,

CAN WE SPOT A FAKE?

SHAHAR MENDELSON, GRIGORIS PAOURIS, AND ROMAN VERSHYNIN

ABSTRACT. The problem of detecting fake data inspires the following seemingly simple mathematical question. Sample a data point X from the standard normal distribution in $\mathbb{R}^n$. An adversary observes X and corrupts it by adding a vector rt, where they can choose any vector t from a fixed set T of adversary's "tricks", and where $r > 0$ is a fixed radius. The adversary's choice of $t = t(X)$ may depend on the true data X. The adversary wants to hide the corruption by making the fake data $X + rt$ statistically indistinguishable from the real data X. What is the largest radius $r = r(T)$ for which the adversary can create an undetectable fake? We show that for highly symmetric sets T, the detectability radius $r(T)$ approximately equals twice the scaled Gaussian width of T. The upper bound actually holds for arbitrary sets T and generalizes to arbitrary, non-Gaussian distributions of real data X. The lower bound may fail for not highly symmetric T, but we conjecture that this problem can be solved by considering the focused version of the Gaussian width of T, which focuses on the most important directions of T.

CONTENTS

1. INTRODUCTION

The ongoing AI boom is revolutionizing human-computer interaction. But not all interaction is good: generative AI has made it easy to create fake data.

> *When can we detect a fake?*

This general question can be nontrivial even in stunningly simple scenarios. Imagine we sample a data point X from the standard normal distribution on $\mathbb{R}^n$. An adversary observes X and corrupts it by adding a vector rt, where they can choose any vector $t = t(X) \in T$ from a fixed set $T \subset \mathbb{R}^n$

of adversary's "tricks", and where $r > 0$ is a fixed radius. Allowing the adversary's choice of $t = t(X)$ to depend on X allows the adversary may to manipulate different data in different ways. The adversary wants to hide the corruption by making the fake data $X + rt$ statistically indistinguishable from the real data X. When can the adversary succeed?

Question 1.1 (Detectability radius). *For a given set $T \subset \mathbb{R}^n$, what is the smallest radius $r(T)$ such that for any $r > r(T)$ we can always detect a fake with high probability?*

If r is very small, the fake can never be detected, since the adversary can always succeed: a tiny shift of the normal distribution is close to the original normal distribution. If r is very large, the fake can always be detected: a hugely shifted vector has an abnormally large length, and this can be easily detected. So where is the phase transition $r(T)$? And what are the optimal strategies for the tester and for the adversary?

1.1. **A geometric viewpoint.** Any test can be encoded by a subset $A \subset \mathbb{R}^n$: for any point in A the tester says "true", and for any point in the complement of A the tester says "fake". A good test recognizes true data as true, and does *not* recognize fake data as true, both with a high probability. Thus, the tester's goal is to find a set A that minimizes the probability of error, so that

$$\mathbb{P}\{X \in A\} \geq 0.9 \quad \text{and} \quad \mathbb{P}\{\exists t \in T : \ X + rt \in A\} \leq 0.1. \qquad (1.1)$$

We have chosen the values 0.9 and 0.1 for illustrative purposes only; they can be arbitrary in general.

Using Minkowski addition, the second inequality in (1.1) can be expressed as $\mathbb{P}\{X \in A - rT\} \leq 0.1$. Thus, denoting by γ_n the standard Gaussian measure on $\mathbb{R}^n$, the tester's goal is to find a subset $A \subset \mathbb{R}^n$ satistying

$$\gamma_n(A) \geq 0.9 \quad \text{and} \quad \gamma_n(A - rT) \leq 0.1. \qquad (1.2)$$

The *detectablity radius* we introduced in Question 1.1 can be formally defined as the largest radius of an undetectable fake:

$$r(T) \coloneqq \sup \{r > 0 : \text{ there exists no set } A \subset \mathbb{R}^n \text{ that satisfies } (1.2)\} .$$

1.2. **Related work.** A lot of effort has been made to compute the detectability radius for an *outsider* adversary whose choice of the vector $t \in T$ that is *not* allowed to depend on X. Instead of (1.2), in the "outsider adversary" scenario the tester is satisfied with the following weaker goal: find a subset $A \subset \mathbb{R}^n$ that satisfies

$$\gamma_n(A) \geq 0.9 \quad \text{and} \quad \sup_{t \in T} \gamma_n(A - rt) \leq 0.1. \qquad (1.3)$$

The study of this "outsider adversary" problem can be traced back to Tuckey's idea of *"higher criticism"* [22], see [11, 12, 13]. Higher criticism was recently revisited from the standpoint of high-dimensional statistics, and was studied for various specific sets T including the spheres in the ℓ^p

metric [15], the indicators of paths and more general clusters in a given graph [2, 3], the set of unit sparse vectors [7, 16, 11, 12, 13, 9, 14, 17], and more generally a set containing all unit vectors with given sparsity patterns [1]. Extensions for sparse regression has been studied in [4, 18, 10, 19]. The only work on the "insider adversary" that we know of is the concurrent work [20], which studies the detectability radius of the discrete cube $T = \{-1, 1\}^n$ and its sparsified versions.

1.3. **Main results.** Returning to our "insider adversary" problem, it turns out that in many situations, the detectability radius $r(T)$ is captured by the quantity that we call the *scaled Gaussian width* of T. It is defined as follows:

$$\bar{w}(T) = \mathbb{E} \sup_{t \in T} \left\langle X, \frac{t}{\|t\|_2^2} \right\rangle \tag{1.4}$$

where X is a standard normal random vector in $\mathbb{R}^n$. This is a scaled version of the more traditional quantity called *Gaussian width*, which is

$$w(T) = \mathbb{E} \sup_{t \in T} \langle X, t \rangle. \tag{1.5}$$

The Gaussian width describes the complexity of the set T. This concept plays an important role in high-dimensional probability, asymptotic convex geometry and high-dimensional inference [23, 5, 6].

The unorthodox scaling in (1.4) is justified by the contribution of the points in T near the origin. Since smaller distortions are easier to hide, having such points in T makes the adversary's job easier and the tester's job harder. So a quantity that captures the detectability radius $r(T)$ should be more sensitive to points in T that are close to the origin than to those that are far from the origin. Such a sensitivity is achieved in (1.4) by dividing by $\|t\|_2^2$.

The following theorem summarizes our main results.

Theorem 1.2 (Detectability radius, informal)**.** *For any set $T \subset \mathbb{R}^n$, we have*

$$r(T) \leq 2\bar{w}(T)\,(1 + o(1)). \tag{1.6}$$

Moreover, for any highly symmetric set T we have

$$r(T) \geq 2\bar{w}(T)\,(1 - o(1)). \tag{1.7}$$

The first part of Theorem 1.2 identifies the regime in which the fake is detectable. A formal, non-asymptotic version of (1.6) is provided in Theorem 2.1. We argue in Section 2.1 that a simple *proximity test* can detect a fake in this regime: if the point is closer to the origin than to any point in T, return "real"; otherwise return "fake".

The second part of Theorem 1.2 identifies the regime in which there exist undetectable fakes. The meaning of "highly symmetric" is explained in Definition 3.1, and a formal, non-asymptotic version of (1.6) is provided in Theorem 3.2. We argue in Section 3.2 that an undetectable fake can be produced in this regime by a *sign flipping strategy*: choose a point $t(X) \in T$

such that adding rt to the outcome X is equivalent to reversing the signs of some coordinates of X.

1.4. **Example: the adversary adds any vector.** To illustrate the intuition behind Theorem 1.2, consider the most basic example where the adversary is allowed to add to X any vector of length at least r. In other words, the set of the adversary's tricks is

$$T = T_n = \{t \in \mathbb{R}^n : \|t\|_2 \geq 1\}. \tag{1.8}$$

The scaled Gaussian width of T equals the Gaussian width of the unit Euclidean ball in $\mathbb{R}^n$, and it is approximately $\sqrt{n}$, see [23, Example 7.5.7]. Thus Theorem 1.2 gives

$$r(T_n) = 2\sqrt{n}(1 + o(1)).$$

There is a simpler way to obtain the same conclusion. The Euclidean norm of the standard normal random vector X is approximately $\sqrt{n}$ [23, Example 7.5.3]. Therefore, if the radius r is significantly larger than $2\sqrt{n}$, adding to X a vector rt for any $t \in T$ would make the norm of the sum significantly larger than $\sqrt{n}$. So the norm of the fake data $X + rt$ must be significantly larger than the typical norm of the real data X. This abnormality can be easily detected.

On the other hand, let us take $r = 2\sqrt{n}$ and simplify reality a little by pretending that the Euclidean norm of X is exactly $\sqrt{n}$. Then the adversary can make an undetectable fake by adding the vector $-2X \in rT$ to X. Indeed, this effectively flips the vector X about the origin, which does not change the distribution of X at all.

1.5. **Example: the adversary adds a sparse vector.** To make the previous example more interesting, imagine that the adversary has limited resources and can change at most s coordinates of X. The choice of which coordinates to change is up to the adversary and may depend on X. In other words, the adversary's set of tricks is

$$T = T_{n,s} = \{t \in \mathbb{R}^n : \|t\|_2 \geq 1, \ \|t\|_0 \leq s\}. \tag{1.9}$$

Here $\|t\|_0$ denotes the *sparsity* of t, which equals the number of nonzero coordinates of t. The set $T_{n,s}$ is a highly symmetric, and its scaled Gaussian width satisfies

$$\bar{w}(T_{n,s}) \asymp \sqrt{s \ln(en/s)},$$

see [23, Section 10.3.3]. Here the sign "$\asymp$" hides absolute constant factors. Thus, Theorem 1.2 gives

$$r(T_{n,s}) \asymp \sqrt{s \ln(en/s)}.$$

The proximity test in this case reduces to checking whether the sum of squares of the largest s coefficients is abnormally large. The sign flipping strategy reduces to reversing the signs of the largest s coefficients.

1.6. **What about non-Gaussian data?** Real world data is rarely Gaussian. The first part of Theorem 1.2, which identifies a regime in which the fake is detectable, can be easily generalized to non-Gaussian random vectors X. A version of this result, Theorem 4.1, holds for a completely arbitrary distribution of X.

As for the second part of Theorem 1.2, which identifies a regime in which there exist undetectable fakes, extending it to general distributions is trickier. Our proof of this result is based on sign flipping strategy, whose validity explores the symmetries of the distribution of X. A more general version this result, Theorem 4.2, holds for any random vector X whose coordinates X_i are independent, symmetric, and bounded random variables. It remains an interesting question to what extent we can relax the assumptions of independence, symmetry and boundedness.

1.7. **What about general sets T?** While the first part of Theorem 1.2 is valid for general sets T, the second part requires T to be highly symmetric, where the rigorous meaning of "highly symmetric" is given in Definition 3.1.

This latter result can fail miserably for general sets T: the scaled Gaussian width can hugely overestimate the detectability radius. A simple example is where T consists of all unit vectors in $\mathbb{R}^n$ whose first coordinate is either $1/2$ or $-1/2$. An elementary calculation carried out in Section 5.1 shows that $\bar{w}(T) \asymp \sqrt{n}$ while $r(T) \asymp O(1)$.

In full generality, we suspect that the detectability radius $r(T)$ must be captured by some geometric quantity $\widetilde{w}(T)$ that is generally smaller than the scaled Gaussian width $\bar{w}(T)$, and which is focused only on the most important directions of T. We can call such quantity the *focused Gaussian width*. In Section 5 we define the focused Gaussian width and show that

$$r(T) \lesssim \widetilde{w}(T) \leq \bar{w}(T).$$

A fascinating problem remains whether the first inequality can be reversed for a general set T, i.e. whether the detectability radius is always equivalent to the focused Gaussian width.

2. When is a fake detectable?

Let us begin by showing that a fake is detectable whenever the radius r is significantly larger than $2\bar{w}(T)$, where $\bar{w}(T)$ is the scaled Gaussian width defined in (1.4). The error term that quantifies "significantly larger" depends on the magnitude of the smallest change the adversary can make. This magnitude is the *inradius* of T, defined as

$$\rho(T) = \inf_{t \in T} \|t\|_2. \tag{2.1}$$

Theorem 2.1 (When the fake is detectable)**.** *Let X be a standard normal random vector in $\mathbb{R}^n$. Let $T \subset \mathbb{R}^n$ be any set and $u > 0$ be any number. Assume that*

$$r \geq 2\bar{w}(T) + \frac{u}{\rho(T)}. \tag{2.2}$$

Then there exists a set $A \subset \mathbb{R}^n$ satisfying

$$\mathbb{P}\{X \in A\} > 1 - e^{-u^2/8} \quad and \quad \mathbb{P}\{X \in A - rT\} < e^{-u^2/8}.$$

Proof. Consider the set

$$K := \left\{ x \in \mathbb{R}^n \mid \langle x, t \rangle < \|t\|_2^2 \text{ for every } t \in T \right\}.$$

We claim that

$$\frac{K}{2} \cap \left(T - \frac{K}{2} \right) = \emptyset. \tag{2.3}$$

Indeed, if this were not true, there would exist vectors $x, y \in K$ and $t \in T$ satisfying $x/2 = t - y/2$, or $x + y = 2t$. Multiplying by t on both sides would give

$$\langle x, t \rangle + \langle y, t \rangle = 2\langle t, t \rangle = 2\|t\|_2^2. \tag{2.4}$$

On the other hand, the definition of K implies $\langle x, t \rangle < \|t\|_2^2$ and $\langle y, t \rangle < \|t\|_2^2$. Adding these two inequalities creates a contradiction with (2.4), and so Claim (2.3) is proved.

We have

$$
\begin{aligned}
\mathbb{P}\left\{ X \in \frac{r}{2}K \right\} &= \mathbb{P}\left\{ \sup_{t \in T} \left\langle X, \frac{t}{\|t\|_2^2} \right\rangle < \frac{r}{2} \right\} \quad \text{(by definition of } K) \\
&\geq \mathbb{P}\left\{ \sup_{t \in T} \left\langle X, \frac{t}{\|t\|_2^2} \right\rangle < \bar{w}(T) + \frac{u}{2\rho(T)} \right\} \quad \text{(by assumption on } r) \\
&> 1 - e^{-u^2/8},
\end{aligned}
$$

where the last step follows by applying the Gaussian concentration inequality (see [8]) for the $1/\rho(T)$-Lipschitz function $f(x) = \sup_{t \in T} \langle x, t/\|t\|_2^2 \rangle$.

Since (2.3) can be rewritten as

$$\frac{r}{2}K \cap \left(rT - \frac{r}{2}K \right) = \emptyset,$$

the previous bound yields

$$\mathbb{P}\left\{ X \in rT - \frac{r}{2}K \right\} \leq 1 - \mathbb{P}\left\{ X \in \frac{r}{2}K \right\} < e^{-u^2/8}.$$

Therefore, the conclusion of the theorem holds for

$$A = \frac{r}{2}K. \quad \square$$

Remark 2.2 (The error term is small). When we look at the assumption (2.2), it is helpful to regard $2\bar{w}(T)$ as the main term and $u/\rho(T)$ as the error term. The error term is typically smaller than the main term, and often much smaller. Indeed, the Gaussian width of any origin-symmetric set T satisfies $w(T) \geq \sqrt{2/\pi} \sup_{t \in T} \|t\|_2$, see [23, Proposition 7.5.2]. Rescaling and rearranging the terms, we conclude that the scaled Gaussian width of any origin-symmetric set T satisfies

$$\frac{1}{\rho(T)} \leq \sqrt{\frac{\pi}{2}} \, \bar{w}(T).$$

Moreover, in most interesting cases $1/\rho(T)$ is much smaller than $\bar{w}(T)$. For example, if T is the unit Euclidean sphere in $\mathbb{R}^n$, then we have

$$\rho(T) = 1 \quad \text{while} \quad \bar{w}(T) \approx \sqrt{n}.$$

Similarly, if $T_{n,s}$ is the set of s-sparse vectors in $\mathbb{R}^n$ of norm at least 1, which we introduced in Section 1.5, then we by [23, Section 10.3.3] we have

$$\rho(T) = 1 \quad \text{while} \quad \bar{w}(T_{n,s}) \asymp \sqrt{s \ln(en/s)}.$$

2.1. **How to spot a fake? A proximity test.** Theorem 2.1 identifies a regime where a fake can be detected: there is a test that can accurately predict whether a data x has been sampled from the standard normal distribution, or alternatively a point from rT has been added to it. What is this test?

This test is encoded by a subset $A \subset \mathbb{R}^n$: upon seeing a data $x \in \mathbb{R}^n$, the tester returns "real" if $x \in A$ and "fake" if $x \in A^c$. In the proof of Theorem 2.1, we made the following choice of A:

$$A = \frac{r}{2}K = \left\{ x \in \mathbb{R}^n \mid \langle x, t \rangle < \frac{r}{2}\|t\|_2^2 \text{ for every } t \in T \right\}.$$

Rewriting the inequality above as $\langle x, rt \rangle < \frac{1}{2}\|rt\|_2^2$, we see that it is simply saying that the point x is closer to the origin than to the point rt. Thus, the proof of Theorem 2.1 yields the following

> **Proximity test.** *If the data point x is closer to the origin than to any point in the set rT, return "true"; otherwise return "fake".*

If the radius r satisfies (2.2), the proximity test succeeds: the probability of a false positive error and the probability of a false negative error are both bounded by $e^{-u^2/8}$.

3. When is a fake undetectable?

Next, let us consider the opposite situation. We will show how the adversary can create an undetectable fake whenever the radius r is significantly larger than $2\bar{w}(T)$, where $\bar{w}(T)$ is the scaled Gaussian width defined in (1.4). The error term that quantifies "significantly larger" will be exactly the same as in Theorem 2.1 and it will depend on the inradius $\rho(T)$ of T, defined in (2.1).

We will only consider highly symmetric sets of tricks T in this regime, and we will demonstrate in Section 5.1 how the result can fail if T is not highly symmetric.

Definition 3.1 (Highly symmetric set). *We say that a set $T \subset \mathbb{R}^n$ is* highly symmetric *if, whenever T contains a point x, the set T must also contain any point $y \in \mathbb{R}^n$ that satisfies $\mathrm{supp}(y) = \mathrm{supp}(x)$ and $\|y\|_2 \geq \|x\|_2$.*

An example of a highly symmetric set is the set T_n of all vectors in $\mathbb{R}^n$ whose norm is bounded below by 1, which we considered in (1.8). A more general example is the set of all s-sparse vectors in $\mathbb{R}^n$ whose Euclidean norm is at least 1, which we considered in (1.9).

Theorem 3.2 (Where the fake is undetectable). *Let X be a standard normal random vector in $\mathbb{R}^n$. Let $T \subset \mathbb{R}^n$ be a highly symmetric set and $u > 0$ be any number. Assume that*

$$0 \le r \le 2\bar{w}(T) - \frac{u}{\rho(T)}. \tag{3.1}$$

Then for any set $A \subset \mathbb{R}^n$ we have

$$\mathbb{P}\{X \in A - rT\} \ge \mathbb{P}\{X \in A\} - e^{-u^2/8}.$$

Proof of Theorem 3.2. Definition 3.1 of a highly symmetric set T implies that the answer to the question "does a given point $x \in \mathbb{R}^n$ belong to T?" depends only on the support of x and whether the norm of x is sufficiently large. Let S denote the family of subsets of $\{1, \ldots, n\}$ that are supports of points in T, that is

$$S := \{\operatorname{supp}(x) : \ x \in T\}.$$

For every set $I \in S$, let $\nu(I)$ denote the smallest Euclidean norm of a vector in T whose support equals I:

$$\nu(I) := \min\{\|x\|_2 : \ x \in T, \ \operatorname{supp}(x) = I\}.$$

Then we can express the set T as follows:

$$T := \{x \in \mathbb{R}^n : \ \exists I \in S \text{ such that } \operatorname{supp}(x) = I, \ \|x\|_2 \ge \nu(I)\}. \tag{3.2}$$

A simple computation shows that the Gaussian width and the inradius of T are

$$\bar{w}(T) = \mathbb{E} \max_{I \in S} \frac{\|X_I\|_2}{\nu(I)}, \quad \rho(T) = \min_{I \in S} \nu(I),$$

where $x_I \in \mathbb{R}^I$ denotes the restriction of a vector $x \in \mathbb{R}^n$ onto the coordinates in I.

Apply the Gaussian concentration inequality (see [8]) for the $1/\rho(T)$-Lipschitz function $f(x) = \max_{I \in S} \|x_I\|_2/\nu(I)$. We obtain that with probability at least $1 - e^{-u^2/8}$, the following holds:

$$\max_{I \in S} \frac{\|X_I\|_2}{\nu(I)} \ge \bar{w}(T) - \frac{u}{2\rho(T)} \ge \frac{r}{2}, \tag{3.3}$$

where the second inequality is due to the assumption on r.

Whenever the event in (3.3) occurs, there exists a set $I = I(X) \in S$ such that

$$\|2X_I\|_2 \ge r\nu(I).$$

By definition of T, this implies that $-2X_I \in rT$, so there exists $t = t(X) \in T$ such that $-2X_I = rt$. Whenever the event (3.3) does not hold, set $I(X) = I_0$ and $t(X) = t_0$ for some arbitrary but fixed $I_0 \in S$ and $t_0 \in T$. Summarizing,

we have constructed a random set $I = I(X) \in S$ and a random point $t = t(X) \in T$ such that

$$\mathbb{P}\{-2X_I = rt\} \geq 1 - e^{-u^2/8}. \qquad (3.4)$$

Moreover, since the quantity $\|X_I\|_2$ in the event (3.3) is determined by the *absolute values* of the coefficients of X, we can arrange that the set I is determined by the absolute values of the coefficients of X.

Adding the vector $-2X_I$ to the vector X is equivalent to reversing the signs of the coefficients of X indexed by I. And since the choice of I is determined by the absolute values of the coefficients of X and is independent of their signs, reversing the coordinates of X indexed by I does not change the distribution of X. (This is a consequence of the symmetry of the Gaussian distribution, which we explain in detail in Lemma 3.3 below.) Hence $X - 2X_I$ has the same distribution as X. Therefore, for any set $A \subset \mathbb{R}^n$ we have

$$\begin{aligned}
\mathbb{P}\{X \in A\} &= \mathbb{P}\{X - 2X_I \in A\} \\
&\leq \mathbb{P}\{X - 2X_I \in A \text{ and } -2X_I = rt\} + \mathbb{P}\{-2X_I \neq rt\} \\
&\leq \mathbb{P}\{X + rt \in A\} + e^{-u^2/8},
\end{aligned}$$

where the in last step we used (3.4). Since $t \in T$, rearranging the terms completes the proof. $\qquad \square$

Let us now explain the crucial fact used in the proof above, namely that sign flipping does not change the distribution.

Lemma 3.3 (Sign flipping). *Let $X = (X_1, \ldots, X_n)$ be a random vector in $\mathbb{R}^n$ whose coordinates X_i are independent and have symmetric distributions.[1] Let $\theta = (\theta_1, \ldots, \theta_n) \in \{-1, 1\}^n$ be a random vector whose value is determined by the absolute values of the coefficients of X. Then the random vector $(\theta_1 X_1, \ldots, \theta_n X_n)$ has the same distribution as X.*

Proof. The random vector of interest has coefficients

$$\theta_i X_i = \theta_i \operatorname{sgn}(X_i)|X_i|.$$

Condition on the absolute values of the coefficients of X. This fixes the values of $|X_i|$ and θ_i, while by symmetry $\operatorname{sgn}(X_i)$ are (conditionally) independent Rademacher random variables. Since the signs θ_i are now fixed, the symmetry of Rademacher distribution implies that

$$\theta_i \operatorname{sgn}(X_i) \equiv \operatorname{sgn}(X_i),$$

where we use the sign "$\equiv$" to indicate the equality of (conditional) joint distributions of the coefficients. Multiplying by the fixed numbers $|X_i|$, we get

$$\theta_i \operatorname{sgn}(X_i)|X_i| \equiv \operatorname{sgn}(X_i)|X_i|.$$

---

[1]A random variable ξ has symmetric distribution if ξ has the same distribution as $-\xi$.

In other words,

$$\theta_i X_i \equiv X_i.$$

Since the conditional distributions are equal almost surely (in fact, deterministically), the original distributions are equal, too. □

3.1. **Undetectability.** Let us explain why the conclusion of Theorem 3.2 can be interpreted as existence of undetectable fakes.

Any fake detection test can be encoded by a subset $A \subset \mathbb{R}^n$: upon seeing a data $x \in \mathbb{R}^n$, the tester returns "real" if $x \in A$ and "fake" if $x \in A^c$. Rewriting the conclusion of Theorem 3.2 as

$$\mathbb{P}\{X \in A^c\} + \mathbb{P}\{X \in A - rT\} \geq 1 - e^{-u^2/8},$$

we see that the two probabilities on the left hand side can not be simultaneously small: we must have

$$\min \left(\mathbb{P}\{X \in A^c\}, \mathbb{P}\{X \in A - rT\} \right) \geq \frac{1}{2} - \frac{1}{2} e^{-u^2/8}.$$

This means that no test can reliably detect a fake: either the false positive rate or the false negative rate must be close to 50% or higher.

3.2. **How to evade detection? A sign flipping strategy.** Theorem 3.2 tells us that for any radius r below a certain threshold, the adversary has a strategy to create an undetectable. What is this strategy?

To create a fake, the adversary looks at the true data x, picks a point $t = t(x) \in T$, and outputs $x + rt$. The proof of Theorem 3.2 contains a recipe for choosing t. We express the set T via (3.2), find a set of indices

$$I = \underset{I \in S}{\mathrm{argmax}} \, \frac{\|x_I\|_2}{\nu(I)}$$

and choose $t \in T$ such that $rt = -2x_I$. (And if such t does not exist, the adversary gives up.) Adding such rt to x is equivalent to reversing the signs of the coefficients of x indexed by I. Thus, the proof of Theorem 3.2 yields the following strategy for the adversary:

> **Sign flipping strategy.** *Given a data point x and a set T as in (3.2), choose a set of indices I that maximizes the ratio $\|x_I\|_2/\nu(I)$ and reverse the signs of the coefficients of x indexed by I. Do this only if such a reversal can be realized as adding some point from rT to x. Otherwise give up.*

If the radius r satisfies (3.1), then the sign flipping strategy succeeds with high probability. It creates fake data that is statistically indistinguishable from the true data. As we pointed out in Section 3.1, for any test either the false positive rate or the false negative rate must be close to 50% or higher.

4. Extensions for non-Gaussian data

Since real world data is rarely Gaussian, it is natural to wonder whether our results generalize to random vectors X with general distributions.

Particularly straightforward is the generalization of Theorem 2.1, which identifies the regime in which fakes are detectable. Consider straightforward generalizations of the concepts of Gaussian width (1.5) and scaled Gaussian width (1.4), in which the standard normal random vector X is replaced with a given arbitrary random vector X taking values in $\mathbb{R}^n$. This leads to the concept of *X-width* and *scaled X-width*, respectively:

$$w_X(T) = \mathbb{E} \sup_{t \in T} \langle X, t \rangle \quad \text{and} \quad \bar{w}_X(T) = \mathbb{E} \sup_{t \in T} \left\langle X, \frac{t}{\|t\|_2^2} \right\rangle.$$

Then Theorem 2.1 generalizes as follows.

Theorem 4.1 (Where the fake is detectable: general distributions). *Let X be any random vector taking values in $\mathbb{R}^n$. Let $T \subset \mathbb{R}^n$ be any origin-symmetric set[2] and $u > 0$ be any number. Assume that*

$$r > 2u \cdot \bar{w}_X(T).$$

Then there exists a set $A \subset \mathbb{R}^n$ such that

$$\mathbb{P}\{X \in A\} > 1 - 1/u \quad \text{and} \quad \mathbb{P}\{g \in A - rT\} < 1/u.$$

Proof. Follow the proof of Theorem 2.1 but use Markov's inequality instead of the Gaussian concentration inequality. We assumed that T is origin-symmetric to make sure that the random variable $\sup_{t \in T} \langle X, t/\|t\|_2^2 \rangle$ takes only non-negative values, which makes Markov's inequality applicable. □

As for Theorem 2.1, which identifies the regime in which fakes are detectable, it is less clear to which non-Gaussian distributions it can be generalized. The proof of Theorem 2.1 uses a sign-flipping strategy that relies crucially on certain symmetries of the Gaussian distribution. It can be extended to any distribution that is sufficiently symmetric:

Theorem 4.2 (Where the fake is undetectable: non-Gaussian data). *Let $X = (X_1, \ldots, X_n)$ be a random vector whose coordinates X_i are independent, symmetric random variables taking values in the interval $[-1, 1]$. Let $T \subset \mathbb{R}^n$ be a highly symmetric set and $u > 0$ be any number. Assume that*

$$r \le 2\bar{w}(T) - \frac{u}{\rho(T)}.$$

Then for any set $A \subset \mathbb{R}^n$ we have

$$\mathbb{P}\{X \in A - rT\} \ge \mathbb{P}\{X \in A\} - 2e^{-u^2/16}.$$

---

[2]The assumption that T is origin symmetric, i.e. that $T = -T$, is a bit annoying, but it is unavoidable in this statement, since a set T consisting of a single point must satisfy $\bar{w}(T) = 0$. However, it is easy to get around this assumption by replacing any set T with the origin-symmetric set $T \cup -T$.

Proof. Follow the proof of Theorem 3.2 but use Talagrand's convex concentration inequality [21, Theorem 6.6] instead of the Gaussian concentration inequality. □

5. Is the detectability radius equivalent to the focused width?

5.1. **The upper bound is not always sharp.** Theorem 2.1 says that a fake can be detected if the radius r is somewhat larger than $2\bar{w}(T)$. Theorem 3.2 demonstrates that this result is optimal for for highly symmetric sets T.

If a set T does not have enough symmetries, Theorem 3.2 can fail. Consider, for example, the set of all unit vectors whose first coordinate equals $\pm 1/2$, that is

$$T = \left\{ t \in \mathbb{R}^n : \|t\|_2 = 1, \ |t_1| = \frac{1}{2} \right\}. \tag{5.1}$$

A simple calculation yields

$$\bar{w}(T) = \bar{w}(T) \asymp \sqrt{n},$$

where the "$\asymp$" sign hides absolute constant factors. Thus, Theorem 2.1 says that fake detection is possible whenever the radius satisfies $r \gtrsim \sqrt{n}$.

However, this result is too conservative. The detection in this example is possible even if the radius is an *absolute constant*. For example, if $r = 100$, the first coordinate of the fake vector $X + rt$ is

$$(X + rt)_1 = X_1 \pm 50.$$

Comparing this to the first coordinate of the real vector X, which is $X_1 \sim N(0,1)$, we see that the first coordinate of any fake vector must have a huge non-zero mean. This can be easily tested. Hence the scaled Gaussian width vastly overestimates the detection radius:

$$r(T) = O(1) \quad \text{while} \quad \bar{w}(T) \asymp \sqrt{n},$$

This violates Theorem 3.2.

5.2. **Focused width.** The example above makes us wonder what causes the scaled Gaussian to overestimate the detection radius, and how to fix this problem. The definition of the Gaussian width takes into account all directions $t \in T$, while not all directions are equally useful to an adversary: some tricks $t \in T$ may be more "revealing" and thus less useful. We wonder if forcing the Gaussian width to *focus* on the most important directions might solve the attention problem.

This motivates the following definition of the *focused Gaussian width*. To make it more broadly applicable, let us define it not just for Gaussian X but for general distributions.

Definition 5.1 (Focused Gaussian width). *The* focused Gaussian width *of T is defined as*

$$\widetilde{w}(T) = \inf_S w(S)$$

where the infimum is over all origin-symmetric sets $S \subset \mathbb{R}^n$ satisfying

$$\forall t \in T \,\exists s \in S : \,\langle t, s \rangle \geq 1. \tag{5.2}$$

In other words, $\widetilde{w}(T)$ is the smallest Gaussian width of an origin-symmetric set whose polar is disjoint from T.

More generally, if X is an arbitrary random vector taking values in $\mathbb{R}^n$, the *focused X-width* is defined as

$$\widetilde{w}_X(T) = \inf_S w_X(S).$$

The set S in this definition encodes the directions of "focus", and taking the infimum over S encourages the width to focus on the most important directions.

To compare with the scaled width, it is convenient to rewrite the definition of the focused X-width equivalently as follows:

$$\widetilde{w}_X(T) = \inf_H \bar{w}_X(H) \tag{5.3}$$

where the infimum is over all measurable sets $H \subset \mathbb{R}^n$ satisfying

$$\forall t \in T \,\exists h \in H : \,\langle t, h \rangle \geq \|h\|_2^2. \tag{5.4}$$

To check the equivalence, choose $h = s/\|s\|_2^2$, or $s = h/\|h\|_2^2$.

Choosing $H = T$ and $h = t$ in (5.4), we immediately obtain:

Lemma 5.2 (Focusing can only reduce the width)**.** *For any random vector X taking values in $\mathbb{R}^n$ and for any set $T \subset \mathbb{R}^n$, we have*

$$\widetilde{w}_X(T) \leq \bar{w}_X(T).$$

5.3. **Does focus help?** The next result states that Theorem 2.1 can be strengthened by replacing the width with the focused width.

Theorem 5.3 (Where the fake is detectable: focused width)**.** *Let X be any random vector taking values in $\mathbb{R}^n$. Let $T \subset \mathbb{R}^n$ be any origin-symmetric set and $u > 0$ be any number. Let*

$$r > 2u \cdot \widetilde{w}_X(T).$$

Then there exists a set $A \subset \mathbb{R}^n$ such that

$$\mathbb{P}\{X \in A\} > 1 - 1/u \quad and \quad \mathbb{P}\{X \in A - rT\} < 1/u.$$

Proof. By the assumption on r and the alternative definition of the focused width (5.3), there exists a set $H \subset \mathbb{R}^n$ satisfying (5.4) and such that

$$r > 2u \cdot \bar{w}_X(H).$$

Consider the set

$$K := \left\{ x \in \mathbb{R}^n \mid \langle x, h \rangle < \|h\|_2^2 \text{ for every } h \in H \right\}.$$

We claim that

$$\frac{K}{2} \cap \left(T - \frac{K}{2} \right) = \emptyset. \tag{5.5}$$

Indeed, if this were not true, there would exist vectors $x, y \in K$ and $t \in T$ such that $x/2 = t - y/2$, or $x + y = 2t$. By (5.4), we can find a vector $h \in H$ satisfying

$$\langle t, h \rangle \geq \|h\|_2^2.$$

Multiplying by h on both sides of the identity $x + y = 2t$ would give

$$\langle x, h \rangle + \langle y, h \rangle = 2\langle t, h \rangle \geq 2\|h\|_2^2. \tag{5.6}$$

On the other hand, by definition of K we have $\langle x, h \rangle < \|h\|_2^2$ and $\langle y, h \rangle < \|h\|_2^2$. This contradicts (5.6), and so Claim (5.5) is proved.

Then

$$\mathbb{P}\left\{X \in \frac{r}{2}K\right\} = \mathbb{P}\left\{\sup_{h \in H}\left\langle X, \frac{h}{\|h\|_2^2}\right\rangle < \frac{r}{2}\right\} \quad \text{(by definition of } K\text{)}$$

$$\geq \mathbb{P}\left\{\sup_{h \in H}\left\langle X, \frac{h}{\|h\|_2^2}\right\rangle < u \cdot \bar{w}_X(H)\right\} \quad \text{(by assumption on } r\text{)}$$

$$> 1 - 1/u,$$

where the last step follows by applying Markov's inequality. Note that the assumption that T is origin-symmetric guarantees that the random variable $\sup_{t \in T}\langle X, t/\|t\|_2^2 \rangle$ takes only non-negative values, which makes Markov's inequality applicable.

Since (5.5) can be rewritten as

$$\frac{r}{2}K \cap \left(rT - \frac{r}{2}K\right) = \emptyset,$$

the previous bound yields

$$\mathbb{P}\left\{X \in rT - \frac{r}{2}K\right\} \leq 1 - \mathbb{P}\left\{X \in \frac{r}{2}K\right\} < 1/u.$$

Therefore, the conclusion of the theorem holds for

$$A = \frac{r}{2}K. \quad \square$$

5.4. **Revisiting a challenging example.** Let us revisit the example from Section 5.1, which shows that the usual Gaussian width can vastly overestimate the detection radius. There we considered the set

$$T = \left\{x \in \mathbb{R}^n : \|x\|_2 = 1, \ |x_1| = \frac{1}{2}\right\}.$$

We noticed a discrepancy: the scaled Gaussian width of T is of the order of $\sqrt{n}$, while the fake detection is possible even when the radius r is around an absolute constant.

Let us now compute the *focused* Gaussian width of T. Consider the set

$$S = \{-2e_1, 2e_1\}$$

where $e_1 = (1, 0, 0, \ldots, 0)$. The condition (5.2) is satisfied, and so

$$\widetilde{w}(T) \leq w(S) = 2\,\mathbb{E}|g| = 2\sqrt{\frac{2}{\pi}}.$$

So the focused Gaussian width of T is $O(1)$ – much smaller than the usual Gaussian width.

As opposed to the Gaussian width, the focused Gaussian width correctly estimates the detectability radius in this example. Indeed, Theorem 5.3 guarantees that the fake detection is possible if the radius is of the order of $O(1)$. In fact, we described such a test in Section 5.1: check if the magnitude of the first coordinate is abnormally large.

5.5. **A conjecture.** Inspired by the example above, one can wonder if the focused Gaussian width is always equivalent to the detectability radius, i.e.

$$r(T) \asymp \widetilde{w}(T).$$

Theorem 5.3 gives $r(T) \lesssim \widetilde{w}(T)$. It remains a question whether this inequality can always be reversed. In other words, does a version of Theorem 3.2 hold for any origin-symmetric set T if we replace the scaled Gaussian width $\bar{w}(T)$ with the focused Gaussian width $\widetilde{w}(T)$?

REFERENCES

[1] L. Addario-Berry, N. Broutin, L. Devroye, G. Lugosi, *On combinatorial testing problems,* Annals of Statistics 38 (2010), 3063–3092.

[2] E. Arias-Castro, E. Candes, H. Helgason, O. Zeitouni, *Searching for a trail of evidence in a maze,* Annals of Statistics 36 (2008), 1726–1757.

[3] E. Arias-Castro, E. Candès, A. Durand, *Detection of an anomalous cluster in a network,* Annals of Statistics 39 (2011), 278–304.

[4] E. Arias-Castro, E. Candes, Y. Plan, *Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism,* Annals of Statistics 39 (2011), 2533–2556.

[5] S. Artstein-Avidan, A. Giannopoulos, V. Milman, *Asymptotic Geometric Analysis, Part I.* Mathematical Surveys and Monographs, 2015.

[6] S. Artstein-Avidan, A. Giannopoulos, V. Milman, *Asymptotic Geometric Analysis, Part II.* American Mathematical Society, 2021.

[7] Y. Baraud, *Non-asymptotic minimax rates of testing in signal detection,* Bernoulli 8 (2002), 577–606.

[8] S. Boucheron, G. Lugosi, P. Massart, *Concentration Inequalities, A nonasymptotic theory of independence.* Clarendon press, Oxford 2012.

[9] T. Cai, J. Jin, M. Low, *Estimation and confidence sets for sparse normal mixtures,* Annals of Statistics 35 (2007), 2421–2449.

[10] A. Carpentier, O. Collier, L. Comminges, A. Tsybakov, Y. Wang, *Minimax rate of testing in sparse linear regression,* Automation and Remote Control 80 (2019), 1817–1834.

[11] D. Donoho, J. Jin, *Higher criticism for detecting sparse heterogeneous mixtures,* Annals of Statistics 32 (2004), 962–994.

[12] D. Donoho, J. Jin, *Higher criticism thresholding: Optimal feature selection when useful features are rare and weak,* Proc. Natl. Acad. Sci. USA 105 (2008), 14790–14795.

[13] D. Donoho, J. Jin, *Feature selection by higher criticism thresholding achieves the optimal phase diagram,* Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 367 (2009), 4449–4470.

[14] P. Hall, J. Jin, *Innovated higher criticism for detecting sparse signals in correlated noise,* Ann. Statist. 38 (2010), 1686–1732.

[15] Yu. Ingster, *Minimax detection of a signal in ℓ_p metrics,* Journal of Mathematical Sciences 68 (1994), 503–515.

[16] Y. Ingster, *Adaptive detection of a signal of growing dimension, I, II,* Math. Methods Statist. 10 (2002), 395–421.

[17] Y. Ingster, C. Pouet, A. Tsybakov, *Classification of sparse high-dimensional vectors,* Philosophical Transactions: Mathematical, Physical and Engineering Sciences 367 (2009), 4427–4448.

[18] Y. Ingster, A. Tsybakov, N. Verzelen, *Detection boundary in sparse regression,* Electronic Journal of Statistics 4 (2010), 1476–1526.

[19] R. Mukherjee, S. Sen, *On minimax exponents of sparse testing,* preprint (2020).

[20] G. Smirnov, *Gaussian volume bounds under hypercube translations and generalizations,* preprint (2024).

[21] M. Talagrand, *A new look at independence,* The Annals of probability (1996), 1–34.

[22] Tukey, J. W. (1976). *T13 N: The higher criticism.* Course Notes, Statistics 411, Princeton Univ.

[23] R. Vershynin, *High dimensional probability. An introduction with applications in Data Science.* Cambridge University Press, 2018.

THE AUSTRALIAN NATIONAL UNIVERSITY
Email address: shahar.mendelson@anu.edu.au

TEXAS A&M UNIVERSITY AND PRINCETON UNIVERSITY
Email address: grigoris@tamu.edu

UNIVERSITY OF CALIFORNIA, IRVINE
Email address: rvershyn@uci.edu