Approximation Algorithms for Outlier Removal in Convex Hulls

Michael Biro ∗  Justine Bonanno†  Roozbeh Ebrahimi †  Lynda Montgomery ∗

Abstract

Given \( n \) points in \( \mathbb{R}^2 \), we give approximation algorithms to find a subset of \( k \) of the points that has minimum-area or minimum-perimeter convex hull. We give algorithms that, for each \( k \), yield a constant-factor approximation to the minimum-perimeter problem in linear time. We also show a 2-approximation for the minimum-area problem in time \( O(\min(n^3 \log n, n^2 \log n + kn(n-k)(n-k+\log k))) \), as well as a heuristic for both problems that appears to work well in practice.

1 Introduction

The problem of finding a subset of size \( k \) from a point set of size \( n \) that has the least-perimeter convex hull was considered in several papers, going back to 1983, with results improving from the original \( O(k^2 n \log n + k^3 n) \) in Dobkin et al. [8] to \( O(n \log n + k^3 n) \) in Datta et al. [7] and Eppstein et al. [4]. Finding the subset of \( k \) points with the minimum-area convex hull was considered in Eppstein [3], and Eppstein et al. [5], where they give \( O(kn^2) \) and \( O(n^2 \log n + k^3 n^2) \) exact algorithms.

These algorithms give the exact solution, but their runtimes can be \( \Omega(n^3) \) and \( \Omega(n^5) \), respectively for perimeter and area, and for large \( k \). Recently, for \( k = n - c \), Atanassov, et al. [2] gave exact \( O(n \log n + (\frac{n}{c}) (3c)^{c+1} n) \) algorithms for both problems, however this still leaves a difficulty of finding exact solutions for \( k \) sufficiently far from \( n \). We give a linear-time, constant-factor approximation to the minimum-perimeter problem, as well as a 2-approximation for the minimum-area problem that runs in \( O(\min(n^3 \log n, n^2 \log n + kn(n-k)(n-k+\log k))) \) time. In addition, we describe a heuristic for choosing the outliers that seems to work well in practice.

2 Minimum-Perimeter Convex Hull

We approximate the \( k \)-outlier minimum-perimeter convex hull by approximating the shape of the convex hull as either a rectangle or circle.

Lemma 2.1. Let \( P \) be a convex set in \( \mathbb{R}^2 \). Then the perimeter of \( P \) is at most a factor \( \sqrt{2} \) away from the perimeter of the minimum-perimeter axis-parallel rectangle containing \( P \).

Lemma 2.2. Let \( P \) be a convex set in \( \mathbb{R}^2 \). Then the perimeter of \( P \) is at most a factor of \( \frac{3}{2} \) away from the perimeter of the minimum disc enclosing \( P \).

We now use recent results of Ahn et al. [1], that finds the minimum-perimeter axis-parallel rectangle in time \( O(n + k^3) \), and results of Har-Peled et al. [9] that finds a \((1 + \epsilon)\)-approximation to the minimum enclosing disk in time \( O(n + n \cdot \min(\frac{1}{\epsilon} \log^2 (\frac{1}{\epsilon}), k)) \). Then we have,

Theorem 2.3. The \( k \)-outlier minimum-perimeter convex hull problem can be approximated by a factor of \( \sqrt{2} \) in time \( O(n + k^3) \), and a factor of \( \frac{3}{2} (1 + \epsilon) \) in time \( O(n + n \cdot \min(\frac{1}{\epsilon} \log^2 (\frac{1}{\epsilon}), k)) \).

These algorithms give constant-factor approximations to the minimum-perimeter problem for a variety of values of \( k \). If \( k = O(n^{1/3}) \) then the rectangle algorithm yields a \( \sqrt{2} \) approximation in linear time, and if \( k = \Omega(n^{1/3}) \), then the disk algorithm gives a \( \frac{3}{2} \) \((1 + \epsilon)\)-approximation in linear time (for constant values of \( \epsilon \)). Therefore, for each value of \( k \), we give constant-factor approximations to the \( k \)-outlier minimum-perimeter problem that run in linear time.

Corollary 2.4. For each \( k \), the \( k \)-outlier minimum-perimeter convex hull problem can be approximated by a constant factor in linear time.

3 Minimum-Area Convex Hull

We approximate the \( k \)-outlier minimum-area convex hull problem by approximating the shape of the convex hull as a rectangle with arbitrary orientation.
Lemma 3.1. Let $P$ be a convex set in $\mathbb{R}^2$. Then the area of $P$ is at most a factor of 2 away from the area of the minimum-area rectangle enclosing $P$.

Proof. Take the longest diagonal $D$ of $P$, and construct a minimal enclosing rectangle $R$ with two sides parallel to $D$. Take $R$ along with the $D$ and the two points defining the perpendicular edges to $D$ of $R$. The area of $P$ is at least the area of the two triangles thus defined, and the two triangles take up exactly half of $R$. Therefore, $R$ has area at most twice the area of $P$, and as the minimum-area rectangle has area at most the area of $R$, it has area at most twice the area of $P$. \hfill \Box

Using the idea in the above proof, we construct an algorithm that checks every possible longest diagonal $D$ of the point set $P$, then computes the minimum-area rectangle containing at least $k$ points among all such diagonals, in time $O(n^2 \log n)$.

1. Examine every pair of points $(p, q)$ and look at the strip defined by lines perpendicular to $pq$ through $p$ and $q$ respectively. Find the points of $P$ that lie in the strip.

2. Sort the points in the strip by distance of $pq$, and for every point in the strip find the corresponding point so that the rectangle defined by the four points contains $k$ points of $P$.

3. Take the minimum-area rectangle among all rectangles constructed.

We combine this with the recent result of Das et al. [6], that finds a minimum-area rectangle in time $O(\log^3 n)$.

Theorem 3.2. The $k$-outlier minimum-area convex hull problem can be 2-approximated in time $O(\min(n^3 \log n, n^2 \log n + kn(n-k)(n-k+\log k)))$.

This approximation is useful if $k$ is $\Theta(n)$, so the optimal solution in [3] runs in time $\Omega(n^5)$, but not $n-c$ for constant $c$, as the optimal solution in [2] runs in time exponential in $c$.

4 Heuristic

In this section, we describe a heuristic for the minimum-area convex hull problem. It runs in time $O(n(n-k) \log n)$ in the worst case, and while there are cases where it gives arbitrarily bad approximations, in practice it has yielded good results.

1. Find the diameter $d$ of the points, say between points $a$ and $b$. Define a lune $L$ by intersecting disks of radius $d$ centered at $a$ and $b$, respectively.

2. Let $O$ be the midpoint of $ab$. Divide $L$ evenly into 4 equal sectors around $O$ and associate the points with their respective sector. Sort the points with respect to distance from $O$.

3. If any sector contains fewer than $n-k$ points, remove all points in that sector.

4. Remove points in order of decreasing distance from $O$ until either $n-k$ points are removed, or one of $a, b$ is removed. If the latter, begin again.

5 Conclusion

For future work, we would like to find a PTAS for minimum-perimeter that runs in linear or near-linear time, as well as improve the running time of the approximation for minimum-area convex hulls.

References


