A Quantization Framework for Smoothed Analysis of Euclidean Optimization Problems

Radu Curticapean[1] and Marvin Künnemann[1,2]

[1] Universität des Saarlandes, Saarbrücken, Germany
curticapean@cs.uni-saarland.de
[2] Max-Planck-Institut für Informatik, Saarbrücken, Germany
marvin@mpi-inf.mpg.de

Abstract. We consider the smoothed analysis of Euclidean optimization problems. Here, input points are sampled according to density functions that are bounded by a sufficiently small smoothness parameter ϕ. For such inputs, we provide a general and systematic approach that allows to design linear-time approximation algorithms whose output is asymptotically optimal, both in expectation and with high probability.

Applications of our framework include maximum matching, maximum TSP, and the classical problems of k-means clustering and bin packing. Apart from generalizing corresponding average-case analyses, our results extend and simplify a polynomial-time probable approximation scheme on multidimensional bin packing on ϕ-smooth instances, where ϕ is constant (Karger and Onak, SODA 2007).

Both techniques and applications of our rounding-based approach are orthogonal to the only other framework for smoothed analysis on Euclidean problems we are aware of (Bläser et al., Algorithmica 2012).

1 Introduction

Smoothed analysis has been introduced by Spielman and Teng [26] to give a theoretical foundation for analyzing the practical performance of algorithms. In particular, this analysis paradigm was able to provide an explanation why the simplex method is observed to run fast in practice despite its exponential worst-case running time.

The key concept of smoothed analysis, i.e., letting an adversary choose worst-case distributions of bounded "power" to determine input instances, is especially well-motivated in a Euclidean setting. Here, input points are typically determined by physical measurements, which are subject to an inherent inaccuracy, e.g., from locating a position on a map. For clustering problems, it is often even implicitly assumed that the points are sampled from unknown probability distributions which are sought to be recovered.

Making the mentioned assumptions explicit, we call a problem *smoothed tractable* if it admits a linear-time algorithm with an approximation ratio that is bounded by $1 - o(1)$ with high probability over the input distribution specified by the adversary. Such an approximation performance is called *asymptotically*

H.L. Bodlaender and G.F. Italiano (Eds.): ESA 2013, LNCS 8125, pp. 349–360, 2013.

optimal. We provide a unified approach to show that several Euclidean optimization problems are smoothed tractable, which sheds light onto the properties that render a Euclidean optimization problem likely to profit from perturbed input.

We employ the *one-step model*, a widely-used and very general perturbation model, which has been successfully applied to analyze a number of algorithms [9,10,11,15]. In this model, an adversary chooses probability densities on the input space, according to which the input instance is drawn. To prevent the adversary from modeling a worst-case instance too closely, we bound the density functions from above by a parameter ϕ. Roughly speaking, for large ϕ, we expect the algorithm to perform almost as bad as on worst-case instances. Likewise, choosing ϕ as small as possible requires the adversary to choose the uniform distribution on the input space, corresponding to an average-case analysis. Thus, the adversarial power ϕ serves as an interpolation parameter between worst and average case.

Formally, given a set of feasible distributions $\mathcal{F}$ that depends on ϕ, and a performance measure t, we define the smoothed performance of an algorithm under the perturbation model $\mathcal{F}$ as $\max_{f_1,\ldots,f_n \in \mathcal{F}} \mathrm{E}_{(X_1,\ldots,X_n)\sim(f_1,\ldots,f_n)}[t(X_1,\ldots,X_n)]$. In this work, we will be concerned with analyzing the smoothed approximation ratio, as well as bounds on the approximation ratio that hold with high probability over the perturbations.

For given ϕ, we require the density functions chosen by the adversary to be bounded by ϕ. For real-valued input, this includes the possibility to add uniform noise in an interval of length $1/\phi$ or Gaussian noise with variance $\sigma \in \Theta(1/\phi)$. In the Euclidean case, the adversary could, e.g., specify for each point a box of volume at least $1/\phi$, in which the point is distributed uniformly.

Related Work. Recently, Bläser, Manthey and Rao [10] established a framework for analyzing the expectation of both running times and approximation ratios for some partitioning algorithms on so-called *smooth* and *near-additive* functionals. We establish a substantially different framework for smoothed analysis on a general class of Euclidean functionals that is disjoint to the class of smooth and near-additive functionals (see Section 7 for further discussion). We contrast both frameworks by considering the maximization counterparts of two problems studied in [10], namely Euclidean matching and TSP. Our algorithms have the advantage of featuring deterministic running times and providing asymptotically optimal approximation guarantees both in expectation and with high probability.

All other related work is problem-specific and will be described in the corresponding sections. As an exception, we highlight the result of Karger and Onak [22], who studied bin packing. To the best of our knowledge, this is the only problem that fits into our framework and has already been analyzed under perturbation. In this paper, a linear-time algorithm for bin packing was given that is asymptotically optimal on instances smoothed with any constant ϕ. We provide a new, conceptually simpler rounding method and analysis that replaces a key step of their algorithm and puts the reasons for its smoothed tractability into a more general context.

Table 1. All (near) linear-time algorithms derived in our framework

problem	running time	restriction on adversary power
MaxM	$O(n)$	$\phi \in o(\sqrt[4]{n})$ or $\phi \in o(n^{\frac{1}{2}\frac{d}{d+2}-\varepsilon})$
MaxTSP	$O(n)$	$\phi \in o(\sqrt[4]{n})$ or $\phi \in o(n^{\frac{1}{2}\frac{d}{d+2}-\varepsilon})$
KMeans	$O(n)$	$k\phi \in o(n^{\frac{1}{2}\frac{1}{kd+1}\frac{d}{d+1}})$
BP$_1$	$O(n \log n)$	$\phi \in o(n^{1-\varepsilon})$
BP$_d$	$O(n)$	$\phi \in o\left(\sqrt[d(d+1)]{\log \log n / \log^{(3)} n}\right)$

Our Results. We provide very fast and simple approximation algorithms on sufficiently smoothed inputs for the following problems: The maximum Euclidean matching problem MaxM, the maximum Euclidean Traveling Salesman problem MaxTSP, the k-means clustering problem KMeans where k denotes the number of desired clusters and is part of the input, and the d-dimensional bin packing problem BP$_d$. The approximation ratio converges to one with high probability over the random inputs. Additionally, all of these algorithms can be adapted to yield asymptotically optimal expected approximation ratios as well. This generalizes corresponding average-case analysis results [14,23].

Almost all our algorithms allow trade-offs between running time and approximation performance: By choosing a parameter p within its feasible range, we obtain algorithms of running time $O(n^p)$, whose approximation ratio converges to 1 as $n \to \infty$, provided that ϕ small enough, where the restriction on ϕ depends on p. The special case of linear-time algorithms is summarized in Table 1.

2 Preliminaries

Given an n-tuple of density functions $f = (f_1, \ldots, f_n)$ and random variables $X = (X_1, \ldots, X_n)$, we write $X \sim f$ for drawing X_i according to f_i for $1 \le i \le n$. We call $Y = (Y_1, \ldots, Y_n)$ a *δ-rounding* of X if $\|X_i - Y_i\| \le \delta$ for all $1 \le i \le n$. For a given X, let $\mathcal{Y}_X^\delta := \{Y \mid \|X_i - Y_i\| \le \delta\}$ be the set of δ-roundings of X.

We will analyze Euclidean functionals $F : ([0,1]^d)^* \to \mathbb{R}$, denoting the dimension of the input space by $d \in \mathbb{N}$. For formalizing the perturbation model, let $\phi : \mathbb{N} \to [1, \infty)$ be an arbitrary function measuring the adversary's power. For better readibility, we usually write ϕ instead of $\phi(n)$. We define $\mathcal{F}_\phi$ to be the set of *feasible* probability density functions $f : [0,1]^d \to [0, \phi]$.

Note that if $\phi = 1$, the set $\mathcal{F}_\phi$ only consists of the uniform distribution on $[0,1]^d$. If however $\phi = n$, the adversary may specify disjoint boxes for each point. Intuitively, to obtain a particular worst-case instance, the adversary would need to specify Dirac delta functions, which corresponds figuratively to setting ϕ to infinity. Observe also that already $\phi \in \omega(1)$ suffices to let all possible locations of a fixed point X_i converge to a single point for n tending to infinity, hence we believe that a superconstant ϕ is especially interesting to analyze.

For a given Euclidean functional F, we analyze the approximation ratio ρ of approximation algorithms ALG. If the functional is induced by an optimization problem, we do not focus on constructing a feasible approximate solution, but rather on computing an approximation of the objective value. However, we adopt this simplification only for clarity of presentation. Each of the discussed algorithms can be tuned such that it also outputs a feasible approximate *solution* for the underlying optimization problem. The approximation ratio on instance X is defined as $\rho(X) = \min\left\{\frac{\mathrm{ALG}(X)}{F(X)}, \frac{F(X)}{\mathrm{ALG}(X)}\right\}$, which allows to handle both maximization and minimization problems at once.

For analyzing running times, we assume the word RAM model of computation and reveal real-valued input by reading in words of at least $\log n$ bits in unit time per word. We call an approximation algorithm a *probable $g_\phi(n)$-approximation* on smoothed instances if $\rho(X) \geq g_\phi(n)$ with high probability, i.e., with probability $1 - o(1)$, when X is drawn from $\mathcal{F}_\phi^n$. The algorithms derived in our framework feature deterministic running times $t(n) \in \mathrm{poly}(n)$ and asymptotically optimal approximation ratios $g_\phi(n)$, i.e., $g_\phi(n) \to 1$ for $n \to \infty$ if ϕ is small enough. For such choices of ϕ, each of our algorithms induces a (non-uniform) probable polynomial-time approximation scheme (PTAS) on smoothed instances.

3 Framework

Our framework builds on the notion of *quantizable* functionals. These are functionals that admit fast approximation schemes on perturbed instances using general rounding strategies. The idea is to round an instance of n points to a *quantized instance* of $\ell \ll n$ points, each equipped with a multiplicity. This quantized input has a smaller problem size, which allows to compute an approximation faster than on the original input. However, the objective function needs to be large enough to make up for the loss due to rounding.

We aim at a trade-off between running time and approximation performance. As it will turn out, varying the number $\ell(n)$ of quantized points on an instance of n points makes this possible. Thus, we keep the function ℓ variable in our definition. On instances of size n, we will write $\ell := \ell(n)$ for short.

Definition 1. *Let $d \geq 1$ and $\mathcal{F}$ be a family of probability distributions $[0,1]^d \to \mathbb{R}_{\geq 0}$. Let $t, R : \mathbb{N} \to \mathbb{R}$ and $Q \in \mathbb{R}$. We say that a Euclidean functional $F : ([0,1]^d)^* \to \mathbb{R}_{\geq 0}$ is* t-time (R,Q)-quantizable *with respect to $\mathcal{F}$, if there is a* quantization algorithm A *and an* approximation functional $g : ([0,1]^d \times \mathbb{N})^* \to \mathbb{R}$ *with the following properties. For any function ℓ satisfying $\ell \in \omega(1)$ and $\ell \in o(n)$,*

1. *The quantization algorithm A maps, in time $O(n)$, a collection of points $X = (X_1, \ldots, X_n) \in [0,1]^{dn}$ to a multiset $A(X) = X' = ((X'_1, n_1), \ldots, (X'_\ell, n_\ell))$, the* quantized input, *with $X'_i \in [0,1]^d$.*
2. *The approximation functional g is computable in time $t(\ell)$ and, for any $f \in \mathcal{F}^n$, fulfills $\Pr_{X\sim f}[|F(X) - g(A(X))| \leq nR(\ell)] \in 1 - o(1)$.*
3. *For any $f \in \mathcal{F}^n$, we have $\Pr_{x\sim f}[F(X) \geq nQ] \in 1 - o(1)$.*

The following theorem states that quantizable functionals induce natural approximation algorithms on smoothed instances. We can thus restrict our attention to finding criteria that make a functional quantizable.

Theorem 2. *Let $\mathcal{F}$ be a family of probability distributions and F be $t(\ell)$-time $(R(\ell), Q)$-quantizable with respect to $\mathcal{F}$. Then for every ℓ with $\ell \in \omega(1)$ and $\ell \in o(n)$, there is an approximation algorithm* ALG *with the following property. For every $f \in \mathcal{F}^n$, the approximation* ALG(X) *on the instance X drawn from f is $(1 - \frac{R(\ell)}{Q})$-close to $F(X)$ with high probability. The approximation can be computed in time $O(n + t(\ell))$.*

For all problems considered here, we also design algorithms whose *expected* approximation ratio converges to optimality in the sense that both $\mathrm{E}[\rho] \to 1$, as already achieved by the framework algorithm, and $\mathrm{E}[\rho^{-1}] \to 1$, the desired guarantee for *minimization* problems, which we ensure using auxiliary algorithms. A sufficient auxiliary algorithm for F is a linear-time algorithm approximating F within a constant factor $0 < c < 1$. Outputting the better solution of our framework algorithm and the c-approximation does not increase the order of the running time, but achieves an approximation ratio of $1 - o(1)$ with probability $1 - o(1)$ due to the previous theorem, yielding $\mathrm{E}[\rho] \to 1$, and still provides a constant approximation ratio on the remaining instances sampled with probability $o(1)$. Thus, additionally $\mathrm{E}[\rho^{-1}] \leq (1 - o(1))\frac{1}{1-o(1)} + o(1)c^{-1} \to 1$ holds.

We respresent multisets of points either as $X' = ((X'_1, n_1), \ldots, (X'_\ell, n_\ell)) \in ([0,1]^d \times \mathbb{N})^\ell$ or expand this canonically to a tuple $X' \in ([0,1]^d)^*$. By $T : ([0,1]^d \times \mathbb{N})^* \to ([0,1]^d)^*$ we denote the transformation that maps the former representation to the latter.

4 Grid Quantization

Our first method for verifying quantizability is grid quantization. Here, the idea is to round the input to the centers of grid cells, where the coarseness of the grid is chosen according to the desired number of distinct points. This method works well for functionals that allow for fast optimal computations on their high-multiplicity version and provide a large objective value on the chosen perturbation model.

Theorem 3. *(Grid quantization) Let $d \geq 1$, $Q \in \mathbb{R}$ and $\mathcal{F}$ be a family of probability distributions $[0,1]^d \to \mathbb{R}_{\geq 0}$. Let $F : ([0,1]^d)^* \to \mathbb{R}_{\geq 0}$ be a Euclidean functional with the following properties.*

1. *On the quantized input $X' = ((X'_1, n_1), \ldots, (X'_\ell, n_\ell))$, the value $F(T(X'))$ can be computed in time $t(\ell) + O(\sum_{i=1}^{\ell} n_i)$.*
2. *There is a constant C such that w.h.p., the functional differs by at most $C\delta n$ on all δ-roundings of an instance X drawn from any $f \in \mathcal{F}^n$. Formally, for each $\delta > 0$ we require $\Pr_{X\sim f}\left[\forall Y \in \mathcal{Y}_X^\delta : |F(X) - F(Y)| \leq C\delta n\right] \in 1 - o(1)$.*
3. *For each $f \in \mathcal{F}^n$, it holds that $\Pr_{X\sim f}\left[F(X) \geq nQ\right] \in 1 - o(1)$.*

Then F is $t(\ell)$-time $(O(\ell^{-\frac{1}{d}}), Q)$-quantizable with respect to $\mathcal{F}$.

In this section, we apply the framework to two Euclidean maximization problems, namely maximum matching and maximum TSP. Both problems have already been analyzed in the average-case world, see, e.g., an analysis of the Metropolis algorithm on maximum matching in [28]. We generalize the result of Dyer et al. [14], who proved the asymptotic optimality of two simple partitioning heuristics for maximum matching and maximum TSP on the uniform distribution in the unit square. However, in contrast to our approach, their partitioning methods typically fail if the points are not identically distributed.

4.1 Maximum Matching and Maximum TSP

Let MaxM(X) denote the maximum weight of a matching of the points $X \subseteq [0,1]^d$, where the weight of a matching M is defined as $\sum_{\{u,v\}\in M} \|u-v\|$. For simplicity, we assume that $|X|$ is even. For the more general problem of finding maximum weighted matchings on *general* graphs with non-integer weights, the fastest known algorithm due to Gabow [19] runs in time $O(mn + n^2 \log n)$.

We aim to apply Theorem 3, for which we only need to check three conditions. The rounding condition (2) is easily seen to be satisfied by a straight-forward application of the triangle inequality. The lower bound condition (3) is satisfied by the following lemma.

Lemma 4. *Let $f \in \mathcal{F}_\phi^n$. Some $\gamma > 0$ satisfies* $\Pr_{X\sim f}\left[\mathrm{MaxM}(X) < \frac{\gamma n}{\sqrt[d]{\phi}}\right] \le e^{-\Omega(n)}$.

We call the task of computing a functional on quantized inputs *quantized version* of the functional. In the case of MaxM, an algorithm for b-matchings from [1] can be exploited, satisfying condition (1).

Lemma 5. *The quantized version of* MaxM *can be computed in time $O(\ell^4 + \ell^3 \log n)$, where $n = \sum_{i=1}^{\ell} n_i$.*

These observations immediately yield the following result.

Theorem 6. MaxM *is $O(\ell^4)$-time $(O(1/\sqrt[d]{\ell}), \Omega(1/\sqrt[d]{\phi}))$-quantizable w. r. t. $\mathcal{F}_\phi$. Hence, for $1 \le p < 4$, there is a $O(n^p)$-time probable $(1 - O(\sqrt[d]{\phi/n^{p/4}}))$-approximation to* MaxM *for instances drawn according to some $f \in \mathcal{F}_\phi^n$. This is asymptotically optimal on smoothed instances with $\phi \in o(n^{p/4})$.*

Interestingly, the restriction on ϕ is independent of the dimension. Note that only $p < 3$ is reasonable, since deterministic cubic-time algorithms for exactly computing MaxM exist. Furthermore, as described in Section 3, an algorithm with an asymptotically optimal expected approximation ratio can be designed. E.g., we might utilize a simple greedy linear-time $\frac{1}{2}$-approximation for MaxM [4].

Similar ideas can be applied to the maximum TSP problem. For $d \ge 2$, define MaxTSP(X) as the maximum weight of a Hamiltonian cycle on $X \subseteq [0,1]^d$, where the weight of a Hamiltonian cycle C is defined as $\sum_{\{u,v\}\in C} \|u-v\|$. The problem is NP-hard (proven for $d \ge 3$ in [7], conjectured for $d = 2$,) but admits a PTAS, cf. [8,7]. According to [16], these algorithms are not practical. They stress the need for (nearly) linear-time algorithms.

Theorem 7. *Let* $1 \leq p \leq 4d/(d+1)$ *and* $f \in \mathcal{F}_\phi^n$ *. On instances drawn from* f*, there is a* $O(n^p)$*-time computable probable* $(1 - O(\sqrt[d]{\phi/n^{p/4}}))$*-approximation for* MaxTSP*. This is asymptotically optimal for* $\phi \in o(n^{p/4})$*.*

Since MaxM is a $\frac{1}{2}$-approximation to MaxTSP, the greedy linear-time computable $\frac{1}{2}$-approximation to MaxM is a $\frac{1}{4}$-approximation to MaxTSP and thus provides an adapted algorithm with asymptotically optimal *expected* approximation ratio for $\phi \in o(n^{p/4})$.

5 Balanced Quantization

Grid quantization proves useful for problems in which algorithms solving the high-multiplicity version are available. To solve even more problems, this section establishes a more careful quantization step yielding *balanced* instances, i.e., instances in which each of the distinct points occurs the same number of times. This has direct applications to k-means clustering and similar problems. In general, this method can be applied to problems in which the objective scales controllably when duplicating all points.

Theorem 8. *Let* $\ell : \mathbb{N} \to \mathbb{N}$ *with* $\ell \in \omega(1)$ *and* $\ell \in o(n)$*. There is a function* $\ell' : \mathbb{N} \to \mathbb{N}$ *such that for each* $n \in \mathbb{N}$ *and* $X = (X_1, \ldots, X_n) \in [0,1]^{dn}$*, we can find, in linear time, a family of* $\ell'(n)$ *cells, i.e., collections of points* $C_1, \ldots, C_{\ell'(n)}$ *with the following properties.*

1. $\frac{\ell'(n)}{\ell(n)} \to 1$ *(we obtain* ℓ *cells asymptotically),*
2. $|C_i| = |C_j|$ *for all* $1 \leq i, j \leq \ell'(n)$ *(all cells are of equal size),*
3. $n - \sum_{i=1}^{\ell'(n)} |C_i| \in O(\frac{n}{\ell^{1/(d+1)}})$ *(almost all points are covered),*
4. $\mathrm{diam}(C_i) \in O(\frac{1}{\ell^{1/(d+1)}})$ *(each element in a cell represents this cell well).*

For some problems, an instance in which every distinct point occurs equally often can be reduced to its distinct points only. In the following, we exploit this property by applying the previous theorem to k-means clustering. The method also allows for improving the results on maximum matching and maximum TSP. We defer the details of this to a full version of this article.

5.1 K-Means Clustering

Let $d \geq 2$ and $k \in \mathbb{N}$. We define KMeans(X, k) to be the k-means objective on the points X where k is the desired number of clusters, i.e.,

$$\mathrm{KMeans}(X, k) = \min_{C_1 \dot\cup \cdots \dot\cup C_k = X} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \text{, where } \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x.$$

K-Means clustering is an important problem in various areas including machine learning and data mining. If either k or d is part of the input, it is NP-hard [13,24]. However, a popular heuristic, the k-means algorithm, usually runs fast on real-world instances despite its worst-case exponential running time. This is substantiated by results proving a polynomial smoothed running time of the k-means method under Gaussian perturbations [3,2]. In terms of solution quality, however, such a heuristic can perform poorly.

Consequently, k-means clustering has also received considerable attention concerning the design of fast deterministic approximation schemes. There exist linear-time asymptotically optimal algorithms, e.g., PTASs with running time $O(nkd+d\text{poly}(k/\varepsilon)+2^{\tilde{O}(k/\varepsilon)})$ [17] and $O(ndk+2^{(k/\varepsilon)^{O(1)}}d^2n^\sigma)$ for any $\sigma > 0$ [12]. Treating the dimension as a constant as we do, [20] showed how to compute a $(1+\varepsilon)$-approximation in time $O(n + k^{k+2}\varepsilon^{-(2d+1)k}\log^{k+1} n \log^k \frac{1}{\varepsilon})$. On a side note, a different perturbation concept has been applied to k-means clusterings in [5]. They restrict their attention to input instances which, when perturbed, maintain the same partitioning of the input points in the optimal clustering.

Define the center of mass of a set C as $\text{cm}(C) := \frac{1}{|C|}\sum_{c\in C} c$ and consider $X' = ((X_1, n_1), \ldots, (X_{\ell'}, n_{\ell'})) = ((\text{cm}(C_1), |C_1|), \ldots, (\text{cm}(C_{\ell'}), |C_{\ell'}|))$, a quantized instance using the cells $C_1, \ldots, C_{\ell'}$ obtained by applying Theorem 8. It holds that $n_1 = n_2 = \cdots = n_{\ell'} = w$. Let $Y = T(X') = (Y_1, \ldots, Y_{n'})$, where Y_i is the rounded version of X_i. Note that the number n' of points in the rounded instance is potentially smaller than n, since points may be lost in the application of Theorem 8.

Lemma 9. *There is a real Δ with $|\text{KMeans}(X,k) - \text{KMeans}(Y,k)| \le \frac{\Delta n}{\ell^{1/(d+1)}}$.*

After establishing that rounding the input does not affect the objective value too much, the following lemma enables us to reduce the instance size significantly.

Lemma 10. *Consider $X = ((X_1, w), \ldots, (X_\ell, w))$ and $Z = ((X_1, 1), \ldots, (X_\ell, 1))$. It holds that $\text{KMeans}(T(X), k) = w\text{KMeans}(T(Z), k)$.*

It is left to give a lower bound on the objective value. Note that for other minimization functionals like minimum Euclidean matching or TSP, already the uniform distribution on the unit cube achieves an objective value of only $O(n^{(d-1)/d})$ [27], making the framework inapplicable in this case (for a more detailed discussion, refer to Section 7).

Lemma 11. *Let $f \in \mathcal{F}_\phi^n$ and $k \in o(\frac{n}{\log n})$. There exists some constant $\gamma > 0$ such that $\Pr_{X\sim f}\left[\text{KMeans}(X,k) < \frac{\gamma n}{(k\phi)^{2/d}}\right] = e^{-\Omega(n)}$.*

For solving the smaller instance obtained by quantization, two approaches are reasonable. The first is to compute an optimal solution in time $O(n^{kd+1})$ [21] and results in the following theorem.

Theorem 12. *For any $k \in o(n/\log n)$, the functional $\text{KMeans}(X,k)$ is $O(\ell^{kd+1})$-time $(O(\ell^{-1/(d+1)}), \Omega((k\phi)^{-2/d}))$-quantizable with respect to $\mathcal{F}_\phi$. Consequently, for $k \in O(\log n/\log\log n)$ and $1 \le p \le kd+1$, there is a $O(n^p)$-time computable probable $\left(1 - O\left(\frac{(k\phi)^{2/d}}{n^{\frac{p}{(d+1)(kd+1)}}}\right)\right)$-approximation for $\text{KMeans}(X,k)$ on smoothed instances.*

Note that this is asymptotically optimal if $\phi \in o(\sqrt[c]{n})$ with $c = 2(1 + 1/d)(kd + 1)/p$ if $k \in O(1)$, or more generally, if $k\phi \in o(n^{\frac{pd}{2(d+1)(kd+1)}})$. Using existing linear-time approximation schemes, also an asymptotically optimal expected approximation ratio can be obtained for the same values of ϕ. Our framework algorithm applies even for large values of k, e.g., $k = \log n / \log\log n$, in which case known deterministic approximation schemes require superlinear time. However, for small k, incorporating such an approximation scheme into our algorithm yields a further improvement of the previous theorem. The details for this second approach are deferred to a full version of this article.

6 Bin Packing

For $X = (X_1, \ldots, X_n) \in [0, 1]^{dn}$, define $\mathrm{BP}_d(X)$ to be the minimum number of bins of volume one needed to pack all elements. An item $X = (x_1, \ldots, x_d)$ is treated as a d-dimensional box, where x_i is its side length in dimension i. Items must not be rotated and must be packed such that their interior is disjoint.

In the following, we extend the result of Karger and Onak [22], who gave linear-time asymptotically optimal approximation algorithms for smoothed instances with $\phi \in O(1)$ and for instances with i.i.d. points drawn from a *fixed* distribution. These tractability results are highly interesting due to the fact that there is not even an asymptotic polynomial-time approximation scheme (APTAS) solving the two-dimensional bin packing problem unless $\mathsf{P} = \mathsf{NP}$, cf. [6].

Karger and Onak's approach appears rather problem-specific, whereas our solution embeds nicely into our framework. The main difference of our approach lies in a much simpler rounding routine and analysis, after which we solve the problem exactly as in their distribution-oblivious algorithm. Note that their algorithm is supplied with a desired approximation performance $\varepsilon > 0$ and suceeds with probability of $1 - 2^{-\Omega(n)}$. Although not stated for this case, we believe that their algorithm may also apply to superconstant choices of ϕ, at a cost of decreasing the success probability. We feel that our analysis offers more insights on the reasons why bin packing is smoothed tractable by putting it into the context of our general framework.

Consider first the case that $d = 1$. Unless $\mathsf{P} = \mathsf{NP}$, BP_1 does not admit a $\frac{3}{2}$-approximation. However, asymptotic polynomial approximation schemes exist [18], i.e., $(1-\varepsilon)$-approximations on instances with a sufficiently large objective value. These approximation schemes have an interesting connection to smoothed analysis due to the following property.

Lemma 13. *For $f \in \mathcal{F}_\phi^n$, there is a $\gamma > 0$ with* $\Pr_{X \sim f}\left[\mathrm{BP}_d(X) < \gamma \frac{n}{\phi^d}\right] \le e^{-\Omega(n)}$.

Using this bound on the objective value, we show an example of how to transform an APTAS into a PTAS on smoothed instances. Plotkin et al. [25] have shown how to compute, in time $O(n \log \varepsilon^{-1} + \varepsilon^{-6} \log^6 \varepsilon^{-1})$, a solution with an objective value of at most $(1+\varepsilon)\mathrm{BP}_1(X) + O(\varepsilon^{-1} \log \varepsilon^{-1})$. We derive

$$\rho = \frac{\mathrm{ALG}}{\mathrm{BP}_1(X)} \le (1+\varepsilon) + O\left(\frac{\varepsilon^{-1}\log\varepsilon^{-1}}{\mathrm{BP}_1(X)}\right) \le (1+\varepsilon) + O\left(\frac{\phi\varepsilon^{-1}\log\varepsilon^{-1}}{n}\right),$$

where the last inequality holds w.h.p. over the perturbation of the input. Setting $\varepsilon := \log n/n^{\delta}$ with some $\delta < 1/6$ yields a running time of $O(n \log n)$ with an approximation ratio $\leq 1 + \log n/n^{\delta} + O(\phi/n^{1-\delta})$. Consequently, there is a linear-time asymptotically optimal approximation algorithm on instances smoothed with $\phi \in o(n^{1-\delta})$ for any $\delta > 0$. Unfortunately, this approach is not easily generalizable to $d > 1$, since already for $d = 2$, no APTAS exists unless $\mathsf{P} = \mathsf{NP}$, as shown in [6]. Nevertheless, the problem is quantizable in our framework.

We say that a single item $X = (x_1, \ldots, x_d)$ fits in a box $B = (b_1, \ldots, b_d)$ if $x_i \leq b_i$ for all $1 \leq i \leq d$. In this case, we write $X \sqsubseteq B$, adopting the notation of [22]. Regarding an item as a box as well, this relation is transitive and any feasible packing containing Y induces a feasible packing by replacing Y with X. Thus, bin packing admits the monotonicity property that for each $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ with $X_i \sqsubseteq Y_i$, it holds that $\mathrm{BP}_d(X) \leq \mathrm{BP}_d(Y)$.

To apply the quantization framework, we require a suitable bound on the rounding errors. Unlike for MaxM and MaxTSP, no deterministic bound of $n\delta$ is possible for a δ-rounding: Let the instance $X^{(n)}$ consist of n copies of $(\frac{1}{2}, \ldots, \frac{1}{2})$. Packing 2^d of the items per bin results in zero waste, hence $\mathrm{BP}_d(X^{(n)}) = n/2^d$. However, for *any* $\delta > 0$, the δ-rounding Y_n consisting of n copies of $(\frac{1}{2} + \frac{\delta}{\sqrt{d}}, \ldots, \frac{1}{2} + \frac{\delta}{\sqrt{d}})$ has an objective value of $\mathrm{BP}_d(Y^{(n)}) = n = 2^d \mathrm{BP}_d(X^{(n)})$. Thus, a smoothed analysis of the rounding error is necessary.

Lemma 14. *For $f \in \mathcal{F}_{\phi}^{n}$ and $t > 0$,*

$$\Pr_{X \sim f}\left[\forall Y \in \mathcal{Y}_X^t : |\mathrm{BP}_d(X) - \mathrm{BP}_d(Y)| > 2ntd\phi\right] \leq 2\exp(-2n(dt\phi)^2).$$

Note that this probability tends to zero if $t \in \omega(\frac{1}{\phi\sqrt{n}})$. Since grid quantization rounds the points to ℓ distincts points by moving each item by at most $t = \sqrt{d}\ell^{-d}$, the requirement $\ell \in o(n)$ even implies that $t \in \omega(\frac{1}{\sqrt{n}})$ for $d \geq 2$.

Solving the high-multiplicity version of the one-dimensional case has been a key ingredient in approximation schemes for this problem since the first APTAS by [18]. The following lemma from [22] solves the multi-dimensional case.

Lemma 15. *Let $X' = ((X_1, n_1), \ldots, (X_\ell, n_\ell))$ be a quantized input with $X_i \in [\delta, 1]^d$. Then $\mathrm{BP}_d(T(X'))$ can be computed in time $O(f(\ell, \delta)\mathrm{polylog}(n))$ where $n := \sum_{i=1}^{\ell} n_i$, $f(\ell, \delta)$ is independent of n and $f(\ell, 1/\sqrt[d]{\ell}) \in 2^{\ell^{O(\ell)}}$.*

Observe that each coordinate of the quantized points obtained by grid quantization is at least $1/(2\sqrt[d]{\ell})$, since these points are the centroids of cubic cells of side length $1/\sqrt[d]{\ell}$. Hence, applying a slightly stronger form of the grid quantization theorem yields the following result using Lemmas 13, 14 and 15.

Theorem 16. *For $d \geq 2$, BP_d is $O(2^{\ell^{O(\ell)}})$-time $(O(\frac{\phi}{\sqrt[d]{\ell}}), \Omega(\frac{1}{\phi^d}))$-quantizable w.r.t. $\mathcal{F}_{\phi}$.*

Consequently, there is a linear-time probable $(1 - O(\phi^{d+1}/\sqrt[d]{\log\log n/\log^{(3)} n}))$-approximation. Hence, BP_d can be computed asymptotically exactly in time

$O(n)$ if $\phi \in o(\sqrt[d(d+1)]{\log\log n / \log^{(3)} n})$. Here, allowing superlinear time has no effect on the admissible adversarial power. Furthermore, since BP_d can be trivially approximated by a factor of n and the success probability of our algorithm is of order $1 - \exp(-\Omega(n^{1-\varepsilon}))$, asymptotically optimal expected approximation ratios can be obtained for the same values of ϕ.

7 Concluding Remarks

Generalizing previous rounding-based approaches, we demonstrate that the general solution technique of quantization performs well on Euclidean optimization problems in the setting of smoothed analysis. We are optimistic that our framework can also be applied to disk covering and scheduling problems.

Note that our approach is orthogonal to the framework for smooth and near-additive Euclidean functionals due to Bläser et al. [10]: A smooth Euclidean functional F on n points can be bounded by $O(n^{1-1/d})$ by definition of smoothness. Hence, it can never compensate for the rounding error of at least $\Omega(\ell^{-1/d})$ per point that our quantization methods induce, as quantization is only reasonable for $\ell \leq n$ and consequently, the total rounding error amounts to $\Omega(n^{1-1/d})$. Conversely, if a functional is large enough to compensate for rounding errors induced by quantization, it cannot be smooth. Thus, for any Euclidean functional, at most one of both frameworks is applicable.

References

1. Anstee, R.P.: A polynomial algorithm for b-matchings: An alternative approach. Information Processing Letters 24(3), 153–157 (1987)
2. Arthur, D., Manthey, B., Röglin, H.: Smoothed analysis of the k-means method. Journal of the ACM 58(5), 19:1–19:31 (2011)
3. Arthur, D., Vassilvitskii, S.: Worst-case and smoothed analysis of the ICP algorithm, with an application to the k-means method. SIAM Journal on Computing 39(2), 766–782 (2009)
4. Avis, D.: A survey of heuristics for the weighted matching problem. Networks 13(4), 475–493 (1983)
5. Awasthi, P., Blum, A., Sheffet, O.: Center-based clustering under perturbation stability. Information Processing Letters 112(1-2), 49–54 (2012)
6. Bansal, N., Correa, J.É.R., Kenyon, C., Sviridenko, M.: Bin packing in multiple dimensions: Inapproximability results and approximation schemes. Mathematics of Operations Research 31, 31–49 (2006)
7. Barvinok, A., Fekete, S.P., Johnson, D.S., Tamir, A., Woeginger, G.J., Woodroofe, R.: The geometric maximum traveling salesman problem. J. ACM 50(5), 641–664 (2003)
8. Barvinok, A.I.: Two algorithmic results for the traveling salesman problem. Mathematics of Operations Research 21(1), 65–84 (1996)
9. Beier, R., Vöcking, B.: Typical properties of winners and losers in discrete optimization. SIAM Journal on Computing 35(4), 855–881 (2006)
10. Bläser, M., Manthey, B., Raghavendra Rao, B.V.: Smoothed Analysis of Partitioning Algorithms for Euclidean Functionals. Algorithmica 66(2), 397–418 (2013)

11. Boros, E., Elbassioni, K., Fouz, M., Gurvich, V., Makino, K., Manthey, B.: Stochastic mean payoff games: Smoothed analysis and approximation schemes. In: Aceto, L., Henzinger, M., Sgall, J. (eds.) ICALP 2011, Part I. LNCS, vol. 6755, pp. 147–158. Springer, Heidelberg (2011)
12. Chen, K.: On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. SIAM Journal on Computing 39(3), 923–947 (2009)
13. Dasgupta, S.: The hardness of k-means clustering. Technical report cs2007-0890, University of California, San Diego (2007)
14. Dyer, M.E., Frieze, A.M., McDiarmid, C.J.H.: Partitioning heuristics for two geometric maximization problems. Operations Research Letters 3(5), 267–270 (1984)
15. Englert, M., Röglin, H., Vöcking, B.: Worst case and probabilistic analysis of the 2-opt algorithm for the TSP: Extended abstract. In: 18th Ann. ACM-SIAM Symp. on Discrete Algorithms, SODA 2007, pp. 1295–1304. SIAM (2007)
16. Fekete, S.P., Meijer, H., Rohe, A., Tietze, W.: Solving a "hard" problem to approximate an "easy" one: Heuristics for maximum matchings and maximum traveling salesman problems. ACM J. on Experimental Algorithmics. 7, 11 (2002)
17. Feldman, D., Monemizadeh, M., Sohler, C.: A PTAS for k-means clustering based on weak coresets. In: 23rd Ann. Symp. on Computational Geometry, SCG 2007, pp. 11–18. ACM (2007)
18. Fernandez de la Vega, W., Lueker, G.: Bin packing can be solved within $1 + \epsilon$ in linear time. Combinatorica 1(4), 349–355 (1981)
19. Gabow, H.N.: An efficient implementation of Edmonds' algorithm for maximum matching on graphs. Journal of the ACM 23(2), 221–234 (1976)
20. Har-Peled, S., Mazumdar, S.: On coresets for k-means and k-median clustering. In: 36th Ann. ACM Symp. on Theory of Computing, STOC 2004, pp. 291–300 (2004)
21. Inaba, M., Katoh, N., Imai, H.: Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering (extended abstract). In: 10th Annual Symp. on Computational Geometry, SCG 1994, pp. 332–339 (1994)
22. Karger, D., Onak, K.: Polynomial approximation schemes for smoothed and random instances of multidimensional packing problems. In: 18th Ann. ACM-SIAM Symp. on Discrete Algorithms, SODA 2007, pp. 1207–1216 (2007)
23. Karp, R.M., Luby, M., Marchetti-Spaccamela, A.: A probabilistic analysis of multidimensional bin packing problems. In: 16th Annual ACM Symp. on Theory of Computing, STOC 1984, pp. 289–298. ACM, New York (1984)
24. Mahajan, M., Nimbhorkar, P., Varadarajan, K.: The planar k-means problem is NP-hard. Theoretical Computer Science 442, 13–21 (2012)
25. Plotkin, S.A., Shmoys, D.B., Tardos, É.: Fast approximation algorithms for fractional packing and covering problems. Mathematics of Operations Research 20(2), 257 (1995)
26. Spielman, D.A., Teng, S.-H.: Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. Journal of the ACM 51(3), 385–463 (2004)
27. Steele, J.M.: Subadditive Euclidean functionals and nonlinear growth in geometric probability. The Annals of Probability 9(3), 365–376 (1981)
28. Weber, M., Liebling, T.M.: Euclidean matching problems and the metropolis algorithm. Mathematical Methods of Operations Research 30(3), A85–A110 (1986)