CSC2414: Discrepancy Theory in Computer Science Fall 2015

Lecture 6 — October 26th, 2015

Aleksandar Nikolov *Scribe: Robert Robere*

1 Introduction

If A is an $m \times n$, real-valued matrix recall that the *discrepancy* of A is

$$\text{disc}(A) := \min_{x \in \{\pm 1\}^n} ||Ax||_\infty.$$

We have seen in previous lectures that matrix discrepancy generalizes normal combinatorial discrepancy (if $\mathcal{F}$ is a set system with m sets, take A to be the 0-1 incidence matrix of $\mathcal{F}$, and the ± 1 vector x is the colouring of the underlying elements), which means that the study of $\text{disc}(A)$ is particularly interesting when the entries of A are bounded in the interval $[-1, 1]$. In this regime we can get an upper bound of $\text{disc}(A) = O(\sqrt{n \log m})$ by taking a uniformly random vector $x \in \{\pm 1\}^n$; the seminal "Six Standard Deviations Suffice" theorem (proved in the last lecture) improves this when m is small:

Theorem 1 (Six Standard Deviations Suffice [3])**.** *Let A be any $m \times n$ matrix with entries bounded by $[-1, 1]$. Then $\text{disc}(A) = O(\sqrt{n \log(m/n)})$.*

In this lecture we give a constructive proof (by way of a simple randomized algorithm) due to Lovett and Mehka [1] of the previous theorem. We also do not require any advanced tools from probability (in particular, the correlation inequality for slabs due to Sidak [2]) — indeed, all we will need is the stability of Gaussians and concentration of measure for a type of Gaussian random walk.

2 The Algorithm

Theorem 2. *Let m, n be positive integers with $m \geq n$ and let A be any $m \times n$ matrix with entries from $[-1, 1]$. There is a randomized algorithm running in time polynomial in m, n which, when given A as input, outputs an $x \in \{\pm 1\}^n$ such that $||Ax||_\infty = O(\sqrt{n \log m/n})$ with high probability.*

In specifying and analyzing the algorithm we take the same geometric viewpoint that is used in the proof from the previous lecture. Let m, n be positive integers with $m \geq n$, and let A be any $m \times n$ matrix with entries from the interval $[-1, 1]$. Let C be a universal constant to be specified later, and consider the convex polytope

$$K = \{x \in \mathbf{R}^n \mid ||Ax||_\infty \leq C\sqrt{n \log(8m/n)}\}.$$

If $a_1, a_2, \ldots, a_m$ are the rows of A we can re-write the previous definition as

$$K = \{x \in \mathbf{R}^n \mid \forall i \in [m], |\langle a_i, x \rangle| \leq C\sqrt{n \log(8m/n)}\},$$

which will be more useful. Intuitively, K is a convex polytope defined as the intersection of a set of "slabs" of the form

$$|\langle a_i, x\rangle| \leq C\sqrt{n\log(8m/n)}.$$

The algorithm will work as follows: we perform (an approximation of) a continuous random walk, starting from the origin, in the convex polytope $K \cap [-1,1]^n$. When the random walk intersects a facet of $K \cap [-1,1]^n$ we restrict further steps of the walk to remain on that facet, until we end up at a vertex of $[-1,1]^n$. Clearly the resulting vector satisfies the conclusion of Theorem 2, but why should this algorithm work?

Intuitively, this works because the facets of K are much further away from the origin (measured by Euclidean distance) than the facets of $[-1,1]^n$: to be precise, the facet of K corresponding to a_i is at a distance

$$\frac{C\sqrt{n\log(8m/n)}}{||a_i||_2} \geq C\sqrt{\log(8m/n)}$$

from the origin since $||a_i||_2 \leq \sqrt{n}$, while the facets of $[-1,1]^n$ are each at distance 1 from the origin (note that we have crucially used the fact that all entries of A are bounded by ± 1 here). It follows that in the continuous random walk above it should be much more likely to hit a facet of $[-1,1]^n$ than a facet of K.

Of course the actual algorithm can only perform an approximation of the continuous random walk as a sequence of very small discretized steps. Here we are again helped by the fact that the facets of $[-1,1]^n$ are close to the origin: by choosing the step-length sufficiently small we will be able to show that after some large (but polynomial) number of steps we will have hit a constant fraction of the facets of $[-1,1]^n$.

For convenience, we will reduce Theorem 2 to the following theorem, which can be viewed as an "algorithmic partial colouring lemma".

Theorem 3. *Let $m \geq n$ be positive integers and let $\delta = 1/\sqrt{n}$. There is a randomized, polynomial-time algorithm and a constant C such that, when given an $m \times n$ matrix A with entries from $[-1,1]$ and a vector $c \in [-1,1]^n$, finds an $x \in [-1,1]^n$ such that the following holds, with probability at least $1/6 - \varepsilon$ for any $1/6 > \varepsilon > 0$.*

1. *For each $i = 1, 2, \ldots, m$ we have $|\langle a_i, x\rangle| \leq C||a_i||_2\sqrt{\log(8m/n)}$*
2. *$|x_i - 1| \leq \delta$ for at least $n/10$ indices i.*

Proof of Theorem 2 from Theorem 3. Note that we can amplify the success probability in the usual way by independent repetition — we do not include it in our analysis, but it adds at a polynomial factor to the run time. Start with $c = \mathbf{0}$, run the algorithm from Theorem 3 and obtain an $x \in [-1,1]^n$. Let x' be the vector obtained by choosing all indices from x for which (2) fails, and apply the algorithm recursively on x' and on the matrix A' obtained by deleting the columns corresponding to the indices satisfying (2). At each recursive step we fix a constant fraction of the coordinates of x, and so we end up with a vector x^* for which all indices satisfy (2) after $S = 10\log n$

recursive steps. The discrepancy of the resulting vector is

$$|\langle a_i, x^* \rangle| < C\sqrt{n}\sqrt{\log(8m/n)} + C\sqrt{n/10}\sqrt{\log(8m/(n/10))} + \cdots + C\sqrt{n/10^S}\sqrt{\log(8m/(n/10^S))}$$
$$< \sqrt{n}\sum_{s=0}^{\infty} \frac{C\sqrt{\log 8m \cdot 10^s/n}}{10^{s/2}} < C'\sqrt{n\log(m/n)}$$

for some constant C'.

Now we round the vector x^* to get a vector $\vec{x} \in \{\pm 1\}^n$. For each coordinate of x^* round the value to the nearest ± 1. For any row a_i, the discrepancy of the resulting vector is

$$|\langle a_i, \vec{x} \rangle| \leq |\langle a_i, x^* \rangle| + |\langle a_i, \vec{x} - x^* \rangle| \leq C'\sqrt{n\log(m/n)} + |\langle a_i, \vec{x} - x^* \rangle|.$$

By (2), each coordinate is distance at most $\delta = 1/\sqrt{n}$ from ± 1. Applying Cauchy-Schwarz we get

$$|\langle a_i, \vec{x} - x^* \rangle| \leq ||a_i||_2 ||\vec{x} - x^*||_2 \leq \sqrt{n} \cdot \left(\sum_{i=1}^{n} \delta^2\right)^{1/2} = \sqrt{n}$$

and so the discrepancy of the resulting vector is $O(\sqrt{n\log(m/n)})$. □

Let $\mathcal{N}(\mu, \sigma^2)$ denote the mean μ Gaussian distribution with variance σ^2. The algorithm is formally described in Algorithm 1. As stated, the algorithm includes several scalar parameters δ, γ, T, C that we fix during the analysis: for now, think of δ, γ as being small reals with, say, $1/\sqrt{n} \geq \delta \gg \gamma > 0$ and T being some large integer on the order of $1/\gamma^2$. The set $\mathcal{D}_t$ contains the facets of K for which

Algorithm 1: Main Algorithm

Input : An $m \times n$ matrix A, with entries from $[-1, 1]$. A vector $c \in [-1, 1]^n$.
Output: A vector $x \in [-1, 1]^n$ satisfying the properties in Theorem 3.
Set $x^0 = c$;
Normalize each row vector a_i in A so that $||a_i||_2 = 1$;
for $t = 1, 2, \ldots, T$ **do**
  Set $\mathcal{D}_t = \{i \in [m] \mid |\langle a_i, x^{t-1} - c \rangle| \geq C\sqrt{\log(8m/n)} - \delta\}$;
  Set $\mathcal{V}_t = \{j \in [n] \mid |x_j^{t-1}| > 1 - \delta\}$;
  Let $\mathcal{W}_t = \{y \in \mathbf{R}^n \mid \forall i \in \mathcal{D}_t, \langle a_i, y \rangle = 0 \text{ and } \forall j \in \mathcal{V}_t, y_j = 0\}$;
  Let $w_1, w_2, \ldots, w_k$ be an orthonormal basis of the subspace $\mathcal{W}_t$;
  Let $g_1, g_2, \ldots, g_k \sim \mathcal{N}(0, 1)$ be sampled i.i.d.;
  Set $\Delta x^t = \sum_{i=1}^{k} g_i w_i$;
  Set $x^t = x^{t-1} + \gamma \Delta x^t$;
end
return x^T

the vector x^{t-1} is "almost tight", and the set $\mathcal{V}_t$ contains the set of facets of $[-1, 1]^n$ for which x^{t-1} is "almost tight". Note that the facets in $\mathcal{D}_t$ are $C\sqrt{\log(8m/n)}$ instead of $C\sqrt{n\log(8m/n)}$ — this is because we have normalized the row vectors a_i. The subspace $\mathcal{W}_t$ contains all vectors orthogonal to the $\mathcal{D}_t$ facets and the $\mathcal{V}_t$ facets. In each iteration of the algorithm, we take the vector x^{t-1} and perturb it by Gaussian random noise in the subspace $\mathcal{W}_t$. By moving in the subspace $\mathcal{W}_t$, we never

increase the discrepancy with respect to the facets in $\mathcal{D}_t$ and we never modify any coordinates that are sufficiently close to ± 1.

The parameters γ, δ should be viewed as tolerance parameters that we must introduce since we are approximating a continuous random walk. The parameter δ defines a small region around the facets of $K \cap [-1,1]^n$ which we use to define when a vector x^t is tight with respect to the facet. By choosing δ to be small enough we are guaranteed that the coordinates are close enough to the ± 1 constraints so that we do not introduce too much extra discrepancy when rounding the fractional coordinates. The parameter γ controls the step-size of our discretized walk — we choose a vector Δx^t of variance-1 Gaussian random noise, projected to the subspace $\mathcal{W}_t$, and make a γ-length step in that direction.

For simplicity, in the rest of this section we fix an $m \times n$ matrix A with entries in $[-1,1]$. Much of the hard work in the analysis is proving the lower bound on the number of fixed coordinates — i.e. the size of $|\mathcal{V}_t|$ — of the output x^T for some appropriately chosen C, δ, γ, T. This essentially follows from our Main Lemma (cf. Lemma 5), which states that if we do not let the random walk continue for too long, then the expected number of facets of K for which x^T is tight will be small. By using $||x^T - c||_2$ as a potential function (which intuitively is a measure of the number of tight facets) we will be able to show the upper bound given by Lemma 5 gives a lower bound on $\mathbb{E}\,|\mathcal{V}_T|$. Applying Markov's inequality finishes the proof.

Gaussian noise enjoys a number of useful properties (e.g. exponential tail bounds), but key to the analysis is the next property that states that a linear combination of samples of Gaussian noise is again Gaussian.

Stability of Gaussians. Let $g_1, g_2, \ldots, g_k$ be i.i.d. samples from $\mathcal{N}(0,1)$, and let $g = (g_1, g_2, \ldots, g_k)$. Then $\langle a, g \rangle \sim \mathcal{N}(0, ||a||_2^2)$ for any $a \in \mathbf{R}^k$.

To prove our Main Lemma we will also use following Azuma-type martingale concentration inequality.

Lemma 4. *Let $\sigma \in \mathbf{R}$ satisfy $0 < \sigma < 1$. Suppose $y_1, y_2, \ldots, y_\ell$ are random variables where $y_1 \sim \mathcal{N}(0, \sigma^2)$, and for all $i > 1$, the conditional distribution of $y_i - y_{i-1}$ given the values of $y_1, y_2, \ldots, y_{i-1}$ is $\mathcal{N}(0, \sigma_i^2)$ for some random variable $0 < \sigma_i < 1$ depending on $y_1, y_2, \ldots, y_{i-1}$. Then*

$$\Pr[|y_\ell| > t\sqrt{\ell}] \leq 2e^{-t^2/2}$$

for any $t > 0$.

Proof. By the definition of the sequence $y_1, y_2, \ldots, y_\ell$, for any $\lambda > 0$

$$\Pr[|y_\ell| > t\sqrt{\ell}] = 2\Pr[e^{\lambda y_\ell} > e^{\lambda t\sqrt{\ell}}]$$

by the symmetry of Gaussians and the monotonicity of e^x. Markov's inequality yields

$$2\Pr[e^{\lambda y_\ell} > e^{\lambda t\sqrt{\ell}}] \leq \frac{2\,\mathbb{E}\,e^{\lambda y_\ell}}{e^{\lambda t\sqrt{\ell}}}.$$

We prove by induction on ℓ that $\mathbb{E}\,e^{\lambda y_\ell} \leq e^{\lambda^2 \ell/2}$. Assuming this, and choosing $\lambda = t/\sqrt{\ell}$, the lemma

follows from a standard calculation:

$$\Pr[|y_\ell| > t\sqrt{\ell}] = 2\Pr[e^{\lambda y_\ell} > e^{\lambda t\sqrt{\ell}}] \le \frac{2\,\mathbb{E}\,e^{\lambda y_\ell}}{e^{\lambda t\sqrt{\ell}}} \le \frac{2e^{\lambda^2\ell/2}}{e^{\lambda t\sqrt{\ell}}} = \frac{2e^{t^2/2}}{e^{t^2}} = 2e^{-t^2/2}.$$

On to the induction. If $\ell = 1$ then $y_1 \sim \mathcal{N}(0,\sigma^2)$ by assumption, so

$$\mathbb{E}\,e^{\lambda y_1} = e^{\lambda^2\sigma^2/2} \le e^{\lambda^2/2}$$

since $\sigma < 1$. So, suppose that the result holds for all $y_1,\ldots,y_{\ell-1}$, and we prove it for y_ℓ. Then

$$\begin{aligned}
\mathbb{E}\,e^{\lambda y_\ell} &= \mathbb{E}_{y_1,y_2,\ldots,y_{\ell-1}}[\mathbb{E}[e^{\lambda y_\ell}|y_1,\ldots,y_{\ell-1}]] \\
&= \mathbb{E}_{y_1,y_2,\ldots,y_{\ell-1}}[e^{\lambda y_{\ell-1}}\,\mathbb{E}[e^{\lambda(y_\ell - y_{\ell-1})}|y_1,\ldots,y_{\ell-1}]] \\
&\le \mathbb{E}_{y_1,\ldots,y_{\ell-1}}[e^{\lambda y_{\ell_1}}e^{\lambda^2/2}] \\
&\le e^{\lambda^2(\ell-1)/2}\cdot e^{\lambda^2/2} = e^{\lambda^2\ell/2},
\end{aligned}$$

where the penultimate inequality follows from the assumption that $y_i - y_{i-1}$ is conditionally a 0-mean Gaussian with variance $\sigma_i^2 < 1$, and the final inequality is the inductive hypothesis. □

The main lemma now follows from the martingale concentration bound and the Stability of Gaussians. It says that if the random walk is not "too long" then, on average, the number of tight discrepancy constraints will be small.

Lemma 5. *If $T = O(1/\gamma^2)$ then there exists a constant C such that*

$$\mathbb{E}\,|\mathcal{D}_{T+1}| = \mathbb{E}\,|\{i \in [m] \mid |\langle a_i, x^T - c\rangle| \ge C\sqrt{n\log(8m/n)}\}| \le n/4.$$

Proof. Let C be a constant that will be fixed later. By linearity of expectation we can write

$$\mathbb{E}\,|\mathcal{D}_{T+1}| = \sum_{i=1}^{m}\Pr[|\langle a_i, x^T - c\rangle| \ge C\sqrt{n\log(8m/n)}].$$

We expand the inner product as

$$\langle a_i, x^T - c\rangle = \sum_{t=1}^{T}\gamma\langle a_i, \Delta x^t\rangle$$

where $\Delta x^t = \sum_{j=1}^{k} g_j w_j$ for i.i.d. Gaussian samples $g_1, g_2, \ldots, g_k \sim \mathcal{N}(0,1)$. The Stability of Gaussians implies that

$$\langle a_i, \Delta x^t\rangle = \sum_{j=1}^{k} g_j\langle a_i, w_j\rangle \sim \mathcal{N}(0,\sigma^2)$$

where $\sigma^2 = \sum_{j=1}^{k}\langle a_i, w_j\rangle^2 \le ||a_i||_2^2 \le 1$ since the a_i vectors are normalized in the algorithm and the basis $w_1,\ldots,w_k$ is an orthonormal basis of a subspace. For each $t = 1,2,\ldots,T$ let $y_t = \sum_{j=1}^{t}\langle a_i, \Delta x^j\rangle$. It follows that the sequence of variables $y_1, y_2, \ldots, y_T$ satisfy the conditions of Lemma 4, thus

$$\Pr[|y_T| > t\sqrt{T}] = \Pr[\gamma|y_T| > t\gamma\sqrt{T}] = \Pr[|\gamma y_T| > t\gamma\sqrt{T}] = \Pr[|\langle a_i, x^T - c\rangle| > t\gamma\sqrt{T}] \le 2e^{-t^2/2}.$$

Since $T = O(1/\gamma^2)$, choosing $t = C\sqrt{\log 8m/n}$ and C any constant such that $C \geq 1/\gamma\sqrt{T}$ yields

$$\Pr[|\langle a_i, x^T - c\rangle| > C\sqrt{\log 8m/n}] \leq 2e^{-\log 8m/n} = \frac{n}{4m}.$$

By summing this inequality over all $i \in [m]$ we get $\mathbb{E}\,|\mathcal{D}_{T+1}| \leq n/4$. □

With this lemma we can prove Theorem 3.

Proof of Theorem 3. Let $x = x^T$ be the output of Algorithm 1. We prove $x^t \in K \cap [-1,1]^n$ for all t with high probability, from which the first property of Theorem 3 follows. Let E_t denote the event that x^t is the first point in the algorithm for which $x^t \notin K \cap [-1,1]^n$ fails, so we have

$$\Pr[x^1, x^2, \ldots, x^T \in K \cap [-1,1]^n] = 1 - \sum_{t=1}^{T} \Pr[E_t].$$

To estimate $\Pr[E_t]$, in iteration t of the algorithm we must violate either a constraint in K or a constraint in $[-1,1]^n$. In either case, by the definition of the algorithm the step length $\gamma||\Delta x^t||_2$ must have been greater than δ. We prove that this happens with very low probability if $\gamma \ll \delta$.

If the event E_t occurs then we have $|\langle w, \gamma\Delta x^t\rangle| > \delta$ for some $w \in \{a_1, \ldots, a_m, e_1, \ldots, e_n\}$, where $e_1, e_2, \ldots, e_n$ is the standard basis (corresponding to the possible variable constraints in $\mathcal{V}_t$). By the Stability of Gaussians we know that $\langle w, \Delta x^t\rangle \sim \mathcal{N}(0, ||w||_2^2)$ and we know $||w||_2 = 1$ since all constraints are normalized. A standard Gaussian tail bound implies that

$$\Pr[|\langle w, \Delta x^t\rangle| \geq \delta/\gamma] \leq 2e^{-(\delta/\gamma)^2/2},$$

and so

$$\Pr[\exists t : x^t \notin K \cap [-1,1]^n] = \sum_{t=1}^{T} \Pr[E_t] \leq \sum_{t=1}^{T} \sum_{w} \Pr[|\langle w, \Delta x^t\rangle| \geq \delta/\gamma] \leq 2nmTe^{-(\delta/\gamma)^2/2}.$$

Choosing $\gamma \leq \delta/\sqrt{D\log(mn/\gamma)}$ for any large constant D and using the fact that $T = O(1/\gamma^2)$ yields an upper bound on the above probability of $1/(mn)^{D-2}$. Thus for all t, $x^t \in K \cap [-1,1]^n$ with probability at least $1 - 1/(mn)^{D-2}$.

We now prove the second property in Theorem 3 by estimating $\mathbb{E}\,|\mathcal{V}_T|$ and using Markov's inequality. By the definition of the algorithm, $\mathbb{E}\,||x^T - c||_2^2 \leq n$ and $x^T - c = \sum_{t=1}^{T} \gamma\Delta x^t$. For any $t = 1, 2, \ldots, T$, if $w_1, w_2, \ldots, w_k$ is the orthonormal basis of $\mathcal{W}_t$ we have

$$\mathbb{E}\,||\Delta x^t||_2^2 = \mathbb{E}\langle \sum_{j=1}^{k} g_j w_j, \sum_{j=1}^{k} g_j w_j\rangle = \sum_{j=1}^{k} \mathbb{E}\, g_j^2 = k = \dim \mathcal{W}_t \geq n - \mathbb{E}\,|\mathcal{D}_t| - \mathbb{E}\,|\mathcal{V}_t|,$$

where $\mathbb{E}\, g_j^2 = 1$ for any j since the Gaussian samples have variance 1. Since the Gaussian samples used by the algorithm are independent and mean 0, and since each of the bases of the subspaces $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_T$ are orthonormal, we have

$$\begin{aligned} n \geq \mathbb{E}\,||x^T - c||_2^2 &= \mathbb{E}\langle \sum_{t=1}^{T} \gamma\Delta x^t, \sum_{t=1}^{T} \gamma\Delta x^t\rangle \\ &= \gamma^2 \sum_{t=1}^{T} \langle \Delta x^t, \Delta x^t\rangle = \gamma^2 \sum_{t=1}^{T} \mathbb{E}\,||x^t||_2^2. \end{aligned}$$

For each $t = 1, 2, \ldots, T$ we have $|\mathcal{D}_t| \leq |\mathcal{D}_{t+1}|$ and $|\mathcal{V}_t| \leq |\mathcal{V}_{t+1}|$, since whenever the random walk is tight to a facet in $\mathcal{D}_t$ or $\mathcal{V}_t$ it remains tight to that facet for all further steps. Using this fact we continue the calculation:

$$n \geq \gamma^2 \sum_{t=1}^{T} \mathbb{E}\,||x^t||_2^2 \geq \gamma^2 \sum_{t=1}^{T} n - \mathbb{E}\,|\mathcal{D}_t| - \mathbb{E}\,|\mathcal{V}_t| \geq \gamma^2(T)(n - \mathbb{E}\,|\mathcal{D}_T| - \mathbb{E}\,|\mathcal{V}_T|).$$

Choose $T = 2/\gamma^2$ and using the fact that $|\mathcal{D}_T| \leq |\mathcal{D}_{T+1}|$ rearrange to get

$$\mathbb{E}\,|\mathcal{V}_T| \geq \frac{1}{2}(n - 2\,\mathbb{E}\,|\mathcal{D}_T|) \geq n/4.$$

Applying Markov's inequality to the random variable $n - |\mathcal{V}_T|$ we get

$$\Pr[n - |\mathcal{V}_T| > 9n/10] = \Pr[n/10 > |\mathcal{V}_T|] \leq \frac{n - \mathbb{E}\,|\mathcal{V}_T|}{9n/10} \leq \frac{3n/4}{9n/10} = \frac{5}{6}$$

and thus $|\mathcal{V}_T| \geq n/10$ with probability at least $1/6$.

We now verify the algorithm runs in polynomial time. This is easy to see — each iteration runs in polynomial time, and the number of iterations is $T = 2/\gamma^2$ where γ is chosen so that

$$\gamma \leq \delta/\sqrt{D\log(mn/\gamma)} = 1/\sqrt{Dn\log(mn/\gamma)},$$

so choosing γ to be, say, $o(1/n)$ will satisfy this property while leaving the running time polynomial.

Now, on a single run of the algorithm, Property (1) of Theorem 3 holds with probability at least $1 - 1/mn^{D-2}$ (where we can choose D to be any large constant, which for a large γ could increase the run-time of the algorithm by a polynomial factor), and Property (2) holds with probability at least $1/6$. It follows that both properties hold with some non-zero probability whenever $1/6 - 1/mn^{D-2} > 0$, and so both properties hold with probability at least $1/6 - \varepsilon$ for any $\varepsilon > 0$ by choosing D sufficiently large relative to m, n. □

References

[1] Shachar Lovett and Raghu Meka. *Constructive discrepancy minimization by walking on the edges.* In the proceedings of FOCS 2012.

[2] Zbyněk Šidák. *Rectangular confidence regions for the means of multivariate normal distributions.* J. Amer. Statist. Assoc. 62: 626-633 (1967).

[3] Joel Spencer. *Six standard deviations suffice.* Transactions of the American Mathematical Society 289(2):679-706 (1985).