Anticoncentration and Polynomial Decompositions

Jacob Denson

September 24th, 2022

Let $X \in \mathbb{R}^N$ be a random vector with independent coordinates. The *invariance principle* says that if $f : \mathbb{R}^N \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ are 'regular', then the quantities $\mathbb{E}[\psi(f(X))]$ depend only on very coarse properties of the distribution of X up to a small error. A basic instance is the central limit theorem, which says that a sum of independent random variables is approximately normally distributed, and thus independent of all properties of those variables but their mean and variance, stated below.

Theorem 1 (Lindeberg)**.** *Let X and A be random vectors in $\mathbb{R}^N$, each having independent coordinates, and sharing the same means and variances. Let*

$$\gamma = \sum_i \mathbb{E}|X_i|^3 + \mathbb{E}|A_i|^3.$$

Let $f(z) = z_1 + \dots + z_N$. Then for any $\psi : \mathbb{R} \to \mathbb{R}$,

$$|\mathbb{E}[\psi(f(X))] - \mathbb{E}[\psi(f(A))]| \leq \|D^3\psi\|_{L^\infty(\mathbb{R})} \cdot \frac{\gamma_3}{6}.$$

One can still exploit this theorem to get transference principles which apply to less regular f, for instance, for quantities with a simple jump discontinuity like if $\psi(t) = \mathbf{I}(t \leq s)$, for which $\mathbb{E}[\psi(f(X))] = \mathbb{P}(f(X) \leq s)$ gives the CDF of the random variable $f(X)$, or $\psi(t) = \text{sgn}(t)$, in which case $\psi(f(X))$ is called a *threshold function.* To get these theorems, we apply an additional *anticoncentration inequality.* Let's see why in Lindeberg's scenario: pick a non-negative $\eta \in C^\infty$ supported on $|t| \leq 1$ and with $\int \eta(x)\ dx = 1$, and we define $\psi_\varepsilon = \psi * \text{Dil}_\varepsilon \eta$, then $\|D^3\psi\|_{L^\infty} \lesssim \varepsilon^{-3}$, and since $\psi(t) \leq \psi_\varepsilon(t - \varepsilon) \leq \psi(t - 2\varepsilon)$,

$$\begin{aligned}
\mathbb{P}(f(A) \leq s) &\leq \mathbb{E}[\psi_\varepsilon(f(A) - \varepsilon)] \\
&\leq \mathbb{E}[\psi_\varepsilon(f(X) - \varepsilon)] + O\left(\varepsilon^{-3}\gamma_3\right) \\
&\leq \mathbf{P}(f(X) \leq s + 2\varepsilon) + O(\varepsilon^{-3}\gamma_3).
\end{aligned}$$

Similarily, one shows $\mathbb{P}(f(A) \leq s) \leq \mathbb{P}(f(X) \leq s - 2\varepsilon) + O(\varepsilon^{-3}\gamma_3)$. Thus to finish this argument and show that $\mathbb{P}(f(A) \leq s) \approx \mathbb{P}(f(X) \leq s)$, we must show that $\mathbb{P}(s - 2\varepsilon \leq f(X) \leq s + 2\varepsilon)$ is small, i.e. that $f(X)$ *does not concentrate.* If for simplicity we assume $f(X)$ and $f(A)$ both have variance one, then we find that $\mathbb{P}(s-2\varepsilon \leq f(X) \leq s+2\varepsilon) \lesssim \varepsilon$, and plugging this in gives that $|\mathbb{P}(f(A) \leq s) - \mathbb{P}(f(X) \leq s)| \lesssim \varepsilon + \varepsilon^{-3}\gamma_3$. Picking $\varepsilon = \gamma_3^{1/4}$ gives an error $O(\gamma_3^{1/4})$. We have thus proved the *Berry-Esseen theorem* by means of an *anticoncentration inequality* for the Gaussian.

The paper [1] we discuss here studies anticoncentration inequalities for random quantities $f(X)$, where f is no longer a linear sum, but a *polynomial* p with an independent vector as inputs. For instance, one might want to study $\psi(f(X))$ with $\psi(x) = \mathrm{sgn}(x)$, quantities called *polynomial threshold functions*. There are results already existing in the literature that give general anticoncentration bounds for general polynomials of a fixed degree (a result of Carbery-Wright), and this result is tight for general polynomials. The paper [1] gives tools indicating a way to identify polynomials for which one can *improve* this anticoncentration result, via a decomposition of this polynomial. One consequence is more sophisticated invariance principles for polynomials, with better error terms if a polynomial has the right decompositions.

1 Notation

- We write X, Y, and Z for standard normal random vectors, and let A and B be Bernoulli random vectors. We write γ for the normal Gaussian distribution on $\mathbb{R}^N$, and β for the Bernoulli distribution on $\{-1,+1\}^N$. We thus have norms $L^p_\gamma(\mathbb{R}^N)$ and $L^p_\beta(\mathbb{R}^N)$ for functions $f : \mathbb{R}^N \to \mathbb{R}$ given by
$$\|f\|_{L^p_\gamma} = \mathbb{E}[|f(X)|^p]^{1/p} \quad \text{and} \quad \|f\|_{L^p_\beta} = \mathbb{E}[|f(A)|^p]^{1/p}.$$
Similarily, we have variances $\mathrm{Var}_\gamma(f)$ and $\mathrm{Var}_\beta(f)$.

- The ith influence $\mathrm{Inf}_i(f)$ is defined as $\|\partial f/\partial x_i\|^2_{L^2_\gamma}$. This agrees with the standard definition of influence that occurs in the analysis of Boolean functions, in the case that f is a multilinear polynomial.

- A k-tensor on $\mathbb{R}^N$ is a quantity of the form
$$\sum A_S dx^{\otimes S}$$
where $\{A_S\}$ are real numbers, and S ranges over $[k]^S$. A k-tensor valued function is
$$A(x) = \sum A_S(x) dx^{\otimes S}.$$
The magnitude $|A|$ of a k-tensor is equal to $(\sum |A_S|^2)^{1/2}$, and using this we can define the L^p_γ and L^p_β norms of a k-tensor valued function in the way you would expect, i.e. as $\mathbb{E}[|A(X)|^p]^{1/p}$ and $[|A(B)|^p]^{1/p}$.

2 The Main Result

For general polynomials p of a fixed degree d, Carbery-Wright showed that

$$\mathbb{P}(|p(X)| \leq \varepsilon \|p\|_{L^2_\gamma}) \lesssim d\varepsilon^{1/d}.$$

This result is tight, for instance, if $p(x) = (x_1 + \cdots + x_N)^d$, or $p(x) = q_1(x)^7 + q_2(x)^7 + q_1(x)^2 q_2(x)^2 + \delta q_3$, where q_1 and q_2 are polynomials of degree d and δ is small. But the $\varepsilon^{1/d}$ error term leads to invariance principles which have poor dependence on d, i.e. the following result.

Theorem 2 (Mossel, O'Donnell, Oleszkiewicz)**.** *If p is a τ-regular multilinear polynomial of degree d, i.e. $Inf_i(p) \leq \tau \, Var_\beta(p)$ for all indices i, then*

$$|\mathbb{P}(p(X) \leq t) - \mathbb{P}(p(A) \leq t| \lesssim d\tau^{1/8d}.$$

Given the poor dependence on d here (tight for general inputs), to obtain better invariance principles it is useful to identify those particular scenarios in which we can improve upon the general result of Carbery-Wright, or equivalently, to identify all obstacles which make the Carbery-Wright inequality tight. Notice that the tight examples to Carbery-Wright are of the form $h(q_1, \dots, q_m)$, where h is a poorly behaved polynomial, and (q_1, q_2) has good anticoncentration results. This is indeed true of all badly behaved counterexamples up to a small error term, which is the main result to be discussed.

We begin with some definitions. We say a vector $q = (q_1, \dots, q_m)$ of polynomial functions $q_i : \mathbb{R}^N \to \mathbb{R}$ is (ε, α) *diffuse* if for any $a \in \mathbb{R}^m$,

$$\mathbb{P}(|q(X) - a| \leq \varepsilon) \leq \varepsilon^m \alpha.$$

Intuitively, this means the probability density of the random vector $q(X)$ has average value at most α on any box of sidelength α. We say a polynomial $p : \mathbb{R}^N \to \mathbb{R}$ has a *decomposition* into $h(q_1, \dots, q_m)$ for $h : \mathbb{R}^m \to \mathbb{R}$ and $q_i : \mathbb{R}^N \to \mathbb{R}$ if $p = h(q_1, \dots, q_m)$, and if, for any monomial $\prod_{i \in \beta} x_i^\beta$ occuring in h, and monomials $x^{\beta_1}, \dots, x^{\beta_m}$ occuring in $q_1, \dots, q_m$ respectively, $\deg(\prod_{i \in \beta} x^{\beta_i}) \leq \deg(p)$. This is to prevent some decomposition where high degree terms in the decomposition cancel each other out, which complicates the analysis of the random variables involved. We can now state the structure result for polynomials, which gives the main result of [1].

Theorem 3. *For any degree d polynomial p, and any $\varepsilon, N, c > 0$, there exists a degree d polynomial p_0, a polynomial vector $q = (q_1, \dots, q_m)$, and a polynomial $h : \mathbb{R}^m \to \mathbb{R}$ such that p_0 has a decomposition into $h(q_1, \dots, q_m)$, and:*

- *$p \approx p_0$ in the quantitative sense that $\|p - p_0\|_{L^2_\gamma} \lesssim_{c,d,N} \varepsilon^N \|p\|_{L^2_\gamma}$.*
- *$(q_1, \dots, q_m)$ is $(\varepsilon, \varepsilon^{-c})$ diffuse, and $m \lesssim_{c,d,N} 1$.*

Remark. The currently known dependence on m on c, d, and N is currently very poor, i.e. that $m \leq A(d + O(1), N/c)$, where A is the Ackerman function. But it is conjectured that one can find a bound which is polynomial in (dN/c).

This structure result is closely related to the characterization of polynomials for which concentration does not occur. A polynomial for which Theorem 2 is tight (in the sense of the dependence of the result on $\tau^{-1/d}$), for an even dimension d, is the multilinear projection q of the polynomial

$$p(x_0, \ldots, x_N) = \tau x_0 + \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i \right)^d = p_0(x) + p_1(x)^d.$$

As $N \to \infty$, $\lim_{N\to\infty} \|p - q\|_{L^2_\gamma} = 0$, so results like hypercontractivity imply that the distributions of $p(X) - q(X)$ are very close, i.e. for any $\delta > 0$, there exists $N_0 > 0$ such that if $N \geq N_0$, then

$$\mathbb{P}(|p(X) - q(X)| \geq \delta A) \lesssim 2^{-c_d A^{2/d}}.$$

Now $\mathrm{Inf}_0(q) = \tau^2$, $\mathrm{Inf}_i(q) \lesssim_d 1/N$, and $\mathrm{Var}_\beta(q) \gtrsim 1 + \tau^2$. Thus q is τ-regular for $\tau \lesssim 1$, and so Theorem 2 applies to p. Note that since A is $\{-1, 1\}^{N+1}$ valued, we always have $p(A) \geq -\tau$. On the other hand, if X is Gaussian, $p_1(X_1, \ldots, X_N)$ then we can guarantee that $|p_1(X_1, \ldots, X_N)| \lesssim \tau^{1/d}$ with probability $\gtrsim \tau^{1/d}$, so $|Lq_1(X_1, \ldots, X_N)| \lesssim \tau^{1/d}$ with probability $\gtrsim \tau^{1/d}$. On the other hand, we guarantee that $\tau X_0 \leq -2\tau$ with probability $\gtrsim 1$. Thus by independence, *both properties* hold with probability $\gtrsim \tau^{1/d}$, and in this case $p(X) \leq -\tau$. Thus

$$|\mathbb{P}(p(X) \leq -\tau) - \mathbb{P}(p(A) \leq -\tau)| = \mathbb{P}(p(X) \leq -\tau) \gtrsim \tau^{1/d}.$$

What happened here? Even though the first coordinate has small influence, the Carbery-Wright result was tight for the polynomial $p_1(X)^d$, i.e. the polynomial concentrated within a $O(\tau)$ neighborhood of zero with probability $\Omega(\tau^{-1/d})$. The Boolean polynomial $p_0(A)^d$ also concentrates within a $O(\tau)$ neighborhood of zero with probability. In this situation, $p(X) \approx \tau X_0$ and $p(A) \approx \tau A_0$, and this causes a problem since X_0 and A_0 are *very* different probability distributions.

Theorem 4. *We say a degree d multilinear polynomial has a $(\tau, \alpha, m, \varepsilon)$ regular decomposition if there exists a polynomial p_0 of degree d such that*

$$\|p - p_0\|_{L^2_\beta} \leq \varepsilon \cdot \mathit{Var}_\gamma(p_0(X))^{1/2}$$

and $p_0 = h(q_1, \ldots, q_m)$, where $(q_1, \ldots, q_m)$ is a vector of multilinear polynomials which is $(\tau^{1/5}, \alpha)$ diffuse and $\mathit{Inf}_j(q_i) \leq \tau$ for all i and j. Under these conditions, for $0 < \tau, \varepsilon < 1/2$, we have

$$|\mathbb{P}(p(A) \leq t) - \mathbb{P}(p(X) \leq t)| \lesssim_{d,m} \tau^{1/5} \alpha \log(1/\tau)^{dm/2+1} + \varepsilon^{1/d} \log(1/\varepsilon)^{1/2}.$$

For $p_0(x) = \tau x_0$ and $p_1(x) = (x_1 + \cdots + x_N)/\sqrt{N}$, the theorem above can only apply with $\alpha = \tau^{-1/5}$, which yields a relatively useful error term of $O(\log(1/\tau))$. Thus the assumptions of this theorem avoid this kind of concentration phenomenon. To recover Theorem 2 from this result, the regularity assumption there implies that the polynomials $(x_0, x_1, \dots, x_n, p)$ are $(\tau^{1/5}, O(d\tau^{(1/d-1)/5}))$ diffuse, and so the theorem above gives that

$$|\mathbb{P}(p(X) \leq t) - \mathbb{P}(p(A) \leq t)| \lesssim_d \tau^{1/5d} \log(1/t)^{d/2+1},$$

which is analogous to Theorem 2 in the sense that we still get a power of $\tau^{1/d}$.

Even if a polynomial does not satisfy the regularity conditions, this might only be true a 'few coordinates are bad', and we can obtain a polynomial by fixing a few values of the polynomial. Results showing this are true are called 'regularity lemmas'. Here is a result applying to the assumptions of Theorem 2.

Theorem 5 (Diakonix, Servedio, Tan, Wan)**.** *If $f = sgn(p(x))$ is a polynomial threshold function of degree d, then there exists a decision tree of depth $\tau^{-1}(d\log(1/\tau))^{O(d)}$ such that a random root of this tree is τ-close to a τ-regular polynomial threshold function of degree d.*

Thus one can make a function τ regular by 'fixing' $\tau^{-1}(d\log(1/\tau))^{O(d)}$ different inputs, for most input values. [1] gets an analogous result which applies to the assumptions of the theorem above, i.e. that there is a decision tree of depth $\tau^{-1}(d\log(1/\tau))^{O(d)}$ such that with probability $1-\tau$, a random root either has a $(\tau, \tau^{-c}, O(1), O(\tau^M))$ regular decomposition, or has variance less than τ^M times the square of it's L^2_γ norm. To the former case, one can apply the theorem above by fixing variables. In the latter case, the function is roughly constant, i.e. it is incredibly highly concentrated, and thus can also be easily understood.

3 The Idea of the Proof

Due to space constraints, we only discuss the idea of the proof of Theorem 3. We emphasize the main principles upon which the proof lies, and the reason for such a poor dependence of m on d, c, and N, without introducing too much numerology, and also concentrating on the case where p is quadratic, since it is characteristic of the more complicated case.

The first principle is a heuristic that the author developed in a previous paper called *strong anticoncentration.* The result says that for a polynomial p, with high probability we have $p(X) \gtrsim \nabla p(X)$. Intuitively this is true because if $p(X)$ is significantly less than $\nabla p(X)$ at some value of X, a shift in the value of X will drastically effect $p(X)$, so that not many points will satisfy $p(X) \lesssim \nabla p(X)$ around this bad point. A significant part of this paper is extending this intuition to *tensors with polynomial coefficients.* For a k tensor $A = \sum A_S dx^{\otimes S}$ with low degree polynomial coefficients, the author shows that with a good probability,

$$|A_1 \otimes \cdots \otimes A_l| \gtrsim |\nabla A_1 \wedge \cdots \wedge \nabla A_l|$$

where we view $\nabla A_i = \sum D_j A_i$.

The main idea is the following. At any stage r of the algorithm, we have a decomposition $p \approx h(q_1, \dots, q_{m_r})$, though not necessarily a diffuse decomposition. Thus, unless our argument is complete, we can find $x \in \mathbb{R}^{m_r}$ such that the diffuse property does not hold for $\mathbb{P}(|q-x| \leq \varepsilon)$. By strong anticoncentration, with large probability we have

$$\prod_{i=1}^{m_r} |q_i(X) - x_i| \gtrsim |\nabla q_1 \wedge \cdots \wedge \nabla q_{m_r}|.$$

Thus with significant probability the quantity $|\nabla q_1 \wedge \cdots \wedge \nabla q_{m_r}|$ is small.

Now let's recall some multi-linear algebra. If a family of vectors $v_1, \dots, v_{m_r}$ is given, and $v_1 \wedge \cdots \wedge v_{m_r}$ is small, then this means these vectors are close to being linearly dependent. And since the $\{q_i\}$ are polynomials, this means, say, we can write q_1 as a function of the other q_j's, plus the products a_i and b_i introduced above, up to some small error. We then remove q_1 from the equation, and introduce the variables $\{a_i\}$ and $\{b_i\}$ into the family of q_i in the algorithm above at the next stage.

How do we ensure that keeping repeating this process will eventually give us the required decomposition? We associate with each stage of the algorithm a tuple of $d+1$ non-negative integers $(a_0, \dots, a_d)$. These integers change on each stage of the algorithm, but it is important that the associated polynomials are *decreasing* at each step, if we give the set of all such polynomials a linear ordering by defining $(a_0, \dots, a_d) \geq (b_0, \dots, b_d)$ if $a_0 \geq b_0$, or $a_0 = b_0$ and $a_1 \geq b_1$, or $a_0 = b_0$, $a_1 = b_1$, and $a_2 \geq b_2$, and so on, i.e. the dictionary ordering. Because the set of all tuples $(a_0, \dots, a_m)$ *has no infinite decreasing subsequence*, like for the non-negative integers, our algorithm must eventually terminate. But for $d > 0$, the number of steps before termination happens is unbounded, i.e. because at each stage of the algorithm we must decrease some a_i term, but we can increase the $a_{i+1}, \dots, a_d$ terms by an arbitrary amount. But we can be slightly careful about quantifying how much this happens, which gives the bounds on the number of iterations involved, and thus the implicit constants in the algorithm, but they still grow quite large in the parameters involved.

References

[1] Daniel Kane, *A Structure Theorem For Poorly Anticoncentrated Gaussian Chaos and Applications To The Study of Polynomial Threshold Functions*, FOCS. (2012), 91-100.

Jacob Denson, UW Madison
email: jcdenson@wisc.edu