PROBLEM SET 4

Due: Tuesday, April 3, beginning of class

Homework policy: I encourage you to try to solve the problems by yourself. However, you may collaborate as long as you do the writeup yourself and list the people you talked with.

Do 4 out of 6.

1. Orthogonal decomposition. Given any $f : \{-1,1\}^n \to \mathbb{R}$, consider $f^S : \{-1,1\}^n \to \mathbb{R}$ defined by $f^S = \hat{f}(S)\chi_S$. We have $f = \sum_{S\subseteq[n]} f^S$ as functions, and this "orthogonal decomposition" has the following three properties:

(i) $f^S(x)$ depends only on the coordinates of x in S;

(ii) $\mathbf{E}_{\boldsymbol{x}}[f^S(\boldsymbol{x})f^T(\boldsymbol{x})] = 0$ if $S \neq T$;

(iii) $\sum_{T\subseteq S} f^T$, denoted $f^{\leq S}$, gives the conditional expectation of f conditioned on the coordinates in S.

(a) Prove property (iii); i.e., $f^{\leq S}(x) = \mathbf{E}[f_{x\to S}]$, where the expectation is over the bits in $\bar{S} = [n] \setminus S$. (Here the notation is that $x \in \{-1,1\}^n$, but in the expression $f_{x\to S}$, we only restrict the S-coordinates of f using the S-bits of x; the $\bar{S}$-bits of x are ignored.)

In the rest of this problem we establish the same kind of decomposition for general real-valued functions on product probability spaces. Specifically, let X be any finite set and let π be a probability distribution on X. We think of the n-fold product set X^n as having the product probability distribution induced by π. All $\mathbf{Pr}[\cdot], \mathbf{E}[\cdot]$ in what follows refer to this product distribution.

(b) We first make property (iii) hold by fiat: For $S \subseteq [n]$, we *define* $f^{\leq S} : X^n \to \mathbb{R}$ to be the function depending only on the coordinates in S giving the conditional expectation; i.e., $f^{\leq S}(x) := \mathbf{E}[f_{x\to S}]$, where the expectation is over the product probability distribution on the coordinates outside S. Now given this definition, explicitly write how we should define the functions f^S so that the equations $f^{\leq S} = \sum_{T\subseteq S} f^T$ hold. Check also that property (i) holds with your definitions. (Hint: inclusion-exclusion.)

(c) Show that $\mathbf{E}_{\boldsymbol{x}}[f^{\leq S}(\boldsymbol{x})f^{\leq T}(\boldsymbol{x})] = \mathbf{E}_{\boldsymbol{x}}[f^{\leq (S\cap T)}(\boldsymbol{x})^2]$, straight from our definition of $f^{\leq S}$.

(d) Now show property (ii), that $\mathbf{E}_{\boldsymbol{x}}[f^S(\boldsymbol{x})f^T(\boldsymbol{x})] = 0$ when $S \neq T$. (Hint: write your definitions of f^S, f^T from (b), and then use (c).)

Remark: This "orthogonal decomposition" of functions f is often a good substitute for Fourier analysis when the domain is a product probability space other than $\{-1,1\}^n$.

2. Logarithmic Sobolev Inequality. Consider the Hypercontractive Theorem with $q = 2$, $p = 2 - 2\epsilon$, and $\rho = \sqrt{1-2\epsilon}$, where $\epsilon \in [0, 1/2]$; if we square it, we get

$$\|T_{\sqrt{1-2\epsilon}} f\|_2^2 \leq \|f\|_{2-2\epsilon}^2$$

for any $f : \{-1,1\}^n \to \mathbb{R}$.

(a) Show that we have equality at $\epsilon = 0$. Explain why we can now conclude that

$$\frac{\partial}{\partial \epsilon} \|T_{\sqrt{1-2\epsilon}} f\|_2^2 \Big|_{\epsilon=0} \le \frac{\partial}{\partial \epsilon} \|f\|_{2-2\epsilon}^2 \Big|_{\epsilon=0}.$$

(b) Show that

$$\frac{\partial}{\partial \epsilon} \|T_{\sqrt{1-2\epsilon}} f\|_2^2 \Big|_{\epsilon=0} = -2\mathbb{I}(f).$$

(c) Show that

$$\frac{\partial}{\partial \epsilon} \|f\|_{2-2\epsilon}^2 \Big|_{\epsilon=0} = -\,\mathbf{Ent}[f^2],$$

where $\mathbf{Ent}[g]$ is the functional defined for nonnegative g by $\mathbf{Ent}[g] = \mathbf{E}[g \ln g] - \mathbf{E}[g] \ln \mathbf{E}[g]$. [1]

We conclude that for all $f : \{-1,1\}^n \to \mathbb{R}$,

$$\mathbf{Ent}[f^2] \le 2\mathbb{I}(f).$$

This is called the "Logarithmic Sobolev Inequality", or the "Entropy-Energy Inequality". (Recall we called $\mathbb{I}(f)$ the "energy" of f in Lecture 1.)

(d) Show that if $f : \{-1,1\}^n \to \{\mathrm{T}, \mathrm{F}\}$ has $p = \mathbf{Pr}[f = \mathrm{T}] \le 1/2$, then

$$2p \ln(1/p) \le \mathbb{I}(f).$$

This significantly improves on the Poincaré Inequality $4p(1-p) \le \mathbb{I}(f)$ for small p.

3. ϵ**-biased sets.** For every positive integer k, there is a field $\mathbb{F}_{2^k}$ with exactly 2^k elements. There is a natural way of encoding the names of the field elements as k-bit strings, $\mathrm{enc} : \mathbb{F}_{2^k} \to \mathbb{F}_2^k$, and this encoding has the property that $\mathrm{enc}(x+y) = \mathrm{enc}(x) + \mathrm{enc}(y)$ for all $x, y \in \mathbb{F}_{2^k}$ and also $\mathrm{enc}(0) = (0, \ldots, 0)$. Further, given $\mathrm{enc}(x)$ and $\mathrm{enc}(y)$, one can compute $\mathrm{enc}(xy)$, $\mathrm{enc}(x/y)$, $\mathrm{enc}(x+y)$, $\mathrm{enc}(x-y)$, in deterministic $\mathrm{poly}(k)$ time.[2]

(a) Let $\boldsymbol{R}$ denote a random string in $\mathbb{F}_2^n$, formed as follows: Pick $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{F}_{2^k}$ independently and uniformly at random; then let the ith bit of $\boldsymbol{R}$ be $\langle \mathrm{enc}(\boldsymbol{a}^i), \mathrm{enc}(\boldsymbol{b}) \rangle_{\mathbb{F}_2}$, where $\langle \cdot, \cdot \rangle_{\mathbb{F}_2}$ denotes *dot product in* $\mathbb{F}_2$. Show that for every nonzero string $S \in \mathbb{F}_2^n$,

$$\frac{1}{2} - \frac{1}{2} \cdot \frac{n}{2^k} \le \mathbf{E}_{\boldsymbol{R}}[\langle \boldsymbol{R}, S \rangle_{\mathbb{F}_2}] \le \frac{1}{2},$$

where in the expectation, we're taking $\langle \boldsymbol{R}, S \rangle_{\mathbb{F}_2}$ (which is in $\mathbb{F}_2$) and reinterpreting it as a real number. (Hint: every nonzero degree-n polynomial over a field has at most n zeroes.)

(b) As needed for Problem 4 on Homework 3, give efficiently constructible ϵ-biased sets for $\{-1,1\}^n$ of size $(n/\epsilon)^2$, whenever n/ϵ is a power of 2.

[1] $0 \ln 0 = 0$.

[2] Specifically, it is known that for every k there is an irreducible polynomial $p(t) \in \mathbb{F}_2[t]$ of degree k; then we may take $\mathbb{F}_{2^k}$ to be the set of polynomials in $\mathbb{F}_2[t]$ modulo $p(t)$. The function enc maps $\sum_{i=0}^{k-1} a_i t^i$ to $(a_0, \ldots, a_{k-1})$. It is known (Shoup, 1990) that one can deterministically find an irreducible p in time $\mathrm{poly}(k)$. (Also, it's very easy to find one in time $2^{O(k)}$ which is pretty much good enough for us.)

4. Tightness of the Hypercontractivity Theorem. It is known[3] that when q is an even positive integer, the largest possible ratio of $\|f\|_q/\|f\|_2$ for degree-d real-valued boolean functions f is achieved, in the limit as $n \to \infty$, by

$$f = \sum_{\substack{S \subseteq [n] \\ |S|=d}} x_S. \tag{1}$$

I don't think there is a closed-form formula for the limiting ratio, but it is known to be $\Theta_q(1) \cdot d^{-1/4} \cdot \sqrt{q-1}^d$. In this problem we show a slightly weaker lower bound for our favorite case, $q = 4$.

(a) Let $f : \{-1,1\}^n \to \boldsymbol{R}$ be the function in (1); assume that d is divisible by 3 and that $n \geq 2d$. Show that

$$\mathbf{E}[f^4] \geq \frac{\binom{n}{d/3,d/3,d/3,d/3,d/3,d/3,n-2d}}{\binom{n}{d}^2} \mathbf{E}[f^2]^2,$$

where the quantity in the numerator is a multinomial coefficient — specifically, the number of ways of choosing six disjoint $d/3$-size subsets of $[n]$. (Hint: given six disjoint $d/3$-size subsets, consider quadruples of d-size sets S that hit each $d/3$-set twice.)

(b) Using Stirling's formula, show that

$$\lim_{n\to\infty} \frac{\binom{n}{d/3,d/3,d/3,d/3,d/3,d/3,n-2d}}{\binom{n}{d}^2} = \Theta(d^{-2} \cdot 9^d).$$

(c) [Extra credit.] Problems (a) and (b) give a lower bound of $\Omega(d^{-1/2} \cdot \sqrt{3}^d)$ for the largest possible ratio of $\|f\|_4/\|f\|_2$ for degree-d f. For extra credit, either: (i) explain Bonami's original 1970 argument which gives an upper bound of $O(d^{-1/8} \cdot \sqrt{3}^d)$; or (ii) show that the f in (1) indeed achieves $\Theta(d^{-1/4} \cdot \sqrt{3}^d)$ in the limit as $n \to \infty$. Or both.

5. Generalized Chernoff bounds. A Chernoff bound says that if $t \geq 1$ and $\sum_{i=1}^n a_i^2 = 1$, then

$$\mathbf{Pr}[|a_1\boldsymbol{x}_1 + \cdots a_n\boldsymbol{x}_n| \geq t] \leq \exp\left(-\Omega(t^2)\right),$$

where the $\boldsymbol{x}_i$'s denote i.i.d. random ± 1 bits, as usual. Prove the following generalization: Let $p(x_1, \ldots, x_n)$ be a multilinear polynomial over the reals of degree at most d, and assume $\mathbf{E}[p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_2)^2] = 1$ (i.e., the sum of squares of p's coefficients is 1). Then for $t \geq 1$,

$$\mathbf{Pr}[|p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)| \geq t] \leq \exp\left(-\Omega(t^{2/d})\right).$$

(Hint: Markov plus $(2, q, (1/\sqrt{q-1})^d)$-hypercontractivity with large q.)

[3] Svante Janson has a published proof; I don't know if it's the earliest.

6. Learning monotone decision trees in "polynomial" time.

(a) Let $f : \{-1,1\}^n \to \{-1,1\}$ be computable by a depth-d decision tree. Show that $\sum_{i=1}^n \hat{f}(i) \leq O(\sqrt{d})$. (Hint: mimic the proof that Majority maximizes $\sum \hat{f}(i)$ for general f; but take the expectation over a random path first.) Conclude that if f is monotone, $\mathbb{I}(f) \leq \sqrt{\text{DT-depth}(f)}$.

(b) Suppose one has access to random examples from a monotone function f. Give a learning algorithm which on input τ, identifies (w.h.p.) a set J which contains all coordinates i with $\mathrm{Inf}_i(f) \geq \tau$. The algorithm should run in time $\mathrm{poly}(n, 1/\tau)$ and the set J identified should have size $O(1/\tau^2)$.

(c) Show that $\mathcal{C} = \{\text{monotone } f : \text{DT-depth}(f) \leq \log n\}$ is learnable from random examples only in time $n^{O(1/\epsilon^2)}$. (Hint: use the Main Lemma that implied Friedgut's Theorem.)