THE MIXING RATE OF MARKOV CHAINS, AN ISOPERIMETRIC INEQUALITY, AND COMPUTING THE VOLUME

László Lovász
Eötvös Loránd University, Budapest
and
Princeton University, Princeton, NJ

Miklós Simonovits
Mathematical Research Institute
Hungarian Academy of Science, Budapest
and
Rutgers University, New Brunswick, NJ

Abstract. Sinclair and Jerrum derived a bound on the mixing rate of time-reversible Markov chains in terms of their conductance. We generalize this result, not assuming time-reversibility and using a weaker notion of conductance. We prove an isoperimetric inequality for subsets of a convex body. These results are combined to simplify an algorithm of Dyer, Frieze and Kannan for approximating the volume of a convex body, and to improve running time bounds.

0. Introduction and preliminaries

Recently Dyer, Frieze and Kannan (1989) designed a polynomial time randomized algorithm to approximate the volume of a convex body K in $\mathbb{R}^n$. A crucial step of the algorithm is to generate a random point in a convex body. This is achieved by making a random walk on the lattice points inside the body. The analysis of the algorithm depends on two factors: a theorem of Sinclair and Jerrum (1988) on the mixing rate of time-reversible Markov chains and on an isoperimetric inequality for subsets of a convex body.

In this paper we improve both of these steps. We generalize the theorem of Sinclair and Jerrum (1988). In fact, we introduce a new proof technique to handle the mixing rate of a Markov chain. This will allow us to handle Markov chains in which the "small" sets need not have large conductance. As a byproduct, we can drop the time-reversibility assumption (this will not help us in the current application, however), and obtain a sharper bound for the mixing rate depending on the starting distribution.

Dyer, Frieze and Kannan point out that an improvement in their isoperimetric inequality, in particular the removal of the step of approximating the body by one with bounded curvature, would result in simpler and faster algorithms. We give a fairly simple proof of such an improved isoperimetric inequality.

Then we sketch how these results can be applied to modify the algorithm of Dyer, Frieze and Kannan. In particular, we improve the running time: to approximate the volume of a convex body in $\mathbb{R}^n$ with relative error less than ε and probability of error less than δ, their algorithm has to solve $O\big(n^{23}(\log n)^5\varepsilon^{-2}\log(1/\varepsilon)\log(1/\delta)\big)$ convex programs; our one needs only $O\big(n^{16}(\log n)^6\log(n/\varepsilon)\log(n/\delta)\varepsilon^{-4}\big)$ simple membership tests. While this is still far from being practical, there is hope of further improvements.

Acknowledgements. We are glad to acknowledge discussions on the topic of this paper with Ravi Kannan, Mike Steele, and Doma Szász. We are particularly greatful to Imre Bárány for pointing out an error in an earlier version of the manuscript.

Preliminaries. A *convex body* is a compact and full-dimensional convex set in $\mathbb{R}^n$. For two convex bodies K_1 and K_2, we consider their Minkowski sum:

$$K_1 + K_2 = \{x_1 + x_2 : \ x_i \in K_i\},$$

and also the (less standard) notation

$$K_1 - K_2 = \{x \in \mathbb{R}^n : \ K_2 + x \subseteq K_1\}.$$

For algorithmic purposes, a standard way to describe a convex body K as an input is a *well-guaranteed weak separation oracle*, which means the following:

Definition (Weak separation oracle). For any $y \in \mathbb{Q}^n$ we may ask the oracle whether y belongs to K or not; together with this query, we also include an error tolerance $\delta > 0$. The answer will be "YES" or "NO". The "YES" answer means that the distance of y from K is at most δ; the "NO" answer means that the distance of y from $\mathbb{R}^n \setminus K$ is less than δ. In this case, we also require a "proof" of this fact, in the form a

hyperplane $c^T x \le \gamma$ through y which almost separates y from K in the sense that

$$\max\{c^T x : x \in K\} \le \gamma + \delta|c|.$$

(If y is near the boundary of K, then either answer is legal.)

In addition, we assume that we know the radius r of some ball contained in K (but not necessarily the center of the ball) and the radius R of another ball with center 0 containing K. (The number of bits needed to describe both of these balls is part of the input size.)

These assumptions about the way convex bodies are given are equivalent (up to polynomial time reductions) with several other natural descriptions; see Grötschel, Lovász and Schrijver (1988). Often we have better descriptions, e.g. membership in K can be tested without error. This is the case e.g. when K is given as the solution set of a system of linear inequalities (with rational coefficients).

For each convex body K, there exists a unique ellipsoid E with minimum volume containing it, called the Löwner–John ellipsoid of the body (see Grötschel, Lovász and Schrijver 1988). An important property of the Löwner–John ellipsoid is the following:

Theorem 0.1. *If we shrink the Löwner-John ellipsoid of a convex body K from its center by a factor of n, we obtain an ellipsoid that is contained in K.*

The Löwner–John ellipsoid itself may be difficult to compute. However, an ellipsoid E with the following somewhat weaker properties can be computed in polynomial time, using a version of the shallow cut ellipsoid method due to Yudin and Nemirovskii (1976); cf. Grötschel, Lovász and Schrijver (1988). We call an ellipsoid E a *weak Löwner–John ellipsoid* for K, if E contains K and if we shrink E from its center by a factor of $n^{3/2}$, we obtain an ellipsoid that is contained in K.

Theorem 0.2. *For every convex body, a weak Löwner–John ellipsoid can be computed using $O(n^4 \log(R/r))$ operations with numbers with $O(n^2(|\log R| + |\log r|))$ digits.*

We remark that if K is given in a more explicit manner (e.g., as the solution set of a system of linear inequalities or the convex hull of a set of vertices) then the factor $n^{3/2}$ can be improved to $2n$.

1. Rapid mixing of Markov chains and random walks on graphs.

Let $M = (v_t : t = 0, 1, \ldots)$ be a finite Markov chain with state set V ($|V| \ge 2$), and transition probabilities p_{ij} ($i, j \in V$). Let $N = |V|$. We shall assume that the Markov chain is irreducible (i.e., every state can be reached from every other) and $p_{ii} \ge 1/2$ (both assumptions are only technical). As well known, the distribution of v_t tends to a probability distribution $q \in \mathbb{R}^V$ over V as $t \to \infty$ (q is the left eigenvector of P belonging to the eigenvalue 1). By the irreducibility, $q(i) > 0$ for every i. The Sinclair–Jerrum theorem estimates the rate of this convergence for time-reversible Markov chains (Markov chains satisfying $p_{ij}q(j) = p_{ji}q(i)$) in terms of the "conductance" of the Markov chain. To strengthen their result, we define a version of "conductance" that disregards small sets. Also, we do not need time-reversibility.

Definition. (Conductance) Let $0 \le \mu \le 1/2$. The *μ-conductance* of the Markov chain is the largest number Φ_μ such that for every set $S \subseteq V$,

$$\sum_{\substack{i \in S \\ j \in V \setminus S}} q(i)p_{ij} \ge \Phi_\mu \min\{\, q(S) - \mu,\ q(V \setminus S) - \mu \,\}.$$

(The left hand side is the probability that if a state is drawn from the final distribution, it belongs to S and the next state will belong to $V \setminus S$.) Note that not only the right hand side is symmetric in S and $V \setminus S$, but so is the right hand side: a simple computation (using that q is a left eigenvector) shows that $\sum_{\substack{i \in S \\ j \in V \setminus S}} q(i)p_{ij} = \sum_{\substack{i \in V \setminus S \\ j \in S}} q(i)p_{ij}$.

The 0-conductance is simply called the *conductance*.

Obviously, the larger μ the larger is the μ-conductance. From the assumption that $p_{ii} \ge 1/2$ it easily follows that the μ-conductance is at most $1/2$, and if $|V| \ge 4$ then it is strictly smaller.

Originally, Sinclair and Jerrum used 0-conductance, which measures a kind of "connectivity" of the Markov chain. The conductance is 0 iff there are two states such that there is no way to get from the first to the second. In some cases, there is a small set of states isolated (or almost isolated) from the rest, and μ-conductance is introduced to measure connectivity more "globally".

Let $p_t(i)$ denote the probability of the event that $v_t = i$ ($i \in V$). Then $p_t(i) \to q(i)$ ($t \to \infty$), and for the difference we have the following bound, which for time-reversible chains (and in a slightly weaker form) was found by Sinclair and Jerrum (1988). Set $q_0 = \min\{q_i : i \in V\}$.

Theorem 1.1. *For an arbitrary starting distribution,*

$$|p_t(i) - q(i)| \le \left(1 - \frac{1}{2}\Phi_0^2\right)^t \sqrt{\frac{q(i)}{q_0}}.$$

If we only know the μ-conductance for some $\mu > 0$, then the Markov chain may have some states which are almost inaccessible and hence we cannot assert anything about the pointwise convergence. But more important for our purposes is the convergence rate in the l_1 metric. To analyse this, we consider for every $0 \le x \le 1$ the "error" $p_t(S) - q(S)$ as a function of $q(S)$. For technical reasons, we introduce the following, slightly different quantity:

$$h_t(x) = \max\left\{\sum_{i\in V}\big(p_t(i) - q(i)\big)w_i\right\}$$

where the maximum is taken over all weights $0 \le w_i \le 1$ with $\sum_i w_i q(i) = x$. Note that $h_t(x)$ is an upper bound on the error we are interested in: if $q(S) = x$ then $p_t(S) - q(S)$ occurs in the maximum when w is the incidence vector of S.

The following alternative definition of $h_t(x)$ gives a certain converse to this observation. Let us order the elements $V = \{v_1, \dots, v_N\}$ so that

$$\frac{p_t(v_1)}{q(v_1)} \ge \frac{p_t(v_2)}{q(v_2)} \ge \dots \ge \frac{p_t(v_N)}{q(v_N)},$$

and let $\lambda_k = \sum_{i=1}^{k} q(v_i)$. Find the index k such that $\lambda_{k-1} \le x < \lambda_k$. Then

$$h_t(x) = \sum_{i=1}^{k-1}\big(p_t(v_i) - q(v_i)\big) + \frac{x - \lambda_{k-1}}{q(v_k)}\big(p_t(v_k) - q(v_k)\big).$$

The proof of the equivalence of the two definitions is left to the reader.

It follows that $h_t(x)$ is a concave piecewise linear function on the interval $[0,1]$. Trivially $0 \le h_t(x) \le 1$ for all t and x and $h_t(0) = h_t(1) = 0$. Note that $\max_x h_t(x)$ is attained at one of the λ_i, and it is exactly half of the l_1-distance of p_t and q.

To motivate the generalization of Theorem 1.1 to the case when only information on Φ_μ (but not on Φ_0) is available, note that we cannot expect to be able to say anything about $h_t(x)$ if $x \le \mu$ or $x \ge 1 - \mu$ (the Markov chain may have a set of states of measure μ isolated from the rest, having too high or too low probability), and hence we shall restrict our attention to the interval $\mu \le x \le 1 - \mu$. Let ℓ_μ be the linear interpolation function

$$\ell_\mu(x) = \frac{1 - \mu - x}{1 - 2\mu}h_0(\mu) + \frac{\mu - x}{1 - 2\mu}h_0(1 - \mu)$$

and C, the constant

$$\max\left\{\frac{h_0(x) - \ell_\mu(x)}{\sqrt{x - \mu}}, \frac{h_0(x) - \ell_\mu(x)}{\sqrt{1 - x - \mu}} : \mu < x < 1 - \mu\right\}$$

(it is easy to see that the maximum exists). Then for the initial distribution and for every $\mu \le x \le 1 - \mu$, we have the inequality

$$h_0(x) \le \ell_\mu(x) + C \cdot \min\{\sqrt{x - \mu}, \sqrt{1 - x - \mu}\}.$$

The following is our main theorem on the mixing rate of Markov chains:

Theorem 1.2. *For every $\mu \le x \le 1 - \mu$ and $t \ge 0$,*

$$h_t(x) \le \ell_\mu(x) + C \cdot \min\{\sqrt{x - \mu}, \sqrt{1 - x - \mu}\}\left(1 - \frac{1}{2}\Phi_\mu^2\right)^t.$$

To derive Theorem 1.1, observe that if $\mu = 0$ then the worst initial distribution is concentrated on a single node, and we have $\ell \equiv 0$ and $C \le 1/\sqrt{q_0}$. In general, we want to choose μ appropriately. The best choice of μ depends on a trade-off between the first and second terms. To make $\ell_\mu(x)$ small, we have to choose a small μ. We can make the second term small by choosing a sufficiently large t; but to get a fast convergence, we want Φ_μ large, and for this we want to choose a large μ.

The following is a somewhat simpler corollary of the theorem, which is valid for all subsets:

Corollary 1.3. *Let $H(\mu) = \max\{h_0(\mu), h_0(1 - \mu)\}$. Then for every $S \subseteq V$ and every $t \ge 0$,*

$$|p_t(S) - q(S)| \le H(\mu) + \left(1 - \frac{1}{2}\Phi_\mu^2\right)^t \frac{1}{\sqrt{q_0}}.$$

The proof of Theorem 1.2 is by induction on t, using the following lemma, which strengthens the fact that $h_t(x) \le h_{t-1}(x)$ for all x.

Lemma 1.4. *Let $t \ge 1$. If $\mu \le x \le 1/2$, then*

$$h_t(x) \le \frac{1}{2}\Big(h_{t-1}\big(x - 2(x - \mu)\Phi_\mu\big) + h_{t-1}\big(x + 2(x - \mu)\Phi_\mu\big)\Big).$$

If $1/2 \le x \le 1 - \mu$, then

$$h_t(x) \le \frac{1}{2}\Big(h_{t-1}\big(x - 2(1 - x - \mu)\Phi_\mu\big) + h_{t-1}\big(x + 2(1 - x - \mu)\Phi_\mu\big)\Big).$$

Proof. We prove the first inequality; the second is analogous. Let the ordering $V = \{v_1, \dots, v_N\}$ and the numbers λ_i be defined as after the definition of $h_t(x)$. By the concavity of h_t, it suffices to prove the inequality for $x = \lambda_k$. Then

$$h_t(x) = \sum_{j=1}^{k}\big(p_t(v_j) - q(v_j)\big).$$

Let $u_i = \sum_{j \le k} p_{ij}$ (where p_{ij} is an abbreviation of $p_{v_i v_j}$). Clearly

$$1 \ge u_i \ge p_{ii} \ge 1/2 \qquad (i \le k)$$

and

$$0 \le u_i \le 1 - p_{ii} \le 1/2 \qquad (i > k).$$

Recall that q is a left eigenvector of the matrix (p_{ij}), i.e., $\sum_{i=1}^{N} q(v_i) p_{ij} = q(v_j)$; furthermore,

$$\sum_{i=1}^{N} p_{t-1}(v_i) p_{ij} = p_t(v_j).$$

Hence $\sum_{i=1}^{N} (p_{t-1}(v_i) - q(v_i)) p_{ij} = p_t(v_j) - q(v_j)$ and we have

$$\begin{aligned} h_t(x) &= \sum_{j=1}^{k} (p_t(v_j) - q(v_j)) \\ &= \sum_{j=1}^{k} \sum_{i=1}^{N} p_{ij} (p_{t-1}(v_i) - q(v_i)) \\ &= \sum_{i=1}^{N} (p_{t-1}(v_i) - q(v_i)) u_i. \end{aligned} \tag{1}$$

Moreover, $0 \le u_i \le 1$ and

$$\begin{aligned} \sum_{i=1}^{N} q(v_i) u_i &= \sum_{i=1}^{N} \sum_{j \le k} q(v_i) p_{ij} \\ &= \sum_{j \le k} \sum_{i=1}^{N} q(v_i) p_{ij} = \sum_{j \le k} q(v_j) = x. \end{aligned} \tag{2}$$

Since, by definition, $h_{t-1}(x)$ is the maximum of

$$\sum_{i=1}^{N} (p_{t-1}(v_i) - q(v_i)) y_i$$

subject to $0 \le y_i \le 1$ and $\sum_{i=1}^{N} q(v_i) y_i = x$, this shows that $h_t(x) \le h_{t-1}(x)$. To get the stronger inequality in the lemma, consider the following numbers:

$$u_i' = \begin{cases} 2u_i - 1, & \text{if } i \le k, \\ 0, & \text{if } i > k, \end{cases}$$

and

$$u_i'' = \begin{cases} 1, & \text{if } i \le k, \\ 2u_i, & \text{if } i > k. \end{cases}$$

Then $0 \le u_i' \le 1$, $0 \le u_i'' \le 1$, and $u_i' + u_i'' = 2u_i$. Set $x' = \sum_{i=1}^{N} q(v_i) u_i'$ and $x'' = \sum_{i=1}^{N} q(v_i) u_i''$, then by (2), $x' + x'' = 2x$ and so by (1),

$$\begin{aligned} h_t(x) &= \sum_{i=1}^{N} (p_{t-1}(v_i) - q(v_i)) u_i = \\ &= \frac{1}{2} \sum_{i=1}^{N} (p_{t-1}(v_i) - q(v_i)) u_i' \\ &\quad + \frac{1}{2} \sum_{i=1}^{N} (p_{t-1}(v_i) - q(v_i)) u_i'' \\ &\le \frac{1}{2} h_{t-1}(x') + \frac{1}{2} h_{t-1}(x''). \end{aligned}$$

It remains to estimate $x - x' = x'' - x$:

$$\begin{aligned} x - x' &= \sum_{i=1}^{N} q(v_i)(u_i - u_i') \\ &= \sum_{i \le k} q(v_i)(1 - u_i) + \sum_{i > k} q(v_i) u_i \\ &= \sum_{i \le k} q(v_i) \left(1 - \sum_{j \le k} p_{ij} \right) + \sum_{i > k} q(v_i) \sum_{j \le k} p_{ij} \\ &= \sum_{i \le k} \sum_{j > k} q(v_i) p_{ij} + \sum_{i > k} \sum_{j \le k} q(v_i) p_{ij} \ge 2(x - \mu) \Phi_\mu \end{aligned}$$

by the definition of Φ_μ. So $x' \le x - 2(x - \mu)\Phi_\mu$ and similarly, $x'' \ge x + 2(x - \mu)\Phi_\mu$. The lemma follows by the concavity of h_{t-1}. ∎

Let $G = (V, E)$ be a graph on N vertices. The *random walk on* G is defined as follows. Let d be at least twice the maximum degree of G. Start at a random node (drawn from some distribution). At each step, if we are at a node with degree r, then we move to each neighbor with probability $1/d$ and stay with probability $(d - r)/d$. Let v_t be the random node we are at after t steps. Clearly, $(v_t : t = 0, 1, \ldots)$ is a symmetric Markov chain. It follows from the basic theory of Markov chains that if G is connected, then the distribution of v_t tends to a uniform distribution over V as $t \to \infty$. (Here we need that d is larger than the minimum degree of G; otherwise, this would not hold for d-regular bipartite graphs. The assumption that d is at least twice the maximum degree is technical.)

Let m be a natural number. The *m-conductance* of the graph G is defined as the μ-conductance of the random walk on G, where $\mu = m/|V|$. This can be cast in graph-theoretical terms as follows. For $S \subseteq V$, we denote by $f(S)$ the total number of edges joining S to its complement $V \setminus S$. The m-conductance is the

minimum of $f(S)/(d(|S|-m))$ over all non-empty sets $S \subseteq V$ with $m < |S| < N/2$.

The conductance measures if there is a bottleneck in the graph: whether we can delete a small set of edges and disconnect the graph into large parts. It is strongly connected to the second largest eigenvalue of the graph (see Alon and Milman (1985), Alon (1986), Dodziuk and Kendall (1986)). Graphs with very good conductance properties are the so-called expander graphs. These are those graphs having conductance larger than a constant $c > 1$; we shall need only that the conductance be at least $p(n)^{-1}$, for some polynomial p.

In terms of random walks on graphs, we can formulate the following consequence of theorem 1.2:

Corollary 1.5. *Let $(v_t : t = 0, 1, \ldots)$ be a random walk on a graph $G = (V, E)$ with N nodes. Assume that for every set $H \subseteq V$ with at most m elements we have*

$$\left|\text{Prob}(v_0 \in H) - \frac{|H|}{N}\right| \leq c.$$

Let Φ_m denote the m-conductance of G. Then for every $S \subseteq V$ we have

$$\left|\text{Prob}(v_t \in S) - \frac{|S|}{N}\right| \leq c + \left(1 - \frac{1}{2}\Phi_m^2\right)^t \sqrt{N}.$$

We shall apply this corollary in a situation when the starting distribution is reasonably spread out so that choosing μ sufficiently small, c will be small.

2. An isoperimetric inequality

Dyer, Frieze and Kannan conjectured that an inequality of the following type must hold.

Theorem 2.1. *Let K be a convex set in $\mathbf{R}^n$ with diameter d. Assume that a surface with $(n-1)$-dimensional measure f splits K into two sets K^* and K^{**}. Then*

$$\min\{\text{vol}(K^*), \text{vol}(K^{**})\} < fd.$$

We in fact prove a slightly stronger result that is easier to state since it avoids the difficulties associated with the notion of $(n-1)$-dimensional measure.

Theorem 2.2. *Let K be a convex set in $\mathbf{R}^n$ with diameter d. Let $K' \cup B \cup K''$ be a decomposition of K into three closed parts such that the distance of K' and K'' is at least t. Then*

$$\min\{\text{vol}(K'), \text{vol}(K'')\} \leq \frac{d}{t}\text{vol}(B).$$

To get theorem 2.1, let B be the closed $(t/2)$-neighborhood of $K^* \cap K^{**}$. We use that

$$\text{vol}(B) \leq ft + o(t).$$

Remove the interior of B from K^* and K^{**}, apply theorem 2.2, and let $t \to 0$: theorem 2.1 follows.

Proof of Theorem 2.2. (sketch). Let E be the Löwner–John ellipsoid of K. First, we remark that if K is "needle–like" in the sense that – for an $\varepsilon > 0$ which later will tend to 0 – all but at most one of the axes of E are shorter than εt, then it is easy to obtain a slightly weaker inequality. Consider the projection of K on the longest axis of E, and let a_1 and a_2 be the endpoints of this projection. Let $r = |a_2 - a_1|$, $e = (a_2 - a_1)/r$, and let H_s be the hyperplane through $a_1 + se$ orthogonal to this axis. Let $f(s)$ be the $(n-1)$-dimensional volume of $H_s \cap K$. By the Brun–Minkowski Theorem (see e.g. Bonnesen and Fenchel (1934)), this function is unimodal in the interval $0 \leq s \leq r$, i.e., it is increasing for some (a_1, a_3) and decreasing for (a_3, a_2).

Assume that $a_1 \in K'$ and $a_2 \in K''$. Assume that (moving from a_1 toward a_2) u_1 is the last s for which $H_s \cap K' \neq \emptyset$ and u_2 is the first one for which $H_s \cap K'' \neq \emptyset$. Then $u_2 - u_1 \geq (1 - 2\varepsilon)t$. By symmetry we may assume that $f(u_1) \leq f(u_2)$. Hence, by the unimodularity, $f(s) \leq f(u_1)$ for $s \in [0, u_1]$. Therefore

$$\text{vol}(K') \leq f(u_1) \cdot r.$$

Similarly,

$$\text{vol}(B) \geq f(u_1) \cdot (1 - 2\varepsilon)t,$$

implying

$$\text{vol}(K') \leq \frac{r}{(1-2\varepsilon)t}\text{vol}(B).$$

This proves the "needle-like" case (with the ε error).

We prove the general case by way of contradiction. By the so-called "Ham-Sandwich" Theorem, there exists a hyperplane that is orthogonal to the plane spanned by the two largest axes of E and cuts both K' and K'' into two parts with equal volume. (If $n = 2$, then the orthogonality condition is vacuous.) If K violates the theorem, then at least one half of it (with K', K'' and B restricted accordingly) also violates it.

What remains to be seen is that repeating this procedure, the second largest axes of the Löwner–John ellipsoids of the bodies tend to 0. Suppose not, then we can find a series of bodies $K^1 \supset K^2 \supset \ldots$ such that K^{m+1} arises from K^m by a series of bisectings as above and the second largest axis of their Löwner–John ellipsoids remains larger than ϵ. We may assume that the planes of the two largest axes of these ellipsoids converge to a plane Π.

Let K_1^m be the projection of K^m on Π. Then by the choice of Π and by theorem 0.1, K_1^m contains a disk with radius ϵ/n. Below we shall derive a contradiction by showing that the area of K_1^m tends to 0.

First of all, the assertion of theorem 2.2 is obvious if $\mathrm{vol}(B) \geq \frac{1}{3}\mathrm{vol}(K)$. So assume that $\mathrm{vol}(B) < \frac{1}{3}\mathrm{vol}(K)$ and the same holds for the corresponding parts after each bisection. One can relatively easily show that if a hyperplane H orthogonal to Π (i.e., orthogonal to some line in Π) cuts K^m into two parts with volume $\geq \frac{1}{3}\mathrm{vol}(K)$ each, then it cuts K_1^m into two parts with area at least $(1/9n^2)$th of the area of K_1^m (e.g. because it follows that the width of K^m orthogonal to H on both sides of H is at least $(1/3n)$th of its total width; then the same holds for K_1^m, which in turn implies that at least $(1/9n^2)$th of the total area of K_1^m is on each side of H).

Since for large enough m, the hyperplane used to cut $K' \cap K^m$ and $K'' \cap K^m$ into equal pieces is almost orthogonal to H, it follows that we have at least $(1/10n^2)$th of the area of K_1^m on both sides of H. Hence the area of K_1^{m+1} is at most $1 - 1/10n^2$ times the area of K_1^m. Hence this area tends to 0 with m, which is a contradiction. ∎

3. The conductance of lattice graphs

We can use the isoperimetric inequality in the previous section to estimate the conductance of lattice graphs.

Recall that a *lattice point* means a point in $\mathbb{R}^n$ with integral coefficients. Let *cube* mean a unit cube whose center is a lattice point and whose edges are parallel to the axes.

Let K be a convex body in $\mathbb{R}^n$. We are interested in lattice points in K; however, we have only a weak separation oracle, and we may not be able to tell exactly which lattice points belong to K. Therefore we shall carry out our arguments for a set V of lattice points such that the centers of cubes contained in K belong to V along with some further centers of cubes intersecting K. The set of lattice points declared as members of K by a weak separation oracle, called with error tolerance $1/2$, will certainly have this property. Define a graph G on V by connecting two nodes (lattice points) if and only if they have distance 1. Let us call G a *weak lattice graph* associated with K. It has (typically) exponentially many nodes (about $\mathrm{vol}(K)$), but it has maximum degree $2n$. G need not be connected; but "typically" it has a giant component H containing all lattice points sufficiently deep inside K.

To be able to apply the Sinclair–Jerrum theorem, we need an estimate on the conductance of G. Unfortunately, we cannot claim that the 0-conductance of this graph is good, but there is a sufficiently small m for which the m-conductance is sufficiently large.

Theorem 3.1. *Assume that the convex body $K \subseteq \mathbb{R}^n$ contains a ball with radius r and is contained in a ball with radius R. Let $G = (V, E)$ be a weak lattice graph associated with K and $m \geq 4n^{3/2}\mathrm{vol}(K)/r$. Then the m-conductance of G is at least $1/(4Rn)$.*

Proof. Let S be any subset of V with $m \leq |S| \leq |V|/2$, and, as before, let $f(S)$ denote the number of edges in the lattice graph G connecting S to $V \setminus S$. Let $U(S)$ and $U(V \setminus S)$ denote the union of cubes with center in S and $V \setminus S$, respectively. Unfortunately, the sets $U(S)$ and $U(V \setminus S)$ do not cover K; but they do cover the convex body $K - Q$, to which we shall be able to apply our isoperimetric inequality. So consider the sets $H_1 = (K-Q) \cap U(S)$ and $H_2 = (K-Q) \cap U(V \setminus S)$. Let σ be the $(n-1)$-dimensional measure of the surface $H_1 \cap H_2$. By our theorem 2.1,

$$\min\{\mathrm{vol}(H_1), \mathrm{vol}(H_2)\} \leq 2R\sigma \leq 2Rf(S).$$

We expect $\mathrm{vol}(H_1)$ to be close to $|S|$ and $\mathrm{vol}(H_2)$ to be close to $|V \setminus S|$. This is not true in general, since $U(S)$ may consist of cubes meeting $K - Q$ only in a very tiny set. However, a simple geometric argument omitted here shows that the number of such cubes is at most $4n^{3/2}\mathrm{vol}(K)/r \leq m$, and hence $\mathrm{vol}(H_1) \geq |S| - m$. Similarly, $\mathrm{vol}(H_2) \geq |V \setminus S| - m \geq |S| - m$. Hence

$$f(S) \geq \frac{1}{2R}\min\{\mathrm{vol}(H_1), \mathrm{vol}(H_2)\} \geq \frac{1}{2R}(|S| - m).$$

∎

4. The volume algorithm

We use basically the same algorithm as Dyer, Frieze and Kannan (1989). However, the results in the previous sections and other observations will enable us to simplify the arguments and improve running times. In particular, we do not have to make the body "smooth".

So we describe a randomised algorithm that, given a convex body K in $\mathbb{R}^n$, (by a weak separation oracle, together with two positive numbers r and R such that K is guaranteed to contain a ball with radius r and is contained in the ball with radius R about the origin), a $\delta > 0$ and an $\varepsilon > 0$, computes a number ζ such that with probability at least $1 - \delta$,

$$(1 - \varepsilon)\mathrm{vol}(K) \leq \zeta \leq (1 + \varepsilon)\mathrm{vol}(K).$$

We may assume that ε, $\delta < 1/10$. The algorithm will be polynomial in n, $1/\varepsilon$, $\log(1/\delta)$, $|\log R|$ and $|\log r|$.

(a) Set

$$k = \lceil 2n \log n \rceil,$$
$$\psi = 2^7 k^2 n^{3/2} \varepsilon^{-1},$$
$$\Psi = 2^7 k^2 n^3 \varepsilon^{-1},$$
$$t = \lceil 2^5 n^3 \Psi^2 \log(n/\varepsilon) \rceil,$$
$$\tau = \lceil 2^6 n^2 (\log n)^2 \log(n/\delta) \varepsilon^{-2} \rceil.$$

We denote by B the unit ball and by Q, the cube with unit edges parallel to the axes, with center at the origin.

(b) To avoid the bodies of extremely elongated forms, apply a linear transformation to K, which makes it **well rounded**: we may assume that we have

$$\psi B \subseteq K \subseteq \Psi B.$$

This is achieved by constructing a weak Löwner-John ellipsoid (Theorem 0.2), and then applying a linear transformation that maps this onto the ball ΨB.

(c) For $i = 0, 1, \ldots, k$, define the convex body K_i as the intersection of K with the ball $\left(1 + (1/n)\right)^i \psi B$. For $i = 0$, we get the ball ψB, while for $i = k$ we get K. This way we succeded in joining the ball to the body K with unknown volume by an increasing sequence of convex bodies "fairly smoothly":

$$1 \leq \frac{\mathrm{vol}(K_i)}{\mathrm{vol}(K_{i-1})} \leq \left(1 + \frac{1}{n}\right)^n < e.$$

(d) Consider the lattice of all integral points. Also consider a cube with edge-length 1 and edges parallel to the axes about each integral point. When we say "cube", we mean one of these. Let V denote the set of all integral points for which the weak separation oracle, called with error bound $1/2$, concludes that they are contained in K (this is essentially $\mathbb{Z}^n \cap K$, except possibly for some lattice points near the boundary of K). For $i = 0, 1, ..., k$, let $V_i = V \cap \left(1 + (1/n)\right)^i \psi B$.

(e) Generate a random point u in ψB from a uniform distribution. (This can be done e.g. as follows. Let $\xi_1, \ldots, \xi_n$ be independent random variables from a standardized normal distribution, and let η be uniformly distributed in $[0, 1]$. Let $\mu_i = \eta^{1/n} \xi_i / \sqrt{\xi_1^2 + \cdots + \xi_n^2}$, then $v = (\mu_1, \ldots, \mu_n)$ is uniformly distributed over E'.

Round u to the nearest integral point to get v_0. If v_0 happens to be outside ψB, return 0 as the volume of K (the probability of this to happen is less than $\delta/(4\tau)$).

(f) We start from v_0 and take a random walk on the integral points as follows. At the jth step, where $(i-1)t \leq j < it$, we are at an integral point v_j in V_i. Toss a coin and if it is head, stay at v_j (this is only a technical detail, needed to be able to apply theorem 1.2). Else, select one of the $2n$ coordinate directions at random, and consider the next integral point in that direction. If this integral point belongs to V_i, then move to it; else, stay at v_j. After it steps, we are allowed to step on any integral point in V_{i+1}. The point $w_i = v_{it}$ will be almost uniformly distributed over V_i (see below); besides providing a random point in V_i, w_i also provides a good starting point for the next phase, when we walk in V_{i+1}, etc. The procedure terminates after kt steps, i.e., when w_k is obtained.

(g) Repeat (e–f) τ times, called *runs*. In each run, check for $1 \leq i \leq k$ whether or not w_i belongs to V_{i-1}. Let b_i denote the number of runs in which we have $w_i \in V_{i-1}$. The random value $\alpha_i = \tau/b_i$ will estimate $|V_i|/|V_{i-1}|$ (if $b_i = 0$, then return 0 as the volume of K; the probability of this to happen will be much less than $\delta/(4k)$).

(h) Estimate $\mathrm{vol}(K) = \mathrm{vol}(K_k)$ by

$$\varsigma = \alpha_1 \cdots \alpha_k \mathrm{vol}(\psi B) = \frac{\alpha_1 \cdots \alpha_k \psi^n \pi^{n/2}}{\Gamma(1 + n/2)}.$$

We owe some explanation here: why do we use this "concatenation" of Markov processes, instead of ordinary Markov processes? We could set out for each i from a fixed point and follow the procedure of Dyer, Frieze and Kannan; however, starting from the wrong point we would get trapped near the border with too large probability. Most likely, this does not happen when starting from a point deep inside the body, but technically it is easier to handle the case when we start from a random point in ψB and generate a nearly uniform distribution on V_i recursively.

5. Analysis of the algorithm

To estimate the error of our algorithm we shall use

$$\frac{\varsigma}{\mathrm{vol}(K)} = \frac{\alpha_1 \cdots \alpha_k \mathrm{vol}(\psi B)}{|V_k|} \cdot \frac{|V_k|}{\mathrm{vol}(K)}$$
$$= \prod_{i=1}^{k} \frac{\alpha_j}{|V_i|/|V_{i-1}|} \cdot \frac{\mathrm{vol}(\psi B)}{|V_0|} \cdot \frac{|V_k|}{\mathrm{vol}(V)}.$$

Therefore the error of our estimate ς comes from three sources:

(E1) $\mathrm{vol}(K)$ is only roughly equal to $|V_k|$, and $\mathrm{vol}(\psi B)$ is only approximately $|V_0|$.

(E2) The integral point w_i generated in step (f) is not really uniformly distributed over V_i.

(E3) We apply a Monte-Carlo calculation in (g); this has a standard deviation.

Of these, (E1) is easy to treat: it follows by elemantary geometric considerations that

$$\exp(-\varepsilon/8)|V_i| \le \mathrm{vol}(K_i) \le \exp(\varepsilon/8)|V_i|.$$

Estimating error (E2) is more difficult, partly because the whole process $(v_j : j = 0, 1, \ldots)$ is not a Markov chain; but in the interval $(i-1)t \le j < it$ it is one, and we can apply our previous results.

Let p_j be the probability distribution of v_j, and let $q_i = p_{it}(V_{i-1})$. Let $m_i = \lceil 4n^{3/2}\mathrm{vol}(K_i)/\psi \rceil$ and let H_i denote the maximum of $\left|p_{(i-1)t}(T) - |T|/|V_i|\right|$ over all sets $T \subseteq V_i$ with $|T| \le m_i$.

By Theorem 3.2, the m_i-conductance of the lattice graph on V_i is at least $1/(4n\Psi)$, and so by Corollary 1.5 we obtain that for every $S \subseteq V_i$,

$$\left|p_{it}(S) - \frac{|S|}{|V_i|}\right| < H_i + \left(1 - \frac{1}{2^5 n^2 \Psi^2}\right)^t \sqrt{|V_i|}. \quad (3)$$

Here the second term is less than $\varepsilon/(16k^2)$ by the choice of t. Further, using the crude estimate that V_k is contained in a cube of size 2Ψ, we get that

$$\left(1 - \frac{1}{2^5 n^2 \Psi^2}\right)^t \sqrt{|V_i|} \le e^{-\frac{1}{2}n(8\log(n/\varepsilon) - 6\log(n/\varepsilon))} = \left(\frac{\varepsilon}{n}\right)^n \le \frac{\varepsilon}{16k^2},$$

if $n \ge 3$ (for $n = 2$ the argument should be slightly modified). The first term in (3) can be estimated by induction on i. An easy argument shows that $|V_i| < 4|V_{i-1}|$ holds for all $1 \le i \le k$. Using this, we have

$$H_1 = \left|\frac{m_1}{|V_0|} - \frac{m_1}{|V_1|}\right| < 3\frac{m_1}{|V_1|} < 4\frac{m_1}{\mathrm{vol}(K_1)} < \frac{4n^{3/2}}{\psi} = \frac{\varepsilon}{32k^2},$$

since v_0 is uniformly distributed over V_0. Now (3) gives that (with the set $T \subseteq V_i$, $|T| \le m_i$ attaining the maximum in the definition of H_i)

$$\begin{aligned} H_i &= \left|p_{i(t-1)}(T) - \frac{|T|}{|V_i|}\right| \\ &\le \left|p_{i(t-1)}(T) - \frac{|T|}{|V_{i-1}|}\right| + 3\frac{|T|}{|V_i|} \\ &\le H_{i-1} + \left(1 - \frac{1}{2^5 n^2 \Psi^2}\right)^t \sqrt{|V_{i-1}|} + 4\frac{m_i}{\mathrm{vol}(K_i)} \\ &< H_{i-1} + \frac{8n^{3/2}}{\psi} = H_{i-1} + \frac{\varepsilon}{16k^2}. \end{aligned}$$

Hence by induction,

$$H_i < \frac{i\varepsilon}{16k^2} \le \frac{\varepsilon}{16k}.$$

Applying (3) again, we get that

$$\left|q_i - \frac{|V_{i-1}|}{|V_i|}\right| < \frac{\varepsilon}{16k} + \frac{\varepsilon}{16k^2} < \frac{\varepsilon}{8k}.$$

Finally, we consider error (E3). For each fixed i, we independently repeat the experiment of generating a random integral point w_i τ times; By the law of large numbers (Chernoff's inequality), we derive that with probability at least $1 - \delta/(2k)$,

$$\exp\left(-\frac{\varepsilon}{2k}\right)\alpha_i \le \frac{|V_{i-1}|}{|V_i|} \le \exp\left(\frac{\varepsilon}{2k}\right)\alpha_i.$$

With probability at least $1 - \delta/2$, this holds simultaneously for all i.

To sum up, we have with probability at least $1 - \delta$ that the value of ζ is not 0 and for every $1 \le i \le k$,

$$\exp\left(-\frac{\varepsilon}{2k}\right)\alpha_i \le \frac{|V_{i-1}|}{|V_i|} \le \exp\left(\frac{\varepsilon}{2k}\right)\alpha_i.$$

If this holds, then

$$\mathrm{vol}(K)(1-\varepsilon) < \zeta < \mathrm{vol}(K)(1+\varepsilon).$$

Hence with probability $1 - \delta$, the number computed by our algorithm does indeed estimate the volume of K with relative error less than ε.

The running time of the algorithm consists of the following.

(i) The time needed to "round" the body in step (b). This is achieved in polynomial time using the "shallow cut" version of the ellipsoid method. This takes $O(n^4 \log(R/r))$ operations with numbers with $O(n^2(|\log R| + |\log r|))$ digits; for details see Grötschel, Lovász and Schrijver (1988). The time needed for this phase is substantial but still negligible compared with the rest, at least as long as R/r is not extremely large.

(ii) The walking time. This takes $kt\tau$ moves, where each move takes the updating of one coordinate one test of membership in K (which we count as one step), and one test in the appropriate ball (which takes one arithmetic operation if we maintain the squared distance of the random walking point from the origin). So this takes $O(kt\tau) = O(n^{16}(\log n)^6 \log(n/\varepsilon)\log(n/\delta)\varepsilon^{-4})$ arithmetic operations with numbers of size $O(\log n + |\log \varepsilon| + |\log \delta|)$.

(iii) The cost of other operations (computing α_i and ζ etc) is negligible compared with (ii).

6. Concluding remarks.

We have formulated a simplest version of the algorithm to show most directly the main ideas (the results in sections 1 and 2). Several methods present themselves to improve the rather inattractive bound on the running time obtained above. Here we mention some of these without going into the details; unfortunately, we have not been able to reduce the running time to anything remotely practical.

1. In phase (a), instead of constructing a weak Löwner–John ellipsoid, it would suffice to construct an ellipsoid E and an ellipsoid E', obtained from E by shrinking from its center by a factor of ϕ, with the following properties: E "almost" contains K in the sense that

$$\mathrm{vol}(K \setminus E) \leq \varepsilon_1 \mathrm{vol}(K)$$

(where ε is a sufficiently small number), and E' is "almost" contained in K in the sense that

$$\mathrm{vol}(E' \setminus K) \leq \varepsilon_1 \mathrm{vol}(E').$$

Using H. W. Lenstra's "dual" version of the ellipsoid method (Lenstra 1983), with a randomized testing of the above conditions, one can construct such a pair with $\phi = 2n$. This saves a factor of n in the running time. It is conceivable that one can construct such a pair of ellipsoids with a substantially smaller ϕ.

A more promising idea is to make larger steps during the random walk inside the body. We mention two such possibilities.

2. One may generate a vector from an appropriate distribution, and try to add it to the current point. If we remain in the body, we move to the new point; else, we stay at the old point. If the distribution of the step is centrally symmetric and sufficiently compact, then the resulting Markov chain has uniform limiting distribution. One expects that the mixing rate is better than for the random walk. We do not know, however, what is a good choice for the distribution and how to estimate the conductivity of the resulting Markov chain. The isoperimetric inequality in section 2 applies specifically to the "continuous" random walk.

3. Let us modify the random walk on lattice points as follows. If we are at a lattice point v then we select a coordinate direction, i.e., a vector $e \in \{e_1, -e_1, \ldots, e_n, -e_n\}$ at random, and find the largest integral s such that $v + 2se \in K$. Then we move to $v + se$. It is not difficult to see that this Markov chain has uniform limit distribution (note, however, that it is not symmetric!). Again, we conjecture that this Markov chain has a better mixing rate.

Frieze proposed a similar procedure for generating a random point in a convex body: he chooses a random coordinate direction and then a random lattice point on the corresponding chord. Again, this Markov chain appears to have a better mixing rate but its analysis seems hard.

References

N. Alon (1986): Eigenvalues and expanders, *Combinatorica* **6**, pp. 83–96.

N. Alon and V. D. Milman (1984): Eigenvalues, expanders, and superconcentrators, *Proc. 25th Ann. Symp. on Found. of Comp. Sci.*, pp. 320–322.

N. Alon and V. D. Milman (1985): λ_1, isoperimetric inequalities for graphs and superconcentrators, *J. Comb. Theory B* **38** pp. 73–88.

I. Bárány and Z. Füredi (1986): Computing the volume is difficult, *Proc. of the 18th Annual ACM Symposium on Theory of Computing*, pp. 442-447.

T. Bonnesen and W. Fenchel (1934): *Theorie der konvexen Körper*, Springer, Berlin.

J. Dodziuk and W. S. Kendall (1986): Combinatorial Laplacians and isoperimetric inequality, in: *From Local Times to Global Geometry, Control and Physics*, (ed. K. D. Ellworthy), Pitman Res. Notes in Math. Series **150**, 68–74.

M. Dyer, A. Frieze and R. Kannan (1989): A Random Polynomial Time Algorithm for Approximating the Volume of Convex Bodies, *Proc. of the 21st Annual ACM Symposium on Theory of Computing*, pp. 375-381.

G. Elekes (1986): A geometric inequality and the complexity of computing volume, *Discrete and Computational Geometry* **1**, pp. 289-292.

M. Grötschel, L. Lovász and A. Schrijver (1988): *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag.

H. W. Lenstra, Jr. (1983), Integer programming with a fixed number of variables, *Oper. Res.* **8**, 538-548.

A. Sinclair and M. Jerrum (1988): Conductance and the rapid mixing property for Markov chains: the approximation of the permanent resolved, *Proc. 20th ACM STOC*, pp. 235–244.

D. B. Yudin and A. S. Nemirovskii (1976), Informational complexity and efficient methods for the solution of convex extremal problems (in Russian), *Ekonomika i Mat. Metody* **12**, pp. 357–369; English translation: *Matekon* **13**, pp. 25–45.