1

Smoothed Analysis of Local Search

Bodo Manthey

Abstract

Local search is a powerful paradigm for finding good solutions to intractable combinatorial optimization problems. However, for many local search heuristics there exist worst-case instances on which they are extremely slow or provide solutions that are far from optimal.

Smoothed analysis is a semi-random input model that has been invented to bridge the gap between poor worst-case and good empirical performance of algorithms. In smoothed analysis, an adversary picks an arbitrary input, which is then slightly randomly perturbed. In particular, smoothed analysis has been applied successfully to local search algorithms in a variety of cases.

We use the 2-opt heuristic for the traveling salesman problem and the k-means method for clustering as examples to explain how local search heuristics can be analyzed in the framework of smoothed analysis. For both algorithm, as for many other local search algorithms, the worst-case running time is exponential in the input size, but polynomial in the framework of smoothed analysis.

1.1 Introduction

Large-scale optimization problems appear in many areas, ranging from engineering to the sciences. Unfortunately, many of these problems are computationally intractable. Thus, finding optimal solutions is very time-consuming. In practice, however, heuristics are often successful in finding close-to-optimal solutions surprisingly quickly. One particularly popular class of such heuristics are local search heuristics, which are often appealing because of their speed and because they are very easy to implement.

A *local search heuristic* for a combinatorial optimization problem is initialized with some solution of the given instance. Then it searches in the neighborhood of the current solution for a solution with better objective value. If successful, the local search heuristic replaces the current solution with this better solution. We call

this a *local improvement step* of the local search heuristic. Then the local search heuristic does the same search again. Here, the neighborhood of a solution are all solutions that can be obtained by slightly modifying this solution. What this means exactly depends on the problem and the local search heuristic.

A local search heuristic terminates if there is no better solution in the neighborhood of the current solution. We call such a solution a *local optimum*. Note that local optima are not necessary globally optimal solutions.

What is striking for many local search algorithms is the discrepancy between worst-case and observed performance: on the one hand, there often exist instances on which they take an exponential number of iterations before reaching a local optimum. It is also often quite easy to come up with instances on which they can converge to local optima that are much worse than global optima. From the complexity-theoretic perspective, finding a local optimum with respect to a given local search algorithm is PLS-complete for many such algorithms. (PLS stands for "polynomial local search" and captures the difficulty of finding local optima of optimization problems with respect to a "neighborhood". Although weaker than NP-completeness, PLS-completeness is widely considered to be a strong evidence of intractability (Schäffer and Yannakakis, 1991).)

On the other hand, this pessimistic view does not seem to reflect reality, where local search algorithms are popular because of their speed. The worst-case examples that show exponential running time are usually fragile constructions that hardly ever occur in practice. Sometimes, local search heuristics even achieve good empirical approximation performance. But even if not, their speed allows to rerun them a number of times with different initializations, which often results in much better performance.

This discrepancy makes local search heuristics a prime candidate for an "analysis beyond the worst case". In particular, smoothed analysis has been applied quite successfully to explain the empirical performance of local search algorithms.

Smoothed analysis is a semi-random input model that has been invented by Spielman and Teng (2004) in order to explain the empirical performance of the simplex method for linear programming. It is a hybrid of worst-case and average-case analysis and interpolates between these two: an adversary specifies an instance, and then this instance is slightly randomly perturbed. The smoothed performance of an algorithm is the maximum expected performance that the adversary can achieve, where the expectation is taken over the random perturbation.

If worst-case instances are isolated in the input space, then it is potentially very unlikely that we obtain such bad instances after perturbation. In principle, smoothed analysis can be applied to any measure of performance, but it has been most successful for the analysis of running times of algorithms that are super-polynomial in the worst-case but fast in practice, such as the two local search heuristics that we discuss in this chapter.

In the following, we explain smoothed analysis of local search algorithms mainly

by means of the 2-opt heuristic for the traveling salesman problem (TSP) and the k-means method for clustering. We will mostly focus on the running time of these algorithms and only briefly touch upon their approximation performance.

1.2 Smoothed analysis of the running time

The goal in this section is to show bounds for the *smoothed number of iterations* of local search algorithms, which is the maximum expected number of iterations, where the expectation is taken over the random perturbation.

We start this section with a simple analysis of the running time of the 2-opt heuristic for the TSP. After that, we sketch how to improve the bound obtained. Finally, we analyze the k-means method as an example of a local search algorithm where the analysis is much less straightforward than for 2-opt.

1.2.1 Main ideas

The main idea behind all smoothed analyses of running times of local search heuristics that have been conducted so far is the following "potential function" approach, where the objective function plays the role of the potential:

- We prove that the objective value of the initial solution is not too big.
- We prove that it is unlikely that iterations improve the objective value by only a small amount.

If the objective value is at most ν in the beginning and there is no iteration that decreases it by less than ε, then the number of iterations can be at most ν/ε.

Note that this approach is still quite pessimistic: first, it is unlikely that we always make the minimal possible improvement. It is more likely that some iterations cause a much larger improvement. Second, often there are several local improvement steps possible. In this case, the approach above assumes that we always make the worst possible choice.

The main advantage of this approach is that it decouples the iterations. If we would analyze iterations depending on earlier iterations, then we would face dependencies that are very hard to deal with.

1.2.2 A simple bound for 2-opt

To illustrate a smoothed analysis of the running time of a local search heuristic, we take the 2-opt heuristic for the TSP as an example. More specifically, we consider the TSP in the Euclidean plane, where the distance between two points $a, b \in \mathbb{R}^2$ is given by $\|a-b\|^2$, i.e., the squared Euclidean distance between the two points. This

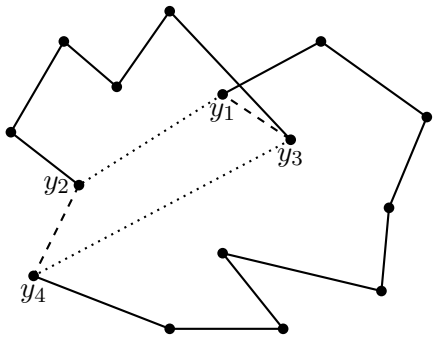

Figure 1.1 An example of a 2-opt step, where the edges $\{y_1, y_2\}$ and $\{y_3, y_4\}$ are replaced by $\{y_1, y_3\}$ and $\{y_2, y_4\}$.

means that – given a set $Y \subseteq \mathbb{R}^2$ of n points – the goal is to compute a Hamiltonian cycle (also called a tour) H through Y that minimizes

$$\sum_{\{a,b\} \in H} \|a - b\|^2.$$

In other words, we want to find a cyclic ordering of the points that minimizes the sum of squared distances of consecutive points.

We chose this problem for two reasons: First, for points in the plane, the concept of "small perturbations" is quite natural. Second, the choice of squared Euclidean distances (compared to the – more natural – Euclidean distances) is because this makes the smoothed analysis of the running time relatively compact, in contrast to many other cases of smoothed analysis, which are quite involved technically.

TSP and the 2-opt heuristic

The 2-opt heuristic for the TSP performs so-called *2-opt steps* to improve a given initial tour as long as possible. A *2-opt step* is the following operation: let H be an arbitrary Hamiltonian tour through the point set Y. Assume that H contains edges $\{y_1, y_2\}$ and $\{y_3, y_4\}$, where the four distinct points y_1, y_2, y_3, and y_4 appear in this order in H. Assume further that $\|y_1 - y_2\|^2 + \|y_3 - y_4\|^2 > \|y_1 - y_3\|^2 + \|y_2 - y_4\|^2$. Then we replace $\{y_1, y_2\}$ and $\{y_3, y_4\}$ by $\{y_1, y_3\}$ and $\{y_2, y_4\}$ to obtain a shorter Hamiltonian tour. See Figure 1.1 for an example.

Initialized with an arbitrary Hamiltonian tour H through the point set Y, the 2-opt heuristic performs 2-opt steps until a local minimum is reached.

Model and approach

We use the following probabilistic input model for the analysis of 2-opt: an adversary specifies a set $X = \{x_1, \ldots, x_n\} \subseteq [0, 1]^2$ consisting of n points from the unit square. Then we obtain the actual input Y by perturbing each point x_i by a random

variable g_i:

$$Y = \{y_i = x_i + g_i \mid i \in \{1, \dots, n\}\}.$$

We assume that $g_1, \dots, g_n$ are independent and follow a 2-dimensional Gaussian distribution with standard deviation σ and mean 0. We call the instance Y a *σ-perturbed point set*.

We mainly exploit two properties of Gaussian distributions in our smoothed analysis: First, their maximum density is bounded. Second, a 2-dimensional Gaussian distribution can be viewed as superposition of two 1-dimensional Gaussian distributions in any two orthonormal directions.

Our approach is as described above: first, we show that the initial tour has a length of $O(n)$ with high probability. Second, we show that the probability that there exists any 2-opt step that decreases the objective function by less than ε is bounded from above by ε times a polynomial in n and $1/\sigma$. Finally, we combine these two ingredients together with the worst-case upper bound of $n!$ for the number of iterations to obtain a smoothed polynomial bound for the number of iterations.

Technical preliminaries and assumptions

In the following, we assume that $\sigma \leq \frac{1}{2\sqrt{n \ln n}}$. This is without loss of generality by Exercise 1.2.

The following lemma is a standard tail bound for Gaussian random variables. A proof can be found in many textbooks on probability theory.

Lemma 1.1 (tail bound for Gaussians) *Let X be a random variable with Gaussian distribution with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$. Then*

$$\mathbb{P}(X \geq \mu + \sigma t) = \mathbb{P}(X \leq \mu - \sigma t) \leq \frac{1}{t\sqrt{2\pi}} \cdot \exp\left(-\frac{t^2}{2}\right).$$

Lemma 1.2 (interval lemma for Gaussians) *Let X be distributed according to a Gaussian distribution with arbitrary mean and standard deviation $\sigma > 0$. Let $t \in \mathbb{R}$, and let $\varepsilon > 0$. Then*

$$\mathbb{P}\big(X \in (t, t + \varepsilon]\big) \leq \frac{\varepsilon}{2\sigma}.$$

Proof This follows from the fact that the density of a Gaussian random variable with standard deviation σ is bounded from above by $\frac{1}{2\sigma}$. □

Upper bound for the initial tour

The following lemma gives a quite trivial upper bound for the length of an initial tour.

Lemma 1.3 *We have $L_{\text{init}} \leq 18n$ with a probability of at least $1 - \frac{1}{n!}$.*

Proof If $Y \subseteq [-1, 2]^d$, then the longest distance between any two points in Y

(measured in squared Euclidean distance) is at most 18. Thus, any tour has a length of at most $18n$ in this case.

If $Y \not\subseteq [-1,2]^2$, then there exists an i such that $\|g_i\|_\infty \geq 1$. Thus, there must exist an $i \in \{1,\dots,n\}$ and a coordinate $j \in \{1,2\}$ such that the absolute value of the j-th entry of g_i is at least 1. We use Lemma 1.1 with $\sigma \leq \frac{1}{2\sqrt{n \ln n}}$ and $t = 1/\sigma$. This yields that the probability for a single entry to be of absolute value at least 1 is bounded from above by

$$\frac{1}{2\sqrt{2\pi n \ln n}} \cdot \exp\bigl(-2n \ln n\bigr) \leq n^{-2n} \leq (n!)^{-2}.$$

A union bound over the n choices of i and the two choices of j and the (very loose) bound $2n \leq n!$ yields the lemma. □

For the remainder of this section, let $\Delta_{a,b}(c) = \|c-a\|^2 - \|c-b\|^2$. The improvement of a 2-opt step, where $\{y_1, y_2\}$ and $\{y_3, y_4\}$ are replaced by $\{y_1, y_3\}$ and $\{y_2, y_4\}$, can thus be written as $\Delta_{y_2,y_3}(y_1) - \Delta_{y_2,y_3}(y_4)$.

Let $\Delta_{\min}$ be the smallest positive improvement by any possible 2-opt step. For the analysis of $\Delta_{\min}$, the following lemma is useful.

Lemma 1.4 *Let $a, b \in \mathbb{R}^2$ with $a \neq b$, and let $c \in \mathbb{R}^2$ be drawn according to a Gaussian distribution with standard deviation σ. Let $I \subseteq \mathbb{R}$ be an interval of length ε. Then*

$$\mathbb{P}\bigl(\Delta_{a,b}(c) \in I\bigr) \leq \frac{\varepsilon}{4\sigma \cdot \|a-b\|}.$$

Proof Since Gaussian distributions are rotationally symmetric and translation invariant, we can assume without loss of generality that $a = (0,0)$ and $b = (\delta, 0)$ with $\delta = \|a-b\|$. Let $c = (c_1, c_2)^{\mathrm{T}}$. Then $\Delta_{a,b}(c) = (c_1^2 + c_2^2) - ((c_1-\delta)^2 + c_2^2) = 2c_1\delta - \delta^2$. Since δ^2 is a constant (independent of a, b, and c), we have $\Delta_{a,b}(c) \in I$ if and only if $2c_1\delta$ falls into an interval of length ε. This is equivalent to c_1 falling into an interval of length $\frac{\varepsilon}{2\delta}$.

Since c_1 is a 1-dimensional Gaussian random variable with a standard deviation of σ, the lemma follows from Lemma 1.2. □

With this lemma, we can bound the probability that any improving 2-opt step yields only a small improvement.

Lemma 1.5 $\mathbb{P}(\Delta_{\min} \leq \varepsilon) = O\bigl(\frac{n^4\varepsilon}{\sigma^2}\bigr)$.

Proof Consider any four distinct points $y_1, y_2, y_3, y_4 \in Y$ and the 2-opt step, where the two edges $\{y_1, y_2\}$ and $\{y_3, y_4\}$ are replaced by $\{y_1, y_3\}$ and $\{y_2, y_4\}$. We prove that the probability that this 2-opt step yields a positive improvement of at most ε is bounded by $O(\varepsilon/\sigma^2)$. Then the lemma follows by a union bound over the choices of the four points $y_1, y_2, y_3, y_4 \in Y$.

The improvement caused by this 2-opt step equals $\Delta_{y_2,y_3}(y_1) - \Delta_{y_2,y_3}(y_4)$. We

use the principle of deferred decision: first, let an adversary fix the position of y_2, y_3, and y_4 arbitrarily. This fixes $\alpha = \Delta_{y_2,y_3}(y_4)$ as well as the distance $\delta = \|y_2 - y_3\|$ between y_2 and y_3. Thus, the improvement caused by this 2-opt step is only in the interval $(0, \varepsilon]$ if $\Delta_{y_2,y_3}(y_1) \in (\alpha, \alpha+\varepsilon]$, which is an interval of size ε. The probability that this happens is bounded from above by $\frac{\varepsilon}{4\sigma\delta}$ according to Lemma 1.4.

Let f be the probability density function of $\delta = \|y_2 - y_3\|$. Then the probability that the 2-opt step considered yields an improvement of at most ε is bounded from above by

$$\int_{\delta=0}^{\infty} \frac{\varepsilon}{4\sigma\delta} \cdot f(\delta)\,\mathrm{d}\delta.$$

Now we observe that the distribution of $1/\delta$ is stochastically dominated by $1/X$, where X is chi-distributed. This is because $\frac{\varepsilon}{4\sigma\delta}$ is decreasing in δ. Thus, the "worst-case" is that x_3 – the unperturbed version of y_3 – is located exactly at y_2. The chi-distribution describes the length of a vector that is distributed according to a Gaussian distribution with mean 0. In the 2-dimensional case, the density of the chi-distribution is given by $\frac{x}{\sigma^2} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right)$.

From this observation, we obtain that we can replace f by the density function of the chi-distribution to get an upper bound for the probability that we have an improvement of at most ε:

$$\int_{\delta=0}^{\infty} \frac{\varepsilon}{4\sigma\delta} \cdot \frac{\delta}{\sigma^2} \cdot \exp\left(-\frac{\delta^2}{2\sigma^2}\right) \mathrm{d}\delta = \int_{\delta=0}^{\infty} \frac{\varepsilon}{4\sigma^3} \cdot \exp\left(-\frac{\delta^2}{2\sigma^2}\right) \mathrm{d}\delta = O\left(\frac{\varepsilon}{\sigma^2}\right).$$

To finish the proof, we take a union bound over the $O(n^4)$ choices for the points y_1, y_2, y_3, and y_4. □

The previous lemma can be turned into a tail bound for the number of iterations that 2-opt needs to converge to a local optimum, which yields our first theorem.

Theorem 1.6 *Let $Y \subseteq \mathbb{R}^2$ be a σ-perturbed point set, and let $\sigma \leq \frac{1}{2\sqrt{n \ln n}}$. Then the expected maximum number of iterations that the 2-opt heuristic needs to compute a locally optimal TSP tour with respect to squared Euclidean distances is bounded from above by $O(n^6 \log n/\sigma^2)$.*

Proof If 2-opt runs for at least t steps, then we must have $L_{\text{init}} \geq 18n$ or $\Delta_{\min} \leq 18n/t$. The probability that any of these events happens is at most $\frac{1}{n!} + O\left(\frac{n^5}{\sigma^2 t}\right)$ by Lemmas 1.3 and 1.5, where the probability is taken over the random perturbation.

Since no TSP tour shows up twice in any run of 2-opt, we know that the number of iterations is upper-bounded by $n!$. Let T be the random variable that is the maximum possible number of iterations that 2-opt can need on the (random) point set Y. Then

$$\mathbb{E}(T) = \sum_{t=1}^{n!} \mathbb{P}(T \geq t) \leq \sum_{t=1}^{n!} \left(\frac{1}{n!} + O\left(\frac{n^5}{\sigma^2 t}\right)\right) = O\left(\frac{n^6 \log n}{\sigma^2}\right).$$

□

1.2.3 Possibilities for improving the bound

Very roughly, the analysis of the running time of 2-opt in the previous section worked as follows:

- We have used the objective function as a potential function, and we have proved a (very simple) upper bound for the length of any initial tour.
- We have divided the possible steps that the algorithm can take into classes. In our case, every 2-opt step is exactly described by four points, namely the four endpoints of the edges involved. The rest of the tour does not play a role.
- For each such class, we have proved that it is unlikely that any iteration of this class yields only a small improvement.
- By taking a union bound over all classes, we turned this into a tail bound for the number of iterations and used this to obtain a bound on the expected number of iterations.

This immediately gives the following options for improving the bound:

- We could show a smaller bound for the length of the initial tour. This would immediately improve the bound.
- We could try to divide the possible iterations of the 2-opt heuristic into fewer classes. Then it would suffice to take the union bound over fewer classes.
- We could try to prove a stronger upper bound for the probability that any iteration of a class yields a small improvement. This would yield a stronger tail bound for the number of iterations and therefore improve the final bound.

In fact, avoiding the naive application of the union bound and instead cleverly partitioning the iterations into groups that can be analyzed simultaneously is usually the key ingredient of a smoothed analysis. In the remainder of this subsection, we sketch ideas how to improve the bound for the smoothed number of iterations of 2-opt.

Improving the initial tour length

So far, we did not make any assumptions on how the initial tour was constructed. In practice, however, one would probably start with some reasonable approximation instead of an arbitrary (bad) tour. For instance, one can find an initial tour of length $O(1)$ (Yukich, 1998), which immediately decreases our bound by a linear factor. (This holds only for squared Euclidean distances in 2-dimensional space. For standard Euclidean distances, one can only guarantee bounds of length $O(\sqrt{n})$.)

Linked pairs of 2-opt steps

The idea behind analyzing so-called "linked pairs of 2-opt steps" is the observation that only taking into account the smallest possible improvement is quite pessimistic. In order to improve the bound, we consider pairs of 2-opt steps that share some vertices. The two 2-opt steps of such a pair do not have to be executed next to each other in a run of 2-opt. This improvement does not fall directly into one of the three possibilities of improvement mentioned above. Indeed, we prove a stronger upper bound for the probability that a class yields a small improvement, but not for single iterations. Instead, we consider two iterations, which increases the number of classes. It turns out that the stronger upper bound more than compensates the increase of the number of different classes.

Sharing only a single vertex leaves the edges of the two 2-opt steps disjoint and, thus, does not help to improve the bound. It turns out that the case in which all four vertices are identical for both 2-opt steps is quite difficult to analyze because of dependencies. Hence, we restrict ourselves to pairs of 2-opt steps that overlap in two or three vertices. As at most six vertices are involved in such pairs, the number of such pairs is at most $O(n^6)$. While this is worse than the number of simple 2-opt steps, which is $O(n^4)$, it is compensated by the fact that the probability that both 2-opt steps yield only a small improvement is much smaller than in the case of a single 2-opt step. Basically, although the improvements obtained from the two 2-opt steps from such a pair are not independent, we can analyze them as if they were. The following lemma, whose formal proof we omit, summarizes this.

Lemma 1.7 *The probability that there exists a linked pair of 2-opt steps that have two or three vertices in common and such that both 2-opt steps improve the tour, but only by at most $\varepsilon > 0$, is at most $O\left(\frac{n^6\varepsilon^2}{\sigma^4}\right)$.*

Crucial for this approach to work is that we encounter sufficiently many linked pairs of 2-opt steps in any run of 2-opt. The following lemma basically states that every sufficiently long sequence of 2-opt steps must contain a constant fraction of 2-opt steps that form disjoint linked pairs. We omit its proof, which is not difficult, but a bit technical.

Lemma 1.8 (Röglin and Schmidt (2018)) *Every sequence of t consecutive 2-opt steps contains at least $(2t-n)/7$ disjoint linked pairs of 2-opt steps that share either two or three vertices.*

Theorem 1.9 *Let $Y \subseteq \mathbb{R}^2$ be a σ-perturbed point set, and let $\sigma \leq \frac{1}{2\sqrt{n \ln n}}$. Then the expected maximum number of iterations that the 2-opt heuristic needs to compute a locally optimal TSP tour with respect to squared Euclidean distances is $O(n^4/\sigma^2)$.*

Proof Let T be the random variable that is the maximum possible number of iterations that 2-opt can need on the (random) point set Y. By Lemma 1.8, there

exist constants $c_1, c_2 > 0$ such that every sequence of at least $t \geq c_1 n^2$ iterations contains at least $c_2 t$ disjoint pairs of linked 2-opt steps sharing two or three vertices.

Then $T \geq t$ only if $t \leq c_1 n^2$ or if $L_{\text{init}} \geq 18n$ or if there is a pair of linked 2-opt steps that yields an improvement of at most $\frac{18n}{c_2 t}$. Thus, there exist constants $c_3, c_4 > 0$ such that, by Lemma 1.7, we have

$$\begin{aligned}\mathbb{E}(T) &\leq c_1 n^2 + \sum_{t \geq c_1 n^2} \mathbb{P}(T \geq t) \leq c_1 n^2 + \sum_{t \geq c_1 n^2} \min\left\{1, c_3 \cdot \frac{n^8}{t^2 \sigma^4}\right\} \\ &\leq c_4 \cdot \frac{n^4}{\sigma^2} + \sum_{t \geq c_4 n^4/\sigma^2} c_3 \cdot \frac{n^8}{t^2 \sigma^4} = O\left(\frac{n^4}{\sigma^2}\right).\end{aligned}$$

□

With the preceding discussion about the initial tour length, we can even improve the bound of Theorem 1.9 to $O(n^3/\sigma^2)$.

1.2.4 A bound for the k-means method

The second example of a local search heuristic whose running time we want to analyze in the framework of smoothed analysis is the k-means method for clustering.

Description of the k-means method

Before doing the smoothed analysis, let us first describe k-means clustering and the k-means method.

We are given a finite set $X \subseteq \mathbb{R}^d$ of n data points and the number $k \in \mathbb{N}$ of clusters. The goal of k-means clustering is to partition these points into k clusters $C_1, \ldots, C_k$. In addition to the clusters, we want to compute cluster centers $c_1, \ldots, c_k \in \mathbb{R}^d$ that are the representatives of their clusters. The centers do not have to be data points. The goal of k-means clustering is to find clusters and centers that minimize the sum of squared distances of data points to cluster centers:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2.$$

If we already know the cluster centers, then this specifies (up to tie-breaking) a clustering: every point is assigned to a cluster whose center is closest to it. The other way round, if we have clusters $C_1, \ldots, C_k$, then each cluster center c_i should be chosen as the center of mass

$$\operatorname{cm}(C_i) = \frac{1}{|C_i|} \cdot \sum_{x \in C_i} x$$

of C_i. This is a direct consequence of the following lemma, the proof of which we leave as Exercise 1.4.

Lemma 1.10 *Let $C \subseteq \mathbb{R}^d$ be a finite set of points, let $c = \mathrm{cm}(C)$ be the center of mass of C, and let $z \in \mathbb{R}^d$ be arbitrary. Then*

$$\sum_{x \in C} \|x - z\|^2 = \sum_{x \in C} \|x - c\|^2 + |C| \cdot \|c - z\|^2.$$

The key idea of the k-means method is to exploit that clustering and centers mutually imply each other: the k-means method alternates between optimizing the clustering based on the given centers and optimizing the centers based on the given clustering. More formally, it works as follows:

(1) Choose initial cluster centers $c_1, \ldots, c_k$.
(2) Construct a clustering $C_1, \ldots, C_k$ from the current cluster centers.
(3) Set $c_i = \mathrm{cm}(C_i)$ for all $i \in \{1, \ldots, k\}$.
(4) If anything changed in Steps 2 or 3, return to Step (2).

The k-means method is one of the most popular clustering algorithms. Its popularity stems from two facts: first, it is very simple. Second, it is very fast on practical data sets. This second fact allows one to rerun it several times with different initializations in order to obtain a good clustering.

However, in contrast to practical performance, the worst-case running time of the k-means method is exponential in the number k of clusters. We can choose $k = \Theta(n)$, which shows that the worst-case number of iterations can be exponential. This lower bound construction works already in the Euclidean plane, i.e., if $d = 2$ is fixed.

The only known worst-case upper bound for the number of iterations is based on counting the number of different clusterings and the trivial fact that no clustering occurs twice in a run of the k-means method. The number of different clusterings of n points in d-dimensional space into k clusters, where clusters have to be separated by hyperplanes, is upper-bounded by n^{3kd}.

Model and approach

In the following, we apply smoothed analysis to the running time of the k-means method. More specifically, our goal is to prove an upper bound that is polynomial in n^k and $1/\sigma$, which removes a factor of d from the exponent compared to the worst-case running time. While such a bound surely does not explain the observed performance of the algorithm, it conveys the basic ideas of the analysis. To keep the analysis relatively simple, we combine the first $\mathrm{poly}(n^k, 1/\sigma)$ bound with techniques that were used later in the proof of a truly polynomial bound.

In comparison to the 2-opt heuristic, we have to address two technical challenges in the smoothed analysis of the k-means method:

- Iterations of the 2-opt heuristic can be compactly represented by the four vertices involved. For the k-means method, such a compact representation of iterations is much less obvious.

- In order to obtain a polynomial bound for the smoothed running time of the 2-opt heuristic, it was sufficient to consider the improvement caused by a single iteration. This does not seem to be the case for the k-means method.

The model that we use for the smoothed analysis is the same as for the 2-opt heuristic: an adversary specifies a set $X \subseteq [0,1]^d$ of n points. Then these points are perturbed by independent Gaussian random variables of standard deviation σ. We call the resulting point set $Y \subseteq \mathbb{R}^d$ again a σ-perturbed point set, and we run the k-means method on this point set Y.

Again, restricting X to the unit hypercube is just a matter of scaling and does not restrict generality. And again, we restrict our analysis to the case $\sigma \leq 1$ because of Exercise 1.2.

In the following, we also make the (natural) assumption that $k, d \leq n$. In many applications, k and d are even considered to be constant. Using the upper bound of n for k and d sometimes simplifies calculations.

The main idea is similar to the 2-opt heuristic: we use the objective function as a potential function and show that it has to decrease sufficiently quickly. However, as noted already at the beginning of this section, there are two issues that make this more difficult than for the 2-opt heuristic: first, a compact description of iterations does not seem to exist for the k-means method. Second, we cannot rule out that there are iterations in which the objective function decreases only by a negligible amount. This makes it necessary to consider longer sequences of iterations, similar to the analysis of linked pairs of 2-opt steps. But while the analysis of linked pairs of 2-opt steps was only necessary to improve the bound, here this seems unavoidable.

For the first issue, it turns out that we can describe iterations by $O(kd)$ points sufficiently precisely. For the second issue, considering sequences of 2^k iterations suffices in order to make it unlikely that all of them yield only a small improvement.

Decrease of the objective function

In order to analyze the decrease of the objective function, we first have to understand what causes it to decrease. The objective function gets smaller (1) by moving cluster centers and (2) by reassigning data points.

Lemma 1.10 implies that moving a cluster center c_i by a distance of ε to the center of mass of its point set C_i decreases the objective value by $\varepsilon^2 \cdot |C_i| \geq \varepsilon^2$.

For a hyperplane H and a point z, we denote by $\operatorname{dist}(z, H)$ the Euclidean distance of z to H. For analyzing the decrease of the objective value caused by reassigning a point, we need the notion of a bisecting hyperplane: for two points $x, y \in \mathbb{R}^d$ with $x \neq y$, we call a hyperplane H the bisector of x and y if H is orthogonal to $x - y$ and $\operatorname{dist}(x, H) = \operatorname{dist}(y, H)$. This means that

$$H = \left\{ z \in \mathbb{R}^d \mid 2z^{\mathrm{T}}(x - y) = (x + y)^{\mathrm{T}}(x - y) \right\}.$$

The decrease of the objective function caused by reassigning a data point to a

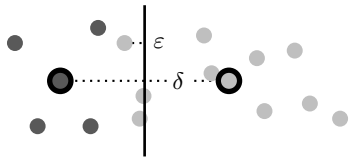

Figure 1.2 If we reassign the leftmost light point – which is at a distance of ε from the bisecting hyperplane – to the dark cluster, then this decreases the objective function by $2\varepsilon\delta$, where δ is the distance between the two centers.

different cluster depends on the distance of this point from the corresponding bisecting hyperplane and the distance between the two cluster centers involved. The following lemma makes this more precise – see also Figure 1.2. We leave its proof as Exercise 1.5.

Lemma 1.11 *Let c_i and c_j be two cluster centers with bisector H, and let $y \in C_i$. If $\|y - c_j\| < \|y - c_i\|$, then reassigning y to C_j decreases the objective value by*

$$2 \cdot \operatorname{dist}(y, H) \cdot \|c_i - c_j\|.$$

The rough idea for the smoothed analysis is as follows: if many points are reassigned to a new cluster, then it is unlikely that all of them are close to their corresponding bisecting hyperplane. If only a few points are reassigned, then at least one cluster center must move significantly. This hope turns out to be wrong for a single iteration, so we have to consider longer sequences of iterations.

Dense iterations

We call an iteration of the k-means method *dense* if there is at least one cluster that gains or loses in total at least $2kd$ points. According to Lemma 1.11, we have to show that in a dense iteration, it is unlikely that all these points are close to their corresponding bisecting hyperplanes or that centers are too close to each other.

We call the point set Y *ε-separated* if, for all hyperplanes $H \subseteq \mathbb{R}^d$, there are less than $2d$ points $y \in Y$ with $\operatorname{dist}(y, H) \leq \varepsilon$. The following lemma quantifies the minimal improvement caused by any dense iteration, provided that Y is ε-separated.

Lemma 1.12 *If Y is ε-separated, then the potential decreases by at least $2\varepsilon^2/n$ in every dense iteration.*

Proof Since the iteration is dense, there must be a cluster C_i that exchanges at least $2kd$ points with other clusters in this iteration. Hence, there must be another cluster C_j with which C_i exchanges at least $2d + 1$ points. Since Y is ε-separated, at least one point $y \in Y$ that switches between C_i and C_j is at a distance of at least ε from the hyperplane bisecting $c_i = \operatorname{cm}(C_i)$ and $c_j = \operatorname{cm}(C_j)$, where C_i and C_j are the clusters before the switch.

In order to bound the decrease of the objective value from below by $2\varepsilon^2/n$,

we need a lower bound of ε/n for $\|c_i - c_j\|$. There exists a hyperplane H' (the bisector from the previous iteration) that separates C_i from C_j. Among all at least $2d+1$ points that want to switch in the current iteration, at least one point y must be at a distance of at least ε from H' since Y is ε-separated. Assume without loss of generality that $y \in C_i$. Then, since (i) $|C_i| \leq n$, (ii) y is at least ε away from H' since Y is ε-separated, and (iii) all points of C_i are on the same side of H', the center of mass $c_i = \mathrm{cm}(C_i)$ must be at least ε/n away from H'. Hence, $\frac{\varepsilon}{n} \leq \mathrm{dist}(c_i, H') \leq \|c_i - c_j\|$. (Note that this argument does not work with the bisector of c_i and c_j instead of H', as some points are on the "wrong side" of this bisector.) □

A simple union bound together the following lemma yields an upper bound for the probability that Y is not ε-separated (Lemma 1.14).

Lemma 1.13 *Let $P \subseteq \mathbb{R}^d$ be any finite set of at least d points, and let $H \subseteq \mathbb{R}^d$ be an arbitrary hyperplane. Then there exists a hyperplane $H' \subseteq \mathbb{R}^d$ that contains at least d points of P such that*

$$\max_{p \in P}\big(\mathrm{dist}(p, H')\big) \leq 2d \cdot \max_{p \in P}\big(\mathrm{dist}(p, H)\big).$$

We skip the proof of Lemma 1.13 and refer to Arthur and Vassilvitskii's paper (Arthur and Vassilvitskii, 2009, Lemma 5.8). The intuition of this lemma is as follows: if there is any hyperplane H such that all points of some set $P \subseteq \mathbb{R}^d$ are close to H, then there is also a hyperplane H' that contains d points of P and all other points in P are close to H'. Lemma 1.13 is useful because of the dependency between the location of the bisecting hyperplanes and the data points. Using it, we can use d points to fix some hyperplane and then use the independent randomness of another d points to show that they are not close to this hyperplane.

Lemma 1.14 *The probability that Y is not ε-separated is at most $n^{2d} \cdot \left(\frac{2d\varepsilon}{\sigma}\right)^d$.*

Proof According to Lemma 1.13 below, it suffices to show that the probability that there are two disjoint sets P and P' consisting of d points of Y each such that all points of P' are $(2d\varepsilon)$-close to the hyperplane through P is bounded by $n^{2d} \cdot \left(\frac{2d\varepsilon}{\sigma}\right)^d$.

Fix any sets P and P' of d points. Using the principle of deferred decisions, we fix the position of the points in P arbitrarily. Then the probability that all points of P' are within distance $2d\varepsilon$ of the hyperplane through P is at most $(2d\varepsilon/\sigma)^d$, because perturbations of the points in P' are independent and the probability that a point is within a distance of $2d\varepsilon$ of a fixed hyperplane is bounded from above by $2d\varepsilon/\sigma$ by Lemma 1.2. The lemma follows by a union bound over the at most n^{2d} choices for P and P'. □

By combining Lemmas 1.12 and 1.14, we obtain the following result about dense iterations.

Lemma 1.15 *For $d \geq 2$ and $\sigma \leq 1$, the probability that there exists a dense iteration in which the potential decreases by less than ε is bounded from above by*

$$\left(\frac{2n^{3.5}\sqrt{\varepsilon}}{\sigma}\right)^d.$$

Proof According to Lemma 1.12, if there is a dense iteration in which the potential decreases by less than ε, then Y is not $\sqrt{n\varepsilon/2}$-separated. By Lemma 1.14 and $d \leq n$, this happens with a probability of at most

$$n^{2d} \cdot \left(\frac{2d\sqrt{n\varepsilon/2}}{\sigma}\right)^d \leq \left(\frac{2dn^{2.5}\sqrt{\varepsilon}}{\sigma}\right)^d \leq \left(\frac{2n^{3.5}\sqrt{\varepsilon}}{\sigma}\right)^d.$$

□

Sparse iterations

We call an iteration *sparse* if every cluster gains or loses in total at most $2kd$ points.

Let C_i^t be the set of points in the i-th cluster in iteration t of the k-means method. We define an *epoch* to be a sequence of consecutive iterations $t, t+1, \ldots, t+\ell$ in which no cluster assumes more than two different point sets. This means that $|\{C_i^a \mid t \leq a \leq t+\ell\}| \leq 2$ for all $i \in \{1, 2, \ldots, k\}$. A trivial upper bound for the length of every epoch is given in the following lemma. In fact, it is possible to show that the length of every epoch is at most 3 (see Exercise 1.6), but this is not needed for the bound that we aim for.

Lemma 1.16 *The length of every epoch is bounded by 2^k.*

Proof After 2^k iterations, at least one cluster must have assumed a third set of points. Otherwise, a clustering would show up a second time. This is impossible as the objective value strictly decreases in every iteration. □

We call a set $Y \subseteq \mathbb{R}^d$ of data points *η-coarse* for some number $\eta > 0$ if, for all triples $P_1, P_2, P_3 \subseteq Y$ of different subsets with $|P_1 \triangle P_2| \leq 2kd$ and $|P_2 \triangle P_3| \leq 2kd$, we have $\|\operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1})\| > \eta$ for at least one $i \in \{1, 2\}$. Here, "$\triangle$" denotes the symmetric difference of two sets.

Lemma 1.17 *Assume that Y is η-coarse, and consider a sequence of 2^k consecutive iterations of the k-means method. If each of these iterations is sparse, then the potential decreases by at least η^2.*

Proof After 2^k iterations, at least one cluster has assumed a third configuration (Lemma 1.16). Since the iterations are sparse, there are sets P_1, P_2, and P_3 such that $|P_1 \triangle P_2|, |P_2 \triangle P_3| \leq 2kd$ such that this cluster switches from point set P_1 to P_2 and later to P_3 (directly from P_2 or after switching back to P_1 – not necessarily in consecutive iterations). Since the instance is η-coarse, we have $\|\operatorname{cm}(P_1) - \operatorname{cm}(P_2)\| > \eta$ or $\|\operatorname{cm}(P_2) - \operatorname{cm}(P_3)\| > \eta$. Thus, the corresponding cluster center must have

moved by at least η in one iteration, which decreases the potential by at least η^2 according to Lemma 1.10. □

Lemma 1.18 *The probability that Y is not η-coarse is bounded from above by $(7n)^{4kd} \cdot (4nkd\eta/\sigma)^d$.*

Proof Let $P_1, P_2, P_3 \subseteq Y$ be three sets with $|P_1 \triangle P_2| \leq \ell$ and $|P_2 \triangle P_3| \leq \ell$. Let $A = P_1 \cap P_2 \cap P_3$, and let B_1, B_2, and B_3 be sets such that $P_i = A \cup B_i$ for $i \in \{1, 2, 3\}$ and B_1, B_2, and B_3 are disjoint to A. We have $|B_1 \cup B_2 \cup B_3| \leq 2\ell$ and $B_1 \cap B_2 \cap B_3 = \emptyset$.

We perform a union bound over the choices of for the sets B_1, B_2, and B_3. The number of possible choice for these sets is upper-bounded by $7^{2\ell} \cdot \binom{n}{2\ell} \leq (7n)^{2\ell}$: We select 2ℓ elements of Y. Then we choose for each element in which of the three sets it should belong. None of these elements belongs to all sets, but there can be elements that belong to no set. We need this possibility since we can have $|B_1 \cup B_2 \cup B_3| < 2\ell$.

For $i \in \{1, 2, 3\}$, we have

$$\operatorname{cm}(P_i) = \frac{|A|}{|A| + |B_i|} \cdot \operatorname{cm}(A) + \frac{|B_i|}{|A| + |B_i|} \cdot \operatorname{cm}(B_i).$$

Hence, for $i \in \{1, 2\}$, we can write $\operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1})$ as

$$\begin{aligned}\operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1}) = &\left(\frac{|A|}{|A| + |B_i|} - \frac{|A|}{|A| + |B_{i+1}|}\right) \cdot \operatorname{cm}(A) \\ &+ \frac{|B_i|}{|A| + |B_i|} \cdot \operatorname{cm}(B_i) - \frac{|B_{i+1}|}{|A| + |B_{i+1}|} \cdot \operatorname{cm}(B_{i+1}).\end{aligned}$$

We distinguish three cases. The first case is that $|B_i| = |B_{i+1}|$ for some $i \in \{1, 2\}$. Then the equation above simplifies to

$$\begin{aligned}\operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1}) &= \frac{|B_i|}{|A| + |B_i|} \cdot \operatorname{cm}(B_i) - \frac{|B_i|}{|A| + |B_i|} \cdot \operatorname{cm}(B_{i+1}) \\ &= \frac{1}{|A| + |B_i|} \cdot \left(\sum_{y \in B_i \setminus B_{i+1}} y - \sum_{y \in B_{i+1} \setminus B_i} y\right).\end{aligned}$$

Since $B_i \neq B_{i+1}$ and $|B_i| = |B_{i+1}|$, there exists a point $y \in B_i \setminus B_{i+1}$.

We use the principle of deferred decisions. We first fix all points in $(B_i \cup B_{i+1}) \setminus \{y\}$ arbitrarily. Then $\| \operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1}) \| \leq \eta$ is equivalent to the event that y assumes a position in a hyperball of radius $(|A| + |B_i|) \cdot \eta \leq n\eta$. The probability that this happens is bounded from above by the maximum density of a Gaussian distribution times the volume of the hyperball, which is at most $(n\eta/\sigma)^d \leq (2n\eta\ell/\sigma)^d$.

The second case is that $A = \emptyset$. This case is in fact identical to the first case.

The third case is that $|B_1| \neq |B_2| \neq |B_3|$. We denote by $\mathcal{B}(c, r) = \{x \in \mathbb{R}^d \mid$

$\|x - c\| \leq r\}$ the hyperball of radius r around c. For $i \in \{1, 2\}$, let

$$r_i = \left(\frac{|A|}{|A| + |B_i|} - \frac{|A|}{|A| + |B_{i+1}|}\right)^{-1} = \frac{(|A| + |B_i|) \cdot (|A| + |B_{i+1}|)}{|A| \cdot (|B_{i+1}| - |B_i|)}$$

and

$$Z_i = \frac{|B_{i+1}|}{|A| + |B_{i+1}|} \operatorname{cm}(B_{i+1}) - \frac{|B_i|}{|A| + |B_i|} \operatorname{cm}(B_i).$$

We observe that the event $\|\operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1})\| < \eta$ is equivalent to the event that $\operatorname{cm}(A) \in \mathcal{B}_i = \mathcal{B}(r_i Z_i, |r_i|\eta)$. Consequently, a necessary condition that the event $\|\operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1})\| < \eta$ occurs for both $i \in \{1, 2\}$ is that the hyperballs $\mathcal{B}_1$ and $\mathcal{B}_2$ intersect.

The two hyperballs intersect if and only if their centers are at a distance of at most $(|r_1| + |r_2|) \cdot \eta$ of each other. Hence,

$$\begin{aligned} & \mathbb{P}\big(\|\operatorname{cm}(P_1) - \operatorname{cm}(P_2)\| \leq \eta \text{ and } \|\operatorname{cm}(P_2) - \operatorname{cm}(P_3)\| \leq \eta\big) \\ \leq\ & \mathbb{P}\left(\|r_1 Z_1 - r_2 Z_2\| \leq \big(|r_1| + |r_2|\big)\eta\right). \end{aligned}$$

With some tedious but not too insightful calculations, we can see that the probability of this event is bounded as desired. □

The main technical problem in the proof of Lemma 1.18 is that we cannot control the position of $\operatorname{cm}(A)$. The reason is that there are too many possible choices for points in A. Because of this, we cannot simply apply a union bound over all possibilities for A.

The first case in the proof of Lemma 1.18 shows that for the case that the same number of points leaves and enters a cluster, it is already quite likely that the potential decreases significantly. In this case, no epochs are needed. The reason is that the influence of $\operatorname{cm}(A)$ cancels out in $\operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1})$ if $|B_i| = |B_{i+1}|$. In this way, the difficulty that we have to say something about $\operatorname{cm}(A)$ disappears.

If $|B_i| \neq |B_{i+1}|$, then $\operatorname{cm}(A)$ shows up with different coefficients in $\operatorname{cm}(C_i)$ and $\operatorname{cm}(C_{i+1})$ and, hence, with a non-zero coefficient in $\operatorname{cm}(P_i) - \operatorname{cm}(P_{i+1})$. This implies that for any position of $\operatorname{cm}(B_i)$ and $\operatorname{cm}(B_{i+1})$, there exists a location for $\operatorname{cm}(A)$ such that $\operatorname{cm}(P_i)$ and $\operatorname{cm}(P_{i+1})$ are close. However, this is only possible if $\operatorname{cm}(A)$ assumes a position in some hyperball of a certain radius. The center of this hyperball depends only on $\operatorname{cm}(B_i)$ and $\operatorname{cm}(B_{i+1})$. We conclude that we can only have $\|\operatorname{cm}(P_1) - \operatorname{cm}(P_2)\| \leq \eta$ and $\|\operatorname{cm}(P_2) - \operatorname{cm}(P_3)\| \leq \eta$ simultaneously if these hyperballs intersect.

Lemmas 1.17 and 1.18 imply the following.

Lemma 1.19 *The probability that there is a sequence of 2^k consecutive sparse improving iterations such that the potential decreases by less than ε over this sequence is at most*

$$(7n)^{4kd} \cdot \left(\frac{4nkd\sqrt{\varepsilon}}{\sigma}\right)^d \leq \left(\frac{c_{\text{sparse}} n^{4k+4}\sqrt{\varepsilon}}{\sigma}\right)^d$$

for some sufficiently large constant c_{sparse}.

Putting everything together

To get a smoothed bound for the number of iterations, we need an upper bound for the objective function of the initial clustering. The proof of the following lemma is almost identical to the proof of Lemma 1.3 and therefore omitted. Here we exploit our assumption that $\sigma \leq 1$.

Lemma 1.20 *Let* $\sigma \leq 1$*, let* $D = 10\sqrt{kd \ln n}$*, and let* Y *be a* σ*-perturbed point set. Then* $\mathbb{P}(Y \not\subseteq [-D, D]^d) \leq n^{-3kd}$.

A consequence of the lemma above is that after the first iteration, the potential is bounded by $ndD^2 = c_{\text{init}} nd^2 k \ln n \leq c_{\text{init}} n^5$ for some constant c_{init}. (The upper bound of $c_{\text{init}} n^5$ is very poor, but simplifies the bounds.)

Theorem 1.21 *For* $d \geq 2$*, the smoothed number of iterations of the* k*-means method is at most* $O(2^k n^{14k+12}/\sigma^2)$.

Proof We choose $\varepsilon = \sigma^2 \cdot n^{-14k-8}$. By Lemma 1.15, the probability that there is a dense iteration that decreases the potential by at most ε is at most cn^{-3kd} for some constant $c > 0$. By Lemma 1.19, the probability that there is a sequence of 2^k consecutive sparse iterations that decrease the potential in total by at most ε is also at most $c'n^{-3kd}$ for some constant $c' > 0$. By Lemma 1.20, the probability that the initial potential is more than $O(n^5)$ is also at most n^{-3kd}.

If any of these events happens nevertheless, we bound the number of iterations by its worst-case bound of n^{3kd} (Inaba et al., 2000). This contributes only $O(1)$ to the expected value. Otherwise, the number of iterations is bounded by $O(2^k n^{14k+13}/\sigma^2)$. □

Towards a truly polynomial bound

The bound obtained in Theorem 1.21 is still quite poor. In particular, it has the number k of clusters in the exponent. It can be shown that the smoothed number of iterations of k-means is bounded by a polynomial in n and $1/\sigma$ (without k or d in the exponent). The idea for this improved analysis is to refine the partitioning of iterations into more types, not only into sparse and dense iterations. However, the analysis becomes technically quite involved, while the analysis presented here already conveys the key ideas.

1.3 Smoothed analysis of the approximation ratio

Local search heuristics are popular not only because they are fast, but also because they succeed relatively often in finding local optima that are not much worse than global optima. In order to understand this theoretically, we would like to analyze

the ratio of the objective value of the local optimum found and of a global optimum. However, there are several issues to this:

- Which local optimum the heuristic finds depends on the initial solution. In fact, a local search heuristic is only fully specified if we also say how the initial solution is computed. For the running time, we have avoided this issue by taking a worst-case approach, i.e., analyzing the maximum running time if we always make the worst possible choice.

  For the approximation ratio, we avoid this issue in the same way by comparing the global optimum with the worst local optimum. However, the downside of this is that we only obtain approximation ratios much worse than the results obtained by using quite simple heuristics to construct the initial solution, rendering the results pure theoretical.
- While local search heuristics often perform very well with respect to speed, their performance in terms of approximation ratio is somewhat mixed. In fact, worst-case examples for the approximation ratio are often quite robust against small perturbations.
- A pure technical issue is that, in order to analyze the approximation ratio, we have to analyze the ratio of two random variables, namely the length of an optimal tour and the length of the tour computed by the algorithm, that are highly dependent.

We consider again the 2-opt heuristic for the TSP, but this time, we use the (standard) Euclidean distances to measure the tour length.

We do not give full proofs in the remainder of this section as most proofs are too lengthy and technical to be presented here. Instead, we restrict ourselves to giving some intuition of the proof ideas.

1.3.1 A simple bound for the approximation ratio of 2-opt

We call a TSP tour through a point set *2-optimal* if it cannot be shortened by a 2-opt step. For a point set Y, we denote by $\mathrm{WLO}(Y)$ (worst local optimum) the length of the longest 2-optimal tour through Y. We denote by $\mathrm{TSP}(Y)$ the length of the shortest TSP tour.

Our goal here is to prove a smoothed approximation ratio of $O(1/\sigma)$. This means that $\mathbb{E}(\mathrm{WLO}(Y)/\mathrm{TSP}(Y)) = O(1/\sigma)$. The idea to prove this is as follows:

- Prove that $\mathrm{TSP}(Y) = \Omega(\sigma \cdot \sqrt{n})$ with high probability.
- Prove that $\mathrm{WLO}(Y) = O(\sqrt{n})$ with high probability.
- If either bound does not hold (which only happens with negligible probability), then we use the trivial upper bound of $n/2$ for the approximation ratio.

The following lower bound for the length of an optimal tour is given without a proof (see also Chapter 8). It follows from concentration of measure results for Euclidean optimization problems.

Lemma 1.22 *There exists a constant $c > 0$ such that $\mathrm{TSP}(Y) \geq c \cdot \sigma\sqrt{n}$ with a probability of at least $1 - \exp(-c'n)$.*

In particular, Lemma 1.22 implies that $\mathbb{E}(\mathrm{TSP}(Y)) = \Omega(\sigma\sqrt{n})$, which we leave as Exercise 1.3.

Next, we state an upper bound for the length of any locally optimal tour. The key insight here is that if a tour is too long, then it must contain two almost parallel edges that are not too far away. These edges can then be replaced by a 2-opt step. Hence, the original tour was not locally optimal.

Lemma 1.23 *Let $Y \subseteq [a,b]^2$ be a set of n points for some $a < b$, and let T be any 2-optimal tour through Y. Then the length $L(T)$ of T is bounded from above by $O\big((b-a) \cdot \sqrt{n}\big)$.*

Combining Lemma 1.23 with the fact that not too many points can be too far outside of the unit hypercube, we obtain the following lemma.

Lemma 1.24 *There exist constants $c, c' > 0$ such that, for all $\sigma \leq 1$, the following holds: the probability that there exists a 2-optimal tour T through Y that has a length of more than $c \cdot \sqrt{n}$ is bounded by $\exp(-c'\sqrt{n})$.*

The upper bound for the length of local optima plus the lower bound for the length of optimal tours together with the trivial worst-case bound of $n/2$ of 2-opt's approximation ratio yield the following result.

Theorem 1.25 *Let $Y \subseteq \mathbb{R}^2$ be σ-perturbed point set. Then*

$$\mathbb{E}\left(\frac{\mathrm{WLO}(Y)}{\mathrm{TSP}(Y)}\right) = O\left(\frac{1}{\sigma}\right).$$

1.3.2 Improved smoothed approximation ratio of 2-opt

In the previous section, we have sketched a bound of $O(1/\sigma)$ for the smoothed approximation ratio of 2-opt. This bound is still far away from explaining the observed approximation performance of 2-opt, which usually finds a solution only a few percent worse than the optimal solution.

The most striking reason that the bound is so poor is the following: we have analyzed the objective value of the globally optimal and locally optimal solution completely independently. The obvious advantage of this is that it avoids all dependencies between the two quantities. The obvious disadvantage is that it only yields a very poor bound: both the upper bound for the length of a locally optimal solution and the lower bound for the length of a globally optimal solution are tight, but the former is achieved if the unperturbed points are spread evenly over $[0,1]^d$, whereas the latter is achieved by putting all unperturbed points at exactly the same location.

By taking the positions of the unperturbed points into account, it is possible to improve the smoothed approximation ratio of 2-opt to $O\big(\log(1/\sigma)\big)$.

This seems to be almost tight, as there exist instances X of n points such that $\mathbb{E}\left(\frac{\mathrm{WLO}(Y)}{\mathrm{TSP}(Y)}\right) = \Omega\left(\frac{\log n}{\log\log n}\right)$ for $\sigma = O(1/\sqrt{n})$. The idea to prove this smoothed lower bound for the approximation ratio is to show that the known worst-case lower bound example for the ratio WLO / TSP of $\Omega(\log n/\log\log n)$ can be made robust against perturbations with $\sigma = O(1/\sqrt{n})$.

However, even the improved bound of $O\big(\log(1/\sigma)\big)$ requires σ to be constant to achieve some constant approximation ratio. Such results are also easily obtained by many simple heuristics for the TSP.

1.4 Discussion and open problems

1.4.1 Running time

Both smoothed analyses that we have presented in Section 1.2 have in common that they are based on analyzing the smallest possible improvement of either a single iteration or a few iterations.

This has been extended to longer sequences of iterations for the flip heuristic for the Max-Cut problem. An instance of Max-Cut is given by an undirected graph $G = (V, E)$ with edge weights $w : E \to [-1, 1]$. The goal is to find a partition $\sigma : V \to \{-1, 1\}$ of the vertices of maximum cut weight

$$\frac{1}{2} \cdot \sum_{e=\{u,v\}\in E} w(e) \cdot \big(1 - \sigma(u)\sigma(v)\big).$$

The flip heuristic for Max-Cut starts with an arbitrary partition. Then it iteratively flips the sign of a vertex if this would increase the cut weight, until it has converged to a local optimum.

The flip heuristic for Max-Cut has been a notoriously difficult problem for a few years because it eluded a smoothed analysis despite its simplicity. In order to make the smoothed analysis of its running time possible, it was necessary to consider much longer sequences of iterations, namely sequences of length linear in the number of vertices. The main challenge then was to find enough independent randomness in such sequences.

In summary, the feature that all smoothed analyses of the running time of local search heuristics have in common seems to be that it is unlikely that iterations cause only very small improvements. In contrast, the worst-case constructions to show an exponential lower bound for the running time of these heuristics are quite fragile. They are usually based on implementing a "binary counter", where each bit is represented by some small gadget. The gadgets for the different bits are scaled

versions of each other, which implies that all but the gadgets for the most significant bits are tiny and easily break under small perturbations.

We conclude this section with three open problems: first, prove that the Lin-Kernighan heuristic for the TSP has polynomial smoothed running time. This heuristic has incredible performance in practice, much better than 2-opt. However, it seems to be difficult to find a compact representation of the iterations. The reason for this is that each iteration replaces an unbounded number of edges.

Second, devise general techniques for the smoothed analysis of local search heuristics. Despite all the similarities, each smoothed analysis of a local search heuristic so far is tailored to the specific algorithm. Is it possible to develop a general framework or general conditions that imply smoothed polynomial running time?

Third, all smoothed analyses of local search heuristics use the decrease of the objective function by showing that it is unlikely that any iteration (or all iterations in some sequence) yields only a small improvement. This still seems to be rather pessimistic, as it is unlikely that a local search heuristic performs very often iterations that only yield the smallest possible improvement. Is it possible to do a smoothed analysis "beyond the smallest improvement" in order to get improved bounds? In particular, the polynomial bounds obtained in the smoothed analyses of the k-means method and the flip heuristic for Max-Cut have quite large degree. We assume that considerably improving these bounds requires new ideas.

1.4.2 Approximation ratio

Given the relatively strong results for local search heuristics with respect to running time and the quite poor results with respect to approximation ratio, the question arises why this is the case. In fact, the results for the approximation ratio that we have presented here are only of purely theoretical interest, as – in case of the TSP – even a simple insertion heuristic achieves an approximation ratio of 2 in the worst case.

For the k-means method, the situation is different: the approximation performance of the k-means method is not very good in the first place. In fact, the main reason why the k-means method is so popular is its speed. This allows us to run it many times on the same data set with different initializations. The hope is that for at least one initialization, we get a good clustering. In general, only very poor guarantees are possible, even in the framework of smoothed analysis (Exercise 1.8).

Thus, the question arises if smoothed analysis is the right tool for analyzing the approximation ratio of algorithms. The few successes – although non-trivial – are merely of theoretical interest. A reason for this could be that the worst-case examples for the approximation ratio seem to be much more robust against small perturbations.

We conclude this section with three open problems: first, prove a non-trivial

bound for the approximation performance of the Lin-Kernighan heuristic for the TSP mentioned in the previous section.

Second, apply smoothed analyses to "hybrid heuristics". The smoothed analysis so far have only been applied to "pure" local search heuristics. However, in particular the approximation ratio depends heavily on a good initialization. Hence, we have to take into account two algorithms (the initialization and the actual heuristic) instead of only one. Is it possible to show improved bounds in this setting? For instance, the k-means method as described in this chapter has a poor approximation performance. Is it possible to prove a good approximation performance, when initialized cleverly?

Third, find a meaningful way to do smoothed analysis of approximation ratios or devise a different approach towards "approximation ratio beyond the worst case" that really explains the approximation performance of such heuristics in practice. One of the strong points of smoothed analysis is that it is a relatively problem-independent semi-random input model. Essentially the only property that is needed for a smoothed analysis is that the concept of "small perturbations" make sense for the problem considered. However, this advantage is also a disadvantage: because of problem independence, smoothed analysis completely ignores any structure that interesting instances might have. Thus, in order to address this question, it might be necessary to come up with more problem-specific input models for "non-worst-case" instances.

1.5 Notes

Smoothed analysis has been introduced by Spielman and Teng (2004) in order to explain the performance of the simplex method for linear programming. Arthur and Vassilvitskii (2009) were the first to apply smoothed analysis to local search heuristics, namely to the k-means method and to the ICP algorithm.

The original smoothed analysis of 2-opt, both for running time and approximation ratio and including the concept of linked pairs, was done by Englert et al. (2014) (see Röglin and Schmidt (2018) for a corrected version of Lemma 1.8). They also provided a Euclidean instance on which 2-opt needs an exponential number of iterations to compute a local optimum. Furthermore, they provided a smoothed analysis of 2-opt for TSP in (non-Euclidean) general graphs (Englert et al., 2016). The analysis of the running time under Gaussian noise using squared Euclidean distances presented here follows a simplified proof by Manthey and Veenstra (2013). The improved smoothed analysis of the approximation ratio is due to Künnemann and Manthey (2015). The absolute length of locally optimal tours is by Chandra et al. (1999). They also proved a worst-case bound of $O(\log n)$ for the approximation ratio of 2-opt. The high-probability statement in Lemma 1.22 follows from Rhee's isoperimetric inequality (Rhee, 1993). Johnson and McGeoch provide experimental

evidence for the performance of 2-opt and the Lin-Kernighan heuristic (Johnson and McGeoch, 1997, 2002).

Arthur and Vassilvitskii (2009) proved a bound polynomial in n^k and $1/\sigma$ for the smoothed running time of the k-means method and a polynomial bound for the so-called ICP algorithm. The bound for the k-means method has been improved to a polynomial bound by Arthur et al. (2011). The proof presented here combines the two proofs to simplify the argument. A weaker bound can be obtained for more general distance measures (Manthey and Röglin, 2013). Vattani (2011) provided an example in 2-dimensional space for which k-means needs exponential time to compute a local optimum. The upper bound for the worst-case running time is by Inaba et al. (2000).

The first smoothed analysis of the running time of the flip heuristic for the Max-Cut problem for graphs of bounded degree has been done by Elsässer and Tscheuschner (2011) (see Exercise 1.9). Etscheid and Röglin (2017) proved a quasi-polynomial bound in general graphs. For the special case of complete graphs, this has been improved by Angel et al. (2017) to a polynomial bound with high probability.

References

Angel, Omer, Bubeck, Sébastien, Peres, Yuval, and Wei, Fan. 2017. Local max-cut in smoothed polynomial time. Pages 429–437 of: *Proc. of the 49th Ann. ACM Symp. on Theory of Computing (STOC)*. ACM.

Arthur, David, and Vassilvitskii, Sergei. 2009. 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.

Arthur, David, Manthey, Bodo, and Röglin, Heiko. 2011. Smoothed Analysis of the k-Means Method. *Journal of the ACM*, **58**(5).

Chandra, Barun, Karloff, Howard, and Tovey, Craig. 1999. New Results on the Old k-Opt Algorithm for the Traveling Salesman Problem. *SIAM Journal on Computing*, **28**(6), 1998–2029.

Elsässer, Robert, and Tscheuschner, Tobias. 2011. Settling the Complexity of Local Max-Cut (Almost) Completely. Pages 171–182 of: Aceto, Luca, Henzinger, Monika, and Sgall, Jiri (eds), *Proc. of the 38th Int. Coll. on Automata, Languages and Programming (ICALP)*. Lecture Notes in Computer Science, vol. 6755. Springer.

Englert, Matthias, Röglin, Heiko, and Vöcking, Berthold. 2014. Worst Case and Probabilistic Analysis of the 2-Opt Algorithm for the TSP. *Algorithmica*, **68**(1), 190–264.

Englert, Matthias, Röglin, Heiko, and Vöcking, Berthold. 2016. Smoothed Analysis of the 2-Opt Algorithm for the General TSP. *ACM Transactions on Algorithms*, **13**(1), 10:1–10:15.

Etscheid, Michael, and Röglin, Heiko. 2017. Smoothed Analysis of Local Search for the Maximum-Cut Problem. *ACM Transactions on Algorithms*, **13**(2), 25:1–25:12.

Inaba, Mary, Katoh, Naoki, and Imai, Hiroshi. 2000. Variance-Based k-Clustering Algorithms by Voronoi Diagrams and Randomization. *IEICE Transactions on Information and Systems*, **E83-D**(6), 1199–1206.

Johnson, David S., and McGeoch, Lyle A. 1997. The Traveling Salesman Problem: A Case Study. Chap. 8 of: Aarts, Emile, and Lenstra, Jan Karel (eds), *Local Search in Combinatorial Optimization*. John Wiley & Sons.

Johnson, David S., and McGeoch, Lyle A. 2002. Experimental Analysis of Heuristics for the STSP. Chap. 9 of: Gutin, Gregory, and Punnen, Abraham P. (eds), *The Traveling Salesman Problem and its Variations*. Kluwer Academic Publishers.

Künnemann, Marvin, and Manthey, Bodo. 2015. Towards Understanding the Smoothed Approximation Ratio of the 2-Opt Heuristic. Pages 859–871 of: Halldórsson, Magnús M., Iwama, Kazuo, Kobayashi, Naoki, and Speckmann, Bettina (eds), *Proc. of the 42nd Int. Coll. on Automata, Languages and Programming (ICALP)*. Lecture Notes in Computer Science, vol. 9134. Springer.

Manthey, Bodo, and Röglin, Heiko. 2013. Worst-Case and Smoothed Analysis of k-Means Clustering with Bregman Divergences. *Journal of Computational Geometry*, **4**(1), 94–132.

Manthey, Bodo, and Veenstra, Rianne. 2013. Smoothed Analysis of the 2-Opt Heuristic for the TSP: Polynomial Bounds for Gaussian Noise. Pages 579–589 of: Cai, Leizhen, Cheng, Siu-Wing, and Lam, Tak-Wah (eds), *Proc. of the 24th Ann. Int. Symp. on Algorithms and Computation (ISAAC)*. Lecture Notes in Computer Science, vol. 8283. Springer.

Rhee, WanSoo T. 1993. A Matching Problem and Subadditive Euclidean Functionals. *The Annals of Applied Probability*, **3**(3), 794–801.

Röglin, Heiko, and Schmidt, Melanie. 2018. *Randomized Algorithms and Probabilistic Analysis*.

Schäffer, Alejandro A., and Yannakakis, Mihalis. 1991. Simple Local Search Problems That are Hard to Solve. *SIAM Journal on Computing*, **20**(1), 56–87.

Spielman, Daniel A., and Teng, Shang-Hua. 2004. Smoothed Analysis of Algorithms: Why the Simplex Algorithm Usually Takes Polynomial Time. *Journal of the ACM*, **51**(3), 385–463.

Vattani, Andrea. 2011. k-Means Requires Exponentially Many Iterations even in the Plane. *Discrete and Computational Geometry*, **45**(4), 596–616.

Yukich, Joseph E. 1998. *Probability Theory of Classical Euclidean Optimization Problems*. Lecture Notes in Mathematics, vol. 1675. Springer.

Exercises

1.1 Consider the following probabilistic model for the TSP: given a finite set V of n vertices, the distance $d(u, v)$ between any $u, v \in V$ is drawn independently and uniformly at random from the interval $[0, 1]$.

Prove that the expected number of iterations that the 2-opt heuristic needs on such instances is at most $O(n^6 \log n)$.

1.2 In the analysis of the 2-opt heuristic and the k-means method, we have restricted ourselves to "reasonably small" σ and claimed that this does not pose

any severe restriction. Justify this by showing that the smoothed number of iterations for both algorithms is monotonically decreasing in σ.

More formally, let $T(n, \sigma)$ denote the smoothed number of iterations of either local search algorithm on instances of n points perturbed by Gaussians of standard deviation σ. Show that $T(n, \sigma)$ is non-increasing in σ.

1.3 Let $X \subseteq \mathbb{R}^2$ be a set of n points in the Euclidean plane and let Y be a perturbation of X as described in Section 1.2.2.

Prove that $\mathbb{E}(\mathrm{TSP}(Y)) = \Omega(\sigma \cdot \sqrt{n})$.

Hint: For any $y \in Y$, estimate the distance to a closest neighbor of y in $Y \setminus \{y\}$.

1.4 Prove Lemma 1.10.

1.5 Prove Lemma 1.11.

1.6 Prove the following stronger version of Lemma 1.16: the length of every epoch is bounded by 3.

1.7 Consider the following variant of the k-means method, which we call "lazy k-means": in every iteration, only one point is reassigned to a new cluster. Ties are broken arbitrarily. After reassigning a single point, the two cluster centers involved are adjusted.

Show that the smoothed running time of lazy k-means is bounded by a polynomial in n and $1/\sigma$, without any exponential dependency on d or k .

Hint: Consider epochs and adjust the concept of η-coarseness appropriately. In order to avoid a factor of 2^k, you have to use the result of Exercise 1.6.

1.8 For the approximation ratio of the k-means method, we consider the ratio of objective value of the worst local optimum divided by objective value of a global optimum.

(a) Give a simple instance that shows that the approximation ratio of the k-means method cannot be bounded by a constant.

(b) Let $\sigma \ll 1$. Show that the smoothed approximation ratio of k-means is not $o(1/\sigma^2)$. Here, smoothed approximation ratio refers again to the expected ratio of the worst local optimum to a global optimum.

1.9 For graphs with maximum degree $O(\log n)$, proving a smoothed polynomial bound for the flip heuristic for Max-Cut is much easier than for general graphs.

For a graph $G = (V, E)$, let Δ be the maximal degree of G, let $n = |V|$, and let $m = |E|$. Let $\phi \geq 1$, and let $f_e : [0, 1] \to [0, \phi]$ be a density function for $e \in E$. Let w_e be drawn according to f_e. We consider the flip heuristic for Max-Cut on the instance (G, w). Let $\delta_{\min}$ be the smallest possible improvement caused by any possible iteration of the flip heuristic. Let T be the maximum number of iterations that the flip heuristic needs on the instance (G, w).

(a) Prove that $\mathbb{P}(\delta_{\min} \leq \varepsilon) \leq 2^{\Delta} n \phi \varepsilon$.

(b) Prove that $\mathbb{P}(T \geq t) \leq 2^{\Delta} nm\phi/t$ for all $t \in \mathbb{N}$ with $t \geq 1$.

(c) Prove that $\mathbb{E}(T) = O(2^{\Delta} n^2 m \phi)$.