An Asymptotic Isoperimetric Inequality

Noga Alon [*] Ravi Boppana [†] Joel Spencer [‡]

Abstract

For a finite metric space V with a metric ρ, let V^n be the metric space in which the distance between $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ is the sum $\sum_{i=1}^n \rho(a_i, b_i)$. We obtain an asymptotic formula for the logarithm of the maximum possible number of points in V^n of distance at least d from a set of half the points of V^n, when n tends to infinity and d satisfies $d \gg \sqrt{n}$.

1 The Main Results

Let V be a finite metric space with metric ρ and with probability measure μ. On the set V^n define naturally the product probability measure

$$\mu_n(a_1, \ldots, a_n) = \prod_{i=1}^n \mu(a_i)$$

and the L^1 metric

$$\rho_n((a_1, \ldots, a_n), (b_1, \ldots, b_n)) = \sum_{i=1}^n \rho(a_i, b_i).$$

For any $S \subseteq V^n$ and $d \geq 0$ define the closed ball

$$B[S, d] = \{u \in V^n : \exists\ v \in S, \quad \rho_n(u, v) \leq d\}.$$

We are interested in minimizing $\mu_n(B[S, d])$ over all sets S with $\mu_n(S) \geq \frac{1}{2}$. In our range of interest this quantity will be nearly 1 so we instead define

$$f_n(d) = \max_{\mu_n(S) \geq \frac{1}{2}} \mu_n(\overline{B[S, d]}\,).$$

[*] Department of Mathematics, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel, and Institute for Advanced Study, Princeton, NJ 08540. Email: noga@math.tau.ac.il. Research supported in part by a USA–Israel BSF grant and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

[†] Department of Computer Science, Courant Institute, New York University, New York, NY 10012. Email: boppana@cs.nyu.edu.

[‡] Departments of Mathematics and Computer Science, Courant Institute, New York University, New York, NY 10012. Email: spencer@cs.nyu.edu.

In this paper we obtain tight bounds on the asymptotic behavior of $f_n(d)$ for every fixed V, ρ, and μ, when n and $d/\sqrt{n}$ tend to infinity. In this range, our estimates provide an asymptotic formula for $\ln f_n(d)$, with the right constant. It is convenient to deal separately with the case $d = o(n)$, which is somewhat simpler, and the case $d = \Omega(n)$. The results are described in the following two subsections.

1.1 Sublinear Distances

Our bounds on $f_n(d)$ for $n \gg d \gg \sqrt{n}$ depend upon a constant c (dependent on V, ρ, μ but not on n) that we call the *spread constant* of (V, ρ, μ). Call $X\colon V \to \mathbf{R}$ *Lipschitz* if

$$|X(v) - X(w)| \leq \rho(v, w)$$

for all $v, w \in V$. Any such function can be considered a random variable over the probability space (V, μ) and as such has a variance $\mathrm{Var}[X]$.

Definition: The spread constant c is the maximum possible $\mathrm{Var}[X]$ over all Lipschitz X.

The spread constant appears to be new and may well be of independent interest. Alekseev [2] and Engel [13, 14] used a somewhat similar constant to estimate the width of a product of partial orders.

We call X *optimal* if it is Lipschitz with maximum possible variance. Since any continuous function attains its maximum in any compact domain, there are optimal X, that is, the supremum of $\mathrm{Var}[X]$ is in fact a maximum. Translating X preserves the Lipschitz condition and the variance so that, when convenient, we may assume the optimal X has mean 0. The following theorem determines the asymptotic behavior of $f_n(d)$ for all d satisfying $n \gg d \gg \sqrt{n}$.

Theorem 1.1 *For $n^{1/2} \ll d \ll n$,*

$$f_n(d) = e^{-\frac{d^2}{2cn}(1+o(1))},$$

with c the spread constant defined above.

1.2 Linear Distances

The asymptotic behavior of $f_n(d)$ for $d = \Theta(n)$ is somewhat more complicated than that for the sublinear case. Define the *maximum average distance* m by

$$m = \max_{\mu(v)>0} \mathrm{E}[\rho(v, w)],$$

where w is a random variable with distribution μ. It is not hard (see Theorem 3.16) to show that $d \geq mn + \sqrt{cn}$ (or even just $d \geq m(n+3)$) implies $f_n(d) = 0$. In other words, for d just a little more than mn, our problem becomes uninteresting. We develop an asymptotic formula for the logarithm of $f_n(d)$ for all d a little less than mn. More precisely, suppose that $d \gg \sqrt{n}$ and $mn - d \gg \sqrt{n}$. For a real number λ, define $L(\lambda)$ to be the maximum of $\ln \mathrm{E}[e^{\lambda X}]$ over all Lipschitz functions $X\colon V \to \mathbf{R}$ with $\mathrm{E}[X] = 0$. For a real number t, define $R(t)$ by

$$R(t) = \sup_{\lambda \in \mathbf{R}} [\lambda t - L(\lambda)]. \tag{1}$$

The following theorem determines the asymptotic behavior of $f_n(d)$ for a wide range of d.

Theorem 1.2 *For $d \gg \sqrt{n}$ and $mn - d \gg \sqrt{n}$, we have*

$$f_n(d) = e^{-R(d/n)n(1+o(1))},$$

where R is defined in Equation (1). In particular, for fixed $t \in (0, m)$ and $d \sim tn$, we have

$$f_n(d) = e^{-R(t)n(1+o(1))}.$$

We note that it is not difficult to show that for all values of n and d,

$$f_n(d) \leq \exp[-\Theta(d^2/n)].$$

See, for example, [20, 5, 22, 21, 11, 27]. The new aspect of the present work is the finding of the correct constant, giving an asymptotic formula *with the right constant* for $\ln f_n(d)$. For a different approach that can yield related results, see [1].

As an example illustrating Theorem 1.2 consider the metric space V consisting of three points with equal probability, where the distance between any two is 1. A simple though tedious computation gives that here $e^{-R(t)} = 2(2-3t)^{t-2/3}(6t+2)^{-1/3-t}$ for all $0 \leq t < 2/3$. This supplies a tight isoperimetric inequality for the space of all vectors of length n over the alphabet $\{1, 2, 3\}$ with the Hamming metric, and can be used, for example, to improve the estimate for ϵ in one of the explicit constructions of [4] of a K_4-free graph on N vertices in which any set of at least $N^{1-\epsilon}$ vertices contains a triangle.

1.3 Large Deviations

Theorems 1.1 and 1.2 are closely related to known results in the theory of Large Deviations, although we do not see any simple way of deriving them from these known results. The statement of Theorem 1.1 resembles the known probabilistic statement that if $Y = X_1 + X_2 + \ldots + X_n$ is the sum of independent, identically distributed, bounded random variables each having mean 0 and variance 1, then for $\sqrt{n} << d << n$,

$$Pr[Y > d] = e^{-d^2/2n(1+o(1))}.$$

(See, e.g., [24], Theorem 5.2.2 or [7], Chapter 6, Theorem 3.1 for a more general result). The main part of our proof of Theorem 1.1 is its reduction to a large deviation inequality for martingales which is very similar to the one above. For the sake of completeness we include a short proof of this inequality as well.

The statement of Theorem 1.2 is closely related to the statement of Cramér's Theorem (see, e.g., [28], pp. 7-10, or [25], p. 35), which asserts that for bounded, identically distributed random variables $X_1, \ldots, X_n$ all having distribution X, their sum Y satisfies

$$Pr(Y > d) = e^{-R(d/n)n(1+o(1))},$$

where

$$R(t) = \sup_{\lambda \in \mathbf{R}} [\lambda t - \ln \mathrm{E}[e^{\lambda X}]].$$

The situation in our case is more complicated, as we are dealing with the maximum over all possible choices of X, and not with a single, given X. Here, too, our presentation is self contained.

It is also worth noting that our proof of Theorem 1.2 contains the assertion of Theorem 1.1, but since the proof of Theorem 1.1 is much simpler we prefer to describe it separately.

2 Sublinear Distances

In this section we consider the case $d = o(n)$, prove Theorem 1.1, and establish several results on the spread constant.

2.1 Graphs

The special case where V is the vertex set of a connected graph G, the metric ρ is the distance metric of G, and μ is uniform on V, is of particular interest and was our original motivation. In this case, the following theorem holds.

Theorem 2.1 *There is an optimal X with $X(v)$ integral for all v. Moreover, there is always an optimal X for which there is a set of vertices U and an assignment of a sign $s(C) \in \{1, -1\}$ to every connected component of $V - U$ so that for every vertex v in U, we have $X(v) = 0$ and for every vertex v in a component C as above, $X(v)$ is the product of $s(C)$ and the distance between v and U.*

Proof. We first prove that there is an integral optimal X. Note that for a metric given by a graph, the assumption that X is Lipschitz is equivalent to assuming that $|X(u) - X(v)| \leq 1$ for every edge $\{u, v\}$. By translating if necessary we find an optimal X with $X(v) = 0$ for some vertex v. Define a graph H on V by letting v, w be adjacent if they are adjacent in G and $X(v) - X(w) = \pm 1$. It suffices to show that H is connected. If H were not connected let $C \subset V$ be a connected component of H and set $X_\epsilon(v) = X(v) + \epsilon$ if $v \in C$; set $X_\epsilon(v) = X(v)$ otherwise. For each subset $S \subseteq V$ let

$X(S)$ denote the sum $\sum_{s\in S} X(s)$. As a function of ϵ,

$$\begin{aligned}\mathrm{Var}[X_\epsilon] &= \mathrm{Var}[X] + \frac{2\epsilon X(C) + \epsilon^2 |C|}{|V|} - \frac{2\epsilon |C| X(V) + \epsilon^2 |C|^2}{|V|^2} \\ &= \mathrm{Var}[X] + \frac{2\epsilon |C|}{|V|}\left(\frac{X(C)}{|C|} - \frac{X(V)}{|V|}\right) + \epsilon^2 \frac{|C|}{|V|}\left(1 - \frac{|C|}{|V|}\right). \qquad (2)\end{aligned}$$

As the coefficient of ϵ^2 is strictly positive, $\epsilon = 0$ cannot be a local maximum. But X_ϵ is Lipschitz for some open interval of ϵ centered at 0, contradicting the optimality of X. This contradiction proves the first part of the theorem.

Next we prove the second part. Note that in the formula (2), the coefficient of the linear term in ϵ is positive iff the average value of X over C exceeds its average over V. Fix an optimal X that attains only integral values, and let s denote the unique integer in the semi-closed interval $[\mathrm{E}[X] - 1/2, \mathrm{E}[X] + 1/2)$. For each integer t, let U_t denote the set of all vertices u for which $X(u) = t$. We claim that for each $t > s$ and each $u \in U_t$, there is a vertex $v \in U_{t-1}$ with uv being an edge. Indeed, otherwise we could modify X to X_ϵ by defining $X_\epsilon(w) = X(w) + \epsilon$ for all w satisfying $X(w) \geq t$, and by leaving $X_\epsilon(w') = X(w')$ for each other vertex w'. Note that if ϵ is a sufficiently small positive real, then X_ϵ is Lipschitz. However, by formula (2) for $\mathrm{Var}[X]$ (where now C is the set of all vertices u satisfying $X(u) \geq t$), if ϵ is positive and sufficiently small, then $\mathrm{Var}[X_\epsilon] > \mathrm{Var}[X]$, contradicting the optimality of X. This contradiction proves the claim.

By repeatedly applying the assertion of the claim we conclude that if $u \in U_t$ with $t > s$, then the distance between u and U_s is at most $t - s$. However, the distance cannot be smaller, as X is Lipschitz, showing that in this case $X(u)$ is precisely s plus the distance between U_s and u. The same argument implies that if $u \in U_t$ and $t < s$ then the distance between u and U_s is precisely $s - t$. Replacing X by $X - s$ (which has the same variance) we obtain an optimal X whose value on the set of vertices $U = U_s$ is 0, such that for every vertex v in $V - U$, the absolute value of $X(v)$ is precisely the distance between v and U. Thus $X(v)$ is either that distance or its negation. It remains to show that there are no two vertices v and v' in a component of $V - U$, with $X(v)$ being the distance between v and U and $X(v')$ being the negation of the distance between v' and U. However, if there are two such vertices then there are two adjacent vertices with this property, contradicting the Lipschitz condition. This contradiction completes the proof. □

Remark. The Laplace matrix of a graph $G = (V, E)$ is the matrix $L = (\ell_{u,v})$ whose rows and columns are indexed by the vertices of G, in which $\ell_{v,v}$ is the degree of v for all $v \in V$ and for each two distinct u, v in V, we have $\ell_{u,v} = -1$ if $uv \in E$ and $\ell_{u,v} = 0$ otherwise. This matrix is symmetric, and thus has real eigenvalues. Its smallest eigenvalue is 0, and its second smallest eigenvalue, denoted by λ_1, is strictly positive iff G is connected. This eigenvalue appears in certain isoperimetric inequalities for G; see [5]. The discussion in [5, p. 76] easily implies that the spread constant c of G satisfies $c \leq |E|/\lambda_1$.

When G is an edge on the vertices $0,1$ an optimal X has $X(0) = 0$, $X(1) = 1$. This is the isoperimetric problem in the Hamming cube, for which the precise result is known for all $|S|$ and all d by the work of Harper [18]; see also [16]. In particular, the set S of $(a_1, \ldots, a_n) \in \{0,1\}^n$ with $\sum_{i=1}^n a_i \leq r$ has the minimal $|B[S,d]|$ among all S of that size. More generally let G be a path on vertices $0, 1, \ldots, k-1$, in that order. It is not difficult to prove that $X(i) = i$ for all i is an optimal X. Here Bollobás and Leader [11] have given the precise isoperimetric result for all $|S|$ and all d. In particular, the set S of $(a_1, \ldots, a_n)$ with $\sum_{i=1}^n X(a_i) \leq r$ has the minimal $|B[S,d]|$ among all S of that size. When G is an even cycle, Bollobás and Leader [10] have again given the precise isoperimetric result.

Our results are more general but less precise. This tradeoff leads to a tantalizing speculation. Given any connected graph G with an optimal X, is it possible that for n sufficiently large and d, r in appropriate ranges the set S of $(a_1, \ldots, a_n)$ with $\sum_{i=1}^n X(a_i) \leq r$ has *the* minimum possible $|B[S,d]|$ among all S of that size?

2.2 The Lower Bound

For convenience, technical results on Large Deviations for both the lower and the upper bounds in Theorem 1.1 have been placed in § 2.4 and § 2.5.

Fix an optimal X with $\mathrm{E}[X] = 0$. Define $W\colon V^n \to \mathbf{R}$ by

$$W(a_1, \ldots, a_n) = \sum_{i=1}^n X(a_i).$$

The function W, viewed as a random variable on the probability space (V^n, μ_n), has the same distribution as $\sum_{i=1}^n X_i$, where the X_i are independent copies of X. Thus W has mean 0 and variance nc. We take either $S = \{a \in V^n : W(a) \leq 0\}$ or $S = \{a \in V^n : W(a) \geq 0\}$, whichever has $\mu_n(S) \geq \frac{1}{2}$. By symmetry assume it's the first case. Since X is Lipschitz, for all $a = (a_1, \ldots, a_n), b = (b_1, \ldots, b_n) \in V^n$, we have

$$|W(a) - W(b)| \leq \sum_{i=1}^n |X(a_i) - X(b_i)| \leq \sum_{i=1}^n \rho(a_i, b_i) = \rho_n(a,b),$$

so that

$$\overline{B[S,d]} \supseteq \{b \in V^n : W(b) > d\},$$

and hence

$$\mu_n(\overline{B[S,d]}\,) \geq \Pr[W > d].$$

The Large Deviation result of § 2.5, Equation (5), gives that this probability is at least $\exp[-\frac{d^2}{2cn}(1+o(1))]$ for $\sqrt{n} \ll d \ll n$, which is the lower bound for Theorem 1.1.

2.3 The Upper Bound

Fix $S \subseteq V^n$ with $\mu_n(S) \geq \frac{1}{2}$ and define a random variable Y on V^n by

$$Y(a) = \rho_n(a, S) = \min_{b \in S} \rho_n(a, b).$$

The random variable Y generates a martingale $Y_0, Y_1, \ldots, Y_n$ exposing one coordinate at a time. Thus $Y_n = Y$, $Y_0 = \mathrm{E}[Y]$, and

$$Y_i(a_1, \ldots, a_n) = \mathrm{E}[Y(a_1, \ldots, a_i, x_{i+1}, \ldots, x_n)],$$

where the x_i are independent, each with distribution μ. Fix i and $a_1, \ldots, a_i$ and consider the distribution of $Z = Y_{i+1} - Y_i$. The function Z depends only on the $i+1$-st coordinate so we may consider $Z\colon V \rightarrow \mathbf{R}$, with $Z(\alpha) = Z(a_1, \ldots, a_i, \alpha, \ldots)$. The martingale property (considering Z as a random variable over the probability space (V, μ)) ensures that $\mathrm{E}[Z] = 0$. *Now for the crucial idea.* Changing the $i+1$-st coordinate from α to β can change Y by at most $\rho(\alpha, \beta)$: the closest point of S from α is at most $\rho(\alpha, \beta)$ further away from β. (Note that we need here that ρ is a metric space.) Therefore

$$|Z(\alpha) - Z(\beta)| \leq \rho(\alpha, \beta).$$

The definition of the spread constant c gives

$$\mathrm{Var}[Z] \leq c.$$

Furthermore $|Z| \leq K$ for all i, $a_1, \ldots, a_i$ where K is the diameter of (V, ρ). The Martingale inequality of § 2.4, Equation (4), now gives

$$\Pr[|Y - \mathrm{E}[Y]| > d] < e^{-\frac{d^2}{2cn}(1+o(1))}$$

for $n^{1/2} \ll d \ll n$. Further, for some large constant K_1,

$$\Pr[|Y - \mathrm{E}[Y]| > K_1 n^{1/2}] < \frac{1}{2}.$$

As $\Pr[Y = 0] = \mu_n(S) \geq \frac{1}{2}$, this inequality implies $\mathrm{E}[Y] \leq K_1 n^{1/2}$. Thus

$$\Pr[Y > d] \leq \Pr[|Y - \mathrm{E}[Y]| \geq d - K_1 n^{1/2}].$$

When $d \gg n^{1/2}$ we can write $d - K_1 n^{1/2} = d(1 + o(1))$ so that

$$\Pr[Y > d] < e^{-\frac{d^2}{2cn}(1+o(1))}$$

as claimed.

2.4 Large Deviation Inequalities via Martingales

The following discussion is similar to the ones in [6, 19, 3]. Let X be a random variable with $\mathrm{E}[X] = 0$, $\mathrm{Var}[X] = 1$, and $|X| \leq K$. For any λ,

$$\mathrm{E}[e^{\lambda X}] = 1 + \frac{\lambda^2}{2} + \sum_{i \geq 3} \frac{\lambda^i}{i!} \mathrm{E}[X^i].$$

Now suppose $\lambda = o(1)$. As $|\mathrm{E}[X^i]| \leq K^i$,

$$\left| \sum_{i \geq 3} \frac{\lambda^i}{i!} \mathrm{E}[X^i] \right| \leq \sum_{i \geq 3} \frac{(K\lambda)^i}{i!} = O(\lambda^3),$$

so that

$$\mathrm{E}[e^{\lambda X}] = 1 + \frac{\lambda^2}{2} + O(\lambda^3) = e^{\frac{\lambda^2}{2}(1+o(1))}. \tag{3}$$

Now let $n^{1/2} \ll d \ll n$ and consider a martingale $\mathrm{E}[Y] = Y_0, Y_1, \ldots, Y_n = Y$ with $|Y_{i+1} - Y_i| \leq K$ and $\mathrm{E}[(Y_{i+1} - Y_i)^2 | Y_i, \ldots, Y_0] \leq 1$. Set, with foresight, $\lambda = \frac{d}{n}$. As $d \ll n$, we have $\lambda = o(1)$. Then

$$\mathrm{E}[e^{\lambda(Y_{i+1} - Y_i)} | Y_i, \ldots, Y_0] \leq e^{\frac{\lambda^2}{2}(1+o(1))},$$

so that

$$\mathrm{E}[e^{\lambda(Y - \mathrm{E}[Y])}] \leq e^{\frac{n\lambda^2}{2}(1+o(1))}$$

and

$$\Pr[Y - \mathrm{E}[Y] > d] < e^{\frac{n\lambda^2}{2}(1+o(1)) - \lambda d} = e^{-\frac{d^2}{2n}(1+o(1))}$$

by our (optimal) choice of λ.

By symmetry the same bound holds for $\Pr[Y - \mathrm{E}[Y] < -d]$. When $\mathrm{Var}[X] = c$ we can apply this result to $Xc^{-1/2}$, resulting in an additional c in the denominator of the exponent:

$$\Pr[Y - \mathrm{E}[Y] > d] \leq e^{-\frac{d^2}{2cn}(1+o(1))}. \tag{4}$$

2.5 Lower Bounds for Large Deviations

Let X have $\mathrm{E}[X] = 0$, $\mathrm{Var}[X] = 1$, and $|X| \leq K$, and set $Y = X_1 + \cdots + X_n$ with the X_i independent copies of X. Let $n^{1/2} \ll d \ll n$. We want to bound $\Pr[Y > d]$ from below. Of course, when $d = Cn^{1/2}$ (for constant C), the Central Limit Theorem says that the limiting probability is the probability that the standard normal distribution is at least C.

For $\lambda = o(1)$, by Equation (3), we have

$$\mathrm{E}[e^{\lambda X_i}] = e^{\frac{\lambda^2}{2}(1+o(1))}.$$

Set $\lambda = \frac{d}{n}$, so that

$$\mathrm{E}[e^{\lambda Y}] = e^{\frac{d^2}{2n}(1+o(1))}.$$

Let $\epsilon > 0$ be arbitrarily small. For $x > d(1+\epsilon)$ we have $e^{\lambda x} \le e^{\lambda(1+\epsilon)x}e^{-\lambda d\epsilon(1+\epsilon)}$, so that

$$\begin{aligned}
\mathrm{E}[e^{\lambda Y}[Y > d(1+\epsilon)]] &\le e^{-\lambda d\epsilon(1+\epsilon)}\mathrm{E}[e^{\lambda(1+\epsilon)Y}] \\
&\le \exp[\frac{d^2}{2n}(1+\epsilon)^2(1+o(1)) - \frac{d^2}{n}\epsilon(1+\epsilon)] \\
&= \exp[\frac{d^2}{2n}(1-\epsilon^2)(1+o(1))].
\end{aligned}$$

(Here $[S]$ denotes the indicator of the Boolean value S.) Similarly for $x < d(1-\epsilon)$ we have $e^{\lambda x} \le e^{\lambda(1-\epsilon)x}e^{\lambda d\epsilon(1-\epsilon)}$, so that

$$\begin{aligned}
\mathrm{E}[e^{\lambda Y}[Y < d(1-\epsilon)]] &\le e^{\lambda d\epsilon(1-\epsilon)}\mathrm{E}[e^{\lambda(1-\epsilon)Y}] \\
&= \exp[\frac{d^2}{2n}(1-\epsilon)^2(1+o(1)) + \lambda d\epsilon(1-\epsilon)] \\
&= \exp[\frac{d^2}{2n}(1-\epsilon^2)(1+o(1))].
\end{aligned}$$

Since $d \gg n^{1/2}$, the contribution to $\mathrm{E}[e^{\lambda Y}]$ is asymptotically all from $d(1-\epsilon) < Y < d(1+\epsilon)$, and so

$$\mathrm{E}[e^{\lambda Y}[d(1-\epsilon) < Y < d(1+\epsilon)]] = e^{\frac{d^2}{2n}(1+o(1))}.$$

In this range $e^{\lambda Y} \le e^{\lambda d(1+\epsilon)}$, so

$$\Pr[d(1-\epsilon) < Y < d(1+\epsilon)] \ge e^{\frac{d^2}{2n}(1+o(1)) - \frac{d^2}{n}(1+\epsilon)}.$$

Replacing d by $d/(1-\epsilon)$ gives

$$\Pr[d < Y] \ge \exp\left[\frac{d^2}{2n(1-\epsilon)^2}(1+o(1)) - \frac{d^2(1+\epsilon)}{n(1-\epsilon)^2}\right].$$

As ϵ is arbitrarily small we absorb it into the $o(1)$ term giving

$$\Pr[d < Y] \ge \exp\left[-\frac{d^2}{2n}(1+o(1))\right]$$

as desired.

A simple linear transformation gives that if $Y = X_1 + \cdots + X_n$ with the X_i independent copies of a random variable X having $\mathrm{E}[X] = 0,\ \mathrm{Var}[X] = c$, and $|X| \le K$, then

$$\Pr[Y > d] \ge e^{-\frac{d^2}{2cn}(1+o(1))} \tag{5}$$

for $n^{1/2} \ll d \ll n$.

3 Linear Distances

In this section we mainly consider the case $d = \Theta(n)$ and prove Theorem 1.2. Along the way we establish several auxiliary results.

3.1 The Log-Moment Function

Let X be a random variable. Define its *log-moment function* $L_X\colon \mathbf{R} \to \mathbf{R}$ by

$$L_X(\lambda) = \ln \mathrm{E}[e^{\lambda X}].$$

The first derivative of L_X is

$$L'_X(\lambda) = \frac{\mathrm{E}[Xe^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]}\,. \tag{6}$$

The second derivative of L_X is

$$L''_X(\lambda) = \frac{\mathrm{E}[X^2e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} - L'_X(\lambda)^2\,.$$

The third derivative of L_X is

$$L'''_X(\lambda) = \frac{\mathrm{E}[X^3e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} - 3\frac{\mathrm{E}[X^2e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]}L'_X(\lambda) + 2L'_X(\lambda)^3\,.$$

In particular, $L_X(0) = 0$, $L'_X(0) = \mathrm{E}[X]$, and $L''_X(0) = \mathrm{Var}[X]$. We can rewrite the formula (6) for the first derivative as

$$\mathrm{E}[(X - L'_X(\lambda))e^{\lambda X}] = 0. \tag{7}$$

By expanding $(X - L'_X(\lambda))^2$, we get another formula for the second derivative:

$$L''_X(\lambda) = \frac{\mathrm{E}[(X - L'_X(\lambda))^2e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]}\,. \tag{8}$$

By expanding $(X - a)^2$, and then minimizing over a, we get yet another formula for the second derivative:

$$L''_X(\lambda) = \min_{a \in \mathbf{R}} \frac{\mathrm{E}[(X - a)^2e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]}\,. \tag{9}$$

By expanding $(X - L'_X(\lambda))^3$, we get another formula for the third derivative:

$$L'''_X(\lambda) = \frac{\mathrm{E}[(X - L'_X(\lambda))^3e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]}\,. \tag{10}$$

By Jensen's inequality (see [17]), if $\mathrm{E}[X] = 0$ then L_X is nonnegative. By Equation (8), the second derivative of L_X is nonnegative, so L_X is convex. By Hölder's inequality (see [17]), the function $X \mapsto L_X(\lambda)$ is convex on $\mathbf{R}^V$. Being convex and finite, the function $X \mapsto L_X(\lambda)$ is automatically continuous (Rockafellar [23, Corollary 10.1.1]).

Let V be a finite set, ρ be a metric on V, and μ be a probability distribution on V. Let $G = (V, \rho, \mu)$. Let Ω be the set of Lipschitz $X\colon V \to R$ with $E[X] = 0$. Define the *log-moment function* $L_G\colon \mathbf{R} \to \mathbf{R}$ as follows: $L_G(\lambda)$ is the maximum of $L_X(\lambda)$ over all $X \in \Omega$. Say that $X \in \Omega$ is λ-*optimal* if $L_X(\lambda) = L_G(\lambda)$.

Observe that the set Ω is a (bounded) convex polytope in $\mathbf{R}^V$. Let Ω_0 denote the *finite* set of extreme points of this polytope. As the function $X \mapsto L_X(\lambda)$ is convex, its supremum over Ω is attained on Ω_0; hence we may give a finite expression

$$L_G(\lambda) = \max_{X \in \Omega_0} L_X(\lambda).$$

In what follows we may restrict λ-optimal X to be from Ω_0. Note further that Ω, and hence Ω_0, depends on (V, ρ) but is essentially independent of the probability distribution μ; only the condition $\mathrm{E}[X] = 0$ creates a dependence on μ. Had we so desired, we could have defined Ω to not depend on μ at all. (Namely, replace the condition $\mathrm{E}[X] = 0$ with $\sum_{v \in V} X(v) = 0$; in the definition of L_G, replace $L_X(\lambda)$ with $L_{X-\mathrm{E}[X]}(\lambda)$.)

Because all L_X vanish at 0, so does L_G. Because all L_X (with $\mathrm{E}[X] = 0$) are nonnegative, so is L_G. Because all L_X are convex, so is L_G. Because X is Lipschitz iff $-X$ is Lipschitz, L_G is even.

Being convex and finite, L_G is continuous and, further, is left-differentiable and right-differentiable (Rockafellar [23, Theorem 23.1]). Because Ω is compact, L_X is convex and differentiable, and $X \mapsto L_X(\lambda)$ is continous, it follows from Dem'yanov and Vasil'ev [12, p. 160] that the right derivative $L_G^{\mathrm{r}}(\lambda)$ is the maximum of $L'_X(\lambda)$ over all λ-optimal functions X. Similarly, the left derivative $L_G^{\ell}(\lambda)$ is the minimum of $L'_X(\lambda)$ over all λ-optimal functions X.

We will need a bound on the first derivative of the log-moment function. First we remind the reader of the definition of maximum average distance. Given a point $v \in V$, define $D_v \colon V \to \mathbf{R}$ by $D_v(w) = \rho(v, w)$. By the triangle inequality, D_v is Lipschitz. Define the *maximum average distance* of G by

$$\mathrm{mad}(G) = \max_{\mu(v)>0} \mathrm{E}[D_v].$$

Say that a point $v \in V$ is *remote* if $\mu(v) > 0$ and $\mathrm{E}[D_v] = \mathrm{mad}(G)$.

Theorem 3.1 *Let $X \colon V \to \mathbf{R}$ be a Lipschitz function with $\mathrm{E}[X] = 0$, and let $\lambda \in \mathbf{R}$. Then*

$$L'_X(\lambda) \le \mathrm{mad}(G)\,.$$

Proof. First we claim that $X \le \mathrm{mad}(G)$ with probability 1. Let $v \in V$ be a point with $\mu(v) > 0$. Because X is Lipschitz, for every point $w \in V$ we have

$$X(v) \le X(w) + D_v(w). \tag{11}$$

Now let w be a random variable with distribution μ. Taking expected values of both sides of Equation (11) shows that

$$X(v) \le \mathrm{E}[X] + \mathrm{E}[D_v] = \mathrm{E}[D_v] \le \mathrm{mad}(G),$$

which was our claim.

Using Equation (6) and the claim, we get

$$\begin{aligned} L'_X(\lambda) &= \frac{\mathrm{E}[Xe^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} \\ &\le \frac{\mathrm{E}[\mathrm{mad}(G)e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} \\ &= \mathrm{mad}(G)\,. \end{aligned}$$

That is the inequality we wanted. □

Next we will prove a bound on the second derivative of the log-moment function. Define the (effective) *diameter* of G by

$$\mathrm{diam}(G) = \max_{\substack{\mu(v)>0 \\ \mu(w)>0}} \rho(v,w).$$

Theorem 3.2 *Let $X\colon V \to \mathbf{R}$ be a Lipschitz function, and let $\lambda \in \mathbf{R}$. Then*

$$L''_X(\lambda) \le \frac{1}{4}\mathrm{diam}(G)^2\,.$$

Proof. Define $m = \min_{\mu(v)>0} X(v)$ and $M = \max_{\mu(v)>0} X(v)$. For every x between m and M, we have

$$(x - \frac{m+M}{2})^2 \le (\frac{M-m}{2})^2\,. \tag{12}$$

Because X is Lipschitz, we have

$$\begin{aligned} M - m &= \max_{\substack{\mu(v)>0 \\ \mu(w)>0}} X(v) - X(w) \\ &\le \max_{\substack{\mu(v)>0 \\ \mu(w)>0}} \rho(v,w) \\ &= \mathrm{diam}(G). \end{aligned} \tag{13}$$

By Equations (9), (12), and (13), we have

$$\begin{aligned} L''_X(\lambda) &\le \frac{\mathrm{E}[(X - \frac{m+M}{2})^2 e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} \\ &\le \frac{\mathrm{E}[(\frac{M-m}{2})^2 e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} \\ &= (\frac{M-m}{2})^2 \\ &\le \frac{\mathrm{diam}(G)^2}{4}\,. \end{aligned}$$

That is the inequality we wanted. □

As a corollary, we get the following bound on the log-moment function itself.

Corollary 3.3 *For every $\lambda \in \mathbf{R}$, we have*

$$L_G(\lambda) \le \frac{\operatorname{diam}(G)^2}{8}\lambda^2 .$$

Proof. Let $X\colon V \to \mathbf{R}$ be a Lipschitz function with mean 0. By Taylor's theorem, there is a real number α such that

$$L_X(\lambda) = L_X(0) + L_X'(0)\lambda + L_X''(\alpha)\frac{\lambda^2}{2} . \tag{14}$$

The first two terms vanish, because $L_X(0) = 0$ and $L_X'(0) = \mathrm{E}[X] = 0$. Simplifying Equation (14) and using Theorem 3.2, we get

$$L_X(\lambda) = L_X''(\alpha)\frac{\lambda^2}{2} \le \frac{\operatorname{diam}(G)^2}{8}\lambda^2 .$$

Maximizing over X finishes the proof. □

Next we will prove a bound on the third derivative of the log-moment function.

Theorem 3.4 *Let $X\colon V \to \mathbf{R}$ be a Lipschitz function, and let $\lambda \in \mathbf{R}$. Then*

$$|L_X'''(\lambda)| \le \frac{1}{4}\operatorname{diam}(G)^3 .$$

Proof. Define $m = \min_{\mu(v)>0} X(v)$ and $M = \max_{\mu(v)>0} X(v)$. We have $m \le X \le M$ with probability 1. It follows from Equation (6) that $m \le L_X'(\lambda) \le M$. Recall from Equation (13) that $M - m \le \operatorname{diam}(G)$. Hence by Equations (10) and (8), and Theorem 3.2, we have

$$\begin{aligned}
|L_X'''(\lambda)| &= \frac{|\mathrm{E}[(X - L_X'(\lambda))^3 e^{\lambda X}]|}{\mathrm{E}[e^{\lambda X}]} \\
&\le \frac{\mathrm{E}[|X - L_X'(\lambda)|^3 e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} \\
&\le \frac{(M-m)\mathrm{E}[(X - L_X'(\lambda))^2 e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} \\
&= (M-m)L_X''(\lambda) \\
&\le \operatorname{diam}(G)L_X''(\lambda) \\
&\le \operatorname{diam}(G)\frac{1}{4}\operatorname{diam}(G)^2 \\
&= \frac{1}{4}\operatorname{diam}(G)^3 .
\end{aligned}$$

That is the inequality we wanted. □

As a corollary, we get the following connection between L_G and the spread constant $c(G)$.

Corollary 3.5

(a) $L_G(\lambda) \le c(G)\lambda^2/2 + \operatorname{diam}(G)^3|\lambda|^3/24$.

(b) $L_G(\lambda) \ge c(G)\lambda^2/2 - \operatorname{diam}(G)^3|\lambda|^3/24$.

(c) $\lim_{\lambda\to 0} L_G(\lambda)/\lambda^2 = c(G)/2$.

Proof.

(a) Let $X: V \to \mathbf{R}$ be a Lipschitz function with mean 0. By Taylor's theorem, there is a real number α such that

$$L_X(\lambda) = L_X(0) + L'_X(0)\lambda + L''_X(0)\frac{\lambda^2}{2} + L'''_X(\alpha)\frac{\lambda^3}{6}.$$

The first two terms vanish, because $L_X(0) = 0$ and $L'_X(0) = \mathrm{E}[X] = 0$. The third coefficient is $L''_X(0) = \mathrm{Var}[X]$. Using Theorem 3.4, we get

$$\begin{aligned} L_X(\lambda) &= \mathrm{Var}[X]\frac{\lambda^2}{2} + L'''_X(\alpha)\frac{\lambda^3}{6} \\ &\leq \mathrm{Var}[X]\frac{\lambda^2}{2} + \frac{\mathrm{diam}(G)^3}{4}\frac{|\lambda|^3}{6} \\ &= \mathrm{Var}[X]\frac{\lambda^2}{2} + \mathrm{diam}(G)^3\frac{|\lambda|^3}{24}. \end{aligned}$$

Maximizing over X gives

$$L_G(\lambda) \leq c(G)\frac{\lambda^2}{2} + \mathrm{diam}(G)^3\frac{|\lambda|^3}{24}.$$

That inequality completes part (a).

(b) Similar to part (a).

(c) Follows from parts (a) and (b). □

Let $G = (V_G, \rho_G, \mu_G)$ and $H = (V_H, \rho_H, \mu_H)$ be two spaces. Define the *product* space $G \times H$ by

$$G \times H = (V_G \times V_H, \rho_G \oplus \rho_H, \mu_G \otimes \mu_H),$$

where $\rho_G \oplus \rho_H$ is the L_1 metric

$$(\rho_G \oplus \rho_H)((a,b),(a',b')) = \rho_G(a,a') + \rho_H(b,b')$$

and $\mu_G \otimes \mu_H$ is the product distribution

$$(\mu_G \otimes \mu_H)(a,b) = \mu_G(a)\mu_H(b).$$

The next theorem shows that the log-moment function "tensorizes".

Theorem 3.6 $L_{G\times H} = L_G + L_H$.

Proof. The lower bound $L_{G\times H} \geq L_G + L_H$ is easy. Let $Y: V_G \to \mathbf{R}$ and $Z: V_H \to \mathbf{R}$ be Lipschitz functions with mean 0. Define $X: V_G \times V_H \to \mathbf{R}$ by

$$X(a,b) = Y(a) + Z(b).$$

It is easy to see that X is also Lipschitz with mean 0. Hence $L_{G\times H} \geq L_X = L_Y + L_Z$. Maximizing over all Y and Z completes the lower bound.

The upper bound $L_{G\times H} \leq L_G + L_H$ is more difficult. Let $X\colon V_G \times V_H \to \mathbf{R}$ be a Lipschitz function of mean 0. Define $Y\colon V_G \to \mathbf{R}$ by

$$Y(a) = \mathrm{E}[X(a,b)],$$

where b is a random variable with distribution μ_H. Because X is Lipschitz with mean 0, so is Y. For each $a \in V_G$, define $Z_a\colon V_H \to \mathbf{R}$ by

$$Z_a(b) = X(a,b) - Y(a). \tag{15}$$

Again it is easy to see that Z_a is Lipschitz with mean 0. Rearranging Equation (15), we get

$$X(a,b) = Y(a) + Z_a(b). \tag{16}$$

Let a and b be independent random variables with distributions μ_G and μ_H, respectively. By Equation (16), we have

$$\begin{aligned}
\mathrm{E}[e^{\lambda X}] &= \mathrm{E}[e^{\lambda X(a,b)}] \\
&= \mathrm{E}[\mathrm{E}[e^{\lambda X(a,b)} \mid a]] \\
&= \mathrm{E}[\mathrm{E}[e^{\lambda(Y(a)+Z_a(b))} \mid a]] \\
&= \mathrm{E}[e^{\lambda Y(a)}\mathrm{E}[e^{\lambda Z_a(b)} \mid a]] \\
&= \mathrm{E}[e^{\lambda Y(a)}e^{L_{Z_a}(\lambda)}] \\
&\leq \mathrm{E}[e^{\lambda Y(a)}e^{L_H(\lambda)}] \\
&= \mathrm{E}[e^{\lambda Y(a)}]e^{L_H(\lambda)} \\
&= e^{L_Y(\lambda)}e^{L_H(\lambda)} \\
&\leq e^{L_G(\lambda)}e^{L_H(\lambda)} \\
&= e^{L_G(\lambda)+L_H(\lambda)}\,.
\end{aligned}$$

Taking logarithms and then maximizing over all X concludes the proof. □

Let $G^n = (V^n, \rho_n, \mu_n)$ be the nth power of the space $G = (V, \rho, \mu)$. As a corollary of the previous theorem, we deduce the following connection between L_{G^n} and L_G.

Corollary 3.7 $L_{G^n} = nL_G$.

Proof. Follows from Theorem 3.6 and induction on n. □

As another corollary, we deduce that the spread constant tensorizes. Bobkov and Houdré [8] proved a similar result.

Corollary 3.8 *$c(G \times H) = c(G) + c(H)$. In particular, $c(G^n) = c(G)n$.*

Proof. Follows from Corollary 3.5, Theorem 3.6, and Corollary 3.7. □

3.2 The Rate Function

Given the space $G = (V, \rho, \mu)$, define the *rate function* $R_G\colon \mathbf{R} \to \mathbf{R} \cup \{\infty\}$ by

$$R_G(t) = \sup_{\lambda \in \mathbf{R}} [\lambda t - L_G(\lambda)].$$

We have borrowed the term "rate function" from the theory of Large Deviations, where a similar concept appears; see Shwartz and Weiss [26] for an introduction. Because L_G is a nonnegative function that vanishes at 0, for $t \geq 0$ we need to take the supremum only over $\lambda \geq 0$. Because L_G is a nonnegative function that vanishes at 0, so is R_G. Because L_G is even, so is R_G. Being the supremum of linear functions, R_G is convex.

Say that λ is *t-extreme* if it attains the supremum in the definition of $R_G(t)$. There might not exist a t-extreme value, because the supremum might not be attained. Even worse, the value of $R_G(t)$ might be ∞. The next theorem will identify the domain on which R_G is finite.

Theorem 3.9 *If $|t| \leq \text{mad}(G)$, then $R_G(t)$ is finite.*

Proof. Let v be a remote point. Let λ be arbitrary. If $\lambda \geq 0$, then define $X = \text{E}[D_v] - D_v$; otherwise, define $X = D_v - \text{E}[D_v]$. Note that X is a Lipschitz function with mean 0. We have

$$\begin{aligned} \lambda t - L_G(\lambda) &\leq \lambda t - \ln \text{E}[e^{\lambda X}] \\ &\leq \lambda t - \ln(\mu(v) e^{|\lambda| \text{mad}(G)}) \\ &= \lambda t - |\lambda| \text{mad}(G) - \ln \mu(v) \\ &\leq -\ln \mu(v). \end{aligned}$$

Taking the supremum over all λ shows that $R_G(t) \leq -\ln \mu(v)$. □

Theorem 3.10 *If $|t| < \text{mad}(G)$, then there is a t-extreme value.*

Proof. Let λ be arbitrary. We have

$$\begin{aligned} \lambda t - L_G(\lambda) &= \lambda[t - \text{mad}(G)] + [\lambda \text{mad}(G) - L_G(\lambda)] \\ &\leq \lambda[t - \text{mad}(G)] + R_G(\text{mad}(G)). \end{aligned}$$

Taking the limit as $\lambda \to \infty$ gives

$$\lim_{\lambda \to \infty} \lambda t - L_G(\lambda) = -\infty.$$

Symmetrically, we have

$$\lim_{\lambda\to-\infty} \lambda t - L_G(\lambda) = -\infty.$$

Thus in finding the supremum of $\lambda t - L_G(\lambda)$, we may restrict λ to a compact interval. Because L_G is continuous, it follows that the supremum is attained; there is a t-extreme value. □

The next theorem provides a general lower bound on the rate function.

Theorem 3.11 *For every $t \in \mathbf{R}$, we have*

$$R_G(t) \geq \frac{2t^2}{\operatorname{diam}(G)^2}\,.$$

Proof. With foresight, choose $\lambda := 4t/\operatorname{diam}(G)^2$. Using Corollary 3.3, we get

$$\begin{aligned} R_G(t) &\geq \lambda t - L_G(\lambda) \\ &\geq \lambda t - \frac{\operatorname{diam}(G)^2}{8}\lambda^2 \\ &= \frac{2t^2}{\operatorname{diam}(G)^2}\,. \end{aligned}$$

That is the inequality we wanted. □

Next, we prove a better estimate on the rate function for $t \ll 1$.

Theorem 3.12 $\lim_{t\to 0} R_G(t)/t^2 = 1/(2c(G))$.

Proof. For each $t \in \mathbf{R}$, define the function $h_t\colon \mathbf{R} \to \mathbf{R}$ by

$$h_t(\lambda) = \lambda t - L_G(\lambda).$$

Using Corollary 3.5, we have

$$\lim_{t\to 0} \frac{h_t(t/c(G))}{t^2} = \frac{1}{2c(G)}\,.$$

Because $R_G(t) = \sup_\lambda h_t(\lambda) \geq h_t(t/c(G))$, we have proved a lower bound on the desired limit.

For the upper bound, we will consider the limit only as t approaches 0 from the right; the left limit is similar. Using Corollary 3.5 again, we have

$$\lim_{t\to 0} \frac{h_t(2t/c(G))}{t^2} = 0\,.$$

In particular, $h_t(t/c(G)) \geq h_t(2t/c(G))$ for all sufficiently small t. Because L_G is convex, h_t is concave. Thus for t small, we may restrict λ in the supremum for $R_G(t)$ to the interval $0 < \lambda \leq 2t/c(G)$.

Using Corollary 3.5 once again, we get

$$\begin{aligned}
\lim_{t\downarrow 0}\frac{R_G(t)}{t^2} &= \lim_{t\downarrow 0}\sup_{0<\lambda\le 2t/c(G)}\frac{\lambda t - L_G(\lambda)}{t^2} \\
&\le \lim_{t\downarrow 0}\sup_{0<\lambda\le 2t/c(G)}\frac{\lambda^2}{4L_G(\lambda)} \\
&= \lim_{\lambda\downarrow 0}\frac{\lambda^2}{4L_G(\lambda)} \\
&= \frac{1}{2c(G)}\,.
\end{aligned}$$

That inequality finishes the upper bound. □

Next, we will argue that R_G has nice "continuity" properties. Suppose that $s < t < \mathrm{mad}(G)$. Because R_G is convex, we have

$$\frac{R_G(t)-R_G(s)}{t-s} \le \frac{R_G(\mathrm{mad}(G))-R_G(s)}{\mathrm{mad}(G)-s}\,.$$

Hence we have

$$\begin{aligned}
R_G(t)-R_G(s) &\le \frac{t-s}{\mathrm{mad}(G)-s}\cdot[R_G(\mathrm{mad}(G))-R_G(s)] \\
&\le \frac{t-s}{\mathrm{mad}(G)-s}\cdot R_G(\mathrm{mad}(G)) \qquad (17)\\
&\le \frac{t-s}{\mathrm{mad}(G)-t}\cdot R_G(\mathrm{mad}(G)).
\end{aligned}$$

3.3 The Second Moment

We will need the following generalization of the usual second-moment method.

Theorem 3.13 *Let X be a random variable, let $\lambda \ge 0$, and let $a > 0$. Then*

$$\Pr[X > L'_X(\lambda) - a] \ge e^{L_X(\lambda)-\lambda[L'_X(\lambda)+L''_X(\lambda)/a]}\cdot\frac{a^2}{L''_X(\lambda)+a^2}\,.$$

Proof. Define the function $g\colon \mathbf{R}\to\mathbf{R}$ by

$$g(x) = \frac{e^{-\lambda x}}{x - L'_X(\lambda) + a}\,.$$

It is easy to check that g is

- negative on $(-\infty, L'_X(\lambda)-a)$ and positive on $(L'_X(\lambda)-a,\infty)$,
- decreasing on $(L'_X(\lambda)-a,\infty)$, and

- convex on $(L'_X(\lambda) - a, \infty)$.

With foresight, choose $t = L'_X(\lambda) + L''_X(\lambda)/a$. (Note that $t > L'_X(\lambda) - a$.) We claim that for every $x \in \mathbf{R}$,

$$[x > L'_X(\lambda) - a] \geq \frac{g(t)}{g(x)} + \frac{g'(t)(x-t)}{g(x)}\,. \tag{18}$$

The proof of the claim divides into three cases.

Case $x > L'_X(\lambda) - a$: Because g is convex on $(L'_X(\lambda) - a, \infty)$, we have

$$g(x) \geq g(t) + g'(t)(x-t).$$

Dividing by $g(x)$ finishes this case.

Case $x = L'_X(\lambda) - a$: This case holds with equality, provided that we interpret $1/g(L'_X(\lambda) - a)$ as 0.

Case $x < L'_X(\lambda) - a$: We have $g(t) > 0,\ g'(t) < 0$, and $g(x) < 0$. Hence both terms on the right side of Equation (18) are negative, which finishes this case.

Having proved the claim, we plug $x := X$ into Equation (18) and take expectations. We get

$$\Pr[X > L'_X(\lambda) - a] \geq g(t) \cdot \mathrm{E}[\frac{1}{g(X)}] + g'(t) \cdot \mathrm{E}[\frac{X-t}{g(X)}]\,. \tag{19}$$

Let us compute the two expected values on the right side of Equation (19). By Equation (7), the first term is

$$\begin{aligned}\mathrm{E}[\frac{1}{g(X)}] &= \mathrm{E}[(X - L'_X(\lambda) + a)e^{\lambda X}] \\ &= a\mathrm{E}[e^{\lambda X}] \\ &= ae^{L_X(\lambda)}\,.\end{aligned}$$

By Equations (7) and (8), the second term cancels out:

$$\begin{aligned}\mathrm{E}[\frac{X-t}{g(X)}] &= \mathrm{E}[(X - L'_X(\lambda) - L''_X(\lambda)/a)(X - L'_X(\lambda) + a)e^{\lambda X}] \\ &= \mathrm{E}[(X - L'_X(\lambda))^2 e^{\lambda X}] - L''_X(\lambda)\mathrm{E}[e^{\lambda X}] \\ &= L''_X(\lambda)\mathrm{E}[e^{\lambda X}] - L''_X(\lambda)\mathrm{E}[e^{\lambda X}] \\ &= 0\,.\end{aligned}$$

Plugging these two values into Equation (19), we get

$$\begin{aligned}\Pr[X > L'_X(\lambda) - a] &\geq g(t) \cdot ae^{L_X(\lambda)} \\ &= e^{-\lambda[L'_X(\lambda)+L''_X(\lambda)/a]} \cdot \frac{1}{L''_X(\lambda)/a + a} \cdot ae^{L_X(\lambda)} \\ &= e^{L_X(\lambda)-\lambda[L'_X(\lambda)+L''_X(\lambda)/a]} \cdot \frac{a^2}{L''_X(\lambda) + a^2}\,.\end{aligned}$$

That is the inequality we wanted. □

As a corollary, we deduce the following well-known inequality, sometimes called the Chebyshev–Cantelli inequality.

Corollary 3.14 (Chebyshev–Cantelli) *Let X be a random variable and let $a \geq 0$. Then*

$$\Pr[X > \mathrm{E}[X] - a] \geq \frac{a^2}{\mathrm{Var}[X] + a^2}\,.$$

Proof. Set $\lambda := 0$ in the previous theorem, and use the identities $L_X(0) = 0$, $L'_X(0) = \mathrm{E}[X]$, and $L''_X(0) = \mathrm{Var}[X]$. □

As an application, we derive a useful bound on the expected distance of a random point from a large set in our metric space $G = (V, \rho, \mu)$. Let S be a nonempty subset of V. Define $D_S \colon V \to \mathbf{R}$ by

$$D_S(v) = \min_{w \in S} \rho(v, w).$$

By the triangle inequality, D_S is Lipschitz.

Theorem 3.15 *If $\mu(S) \geq \frac{1}{2}$, then $\mathrm{E}[D_S] \leq \sqrt{c(G)}$.*

Proof. Because D_S is Lipschitz, $\mathrm{Var}[D_S] \leq c(G)$. Let us apply the Chebyshev–Cantelli inequality with $X := D_S$ and $a := \mathrm{E}[D_S]$. We get

$$\begin{aligned} \frac{1}{2} &\geq 1 - \mu(S) \\ &= \Pr[D_S > 0] \\ &\geq \frac{\mathrm{E}[D_S]^2}{\mathrm{Var}[D_S] + \mathrm{E}[D_S]^2}\,. \end{aligned}$$

Solving for $\mathrm{E}[D_S]$ gives

$$\mathrm{E}[D_S] \leq \sqrt{\mathrm{Var}[D_S]} \leq \sqrt{c(G)}\,,$$

which completes the proof. □

As another application, we show that $f_n(d) = 0$ for d just a little larger than $\mathrm{mad}(G)n$.

Theorem 3.16 *If $d \geq \mathrm{mad}(G)n + \sqrt{c(G)n}$, then $f_n(d) = 0$.*

Proof. Suppose that $d > \mathrm{mad}(G)n + \sqrt{c(G)n}$. (The case $d = \mathrm{mad}(G)n + \sqrt{c(G)n}$ will follow from the right-continuity of f_n.) Let S be a subset of V^n with $\mu_n(S) \geq \frac{1}{2}$. We will show that

$\mu_n(\overline{B[S,d]}\,) = 0$. To do so, let $v \in V^n$ be an arbitrary point with $\mu_n(v) > 0$; we will show that $v \in B[S,d]$. By the definition of $\mathrm{mad}(G)$, we have

$$\begin{aligned} \mathrm{E}[D_v] &= \sum_{i=1}^{n} \mathrm{E}[D_{v_i}] \\ &\leq \sum_{i=1}^{n} \mathrm{mad}(G) \\ &= \mathrm{mad}(G)n. \end{aligned} \tag{20}$$

Also, because D_v is Lipschitz, we have

$$\mathrm{Var}[D_v] \leq c(G^n) = c(G)n. \tag{21}$$

We will apply the Chebyshev–Cantelli inequality with $X := -D_v$ and $a := d - \mathrm{E}[D_v]$. By Equations (20) and (21), we have $a^2 > c(G)n \geq \mathrm{Var}[D_v]$. Hence we get

$$\begin{aligned} \mu_n(B[v,d]) &= \Pr[D_v \leq d] \\ &= \Pr[D_v \leq \mathrm{E}[D_v] + a] \\ &\geq \frac{a^2}{\mathrm{Var}[D_v] + a^2} \\ &> \frac{1}{2}. \end{aligned}$$

Because $\mu_n(B[v,d]) > \frac{1}{2}$ and $\mu_n(S) \geq \frac{1}{2}$, we have $S \cap B[v,d] \neq \emptyset$, which is equivalent to $v \in B[S,d]$. That is what we wanted to show. □

In this theorem, we can weaken the hypothesis to $d \geq \mathrm{mad}(G)(n+3)$, using the Berry–Esseen theorem (see Feller [15, Section XVI.5]). We omit the details.

3.4 Upper Bound

Let d be such that $d \gg \sqrt{n}$ and $\mathrm{mad}(G)n - d \gg \sqrt{n}$. Our goal is to show that $\ln f_n(d) \sim -nR_G(d/n)$. In this section we will prove the upper bound on $\ln f_n(d)$.

Let S be a subset of V^n with $\mu_n(S) \geq \frac{1}{2}$. Let $\lambda \geq 0$ be arbitrary. By Theorem 3.15 and Corollaries 3.7 and 3.8, we have

$$\begin{aligned} \mu_n(\overline{B[S,d]}\,) &= \Pr[D_S > d] \\ &\leq e^{-\lambda d}\mathrm{E}[e^{\lambda D_S}] \\ &= e^{-\lambda d} e^{\lambda \mathrm{E}[D_S]} \mathrm{E}[e^{\lambda(D_S - \mathrm{E}[D_S])}] \\ &\leq e^{-\lambda d} e^{\lambda \mathrm{E}[D_S]} e^{L_{G^n}(\lambda)} \\ &\leq e^{-\lambda d} e^{\lambda\sqrt{c(G^n)}} e^{L_{G^n}(\lambda)} \\ &\leq e^{-\lambda d} e^{\lambda\sqrt{c(G)n}} e^{L_G(\lambda)n} \\ &= e^{-n[\lambda(d/n - \sqrt{c(G)/n}\,) - L_G(\lambda)]}. \end{aligned}$$

Taking the infimum over all $\lambda \geq 0$ gives

$$\mu_n(\overline{B[S,d]}\,) \leq e^{-nR_G(d/n-\sqrt{c(G)/n}\,)}.$$

Set $s = d/n - \sqrt{c(G)/n}$ and $t = d/n$. We just need to argue that $R_G(s) \sim R_G(t)$. The proof divides into two cases.

Case $d \leq n^{5/6}$**:** This case follows from Theorem 3.12.

Case $d \geq n^{5/6}$**:** By Equation (17), we have

$$R_G(t) - R_G(s) \leq \sqrt{\frac{c(G)}{n}} \cdot \frac{R_G(\mathrm{mad}(G))}{\mathrm{mad}(G) - t}\,. \tag{22}$$

The conditions $d \geq n^{5/6}$ and $\mathrm{mad}(G)n - d \gg \sqrt{n}$ imply $t^2(\mathrm{mad}(G) - t) \gg 1/\sqrt{n}$. Plugging back into Equation (22), and using Theorem 3.11, we get

$$R_G(t) - R_G(s) \ll t^2 \leq \mathrm{diam}(G)^2 R_G(t)/2.$$

Hence $R_G(s) \sim R_G(t)$, which proves the upper bound.

3.5 Lower Bound

Again, let d be such that $d \gg \sqrt{n}$ and $\mathrm{mad}(G)n - d \gg \sqrt{n}$. We will prove a lower bound on $\ln f_n(d)$. Choose

$$t = \frac{d}{n} + \frac{\sqrt{c(G)}}{\sqrt{n}} + \frac{\mathrm{diam}(G)}{2\sqrt{n}}\,.$$

Note that $0 < t < \mathrm{mad}(G)$ for sufficiently large n. Hence, by Theorem 3.10, there is a t-extreme value; call it λ. Because the function $\alpha \mapsto \alpha t - L_G(\alpha)$ has a maximum at λ, its right derivative at λ is at most 0, and its left derivative is at least 0. In other words, we have

$$L_G^{\ell}(\lambda) \leq t \leq L_G^{\mathrm{r}}(\lambda). \tag{23}$$

Because $L_G^{\mathrm{r}}(\lambda) \geq t$, there must be a λ-optimal function $Y\colon V \to \mathbf{R}$ such that $L_Y'(\lambda) \geq t$. Because $L_G^{\ell}(\lambda) \leq t$, there must be a λ-optimal function $Z\colon V \to \mathbf{R}$ such that $L_Z'(\lambda) \leq t$. Hence there is a number p between 0 and 1 such that

$$pL_Y'(\lambda) + (1-p)L_Z'(\lambda) = t. \tag{24}$$

Choose $k = \lceil pn \rceil$. Define $W\colon V^n \to \mathbf{R}$ by

$$W(v) = \sum_{i=1}^{k} Y(v_i) + \sum_{i=k+1}^{n} Z(v_i). \tag{25}$$

Because Y and Z are Lipschitz, so is W. Because Y and Z both have mean 0, so does W. Because Y and Z (being Lipschitz) both have variance at most $c(G)$, the variance of W is at most $c(G)n$.

Consider the set

$$S = \{\, v \in V^n : W(v) < \sqrt{c(G)n}\,\}. \tag{26}$$

By the Chebyshev–Cantelli inequality (Corollary 3.14) applied to $X := -W$ and $a := \sqrt{c(G)n}$, we have

$$\begin{aligned}\mu_n(S) &= \Pr[W < \sqrt{c(G)n}\,] \\ &\geq \frac{c(G)n}{\mathrm{Var}[W] + c(G)n} \\ &\geq \frac{c(G)n}{c(G)n + c(G)n} \\ &= \frac{1}{2}\,.\end{aligned}$$

Because W is Lipschitz, Equation (26) implies that

$$B[S,d] \subseteq \{\, v \in V^n : W(v) < \sqrt{c(G)n} + d\,\},$$

and so

$$\overline{B[S,d]} \supseteq \{\, v \in V^n : W(v) \geq \sqrt{c(G)n} + d\,\}.$$

Hence

$$\mu_n(\overline{B[S,d]}\,) \geq \Pr[W \geq \sqrt{c(G)n} + d]. \tag{27}$$

To estimate this probability, we first obtain estimates on $L_W(\lambda)$, $L'_W(\lambda)$, and $L''_W(\lambda)$. From Equation (25), note that

$$L_W = kL_Y + (n-k)L_Z. \tag{28}$$

In particular, because Y and Z are λ-optimal, we have

$$\begin{aligned}L_W(\lambda) &= kL_Y(\lambda) + (n-k)L_Z(\lambda) \\ &= kL_G(\lambda) + (n-k)L_G(\lambda) \\ &= nL_G(\lambda). \end{aligned}\tag{29}$$

By Equations (28) and (24), and the definition of k, we have

$$\begin{aligned}L'_W(\lambda) &= kL'_Y(\lambda) + (n-k)L'_Z(\lambda) \\ &= pnL'_Y(\lambda) + (1-p)nL'_Z(\lambda) + (k-pn)[L'_Y(\lambda) - L'_Z(\lambda)] \\ &= nt + (\lceil pn \rceil - pn)[L'_Y(\lambda) - L'_Z(\lambda)]\,. \end{aligned}\tag{30}$$

From the paragraph following Equation (23), we have $L'_Y(\lambda) \geq L'_Z(\lambda)$. Because L_Z is convex, we have $L'_Z(\lambda) \geq L'_Z(0) = \mathrm{E}[Z] = 0$. Also, by Theorem 3.1, we have $L'_Y(\lambda) \leq \mathrm{mad}(G)$. Summarizing, we have

$$0 \leq L'_Y(\lambda) - L'_Z(\lambda) \leq \mathrm{mad}(G).$$

Plugging back into Equation (30) gives

$$nt \leq L'_W(\lambda) \leq nt + \mathrm{mad}(G). \tag{31}$$

By Theorem 3.2, we have

$$\begin{aligned} L''_W(\lambda) &= kL''_Y(\lambda) + (n-k)L''_Z(\lambda) \\ &\leq k\frac{\mathrm{diam}(G)^2}{4} + (n-k)\frac{\mathrm{diam}(G)^2}{4} \\ &= \frac{\mathrm{diam}(G)^2}{4}n\,. \end{aligned} \tag{32}$$

Define $a := \mathrm{diam}(G)\sqrt{n}/2$. Note that $L''_W(\lambda) \leq a^2$ by Equation (32). We will apply Theorem 3.13 with $X := W$. Using Equations (27), (31), and (29), we get

$$\begin{aligned} \mu_n(\overline{B[S,d]}\,) &\geq \Pr[W > \sqrt{c(G)n} + d] \\ &= \Pr[W > nt - a] \\ &\geq \Pr[W > L'_W(\lambda) - a] \\ &\geq e^{L_W(\lambda) - \lambda[L'_W(\lambda) + L''_W(\lambda)/a]} \cdot \frac{a^2}{L''_W(\lambda) + a^2} \\ &\geq e^{nL_G(\lambda) - \lambda[nt + \mathrm{mad}(G) + \mathrm{diam}(G)\sqrt{n}/2]} \cdot \frac{1}{2} \\ &\geq \frac{1}{2}e^{-nR_G(t + \mathrm{diam}(G)/(2\sqrt{n}\,) + \mathrm{mad}(G)/n)}\,. \end{aligned}$$

We just need to argue that

$$R_G(t + \frac{\mathrm{diam}(G)}{2\sqrt{n}} + \frac{\mathrm{mad}(G)}{n}) \sim R_G(\frac{d}{n}).$$

The proof of this asymptotic equation is essentially the same as the continuity argument at the end of our proof of the upper bound. That observation completes the proof of the lower bound.

We can weaken the hypothesis $\mathrm{mad}(G)n - d \gg \sqrt{n}$ to $d \leq \mathrm{mad}(G)(n-10)$ and still prove the asymptotic estimate $\ln f_n(d) \sim -R_G(d/n)n$. The proof replaces our second-moment method with the Berry–Esseen theorem (see Feller [15, Section XVI.5]). We omit the details.

3.6 Game Theory

Our original solution of the linear-distance case was in terms of a game. Although our current solution does not use this game, we describe the game here for its possible independent interest.

We call our game the Bernstein–Lipschitz game. Assume the space $G = (V, \rho, \mu)$ has been fixed. The game involves a parameter $t \in \mathbf{R}$. There are two players, Bernstein and Lipschitz. Bernstein chooses a real number λ. Lipschitz chooses a Lipschitz function $X: V \to \mathbf{R}$ such that $\mathrm{E}[X] = 0$. Lipschitz pays to Bernstein the amount $\lambda t - \ln \mathrm{E}[e^{\lambda X}]$.

The Bernstein–Lipschitz game is a two-person, zero-sum game. In general, such games have to allow randomized strategies for both players, but our game is special. Note that the payoff function is concave in λ. Hence Bernstein has an optimal strategy that is deterministic. Using that observation, it is easy to see that the value of the game is exactly $R_G(t)$.

Note that Bernstein's strategy domain is the 1-dimensional set $\mathbf{R}$. From the main result of Bohnenblust, Karlin, and Shapley [9], it follows that Lipschitz must have an optimal strategy that is the mixture of only 2 deterministic strategies. In the proof of our lower bound, we have borrowed some ideas from that paper, namely in our choice of the functions Y and Z and the "mixture" W.

We named Bernstein in honor of Sergei N. Bernstein (1880–1968), who was apparently the first person to use the exponential-moment method to bound the tail of a random variable. We named Lipschitz in honor of Rudolf O. S. Lipschitz (1832–1903), who developed the concept of Lipschitz continuity.

Acknowledgments

We thank Jeong Han Kim and Charles Epstein for helpful comments. We thank Rodney Canfield for pointing out the references [2, 13, 14], Prasad Tetali for pointing out the reference [8], and an anonymous referee for pointing out the connection to Cramér's Theorem.

References

[1] R. Ahlswede, E. Yang and Z. Zhang, *Identification via compressed data*, to appear.

[2] V. B. Alekseev, *The number of monotone k-valued functions*, Problemy Kibernet. 28 (1974), 5–24 (in Russian). Correction in Problemy Kibernet. 29 (1974), 248.

[3] N. Alon, J. H. Kim, and J. Spencer, *Nearly perfect matchings in regular simple hypergraphs*, Israel J. Mathematics 100 (1997), 171-187.

[4] N. Alon and M. Krivelevich, *Constructive bounds for a Ramsey-type problem*, Graphs and Combinatorics 13 (1997), 217-225.

[5] N. Alon and V. D. Milman, λ_1*, isoperimetric inequalities for graphs, and superconcentrators*, J. Combinatorial Theory Ser. B 38 (1985), 73–88.

[6] N. Alon and J. Spencer, **The Probabilistic Method**, Wiley, 1992.

[7] T. V. Arak and A. Y. Zaĭtsev, **Uniform Limit Theorems for Sums of Independent Random Variables**, Proc. of Steklov Institute of Math. 174, AMS, 1988.

[8] S. G. Bobkov and C. Houdré, *Variance of Lipschitz functions and an isoperimetric problem for a class of product measures*, Bernoulli 2 (1996), 249–255.

[9] H. F. Bohnenblust, S. Karlin, and L. S. Shapley, Games with continuous, convex pay-off, in *Contributions to the Theory of Games vol. 1*, Annals of Mathematics Studies no. 24, (H. W. Kuhn and A. W. Tucker, eds.), Princeton Univ. Press (1950), 181–192.

[10] B. Bollobás and I. Leader, *An isoperimetric inequality on the discrete torus*, SIAM J. Discrete Math. 3 (1990), 32–37.

[11] B. Bollobás and I. Leader, *Compressions and isoperimetric inequalities*, J. Combinatorial Theory Ser. A 56 (1991), 47–62.

[12] V. F. Dem'yanov and L. V. Vasil'ev, **Nondifferentiable Optimization**, Optimization Software, Inc., Publications Division, New York, 1985.

[13] K. Engel, *Optimal representations of partially ordered sets and a limit Sperner theorem*, European J. Combinatorics 7 (1986), 287–302.

[14] K. Engel, **Sperner Theory**, Cambridge Univ. Press, 1997.

[15] W. Feller, **An Introduction to Probability Theory and Its Applications** Volume II (Second Edition), Wiley, 1971.

[16] P. Frankl and Z. Füredi, *A short proof of a theorem of Harper about Hamming spheres*, Discrete Math. 34 (1981), 311–313.

[17] G. H. Hardy, J. E. Littlewood, and G. Pólya, **Inequalities** (Second Edition), Cambridge Univ. Press, 1952.

[18] L. Harper, *Optimal numbering and isoperimetric problems on graphs*, J. Combinatorial Theory 1 (1966), 385–394.

[19] J. Kahn, *Asymptotically good list-colorings*, J. Combinatorial Theory Ser. A 73 (1996), 1–59.

[20] B. Maurey, *Construction de suites symétriques*, Compt. Rend. Acad. Sci. Paris 288 (1979), 679–681 (in French).

[21] C. J. H. McDiarmid, On the method of bounded differences, in *Surveys in Combinatorics 1989*, London Math. Society Lecture Notes Series 141 (J. Siemons, ed.), Cambridge Univ. Press (1989), 148–188.

[22] V. D. Milman and G. Schechtman, **Asymptotic Theory of Finite Dimensional Normed Spaces**, Lecture Notes in Mathematics 1200, Springer–Verlag, Berlin and New York, 1986.

[23] R. T. Rockafellar, **Convex Analysis**, Princeton Univ. Press, 1970.

[24] W. F. Stout, **Almost Sure Convergence**, Academic Press, New York, 1974.

[25] D. W. Stroock, **An Introduction to the Theory of Large Deviations**, Springer Verlag, Berlin, 1984.

[26] A. Shwartz and A. Weiss, **Large Deviations for Performance Analysis**, Chapman & Hall, London, 1995.

[27] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Hautes Études Sci. Publ. Math. 81 (1995), 73–205.

[28] S. R. S. Varadhan, **Large Deviations and Applications**, SIAM, Philadelphia, 1984.