Lecture 11: Concentration on Influential Variables

Instructor: Dieter van Melkebeek Scribe: Seeun Umboh

Today we show some applications of the Hypercontractivity Theorem that we proved last lecture, involving the concentration of Boolean functions on their influential variables.

1 Hypercontractivity of the Noise Operator

Last time, we proved that the noise operator T_α is hypercontractive. Recall that the noise operator $T_\alpha f$ is defined as

$$(T_\alpha f)(x) = \mathrm{E}_{y \sim_\epsilon x}[f(y)] \tag{1}$$

where $\epsilon = \frac{1-\alpha}{2}$, and $y \sim_\epsilon x$ means that we choose y_i to be x_i with probability $1-\epsilon$ and $-x_i$ with probability ϵ. We have also shown that it has Fourier expansion $T_\alpha f = \sum_{S \subseteq [n]} \alpha^{|S|} \hat{f}(S) \chi_S$.

The hypercontractivity of the noise operator states that not only is the norm of $T_\alpha f$ is no more than the norm of f, but that this holds even if we use the q-norm for $T_\alpha f$ and the p-norm for f where $p < q$. The latter is a stronger statement since the q-norm of a function is no smaller than its p-norm if $p < q$. Formally,

Theorem 1 (Hypercontractivity)**.** *For all p, q, α such that $1 \leq p \leq q \leq \infty$ and $|\alpha| \leq \sqrt{\frac{p-1}{q-1}}$, and for all functions $f : \{-1,1\}^n \to \mathbb{R}$*

$$||T_\alpha g||_q \leq ||g||_p.$$

Today, we will use a corollary of this where $q = 2$, $p = 1 + \alpha^2$.

Corollary 1. *For all $-1 \leq \alpha \leq 1$ and functions $f : \{-1,1\}^n \to \mathbb{R}$*

$$||T_\alpha g||_2 \leq ||g||_{1+\alpha^2}.$$

The corollary allows us to prove a lemma which we will use to prove the following theorems. The first of these deals with approximating Boolean functions by (small) juntas.

Theorem 2. *For all Boolean functions f, and for all $\epsilon > 0$, f is ϵ-close to an r-junta, where $r = 2^{O(I(f)/\epsilon)}$.*

We will also prove a lower bound on the maximum influence of a variable,

Theorem 3. *For all balanced functions f, there exists a variable index i such that*

$$I_i \geq \Omega\left(\frac{\log n}{n}\right).$$

Note that this is stronger than the trivial bound of $1/n$ for balanced functions, and that it is tight – recall that we have shown there exists such a variable for the Tribes function.

Also note that one approach to proving theorem 2 is by successively applying a generalized version of theorem 3. Intuitively, we can use the latter to obtain some variables of high influence, and use these to form the r-junta. We will use a different approach, as we derive both results as corollaries to a key main lemma. The lemma states that functions of low influence are concentrated on small subsets consisting only of influential variables. It is a strengthening of our previous results on learning Boolean functions through Fourier analysis, which only showed that we can focus on small subsets.

2 Main Lemma

Recall that if f has small influence $I(f)$, then we can approximate it well by only caring about χ_S for which S is small, say $|S| \leq d$. That is,

$$\sum_{|S|\geq d} \left(\hat{f}(S)\right)^2 \leq \frac{1}{d} \sum_{S\subseteq[n]} |S| \left(\hat{f}(S)\right)^2 = \frac{I(f)}{d} \leq \epsilon, \tag{2}$$

where the last inequality holds provided that $d \leq \frac{I(f)}{\epsilon}$. So if $I(f)$ is small, d does not need to be big to get a small ϵ.

This was what we used to develop learning algorithms for functions with low influence such as DNFs, and constant-depth circuits. The main lemma we prove today shows that, in fact, not only can we restrict our attention to the small subsets but we only need to consider subsets that contain variables of high influence. In particular, we will show that with appropriate settings of d and threshold τ, it is sufficient to consider only subsets $S \subseteq R$ of size $|S| \leq d$ where $R = \{i \in [n] : I_i(f) \geq \tau\}$, the set of *relevant* variables with influence at least τ.

That is, we want to show that for $\mathcal{S} = \{S \subseteq R : |S| \leq d\}$ where d is small and τ is relatively large,

$$\sum_{S\notin\mathcal{S}} \left(\hat{f}(S)\right)^2 \leq \epsilon. \tag{3}$$

Now, if $S \notin \mathcal{S}$, then S is either a big subset, or a small subset containing at least one variable of small influence. So, since we have already bounded the contribution by big subsets with (2), what remains is bounding the contribution by the latter, and this is where the Hypercontractivity Theorem comes in. In particular, using Corollary 1, we can bound the contribution of small subsets containing variable i

$$\sum_{\substack{S\ni i\\|S|\leq d}} \left(\hat{f}(S)\right)^2 \tag{4}$$

by a power of I_i. Summing this over the irrelevant variables $i \notin R$, we get a bound on the contribution of small subsets containing irrelevant variables. Combining this with (2), we get the desired bound (3).

2.1 Small Subsets Containing Low Influence Variables

Recall that $I_i = ||D_i f||_p^p = \mathrm{E}[|D_i|]$, for each $p \geq 1$. Let $g = D_i f$. Using the definition of norms, we know that

$$||g||^2_{1+\alpha^2} = \left(||g||^{1+\alpha^2}_{1+\alpha^2}\right)^{\frac{2}{1+\alpha^2}} = I_i^{\frac{2}{1+\alpha^2}}. \tag{5}$$

We also have

$$||T_\alpha g||_2^2 = \sum_{S\subseteq[n]} \left(\alpha^{|S|}\hat{g}(S)\right)^2 = \sum_{S\ni i} \left(\alpha^{|S|-1}\hat{f}(S)\right)^2, \tag{6}$$

where the first equality follows from the Fourier expansion of $T_\alpha g$, and the second from the fact that $I_i = \sum_{S\ni i}\left(\hat{f}(S)\right)^2$, which we established in lecture 6.

Now, Corollary 1 tells us that (6) $\leq$ (5), and so

$$\sum_{S\ni i} \left(\alpha^{|S|-1}\hat{f}(S)\right)^2 \leq I_i^{\frac{2}{1+\alpha^2}}. \tag{7}$$

Note that since $|\alpha| \leq 1$, most of the contribution to the LHS of (7) comes from small subsets containing i, and is bounded by the RHS of (7) which is a power of the influence of i.

Using (7), we get that for any variable i, the contribution of small subsets containing i is bounded by

$$\sum_{\substack{S\ni i\\|S|\leq d}} \left(\hat{f}(S)\right)^2 \leq \left(\frac{1}{\alpha^2}\right)^{d-1} \sum_{S\ni i|S|\leq d} \left(\alpha^{|S|-1}\hat{f}(S)\right)^2 \leq \left(\frac{1}{\alpha^2}\right)^{d-1} I_i^{\frac{2}{1+\alpha^2}} \tag{8}$$

where the last inequality follows from (7).

This allows us to bound the contribution of small subsets containing irrelevant variables,

$$\sum_{\substack{S\not\subseteq R\\|S|\leq d}} \left(\hat{f}(S)\right)^2 \leq \sum_{i\notin R}\sum_{\substack{S\ni i\\|S|\leq d}} \left(\hat{f}(S)\right)^2 \leq \left(\frac{1}{\alpha^2}\right)^{d-1} \sum_{i\notin R} I_i^{\frac{2}{1+\alpha^2}} \leq \left(\frac{1}{\alpha^2}\right)^{d-1} \tau^{\frac{1-\alpha^2}{1+\alpha^2}} \cdot I(f) \tag{9}$$

The first inequality is from the fact that if $S \not\subseteq R$, then S contains at least one element not in R, and so S appears on the RHS of that at least once. The second is from (8). The third is because

$$\sum_{i\notin R} I_i \cdot I_i^{\frac{1-\alpha^2}{1+\alpha^2}} \leq \tau^{\frac{1-\alpha^2}{1+\alpha^2}} \cdot \sum_{i\notin R} I_i \leq \tau^{\frac{1-\alpha^2}{1+\alpha^2}} \cdot I(f).$$

2.2 Combining Bounds

In the previous section, we established a bound on the contribution to the Fourier spectrum due to small sets containing at least 1 variable of low influence. Now, we are ready to combine our previous result bounding the contribution of large sets and our new result to try to get equation (3). Since sets not in $\mathcal{S}$ are either large or they are small and contain at least 1 variable of low influence, we have

$$\sum_{S\notin\mathcal{S}} \left(\hat{f}(S)\right)^2 \leq \underbrace{\sum_{|S|\geq d} \left(\hat{f}(S)\right)^2}_{(*)} + \underbrace{\sum_{\substack{S\not\subseteq R\\|S|\leq d}} \left(\hat{f}(S)\right)^2}_{(**)} \leq \epsilon$$

where the last inequality holds provided that both $(*)$, $(**)$ are at most $\epsilon/2$. We know that $(*) \leq \frac{I(f)}{d}$ by (2) and $(**) \leq \left(\frac{1}{\alpha^2}\right)^{d-1} \tau^{\frac{1-\alpha^2}{1+\alpha^2}} I(f)$ by (9). To get $(*) \leq \epsilon/2$, it is sufficient to have $d \geq \frac{2I(f)}{\epsilon}$. To get $(**) \leq \epsilon/2$, it is sufficient to have $d\left(\frac{1}{\alpha^2}\right)^{d-1}\tau^{\frac{1-\alpha^2}{1+\alpha^2}} \leq 1$. Rewriting the last inequality, we need $\tau \leq \left(\frac{1}{d}(\alpha^2)^{d-1}\right)^{\frac{1+\alpha^2}{1-\alpha^2}}$. Since $1/2^{d-1} \leq 1/d$, it suffices to have $\tau \leq \left((\alpha^2/2)^{d-1}\right)^{\frac{1+\alpha^2}{1-\alpha^2}}$. So, rearranging terms, this tells us that we need to choose $\tau \leq \beta^d$ for some β, where $0 < \beta < 1$, depending on α. We will not attempt to optimize β.[1]

To recap, we have shown that in order to ϵ-approximate Boolean functions f with low influence, we only need to consider Fourier coefficients of sets of size at most d and whose variables have influence at least τ. This is formalized in the following lemma

Lemma 1 (Main Lemma)**.** *There exists a constant $0 < \beta < 1$, such that for all $\epsilon > 0$,*

$$\sum_{S \notin \mathcal{S}} (\hat{f}(S))^2 \leq \epsilon,$$

where $\mathcal{S} = \{S \subseteq R : |S| \leq d\}$ and $R = \{i \in [n] : I_i(f) \geq \tau\}$ and $d = \frac{2I(f)}{\epsilon}$ and $\tau = \beta^d$.

Note that since for each $i \in R$, we have $I_i(f) \geq \tau$ and $I(f) = \sum_i I_i(f)$, we can upper bound the size of R by

$$|R| \leq \frac{I(f)}{\tau} = \frac{I(f)}{\beta^d} \leq 2^{O(\frac{I(f)}{\epsilon})}. \tag{10}$$

We also note that this lemma is a stronger result than the one we used for learning functions with low influence.

3 Applying The Lemma

We now use the Main Lemma to prove the two results we promised earlier.

3.1 Approximation By Juntas

First, we use the Main Lemma to obtain a real function g on the Boolean cube where

$$g = \sum_{S \in \mathcal{S}} \hat{f}(S)\chi_S,$$

and $\sum_{S \notin \mathcal{S}} \left(\hat{f}(S)\right)^2 \leq \epsilon$. Then, to approximate f, we use $h(x) = sign(g(x))$.

The Main Lemma tells us that for g, we lose no more than ϵ of the Fourier spectrum, and using an argument similar to the one we used for learning with samples, we can show that if we take the sign of g, this gives us a Boolean function at most ϵ away from f. Using that argument we have

$$\begin{aligned}
\Pr_x[h(x) \neq f(x)] &\leq \Pr_x[|g(x) - f(x)| \geq 1] \\
&\leq \mathrm{E}_x[(g(x) - f(x))^2] \\
&= \sum_{S \notin \mathcal{S}} \left(\hat{f}(S)\right)^2 \\
&\leq \epsilon
\end{aligned}$$

Since g depends only on variables in R, this gives us a $|R|$-junta ϵ-approximation of f, where (10) gives us the bound on $|R|$ we need.

[1] This means we really only need the fact that the noise operator is hypercontractive for $q = 2$, but don't care about the precise parameters.

3.2 Lower Bound on Maximum Influence of a Variable

We will actually prove a generalized version of theorem 3, since we will need it for our application to social choice theory in the next lecture. Instead of proving the result for balanced functions, we prove it for general Boolean functions.

Theorem 4 (Generalized Theorem 3)**.** *There exists $c > 0$, such that for any $f : \{-1,1\}^n \to \{-1,1\}$,*

$$\exists i \in [n] : I_i \geq c \cdot \sigma^2(f) \cdot \frac{\log n}{n}.$$

Note that for balanced function $\sigma^2(f) = 1$, giving us theorem 3. Also, note that the $I(f) \geq \sigma^2(f)$, so without the log factor, the theorem is trivially true.

Before we start with the proof, we give a sketch of the proof. It follows by taking the better of the trivial lower bound of $I(f)/n$ for the maximum(over $i \in [n]$) of I_i, and a similar bound obtained by applying the Main Lemma and focusing on the relevant variables. If f has low influence, such that (10) gives a good bound on the number of relevant variables $|R|$, then the latter is better. On the other hand, if f has high influence, the former gives a better bound.

Proof. Trivially, we know that there exists an $i \in [n]$ with $I_i \geq I/n$. By the Main Lemma, if the total influence of f is small, we can restrict ourselves to a small number of relevant variables while maintaining most of the spectrum and total influence. The sum of the influences of the relevant variables divided by the number of relevant variables, gives us a lower bound for the maximum influence of a variable.

We lower bound the total influence of the relevant variables as follows

$$\sum_{i \in R} I_i = \sum_{i \in R} \sum_{S \ni i} (\hat{f}(S))^2 \geq \underbrace{\sum_{S \neq \emptyset} (\hat{f}(S))^2}_{(+)} - \underbrace{\sum_{S \not\subseteq R} (\hat{f}(S))^2}_{(++)} \geq \sigma^2(f) - \epsilon,$$

where the second inequality is because every non-empty set that is fully contained in R appears in the LHS at least once, and the third because we proved in lecture 6 that $(+) = \sigma^2(f)$ and an application of the Main Lemma gives us $(++) \leq \epsilon$, for some ϵ that we will determine later.

Using the lower bound and the fact that $|R| \leq 2^{O(I(f)/\epsilon)}$, we have that there exists $i \in R$ such that

$$I_i \geq \frac{\sum_{j \in R} I_j}{|R|} \geq \frac{\sigma^2(f) - \epsilon}{2^{O(I(f)/\epsilon)}}$$

Now, we want to maximize the RHS of the inequality. We are allowed to tweak ϵ, the parameter for our application of the Main Lemma. Both the numerator and the denominator increase as ϵ decreases, but the denominator depends exponentially on ϵ. So we pick $\epsilon = \sigma^2(f)/2$, because then we lose only a constant fraction in the numerator and the denominator remains small.

Now, we have two bounds for I_i in terms of I:

$$I_i \geq \frac{\sigma^2(f)}{2^{bI(f)/\sigma^2(f)}}, \tag{11}$$

for some positive constant b, and

$$I_i \geq I/n. \tag{12}$$

Let us now compare the two bounds. When I is big, (11) does not give us anything good since the denominator will be huge and so the resulting lower bound is smaller than what we get if we use (12). On the other hand, when I is small, the denominator of (11) is much smaller than n, and (11) gives us the better lower bound. The break-even point roughly occurs when the denominator in (11) equals the one in (12), i.e., when $\frac{I(f)}{\sigma^2(f)} = \Theta(\log n)$.

Let us make this reasoning rigorous. If $\frac{I(f)}{\sigma^2(f)} \geq c \cdot \log n$, then (12) immediately gives us a lower bound of $c \cdot \sigma^2(f) \cdot \frac{\log n}{n}$. On the other hand, if $\frac{I(f)}{\sigma^2(f)} \leq c \cdot \log n$ and $c < 1/b$, then $2^{bI(f)/\sigma^2(f)} < n/\log n$, so by (11) we have that $I_i \geq \sigma^2(f) \cdot \frac{\log n}{n}$.

So, we have shown that taking the better of (12) and (11) gives us $I_i = \Omega\left(\sigma^2(f) \cdot \frac{\log n}{n}\right)$, as desired. $\square$