18.218 Topics in Combinatorics Spring 2021 – Lectures 8 - 10

Dor Minzer

In this lecture, we will introduce the notion of noise stability. We will asymptotically calculate the noise stability of majority, as well as use it to prove Arrow's impossibility theorem from social choice theory.

1 Noise stability and noise sensitivity

Consider the voting scheme interpretation we had for Boolean functions, wherein n-voters cast their votes $x = (x_1, \dots, x_n)$, and a function $f\colon \{-1,1\}^n \to \{-1,1\}$ is used to aggregate their votes to a final decision. It is often the case that the votes that we get are not the actual votes, but rather a noisy version of them $y = (y_1, \dots, y_n)$ (due to errors in the communication channel for example), wherein $y_i = x_i$ for $1-\varepsilon$ fraction of the voters, but y_i may be different for an ε fraction of the voter. We would like this noise to be unlikely to affect the final decision, i.e. we would like that $f(x) = f(y)$ with as high probability as possible. What functions satisfy this? Towards this end, we define the notions of noise stability.

Definition 1.1 (Noise Stability)**.** *Let $f\colon \{-1,1\}^n \to \mathbb{R}$, and $\rho \in [0,1]$. The stability of f with parameter ρ is defined as $\mathsf{Stab}_\rho(f) = \langle f, \mathrm{T}_\rho f\rangle$.*

For ± 1 valued functions, we get that

$$\begin{aligned}\mathsf{Stab}_\rho(f) = \mathop{\mathbb{E}}_{(x,y)\ \rho\ \text{correlated}}[f(x)f(y)] &= \Pr_{(x,y)\ \rho\ \text{correlated}}[f(x)=f(y)] - \Pr_{(x,y)\ \rho\ \text{correlated}}[f(x)\neq f(y)] \\ &= 2\Pr_{(x,y)\ \rho\ \text{correlated}}[f(x)=f(y)] - 1,\end{aligned}$$

so indeed functions with high noise stability are precisely functions for which the noise is the least likely to make change. Let us think for a moment that $\rho = 1-\varepsilon$, and that f is a balanced Boolean function. Among this class of functions, what is the most stable function there is?

Intuitively, the more coordinates f depends on the more likely it is to change its value due to noise. I.e., one may expect that among all balanced Boolean functions, dictators would maximize noise stability, and this is indeed the case. We will establish this fact soon.

How about functions that very much do not look like dictatorships? It turns out that then, majority is the best there is.

Theorem 1.2. *For all $\rho > 0$, $\delta > 0$ there is $\tau > 0$ such that if $f\colon \{-1,1\}^n \to \{-1,1\}$ is balanced, and $I_i[f] \leqslant \tau$ for all $i \in [n]$, then*
$$\mathsf{Stab}_\rho(f) \leqslant \mathsf{Stab}_\rho(\mathsf{Majority}) + \delta.$$

We will prove this theorem in a few lectures, and see one application for it (which was in fact the original motivation for the formulation of this result) in hardness of approximation. In this lecture and the next one, we shall focus our attention on several more basic applications of stability.

In particular, today we will compute the noise stability of majority asymptotically, which will give us some geometric intuition about the problem and exhibit the relation to Gaussian space that we will elaborate on later on in the course. We will then give an unrelated early application of noise stability in social choice theory, which is Kalai's proof of Arrow's impossibility theorem.

To get started, let's get a Fourier analytic formula for the noise stability of a function.

Claim 1.3. *Let* $f\colon \{-1,1\}^n \to \mathbb{R}$, $\rho \in [0,1]$. *Then* $\mathsf{Stab}_\rho(f) = \sum_S \rho^{|S|} \widehat{f}(S)^2$.

Proof. Plancherel. $\square$

Now that we have a Fourier analytic expression for stability and noise sensitivity, we can note that dictators indeed maximize the stability among all balanced functions.Indeed, it follows that $\mathsf{Stab}_\rho(f) \leqslant \rho$, and equality is achieved if and only if f is degree 1 (in which case f must be a dictatorship).

1.1 The noise stability of Majority

Let $f(x_1,\dots,x_n) = \mathsf{Majority}(x_1,\dots,x_n) = \mathsf{sign}\left(\frac{x_1+\ldots+x_n}{\sqrt{n}}\right)$, $\rho \in [0,1]$. Throughout, we let x, y be sampled according to the ρ-correlated distribution over $\{-1,1\}^n$, and our goal is to compute $\mathbb{E}_{x,y}\left[f(x)f(y)\right]$. We shall be somewhat imprecise, as our main goal is to deliver geometric intuition.

Moving to Gaussians. Consider the distributions of $X = \frac{x_1+\ldots+x_n}{\sqrt{n}}$ and $Y = \frac{y_1+\ldots+y_n}{\sqrt{n}}$. By the central limit theorem, X and Y each convergence in distribution to a standard random variable $N(0,1)$, say $X \to G_1$, $Y \to G_2$. By the multi-dimensional version of the CLT, since the covariance of X, Y is

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i + \frac{1}{n}\sum_{i\neq j} x_i y_j\right] = \rho,$$

the covariance of G_1, G_2 is also ρ. Hence, we get that

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{x,y}\left[f(x)f(y)\right] = \mathop{\mathbb{E}}_{G_1,G_2}\left[\mathsf{sign}(G_1)\mathsf{sign}(G_2)\right]$$

The computation in Gaussian space. There is a neat what to compute the last expectation which we next present. How can we generate standard Gaussians G_1, G_2 whose covariance is ρ? Well, one way to do so it to take a multi-dimensional standard Gaussian, say $h \sim N(0, I)$ in $\mathbb{R}^2$, take $u, v \in \mathbb{R}^2$ unit vectors whose inner product is ρ, and define $G_1 = \langle u, h\rangle$, $G_2 = \langle v, h\rangle$. Indeed, it is a standard fact that G_1, G_2 are distributed as Gaussians in this case, and

$$\mathbb{E}\left[G_1 G_2\right] = \sum_{i=1}^{n} u_i v_i h_i^2 + \sum_{i\neq j} u_i v_j h_i h_j = \langle u, v\rangle = \rho.$$

Thus,

$$\mathop{\mathbb{E}}_{G_1,G_2}\left[\mathsf{sign}(G_1)\mathsf{sign}(G_2)\right] = 1 - 2\Pr\left[\mathsf{sign}(G_1) \neq \mathsf{sign}(G_2)\right] = 1 - 2\Pr_h\left[\mathsf{sign}(\langle u, h\rangle) \neq \mathsf{sign}(\langle v, h\rangle)\right].$$

Let us think of this geometrically now. Consider the normal lines to u and v, call them ℓ_u, ℓ_v. Note that ℓ_u divides the plane into vectors that give positive inner product with u, and vectors that give negative inner product with u, and the same goes for v. Thus, the region between these lines is precisely the region of vectors that give different signs in these inner products. As the angle between ℓ_u and ℓ_v is the same, the probability a random Gaussian vector falls in this region is $\frac{2\sphericalangle(u,v)}{2\pi} = \frac{\mathsf{arccos}(\langle u,v\rangle)}{\pi} = \frac{\mathsf{arccos}(\rho)}{\pi}$. Thus, we get that

$$\mathop{\mathbb{E}}_{G_1,G_2}\left[\mathsf{sign}(G_1)\mathsf{sign}(G_2)\right] = 1 - \frac{2}{\pi}\mathsf{arccos}(\rho).$$

We thus get:

Theorem 1.4 (Sheppards Formula)**.** $\mathsf{Stab}_\rho(\mathsf{Majority}_n) = 1 - \frac{2}{\pi}\mathsf{arccos}(\rho) + o(1)$.

Exercise to think about: does this formula tell you anything about the Fourier weight distribution of Majority? What is $W^{\geqslant k}[\mathsf{Majroity}]$ asymptotically in k?

2 Arrow's impossibility theorem

2.1 Set up

Suppose we have elections between 3 candidates, A, B and C. Each one of n voters must declare their ranking among the 3 candidates. We would like to interpret it as a collection of bits, and thus encode the vote of a participant i as $x_i, y_i, z_i \in \{-1, 1\}$, where $x_i = 1$ if $A > B$ in their eyes (i.e. if they prefer A over B), and otherwise $x_i = -1$; $y_i = 1$ if $B > C$, and otherwise $y_i = -1$; $z_i = 1$ if $C > A$, and otherwise $z_i = -1$. We note that then a vote of a participant is a vector from

$$\{(1, 1, -1), (1, -1, 1), (-1, 1, 1), (1, -1, -1), (-1, 1, -1), (-1, -1, 1)\},$$

as the votes $(1, 1, 1)$, $(-1, -1, -1)$ do not represent a valid ranking. In other words, these are all of the assignments in the support of $\mathsf{NAE}_3 \colon \{-1, 1\}^3 \to \{0, 1\}$ defined as $\mathsf{NAE}_3(a, b, c) = 1 - 1_{a=b=c}$.

Given the vectors $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, $z = (z_1, \ldots, z_n)$ that encode the preferences of the votes, we wish to use a function $f \colon \{-1, 1\}^n \to \{-1, 1\}$ in order to determine the overall ranking between the 3 candidates. To do that, we will compute $f(x)$ to determine the preference between A and B, and similarly compute $f(y)$ and $f(z)$ to determine the preference between B and C, A and C respectively. We thus get $(f(x), f(y), f(z)) \in \{-1, 1\}^3$, and for that to represent a valid ranking we must have that $\mathsf{NAE}_3(f(x), f(y), f(z)) = 1$. In this case, we say that the elections have a *Condorcet winner*.

Condorcet himself noted that the majority function does not always yield a Condorcet winner, which raises the question of whether there is a voting rule f that avoids such paradoxes.

Theorem 2.1. *[Arrow's impossibility theorem] Suppose $f \colon \{-1, 1\}^n \to \{-1, 1\}$ is an unanimous voting rule, i.e. such that $f(\vec{1}) = 1$, $f(-\vec{1}) = -1$. If in 3-candidate election f always has a Condorcet winner, then f is a dictatorship.*

2.2 Preliminary facts

To prove Theorem 2.1, we first need the Fourier expansion of NAE_3.

Claim 2.2. $\mathsf{NAE}_3(a, b, c) = \frac{3}{4} - \frac{1}{4}(ab + bc + ac)$.

Proof.

$$\begin{aligned}
\mathsf{NAE}_3(a, b, c) &= 1 - \mathsf{Eq}(a, b, c) \\
&= 1 - \frac{(1-a)(1-b)(1-c)}{8} - \frac{(1+a)(1+b)(1+c)}{8} \\
&= 1 - \frac{1 - a - b - c + ab + ac + bc - abc}{8} - \frac{1 + a + b + c + ab + ac + bc + abc}{8} \\
&= \frac{3}{4} - \frac{1}{4}(ab + bc + ac).
\end{aligned}$$

□

We also need to extend the definition of ρ-correlated inputs to negative ρ's.

Definition 2.3. *Let $-1 \leqslant \rho < 0$. The distribution of ρ-correlated inputs is defined as the joint distribution of $(a, b) \in \{-1, 1\}^2$ such that marginally each one of a and b is distributed uniformly, and $\mathbb{E}[ab] = \rho$.*

An alternative way to define this distribution, more along the lines of our definition of $\rho \geqslant 0$, is to say that given a, the distribution of ρ-correlated inputs with a is the distribution that sampled a $-\rho$-correlated input with $-a$, i.e. with probability $-\rho$ we take $b = -a$, and otherwise we resample $b \in \{-1, 1\}$. It is easy to check that the two definitions coincide.

Just as in the case of $\rho > 0$, we can define the averaging operator T_ρ according to the ρ-correlated distribution, and the stability $\mathsf{Stab}_\rho = \mathbb{E}_{(x,y)\ \rho\text{-correlated}}[f(x)f(y)] = \langle f, \mathrm{T}_\rho f \rangle$.

2.3 Proof of Theorem 2.1

Let us sample $(x_i, y_i, z_i) \sim \mathsf{NAE}_3^{-1}(1)$ for each $i = 1, \dots, n$ independently. By assumption, we always have that $\mathsf{NAE}_3(f(x), f(y), f(z)) = 1$, and we next compute the expectation of it in a different way. Using Claim 2.2,

$$\begin{aligned}\mathop{\mathbb{E}}_{x,y,z}[\mathsf{NAE}_3(f(x), f(y), f(z))] &= \mathop{\mathbb{E}}_{x,y,z}\left[\frac{3}{4} - \frac{1}{4}(f(x)f(y) + f(y)f(z) + f(x)f(z))\right] \\ &= \frac{3}{4} - \frac{3}{4}\mathop{\mathbb{E}}_{x,y,z}[f(x)f(y)],\end{aligned}$$

where in the last equality we used symmetry. We now inspect the joint distribution of x, y. Clearly, (x_i, y_i) are independently picked for each i, and inspecting the marginal of each one of x_i, y_i, we see that they are uniformly distributed. Also,

$$\mathbb{E}[x_i y_i] = \frac{1}{6}(2 - 4) = -\frac{1}{3},$$

so (x_i, y_i) is ρ-correlated with $\rho = -\frac{1}{3}$. Thus, $\mathbb{E}_{x,y,z}[f(x)f(y)] = \mathsf{Stab}_{-1/3}(f)$, and we get the identity

$$1 = \mathop{\mathbb{E}}_{x,y,z}[\mathsf{NAE}_3(f(x), f(y), f(z))] = \frac{3}{4} - \frac{3}{4}\mathsf{Stab}_{-1/3}(f).$$

Rearranging, we get that $\mathsf{Stab}_{-1/3}(f) = -\frac{1}{3}$. We now use Claim 1.3 to extract from this information about the Fourier spectrum of f. Note that

$$-\frac{1}{3} = \mathsf{Stab}_{-1/3}(f) = \sum_{k=0}^{n}\left(-\frac{1}{3}\right)^k W^{=k}[f],$$

and that trivially

$$\sum_{k=0}^{n}\left(-\frac{1}{3}\right)^k W^{=k}[f] \geqslant -\frac{1}{3}\sum_{k=0}^{n} W^{=k}[f] = -\frac{1}{3}\|f\|_2^2 = -\frac{1}{3}.$$

Furthermore, unless all of the Fourier weight of f likes on the first level (the only level which is multiplied by $(-1/3)$ and not something larger), this inequality is strict. Thus, we conclude that $W^{=1}[f] = 1$, and by the homework exercise f is either a dictatorship or an anti-dictatorship. The unanimity now implies it is a dictatorship, finishing the proof. □

2.4 Robust Arrow's theorem

The proof we have just seen is not the original proof of Arrow, and was given by Gil Kalai in 2002. It has the added benefit that it is able to establish a more robust version of the result, that reads as follows.

Theorem 2.4. *Suppose $f\colon \{-1,1\}^n \to \{-1,1\}$ is a voting rule such that the probability of reaching a Condorcet paradox when sampling the votes x, y, z as $(x_i, y_i, z_i) \sim \mathsf{NAE}_3^{-1}(1)$ for each i independently is at most ε. Then, f is ε-close to a dictatorship or an anti-dictatorship.*

Proof. Running Kalai's argument, we get that

$$1 - \varepsilon \leqslant \frac{3}{4} - \frac{3}{4}\mathsf{Stab}_{-1/3}(f),$$

and rearranging $\mathsf{Stab}_{-1/3}(f) + \frac{1}{3} \leqslant \varepsilon$. Note that

$$\mathsf{Stab}_{-1/3}(f) + \frac{1}{3} = \sum_{k=0}^{\infty}\left(\frac{1}{3} + \left(-\frac{1}{3}\right)^k\right) W^{=k}[f] \geqslant \sum_{k \neq 1}\left(\frac{1}{3} - \frac{1}{27}\right) W^{=k}[f] = \frac{8}{27}\|f - f^{=1}\|_2^2,$$

so we get $\|f - f^{=1}\|_2^2 \leqslant \frac{27}{8}\varepsilon$. The result now follows from the FKN theorem. □

3 Noise sensitivity

Today we will focus on a notion called noise sensitivity. We will give a characterization of noise sensitive monotone functions due to Bejamini Kalai and Schramm, and en route introduce basic results and techniques such as the level d inequalities and decoupling.

Definition 3.1 (Noise Sensitivity)**.** *Let $f\colon \{-1,1\}^n \to \mathbb{R}$, and $\varepsilon > 0$. The noise sensitivity of f with parameter ρ is defined as* $\mathsf{NS}_\varepsilon(f) = \frac{1}{2} - \frac{1}{2}\mathsf{Stab}_{1-2\varepsilon}(f)$.

For ± 1 valued functions, we get that $\mathsf{NS}_\varepsilon(f) = \Pr_{\substack{(x,y)\\ 1-2\varepsilon \text{ correlated}}}[f(x) \neq f(y)]$. Note that if we sample x, y independently, the probability that $f(x) \neq f(y)$ is $2\Pr_x[f(x) = -1]\Pr_y[f(y) = 1] = \frac{1}{2}\mathsf{var}(f)$. Informally, we say that f is noise sensitive if $f(x), f(y)$ behave independently when x, y are sampled in $(1-2\varepsilon)$-correlated manner. I.e., we say f is (ε, ξ) noise sensitive if $\left|\mathsf{NS}_\varepsilon(f) - \frac{1}{2}\mathsf{var}(f)\right| \leqslant \xi$.

The main question that will concern us today is to characterize functions that are noise sensitive. To gain some intuition into this question, we first give Fourier analytic formula for noise sensitivity.

Claim 3.2. *Let $f\colon \{-1,1\}^n \to \mathbb{R}$. Then* $\mathsf{NS}_\varepsilon(f) = \frac{1}{2}\sum\limits_{S}(1-(1-2\varepsilon)^{|S|})\widehat{f}(S)^2$.

Proof. Plug in Claim 1.3 into the definition of NS_ε. $\square$

Note that

$$\frac{1}{2}\mathsf{var}(f) - \mathsf{NS}_\varepsilon(f) = \frac{1}{2}\sum_{S\neq\emptyset}((1-2\varepsilon)^{|S|})\widehat{f}(S)^2 = \frac{1}{2}\sum_{k=1}^{n}(1-2\varepsilon)^k W^{=k}[f].$$

Thus, this quantity is always non-negative. Also, we see that a function f is noise sensitive if and only if almost all of its Fourier weight lies on high levels; for example, if all but δ of the Fourier mass lies above level $T \gg 1/\varepsilon \log(1/\delta)$, then the above difference is at most $\delta + (1-2\varepsilon)^T \leqslant 2\delta$.

Our question is therefore: which Boolean functions have only negligible weight on the low levels? For general functions, this question is too difficult to answer and one can only give a sufficient condition. For the class of monotone functions, this answer is also a necessary condition.

4 The BKS theorem

To answer this question, we define the parameter $M(f) = \sum\limits_{i=1}^{n} I_i[f]^2$. Before we state the theorem, we make some sense of this parameter. If f is monotone, as you have seen in the homework problem, $\widehat{f}(\{i\}) = I_i[f]$, and so $M(f) = W^{=1}[f]$. In general however, one only has that $W^{=1}[f] \leqslant M(f)$, and $M(f)$ itself may be very large – for example, it is as large as n for $f = \prod\limits_{i=1}^{n} x_i$.

Nevertheless, it turns out that when $M(f)$ is small, it immediately provides bounds on the weight of the function f on all of the low-levels, as follows.

Theorem 4.1. *There exists an absolute constant $C > 0$, such that for all $k \in \mathbb{N}$,*

$$W^{=k}[f] \leqslant \left(\frac{C}{k}\right)^k M(f)\log\left(\frac{k}{M(f)}\right)^{k-1}.$$

Most of our effort today will be devoted into proving this theorem. We first show several consequences of it, and in particular use it to show a criteria for noise sensitivity. For $k \leqslant \log(1/M(f))$ we get that

$$W^{=k}[f] \leqslant M(f)\left(\frac{C}{k}\right)^k \left(2\log\left(\frac{1}{M(f)}\right)\right)^k \leqslant M(f)\left(\frac{C\log(\frac{1}{M(f)})}{k}\right)^k$$

Inspecting the last expression as a function of k, we see that it is increasing up to $k = \frac{C}{2}\log(1/M(f))$, hence for $k \leqslant aC\log(1/M(f))$ for small a we get that

$$W^{=k}[f] \leqslant M(f)a^{-aC\log(1/M(f))} = M(f)e^{a\log(1/a)C\log(1/M(f))},$$

so for small enough absolute constant a we get that $W^{=k}[f] \leqslant \sqrt{M(f)}$. We conclude:

Corollary 4.2. *There exists an absolute constant $c > 0$, such that for all $k \leqslant c\log(1/M(f))$ we have that*

$$W^{=k}[f] \leqslant \sqrt{M(f)}.$$

Corollary 4.3. *There exists an absolute constant $\alpha > 0$ such that for all $\varepsilon > 0$ and $f\colon \{-1,1\}^n \to \{-1,1\}$,*

$$\left|\frac{1}{2}\mathsf{var}(f) - \mathsf{NS}_\varepsilon(f)\right| \leqslant M(f)^{\alpha\varepsilon}.$$

Proof. Let $c > 0$ be from the previous corollary, and let $T = c\log(1/M(f))$ As seen earlier, the left hand side is equal to

$$\frac{1}{2}\sum_{k=1}^{n}(1-2\varepsilon)^k W^{=k}[f] \leqslant \frac{1}{2}(W^{\leqslant T}[f - \mathbb{E}[f]] + (1-2\varepsilon)^T).$$

For the first term, by the previous corollary, we have $W_{\leqslant T}[f - \mathbb{E}[f]] \leqslant T\sqrt{M(f)} \leqslant M(f)^\alpha$ for sufficiently small $\alpha > 0$. For the second term, $(1-2\varepsilon)^T \leqslant e^{-2\varepsilon T} = e^{-2\varepsilon c\log(1/M(f))} = M(f)^{2\varepsilon c} \leqslant M(f)^{\varepsilon\alpha}$. Plugging these two bounds finishes the proof. □

Thus, we see that if $M(f) \leqslant \xi^{1/(\alpha\varepsilon)}$, then f is (ε, ξ) noise sensitive. For monotone functions, we have a two sided connection.

Theorem 4.4. *There exists $\alpha > 0$ such that for all monotone $f\colon \{-1,1\}^n \to \{-1,1\}$,*

$$(1-\varepsilon)M(f) \leqslant \left|\frac{1}{2}\mathsf{var}(f) - \mathsf{NS}_\varepsilon(f)\right| \leqslant M(f)^{\alpha\varepsilon}.$$

Thus, f is $(\varepsilon, o(1))$ noise sensitive if and only if $M(f) = o(1)$.

The rest of this lecture is devoted to the proof of Theorem 4.1. For technical reasons, we will prove the following quantitatively weaker statement:

Theorem 4.5. *There exists an absolute constant $C > 0$, such that for all $k \in \mathbb{N}$,*

$$W^{=k}[f] \leqslant C^k M(f)\log\left(\frac{1}{M(f)}\right)^{k-1}.$$

The proof of this theorem is already fairly technical as is, and getting Theorem 4.1 requires additional effort.

5 A first attempt

Inspired by the proof of the KKL theorem and the Friedgut junta theorem, one may hope to divide the level k weight of f according to the contribution of each $i \in [n]$, use hypercontractivity and then make a conclusion. This approach almost works, and we will present it anyway since it is illuminating.

5.1 The level k inequalities

Recall that in earlier lectures, we have seen that if $g\colon \{-1,1\}^n \to \{-1,0,1\}$ is non-zero with probability at most δ, then most of its weight lies above level $\log(1/\delta)$. The precise bounds we got there are that $W^{\leqslant \log(1/\delta)/20} \leqslant \delta^{24/20}$. Below, we improve that for constant levels showing that $W^{\leqslant k}[f] \leqslant \delta^2 \mathsf{polylog}(1/\delta)$.

Lemma 5.1. *[Level k inequality] Suppose $g\colon \{-1,1\}^n \to \{-1,0,1\}$ is non-zero with probability δ, and let $k \in \mathbb{N}$. Then*

$$W^{\leqslant k}[g] \leqslant \delta^2 (e\log(2/\delta))^k.$$

Proof. Let $q \in \mathbb{N}$ to be determined later. Note that

$$W^{\leqslant k}[g] = \langle g^{\leqslant k}, g\rangle \leqslant \|g^{\leqslant k}\|_q \|g\|_{q/(q-1)}.$$

By hypercontractivity,

$$\|g^{\leqslant k}\|_q \leqslant \sqrt{q-1}^k \|g^{\leqslant k}\|_2 = \sqrt{q-1}^k \sqrt{W^{\leqslant k}[g]},$$

so rearranging $W^{\leqslant k}[g] \leqslant (q-1)^k \|g\|_{q/(q-1)}^2$. Also, $\|g\|_{q/(q-1)} = \delta^{(q-1)/q}$, so we get the bound

$$W^{\leqslant k}[g] \leqslant (q-1)^k \delta^{2\frac{q-1}{q}}.$$

Choosing $q = \log(4/\delta)$, we get

$$W^{\leqslant k}[g] \leqslant (\log(2/\delta))^k \delta^{2-\frac{1}{\log(4/\delta)}} = (e\log(2/\delta))^k \delta^2$$

□

Remark 5.2. *We note the level 0 weight of g is δ^2. Lemma 5.1 thus tells us that the level 1 weight of g can only jump multiplicatively by a logarithmic factor, and that this extends to all $k \in \mathbb{N}$. As we will see below, this improvement will be essential in our mock BKS theorem, but we remark that there are several other applications of the level k inequality that we will not show in this course in which this improvement is essential.*

Armed with Lemms 5.1, we may now attempt to prove the BKS theorem. Note that

$$W^{=k}[f] = \sum_{|S|=k} \widehat{f}(S)^2 = \frac{1}{k}\sum_{i=1}^{n} \sum_{|S|=k, i\in S} \widehat{f}(S)^2 = \frac{1}{k}\sum_{i=1}^{n} W^{=k-1}[\partial_i f].$$

The function $\partial_i f$ gets the valued $0,-1,1$ and is non-zero with probability $I_i[f]$. Hence, by Lemma 5.1 we have that $W^{=k-1}[\partial_i f] \leqslant 10^k I_i[f]^2 \log^{k-1}(1/I_i[f])$. Plugging this into the inequality above gives that

$$W^{=k}[f] \leqslant \frac{10^k}{k}\sum_{i=1}^{n} I_i[f]^2 \log^{k-1}(1/I_i[f]).$$

This is very similar to the bound we want, which is $C^k \sum\limits_{i=1}^{n} I_i[f]^2 \log^{k-1}(1/M(f))$. Alas, it is weaker, and it is not clear how to improve it.

6 Proof of Theorem 4.5

In a sense, the main deficiency in the above argument is that the level k inequality has been applied to each one of the derivatives by themselves. This sort of argument assumes that in the worst case, the level k inequality is tight for each one of the derivatives, which cannot be the case.

To utilize this point, the proof of the theorem uses a useful idea (not only in the context of Boolean functions) called decoupling. Let $k \geqslant 2$, and consider the level k weight of f. It would be useful if we could partition $[n]$ into two sets, I and J such that each $S \subseteq [n]$ of size k in the support of $\widehat{f}$ would have one variable from I, and $k-1$ variables from J. This would then be useful, as randomly restricting the coordinates of J, we would reduce the level k weight of the function to the level 1 weight of the restricted function, which is directly related to influences of the restricted function.

We first need to find such partition I, J of $[n]$. The condition we are looking for is too strong, and we have to settle for a more modest one:

Claim 6.1. *There is a partition (I, J) of $[n]$ such that*

$$\sum_{\substack{|S|=k \\ |S \cap I|=1}} \widehat{f}(S)^2 \geqslant \frac{1}{e} W^{=k}[f].$$

Proof. Choose the partition (I, J) randomly by including each $i \in [n]$ in I with probability $1/k$, and otherwise in J.

$$\mathop{\mathbb{E}}_{I,J}\left[\sum_{\substack{|S|=k \\ |S \cap I|=1}} \widehat{f}(S)^2\right] = \mathop{\mathbb{E}}_{I,J}\left[\sum_{|S|=k} \widehat{f}(S)^2 1_{|S\cap I|=1}\right] = \sum_{|S|=k} \widehat{f}(S)^2 \mathop{\mathbb{E}}_{I,J}\left[1_{|S\cap I|=1}\right].$$

Note that

$$\mathop{\mathbb{E}}_{I,J}\left[1_{|S\cap I|=1}\right] = \Pr\left[|S \cap I| = 1\right] = \binom{k}{1}\frac{1}{k}\left(1-\frac{1}{k}\right)^{k-1} = \left(1-\frac{1}{k}\right)^{k-1} \geqslant \frac{1}{e},$$

so the expectation is at least $\frac{1}{e} W^{=k}[f]$, and in particular there is a partition (I, J) as desired. $\square$

Fix (I, J) as in the claim. For the rest of the proof, we write as input x in the hypercube as $x = (y, z)$, where y is the I-part of x and z is the J-part of x. We now partition the sum on the left hand side of the claim according to which coordinate in I is involved, and for that for each $i \in I$ we define $f_i \colon \{-1, 1\}^J \to \mathbb{R}$ by

$$f_i'(z) = \sum_{T \subseteq J, |T|=k-1} \widehat{f}(T \cup \{i\}) \chi_T(z).$$

Thus,

$$\langle y_i f_i', f\rangle = \sum_{|T|=k-1, T\subseteq J} \widehat{f}(T \cup \{i\})^2 = \|f_i'\|_2^2.$$

It is convenient to normalize f_i' so that it has 2-norm 1, $f_i = f_i'/\|f_i'\|_2$, in which case we get

$$\langle y_i f_i, f\rangle^2 = \sum_{|T|=k-1, T\subseteq J} \widehat{f}(T \cup \{i\})^2.$$

Thus,

$$\sum_{\substack{|S|=k \\ |S \cap I|=1}} \widehat{f}(S)^2 = \sum_{i \in I} \langle y_i f_i, f\rangle^2,$$

and our goal is to bound each one of the inner products on the right hand side. Write

$$\langle f_i, y_i f\rangle^2 = \mathop{\mathbb{E}}_{y,z}\left[f_i(z) y_i f(y, z)\right]^2 = \mathop{\mathbb{E}}_{z}\left[f_i(z) \mathop{\mathbb{E}}_{y}\left[y_i f(y, z)\right]\right]^2 \tag{1}$$

The idea is now to use the fact that f_i is a low-degree function, and hence it behaves as if it is a bounded function. If indeed it was bounded, say by $T = O(1)$, we would be able to say that

$$\langle f_i, y_i f\rangle^2 \leqslant T^2 \mathop{\mathbb{E}}_{z}\left[\left|\mathop{\mathbb{E}}_{y}\left[y_i f(y,z)\right]\right|\right]^2 \leqslant T^2 \mathop{\mathbb{E}}_{z,y_{-i}}\left[\left|\mathop{\mathbb{E}}_{y_i}\left[y_i f(y,z)\right]\right|\right]^2 = T^2 \mathop{\mathbb{E}}_{z,y}\left[1_{f(y,z)\neq f(y,z+e_i)}\right]^2 = T^2 I_i[f]^2,$$

which is the sort of bound we are after (summing over i gets $\leqslant T^2 M(f)$).

The situation is of course not as simple, and f_i is not really a bounded function. To get around this issue, we introduce a threshold parameter T to be determined later and analyze separately cases where $|f_i(z)| \geqslant T$ and cases where $|f_i(z)| \leqslant T$. First, write

$$\begin{aligned}(1) &\leqslant \mathop{\mathbb{E}}_{z}\left[|f_i(z)|\left(1_{|f_i(z)|<T} + 1_{|f_i(z)|\geqslant T}\right)\left|\mathop{\mathbb{E}}_{y}\left[y_i f(y,z)\right]\right|\right]^2 \\ &\leqslant \underbrace{2\mathop{\mathbb{E}}_{z}\left[|f_i(z)|\, 1_{|f_i(z)|<T}\left|\mathop{\mathbb{E}}_{y}\left[y_i f(y,z)\right]\right|\right]^2}_{(I)} + \underbrace{2\mathop{\mathbb{E}}_{z}\left[|f_i(z)|\, 1_{|f_i(z)|\geqslant T}\left|\mathop{\mathbb{E}}_{y}\left[y_i f(y,z)\right]\right|\right]^2}_{(II)},\end{aligned}$$

where we used $(a+b)^2 \leqslant 2a^2 + 2b^2$. We may bound the first expression as in the simplistic case by

$$(I) \leqslant 2T^2 \mathop{\mathbb{E}}_{z}\left[\left|\mathop{\mathbb{E}}_{y}\left[y_i f(y,z)\right]\right|\right]^2 \leqslant 2T^2 I_i[f]^2.$$

As for the second one, using Cauchy-Schwarz we get that

$$(II) \leqslant 2\mathop{\mathbb{E}}_{z}\left[|f_i(z)|^2\, 1_{|f_i(z)|\geqslant T}\right]\mathop{\mathbb{E}}_{z}\left[\left|\mathop{\mathbb{E}}_{y}\left[y_i f(y,z)\right]\right|^2\right],$$

and we bound each one of the expectations separately. For the first one, applying Cauchy-Schwarz

$$\mathop{\mathbb{E}}_{z}\left[|f_i(z)|^2\, 1_{|f_i(z)|\geqslant T}\right] \leqslant \|f_i\|_4^2 \sqrt{\Pr\left[|f_i(z)| \geqslant T\right]}.$$

The norm is at most $3^{k-1}\|f_i\|_2^2 = 3^{k-1}$ by hypercontractivity. The probability is at most $e^{-\frac{1}{2}T^{2/(k-1)}}$ by the tail bound from Lecture 5. For the second expectation, we only note that $\left|\mathbb{E}_y\left[y_i f(y,z)\right]\right| = \left|\widehat{f_{J\to z}}(\{i\})\right|$. Therefore,

$$(II) \leqslant 2\cdot 3^{k-1} e^{-\frac{1}{4}T^{2/(k-1)}} \mathop{\mathbb{E}}_{z}\left[\left|\widehat{f_{J\to z}}(\{i\})\right|^2\right].$$

Thus,

$$\langle f_i, y_i f\rangle^2 \leqslant 2T^2 I_i[f]^2 + 2\cdot 3^{k-1} e^{-\frac{1}{4}T^{2/(k-1)}} \mathop{\mathbb{E}}_{z}\left[\left|\widehat{f_{J\to z}}(\{i\})\right|^2\right],$$

and summing over i gives that

$$\sum_{i\in I}\langle f_i, y_i f\rangle^2 \leqslant 2T^2 M(f) + 2\cdot 3^{k-1} e^{-\frac{1}{4}T^{2/(k-1)}} \mathop{\mathbb{E}}_{z}\left[W^{=1}[f_{J\to z}]\right] \leqslant 2T^2 M(f) + 2\cdot 3^{k-1} e^{-\frac{1}{4}T^{2/(k-1)}}.$$

We now choose $T = 100\log(1/M(f))^{(k-1)/2}$ and get that

$$\sum_{i\in I}\langle f_i, y_i f\rangle^2 \leqslant 2\cdot 100^2 M(f)\log(1/M(f))^{k-1} + 2\cdot 3^{k-1} e^{-25\log(1/M(f))} \leqslant C^k M(f)\log(1/M(f))^{k-1},$$

and we are done by the choice of I.