# A combinatorial measure of closeness in point sets

**Alexander Pilz**[*1] **and Patrick Schnider**[1]

1 **Department of Computer Science, ETH Zürich. Zürich, Switzerland.**
{alexander.pilz,patrick.schnider}@inf.ethz.ch

**Abstract**

We introduce *stripe closeness* and *stripe remoteness*, two combinatorial measures that capture how close together or far apart a set of query points lies within another set of points. The idea behind these concepts is that we look at all possible projections of the point set to a line and count the number of points that lie between the query points. For two points in a point set, the notion of stripe closeness can be seen as a combinatorial distance measure. We give bounds on the stripe closeness of two closest points. Further, we analyze stripe remoteness for triples in point sets and show that there are always three points that have high stripe remoteness.

## 1 Introduction

Let $P$ be a set of points in general position (i.e., no three points on a line) and let $a$ and $b$ two points of $P$. How far is $a$ from $b$? The common answer to this question is of course the Euclidean distance of $a$ and $b$. However, this distance depends on the embedding of $P$ and is not invariant under affine transformations. In many settings where point sets occur, we are not interested in actual coordinates of the points, but only their combinatorial structure (e.g., their allowable sequence or their order type). In these settings it seems natural to define a distance measure that only depends on the combinatorial structure of the point set.

For points on a line, there is a natural combinatorial distance measure: For any two points $a$ and $b$ in $P$ we define the distance of $a$ and $b$ as the number of points of $P$ that lie between $a$ and $b$. Alternatively, we can also count the points that are not between $a$ and $b$ and define the distance between the total number of points and the number of these points. We can extend this idea for more query points. We define the *remoteness* of a one-dimensional point set $Q$ with respect to a one-dimensional point set $P$ as follows: Let $a$ and $b$ be the two extreme points of $Q$. Then the remoteness of $Q$ with respect to $P$ is the number of points of $P$ that are between $a$ and $b$. In particular, if $Q$ consists of two points, then a small remoteness can be considered to have a small combinatorial distance. We generalize this concept to $\mathbb{R}^2$ by considering all projections of the two-dimensional point set to a line.

A *stripe* $s = (\ell_1, \ell_2)$ is a pair of two parallel lines $\ell_1, \ell_2$ in the plane. We say that the area of the plane that lies between the two lines is *inside* the stripe and denote it by $\mathsf{int}(s)$, while the rest of the plane is *outside* of the stripe, denoted by $\mathsf{out}(s)$. We consider $\ell_1$ and $\ell_2$ to be both inside and outside of $s$, that is, $\mathsf{int}(s) \cap \mathsf{out}(s) = \ell_1 \cup \ell_2$. Let $P$ be a planar point set in general position. For any stripe $s$ we define $i_P(s) := |\{p \in P \mid p \in \mathsf{int}(s)\}|$ and $o_P(s) := |\{p \in P \mid p \in \mathsf{out}(s)\}|$ as the number of points of $P$ inside and outside of $s$, respectively. Let $Q$ be another set of points in general position. We define the *stripe remoteness* of $Q$ with respect to $P$ as follows: consider all the stripes for which all of $Q$ lies inside. Among those, pick one that has the smallest number of points of $P$ inside. The stripe remoteness of $Q$ with respect to $P$, denoted by $\mathsf{instripe}_P(Q)$, is the number of points of $P$

* Supported by a Schrödinger fellowship of the Austrian Science Fund (FWF): J-3847-N35.

34th European Workshop on Computational Geometry, Berlin, Germany, March 21–23, 2018.
This is an extended abstract of a presentation given at EuroCG'18. It has been made public for the benefit of the community and should be considered a preprint rather than a formally reviewed paper. Thus, this work is expected to appear eventually in more final form at a conference with formal proceedings and/or in a journal.

inside this stripe. Using the notation above, we can write this as

$$\mathsf{instripe}_P(Q) := \min_{s:Q\subset\mathsf{int}(s)} i_P(s) \ .$$

Note that if $|Q| = 2$ and $Q \subset P$, then $\mathsf{instripe}(Q) = 2$ as we can always choose both $\ell_1$ and $\ell_2$ to be the line through the two points. Hence, stripe remoteness is not a good candidate for a combinatorial distance measure. But for $|Q| > 2$ the situation is non-trivial. A point set $Q$ having high stripe remoteness in $P$ can be interpreted in the following way: In every projection to a line, there are two points of $Q$ that have high distance w.r.t. $P$ (i.e., the number of projected points of $P$ between them is large). We show that for every point set $P$ in general position, there are 3 points in $P$ whose stripe remoteness is in $\Omega(|P|)$.

Note that in the one-dimensional case we might as well count the number of points that are not between $a$ and $b$. The fewer such points there are, the further $a$ is from $b$. We can also extend this idea to more points and two dimensions: We say that a stripe $s$ is *between* $Q$ if all points of $Q$ are in $\mathsf{out}(s)$ and each connected component of $\mathsf{out}(s)$ contains at least one point of $Q$; this is denoted by $Q \prec s$. We define the *stripe closeness* of $Q$ with respect to $P$ as the minimal number of points of $P$ outside a stripe that is between $Q$, i.e.,

$$\mathsf{outstripe}_P(Q) := \min_{s:Q\prec s} o_P(s) \ .$$

This measure is non-trivial already for $|Q| = 2$. We may define

$$d_P(a,b) := |P| - \mathsf{outstripe}_P(\{a,b\}) + 1 \ .$$

This corresponds to the maximum number of points strictly inside a stripe s.t. one of $a$ and $b$ is on each of the two lines defining the stripe. It is not too hard to check that $d_P$ is a metric when defining $d_P(a,a) := 0$; note that (somewhat counterintuitive) the additive 1 is needed for the triangle inequality. Also note that any point set $P$ contains two points $a$ and $b$ such that $d_P(a,b) = |P| - 1$. A point set $Q$ having high stripe closeness can be interpreted in the following way: In every projection to a line, there are two points of $Q$ with no other point of $Q$ and only few points of $P$ between them. We show that for every point set $P$ in general position, we can find a subset of 2 points that have linear stripe remoteness. For the combinatorial distance measure $d_P(a,b)$, this implies that there are always two points of distance at most $(1-c)|P|$, for some constant $c$. We will see that this is asymptotically tight.

Projections of multivariate data to one and two dimensions is common in data analysis (see, e.g., [5]). However, distances usually do not only depend on the combinatorial properties of the point set. Combinatorial distance measures for two points $a$ and $b$ in a finite set $P$ may be defined via the size of the intersection $P \cap R$ of $P$ and a region $R$ that contains $a$ and $b$. More specifically, so-called *region-counting* distance functions have been used [3,6]: we are given two points $p$ and $q$ as well as a region $R$; translate, rotate, and scale both the points and $R$ such that $p$ and $q$ coincide with $a$ and $b$, and measure the distance as the size of the intersection with $P$ and the transform of $R$. These measures have been used for point searching and nearest-neighbor problems [3,4,6]. However, they are not invariant under affine transformations. This is a property fulfilled by our approach; the distances are equivalent for all point sets with the same allowable sequence (i.e., there is a bijection between the point sets such that the order of the slopes defined by all point pairs is preserved). Our approach of taking the minimum or maximum is inspired by combinatorial properties of single points w.r.t. the point set, e.g., the *Tukey depth* [7] of a point $p$ (which is the minimal number of points contained in a half-plane that also contains $p$). See [2] for a survey on depth measures.

## 2 Stripe Remoteness

In this section we prove the following result:

▶ **Theorem 2.1.** *Let $\alpha$ be the unique zero of $f(x) = 2x^3 - 3x^2 - 6x + 1$ in $[0, 1]$ ($\alpha \approx 0.155792$). Then for every $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that for every $n > n_0$ any set $P$ of $n$ points in general position contains three points $p_1, p_2, p_3$ for which $\mathsf{instripe}_P(\{p_1, p_2, p_3\}) \geq (\alpha - \epsilon)n$.*

This result can also be phrased in a slightly less technical manner:

▶ **Corollary 2.2.** *Let $P$ be a set of $n$ points in general position. Then $P$ contains three points $p_1, p_2, p_3$ such that any stripe with $p_1, p_2, p_3$ inside has $\Omega(n)$ points of $P$ inside.*

**Proof of Theorem 2.1.** Let $P$ be a point set of $n$ points in general position. We want to show that there are 3 points such that every stripe with them inside has at least $c$ points inside, where $c = (\alpha - \epsilon)n$ for any $\epsilon > 0$. Let $A$ be the dual line arrangement of $P$ under the point-line-duality, in which we map a point $p = (x_p, y_p)$ to a line $p^* : y = x_p x + y_p$ (we may assume w.l.o.g. that no two points have the same $x$-coordinate). In the dual setting, a stripe translates into a vertical line segment, and the points that lie inside the stripe correspond to the lines intersected by the segment. Hence, in the dual setting, we want to find three lines such that any vertical line segment intersecting these three lines intersects at least $c$ lines. We will show that there are three such lines in the following way: for every triple $T$ of lines, we look at the shortest vertical line segment $s_T(-\infty)$ that intersects the three lines and lies on an (arbitrary) vertical line to the left of the leftmost crossing of the arrangement. We list all the lines crossed by $s_T(-\infty)$ in a list $L_T(-\infty)$. We then sweep the arrangement from left to right, always looking at the shortest vertical line segment $s_T(x)$ intersecting the triple and update the list $L_T(x)$ of lines crossed by it, whenever necessary. Clearly, an update is only necessary after the sweeping line passes over a crossing, so we only need to consider one $x$-coordinate between any two consecutive crossings $c_i$ and $c_{i+1}$, which we denote by $x_i$. Additionally, let $x_0$ be an $x$-coordinate to the left of the first crossing $c_1$. We say that the triple $T$ of lines is *valid after crossing* $c_i$ if during the whole movement from left infinity to $x_i$ (or, equivalently, shortly after the $i$th crossing), the list of crossed lines $L_T$ contains at least $c$ lines, that is, $|L_T(x_j)| \geq c$ for all $j \in \{0, \dots, i\}$. In particular, any triple that is valid after the last crossing satisfies the desired properties.

The triples $T$ that are valid before the first crossing are the ones for which $s_T(x_0)$ intersects at least $k$ lines, for $k \geq c$. As $s_T(x_0)$ is the shortest line segment intersecting $T$, the topmost and bottommost intersected lines have to be lines of $T$, the third line of $T$ being one of the remaining $k - 2$. There are $n + 1 - k$ pairs of lines with exactly $k - 2$ lines between them, thus the total number of triples that are valid before the first crossing is

$$\sum_{k=c}^{n} (k-2)(n+1-k) = \frac{1}{6}(c-n-1)(2c^2 - cn - 10c - n^2 + 4n + 12) \ .$$

Moving over a crossing $c_i$, a triple $T$ becomes invalid only if $s_T(x_{i-1})$ intersects exactly $c$ lines, the topmost or the bottommost line is one of the lines of crossing $i$ and the second line in the crossing is not in $T$. More precisely, let $c_i$ be the crossing of two lines $a$ and $b$ where $a$ is above $b$ before the crossing. Let $T$ be a triple that is valid after crossing $c_{i-1}$ and that contains $a$ where $a$ is the topmost line intersected by $s_T(x_{i-1})$. If $b$ is also in $T$, then $s_T(x_i)$ intersects the same lines as $s_T(x_{i-1})$, only that $a$ and $b$ have switched places, so $T$ is still valid after crossing $c_i$. If $s_T(x_{i-1})$ intersects more than $c$ lines, $T$ is still valid after the crossing $c_i$. As the bottommost intersected line has to be in $T$, and a triple with $b$ does not

become invalid, at most $(c-3)$ triples can become invalid at crossing $c_i$, as right before it there are $c-2$ lines between $a$ and the bottommost line, one of them being $b$. The same arguments hold if $b$ is the bottommost line intersected by $s_T(x_{i-1})$, so the number of triples that become invalid at crossing $c_i$ is at most $2(c-3)$. Thus the total number of triples that are still valid after sweeping over all $\binom{n}{2}$ crossings of $A$ is at least

$$\begin{aligned}\beta(n) &:= \sum_{k=c}^{n}(k-2)(n+1-k) - 2(c-3)\binom{n}{2} \\ &= \frac{c^3}{3} - \frac{c^2 n}{2} - 2c^2 - cn^2 + \frac{7cn}{2} + \frac{11c}{3} + \frac{n^3}{6} + \frac{5n^2}{2} - \frac{17n}{3} - 2 \ .\end{aligned}$$

In particular, if $c$ is linear in $n$, that is, $c = \gamma n$, this is a polynomial of degree 3 with leading coefficient $\frac{\gamma^3}{3} - \frac{\gamma^2}{2} - \gamma + \frac{1}{6}$. This leading coefficient is larger than 0 in the interval $(0,1)$ if and only if $\gamma < \alpha$. Hence, for $\gamma = \alpha - \epsilon$, we have that $\lim_{n\to\infty}(\beta(n)) = \infty$, so for $n$ large enough, there is at least one triple that is still valid after the last crossing. ◀

On the other hand, there are point sets where $\mathsf{instripe}(\{p_1,p_2,p_3\}) \leq (1-\epsilon)n$ for any three points $p_1, p_2, p_3$: Let $P$ be a set of points in convex position and let $Q = \{p_1,p_2,p_3\}$ be any three points of $P$. Let $h_1$, $h_2$ and $h_3$ be the number of points along the boundary of the convex hull between $p_1$ and $p_2$, $p_2$ and $p_3$ and $p_3$ and $p_1$, respectively. By the pigeonhole principle, one of these number, without loss of generality $h_1$, is at least $\frac{n-1}{3}$. Consider the stripe defined by the line through $p_1$ and $p_2$ and the parallel line through $p_3$. This stripe has $Q$ inside, but all the points between $p_1$ and $p_2$, of which there are at least $\frac{n-1}{3}$ many outside.

## 3 Stripe Closeness

We proceed with showing our bounds on the stripe closeness of two points.

▶ **Theorem 3.1.** *Let $\alpha = \sqrt{5} - 2 \approx 0.23607$. Then for every $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that for every $n > n_0$ any set $P$ of $n$ points in general position contains two points $p_1, p_2$ for which $\mathsf{outstripe}_P(\{p_1,p_2\}) \geq (\alpha-\epsilon)n + 2$.*

Again, the result can be phrased in a less technical manner:

▶ **Corollary 3.2.** *Let $P$ be a set of $n$ points in general position. Then $P$ contains two points $p_1, p_2$ such that any stripe with $p_1, p_2$ outside has $\Omega(n)$ points of $P$ outside.*

**Proof of Theorem 3.1.** Let $P$ be a set of $n$ points in general position. As in the proof of Theorem 2.1, we will work in the dual setting, only that now we want to find a pair of lines $\ell_1$ and $\ell_2$ such that every shortest vertical line segment intersecting the pair only intersects few lines, namely at most $c := (1-\alpha+\epsilon)n$ many. As the points outside of a stripe correspond to the lines not intersected by the dual vertical line segment, the existence of such a pair of lines shows the claimed result, as for a shortest vertical line segment intersecting $\ell_1$ and $\ell_2$, these two lines must be the topmost and bottommost intersection, and we can thus shorten every segment slightly, such that at least $n - c + 2 = n - (1-\alpha+\epsilon)n + 2 = (\alpha-\epsilon)n + 2$ many lines are not intersected.

We will again show the existence of such a pair of lines using a sweeping argument. For any pair $R$ of lines, let $s_R(x)$ be the shortest vertical line segment intersecting $R$ at that $x$-coordinate and let $L_R(x)$ be the respective list of intersected lines. Again, let $x_i$ be an $x$-coordinate after the crossing $c_i$ and before the crossing $c_{i+1}$. Analogously to triples in

the previous section, we call a pair of lines $R$ *valid after crossing* $c_i$ if during the whole movement from left infinity to $x_i$, the list of crossed lines $L_T$ contains at most $c$ lines, that is, $|L_T(x_k)| \leq c$ for all $k \in \{0, \dots, i\}$.

The number of pairs that are valid before the first crossing can be computed as

$$\sum_{k=2}^{c}(n+1-k) = \frac{1}{2}(c-1)(2n-c) \ .$$

Let now $c_i$ be the crossing of two lines $a$ and $b$ where $a$ is above $b$ before the crossing. Let $\ell_1, \ell_2$ be a pair of lines and let $\ell_1$ be above $\ell_2$ at the $x$-coordinate of the crossing $c_i$. Assume that the pair $\ell_1, \ell_2$ becomes invalid at the crossing $c_i$. Then $s_{\{\ell_1,\ell_2\}}(x_{i-1})$ crosses exactly $c$ lines and either $\ell_1 = b$ or $\ell_2 = a$, as in all other cases the segment $s_{\{\ell_1,\ell_2\}}(x_i)$ crosses $c$ or more lines. In particular, at each crossing at most two pairs become invalid. However, the number of initially valid pairs is strictly smaller than two times the number of crossings, and we therefore need to be more thorough. Note that if $\ell_1 = b$, the pair can only become invalid if there are at least $c-1$ lines under the crossing, that is, if the crossing is above the $c$-level of the arrangement. Similarly, if $\ell_2 = a$, the pair only becomes invalid if the crossing is below the $(n-c-1)$-level. Alon and Györi [1] have shown that the number of crossings below the $(n-c-1)$-level is at most $(n-c-1)n$, if $(n-c-1) < \frac{n}{2}$. Indeed $n-c-1 = n-(1-\alpha+\epsilon)n-1 = \alpha n - \epsilon n - 1 < \frac{n}{2}$ as $\alpha < \frac{1}{2}$, thus by symmetry we get that in total at most $2(n-c-1)n$ pairs become invalid. So the number of pairs that are still valid after the last crossing is at least

$$\beta(n) := \sum_{k=2}^{c}(n+1-k) - 2(n-c-1)n = -\frac{c^2}{2} + 3cn + \frac{c}{2} - 2n^2 + n \ .$$

If $c$ is linear in $n$, that is, $c = \gamma n$, this is a quadratic polynomial with leading term $-\frac{\gamma^2}{2}+3\gamma-2$. This leading term is larger than 0 in the interval $(0,1)$ if and only if $\gamma > 3-\sqrt{5} = 1-\alpha$. Hence for $\gamma = 1-\alpha+\epsilon$, we have that $\lim_{n\to\infty}(\beta(n)) = \infty$, so for $n$ large enough, there is at least one pair that is still valid after the last crossing. ◀

On the other hand, we may have $\mathsf{outstripe}(\{p_1, p_2\}) \leq (1-\epsilon)n$ for any two points $p_1, p_2$.

▶ **Theorem 3.3.** *For any $n \geq 12$ there exist point sets $P$ of $n$ points in general position such that for every pair of points $p_1, p_2$ in $P$ there is a stripe with $p_1, p_2$ outside but at least $\lfloor \frac{n}{3} \rfloor$ points of $P$ inside.*

**Proof Sketch.** See Figure 1 for an accompanying illustration. Let $m$ be the smallest integer that is divisible by 3 such that $m \geq n \geq 12$. Let $a$, $b$ and $c$ be three points that span an equilateral triangle with side length $\frac{1}{18}m^2$. Let $a_1$ and $a_2$ be the two points on the line through $a$ and $b$ that have distance $\frac{1}{6}m$ from $a$, with $a_1$ being closer to $b$. Similarly, let $b_1$ and $b_2$ be the two points on the line through $b$ and $c$ that have distance $\frac{1}{6}m$ from $b$, with $b_1$ being closer to $c$ and let $c_1$ and $c_2$ be the two points on the line through $c$ and $a$ that have distance $\frac{1}{6}m$ from $c$, with $c_1$ being closer to $a$. Place $n$ points as follows: Place $\frac{m}{3}$ points on the line segment between $a_1$ and $a_2$ such that the first point has distance $\frac{1}{2}$ to $a_1$ and any two consecutive points have distance 1 and call this point set $A$. Do the same for the line segments between $b_1$ and $b_2$ and between $c_1$ and $c_2$ to get point sets $B$ and $C$, respectively. If necessary, take away 1 point from $B$ and possibly another one from $C$. Finally, wiggle the point set slightly so that it is in general position.

It can be shown that if we project the segment $c_1c_2$ onto the line through $a$ and $b$ such that the image lies entirely in the segment $a_1a_2$, then the length of this image is smaller

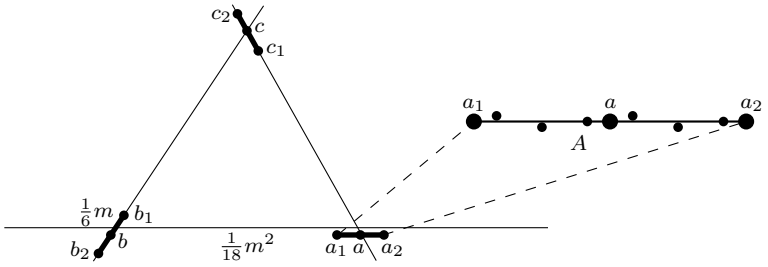


**Figure 1** Construction for Theorem 3.3.

than 1. By symmetry, the lengths of the projections of $a_1a_2$ onto $b_1b_2$ and of $b_1b_2$ onto $c_1c_2$ are also smaller than 1. If $p_1$ and $p_2$ are in the same set, w.l.o.g. $A$, it thus follows that there is a stripe with $p_1, p_2$ outside and $C$ inside. On the other hand, if $p_1$ and $p_2$ are in different sets, w.l.o.g. $A$ and $B$, it can be easily argued that there is also such a stripe. ◀

## 4 Conclusion

We defined a combinatorial distance measure on point sets and showed that there are always points which are sufficiently close. The approach gives rise to several open problems.

The distances are equivalent for all point sets with the same allowable sequence. However, there can be two point sets with the same order type (i.e., there is a bijection between them such that the corresponding triples are oriented in the same way) for which the distance is different for two corresponding pairs. Reasonable generalizations of stripes for this setting could be double wedges. Can we get analogous bounds there?

Our result gives an upper bound on the distance between closest points when the stripe is required to be orthogonal to the line defined by the points. (This corresponds to a region-counting distance function with a stripe orthogonal and between its two reference points.) Is there a linear lower bound in that setting?

While our bounds are asymptotically tight, the gap between the constant factors are large. A natural open problem is to close these gaps.

## References

1 Noga Alon and E Györi. The number of small semispaces of a finite set of points in the plane. *Journal of Combinatorial Theory, Series A*, 41(1):154 – 157, 1986.

2 Greg Aloupis. Geometric measures of data depth. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, volume 72 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 147–158, 2003.

3 Erik D. Demaine, John Iacono, and Stefan Langerman. Proximate point searching. *Comput. Geom.*, 28(1):29–40, 2004.

4 Jonathan Derryberry, Don Sheehy, Maverick Woo, and Danny Dominic Sleator. Achieving spatial adaptivity while finding approximate nearest neighbors. In *Proc. 20th Canadian Conference on Computational Geometry (CCCG 2008)*, 2008.

5 Jerome H. Friedman and John W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Computers*, 23(9):881–890, 1974.

6 John Iacono and Stefan Langerman. Proximate planar point location. In *Proc. 19th Symposium on Computational Geometry (SoCG 2003)*, pages 220–226. ACM, 2003.

7 J. W. Tukey. Mathematics and the picturing of data. In *Proc. International Congress of Mathematicians*, pages 523–531, 1975.