

Anti-concentration Inequalities

Roman Vershynin Mark Rudelson

University of California, Davis

University of Missouri-Columbia

Phenomena in High Dimensions

Third Annual Conference

Samos, Greece

June 2007

Concentration and Anti-concentration

- **Concentration phenomena:** Nice random variables X are concentrated about their means.
- Examples:
 1. **Probability theory:** $X =$ sum of independent random variables (concentration inequalities: Chernoff, Bernstein, Bennett, ...; large deviation theory).
 2. **Geometric functional analysis:** $X =$ Lipschitz function on the Euclidean sphere.
- *How strong* concentration should one expect?
No stronger than a Gaussian (Central Limit Theorem).
- **Anti-concentration phenomena:** nice random variables S concentrate *no stronger* than a Gaussian.
(Locally well spread).

Concentration and Anti-concentration

- **Concentration inequalities:**

$$\mathbb{P}(|X - \mathbb{E}X| > \varepsilon) \leq ?$$

- **Anti-concentration inequalities:** for a given (or all) v ,

$$\mathbb{P}(|X - v| \leq \varepsilon) \leq ?$$

- Concentration is better understood than anti-concentration.

Anti-concentration

Problem

Estimate *Lévy's concentration function* of a random variable X :

$$p_\varepsilon(X) := \sup_{v \in \mathbb{R}} \mathbb{P}(|X - v| \leq \varepsilon).$$

1. Probability Theory.

- For *sums of independent random variables*, studied by [Lévy, Kolmogorov, Littlewood-Offord, Erdős, Esséen, Halasz, . . .]
- For *random processes* (esp. Brownian motion), see the survey [Li-Shao]

Anti-concentration

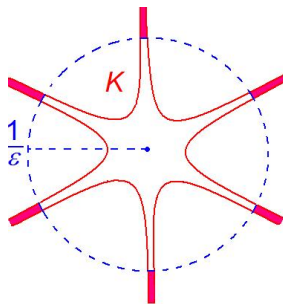
2. Geometric Functional Analysis. For Lipschitz functions:

Small Ball Probability Theorem

Let f be a convex even function on the unit Euclidean sphere (S^{n-1}, σ) , whose average over the sphere = 1 and Lipschitz constant = L . Then

$$\sigma(\mathbf{x} : |f(\mathbf{x})| \leq \varepsilon) \leq \varepsilon^{c/L^2}.$$

- Conjectured by V.; [Latała-Oleszkiewicz] deduced the Theorem from the *B-conjecture*, solved by [Cordero-Fradelizi-Maurey].
- **Interpretation.** $K \subseteq \mathbb{R}^n$: convex, symmetric set; $f(\mathbf{x}) = \|\mathbf{x}\|_K$.
SBPT: asymptotic “dimension” of the spikes (parts of K far from the origin) is $\gtrsim 1/L^2$.
- Applied to Dvoretzky-type thms in [Klartag-V.]

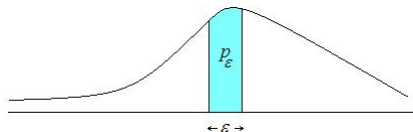


Anti-concentration

$$p_\varepsilon(X) := \sup_{v \in \mathbb{R}} \mathbb{P}(|X - v| \leq \varepsilon).$$

- What estimate can we *expect*?
- For every random variable X **with density**, we have

$$p_\varepsilon(X) \sim \varepsilon.$$



- If X is **discrete**, this fails for small ε (because of the atoms), so we can only expect

$$p_\varepsilon(X) \lesssim \varepsilon + \text{measure of an atom.}$$

Anti-concentration

- **Classical example:** Sums of independent random variables

$$S := \sum_{k=1}^n a_k \xi_k$$

where ξ_1, \dots, ξ_n are i.i.d. (we can think of ± 1),
and $\mathbf{a} = (a_1, \dots, a_n)$ is a fixed vector of real coefficients

- An *ideal estimate* on the concentration function would be

$$p_\varepsilon(\mathbf{a}) := p_\varepsilon(S) \lesssim \varepsilon / \|\mathbf{a}\|_2 + e^{-cn},$$

where e^{-cn} accounts for the size of atoms of S .

Anti-concentration

- Ideal estimate:

$$p_\varepsilon(\mathbf{a}) = \sup_{\mathbf{v} \in \mathbb{R}} \mathbb{P}(|\mathbf{S} - \mathbf{v}| \leq \varepsilon) \lesssim \varepsilon / \|\mathbf{a}\|_2 + e^{-cn}.$$

- Trivial example: **Gaussian sums**,
with $\xi_k =$ standard normal i.i.d. random variables.
The ideal estimate holds even without the exponential term.
- Nontrivial example: **Bernoulli sums**,
with $\xi_k = \pm 1$ symmetric i.i.d. random variables.
- The problem for Bernoulli sums is nontrivial even for $\varepsilon = 0$,
i.e. estimate the **size of atoms** of \mathbf{S} .
This is the *most studied case* in the literature.

Application: Random matrices

This was our main motivation.

- A : an $n \times n$ matrix with i.i.d. entries.
What is the probability that A is singular?

Ideal answer: e^{-cn} .

- Geometric picture.

Let X_k denote the column vectors of A .

A nonsingular $\Rightarrow X_1 \notin \text{span}(X_2, \dots, X_n) := H$

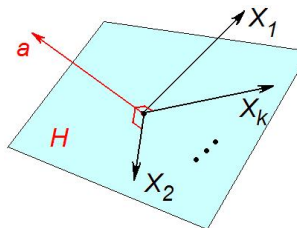
- We condition on H (i.e. on X_2, \dots, X_n); let a be the normal of H .
 A nonsingular $\Rightarrow \langle a, X_1 \rangle \neq 0$.

Write this in coordinates for $a = (a_k)_1^n$ and $X = (\xi_k)_1^n$ (i.i.d):

$$A \text{ is nonsingular} \Rightarrow \sum_{k=1}^n a_k \xi_k \neq 0.$$

$$\mathbb{P}(A \text{ is singular}) \geq p_0(a).$$

- Thus, in order to solve the invertibility problem, we *have* to prove an anti-concentration inequality. See Mark Ridelson's talk.



Anti-concentration: the Littlewood-Offord Problem

Littlewood-Offord Problem.

For Bernoulli sums $S = \sum a_k \xi_k$, estimate the concentration function

$$p_\varepsilon(\mathbf{a}) = \sup_{v \in \mathbb{R}} \mathbb{P}(|S - v| \leq \varepsilon).$$

- For **concentrated vectors**, e.g. $\mathbf{a} = (1, 1, 0, \dots, 0)$,
 $p_0(\mathbf{a}) = \frac{1}{2} = \text{const.}$
There are lots of cancelations in the sum $S = \pm 1 \pm 1$.
- For **spread vectors**, the small ball probability gets better:
for $\mathbf{a} = (1, 1, 1, \dots, 1)$, we have $p_0(\mathbf{a}) = \binom{n}{n/2} / 2^n \sim n^{-1/2}$.
- This is a general fact:

If $\mathbf{a} \geq 1$ pointwise, then $p_0(\mathbf{a}) \leq p_0(1, 1, \dots, 1) \sim n^{-1/2}$.
[Littlewood-Offord], [Erdős, 1945].

- Still *lots of cancelations* in the sum $S = \pm 1 \pm 1 \dots \pm 1$.
How can one prevent cancelations?

Anti-concentration: the Littlewood-Offord Problem

Littlewood-Offord Problem.

For Bernoulli sums $S = \sum a_k \xi_k$, estimate the concentration function

$$p_\varepsilon(\mathbf{a}) = \sup_{v \in \mathbb{R}} \mathbb{P}(|S - v| \leq \varepsilon).$$

- Will be less cancelations if the coefficients are **essentially different**:
For $\mathbf{a} = (1, 2, 3, \dots)$, we have $p_0(\mathbf{a}) \sim n^{-3/2}$.
- This is a general fact:

If $|a_j - a_k| \geq 1$ for $k \neq j$, then $p_1(\mathbf{a}) \lesssim n^{-3/2}$.

[Erdős-Moser, 1965], [Sárközi-Szemerédi, 1965], [Hálasz, 1977].

- Still *lots of cancelations* in the sum $S = \pm 1 \pm 2 \cdots \pm n$.
- **Question.** *How to prevent cancelations in random sums?*
For what vectors \mathbf{a} is the concentration function $p_0(\mathbf{a})$ small?
E.g. *exponential* rather than polynomial.

Anti-concentration: the Littlewood-Offord Phenomenon

- [Tao-Vu, 2006] proposed an explanation for cancelations, which they called the *Inverse Littlewood-Offord Phenomenon*:
- The only source of cancelations in random sums $S = \sum \pm a_k$ is a rich **additive structure** of the coefficients a_k .
- Cancelations can only occur when the coefficients a_k are *arithmetically commensurable*. Specifically, if there are lots of cancelations, then the coefficients a_k can be embedded into a **short arithmetic progression**.

The Inverse Littlewood-Offord Phenomenon

If the small ball probability $p_\varepsilon(a)$ is large, then the coefficient vector a can be embedded into a short arithmetic progression.

Anti-concentration: the Littlewood-Offord Phenomenon

Theorem (Tao-Vu)

Let a_1, \dots, a_n be integers, and let $A \geq 1$, $\delta \in (0, 1)$. Suppose for the random Bernoulli sums one has

$$p_0(\mathbf{a}) \geq n^{-A}.$$

Then all except $O_{A,\varepsilon}(n^\delta)$ coefficients a_k are contained in the Minkowski sum of $O(A/\delta)$ arithmetic progressions of lengths $n^{O_{A,\delta}(1)}$.

- **Usefulness.** One can reduce the small ball probability to an **arbitrary polynomial order** by controlling the additive structure of a .
- **Shortcomings.** **1.** We often have *real coefficients* a_k (not \mathbb{Z}).
2. We are interested in *general small ball probabilities* $p_\varepsilon(\mathbf{a})$ rather than the measure of atoms $p_0(\mathbf{a})$.
- **Problem.** **Develop the Inverse L.-O. Phenomenon over \mathbb{R} .**

Essential integers

- For *real* coefficient vectors $a = (a_1, \dots, a_n)$, the embedding into an arithmetic progression must clearly be *approximate* (*near* an arithmetic progression).
- Thus we shall work over the **essential integer** vectors: *almost* all their coefficients (99%) are *almost* integers (± 0.1).

Embedding into arithmetic progressions via LCD

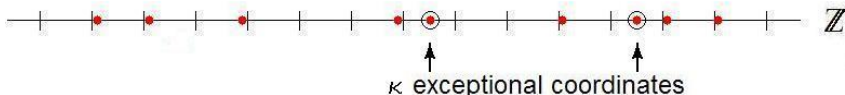
- **Goal:** embed a vector $a \in \mathbb{R}^n$ into a short arithmetic progression (essentially). What is its length?
- Bounded by the **essential least common denominator (LCD)** of a :

$$D(a) = D_{\alpha, \kappa}(a) = \inf\{t > 0 : ta \text{ is a nonzero essential integer}\}$$

(all except κ coefficients of ta are of dist. α from nonzero integers).

- For $a \in \mathbb{Q}^n$, this is the usual LCD.

Coordinates of $D(a)a$



- The vector $D(a)a$ (and thus a itself) essentially embeds into an arithmetic progression of length $\|D(a)a\|_{\infty} \lesssim D(a)$.
So, $D(a)$ being small means that a has rich **additive structure**.
- Therefore, the Inverse L.-O. Phenomenon should be:
if the small ball probability $p_{\varepsilon}(a)$ is large, then $D(a)$ is small.

Anti-concentration: the Littlewood-Offord Phenomenon

Theorem (Anti-Concentration)

Consider a sum of independent random variables

$$S = \sum_{k=1}^n a_k \xi_k$$

where ξ_k are i.i.d. with third moments and $C_1 \leq |a_k| \leq C_2$ for all k .
Then, for every $\alpha \in (0, 1)$, $\kappa \in (0, n)$ and $\varepsilon \geq 0$ one has

$$p_\varepsilon(S) \lesssim \frac{1}{\sqrt{\kappa}} \left(\varepsilon + \frac{1}{D_{\alpha, \kappa}(\mathbf{a})} \right) + e^{-c\alpha^2 \kappa}.$$

Recall: $D_{\alpha, \kappa}(\mathbf{a})$ is the essential LCD of \mathbf{a} ($\pm\alpha$ and up to κ coefficients).

Partial case:

Anti-concentration: the Littlewood-Offord Phenomenon

$$p_\varepsilon(\mathbf{a}) \lesssim \frac{1}{\sqrt{\kappa}} \left(\varepsilon + \frac{1}{D_{\alpha, \kappa}(\mathbf{a})} \right) \quad \text{if all } |a_k| \sim \text{const.} \quad (\text{ILO})$$

Partial case:

- $\varepsilon = 0$; thus $p_0(\mathbf{a})$ is the measure of atoms
- accuracy $\alpha = 0.1$
- number of exceptional coefficients $\kappa = 0.01n$:

Inverse Littlewood-Offord Phenomenon

99% of the coefficients of \mathbf{a} are within 0.1 of an arithmetic progression of length $\sim n^{-1/2}/p_0(\mathbf{a})$.

- By controlling the additive structure of \mathbf{a} (removing progressions), we can **force the concentration function to arbitrarily small level**, up to exponential in n .

Examples:

Anti-concentration: the Littlewood-Offord Phenomenon

$$p_\varepsilon(\mathbf{a}) \lesssim \frac{1}{\sqrt{\kappa}} \left(\varepsilon + \frac{1}{D_{\alpha, \kappa}(\mathbf{a})} \right) \quad \text{if all } |a_k| \sim \text{const.} \quad (\text{ILO})$$

Examples. $\varepsilon = 0$, accuracy $\alpha = 0.1$, exceptional coeffs $\kappa = 0.01n$:

- $\mathbf{a} = (1, 1, \dots, 1)$. Then $D(\mathbf{a}) \gtrsim \text{const.}$ Thus (ILO) gives

$$p_0(\mathbf{a}) \lesssim n^{-1/2}. \quad \text{Optimal (middle binomial).}$$

- $\mathbf{a} = (1, 2, \dots, n)$. To apply (ILO), we normalize and truncate:

$$p_0(\mathbf{a}) = p_0\left(\frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\right) \leq p_0\left(\frac{n/2}{n}, \frac{n/2+1}{n}, \dots, \frac{n}{n}\right)$$

The LCD of such vector is $\gtrsim n$. Then (ILO) gives

$$p_0(\mathbf{a}) \lesssim n^{-3/2}. \quad \text{Optimal.}$$

- \mathbf{a} more irregular \Rightarrow can reduce $p_0(\mathbf{a})$ **further**.

Soft approach

- We will sketch the **proof**.
There are two approaches, soft and ergodic.
- **Soft approach**: deduce anti-concentration inequalities from **Central Limit Theorem**. [Litvak-Pajor-Rudelson-Tomczak].
- By CLT, the random sum

$$S \approx \text{Gaussian.}$$

Hence can approximate the concentration function

$$p_\varepsilon(S) \approx p_\varepsilon(\text{Gaussian}) \sim \varepsilon.$$

- For this, one uses a *non-asymptotic* version of CLT [Berry-Esséen]:

Soft approach

Theorem (Berry-Esséen's Central Limit Theorem)

Consider a sum of independent random variables $S = \sum a_k \xi_k$, where ξ_k are i.i.d. centered with variance 1 and finite third moments. Let g be the standard normal random variable. Then

$$|\mathbb{P}(S/\|a\|_2 \leq t) - \mathbb{P}(g \leq t)| \lesssim \left(\frac{\|a\|_3}{\|a\|_2} \right)^3 \quad \text{for every } t.$$

- The more *spread* the coefficient vector a , the better (RHS smaller). RHS minimized for $a = (1, 1, \dots, 1)$, for which it is $\left(\frac{n^{1/3}}{n^{1/2}}\right)^3 = n^{-1/2}$. Thus the best bound the soft approach gives is $p_0(a) \leq n^{-1/2}$.
- **Anti-concentration inequalities can not be based on ℓ_p norms** of the coefficient vector a (which works nicely for the concentration inequalities, e.g. Bernstein's!).
- The ℓ_p norms do not distinguish between $(1, 1, \dots, 1)$ and $(1 + \frac{1}{n}, 1 + \frac{2}{n}, \dots, 1 + \frac{n}{n})$, for which concentration functions are different. The norms *feel the bulk* and *ignore the fluctuations*.

Ergodic approach

Instead of applying Berry-Esséen's CLT directly, use a tool from its proof: Esséen's inequality. This method goes back to [Halasz, 1977].

Proposition (Esséen's Inequality)

The concentration function of any random variable S is bounded by the L^1 norm of its characteristic function $\phi(t) = \mathbb{E} \exp(iSt)$:

$$p_\varepsilon(S) \lesssim \int_{-\pi/2}^{\pi/2} |\phi(t/\varepsilon)| dt.$$

- **Proof:** take Fourier transform.
- We use Esséen's Inequality for the random sum $S = \sum_1^n a_k \xi_k$. We work with the example of Bernoulli sums ($\xi_k = \pm 1$). By the independence, the characteristic function of S factors

$$\phi(t) = \prod_1^n \phi_k(t), \quad \phi_k(t) = \mathbb{E} \exp(ia_k \xi_k t) = \cos(a_k t).$$

Ergodic approach

Then

$$|\phi(t)| = \prod_1^n |\cos(a_k t)| \leq \exp(-f(t)),$$

where

$$f(t) = \sum_1^n \sin^2(a_k t).$$

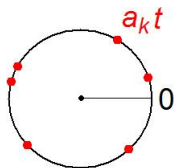
By Esséen's Inequality,

$$\begin{aligned} \rho_\varepsilon(\mathcal{S}) &\lesssim \int_{-\pi/2}^{\pi/2} |\phi(t/\varepsilon)| dt \leq \int_{-\pi/2}^{\pi/2} \exp(-f(t/\varepsilon)) dt \\ &\sim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt. \end{aligned}$$

Ergodic approach

$$p_\varepsilon(\mathcal{S}) \lesssim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt, \quad \text{where } f(t) = \sum_1^n \sin^2(a_k t).$$

- **Ergodic approach:** regard t as *time*; $\varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} =$ long term average.
- A system of n particles $a_k t$ that move along \mathbb{T} at speeds a_k :



- The estimate is *poor* precisely when $f(t)$ is small
 \Leftrightarrow most particles *return to the origin*, making $\sin^2(a_k t)$ small.
- We are thus interested in the **recurrence properties** of the system.
How often do most particles return to the origin?

Ergodic approach

$$p_\varepsilon(\mathcal{S}) \lesssim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt, \quad \text{where } f(t) = \sum_1^n \sin^2(a_k t).$$

- We need to understand *how particles can move in the system*.
- Two extreme types of systems (common in ergodic theory):
 1. **Quasi-random** (“mixing”). Particles move as if independent.
 2. **Quasi-periodic**. Particles “stick together”.

Ergodic approach

$$p_\varepsilon(\mathbf{S}) \lesssim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt, \quad \text{where } f(t) = \sum_1^n \sin^2(a_k t).$$

1. Quasi-random systems.

- By “independence”, the event that *most particles are near the origin* is exponentially rare (frequency e^{-cn}).
- Away from the origin, $\sin^2(a_k t) \geq \text{const}$, thus $f(t) \sim cn$.
- This leads to the bound

$$p_\varepsilon(\mathbf{S}) \lesssim \varepsilon + e^{-cn}.$$

(ε is due to a constant initial time to depart from the origin).

- This is an ideal bound. Quasi-random systems are *good*.

Ergodic approach

$$p_\varepsilon(\mathcal{S}) \lesssim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt, \quad \text{where } f(t) = \sum_1^n \sin^2(a_k t).$$

2. Quasi-periodic systems.

- **Example.** $\mathbf{a} = (1, 1, \dots, 1)$. Move as one particle.
Thus $f(t) \sim n \sin^2 t$, and integration gives $p_\varepsilon(\mathcal{S}) \lesssim n^{-1/2}$.
- **More general example.** Rational coefficients with **small LCD**. Then ta_k often becomes an integer, i.e. the particles often return to the origin together.
- **Main observation.** Small LCD is the *only* reason for the almost periodicity of the system:

Ergodic approach

$$p_\varepsilon(\mathcal{S}) \lesssim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt, \quad \text{where } f(t) = \sum_1^n \sin^2(a_k t).$$

Observation (Quasi-periodicity and LCD)

If a system (ta_k) is quasi-periodic then essential LCD of (a_k) is small.

- **Proof.** Assume most of ta_k often return near the origin together – say, with **frequency** ω (i.e. spend portion of time ω near the origin).
- Equivalently, ta becomes an *essential integer* with frequency ω .
- Thus ta becomes essential integer **twice within time** $\sim \frac{1}{\omega}$.
 \exists two instances $0 < t_1 - t_2 < 1/\omega$ in which $t_1 a$ and $t_2 a$ are different essential integers.
- Subtract $\Rightarrow (t_2 - t_1)a$ is also an essential integer.
By the definition of the essential LCD,

$$D(a) \leq t_2 - t_1 < \frac{1}{\omega}.$$

□

Ergodic approach

$$p_\varepsilon(\mathcal{S}) \lesssim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt, \quad \text{where } f(t) = \sum_1^n \sin^2(a_k t).$$

- Conclusion of the proof.

1. If the essential LCD $D(\mathbf{a})$ is large, then the system is *not* quasi-periodic \Rightarrow closer to *quasi-random*.

2. For quasi-random systems, the concentration function $p_\varepsilon(\mathcal{S})$ is small.

- Ultimately, the argument gives

$$p_\varepsilon(\mathbf{a}) \lesssim \frac{1}{\sqrt{n}} \left(\varepsilon + \frac{1}{D(\mathbf{a})} \right) + e^{-cn}.$$



Improvements

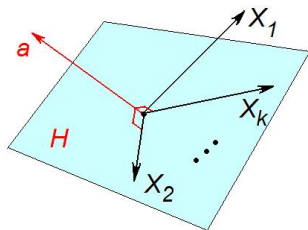
[O.Friedland-S.Sodin] recently simplified the argument:

- Used a more convenient notion of essential integers as *vectors in \mathbb{R}^n that can be approximated by integer vectors within $\alpha\sqrt{n}$ in Euclidean distance.*
- Bypassed *Halasz's regularity argument* (which I skipped) using a direct and simple analytic bound.

Using the anti-concentration inequality

$$p_\varepsilon(\mathbf{a}) \lesssim \frac{1}{\sqrt{n}} \left(\varepsilon + \frac{1}{D(\mathbf{a})} \right) + e^{-cn}.$$

- In order to use the anti-concentration inequality, we need to know that LCD of \mathbf{a} is *large*.
- Is LCD large for *typical* (i.e. random) coefficient vectors \mathbf{a} ?
- For random matrix problems, \mathbf{a} = normal to the random hyperplane spanned by $n - 1$ i.i.d. vectors X_k in \mathbb{R}^n :



- **Random Normal Theorem:** $D(\mathbf{a}) \geq e^{cn}$ with probability $1 - e^{-cn}$.