Hypercontractivity and its Applications

Punyashloka Biswal

February 9, 2010

Abstract

Hypercontractive inequalities are a useful tool in dealing with extremal questions in the geometry of high-dimensional discrete and continuous spaces. In this survey we trace a few connections between different ways of stating hypercontractivity, and also present some relatively recent applications of these techniques in computer science.

1 Preliminaries and notation

Fourier analysis on the hypercube. The uniform probability measure on the Boolean cube $\{-1,1\}^n$ gives rise to a natural inner product $\langle f, g\rangle = \mathbb{E}_x f(x)g(x)$ on functions $f, g\colon \{-1,1\}^n \to \mathbb{R}$. The multilinear polynomials $\chi_S(x) = \prod_{i\in S} x_i$ (where S ranges over subsets of $[n]$) form an orthogonal basis under this inner product; they are called the Fourier basis. Thus, for any function $f\colon \{-1,1\}^n \to \mathbb{R}$, we have $f = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S(x)$, where the Fourier coefficients $\hat{f}(S) = \langle f, \chi_S\rangle$ obey Plancherel's relation $\sum \hat{f}(S)^2 = 1$. It is easy to verify that $\mathbb{E}_x f(x) = \hat{f}(0)$ and $\mathrm{Var}_x f(x) = \sum_{S\neq\emptyset} \hat{f}(S)^2$.

Norms. For $1 \le p < \infty$, define the ℓ_p norm $\|f\|_p = (\mathbb{E}_x |f(x)|^p)^{1/p}$. These norms are monotone in p: for every function f, $p \ge q$ implies $\|f\|_p \ge \|f\|_q$. For a linear operator M carrying functions $f\colon \{-1,1\}^n \to \mathbb{R}$ to functions $Mf = g\colon \{-1,1\}^n \to \mathbb{R}$, we define the p-to-q operator norm $\|M\|_{p\to q} = \sup_f \|Mf\|_q/\|f\|_q$. M is said to be a contraction from ℓ_p to ℓ_q when $\|M\|_{p\to q} \le 1$. Because of the monotonicity of norms, a contraction from ℓ_p to ℓ_p is automatically a contraction from ℓ_p to ℓ_q for any $q < p$. When $q > p$ and $\|M\|_{p\to q} \le 1$, then M is said to be hypercontractive.

Convolution operators. The convolution $(f*g)(x) = \mathbb{E}_y f(x)g(xy)$ of two functions $f, g\colon \{-1,1\}^n \to \mathbb{R}$ defines a linear operator $f \mapsto f * g$, where xy is the coordinate-wise product. Convolution is commutative and associative, and the Fourier coefficients of a convolution satisfy the useful property $\widehat{f * g} = \hat{f}\hat{g}$. We shall be particularly interested in the convolution properties of the following functions

- The Dirac delta $\delta\colon \{-1,1\}^n \to \mathbb{R}$, given by $\delta(1,\ldots,1) = 1$ and $\delta(x) = 0$ otherwise. It is the identity for convolution and has $\hat{\delta}(S) = 1$ for all $S \subseteq [n]$.

- The edge functions $h_i\colon \{-1,1\}^n \to \mathbb{R}$ given by

$$h_i(x) = \begin{cases} 1/2 & x = (1,\dots,1) \\ -1/2 & x_i = -1, x_{[n]\setminus\{i\}} = (1,\dots,1) \\ 0 & \text{otherwise.} \end{cases}$$

  $\hat{h}_i(S)$ is 1 or 0 according as S contains or does not contain i, respectively. For any function $f\colon \{-1,1\}^n \to \mathbb{R}$, $(f * h_i)(x) = (f(x) - f(y))/2$, where y is obtained from x by flipping just the ith bit. Convolution with h_i acts as an orthogonal projection (as we can easily see in the Fourier domain), so for any functions $f, g\colon \{-1,1\}^n \to \mathbb{R}$, we have $\langle f * h_i, g\rangle = \langle f, h_i * g\rangle = \langle f * h_i, g * h_i\rangle$

- The Bonami-Gross-Beckner noise functions $\mathrm{BG}_\rho\colon \{-1,1\}^n \to \mathbb{R}$ for $0 \le \rho \le 1$, where $\widehat{\mathrm{BG}}_\rho(S) = \rho^{|S|}$ and we define $0^0 = 1$. These operators form a semigroup, because $\mathrm{BG}_\sigma * \mathrm{BG}_\rho = \mathrm{BG}_{\sigma\rho}$ and $\mathrm{BG}_1 = \delta$. Note that $\mathrm{BG}_\rho(x) = \sum_S \rho^{|S|}\chi_S(x) = \prod_i (1 + \rho x_i)$. We define the noise operator T_ρ acting on functions on the discrete cube by $T_\rho f = \mathrm{BG}_\rho * f$. In combinatorial terms, $(T_\rho f)(x)$ is the expected value of $f(y)$, where y is obtained from x by independently flipping each bit of x with probability $1 - \rho$.

Lemma 1. $\frac{d}{d\rho}\mathrm{BG}_\rho = \frac{1}{\rho}\mathrm{BG}_\rho * \sum h_i$

Proof. This is easy in the Fourier basis:

$$\widehat{\mathrm{BG}}'_\rho = (\rho^{|S|})' = |S|\rho^{|S|-1} = \sum_{i\in[n]} \hat{h}_i \frac{\widehat{\mathrm{BG}}_\rho}{\rho}. \qquad \square$$

2 The Bonami-Gross-Beckner Inequality

2.1 Poincaré and Log-Sobolev inequalities

The Poincaré and logarithmic Sobolev inequalities both relate how far a function $f\colon \{-1,1\}^n \to \mathbb{R}$ is from constant to how fast it changes "locally". The amount of local change is quantified by the *energy* $\mathbb{D}(f,f)$, where the Dirichlet form $\mathbb{D}$ is defined as

$$\mathbb{D}(f,g) = \tfrac{1}{2} \mathop{\mathbb{E}}_{xy\in E}(f(x) - f(y))(g(x) - g(y))$$

(E is the set of pairs x, y that differ in a single coordinate). In terms of the edge functions h_i, observe that $\mathbb{D}(f,g) = \frac{2}{n}\sum_i \langle f * h_i, g * h_i\rangle$.

In the case of the Poincaré inequality, we measure the distance of f to a constant by its variance $\mathrm{Var}(f) = \mathbb{E}(f - \mathbb{E} f)^2 = \mathbb{E} f^2 - (\mathbb{E} f)^2$. Then the Poincaré constant (of the discrete cube) is the largest λ such that the inequality

$$\mathbb{D}(f,f) \ge \lambda \operatorname{Var}(f)$$

holds for all $f\colon \{-1,1\}^n \to \mathbb{R}$. This quantity is also the smallest nonzero eigenvalue of the Laplacian of the discrete cube, viewed as a graph (i.e., its spectral expansion).

Another way of measuring the non-constantness of a function is to consider its entropy $\mathrm{Ent}(f) = \mathbb{E}[f \log \frac{f}{\mathbb{E} f}]$ (where we assume $f \ge 0$ and use the convention that $0 \log 0 = 0$). Note that $\mathrm{Ent}(cf) =$

$c\,\mathrm{Ent}(f)$ for any $c \geq 0$, so the entropy is homogenous of degree 1 in its argument. Because we are comparing the entropy with the energy (which is homogenous of degree 2) we use the entropy of the *square* of the function to define the Log-Sobolev constant: the largest α such that the inequality

$$\mathbb{D}(f,f) \geq \alpha\,\mathrm{Ent}(f^2)$$

holds for all $f\colon \{-1,1\}^n \to \mathbb{R}$. For the discrete cube $\{-1,1\}^n$, we have $\lambda = 2/n$ and $\alpha = 1/n$, as we shall see below. It is interesting to ask how these quantities are related when we consider other probability spaces equipped with a suitable Dirichlet form (for example, d-regular graphs with $\mathbb{D}(f,g) = \mathbb{E}_{xy \in E}(f(x)-f(y))(g(x)-g(y))$, where the expectation is taken over all edges). Set $f = 1 + \epsilon g$ for a sufficiently small ϵ and observe that $\mathrm{Var}(f) = \epsilon^2\,\mathrm{Var}(g)$ and $\mathbb{D}(f,f) = \epsilon^2\,\mathbb{D}(g,g)$, whereas

$$\begin{aligned}\mathrm{Ent}(f^2) &= \mathbb{E}\left[(1+\epsilon g)^2(2\log(1+\epsilon g) - \log\mathbb{E}[(1+\epsilon g)^2])\right] \\ &= 2\epsilon^2\,\mathrm{Var}(g) + O(\epsilon^3)\end{aligned}$$

This shows that $\alpha \leq \lambda/2$, which is tight in the case of the cube. However, for constant-degree expander families (in particular, for random d-regular graphs with high probability) we have [DSC96, Example 4.2] $\lambda = \Omega(1)$ but $\alpha = O(\log\log n/\log n) \ll \lambda$.

2.2 Hypercontractivity and the log-Sobolev inequality

When $\rho \in [0,1]$, the noise operator T_ρ is easily seen to contract ℓ_2: for any $f\colon \{-1,1\}^n \to \mathbb{R}$, we have $\|T_\rho f\|_2^2 = \sum_S \rho^{|S|}\hat{f}(S)^2 \leq \sum_S \hat{f}(S)^2 = \|f\|_2^2$. Now consider its behavior from ℓ_2 to ℓ_q for some $q > 2$. When $\rho = 1$, we have $T_1 f = f$; in particular, for $g(x) = (1+x_1)/2$, $\|g\|_q = 1/2^{1/q} > 1/2^{1/2} = \|g\|_2$. On the other hand, $T_0 f = \mathbb{E} f$, so $\|T_0 f\|_q = |\mathbb{E} f| \leq \|\mathbb{E} f^2\|^{1/2}$. By the intermediate value theorem, there must be some $\rho \in (0,1)$ such that $\|T_0\|_{2\to q} = 1$. A theorem of Gross [Gro75] connects this critical ρ with the Log-Sobolev constant α of the underlying space (we shall use the language of [DSC96], which is more convenient in the finite setting):

Theorem 2. $\|T_\rho\|_{p\to q} \leq 1$ *if and only if* $\rho^{-2\alpha n} \geq \frac{q-1}{p-1}$.

Stated differently, $\|T_{1-\epsilon} f\|_q \leq \|f\|_2$ when $q \leq (1-\epsilon)^{-2} + 1 \approx 2 + 2\epsilon$. Thus to prove hypercontractive inequalities on the discrete cube, it suffices to bound the log-Sobolev constant. We shall prove this claim for $p = 2$, which turns out to imply the general version.

Proof. We shall prove that $\|T_\rho f\|_q \leq \|f\|_2$ for $q = 1 + \rho^{-2\alpha n}$; the remainder of the theorem can be shown using similar techniques. As we observed before, this inequality is tight when $\rho = 1$, so it suffices to show that $\frac{d}{d\rho}\|T_\rho f\|_q \geq 0$ for $0 \leq \rho \leq 1$. For notational convenience, let $G = \|T_\rho f\|_q^q$. Then

$$\|T_\rho f\|_q' = (G^{1/q})' = q^{-2} G^{(1/q)-1}\left(qG' - q' G\log G\right).$$

Now we use the fact that $G = \mathbb{E}(T_\rho f)^q$ to get

$$G' = q\,\mathbb{E}\left[(T_\rho f)^{q-1}(T_\rho f)'\right] + q'\,\mathbb{E}\left[(T_\rho f)^q \log(T_\rho f)\right].$$

Applying Lemma 3 and simplifying, we get

$$qG' - q' G\log G = q'\,\mathrm{Ent}\left((T_\rho f)^q\right) + \frac{nq^2}{2\rho}\,\mathbb{D}\left((T_\rho f)^{q-1}, T_\rho f\right).$$

We use Lemma 4 to handle the second term, and plug in $q = 1 + \rho^{-2\alpha n}$ to get

$$qG' - q'G \log G = n\rho^{-2\alpha n - 1}\big[\mathbb{D}\big((T_\rho f)^{q/2}, (T_\rho f)^{q/2}\big) - \mathrm{Ent}\left((T_\rho f)^q\right)\big],$$

whose positivity we are guaranteed by the log-Sobolev inequality applied to $(T_\rho f)^{(q-1)/2}$. $\square$

Lemma 3. *For any* $f, g\colon \{-1,1\}^n \to \mathbb{R}$, $\langle g, \frac{d}{d\rho}(T_\rho f)\rangle = \frac{n}{2\rho}\mathbb{D}(g, T_\rho f)$.

Proof. Recalling Lemma 1 and the projection property of the h_is, we have

$$\langle g, (T_\rho f)'\rangle = \langle g, \mathrm{BG}'_\rho * f\rangle = \left\langle g, \frac{1}{\rho}\mathrm{BG}_\rho * f * \sum_i h_i \right\rangle = \frac{1}{\rho}\sum_i \langle g * h_i, \mathrm{BG}_\rho * f\rangle = \frac{n}{2\rho}\mathbb{D}(g, T_\rho f). \quad \square$$

Lemma 4. *For any* $f\colon \{-1,1\}^n \to \mathbb{R}$ *and* $q \geq 2$, $\mathbb{D}(f, f^{q-1}) \geq \frac{4(q-1)}{q^2}\mathbb{D}\left(f^{q/2}, f^{q/2}\right)$.

Proof. It suffices to show that $(a^{q-1} - b^{q-1})(a-b) > \frac{4(q-1)}{q^2}(a^{q/2} - b^{q/2})^2$ for all $a > b \geq 0$ and $q \geq 2$. But observe that

$$\left(\int_a^b t^{q/2-1}dt\right)^2 = \frac{4}{q^2}(a^{q/2} - b^{q/2})^2$$
$$\int_a^b t^{q-2}dt \int_a^b dt = \frac{1}{q-1}(a^{q-1} - b^{q-1})(a-b)$$

and the inequality between the integrals follows from convexity. $\square$

2.3 Tensoring property

Let α_n be the log-Sobolev constant for an n-dimensional cube. Then

Theorem 5. $\alpha_{2n} = \alpha_n/2$

When n is a power of 2, we can conclude inductively that $\alpha_n = \alpha_1/n$; a proof along similar lines works for arbitrary n as well.

Proof. For any $f\colon \{-1,1\}^n \times \{-1,1\}^n \to \mathbb{R}$, set $g(x) = \|f(x,\cdot)\|_2$. Then by the conditional entropy formula,

$$\mathrm{Ent}(f^2) \leq \mathrm{Ent}(g^2) + \mathop{\mathbb{E}}_x \mathop{\mathrm{Ent}}_y(f(x,y)^2) \leq (\mathbb{D}(g) + \mathop{\mathbb{E}}_x \mathbb{D}_y(f(x,y))/\alpha_n$$

and by convexity,

$$\mathbb{D}(G) = \tfrac{1}{2}\mathop{\mathbb{E}}_{x,x'}(G(x) - G(x'))^2 \leq \tfrac{1}{2}\mathop{\mathbb{E}}_{x,x'}\mathop{\mathbb{E}}_y\left[(f(x,y) - f(x',y))^2\right] = \mathop{\mathbb{E}}_y \mathbb{D}(f(\cdot,y))$$

where x, x' range over edges of$\{-1,1\}^n$. Taken together, these give

$$\mathrm{Ent}(f^2) \leq \frac{\mathbb{E}_x \mathbb{D}(f(x,\cdot)) + \mathbb{E}_y \mathbb{D}(f(\cdot,y))}{\alpha_n} \leq \frac{2\,\mathbb{D}(f)}{\alpha_n} = \frac{\mathbb{D}(f)}{\alpha_n/2}.$$

as claimed. $\square$

2.4 Two-point inequality

It remains to show that the log-Sobolev inequality holds for the uniform distribution on the two-point space $\{-1, 1\}$ with $\alpha = 2$. Without loss of generality, consider $f(x) = 1 + sx$. Then

$$\mathrm{Ent}(f^2) = \tfrac{1}{2}(1+s)^2 \log(1+s)^2 + \tfrac{1}{2}(1-s)^2 \log(1-s)^2 - (1+s^2)\log(1+s^2)$$

and $\mathbb{D}(f, f) = 2s^2$. We shall show that $\phi(s) = \mathbb{D}(f, f) - \alpha\,\mathrm{Ent}(f^2)$ is non-negative for $-1 \leq s \leq 1$. By symmetry it suffices to consider $s \geq 0$. But $\phi(0) = 0$ and

$$\phi'(s) = 4s + 2s\log(1+s^2) + 2(1-s)\log(1-s) - 2(1+s)\log(1+s),$$

which is non-negative because $\phi'(0) = 0$ and

$$\phi''(s) = \frac{4s^2}{s^2+1} + 2\log\frac{1+s^2}{1-s^2} \geq 0.$$

2.5 Non-product groups

We have kept the exposition simple and self-contained by stating our definitions in terms of the hypercube $\{-1, 1\}^n$ equipped with Hamming edges and the uniform distribution. The relationship between logarithmic Sobolev inequalities and hypercontractive noise operator subgroups, as stated by Gross [Gro75], holds for a much wider class of spaces. Diaconis and Saloff-Coste [DSC96] explored an intermediate between these two extremes of specialization to give improved mixing time results for Markov chains on various graphs.

One of the first discrete applications of hypercontractivity was a celebrated theorem of Kahn, Kalai and Linial [KKL88] relating the maximum influence of a function on the hypercube to its variance. In recent work [OW09b], O'Donnell and Wimmer have generalized the KKL theorem to apply to the wider class of Schreier graphs associated with group actions (defined below). An action of a group G on a set X is a homomorphism from G to the group of bijections on X, and we write x^g for the image of x under the bijection for g. If S is a set of generators for G, then the Schreier graph $\mathrm{Sch}(G, S, X)$ has vertex set X and edges (x, x^g) for all $x \in X$ and $g \in S$. It is known that every regular graph can be obtained in this way. The definition of the Dirichlet form $\mathbb{D}$ generalizes without change, but to be able to derive a log-Sobolev inequality for this space, we must define the noise operator T_ρ in an appropriate fashion to satisfy the claim of Lemma 1: $\langle g, \frac{d}{d\rho}(T_\rho f)\rangle \propto \frac{1}{\rho}\mathbb{D}(g, T_\rho f)$.

3 Boolean-Valued Functions

3.1 Influences

Write x_{-i} for the collection of random variables $\{x_1, \ldots, x_n\} \setminus \{x_i\}$. The influence of the ith coordinate on a function $f\colon \{-1, 1\}^n \to \mathbb{R}$ is given by $\mathrm{Inf}_i(f) = \mathbb{E}_{x_{-i}} \mathrm{Var}_{x_i} f(x) = \mathbb{E}_{x_{-i}}\left[\mathbb{E}_{x_i} f(x)^2 - (\mathbb{E}_{x_i} f(x))^2\right]$. When f is Boolean-valued, this quantity is just the probability that changing x_i changes $f(x)$. Writing f in the Fourier basis, we have $\mathbb{E}_{x_{-i}} \mathbb{E}_{x_i} f(x)^2 = \mathbb{E}_x f(x)^2 = \sum_S \hat{f}(S)^2$ and $\mathbb{E}_{x_{-i}}(\mathbb{E}_{x_i} f(x))^2 = \sum_{S \not\ni i} \hat{f}(S)^2$, so that $\mathrm{Inf}_i(f) = \sum_{S \ni i} \hat{f}(S)^2 = \mathbb{E}(f * h_i)^2$. In addition, we define the total influence $\mathrm{Inf}(f) = \sum_i \mathrm{Inf}_i(f) = \sum_S |S|\hat{f}(S)^2$.

3.2 Structural results

Boolean functions are natural combinatorial objects, but they were first studied from an analytical viewpoint in work on voting and social choice. In this setting, a function $f\colon \{-1,1\}^n \to \{-1,1\}$ is viewed as a way to combine the preferences of n voters to yield the result of the election. This explains the notions of dictator or junta functions, which depend on only one or a few of their coordinates, respectively. In this context it is also natural to consider functions where no coordinate ("voter") has a very large influence. Kahn, Kalai, and Linial [KKL88] first introduced the Fourier analysis of Boolean functions as a technique in computer science. Their theorem establishes that if a function is far from a constant (i.e., has super-constant variance), then it must have a variable of influence $\Omega(\frac{\log n}{n})$. We state a strengthening of their original inequality due to Talagrand [Tal95]:

Theorem 6 ([KKL88, Tal95]). *For any* $f\colon \{-1,1\}^n \to \{-1,1\}$,

$$\sum_i \frac{\mathrm{Inf}_i(f)}{\log(1/\mathrm{Inf}_i(f))} \geq \Omega(1)\cdot \mathrm{Var}(f).$$

We can compare this to the Poincaré inequality on the cube, which can be stated as

$$\sum_i \mathrm{Inf}_i(f) \geq \Omega(1)\cdot \mathrm{Var}(f)$$

(i.e., there exists a variable of influence $\Omega\left(\frac{1}{n}\right)$.) The KKL theorem is a stronger result of the same form: it is a comparison between a local and a global measure of variation. The proofs of KKL and Talagrand used the hypercontractivity of the cube, but we present here a more recent proof due to Rossignol that uses the log-Sobolev inequality instead. For simplicity we'll just show the weaker statement that the maximum influence is $\Omega\left(\frac{\log n}{n}\right)$.

Proof. Write $f - \mathbb{E} f = f_1 + \cdots + f_n$, where $f_j = \sum_{S:\max S = j} \hat{f}(S)\chi_S$. For each f_j, the log-Sobolev inequality states that $\mathbb{D}(f_j, f_j) \geq \alpha \,\mathrm{Ent}(f_j^2) = \frac{1}{n}\mathrm{Ent}(f_j^2)$. By writing $\mathbb{D}(f_j, f_j)$ in terms of the Fourier coefficients $\hat{f}(S)$, we can check that $\mathbb{D}(f,f) = \sum_{j=0}^n \mathbb{D}(f_j, f_j)$, so that we can sum all these inequalities to obtain

$$n\,\mathbb{D}(f,f) \geq \sum_j \mathrm{Ent}(f_j^2) = \underbrace{\sum_j \mathbb{E}\left[f_j^2 \log(f_j^2)\right]}_{A} + \underbrace{\sum_j \mathbb{E}\, f_j^2 \log\frac{1}{\mathbb{E}\, f_j^2}}_{B}.$$

In order to bound B, we begin by noting that

$$\mathbb{E}\, f_j^2 = \sum_{S:\max S = j} \hat{f}(S)^2 \leq \sum_{S \ni j} \hat{f}(S)^2 = \mathbb{E}(f * h_j)^2$$

where the h_js are the edge functions we defined earlier. Letting $M(f) = \max_j \mathbb{E}(f * h_j)^2 = \max_j \mathrm{Inf}_j(f)$, we have

$$B = \sum_j \mathbb{E}\, f_j^2 \log\frac{1}{\mathbb{E}\, f_j^2} \geq \sum_j \mathbb{E}\, f_j^2 \log\frac{1}{M(f)} = \mathrm{Var}(f)\log\frac{1}{M(f)}$$

where we have used the orthogonality of the f_js and the fact that $\mathrm{Var}(f) = \sum_{S\neq\emptyset} \hat{f}(S)^2$.

To bound A, we split it up further:

$$A = \underbrace{\sum_j \mathbb{E}\left[f_j^2 \log(f_j^2) \cdot 1_{f_j^2 \le t}\right]}_{A_1} + \underbrace{\sum_j \mathbb{E}\left[f_j^2 \log(f_j^2) \cdot 1_{f_j^2 > t}\right]}_{A_2}.$$

For $0 \le t \le 1/e^2$, we have that $\sqrt{t}\log\sqrt{t}$ is a nonpositive decreasing function and therefore,

$$A_1 = 2\sum_j \mathbb{E}\left[|f_j|\log|f_j| \cdot |f_j| 1_{f_j^2 \le t}\right] \ge 2\sqrt{t}\log\sqrt{t}\sum_j \mathbb{E}\,|f_j \cdot 1_{f_j^2 \le t}| \ge \sqrt{t}\log t \sum_j \mathbb{E}\,|f_j|.$$

By comparing Fourier coefficients, it is easy to verify that $f_j = \mathbb{E}_{x_{j+1},\ldots,x_n}(f * h_j)$. Therefore, by convexity, $\mathbb{E}\,|f_j| \le \mathbb{E}\,|f * h_j|$.

Until now, the proof has made no use of the fact that f takes on only Boolean values. Now we argue that because $f(x) \in \{-1,1\}$, we must have $(f * h_j)(x) \in \{-1,0,1\}$, so that $\mathbb{E}\,|f * h_j| = \mathbb{E}(f * h_j)^2$. Plugging this into our bound for A_1 yields

$$A_1 \ge \sqrt{t}\log t \sum_j \mathbb{E}(f * h_j)^2 = \frac{n}{2}\sqrt{t}\log t \cdot \mathbb{D}(f,f).$$

For A_2, note that $\log(\cdot)$ is increasing, so

$$A_2 \ge \log t \sum_j \mathbb{E}\, f_j^2 = \log t \operatorname{Var} f.$$

Summing all these bounds gives us

$$n\,\mathbb{D}(f,f) \ge \log\frac{1}{M(f)}\operatorname{Var}(f) + \frac{n}{2}\sqrt{t}\log t \cdot \mathbb{D}(f,f) + \log t \cdot \operatorname{Var}(f).$$

By the Poincaré inequality, $\mathbb{D}(f,f) \ge \frac{2}{n}\operatorname{Var}(f)$, so we can set $t = \left(\frac{2\operatorname{Var}(f)}{ne\,\mathbb{D}(f,f)}\right)^2 \le 1/e^2$. With this substitution, the above inequality becomes

$$\frac{2}{e\sqrt{t}} \ge \log\frac{t^{1+1/e}}{M(f)}.$$

INCOMPLETE □

We are now in a position to state the recent result of O'Donnell and Wimmer [OW09b] generalizing the KKL theorem to Schreier graphs satisfying a certain technical property.

Theorem 7 ([OW09b])**.** *Let G be a group acting on a set X, $U \subseteq X$ be a union of conjugacy classes that generates G, and α be the log-Sobolev constant of $\operatorname{Sch}(G,X,U)$. Then for any $f\colon X \to \{-1,1\}$,*

$$\frac{\sum_U \operatorname{Inf}_u(f)}{\log(1/\max_U \operatorname{Inf}_u(f))} \ge \Omega(\alpha \operatorname{Var}(f)).$$

In particular, there is some $u \in U$ such that $\operatorname{Inf}_u(f) \ge \Omega(\rho \log\frac{1}{\rho})\operatorname{Var}(f)$.

For an Abelian group such as $\mathbb{Z}_2^n$ (the cube), every group element is in a conjugacy class by itself, so the extra condition on U is vacuous. Using $\alpha = \Omega(\frac{1}{n})$ for the cube, we recover the original KKL theorem. O'Donnell et al. apply the generalized result to the non-Abelian group S_n of permutations on $[n]$, generated by transpositions and acting on the family $\binom{[n]}{k}$ of k-subsets of $[n]$. By viewing these families as sets of n-bit strings, they recover a "rigidity" version of the Kruskal-Katona theorem that states (roughly) that if a subset of a layer of a cube has a small expansion to the layer above it, then it must be correlated to some dictator function.

Coding theoretic interpretation. In the *long code*, an integer $i \in [n]$ is encoded as the dictator function $(x_1, \ldots, x_n) \mapsto x_i$. By using many more bits (n rather than $\log n$) of redundant storage, we hope to be able to recover from corruptions in the data. The theorem tells us that as long as the corrupted version of an encoding is far from a constant function, it can be decoded to a coordinate whose influence is $\Omega(\log n)$ times the average influence. Clearly there can be only $O(\log n)$ such coordinates, so we have a "small" set of candidate long codes to which we might decode the word. To complete this picture, we'd like to understand how far the word can be from functions that depend only on these coordinates; the following two theorems of Friedgut, which we state without proof, furnish this information.

Theorem 8 ([Fri98]). *For every $f\colon \{-1,1\}^n \to \{-1,1\}$ and $0 < \epsilon < 1$, there is a function $g\colon \{-1,1\}^n \to \{-1,1\}$ depending on at most $\mathrm{Inf}(f)/\epsilon$ variables such that $\mathbb{E}\,|f-g| \leq \epsilon$.*

If $f\colon \{-1,1\}^n \to \{-1,1\}$ has the property that $\sum_{|S|>1} \hat{f}(S)^2 \leq \epsilon$, then there is a function $g\colon \{-1,1\}^n \to \{-1,1\}$ depending on at most one variable such that $\mathbb{E}\,|f-g| \leq \epsilon$.

4 Gaussian isoperimetry and an algorithmic application

Hypercontractive inequalities were first investigated in the context of Gaussian probability spaces, for their applications to quantum field theory. The following simple proof reduces the continuous Gaussian hypercontractive inequality to its discrete counterpart on the cube.

4.1 From the central limit theorem to Gaussian hypercontractivity

Theorem 9 ([Gro75]). *Let $x \in \mathbb{R}$ be normally distributed, i.e.,*

$$\Pr[x \in A] = \frac{1}{\sqrt{2\pi}} \int_A \exp\left(-\frac{x^2}{2}\right) dx.$$

Then for a smooth function $f\colon \mathbb{R} \to \mathbb{R}$, the random variable $F = f(x)$ satisfies

$$\mathbb{D}(F,F) \geq \alpha \,\mathrm{Ent}(F^2)$$

with $\alpha = 1$ and

$$\mathbb{D}(F,G) = \left\langle \frac{dF}{dx}, \frac{dG}{dx} \right\rangle.$$

Proof. We shall approximate the Gaussian distribution by a weighted sum of Bernoulli variables. Let $y \in \{-1,1\}^k$ be uniformly distributed, and set $g(y) = \frac{y_1 + \cdots + y_k}{\sqrt{k}}$. By the log-Sobolev inequality applied to $f \circ g(y)$, we have $\mathbb{D}\left(f \circ g(y), f \circ g(y)\right) \geq \mathrm{Ent}(f \circ g(y)^2)$. By the central limit theorem, the right side converges to $\mathrm{Ent}(f(x)^2) = \mathrm{Ent}(F^2)$ as $k \to \infty$, so it remains to show that the left side converges to $\mathbb{D}(F,F)$ as well. Let $y|_{y_i=\theta}$ be the value obtained by replacing the ith coordinate of y with the value θ, and observe that $g(y|_{y_i=1}) - g(y|_{y_i=-1}) = 2/\sqrt{k}$. Then, using the smoothness of f, we have

$$|\left(h_i * (f \circ g)\right)(y)| = \frac{1}{2}\,|f \circ g(y|_{y_i=1}) - f \circ g(y|_{y_i=-1})| = \frac{1}{\sqrt{k}}\left|f' \circ g(y)\right| + o\left(\frac{1}{\sqrt{k}}\right),$$

so that

$$\mathbb{D}\left(f \circ g(y), f \circ g(y)\right) = \mathop{\mathbb{E}}_y\left[\sum_i \left(h_i * (f \circ g)\right)(y)^2\right] = \mathop{\mathbb{E}}_y\left[f' \circ g(y)^2 + o(1)\right].$$

The second term vanishes as $k \to \infty$, and the first term converges to $\mathbb{D}(F, F)$ by the Central Limit Theorem. ☐

The tensoring property of log-Sobolev inequalities lets us extend this result to Gaussian distributions over $\mathbb{R}^d$. We are also interested in the corresponding noise operator S_ρ, given by $S_\rho f(x) = \mathbb{E}_y f(y)$, where $y \in \mathbb{R}^d$ is a Gaussian random variable that is ρ-correlated with x (it is called the Ornstein-Uhlenbeck operator). By the analog of Theorem 2 in this setting, we conclude that every function $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfies $\|T_\rho f\|_q \leq \|f\|_p$ where $q > p \geq 1$ and $\rho^{-4} \geq (p-1)/(q-1)$.

4.2 Reverse hypercontractivity and isoperimetry

In 1982, Borell showed a *reversed* inequality of a similar form when $q < p < 1$:

Theorem 10 (Reverse hypercontractivity, [Bor82])**.** *Fix $q < p \leq 1$ and $\rho \geq 0$ such that $\rho^{-4} \geq (p-1)/(q-1)$. Then for any positive-valued function $f \colon \mathbb{R}^d \to \mathbb{R}^+$, we have $\|T_\rho f\|_q \geq \|f\|_p$.*

Note that the expressions $\|\cdot\|_p$ are not norms when $p < 1$; in particular, they are not convex. However, this theorem can be proved by means similar to our proof for the Gaussian log-Sobolev inequality: we start with a base result for the 2-point space, proceed by tensoring to the hypercube, and use the central limit theorem to cover Gaussian space.

As an application of Borell's result, consider the following strong isoperimetry theorem for Gaussian space (due to Sherman).

Theorem 11 (Gaussian isoperimetry, [She09])**.** *Let $u, u' \in \mathbb{R}^d$ be ρ-correlated Gaussian random variables. Then for any set $A \subseteq \mathbb{R}^d$ and any $\tau > 0$, we have*

$$\Pr_u \left[\Pr_{u'}[u' \in A] \leq \tau \right] \leq \frac{\tau^{1-\rho}}{\mu(A)}$$

Proof. When $\mu(A) \leq \tau^{1-\delta}$, there is nothing to prove. Otherwise, let f be the indicator function of A and observe that $\Pr_{u,u'}[u' \in A] = S_\rho f(u)$. Therefore, for $q = 1 - 1/\rho < 0$, we have

$$\begin{aligned} \Pr_u \left[\Pr_{u'}[u' \in A] \leq \tau \right] &= \Pr_u [S_\rho f(u) \leq \tau] \\ &= \Pr_u [S_\rho f(u)^q \geq \tau^q] \\ &\leq \frac{\mathbb{E}_u (S_\rho f(u))^q}{\tau^q} \end{aligned}$$

by an application of Markov's inequality. But $\mathbb{E}_u (S_\rho f(u))^q$ is just $\|S_\rho f\|_q^q$, and we know by Borell's theorem that $\|S_\rho f\|_q \geq \|f\|_p$ for $p = 1 - \rho$. Thus

$$\Pr_u \left[\Pr_{u'}[u' \in A] \leq \tau \right] \leq \frac{\|f\|_p^q}{\tau^q} = \frac{\mu(A)^{q/p}}{\tau^q} = \left(\frac{\tau^{1-\rho}}{\mu(A)} \right)^{1/\rho} \leq \frac{\tau^{1-\rho}}{\mu(A)}$$

where we have used the facts that $q < 0$ and $\rho \leq 1$. ☐

4.3 Fast graph partitioning and the constructive Big Core Theorem

Problem and SDP rounding algorithm. In the c-balanced separator problem, we are given a graph G on n vertices and asked to find the smallest set of edges such that their removal disconnects the graph into pieces of size at most cn. The problem is NP-hard, and the best known approximation ratio[1] is $\Theta(\sqrt{\log n})$.

The first algorithm to achieve this bound was based on a semidefinite program that assigns a unit vector to each vertex and minimizes the total embedded squared length of the edges subject to the constraint that every large (compared to cn) set of vertices is spread out, and that the squared distances between the points form a metric. To round this SDP, Arora, Rao and Vazirani [ARV09] pick a random direction u and project all the points along u. They then define sets A and B consisting of points x whose projections are sufficiently large, i.e., $A = \{x \mid \langle x, u\rangle < -K\}$ and similarly $B = \{x \mid \langle x, u\rangle > K\}$, where K is chosen to make A and B have size $\Theta(n)$ with high probability. Next, they discard points $a \in A, b \in B$ such that $\|a - b\|$ is much smaller than expected for a pair whose projections are $\geq 2K$ apart. Finally, if the resulting pruned sets $A' \subset A$ and $B' \subset B$ are large enough, they show that greedily growing A yields a good cut.

Matchings and cores. The key step in making this argument work is to ensure that not too many pairs (a, b) are removed in the pruning step. To bound the probability of this bad event, we consider the possibility that for a large fraction $\delta = \Omega(1)$ of directions u, there exists a matching of points M_u such that each pair $(a, b) \in M_u$ is short (i.e., $\|a - b\| \leq \ell = O(1/\sqrt{\log n})$) but stretched along u (i.e., $|\langle a - b, u\rangle| \geq \sigma = \Omega(1)$). Such a set of points is called a (σ, δ, ℓ)-*core*. The big core theorem (first proved with optimal parameters by Lee [Lee05]) asserts that this situation can't arise: for a fixed σ, δ, and ℓ, we must have $n \gg \exp(\sigma^6/\ell^4 \log^2(1/\delta))$, which is a contradiction for our chosen values of σ, δ, ℓ.

In order to prove the big core theorem, Lee concatenates pairs that share a point and belong in matchings for nearby directions. The existence of a long chain of such concatenations is what leads to a contradiction: if we consider the endpoints a, b of a chain of length p, the projection $|\langle a - b, u\rangle|$ grows linearly in p whereas the distance $\|a - b\|$ grows only as $\sqrt{p}$ (recall that the SDP constrained the *squared* distances to form a metric).

Boosting. The matching chaining argument we have just presented in its simple form doesn't work, for the following reason. At each chaining step, the fraction of nearby directions available for our use reduces by roughly $1 - \delta$ (by a union bound) so that we are rapidly left with no direction to move in. To remedy this situation, we need to boost the fraction of usable directions at each step, say from $\delta/2$ to $1 - \delta/2$, so that we can carry on chaining in spite of a $1 - \delta$ loss. Lee's proof uses the standard isoperimetric inequality for the sphere to show that this boosting can be performed with no change in ℓ and a very small penalty in σ. In other words, we take advantage of the fact that a very small dilation of a set of constant measure (i.e., the set of available directions) has measure close to 1.

Faster algorithms. Lee's big core theorem is non-constructive in the sense that it only shows the *existence* of such a long chain of matched pairs in order to give a contradiction. While this

[1] For technical reasons, it is actually a pseudo-approximation: the algorithm's output for c is compared to the optimal value for $c' \neq c$.

form suffices to bound the approximation ratio of the ARV rounding scheme, other variants of their technique require a way to *efficiently sample* long chains, not just show their existence. Sherman shows how to perform this sampling in a way that is oblivious to the point set, and uses it to obtain the fastest known algorithm for sparsest cut and graph partitioning.

Theorem 12 (Constructive big core [She09])**.** *For any $1 \le R \le \Theta(\sqrt{\log n})$, there is $P \ge \Theta(R^2/\log n)$ and an efficiently sampleable distribution μ over $(\mathbb{R}^d)^{\le P}$ such that: for any (σ, δ, ℓ)-core M, if the string of directions is sampled from μ, the expected number of chains whose endpoints are $\ge P\ell$ apart is at least $\exp(-O(P^2)n)$.*

We sketch some of the ideas of the proof here. Sherman constructs two sequences of Gaussian directions $u_1, \ldots, u_P$ and $w_1, \ldots, w_P$. Each w_i is an independent Gaussian vector, whereas each u_i for $i > 1$ is a Gaussian vector ρ-correlated with u_{i-1}. Finally, the distribution μ is given by randomly shuffling together the u_i and w_i, picking a uniformly random R between 1 and P, and returning the first R elements of the shuffled sequence. The correlated directions u_i correspond to the steps in which Lee's proof chained pairs from similar directions, whereas the independent w_i correspond to the region-growing steps necessary for boosting. By randomly interleaving these two types of moves, Sherman's sampling algorithm can be oblivious to the actual point set it is acting on.

5 Complexity theoretic applications

5.1 Dictatorship testing with perfect completeness

Definitions. A function $f\colon \{-1,1\}^n \to \mathbb{R}$ is said to be (ϵ, δ)-quasirandom if $\hat{f}(S) \le \epsilon$ whenever $|S| \le 1/\delta$. In order to show that a given problem is hard to approximate, we often need to design a test that

- performs q *queries* on a black-box function f,
- accepts every dictator function with probability $\ge c$ (the *completeness* probability), and
- accepts every (ϵ, δ)-quasirandom function with probability $\le s$ (the *soundness* probability).

A test is said to be *adaptive* if each query is allowed to depend on the result of the queries so far.

While dictatorship tests for the $c < 1$ setting have been known for over a decade (first from the work of Håstad and more recently via the Unique Games Conjecture of Khote), there were no nontrivial bounds for $c = 1$ until some recent results of O'Donnell and Wu. Their analysis, which we show below, relies heavily on the hypercontractive inequality.

Theorem 13 ([OW09a])**.** *For every $n > 0$, there is a 3-query non-adaptive test that accepts every dictator function $(x_1, \ldots, x_n) \mapsto x_i$ with probability $c = 1$ but accepts any $(\delta, \delta/\log(1/\delta))$-quasirandom odd function $f\colon \{-1,1\}^n \to [-1,1]$ with probability $\le s = 5/8 + O(\sqrt{\delta})$.*

The proof uses the following strengthening of the hypercontractive inequality for restricted parameter values.

Lemma 14. *If $0 \le \rho \le 1$, $q \ge 1$, and $0 \le \lambda \le 1$ satisfy $\rho^\lambda \le 1/\sqrt{q-1}$, then for all $f\colon \{-1,1\}^n \to \mathbb{R}$, $\|T_\rho f\|_q \le \|T_\rho f\|_2^{1-\lambda} \|f\|_2^{\lambda}$.*

Proof.

$$\begin{aligned}\|T_\rho f\|_q^2 &= \|T_{\rho^\lambda} T_{\rho^{1-\lambda}} f\|_q^2 \\ &= \|T_{\rho^{1-\lambda}} f\|_2^2 \\ &= \sum_S |\rho \hat{f}(S)|^{2(1-\lambda)} |\hat{f}(S)|^{2\lambda} \\ &= \|T_\rho f\|_2^{2(1-\lambda)} \|f\|_2^{2\lambda} \square\end{aligned}$$

Proof of Theorem 13. Define the "not-two" predicate $\texttt{NTW}\colon \{-1,1\}^3 \to \{-1,1\}$ as follows: $\texttt{NTW}(a,b,c) = 1$ if exactly two of a,b,c equal -1, and $\texttt{NTW}(a,b,c) = -1$ otherwise. Explicitly,

a	-1	-1	-1	-1	1	1	1	1
b	-1	-1	1	1	-1	-1	1	1
c	-1	1	-1	1	-1	1	-1	1
$\texttt{NTW}(a,b,c)$	-1	1	1	-1	1	-1	-1	-1

Let $\delta \in [0,1]$ be a parameter to be fixed later. For $i = 1,\ldots,n$, we pick bits $x_i, y_i, z_i \in \{-1,1\}$ as follows:

- with probability $1-\delta$: we choose x_i, y_i uniformly and independently, then set $z_i = -x_i y_i$;
- with probability δ: we choose x_i uniformly, then set $y_i = z_i = x_i$.

Note that for $i \neq j$, (x_i, y_i, z_i) is independent of (x_j, y_j, z_j). We accept if $\texttt{NTW}(f(x), f(y), f(z)) = -1$. It is immediate from the construction of x_i, y_i, z_i that $\texttt{NTW}(x_i, y_i, z_i) = -1$ for $i = 1,\ldots,n$. Therefore, if f is a dictator function, it follows that $\texttt{NTW}(f(x), f(y), f(z))$ must also equal -1.

Soundness. It remains to analyze the test when f is pseudorandom. We begin by writing $\texttt{NTW}$ in the Fourier basis: $\texttt{NTW} = -\frac{1}{4}\chi_\emptyset - \frac{1}{4}(\chi_{\{1\}} + \chi_{\{2\}} + \chi_{\{3\}}) - \frac{1}{4}(\chi_{\{1,2\}} + \chi_{\{2,3\}} + \chi_{\{1,3\}}) + \frac{3}{4}\chi_{\{1,2,3\}}$. Therefore, by symmetry,

$$\mathop{\mathbb{E}}_{x,y,z} \texttt{NTW}(f(x), f(y), f(z)) = -\tfrac{1}{4} - \tfrac{3}{4} \mathop{\mathbb{E}}_{x} f(x) - \tfrac{3}{4} \mathop{\mathbb{E}}_{x,y} f(x)f(y) + \tfrac{3}{4} \mathop{\mathbb{E}}_{x,y,z} f(x)f(y)f(z).$$

We shall systematically rewrite the right-hand side in terms of the Fourier coefficients of f. By our assumption that f is odd, we have $\hat{f}(S) = 0$ whenever S has even cardinality. Therefore $\mathbb{E} f(x) = \hat{f}(\emptyset) = 0$. Also,

$$\mathop{\mathbb{E}}_{x,y} f(x)f(y) = \sum_{S,T} \hat{f}(S)\hat{f}(T) \mathop{\mathbb{E}}_{x,y} \chi_S(x)\chi_T(y).$$

Consider a summand where $S \neq T$, and without loss of generality fix $i \in S \setminus T$. It is easy to see that the contributions due to $x_i = \pm 1$ cancel each other. Thus, the only terms that remain are of the form $S = T$, i.e.,

$$\mathop{\mathbb{E}}_{x,y} f(x)f(y) = \sum_S \hat{f}(S)^2 \mathop{\mathbb{E}}_{x,y} \chi_S(x)\chi_S(y) = \sum_S \hat{f}(S)^2 \left(\mathop{\mathbb{E}}_{x_i,y_i} x_i y_i\right)^{|S|} = \sum_S \hat{f}(S)^2 \delta^{|S|},$$

where we have used the fact that $\mathbb{E}(x_i y_i) = (1-\delta)\cdot 0 + \delta \cdot 1 = \delta$. But $\hat{f}(S)$ is nonzero only for $|S|$ odd, and $\sum_S \hat{f}(S)^2 = 1$, so we can upper-bound the above sum by δ.

Bounding the cubic term. We proceed similarly:

$$\mathop{\mathbb{E}}_{x,y,z} f(x)f(y)f(z) = \sum_{S,T,U} \hat{f}(S)\hat{f}(T)\hat{f}(U) \mathop{\mathbb{E}}_{x,y,z} \chi_S(x)\chi_T(y)\chi_U(z). \tag{1}$$

Each of the expectations can be written as a product over coordinates $i \in [n]$ using the fact that individual coordinates of x, y, z are chosen independently. When i belongs to exactly one of S, T, U (say S), then it contributes a factor $\mathbb{E}\, x_i = 0$, making the product zero. Similarly, when i belongs to two of the sets (say S, T), then the contribution is $\mathbb{E}\, x_i y_i = \delta$ by our earlier calculation. Finally, when i belongs to all three of the sets, we have $\mathbb{E}\, x_i y_i z_i = (1-\delta)\cdot(-1) + \delta\cdot(0) = -(1-\delta)$. In light of this calculation, any triple S, T, U that makes a nonzero contribution to the sum (1) must be of the form

$$S = A \cup B \cup C \qquad\qquad T = A \cup C \cup D \qquad\qquad U = A \cup D \cup B$$

for suitable sets $A, B, C, D \subseteq [n]$ where A is disjoint from B, C, D. Also $|S|, |T|, |U|$ must be odd, from which we can show that $|A|$ must be odd. In terms of these new sets we can rewrite

$$\mathop{\mathbb{E}}_{x,y,z} f(x)f(y)f(z) = -\sum_{\substack{B,C,D \text{ disj. from } A\\ |A| \text{ odd}}} \hat{f}(A\cup B\cup C)\hat{f}(A\cup C\cup D)\hat{f}(A\cup D\cup B)(1-\delta)^{|A|}\delta^{|B|+|C|+|D|}.$$

For a fixed A, define the function $g_A\colon \{-1,1\}^{[n]\setminus A} \to \mathbb{R}$ by $\hat{g}_A(X) = \hat{f}(A \cap X)$. Then we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x,y,z} f(x)f(y)f(z)\\
&= -\sum_{|A| \text{ odd}} (1-\delta)^{|A|} \sum_{\substack{B,C,D\\ \text{disj. from } A}} \hat{g}_A(B\cup C)\sqrt{\delta}^{|B\cup C|}\cdot \hat{g}_A(C\cup D)\sqrt{\delta}^{|C\cup D|}\cdot \hat{g}_A(D\cup B)\sqrt{\delta}^{|D\cup B|}\\
&= -\sum_{|A| \text{ odd}} (1-\delta)^{|A|} \sum_{\substack{B,C,D\\ \text{disj. from } A}} \widehat{T_{\sqrt{\delta}}g_A}(B\cup C)\cdot \widehat{T_{\sqrt{\delta}}g_A}(C\cup D)\cdot \widehat{T_{\sqrt{\delta}}g_A}(D\cup B)\\
&= -\sum_{|A| \text{ odd}} (1-\delta)^{|A|} \|T_{\sqrt{\delta}} g_A\|_3^3.
\end{aligned}$$

Write $g_A(u) = \mathbb{E}_x\, g_A(u) + \tilde{g}_A(u) = \hat{f}(A) + \tilde{g}_A(u)$. Then, using the inequality $|a+b|^3 \le 4(|a|^3 + |b|^3)$, we have

$$\|T_{\sqrt{\delta}} g_A\|_3^3 = \|\hat{f}(A) + \tilde{g}_A\|_3^3 \le 4|\hat{f}(A)|^3 + 4\|\tilde{g}_A\|_3^3$$

and therefore,

$$\sum (1-\delta)^{|A|}\|T_{\sqrt{\delta}} g_A\|^3 = 4\sum (1-\delta)^{|A|}|\hat{f}(A)|^3 + 4\sum (1-\delta)^{|A|}\|\tilde{g}_A\|_3^3.$$

To bound the first term, note that $\sum (1-\delta)^{|A|}|\hat{f}(A)|^3 \le \sum \hat{f}(A)^2 \cdot \max\{(1-\delta)^{|A|}|\hat{f}(A)|)\}$. The sum of the squared Fourier coefficients is just 1 (by Parseval's identity) and we can use the $(\delta, \frac{\delta}{\log(1/\delta)})$-pseudorandomness property to bound the quantity in the maximum: when $|A| < \frac{1}{\delta}\log\frac{1}{\delta}$, then $|\hat{f}(A)| \le \sqrt{\delta}$ and when $|A| \ge \frac{1}{\delta}\log\frac{1}{\delta}$ then $(1-\delta)^{|A|} \le \delta$. Thus the entire first summand is $O(\sqrt{\delta})$.

Hypercontractivity. It remains to bound $\sum(1-\delta)^{|A|}\|\tilde{g}_A\|_3^3$. Fix $\lambda = \frac{\log 2}{\log(1/\delta)}$ and apply the modified hypercontractive inequality:

$$\sum(1-\delta)^{|A|}\|T_{\sqrt{\delta}}\tilde{g}_A\|_3^3 \leq \sum(1-\delta)^{|A|}\|T_{\sqrt{\delta}}\tilde{g}_A\|_2^{3-3\lambda}\|\tilde{g}_A\|_2^{3\lambda}$$

Now, $\|\tilde{g}_A\|_2^{3\lambda} \leq 1$ and $\|T_{\sqrt{\delta}}\tilde{g}_A\|_2^{3-3\lambda} = O(\sqrt{\delta})\sum_{\emptyset\neq B\subseteq\overline{A}}\delta^{|B|}\hat{f}(A\cup B)^2$. The contribution of the corresponding term to the sum we were trying to bound is $O(\sqrt{\delta})\cdot\hat{f}(A\cup B)^2\cdot(1-\delta)^{|A|}\delta^{|B|}$. For each choice of $A\cup B$, the $(1-\delta)^{|A|}\delta^{|B|}$ terms sum to at most one, and all the $\hat{f}(A\cup B)^2$ terms themselves sum to at most one. Therefore, we have bounded the entire sum by $O(\sqrt{\delta})$ as desired. □

6 Integrality gap for Unique Label Cover SDP

Problem and SDP relaxation. In the Unique Label Cover problem, we are given a label set L and a weighted multigraph $G=(V,E)$ whose edges are labeled by permutations $\{\pi_e\colon L\to L\}_{e\in E}$, and are asked to find an assignment $f\colon V\to L$ of labels to edges that maximizes the fraction of edges $e\{u,v\}$ that are "consistent" with our labeling, i.e., $\pi_e(f(u))=f(v)$. If there exists a labeling that satisfies all the edges, then it is easy to find such a labeling. However, when all we can guarantee is that 99% fraction of the edges can be satisfied, it is not known how to find a labeling satisfying even 1% of them. At the same time, present techniques cannot show that finding a 1%-consistent labeling is NP-hard.

One approach to solving this problem is to use an extension of the Goemans-Williamson SDP for Max-Cut, where we set up a vector v_i for every vertex v and label i:

$$\begin{aligned}
\text{maximize } \; & \mathbb{E}_{e\{u,v\}}\textstyle\sum_{i\in L}\langle u_i, v_{\pi_e(i)}\rangle & \\
\text{subject to } \; & \langle u_i, v_j\rangle \geq 0 & \forall u,v\in V, \forall i,j\in L\\
& \textstyle\sum_{i\in L}\langle v_i, v_i\rangle = 1 & \forall v\in V\\
& \langle \textstyle\sum_{i\in L} u_i, \sum_{j\in L} v_j\rangle = 1 & \forall u,v\in L\\
& \langle v_i, v_j\rangle = 0 & \forall v\in V, \forall i\neq j\in L
\end{aligned}$$

(The expectation in the objective is over a distribution where $e\{u,v\}$ is picked with probability proportional to its weight.) The intent is that $\|v_i\|^2$ should be the probability that v receives label i, and $\langle u_i, v_j\rangle$ should be the corresponding joint probability. It is easy to see that this SDP is a relaxation of the original problem.

Gap instance. In an influential paper, Khot and Vishnoi [KV05] constructed an integrality gap for this SDP: for a label set of size 2^k and an arbitrary parameter $\eta\in[0,\frac{1}{2}]$, a graph whose optimal labeling satisfies $\leq 1/2^{\eta k}$ fraction of the edges, but for which the SDP optimum is at least $1-\eta$. The hypercontractive inequality plays a central role in the soundness analysis, which we present below.

Let $\tilde{V}$ be the set of all functions $f\colon\{-1,1\}^k\to\{-1,1\}$ and L be the Fourier basis $\{\chi_S \mid S\subseteq[k]\}$; clearly, $|L|=2^k$. Observe that $\tilde{V}$ is an Abelian group under pointwise multiplication, and L is a subgroup. We take the quotient $V=\tilde{V}/L$ to be the vertex set. Fix an arbitrary representative for each coset and write $V=\{f_1L, f_2L,\ldots,f_{|V|}L\}$. We shall define a weighted edge between every pair of these representative functions, then show how to extend this definition to all pairs of functions, and finally map these edges to edges between cosets.

- The edge $\tilde{e}\{f, g\}$ has weight equal to $\Pr_{h,h'}[(f, g) = (h, h')]$, where $h, h' \in V$ are drawn to be ρ-correlated on every bit with uniform marginals, where $\rho = 1 - 2\eta$.
- With every edge $\tilde{e}\{f_i, f_j\}$ between representative functions, we associate the identity permutation.
- A non-representative function acts as if its label is assigned according to its coset's representative. Thus, the permutation associated with $\tilde{e}\{f_i\chi_S, f_j\chi_T\}$ is $\chi_U\chi_S \mapsto \chi_U\chi_T$.
- In the actual graph under consideration, every edge $\tilde{e}\{f_i\chi_S, f_j\chi_T\}$ appears as an edge $e\{f_iL, f_jL\}$ (with the same permutation and weight).

Soundness analysis. Given a labeling $R\colon V \to L$ on the cosets, we consider the induced labeling $\tilde{R}\colon \tilde{V} \to L$ given by $\tilde{R}(f_i\chi_S) = R(f_iL)\chi_S$. From our definitions, it is clear that the objective value attained by $\tilde{R}$ is precisely $\Pr_{h,h'}[\tilde{R}(h) = \tilde{R}(h')]$, where h, h' are chosen as before. Fix any label χ_S and consider the indicator function $\phi\colon \tilde{V} \to \{0, 1\}$ of functions that $\tilde{R}$ labels with χ_S. Since exactly one function in each coset gets labeled χ_S, we know that $\mathbb{E}\,\phi = 1/2^k$. Therefore,

$$\Pr_{h,h'}[\tilde{R}(h) = \tilde{R}(h') = \chi_S] = \mathop{\mathbb{E}}_{h,h'}[\phi(h)\phi(h')] = \langle h, T_\rho h\rangle = \|T_{\sqrt{\rho}}h\|_2^2,$$

which we can upper-bound (using hypercontractivity) by $\|h\|_{1+\rho}^2 = 1/2^{\frac{2k}{1+\rho}} \leq 1/2^{\eta k}$.

References

[ARV09] Sanjeev Arora, Satish Rao, and Umesh V. Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56(2), 2009.

[Bor82] C. Borell. Positivity improving operators and hypercontractivity. *Mathematische Zeitschrift*, 180(3):225–234, 1982.

[DSC96] P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

[Fri98] E. Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):27–35, 1998.

[Gro75] L. Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[KKL88] J. Kahn, G. Kalai, and N. Linial. The influence of variables on Boolean functions. In *Proceedings of 29th IEEE Symp. Foundations of Computer Science (FOCS)*, 1988.

[KV05] Subhash Khot and Nisheeth K. Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into ℓ_1. In *FOCS*, pages 53–62. IEEE Computer Society, 2005.

[Lee05] James R. Lee. On distance scales, embeddings, and efficient relaxations of the cut cone. In *SODA*, pages 92–101. SIAM, 2005.

[OW09a] Ryan O'Donnell and Yi Wu. 3-bit dictator testing: 1 vs. 5/8. In Claire Mathieu, editor, *SODA*, pages 365–373. SIAM, 2009.

[OW09b] R. ODonnell and K. Wimmer. KKL, Kruskal-Katona, and monotone nets. In *FOCS*, 2009.

[She09] J. Sherman. Breaking the Multicommodity Flow Barrier for $O(\sqrt{\log n})$-Approximations to Sparsest Cut. *University of California, Berkeley*, 2009.

[Tal95] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathematiques de l'IHES*, 81(1):73–205, 1995.