions, with slight abuse of notation, x is shorthand for $\mathrm{dist}(p,q)$. The parameter r is effectively a distance threshold for outliers):

$$D_r^{\mathrm{Huber}}(\mathrm{dist}(p,q)) = D_r^{\mathrm{Huber}}(x) = \min\{x^2/2, r(x - r/2)\} \tag{14}$$

$$D_r^{\mathrm{Tukey}}(\mathrm{dist}(p,q)) = D_r^{\mathrm{Tukey}}(x) = \min\{r^2/6, \frac{3r^4x^2 - 3r^2x^4 + x^6}{6r^4}\} \tag{15}$$

It is straightforward to show that the families $(M, D_1^{\mathrm{Huber}})$ and $(M, D_1^{\mathrm{Tukey}})$ are tractable, with a proof very similar to that in the preceding subsection. The shatter function is again that of balls in the metric space. Consequently, Theorem 5.1 shows that we can perform data reduction for k-clustering (with weights) for these loss functions.

Other applications. As discussed in the introduction, our method provides the smallest core-set for the k-line-median problem.

Our method also enables an approach to the Robust k-median with m outliers problem: discarded outliers can be treated simply as infinite-weight centers, so our method can handle a constant number of discarded outliers. This causes an exponential dependence of the run-time on m, but is still the only known near-linear-time (in n) $(1+\varepsilon)$-approximation for the problem, even for $k = 1$, $d = 2$.

References

[1] P. Agarwal and J. Phillips. An efficient algorithm for 2d euclidean 2-center with outliers. *Algorithms-ESA 2008*, pages 64–75, 2008.

[2] P. K. Agarwal, C. M. Procopiuc, and K. R. Varadarajan. Approximation algorithms for k-line center. In *Proc. 10th Annu. European Symp. on Algorithms (ESA)*, volume 2461 of *Lecture Notes in Computer Science*, pages 54–63. Springer, 2002.

[3] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, July 1997.

[4] N. Alon and B. Sudakov. On two segmentation problems. *J. of Algorithms*, 33:173–184, 1999.

[5] V. Angadi and S. Patel. Hough transformation. Technical report, 2010.

[6] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004.

[7] P. L. Bartlett and P. M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *J. Comput. System Sci.*, 56:174–190, 1998.

[8] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *J. Comput. System Sci.*, 52:434–452, 1996.

[9] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *J. of Comput. Syst. Sci.*, 50(1):74–86, 1995.

[10] S. Ben-David, N. Cesa-Bianchi, and P. M. Long. Characterizations of learnability for classes of {0, ..., n}-valued functions. In *Proc. 5'th Annual ACM COLT*, pages 333–340, 1992.

[11] A. Benczúr and D. Karger. Approximating s–t minimum cuts in $\tilde{O}(n^2)$ time. In *Proc. 28th STOC*, pages 47–55, 1996.

[12] Daniel Bienstock, Michel X. Goemans, David Simchi-Levi, and David P. Williamson. A note on the prize collecting traveling salesman problem. *Math. Program*, 59:413–420, 1993.

[13] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, October 1989.

[14] M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, SODA '01, pages 642–651, Philadelphia, PA, USA, 2001. Society for Industrial and Applied Mathematics.

[15] K. Chen. A constant factor approximation algorithm for k -median clustering with outliers. In Shang-Hua Teng, editor, *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008, San Francisco, California, USA, January 20-22, 2008*, pages 826–835. SIAM, 2008.

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

[17] A. Deshpande and K. R. Varadarajan. Sampling-based dimension reduction for subspace approximation. In *Proc. 39th Annu. ACM Symp. on Theory of Computing (STOC)*, pages 641–650, 2007.

[18] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. In *Machine Learning*, volume 56, pages 9–33, 2004.

[19] M. Effros and L. J. Schulman. Deterministic clustering with data nets. ECCC TR04-085, http://eccc.hpi-web.de/eccc-reports/2004/TR04-050/index.html, 2004.

[20] T. Feder and D. Greene. Optimal algorithms for approximate clustering. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, STOC '88, pages 434–444, New York, NY, USA, 1988. ACM.

[21] D. Feldman, A. Fiat, and M. Sharir. Coresets for weighted facilities and their applications. In *Proc. 47th IEEE Annu. Symp. on Foundations of Computer Science (FOCS)*, pages 315–324, 2006.

[22] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In L. Fortnow and S. P. Vadhan, editors, *STOC*, pages 569–578. ACM, 2011. Full version at arXiv:1106.1379; theorem citations follow the full version.

[23] D. Feldman, M. Monemizadeh, and C. Sohler. A PTAS for k-means clustering based on weak coresets. In *Proc. 23rd ACM Symp. on Computational Geometry (SoCG)*, pages 11–18, 2007.

[24] B. Foster, S. Mahadevan, and R. Wang. A gpu-based approximate svd algorithm. Technical report, U. Massachusetts, 2010.

[25] M. X. Goemans and D. P. Williamson. A general approximation technique for constrained forest problems. *SIAM Journal on Computing*, 24(2):296–317, 1995.

[26] F. Hampel, C. Hennig, and E. Ronchetti. A smoothing principle for the huber and other location m-estimators. *Computational Statistics & Data Analysis*, 55(1):324–337, 2011.

[27] S. Har-Peled. No coreset, no cry. In *Proc. 24th Int. Conf. Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, volume 3328 of *Lecture Notes in Computer Science*, pages 324–335. Springer, 2004.

[28] S. Har-Peled. Coresets for discrete integration and clustering. In *Proc. 26th Int. Conf. Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, volume 4337 of *Lecture Notes in Computer Science*, pages 33–44. Springer, 2006.

[29] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *Proc. 36th Annu. ACM Symp. on Theory of Computing (STOC)*, pages 291–300, 2004.

[30] J. Hardin, A. Mitani, L. Hicks, and B. VanKoten. A robust measure of correlation between two genes on a microarray. *BMC bioinformatics*, 8(1):220, 2007.

[31] D. Haussler and E. Welzl. Epsilon-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987.

[32] D.C. Hoaglin, F. Mosteller, and J.W. Tukey. *Understanding robust and exploratory data analysis*, volume 3. Wiley New York, 1983.

[33] P. J. Huber. *Robust Statistics.* Wiley, 1981.

[34] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. System Sci.*, 48:464–497, 1994.

[35] J. Kleinberg, C. Papadimitriou, and P. Raghavan. Approximation algorithms for segmentation problems. In *Proc. 30th STOC*, 1998.

[36] M. Langberg and L. J. Schulman. Universal epsilon-approximators for integrals. In M. Charikar, editor, *SODA*, pages 598–607. SIAM, 2010.

[37] J.H. Lin and J.S. Vitter. e-approximations with minimum packing constraint violation (extended abstract). In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 771–782. ACM, 1992.

[38] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is NP-hard. In S. Das and R. Uehara, editors, *Proc. WALCOM, LNCS 5431*, pages 274–285. Springer-Verlag, 2009.

[39] J. Matousek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.

[40] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

[41] S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Arxiv preprint arXiv:1010.2731*, 2010.

[42] M. Owen. Tukeys biweight correlation and the breakdown, m.sc. thesis, pomona college. 2010.

[43] D. Pollard. *Convergence of stochastic processes.* Springer, 1984.

[44] P.J. Rousseeuw, A.M. Leroy, and J. Wiley. *Robust regression and outlier detection*, volume 3. Wiley Online Library, 1987.

[45] N. Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13(1):145–147, 1972.

[46] L. J. Schulman. Clustering for edge-cost minimization. In *Proc. 32nd STOC*, pages 547–555, 2000.

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

[48] V. N. Vapnik. Inductive principles of the search for empirical dependences. In *Proc. 2nd Annual Workshop on Computational Learning Theory*, 1989.

[49] 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.

[50] Z. Zhang. M-estimators. http://research.microsoft.com/en-us/um/people/zhang/INRIA/Publis/Tutorial-Estim/node20.html.