Lectures 14–15: Log-Sobolev inequalities, hypercontractivity, and random walks
CSE 599S: Entropy optimality, Winter 2016
Instructor: James R. Lee **Last updated:** March 13, 2016

Our goal in this lecture is to prove something about (anti-)concentration properties of short random walks in the hypercube. We begin with a brief motivation coming from locality-sensitive hashing before stating our main objective Lemma 1.2. By diagonalizing the random walk operator, we will arrive at a spectral formulation of our objective. Finally, we will see that *hypercontractivity* gives the spectral information we need, by providing bounds on the eigenvalues of indicator functions of subsets of the hypercube.

1 Locality-sensitive hashing

Consider the following representative setting for *locality-sensitive hashing*. Fix the dimension $d \geqslant 1$. A random mapping $\mathcal{H} : \{-1,1\}^d \to \mathbb{N}$ is called an (r, R, p, q)-*LSH family* if for any pair $x, y \in \{-1,1\}^d$, it holds that

1. $\|x-y\|_1 \leqslant r \implies \mathbb{P}\left[\mathcal{H}(x) = \mathcal{H}(y)\right] \geqslant p$,
2. $\|x-y\|_1 > R \implies \mathbb{P}\left[\mathcal{H}(x) = \mathcal{H}(y)\right] \leqslant q$.

Given an approximation parameter $c > 0$, we would like to know what quality of (r, cr, p, q)-LSH families exist. The value

$$\rho(c) = \max_{1 \leqslant r \leqslant d} \inf \left\{ \frac{\log(1/p)}{\log(1/q)} : \exists (r, cr, p, q)\text{-LSH} \right\}$$

is a crucial parameter in determining the time and space complexity of nearest-neighbor search algorithms based on locality-sensitive hashing.

Exercise (1 point) 1.1. Give an explanation for the appearance of the ratio $\log(1/p)/\log(1/q)$ along the following lines. You may assume that $0 < p < q < 1$ are constants independent of n. Suppose that properties (1) and (2) hold for a random map $\mathcal{H}$. Consider an integer $k \in \mathbb{N}$, and the hash function $\mathcal{H}^{\otimes k} : \{-1,1\}^d \to \mathbb{N}^k$ given by $\mathcal{H}^{\otimes k}(x) = (\mathcal{H}_1(x), \ldots, \mathcal{H}_k(x))$ where $\{\mathcal{H}_i\}$ are i.i.d. copies of $\mathcal{H}$.

If our database $\mathcal{D} \subseteq \{-1,1\}^d$ has $|\mathcal{D}| = n$. How large should we choose k so that $\mathcal{H}^{\otimes k}$ is an $(r, R, p(k), 1/n)$-LSH family for some $p(k)$? (Here we have extended the notation of an LSH family to allow for a range that is not $\mathbb{N}$.) What value of $p(k)$ arises? Intuitively, to construct a nearest-neighbor data structure, we will need to sample at least $1/p(k)$ times so that two points with distance r end up in the same bucket with high probability.

Assuming that the values of a single hash function on $\mathcal{D}$ can be stored in $O(n)$ space, what is the total space consumption of our data structure? □

The paper [Motwani-Naor-Panigrahi 2007] shows how fairly tight lower bounds on $\rho(c)$ can be proved by analyzing random walks. Intuitively, if we consider a random hash bucket $\mathcal{H}^{-1}(x)$ for $x \in \{-1,1\}^d$ chosen uniformly at random, then property (2) suggests this bucket should not be too large (since most points are "far" from x, and far points hash to different buckets with probability at least q). On the other hand, if we do a random walk of length r from a random point $z \in \mathcal{H}^{-1}(x)$, then property (1) suggests that the walk should often end back in $\mathcal{H}^{-1}(x)$ with probability at least p.

A question on random walks. For a point $x \in \{-1,1\}^d$, let $W_r(x)$ be a random variable denoting the outcome of a random walk from x of length r (one step of the random walk involves choosing a uniformly random $i \in \{1,2,\ldots,d\}$ and flipping the ith bit).

Lemma 1.2. *Let r be an odd integer. For any subset $B \subseteq \{-1,1\}^d$, it holds that*

$$\mathop{\mathbb{P}}_{x\in B}[W_r(x) \in B] \leqslant \left(\frac{|B|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}} .$$

The preceding lemma asserts an upper bound on the probability of a random walk starting from a uniformly random $x \in B$ to end back in B. Observe that if $r \gg d \log d$, then the random walk mixes, and $W_r(x)$ is close to a uniformly random point of $\{-1,1\}^d$, in which case we expect the bound $|B|/2^d$. Lemma 1.2 gives a non-trivial bound even for much shorter random walks.

For $r = \varepsilon d$, with $\varepsilon \ll 1$, we have

$$\frac{e^{2r/d}-1}{e^{2r/d}+1} \approx \frac{2\varepsilon}{2+2\varepsilon} = \frac{\varepsilon}{1+\varepsilon} .$$

Our goal in these lecture is to develop the tools necessary to prove Lemma 1.2.

1.1 Diagonalizing the random walk operator

Let us denote by P the transition matrix of the random walk on $\{-1,1\}^d$, so that $P_{xy} = \frac{1}{d}$ if x and y differ in exactly one coordinate, and $P_{xy} = 0$ otherwise. Recall the space $L^2(\{-1,1\}^d)$ of functions $f : \{-1,1\}^d \to \mathbb{R}$ equipped with the inner product $\langle f, g\rangle = \mathbb{E}_{x\in\{-1,1\}^d} f(x)g(x)$.

Exercise (1 point) 1.3. Prove the formula

$$\mathop{\mathbb{P}}_{x\in B}[W_r(x) \in B] = \frac{2^d}{|B|}\langle \mathbf{1}_B, P^r \mathbf{1}_B\rangle , \tag{1.1}$$

where $\mathbf{1}_B$ denotes the characteristic function of B.

Given the representation (1.1), it makes sense to diagonalize P. For $S \subseteq \{1,\ldots,d\}$, let $u_S \in L^2(\{-1,1\}^d)$ denote the function $u_S(x) = \prod_{i\in S} x_i$.

Exercise (1 point) 1.4. Prove that the functions $\{u_S : S \subseteq \{1,\ldots,d\}\}$ form a complete orthonormal basis of eigenfunctions for P, and

$$P u_S = \left(1 - \frac{2|S|}{d}\right) u_S .$$

Recall that given a function $f \in L^2(\{-1,1\}^d)$, we denote its Fourier coefficients by $\hat{f}(S) = \langle f, u_S\rangle$. We can use the preceding exercise to give a spectral expression for (1.1):

$$\langle \mathbf{1}_B, P^r \mathbf{1}_B\rangle = \sum_{S\subseteq[d]} \hat{\mathbf{1}}_B(S)^2 \left(1-\frac{2|S|}{d}\right)^r \leqslant \sum_{|S|\leqslant d/2} \hat{\mathbf{1}}_B(S)^2 \left(1-\frac{2|S|}{d}\right)^r \leqslant \sum_{S\subseteq[d]} \hat{\mathbf{1}}_B(S)^2 e^{-2r|S|/d} , \tag{1.2}$$

where in the first inequality we have used the fact that r is odd to drop the negative terms, and in the final inequality we have used $1 - x \leqslant e^{-x}$.

1.2 The noise operator

In order to understand the final expression, let's consider more generally an expression of the form

$$\sum_{S\subseteq[d]} \hat{\mathbf{1}}_B(S)^2 \varepsilon^{2|S|}$$

for some $\varepsilon \in [0,1]$. For the next exercise, recall the definition of the $L^p(\{-1,1\}^d)$ norm:

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

Exercise (1 point) 1.5. Consider the operator $T_\varepsilon : L^2(\{-1,1\}^d) \to L^2(\{-1,1\}^d)$ defined as follows:

$$T_\varepsilon f(x_1,\ldots,x_d) = \mathbb{E}[f(X_1^\varepsilon,\ldots,X_d^\varepsilon)],$$

where $\{X_i^\varepsilon\}$ are independent random variables satisfying $X_i^\varepsilon = X_i$ with probability $\frac{1}{2}(1+\varepsilon)$, and $X_i^\varepsilon = -X_i$ with probability $\frac{1}{2}(1-\varepsilon)$. Prove that for any function $f \in L^2(\{-1,1\}^d)$, we have

$$\|T_\varepsilon f\|_2^2 = \sum_{S\subseteq[d]} \hat{f}(S)^2 \varepsilon^{2|S|}. \tag{1.3}$$

Combining the preceding exercise with (1.2) gives us

$$\langle \mathbf{1}_B, P^r \mathbf{1}_B\rangle \leqslant \|T_\varepsilon \mathbf{1}_B\|_2^2 \tag{1.4}$$

with $\varepsilon = e^{-r/d}$.

In fact, we have not done much yet. We simply replaced our discrete-time random walk P^r with a continuous time random walk T_ε. For reasons we have already seen, the continuous-time random walk will be somewhat easier to work with, but as the preceding arguments show, they are closely related from a spectral viewpoint.

2 Hypercontractivity

To finish the proof of Lemma 1.2, we need to study the quantity $\|T_\varepsilon \mathbf{1}_B\|_2^2$. Of course from (1.3), we know this is "simply" a matter of understanding the Fourier spectrum $\{\hat{\mathbf{1}}_B(S) : S \subseteq [d]\}$. But we need to use the fact that $\mathbf{1}_B$ is the indicator of a small subset.

The problem is that being a $\{0,1\}$-valued function with small support is not such an easy property to exploit from an analytic standpoint. Instead, it makes sense to look at an analytic property that such functions have. For instance, suppose that $|B| = \varepsilon 2^d$ and we consider the L^p norms of the normalized indicator $\mathbf{1}_B/\varepsilon$:

$$\left\|\frac{\mathbf{1}_B}{\varepsilon}\right\|_p = \varepsilon^{\frac{1-p}{p}}.$$

As $p \to \infty$, we have $\|\mathbf{1}_B/\varepsilon\|_p \to \|\mathbf{1}_B/\varepsilon\|_\infty = 1/\varepsilon$. Of course when $p = 1$, we have $\|\mathbf{1}_B/\varepsilon\|_1 = 1$, but notice that the size of the set is also captured near $p = 1$ if we calculate

$$\frac{d}{dp}\Big|_{p=1} \varepsilon^{\frac{1-p}{p}} = \log\frac{1}{\varepsilon}.$$

This provides some motivation for trying to compare $\|T_\varepsilon \mathbf{1}_B\|_2^2$ to $\|\mathbf{1}_B\|_p$ for some other values of p, Since T_ε is an averaging operator and norms are convex, we always have $\|T_\varepsilon f\|_p \leqslant \|f\|_p$ for any $f \in L^p(\{-1,1\}^d)$. If we can achieve a stronger bound of the form $\|T_\varepsilon f\|_q \leqslant \|f\|_p$ for $q > p$, this phenomenon is referred to as *hypercontractivity.*

The following result is due independently to Bonami and Gross (see O'Donnell's book).

Theorem 2.1. *For every $f \in L^2(\{-1,1\}^d)$ and $q \geqslant p \geqslant 1$, it holds that*

$$\|T_\varepsilon f\|_q \leqslant \|f\|_p \quad \textit{whenever} \quad \varepsilon \leqslant \sqrt{\frac{q-1}{p-1}}\,.$$

In particular, for every $\varepsilon \in [0,1]$,

$$\|T_\varepsilon f\|_2 \leqslant \|f\|_{1+\varepsilon^2}\,.$$

First let us observe how this theorem can be used to finish the proof of Lemma 1.2. We calculate

$$\|T_\varepsilon \mathbf{1}_B\|_2^2 \leqslant \|\mathbf{1}_B\|_{1+\varepsilon^2}^2 = \left(\frac{|B|}{2^d}\right)^{\frac{2}{1+\varepsilon^2}}$$

Plugging in $\varepsilon = e^{-r/d}$ and recalling (1.1) and (1.4) yields

$$\mathop{\mathbb{P}}_{x\in B}\left[W_r(x) \in B\right] \leqslant \frac{2^d}{|B|}\left(\frac{|B|}{2^d}\right)^{\frac{2}{1+e^{-2r/d}}} = \left(\frac{|B|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}}$$

Exercise (1 point) 2.2. Recall that when $r = \gamma d$ and $\gamma \ll 1$, the exponent is approximately $\frac{\gamma}{1+\gamma}$. For every $\eta < 1$, give an example of a subset $B \subseteq \{-1,1\}^d$ with $|B| \approx 2^{(1-\eta)d}$ and such that

$$\frac{\log \mathbb{P}_{x\in B}\left[W_r(x) \in B\right]}{\log\left(2^{-d}|B|\right)} \approx \gamma\,,$$

showing that the linear dependence on γ is tight.

2.1 Log-Sobolev inequalities

As we have seen in Lecture 13 (refer to Chapters 1 and 4 of the Montenegro-Tetali book), by work of Gross we know that Theorem 2.1 is equivalent to the log-Sobolev inequality:

$$\mathrm{Ent}(f) \leqslant 2n \cdot \mathcal{E}\left(\sqrt{f}, \sqrt{f}\right) \quad \text{for all} \quad f : \{-1,1\}^d \to \mathbb{R}_+\,, \tag{2.1}$$

where

$$\mathrm{Ent}(f) = \mathbb{E}\left[f \log \frac{f}{\mathbb{E} f}\right]$$

$$\mathcal{E}\left(\sqrt{f}, \sqrt{f}\right) = \frac{1}{2n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{x\in\{-1,1\}^d}\left(\sqrt{f(x)} - \sqrt{f(x \oplus e_i)}\right)^2.$$

Let us prove (2.1). We may assume that $\mathbb{E} f = 1$. Let $X = (X_1, \ldots, X_d) \in \{-1,1\}^d$ be a random vector with density f, and let $B = (B_1, \ldots, B_d) \in \{-1,1\}^d$ be a uniformly random point. Then by the chain rule for relative entropy:

$$\mathrm{Ent}(f) = D\left(X_1, \ldots, X_n \,\|\, B_1, \ldots, B_n\right) = \sum_{i=1}^{d} \mathbb{E}\left[D(X_i \,\|\, B_i \mid X_1, \ldots, X_{i-1})\right] ,$$

where we have additional used the fact that the coordinates of B are independent.

Denote $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$. By convexity of the relative entropy, we have

$$\mathbb{E}\left[D(X_i \,\|\, B_i \mid X_1, \ldots, X_{i-1})\right] \leqslant \mathbb{E}\left[D(X_i \,\|\, B_i \mid X_{-i})\right] .$$

To complete the proof, we are left to show that

$$\mathbb{E}\left[D(X_i \,\|\, B_i \mid X_{-i})\right] \leqslant \mathbb{E}\left[\left(\sqrt{f(B)} - \sqrt{f(B \oplus e_i)}\right)^2\right] \tag{2.2}$$

Let use use $(\frac{1}{2}, \frac{1}{2})$ to denote the law of B_i. Then since f is the density of X, we have

$$\begin{aligned}
\mathbb{E}[D(X_i \,\|\, B_i \mid X_{-i})] &= \mathbb{E}[f(B) \cdot D(X_i \,\|\, (\tfrac{1}{2}, \tfrac{1}{2}) \mid X_{-i} = B_{-i})] \\
&= \mathbb{E}\left[\frac{f(B) + f(B \oplus e_i)}{2} \cdot D(X_i \,\|\, (\tfrac{1}{2}, \tfrac{1}{2}) \mid X_{-i} = B_{-i})\right] .
\end{aligned}$$

If we condition on B and set $a = f(B)$ and $b = f(B \oplus e_i)$, we can write the expression in brackets as

$$\begin{aligned}
\frac{a+b}{2}\left(\frac{a}{a+b}\log\frac{2a}{a+b} + \frac{b}{a+b}\log\frac{2b}{a+b}\right) &= \frac{a}{2}\log\left(1 + \frac{a-b}{a+b}\right) + \frac{b}{2}\log\left(1 + \frac{b-a}{a+b}\right) \\
&\leqslant \frac{a}{2} \cdot \frac{a-b}{a+b} + \frac{b}{2} \cdot \frac{b-a}{a+b} \\
&= \frac{1}{2}\frac{(a-b)^2}{a+b} ,
\end{aligned}$$

where the inequality employed was simply $\log(1+x) \leqslant x$ for $x > -1$.

The last observation is that

$$\frac{1}{2}\frac{(a-b)^2}{a+b} \leqslant (\sqrt{a} - \sqrt{b})^2 ,$$

yielding (2.2).