18.218 Topics in Combinatorics Spring 2021 – Lecture 5

Dor Minzer

In this lecture, we will discuss one of the most important tools in analysis of Boolean functions, known as the hypercontractive inequality. In essence, the inequality tells us useful information about the behaviour of the values of low-degree real-valued functions on the hypercube.

Definition 0.1. *For any $p \geqslant 1$ we define the L^p norm of $f\colon \{-1,1\}^n \to \mathbb{R}$ as*

$$\|f\|_p = \left(\mathop{\mathbb{E}}_{x\sim\{-1,1\}^n}[|f(x)|^p]\right)^{1/p}.$$

1 Motivation – degree 1 functions

To begin the discussion, consider the case of degree 1 functions. That is, suppose we have a function $f\colon \{-1,1\}^n \to \mathbb{R}$ of the form $f(x) = \sum_{i=1}^{n} a_i x_i$, where we normalize the coefficients a_i so that $\sum_{i=1}^{n} a_i^2 = 1$. What can we say about the distribution of $f(x)$?

Well, by Parseval we clearly have that $\mathbb{E}_x\left[|f(x)|^2\right] = 1$, but because f has degree 1 we are able to say much more. Roughly speaking, if each one of the coefficients a_i is small individually, i.e. $|a_i| \leqslant \varepsilon$ for each i, then the distribution of $f(x)$ is similar to that of a standard Gaussian random variable $N(0,1)$. In particular, we are able to conclude that the moments of $f(x)$ are similar to those of a Gaussian random variable, i.e. $\mathbb{E}\left[|f(x)|^{2m}\right] \approx (2m-1)!!$

If, on the other hand some of the coefficients are large, then we can partition f into $H + L$, where H is the part of f with large coefficients (which in particular depends only on a few coordinates), and L which is the part of f with low coefficients (which then intuitively behaves like a Gaussian). Using this information, one again may prove moment bounds on f. In particular, a direct computation shows:

Lemma 1.1. *If $f\colon \{-1,1\} \to \mathbb{R}$ has degree 1, then $\|f\|_4 \leqslant \sqrt{3}\|f\|_2$. More generally, for any $q \geqslant 2$, $\|f\|_q \leqslant \sqrt{q-1}\|f\|_2$.*

This inequality is good enough for many purposes – for example, it is good enough in order to show that the value of $|f(x)|$ is relatively concentrated around $\|f\|_2$.

2 The hypercontractive inequality

The case that f is a higher degree function, even 2, is significantly more complex to understand. In particular, $f(x)$ need not behave like a Gaussian, even if all its coefficients are small in magnitude. Thus, our understanding of the distribution of $f(x)$ is significantly weaker. Yet, we can prove moment bounds for it (and get as a corollary tail bounds and such).

2.1 Low-degree function formulation

Theorem 2.1. *If $f\colon \{-1,1\} \to \mathbb{R}$ has degree d, and $q \geqslant 2$ then $\|f\|_q \leqslant \sqrt{q-1}^d \|f\|_2$.*

Proof. We will prove the statement for $q = 4$; a similar argument works for all even q (which by itself is good enough for all applications that we'll see). To prove the statement for other q's, one needs a different argument.

The proof proceeds by induction on d and n. The proof for $d = 0$ is trivial, so assume $d \geqslant 1$. We may take $i \in [n]$, and then write $x = (y, x_n)$ and $f(x) = g(y) + x_n h(y)$, where $g(y) = \sum_{S \not\ni i} \widehat{f}(S)\chi_S(y)$, and $h(y) = \partial_i f(y)$. Then

$$\mathop{\mathbb{E}}_{x}\left[f(x)^4\right] = \mathop{\mathbb{E}}_{y,x_n}\left[g(y)^4 + \binom{4}{1}g(y)^3 x_n h(y) + \binom{4}{2}g(y)^2 x_n^2 h(y)^2 + \binom{4}{3}g(y) x_n^3 h(y)^3 + x_n^4 h(y)^4\right].$$

Note that $x_n^2 = 1$, and $x_n = x_n^3$ have expectation 0, so

$$\|f\|_4^4 = \|g\|_4^4 + 6\mathop{\mathbb{E}}_{y}\left[g(y)^2 h(y)^2\right] + \|h\|_4^4.$$

By Cauchy-Schwarz, $\mathbb{E}_y\left[g(y)^2 h(y)^2\right] \leqslant \|g\|_4^2 \|h\|_4^2$, and by the inductive hypothesis $\|g\|_4 \leqslant \sqrt{3}^d \|g\|_2$ and $\|h\|_4 \leqslant \sqrt{3}^{d-1}\|h\|_2$. Plugging that in we get that

$$\|f\|_4^4 \leqslant 9^d \|g\|_2^4 + 6 \cdot 3^d \|g\|_2^2 \cdot 3^{d-1}\|h\|_2^2 + 9^{d-1}\|h\|_2^4 \leqslant 9^d(\|g\|_2^4 + 2\|g\|_2^2\|h\|_2^2 + \|h\|_2^4).$$

To finish the proof, note that $\|f\|_2^2 = \|g\|_2^2 + \|h\|_2^2$, and the right hand side about is just 9^d times the square of $\|f\|_2^2$. $\square$

2.2 Noise operator formulation

The hypercontractive inequality has yet another useful and equivalent formulation. To state it, we need to introduce the noise operator, T_ρ.

Definition 2.2. *Let $x \in \{-1,1\}^n$, and let $\rho \in [0,1]$. The distribution of ρ-correlated inputs with x, denoted as $y \sim \mathrm{T}_\rho x$, is defined as: for each $i \in [n]$ independently, set $y_i = x_i$ with probability ρ, and otherwise resample y_i according to the uniform distribution over $\{-1,1\}$.*

Intuitively, one may think of $y \sim \mathrm{T}_\rho x$ as a point obtained after performing a random walk of length $(1-\rho)n/2$ from x. With this definition in place, we can define the averaging operator T_ρ acting on functions, i.e. $\mathrm{T}_\rho : L^2(\{-1,1\}^n) \to L^2(\{-1,1\}^n)$, as follows: given $f\colon \{-1,1\}^n \to \mathbb{R}$, define

$$\mathrm{T}_\rho f(x) = \mathop{\mathbb{E}}_{y \sim \mathrm{T}_\rho x}[f(y)].$$

The source of the name the "hypercontractive inequality" really lies in this operator. First, one may easily show that for each $\rho \in [0,1]$, the operator T_ρ is a contraction – it can only shrink norms. That is, for all $q \geqslant 1$, $\|\mathrm{T}_\rho f\|_q \leqslant \|f\|_q$. It turns out that in fact a much stronger result holds

Theorem 2.3 (The hypercontractive inequality)**.** *For all $f\colon \{-1,1\} \to \mathbb{R}$, $1 \leqslant p \leqslant q$ and $0 \leqslant \rho \leqslant \sqrt{\frac{p-1}{q-1}}$ it holds that $\|\mathrm{T}_\rho f\|_q \leqslant \|f\|_p$.*

We will not include here the proof, and defer the interested reader to Ryan's book. It is a good exercise however to work out the case that $q = 4$ and $p = 2$, and see how to adapt the proof from the previous section to this case. For that, the following claim is useful, showing the effect of T_ρ on the Fourier transform.

Claim 2.4. *For all $f\colon \{-1,1\} \to \mathbb{R}$ and $\rho \in [0,1]$ we have that*

$$\mathrm{T}_\rho f(x) = \sum_{S\subseteq[n]} \rho^{|S|} \widehat{f}(S)\chi_S(x).$$

Proof. Note that as the operator T_ρ is linear, it is enough to show that for each character χ_S, it holds that $(\mathrm{T}_\rho\chi_S)(x) = \rho^{|S|}\chi_S(x)$. Indeed, note that

$$(\mathrm{T}_\rho\chi_S)(x) = \mathop{\mathbb{E}}_{y\sim \mathrm{T}_\rho x}[\chi_S(y)] = \mathop{\mathbb{E}}_{y\sim \mathrm{T}_\rho x}\left[\prod_{i\in S} y_i\right] = \prod_{i\in S} \mathop{\mathbb{E}}_{y_i\sim \mathrm{T}_\rho x_i}[y_i],$$

where the last transition is by independence. With probability ρ, we have $y_i = x_i$, and otherwise we resample y_i uniformly in $\{-1,1\}$, in which case the contribution to the expectation is 0 therefore, $\mathbb{E}_{y_i\sim \mathrm{T}_\rho x_i}[y_i] = \rho x_i$, and plugging it above finishes the proof. □

3 Hypercontractivity – basic applications

We begin by showing a few simple yet instructive applications of the hypercontractive inequality. In the next lecture we will see more substantial ones.

3.1 Small Set Expansion

Definition 3.1 (Noisy hypercube)**.** *For $\rho \in [0,1]$, the ρ-noisy hypercube graph is the graph on the vertex set $\{-1,1\}^n$, whose edges are sampled according to the T_ρ process. Namely, the distribution over the neighbours of a given vertex $x \in \{-1,1\}^n$ is given by $\mathrm{T}_\rho x$.*

Definition 3.2 (Edge expansion)**.** *Let $G = (V,E,w)$ be a weighted regular graph, and let $S \subseteq V$ be a vertex set. The expansion of S is defined as*

$$\Phi_G(S) = \Pr_{x\in S, y\sim_w N(x)}[y \notin S].$$

Expander graphs, that play important role in discrete mathematics, are graphs G in which $\Phi_G(S) \geqslant c$ for all subsets S containing at most half of the vertices of G, where c is some absolute constant. Intuitively, the larger c is the better the expansion and mixing of the graph is; however, since this is a requirement for all $|S| \leqslant |V|/2$, it is easily seen that one cannot hope to get c that is close to 1.

For this purpose, one sometimes considers the notion of small set expansion. Here, the point is that one may require that the expansion of sets much smaller than $n/2$ have expansion close to 1.

Definition 3.3. *A graph $G = (V,E,w)$ is called an (ε,δ)-small set expander if for any $S \subseteq V$ of size at most δn, it holds that $\Phi_G(S) \geqslant 1-\varepsilon$.*

Informally, when we say a graph is a small set expander, what we really mean is that we (often implicitly) have in mind a sequence of graphs $(G_n)_{n\in\mathbb{N}}$, such that for every $\varepsilon > 0$ there is $\delta > 0$ such that for large enough n, G_n is an (ε,δ)-small set expander.

Claim 3.4. *For $\rho = \frac{1}{\sqrt{3}}$, the noisy hypercube graph is a small-set expander.*

Proof. Fix $\delta > 0$, and let $S \subseteq \{-1,1\}^n$ be a set of vertices of size at most $\delta 2^n$. Let $f = 1_S$ be the indicator set of S. Note that

$$\langle 1_S, \mathrm{T}_\rho 1_S \rangle = \Pr_{\substack{x \sim \{-1,1\}^n \\ y \sim \mathrm{T}_\rho x}} [x \in S, y \in S],$$

so

$$\frac{1}{\mu(S)} \langle 1_S, \mathrm{T}_\rho 1_S \rangle = \Pr_{\substack{x \sim S \\ y \sim \mathrm{T}_\rho x}} [y \in S] = 1 - \Phi_G(S).$$

Thus, so show that $\Phi_G(S)$ is close to 1, we must upper bound the left hand side. For that, we use a useful Hölder-inequality trick:

$$\frac{1}{\mu(S)} \langle 1_S, \mathrm{T}_\rho 1_S \rangle \leqslant \frac{1}{\mu(S)} \|1_S\|_{4/3} \|\mathrm{T}_\rho 1_S\|_4 = \frac{1}{\mu(S)} \mu(S)^{3/4} \|\mathrm{T}_\rho 1_S\|_4 \leqslant \frac{1}{\mu(S)} \mu(S)^{3/4} \|1_S\|_2 = \mu(S)^{1/4},$$

which is at most $\delta^{1/4}$. In the penultimate inequality we used hypercontractivity. □

Remark 3.5. *There is nothing special about $\rho = 1/\sqrt{3}$, and the noisy hypercube is a small-set expander for any ρ bounded away from 1. The proof is an easy adaptation of the proof above, and is left to the reader.*

3.2 A concentration inequality for low-degree functions

One simple application of the hypercontractive inequality is a concentration bound for low-degree functions, which is somewhat similar to Chernoff's inequality for linear functions, and can be seen as a generalization of it for higher degrees.

Theorem 3.6. *Suppose $f\colon \{-1,1\}^n \to \mathbb{R}$ is a function of degree at most d, and let $t \geqslant 2^d$. Then*

$$\Pr_x [|f(x)| \geqslant t\|f\|_2] \leqslant e^{-\frac{t^{2/d}}{2}}.$$

Proof. Let $q \geqslant 2$ be a parameter to be chosen later. We have

$$\Pr_x [|f(x)| \geqslant t\|f\|_2] = \Pr_x [|f(x)|^q \geqslant t^q \|f\|_2^q] \leqslant \frac{\mathbb{E}_x [|f(x)|^q]}{t^q \|f\|_2^q} = \frac{\|f\|_q^q}{t^q \|f\|_2^q},$$

where we used Markov's inequality. By hypercontractivity, $\|f\|_q \leqslant \sqrt{q-1}^d \|f\|_2$, so we get that

$$\Pr_x [|f(x)| \geqslant t\|f\|_2] \leqslant \frac{\sqrt{q-1}^{dq} \|f\|_2^q}{t^q \|f\|_2^q} = e^{\frac{d}{2} q \log(q-1) - q \log t}.$$

Optimizing, we set $q = \frac{t^{2/d}}{2}$ and get that

$$\Pr_x [|f(x)| \geqslant t\|f\|_2] \leqslant e^{-\frac{t^{2/d}}{2}}.$$ □

Remark 3.7. *The exponent $t^{2/d}$ is tight.*

3.3 An anti-concentration for low-degree functions

On the other hand, one may ask if the value of $f(x)$ is non-trivial with non-trivial probability (as opposed to being 0 almost always, and very rarely huge). The following inequality asserts that this is not the case.

Theorem 3.8. *Suppose $f\colon \{-1,1\}^n \to \mathbb{R}$ is a function of degree at most d, and let $0<\theta<1$. Then*

$$\Pr_x\left[|f(x)| \geqslant \theta\|f\|_2\right] \geqslant \frac{(1-\theta^2)^2}{9^d}.$$

Proof. By definition,

$$\|f\|_2^2 = \mathop{\mathbb{E}}_x\left[f(x)^2\right] = \mathop{\mathbb{E}}_x\left[f(x)^2 1_{|f(x)|\geqslant\theta\|f\|_2}\right] + \mathop{\mathbb{E}}_x\left[f(x)^2 1_{|f(x)|\leqslant\theta\|f\|_2}\right].$$

We upper bound each expectation on the right hand side separately. For the first one, we use Cauchy-Schwarz:

$$\mathop{\mathbb{E}}_x\left[f(x)^2 1_{|f(x)|\geqslant\theta\|f\|_2}\right] \leqslant \mathop{\mathbb{E}}_x\left[f(x)^4\right]^{1/2} \mathop{\mathbb{E}}_x\left[1^2_{|f(x)|\geqslant\theta\|f\|_2}\right]^{1/2} = \|f\|_4^2\sqrt{\Pr\left[|f(x)| \geqslant \theta\|f\|_2\right]}.$$

Using hypercontractivity now we get that $\|f\|_4 \leqslant \sqrt{3}^d\|f\|_2$, so the second expectation is at most

$$3^d\|f\|_2^2\sqrt{\Pr\left[|f(x)| \geqslant \theta\|f\|_2\right]}.$$

For the second expectation, clearly

$$\mathop{\mathbb{E}}_x\left[f(x)^2 1_{|f(x)|\geqslant\theta\|f\|_2}\right] \leqslant \theta^2\|f\|_2^2.$$

Plugging the two estimates above, we get

$$\|f\|_2^2 \leqslant 3^d\|f\|_2^2\sqrt{\Pr\left[|f(x)| \geqslant \theta\|f\|_2\right]} + \theta^2\|f\|_2^2,$$

and rearranging yield that $\Pr\left[|f(x)| \geqslant \theta\|f\|_2\right] \geqslant \frac{(1-\theta^2)^2}{9^d}$. □

3.4 The 1-norm trick

Theorem 2.1 tells us that for low-degree functions f, the q-norm of f is comparable to the 2-norm of f. This raises the question of whether one can more generally relate the q-norm of f to the p-norm of f for any $q>p\geqslant 1$ in this case. If $p\geqslant 2$, then $\|f\|_2\leqslant\|f\|_p$, so one gets that for free. It turns out that one can get a result for any $p\geqslant 1$ using a simple trick.

Lemma 3.9. *Let $f\colon\{-1,1\}^n\to\mathbb{R}$ be a function of degree at most d. Then $\|f\|_2\leqslant 3^d\|f\|_1$.*

Proof. Note that by Hölder's inequality

$$\|f\|_2^2 = \mathop{\mathbb{E}}_x\left[|f(x)|^{4/3}\,|f(x)|^{2/3}\right] \leqslant \mathop{\mathbb{E}}_x\left[|f(x)|^4\right]^{1/3}\mathop{\mathbb{E}}_x\left[|f(x)|\right]^{2/3} = \|f\|_4^{4/3}\|f\|_1^{2/3}.$$

By Theorem 2.1, $\|f\|_4 \leqslant \sqrt{3}^d\|f\|_2$, and plugging that in gives that $\|f\|_2^2 \leqslant \sqrt{3}^{4d/3}\|f\|_2^{4/3}\|f\|_1^{2/3}$. Rearranging finishes the proof. □

3.5 Next lecture

In the next lecture we will show more applications of the hypercontractive inequality.

Degree 1 functions that are close to Boolean. Recall that in the HW assignment, you have proved that degree 1 functions that are Boolean can only be dictatorships (or anti-dictatorships). What can one say if a degree 1 function $f(x) = \sum_{i=1}^{a_i x_i}$ is nearly Boolean, i.e. close to a Boolean function in L^2 distance?

The Fourier spectrum of small sets. What can one say regarding the Fourier spectrum of small-sets? Can their indicator function be a low-degree polynomial (or "close" to one)? We will study this question; you are encouraged to think of how would such statement align with the small-set expansion property we have seen in this lecture.

The KKL theorem and the Friedgut Junta theorem. Moving on from low-degree functions, one may ask about the structure of functions that have small average degree, i.e. $I[f] \leqslant K$ for K thought of as small. What can one prove about such functions? How does it all relate to the study of low-degree functions?

18.218 Topics in Combinatorics: Analysis of Boolean Functions
Spring 2021