When Every Point Is Extreme

Bogdan Grechuk · 27 Nov 2026

https://functor.network/user/3333/entry/2102

This post continues my series on favourite theorems of the twenty-first century. My choice for 2006 in Probability and Statistics is a theorem of David Donoho and Jared Tanner about separating random points by hyperplanes. Their paper first appeared online on July 15, 2006, and was published in the Journal of the American Mathematical Society in 2009 (Donoho and Tanner 2009).

Imagine a large cloud of randomly sampled points in the plane. Many points will lie inside triangles formed by other points. Such an interior point cannot be separated from the rest of the cloud by a line.

In high dimensions, the situation can be dramatically different. Every point can be separated from all the others. More surprisingly, the same can hold for every sufficiently small subset of the cloud.

Let us make this precise.

Consider independent random points

$$x_1, \dots, x_n \in \mathbb{R}^d, \qquad X = \{x_1, \dots, x_n\}.$$

Initially, suppose that they have the standard Gaussian distribution: all their coordinates are independent, and each coordinate has density

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2).$$

A *hyperplane* in $\mathbb{R}^d$ is a set

$$\{x : \langle v, x \rangle = t\},$$

where $v \neq 0$, $t \in \mathbb{R}$, and

$$\langle v, x \rangle = \sum_{j=1}^{d} v_j x_j$$

is the usual inner product. In two dimensions a hyperplane is a line; in three dimensions it is a plane.

A subset $S \subseteq X$ can be *strictly separated* from the remaining points if there are $v \neq 0$ and t such that

$$\langle v, x \rangle > t \quad (x \in S), \qquad \langle v, x \rangle < t \quad (x \in X \setminus S).$$

Thus one linear inequality distinguishes the selected points from every other point in the sample.

Here is a precise asymptotic form of Donoho and Tanner's result.

> **Theorem.** Let $n = n(d)$ satisfy
>
> $$d \longrightarrow \infty, \qquad \frac{n}{d} \longrightarrow \infty, \qquad \frac{\log n}{d} \longrightarrow 0.$$
>
> Let X consist of n independent standard Gaussian points in $\mathbb{R}^d$. For every fixed $0 < \varepsilon < 1$, the probability tends to one that every subset $S \subseteq X$ with
>
> $$1 \leq |S| \leq \left\lfloor (1-\varepsilon) \frac{d}{2\mathrm{e} \log(n/d)} \right\rfloor$$
>
> can be strictly separated from $X \setminus S$ by a hyperplane.

Here $\log$ is the natural logarithm, e is its base, and $\lfloor u \rfloor$ denotes the greatest integer not exceeding u.

The conclusion holds simultaneously for every subset within the stated size bound. The separating hyperplane may depend on the subset, but one randomly sampled cloud has all these separation properties at once.

The result also holds when the points come from any common Gaussian distribution with nonsingular covariance. Such a distribution is an invertible affine image of the standard Gaussian distribution, and invertible affine maps preserve hyperplane separation.

What does the bound allow?

If $n = d^2$, it becomes

$$\left\lfloor (1-\varepsilon) \frac{d}{2\mathrm{e} \log d} \right\rfloor.$$

Thus, although the number of observations is much larger than the dimension, every subset containing up to this many observations can be isolated by a single linear inequality.

The permitted sample sizes can grow much faster than any polynomial. For example,

$$n = \left\lfloor \exp(\sqrt{d}) \right\rfloor$$

satisfies the assumptions, and the allowed subset size grows proportionally to $\sqrt{d}$.

These are asymptotic statements: the probability of the simultaneous separation property approaches one along the specified sequences.

There is an appealing geometric interpretation.

The *convex hull* of X consists of all weighted averages

$$\sum_{i=1}^{n} \lambda_i x_i, \qquad \lambda_i \geq 0, \qquad \sum_{i=1}^{n} \lambda_i = 1.$$

A point of X is a vertex of this convex hull precisely when it does not belong to the convex hull of the other points.

Taking subsets of size one in the theorem shows that every sampled point is a vertex, with probability tending to one. All these extreme points are ordinary observations from the same Gaussian distribution.

The underlying result is stronger still: the convex hull of every sufficiently small subset is a face of the full convex hull. A face is the intersection with a supporting hyperplane, meaning a hyperplane that meets the convex hull while leaving it entirely on one side.

Consequently, sufficiently small groups of points sit together on the boundary. In particular, every pair forms an edge once the guaranteed subset size is at least two.

The proof connects the random cloud with a familiar deterministic object.

Let $e_1, \ldots, e_n$ be the standard coordinate vectors in $\mathbb{R}^n$. Their convex hull is a simplex. If A is a d by n matrix with independent standard Gaussian entries, then

$$Ae_i = x_i$$

are independent Gaussian points. Their convex hull is therefore a random linear image of the simplex.

Donoho and Tanner estimate how many faces are lost under this mapping. Below the stated threshold, the expected number of lost small faces tends to zero, which forces the probability of losing even one to tend to zero. The estimates balance the number of candidate faces against probabilities expressed through Gaussian angles.

The same paper connects this geometry with sparse recovery: reconstructing a vector with few nonzero coordinates from a much smaller collection of linear measurements. The reconstruction minimizes the sum of absolute coordinate values among vectors consistent with the measurements.

The subject also connects with my own work. In later joint papers (Gorban et al. 2018; Grechuk et al. 2021), we developed related stochastic separation theorems for broad classes of non-Gaussian distributions and studied their use in correcting AI systems.

The basic idea is that a linear test can isolate a recorded error, represented by a vector, from other stored data. A corrector can then change the decision on that error while leaving decisions on the other stored vectors unchanged.

What I particularly like about Donoho and Tanner's theorem is this useful simplicity produced by high dimension. A random cloud can have vastly more points than coordinates, yet every small selected group remains accessible by one linear inequality. Probability turns a complicated collection of observations into a remarkably structured geometric object.

References

Donoho, David L., and Jared Tanner. 2009. "Counting Faces of Randomly Projected Polytopes When the Projection Radically Lowers Dimension." *Journal of the American Mathematical Society* 22 (1): 1–53. https://doi.org/10.1090/S0894-0347-08-00600-0.

Gorban, A. N., A. Golubkov, Bogdan Grechuk, E. M. Mirkes, and I. Y. Tyukin. 2018. "Correction of AI Systems by Linear Discriminants: Probabilistic Foundations." *Information Sciences* 466: 303–22. https://doi.org/10.1016/j.ins.2018.07.040.

Grechuk, Bogdan, Alexander N. Gorban, and Ivan Y. Tyukin. 2021. "General Stochastic Separation Theorems with Optimal Bounds." *Neural Networks* 138: 33–56. https://doi.org/10.1016/j.neunet.2021.01.034.