Sparsifying sums of norms

Arun Jambulapati[*] James R. Lee[†] Yang P. Liu[‡] Aaron Sidford[§]

Abstract

For any norms $N_1, \ldots, N_m$ on $\mathbb{R}^n$ and $N(x) := N_1(x) + \cdots + N_m(x)$, we show there is a sparsified norm $\tilde{N}(x) = w_1 N_1(x) + \cdots + w_m N_m(x)$ such that $|N(x) - \tilde{N}(x)| \leqslant \varepsilon N(x)$ for all $x \in \mathbb{R}^n$, where $w_1, \ldots, w_m$ are non-negative weights, of which only $O(\varepsilon^{-2} n \log(n/\varepsilon)(\log n)^{2.5})$ are non-zero. Additionally, we show that such weights can be found with high probability in time $O(m(\log n)^{O(1)} + \mathrm{poly}(n))T$, where T is the time required to evaluate a norm $N_i(x)$, assuming that $N(x)$ is poly(n)-equivalent to the Euclidean norm. This immediately yields analogous statements for sparsifying sums of symmetric submodular functions. More generally, we show how to sparsify sums of pth powers of norms when the sum is p-uniformly smooth.

Contents

[*]University of Washington, jmblpati@uw.edu
[†]University of Washington, jrl@cs.washington.edu
[‡]Stanford University, yangpliu@stanford.edu
[§]Stanford University, sidford@stanford.edu

1 Introduction

Consider a collection $N_1, \dots, N_m : \mathbb{R}^n \to \mathbb{R}_+$ of semi-norms[1] on $\mathbb{R}^n$ and the semi-norm defined by

$$N(x) := N_1(x) + \cdots + N_m(x)\,.$$

It is natural to ask whether N can be *sparsified* in the following sense. Given nonnegative weights $w_1, \dots, w_m$, define the approximator $\tilde{N}(x) := w_1 N_1(x) + \cdots + w_m N_m(x)$. We say that $\tilde{N}$ is *s-sparse* if at most s of the weights $\{w_i\}$ are non-zero, and that $\tilde{N}$ is an *ε-approximation of N* if it holds that

$$|N(x) - \tilde{N}(x)| \leqslant \varepsilon N(x)\,, \quad \forall x \in \mathbb{R}^n\,. \tag{1.1}$$

A prototypical example occurs for cut sparsifiers of weighted graphs. In this case, one has an undirected graph $G = (V, E, c)$ with nonnegative weights $\{c_e : e \in E\}$, with $n = |V|$ and $N(x) := \sum_{uv \in E} c_{uv} |x_u - x_v|$. A *weighted cut sparsifier* is given by nonnegative edge weights $\{w_e : e \in E\}$. Defining $\tilde{N}(x) := \sum_{uv \in E} w_{uv} c_{uv} |x_u - x_v|$, the typical approximation criterion is that

$$|N(x) - \tilde{N}(x)| \leqslant \varepsilon N(x)\,, \quad \forall x \in \{0,1\}^V\,, \tag{1.2}$$

where $x \in \{0,1\}^V$ naturally indexes cuts in G. A straightforward ℓ_1 variant of the discrete Cheeger inequality shows that (1.2) is equivalent to (1.1) in the setting of weighted graphs.

Benczúr and Karger [BK96] showed that for every graph G and every $\varepsilon > 0$, one can construct an s-sparse ε-approximate cut sparsifier with $s \leqslant O(\varepsilon^{-2} n \log n)$. Their result addresses the case when each N_i is a 1-dimensional semi-norm of the form $N_i(x) = c_{uv}|x_u - x_v|$. We show that one can obtain similar sparsifiers in substantial generality. Further, we show how to compute such sparsifiers efficiently when the norm N *is (r, R)-rounded*, i.e. $r\|x\|_2 \leqslant N(x) \leqslant R\|x\|_2$ for all $x \in \mathbb{R}^n$.

Theorem 1.1. *Consider a collection $N_1, \dots, N_m$ of semi-norms on $\mathbb{R}^n$ and $N(x) := N_1(x) + \cdots + N_m(x)$. For every $\varepsilon > 0$, there is an $O(\varepsilon^{-2} n \log(n/\varepsilon)(\log n)^{2.5})$-sparse ε-approximation of N. Further, if the norm N is (r, R)-rounded, then weights realizing the approximation can be found in time $O(m(\log n)^{O(1)} + \mathrm{poly}(n))\mathcal{T}_{\mathrm{eval}} \log(mR/r)$ with high probability if each N_i can be evaluated in time $\mathcal{T}_{\mathrm{eval}}$.*

Application to symmetric submodular functions. A function $f : 2^V \to \mathbb{R}_+$ is *submodular* if

$$f(S \cup \{v\}) - f(S) \geqslant f(T \cup \{v\}) - f(v), \quad \forall S \subseteq T \subseteq V, v \in V \setminus T\,.$$

A submodular function is *symmetric* if $f(S) = f(V \setminus S)$ for all $S \subseteq V$.

Consider submodular functions $f_1, \dots, f_m : \{0,1\}^V \to \mathbb{R}_+$ and denote $F(S) := f_1(S) + \cdots + f_m(S)$. Given nonnegative weights $w_1, \dots, w_m$, define $\tilde{F}(S) := w_1 f_1(S) + \cdots + w_m f_m(S)$. We say that $\tilde{F}$ is an *s-sparse ε-approximation for F* if it holds that at most s of the weights $\{w_i\}$ are non-zero and

$$|F(S) - \tilde{F}(S)| \leqslant \varepsilon F(S)\,, \quad \forall S \subseteq V\,.$$

Motivated by the ubiquity of submodular functions in machine learning and data mining, Rafiey and Yoshida [RY22] established in this setting that, even if the f_i are asymmetric, for every $\varepsilon > 0$,

[1] A semi-norm N is nonnegative and satisfies $N(\lambda x) = |\lambda| N(x)$ and $N(x+y) \leqslant N(x) + N(y)$ for all $\lambda \in \mathbb{R}$, $x, y \in \mathbb{R}^n$, though possibly $N(x) = 0$ for $x \neq 0$.

there is an $O(Bn^2/\varepsilon^2)$-sparse ε-approximation for F, where $n := |V|$ and B is the maximum number of vertices in the base polytope of any f_i. In the case $B \leqslant O(1)$, their result is tight for (directed) cuts in directed graphs [CKP+17].

However, for *symmetric* submodular functions, the situation is better. For such functions $f : 2^V \to \mathbb{R}_+$ with $f(\emptyset) = 0$, the Lovász extension [Lov83] of f is a semi-norm on $\mathbb{R}^V$ (see Section 3.4.1). Therefore, Theorem 1.1 immediately yields an analogous sparsification result in this setting. In comparison to [RY22], in this symmetric setting, we have no dependence on B, and the quadratic dependence on n improves to nearly linear.

Corollary 1.2 (Symmetric submodular functions)**.** *If $f_1, \ldots, f_m : 2^V \to \mathbb{R}_+$ are symmetric submodular functions with $f_1(\emptyset) = \cdots = f_m(\emptyset) = 0$, and $F(S) := f_1(S) + \cdots + f_m(S)$, then for every $\varepsilon > 0$, there is an $O(\varepsilon^{-2} n \log(n/\varepsilon)(\log n)^{2.5})$-sparse ε-approximation of F where $n = |V|$.*

Additionally, if the functions f_i are integer-valued with $\max_{i\in[m],S\subseteq V} f_i(S) \leqslant R$, then the weights realizing the approximation can be found in time $O(mn(\log n)^{O(1)} + \mathrm{poly}(n))\mathcal{T}_{\mathrm{eval}} \log(mR)$, with high probability, assuming each f_i can be evaluated in time $\mathcal{T}_{\mathrm{eval}}$.

The deduction of Corollary 1.2 from Theorem 1.1 appears in Section 3.4.1.

Sums of higher powers. In the setting of graphs, *spectral sparsification* [ST11], a notion stronger than (1.2), has been extensively studied. Given semi-norms $N_1, \ldots, N_m$ on $\mathbb{R}^n$, define a semi-norm via their ℓ_2-sum as

$$N(x)^2 := N_1(x)^2 + \cdots + N_m(x)^2 .$$

If $w_1, \ldots, w_m$ are nonnegative weights and $\tilde{N}(x)^2 := w_1 N_1(x)^2 + \cdots + w_m N_m(x)^2$, we say that $\tilde{N}^2$ is an *s-sparse ε-approximation for* N^2 if it holds that at most s of the weights $\{w_i\}$ are non-zero and

$$\left|N(x)^2 - \tilde{N}(x)^2\right| \leqslant \varepsilon N(x)^2 , \quad \forall x \in \mathbb{R}^n . \tag{1.3}$$

When $G = (V, E, c)$ is a weighted graph and each $N_i(x)$ is of the form $\sqrt{c_{uv}}|x_u - x_v|$ for some $uv \in E$, (1.3) is called an *ε-spectral sparsifier of G*. In this setting, a sequence of works [ST11, SS11, BSS12] culminates in the existence of $O(n/\varepsilon^2)$-sparse ε-approximations for every $\varepsilon > 0$. These results generalize [Rud99, BSS14] to the setting of arbitrary 1-dimensional semi-norms, where

$$N_1(x) = |\langle a_1, x\rangle|, \ldots, N_m(x) = |\langle a_m, x\rangle|, \qquad a_1, \ldots, a_m \in \mathbb{R}^n . \tag{1.4}$$

We establish the existence of near-linear-size sparsifiers for sums of powers of a substantially more general class of higher-dimensional norms. Recall that a semi-norm N on $\mathbb{R}^n$ is said to be *p-uniformly smooth with constant S* if it holds that

$$\frac{N(x+y)^p + N(x-y)^p}{2} \leqslant N(x)^p + N(Sy)^p \qquad x, y \in \mathbb{R}^n . \tag{1.5}$$

Note that when $N_i(x) = |\langle a_i, x\rangle|$, then N is 2-uniformly smooth with constant 1. We say that two semi-norms N_X and N_Y are *K-equivalent* if there is a number $\lambda > 0$ such that $N_Y(z) \leqslant \lambda N_X(z) \leqslant K N_Y(z)$ for all $z \in \mathbb{R}^n$. Every norm is 1-uniformly smooth with constant 1 by the triangle inequality, so the next theorem generalizes Theorem 1.1.

Theorem 1.3 (Sums of pth powers of uniformly smooth norms)**.** *Consider $p \geqslant 1$ and semi-norms $N_1, \dots, N_m$ on $\mathbb{R}^n$. Denote $N(x)^p := N_1(x)^p + \cdots + N_m(x)^p$, and suppose that for some numbers $K, S > 1$ the norm N is K-equivalent to a norm which is $\min(p,2)$-uniformly smooth with constant S. Then for every $\varepsilon \in (0,1)$, there is an $O(s)$-sparse ε-approximation to N^p such that*

$$s \leqslant \begin{cases} \frac{K^{2p}}{\varepsilon^2} n \left(S\psi_n \log(n/\varepsilon)\right)^p (\log n)^2 & 1 \leqslant p \leqslant 2 \\ \frac{K^{2p} S^p p^2}{\varepsilon^2} \left(\frac{n+p}{2}\right)^{p/2} \left(\psi_n \log(n/\varepsilon)\right)^2 (\log n)^2 & p \geqslant 2\,. \end{cases}$$

Above, we use $\psi_n \leqslant O(\sqrt{\log n})$ [Kla23] to denote the KLS constant on $\mathbb{R}^n$ (see Theorem 1.7 below).

Note that for $N(x)$ to be $\min(p,2)$-uniformly smooth with constant $O(S)$, it suffices that each N_i is $\min(p,2)$-uniformly smooth with constant S [Fig76]. To see the relevance of this theorem in the case $p = 2$, note that by John's theorem, every d-dimensional semi-norm is $\sqrt{d}$-equivalent to a Euclidean norm (which is 2-uniformly smooth with constant 1). So if $A_1, \dots, A_m \in \mathbb{R}^{d\times n}$ and $\hat{N}_1, \dots, \hat{N}_m$ are *arbitrary* norms on $\mathbb{R}^d$, then taking $N_i(x) := \hat{N}_i(A_i x)$, we obtain an $O(d\varepsilon^{-2} n (\log(n/\varepsilon))^2 (\log n)^3)$-sparse ε-approximation to N^2, substantially generalizing the setting of (1.4) (albeit with an extra $d(\log(n/\varepsilon))^{O(1)}$ factor in the sparsity).

Unlike in the setting of graph sparsifiers where spectral sparsification is a strictly stronger notion (due to the equivalence of (1.2) and (1.1)), the notions of approximation guaranteed by Theorem 1.1 and Theorem 1.3 for $p > 1$ are, in general, incomparable. For example, even if $\|\tilde{A}x\|_2 \approx \|Ax\|_2$ for all $x \in \mathbb{R}^n$, it is not necessarily true that $\|\tilde{A}x\|_1 \approx \|Ax\|_1$ for all $x \in \mathbb{R}^n$.

Let us now discuss some consequences of Theorem 1.3.

Dimension reduction for ℓ_p sums. Fix $1 \leqslant p \leqslant 2$ and a subspace $X \subseteq \ell_p^m$ with $\dim(X) = n$. It is known that for any $\varepsilon > 0$, there is a subspace $\tilde{X} \subseteq \ell_p^d$ with $d \leqslant O(\varepsilon^{-2} n \log(n)(\log\log n)^2)$ such that the ℓ_p norms on X and $\tilde{X}$ are $(1+\varepsilon)$-equivalent [Tal95]. For $p = 1$, this can be improved to $d \leqslant O(\varepsilon^{-2} n \log n)$ [Tal90].

Consider the following more general setting. Suppose $Z_1, \dots, Z_m$ are each p-uniformly smooth Banach spaces with their smoothness constants bounded by S. Let us write $(Z_1 \oplus \cdots \oplus Z_m)_p$ for the Banach space $Z = Z_1 \oplus \cdots \oplus Z_m$ equipped with the norm

$$\|x\|_Z := \left(\|x\|_{Z_1}^p + \cdots + \|x\|_{Z_m}^p\right)^{1/p}.$$

Theorem 1.3 shows the following: For any n-dimensional subspace $X \subseteq Z$ and $\varepsilon > 0$, there are indicies $i_1, \dots, i_d \in \{1, \dots, m\}$ with $d \leqslant O((S/\varepsilon)^{-2} n (\log(n/\varepsilon))^p (\log n)^{2+p/2})$ and a subspace $\tilde{X} \subseteq (Z_{i_1} \oplus \cdots \oplus Z_{i_d})_p$ that is $(1+\varepsilon)$-equivalent to X. The aforementioned results for subspaces of ℓ_p^m correspond to the setting where each Z_i is 1-dimensional. The case $p \geqslant 2$ of Theorem 1.3 similarly generalizes [BLM89].

Application to spectral hypergraph sparsifiers. Consider a weighted hypergraph $H = (V, E, c)$, where $\{c_e : e \in E\}$ are nonnegative weights. To every hyperedge $e \in E$, one can associate the semi-norm $N_e(x) := \sqrt{c_e}\max_{u,v\in e} |x_u - x_v|$, and the hypergraph energy

$$N(x)^2 := \sum_{e\in E} N_e(x)^2\,.$$

Soma and Yoshida [SY19] formalized the notion of spectral sparsification for hypergraphs; it coincides with the notion of approximation expressed in (1.3). In this setting, a sequence of works [SY19, BST19, KKTY21b, KKTY21a, JLS22, Lee22] culminates in the existence of $O(\varepsilon^{-2}n(\log n)^2)$-sparse ε-approximations to N^2 for every $\varepsilon > 0$.

One can obtain a similar result via an application of Theorem 1.3, as follows. We can express each hyperedge norm as $N_e(x) = \|A_e x\|_\infty$, where $A_e : \mathbb{R}^n \to \mathbb{R}^{\binom{|e|}{2}}$ is defined by $(A_e x)_{uv} = x_u - x_v$ for all $\{u, v\} \in \binom{e}{2}$. The ℓ_∞ norm on $\mathbb{R}^d$ is K-equivalent to the $\ell_{\lceil \log d \rceil}$ norm with $K = O(1)$, and the ℓ_p norm on $\mathbb{R}^n$ is 2-uniformly smooth with constant $S \leqslant O(\sqrt{p})$ [Han56]. Applying Theorem 1.3 with $S \leqslant O(\sqrt{\log n})$ and $K \leqslant O(1)$ yields $O(\varepsilon^{-2}n(\log(n/\varepsilon))^2(\log n)^4)$-sparse ε-approximators in this special case, nearly matching the known results on spectral hypergraph sparsification. Additionally, Theorem 1.3 can be applied to give nontrivial sparsification results in substantially more general settings, as the next example shows.

Example 1.4 (Sparsification for matrix norms)**.** Consider a matrix generalization of this setting: $X \in \mathbb{R}^{d \times d}$, and matrices $S_1, \dots, S_m$ with $S_i \in \mathbb{R}^{d_i \times d}$, and $T_1, \dots, T_m$ with $T_i \in \mathbb{R}^{d \times e_i}$. Define $N_i(X) := \|S_i X T_i\|_{op}$, where $\|\cdot\|_{op}$ denotes the operator norm. Then the norm given by $N(X) := (\|S_i X T_i\|_{op}^2 + \cdots + \|S_m X T_m\|_{op}^2)^{1/2}$ can be sparsified down to $O((d/\varepsilon)^2(\log(d/\varepsilon))^2(\log d)^4)$ terms. This follows because the Schatten p-norm of an operator is 2-uniformly smooth with constant $O(\sqrt{p})$ [BCL94], and for rank d matrices, the Schatten p-norm is $O(1)$-equivalent to the operator norm when $p \asymp \log d$.

Further results and open questions for sums of squared norms. The *rank of a hypergraph H* is defined as the quantity $r := \max_{e \in E} |e|$. The best-known result for spectral hypergraph sparsification is due to [JLS22, Lee22]: For every $\varepsilon > 0$, there is an $O(\varepsilon^{-2}\log(r) \cdot n \log n)$-sparse ε-approximation to N^2. In Section 4, we obtain the following generalization.

Theorem 1.5 (Sums of squares of ℓ_p norms)**.** *Consider a family of operators $\{A_i : \mathbb{R}^n \to \mathbb{R}^{k_i} : i \in [m]\}$, and $2 \leqslant p_1, \dots, p_m \leqslant p$. Suppose that $N_1, \dots, N_m$ are semi-norms on $\mathbb{R}^n$ and that $N_i(x)$ is K-equivalent to $\|A_i x\|_{p_i}$ for all $i \in [m]$. Then for every $\varepsilon > 0$, there is an $O((K^3/\varepsilon)^2 pn \log(n/\varepsilon))$-sparse ε-approximation to N^2 where $N(x)^2 := N_1(x)^2 + \cdots + N_m(x)^2$.*

In particular, if $k_1, \dots, k_m \leqslant r$, then each $\|A_i x\|_\infty$ is $O(1)$-equivalent to $\|A_i x\|_p$ for $p \asymp \log r$, and thus Theorem 1.5 generalizes the aforementioned result for spectral hypergraph sparsifiers. One should note that, for any fixed $p \geqslant 2$, Theorem 1.5 is tight for methods based on independent sampling, by the coupon collector bound. (Although it is known that in some settings [BSS12] the $\log(n)$ factor can be removed by other methods.)

It is a fascinating open question whether the assumption of p-uniform smoothness can be dropped from Theorem 1.3. In Section 4.2, we show that it is possible to obtain a non-trivial result for sums of pth powers of general norms for $p \in [1, 2]$.

Theorem 1.6 (General sums of pth powers)**.** *If $N_1, \dots, N_m$ are arbitrary semi-norms on $\mathbb{R}^n$, $1 \leqslant p \leqslant 2$, and $N(x)^p := N_1(x)^p + \cdots + N_m(x)^p$, then for every $\varepsilon > 0$, there is an s-sparse ε-approximation to N^p with*

$$s \lesssim \varepsilon^{-2}\left(n^{2-1/p}\log(n/\varepsilon)(\log n)^{1/2} + n\log(n/\varepsilon)^p(\log n)^{2+p/2}\right).$$

Note that in the $p = 2$ case, applying Theorem 1.3 directly for $K = \sqrt{n}, S = 1, p = 2$ results in a worse sparsity bound of $O(\varepsilon^{-2}n^2\log(n/\varepsilon)^2(\log n)^3)$.

1.1 Importance sampling for general norms

Let us now fix semi-norms $N_1, \ldots, N_m$ on $\mathbb{R}^n$ and define $N(x) := N_1(x) + \cdots + N_m(x)$ for all $x \in \mathbb{R}^n$, as in the setting of Theorem 1.1. Our method for constructing sparsifiers is simply independent sampling: Consider a probability distribution $\rho = (\rho_1, \ldots, \rho_m)$ on $\{1, \ldots, m\}$, and then sample M indices $i_1, \ldots, i_M$ independently from ρ and take

$$\tilde{N}(x) := \frac{1}{M}\left(\frac{N_{i_1}(x)}{\rho_{i_1}} + \cdots + \frac{N_{i_m}(x)}{\rho_{i_m}}\right).$$

We have $\mathbb{E}[N_{i_1}(x)/\rho_{i_1}] = N(x)$, and therefore $\mathbb{E}[\tilde{N}(x)] = N(x)$ for any fixed x.

In order for these unbiased estimators to be suitably concentrated, it is essential to choose a suitable distribution ρ. To indicate the subtlety involved, we recall two choices for the case of graphs. Suppose that G consists of edges $\{u_1, v_1\}, \ldots, \{u_m, v_m\}$ and $N_i(x) = |x_{u_i} - x_{v_i}|$ for each $i \in [m]$. Benczúr and Karger [BK96] define ρ_i to be inversely proportional to the largest k such that the edge $\{u_i, v_i\}$ is contained in a maximal induced k-edge-connected subgraph. Spielman and Srivastava [SS11] define ρ_i as proportional to the effective resistance across the edge $\{u_i, v_i\}$ in G.

Let μ denote the probability measure on $\mathbb{R}^n$ whose density is proportional to $e^{-N(x)}$. We will take ρ_i proportional to the average of $N_i(x)$ under this measure:

$$\rho_i := \frac{\int_{\mathbb{R}^n} N_i(x)\, e^{-N(x)}\, dx}{\int_{\mathbb{R}^n} N(x)\, e^{-N(x)}\, dx}. \tag{1.6}$$

To motivate this choice of $\rho = (\rho_1, \ldots, \rho_m)$, let us now explain the general framework for analyzing sparsification by i.i.d. random sampling and chaining.

Symmetrization. For any norm N on $\mathbb{R}^n$, we use the notation $B_N := \{x \in \mathbb{R}^n : N(x) \leqslant 1\}$. Our goal is to control the maximum deviation

$$\mathbb{E} \max_{x \in B_N} \left|\tilde{N}(x) - \mathbb{E}[\tilde{N}(x)]\right|.$$

By a standard symmetrization argument (see Section 2.2), to bound this quantity by $O(\delta)$, it suffices to prove that for every *fixed* choice of indices $i_1, \ldots, i_M$, we have

$$\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_M} \frac{1}{M} \sum_{j=1}^{M} \varepsilon_i \frac{N_{i_j}(x)}{\rho_{i_j}} \leqslant \delta \left(\max_{x \in B_N} \tilde{N}(x)\right)^{1/2}, \tag{1.7}$$

where $\varepsilon_1, \ldots, \varepsilon_M \in \{-1, 1\}$ are uniformly random signs.

Chaining and entropy estimates. If we define $V_x := \frac{1}{M}(\varepsilon_1 N_{i_1}(x)/\rho_{i_1} + \cdots + \varepsilon_M N_{i_M}(x)/\rho_{i_M})$, then $\{V_x : x \in \mathbb{R}^n\}$ is a subgaussian process, and $\mathbb{E}\max\{V_x : x \in B_N\}$ can be controlled via standard chaining arguments (see Section 2.1 for background on subgaussian processes, covering numbers, and chaining upper bounds). Define the distance

$$d(x, y) := \left(\mathbb{E}\,|V_x - V_y|^2\right)^{1/2} = \frac{1}{M}\left(\sum_{j=1}^{M}\left(\frac{N_{i_j}(x) - N_{i_j}(y)}{\rho_{i_j}}\right)^2\right)^{1/2}.$$

and let $\mathcal{K}(B_N, d, r)$ denote the minimum number K such that B_N can be covered by K balls of radius r in the metric d. Then Dudley's entropy bound (Lemma 2.3) asserts that

$$\mathbb{E} \max_{x \in B_N} V_x \lesssim \int_0^\infty \sqrt{\log \mathcal{K}(B_N, d, r)}\, dr\,, \tag{1.8}$$

Our goal, then, is to choose sampling probabilities $\rho_1, \ldots, \rho_m$ so as to make the covering numbers $\mathcal{K}(B_N, d, r)$ suitably small.

In order to get a handle on the distance d, let us define

$$\mathcal{N}^\infty(x) := \max_{j \in [M]} \frac{N_{i_j}(x)}{\rho_{i_j}}, \quad \text{and} \quad \kappa := \max\{\mathcal{N}^\infty(x) : x \in B_N\}\,.$$

Then we can bound

$$\begin{aligned} d(x,y) &\leqslant M^{-1/2} \sqrt{\mathcal{N}^\infty(x-y)} \left(\frac{1}{M} \sum_{j=1}^{M} \frac{|N_{i_j}(x) - N_{i_j}(y)|}{\rho_{i_j}} \right)^{1/2} \\ &\leqslant M^{-1/2} \sqrt{\mathcal{N}^\infty(x-y)} \left(2 \max_{x \in B_N} \tilde{N}(x) \right)^{1/2}. \end{aligned}$$

Using this in (1.8) gives the upper bound

$$\begin{aligned} \mathbb{E} \max_{x \in B_N} V_x &\lesssim M^{-1/2} \left(\max_{x \in B_N} \tilde{N}(x) \right)^{1/2} \int_0^\infty \sqrt{\log \mathcal{K}(B_N, (\mathcal{N}^\infty)^{1/2}, r)}\, dr \\ &= M^{-1/2} \left(\max_{x \in B_N} \tilde{N}(x) \right)^{1/2} \int_0^{\sqrt{\kappa}} \sqrt{\log \mathcal{K}(B_N, \mathcal{N}^\infty, r^2)}\, dr\,, \end{aligned} \tag{1.9}$$

where we have used that the last integrand vanishes above $\sqrt{\kappa}$ since $B_N \subseteq \kappa B_{\mathcal{N}^\infty}$.

Dual-Sudakov inequalities. In order to bound the entropy integral (1.9), let us recall the dual-Sudakov inequality (see [PTJ85] and [LT11, (3.15)]) which allows one to control covering numbers of the Euclidean ball. Let B_2^n denote the Euclidean ball in $\mathbb{R}^n$. Then for any norm N on $\mathbb{R}^n$, it holds that

$$\sqrt{\log \mathcal{K}(B_2^n, N, r)} \lesssim \frac{1}{r} \mathbb{E}\,[N(\boldsymbol{g})]\,, \tag{1.10}$$

where $\boldsymbol{g}$ is a standard n-dimensional Gaussian.

An adaptation of the Pajor-Talagrand proof of (1.10) (see Lemma 3.2) allows one to show that for any norms N and $\hat{N}$ on $\mathbb{R}^n$,

$$\log \mathcal{K}(B_N, \hat{N}, r) \lesssim \frac{1}{r} \mathbb{E}\left[\hat{N}(\boldsymbol{Z})\right]\,, \tag{1.11}$$

where $\boldsymbol{Z}$ has law $\mu \propto e^{-N(x)}$. A closely related estimate was proved by Milman and Pajor [MP89]; see the remarks after (1.16).

In particular, we can apply this with $\hat{N} = \mathcal{N}^\infty$, yielding

$$\log \mathcal{K}(B_N, \mathcal{N}^\infty, r) \lesssim \frac{1}{r} \mathbb{E}\,[\mathcal{N}^\infty(\boldsymbol{Z})] = \frac{1}{r} \mathbb{E} \max_{j \in [M]} \frac{N_{i_j}(\boldsymbol{Z})}{\rho_{i_j}}\,. \tag{1.12}$$

At this point, it is quite natural to hope that $N_j(\boldsymbol{Z})$ is concentrated around its mean, in which case the choice $\rho_j \propto \mathbb{E}[N_j(\boldsymbol{Z})]$ seems appropriate. Indeed, this is the first point at which we will employ convexity in an essential way. The density $e^{-N(x)}$ is log-concave, and therefore $\boldsymbol{Z}$ is a log-concave random variable. By recent progress on the KLS conjecture, we know that Lipschitz functions of isotropic log-concave vectors concentrate tightly around their mean.

Let ψ_n denote the KLS constant in dimension n. In the past few years there has been remarkable progress on bounding ψ_n [Che21, KL22, JLV22, Kla23]. In particular, Klartag and Lehec established that $\psi_n \leqslant O((\log n)^5)$, and the best current bound is $\psi_n \leqslant O(\sqrt{\log n})$ [Kla23].

Exponential concentration and the KLS conjecture. The next lemma expresses a classical connection between exponential concentration and Poincaré inequalities [GM83]. Say that $\varphi : \mathbb{R}^n \to \mathbb{R}$ is *L-Lipschitz* if $\|\varphi(x) - \varphi(y)\|_2 \leqslant L\|x-y\|_2$ for all $x, y \in \mathbb{R}^n$.

Theorem 1.7. *There is a constant $c > 0$ such that the following holds. Suppose $\boldsymbol{X}$ is a random variable on $\mathbb{R}^n$ whose law is isotropic and log-concave. Then for every L-Lipschitz function $\varphi : \mathbb{R}^n \to \mathbb{R}$ and $t > 0$,*

$$\mathbb{P}\left(|\varphi(\boldsymbol{X}) - \mathbb{E}[\varphi(\boldsymbol{X})]| > t\right) \leqslant 2e^{-ct/(\psi_n L)}\,.$$

In Section 2.3, we prove the following consequence.

Corollary 1.8. *There is a constant $c > 0$ such that the following holds. Consider a norm $\mathcal{N}$ on $\mathbb{R}^n$ and a random vector $\boldsymbol{Z}$ whose distribution is symmetric and log-concave. Then for any $t > 0$,*

$$\mathbb{P}\left(\left|\mathcal{N}(\boldsymbol{Z}) - \mathbb{E}[\mathcal{N}(\boldsymbol{Z})]\right| > t\right) \leqslant 2\exp\left(-\frac{c}{\psi_n}\frac{t}{\mathbb{E}[\mathcal{N}(\boldsymbol{Z})]}\right).$$

With this in hand, we can immediately use a union bound (see Lemma 3.7) to obtain

$$\mathbb{E}\max_{j\in[M]} \frac{N_{i_j}(\boldsymbol{Z})}{\mathbb{E}[N_{i_j}(\boldsymbol{Z})]} \lesssim \psi_n \log M\,.$$

To make ρ a probability measure, we take $\rho_j := \mathbb{E}[N_j(\boldsymbol{Z})]/\mathbb{E}[N(\boldsymbol{Z})]$ for $j = 1, \ldots, m$, and then (1.12) becomes

$$\log \mathcal{K}(B_N, \mathcal{N}^\infty, r) \lesssim \frac{1}{r}(\psi_n \log M)\,\mathbb{E}[N(\boldsymbol{Z})] = \frac{1}{r} n\psi_n \log(M)\,, \tag{1.13}$$

where the last inequality uses $\mathbb{E}[N(\boldsymbol{Z})] = n$, which follows from a straightforward integration using that the law of $\boldsymbol{Z}$ has density proprtional to $e^{-N(x)}$ (Lemma 3.14). Thus we have

$$\int_{\sqrt{\kappa}/n^2}^{\sqrt{\kappa}} \sqrt{\log \mathcal{K}(B_N, \mathcal{N}^\infty, r^2)}\,dr \lesssim \left(n\psi_n \log M\right)^{1/2} \int_{\sqrt{\kappa}/n^2}^{\sqrt{\kappa}} \frac{1}{r}\,dr \lesssim \left(n\psi_n \log M\right)^{1/2} \log n\,. \tag{1.14}$$

Standard volume arguments in $\mathbb{R}^n$ (Lemma 2.4) allow us to control the rest of the integral:

$$\int_0^{1/n^2} \sqrt{\log \mathcal{K}(B_{\mathcal{N}^\infty}, \mathcal{N}^\infty, r^2)}\,dr \lesssim 1\,,$$

and therefore

$$\int_0^{\sqrt{\kappa}/n^2} \sqrt{\log \mathcal{K}(B_N, \mathcal{N}^\infty, r^2)}\,dr \leqslant \sqrt{\kappa}\int_0^{1/n^2} \sqrt{\log \mathcal{K}(B_{\mathcal{N}^\infty}, \mathcal{N}^\infty, r^2)}\,dr \lesssim \sqrt{\kappa}\,.$$

Plugging this and (1.14) into (1.9) gives

$$\mathbb{E} \max_{x \in B_N} V_x \lesssim M^{-1/2} \left(\max_{x \in B_N} \tilde{N}(x) \right)^{1/2} \left(\sqrt{\kappa} + \left(n\psi_n \log M \right)^{1/2} \log n \right) .$$

Finally, observe that (1.13) gives the bound $\kappa \lesssim n\psi_n \log(M)$, resulting in

$$\mathbb{E} \max_{x \in B_N} V_x \lesssim \left(\frac{n\psi_n \log(M)(\log n)^2}{M} \right)^{1/2} \left(\max_{x \in B_N} \tilde{N}(x) \right)^{1/2} .$$

Choosing $M \asymp \delta^{-2} n (\log n)^2 \psi_n \log(n/\delta)$ yields our desired goal (1.7).

Modifications for sums of pth powers. In order to apply these methods to sums of pth powers $N(x)^p = N_1(x)^p + \cdots + N_m(x)^p$ for $1 \leqslant p \leqslant 2$, we use the natural analog of (1.6):

$$\rho_i := \frac{\int_{\mathbb{R}^n} N_i(x)^p \, e^{-N(x)^p} \, dx}{\int_{\mathbb{R}^n} N(x)^p \, e^{-N(x)^p} \, dx} . \tag{1.15}$$

Note that if $p = 2$ and one defines $N_i(x) := |(Ax)_i|$ for a matrix $A \in \mathbb{R}^{m \times n}$, then $2\rho_i$ are exactly the leverage scores of A. For $p > 2$, we choose ρ_i proportional to $\int_{\mathbb{R}^n} N_i(x)^p \, e^{-N(x)^2} \, dx$.

The main hurdle in this setting is that we only establish the analog of (1.11) for p-uniformly smooth norms: As shown in Lemma 3.2, if $\boldsymbol{Z}$ has the law whose density is proportional to $e^{-N(x)^p}$ and N is p-uniformly smooth, then for any norm $\hat{N}$,

$$(\log \mathcal{K}(B_N, B_{\hat{N}}, r))^{1/p} \lesssim \frac{1}{r} \mathbb{E}[\hat{N}(\boldsymbol{Z})] . \tag{1.16}$$

A closely-related estimate is mentioned in [MP89, Eq. (9)], where instead the distribution of $\boldsymbol{Z}$ is uniform on B_N.

General norms and block Lewis weights. To obtain Theorem 1.6 for general norms, we resort to a dimension-dependent version of (1.16) (see Lemma 4.7). Moreover, we need to augment the sampling probabilities in (1.15) in order to effectively bound the diameter $\mathrm{diam}(B_N, \mathcal{N}^\infty)$. For this, as well as for sums of squares of ℓ_p norms (Theorem 1.5), in Section 4 we formulate a generalization of ℓ_p Lewis weights, motivated by the construction of weights in [KKTY21a, JLS22, Lee22].

For a collection of vectors $a_1, \ldots, a_k \in \mathbb{R}^n$, the ℓ_p Lewis weights [Lew78, Lew79] result from consideration of the optimization

$$\max\{|\det(U)| : \alpha(U) \leqslant 1\} , \tag{1.17}$$

where α is the norm on linear operators $U : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$\alpha(U) = \left(\sum_{i=1}^{k} \|U a_i\|_2^p \right)^{1/p} .$$

In Section 4, we consider a substantial generalization of this setting where $S_1 \cup \cdots \cup S_m = \{1, \ldots, k\}$ is a partition of the index set. Given $p_1, \ldots, p_m \geqslant 2$ and $q \geqslant 1$, we define the norm

$$\alpha(U) := \left(\sum_{j=1}^{m} \left(\sum_{i \in S_j} \|U a_i\|_2^{p_j} \right)^{q/p_j} \right)^{1/q} ,$$

and establish properties of the corresponding optimizer of (1.17).

1.2 Computing the sampling weights via homotopy

In Section 3.4, we present an algorithm constructing a sparsifier for $N(x) = N_1(x) + \cdots + N_m(x)$ that runs in time $n^{O(1)}$ plus the time required to do $m(\log n)^{O(1)} + n^{O(1)}$ total evaluations of norms $N_i(y)$ for various $i \in [m]$ and $y \in \mathbb{R}^n$. It employs a homotopy-type method that has been used for efficient sparsification in a range of settings (see, e.g., [MP12, KLM+17, JSS18, AJSS19]).

To compute reasonable overestimates of the sampling weights $\{\rho_i\}$ from (1.6), we must sample from the probability measure μ with density proportional to $e^{-N(x)}$. Sampling from a log-concave distribution is a well-studied task, and can be done in $n^{O(1)}$ time; see [LV07] and Theorem 3.20. If a norm evaluation can be performed in time $\mathcal{T}_{\text{eval}}$, this would naively require time $mn^{O(1)}\mathcal{T}_{\text{eval}}$, whereas we would like our algorithms to run in nearly "input-sparsity time," as expressed before.

We first observe that one need only sample from a distribution with density $\propto e^{-\tilde{N}(x)}$ for some norm $\tilde{N}$ that is $O(1)$-equivalent to $N(x)$. Given this fact, let us suppose that our norm is (r, R)-rounded in the sense that $r\|x\|_2 \leqslant N(x) \leqslant R\|x\|_2$ for all $x \in \mathbb{R}^n$. Define the family of norms $N_t(x) := N(x) + t\|x\|_2$. For $t = R$, it holds that $N_R(x)$ is 2-equivalent to the norm $R\|x\|_2$, and sampling from the distribution with density $\propto e^{-R\|x\|_2}$ is trivial. Therefore we can construct an $n(\log n)^{O(1)}$-sparse 1/2-approximation $\tilde{N}_R(x)$ to $N_R(x)$.

Now assuming we have an $n(\log n)^{O(1)}$-sparse 1/2-approximation $\tilde{N}_t$ to N_t for $r \leqslant t \leqslant R$, we construct a sparsifier for $N_{t/2}(x)$ by sampling from the measure with density $\propto e^{-\tilde{N}_t(x)}$. This works because $\tilde{N}_t$ is 2-equivalent to N_t, which is 2-equivalent to $N_{t/2}$. After $O(\log(R/r))$ iterations, we arrive at sparse norm $\tilde{N}$ that is $O(1)$-equivalent to N, and then by sampling from the distribution with density $\propto e^{-\tilde{N}(x)}$, we are able to construct a sparse ε-approximation to N itself.

2 Preliminaries

Let us denote $[n] := \{1, 2, \ldots, n\}$. All logarithms are taken with base e unless otherwise indicated. We use the notation $a \lesssim b$ if there exists a universal constant $C > 0$ such that $a \leqslant Cb$, and the notation $a \asymp b$ for the conjunction of $a \lesssim b$ and $b \lesssim a$. We say that an algorithm succeeds *with high probability* if for everything constant $C > 0$, there is a choice of constants in the algorithm that makes it have success probability at least $1 - n^{-C}$.

In this paper, there is no argument that distinguishes norms and semi-norms, and for simplicitly we use "norm" to refer to both.

2.1 Covering numbers, chaining, and subgaussian processes

Consider a metric space (T, d). A random process $\{V_x : x \in T\}$ is said to be *subgaussian with respect to d* if there is a number $\alpha > 0$ such that

$$\mathbb{P}\left(|V_x - V_y| > t\right) \leqslant \exp\left(\frac{-t^2}{\alpha^2\, d(x,y)^2}\right), \qquad t > 0\,. \tag{2.1}$$

Say that $\{V_x : x \in T\}$ is *centered* if $\mathbb{E}[V_x] = 0$ for all $x \in T$.

Given a metric space (T, d), define the ball $B(x, r) := \{y \in T : d(x, y) \leqslant r\}$.

Definition 2.1 (Covering and entropy numbers)**.** For a number $r > 0$, we define the *covering number* $\mathcal{K}(T, d, r)$ as the smallest number of balls $\{B(x_i, r) : i \geqslant 1\}$ required to cover T. Define the *entropy numbers* $e_h(T, d) := \inf\{r > 0 : \mathcal{K}(T, d, r) \leqslant 2^{2^h}\}$ for $h \geqslant 0$.

If $T \subseteq \mathbb{R}^n$ and N is a norm on $\mathbb{R}^n$, it induces a natural distance $d(x, y) := N(x - y)$ on T. In this case we use the notation $\mathcal{K}(T, N, r)$ and $e_h(T, N)$ to denote the associated covering and entropy numbers respectively. We additionally write $B_N := \{x \in \mathbb{R}^n : N(x) \leqslant 1\}$ for the unit ball of N.

The generic chaining functional. Recall Talagrand's generic chaining functional [Tal14, Def. 2.2.19]:

$$\gamma_2(T, d) := \inf_{\{\mathcal{A}_h\}} \sup_{x \in T} \sum_{h=0}^{\infty} 2^{h/2} \mathrm{diam}(\mathcal{A}_h(x), d)\,, \tag{2.2}$$

where the infimum runs over all sequences $\{\mathcal{A}_h : h \geqslant 0\}$ of partitions of T satisfying $|\mathcal{A}_h| \leqslant 2^{2^h}$ for each $h \geqslant 0$. Note that we use the notation $\mathcal{A}_h(x)$ for the unique set of $\mathcal{A}_h$ that contains x. The next theorem constitutes the generic chaining upper bound; see [Tal14, Thm 2.2.18].

Theorem 2.2. *If $\{V_x : x \in T\}$ is a centered subgaussian process satisfying (2.1) with respect to distance d, then*

$$\mathbb{E} \sup_{x \in T} V_x \lesssim \alpha \gamma_2(T, d)\,. \tag{2.3}$$

A classical way of controlling $\gamma_2(T, d)$ is given by Dudley's entropy bound (see, e.g., [Tal14, Prop 2.2.10]). The follow two upper bounds are equivalent up to universal constant factors.

Lemma 2.3 (Dudley)**.** *For any metric space (T, d), it holds that*

$$\gamma_2(T, d) \lesssim \sum_{h \geqslant 0} 2^{h/2} e_h(T, d) \tag{2.4}$$

$$\gamma_2(T, d) \lesssim \int_0^\infty \sqrt{\log \mathcal{K}(T, d, r)}\, dr\,. \tag{2.5}$$

The next lemma follows from a straightforward volume argument.

Lemma 2.4. *If N is a norm on $\mathbb{R}^n$, then for any $\varepsilon > 0$ and $h \geqslant 0$,*

$$\mathcal{K}(B_N, N, \varepsilon) \leqslant \left(\frac{2}{\varepsilon}\right)^n \quad \text{and} \quad e_h(B_N, N) \leqslant 2 \cdot 2^{-2^h/n}\,.$$

To show that our sampling algorithms succeed with high probability, as opposed to only in expectation, we use the following refinement of Theorem 2.2.

Theorem 2.5 ([Tal14, Thm 2.2.27])**.** *Suppose $\{V_x : x \in T\}$ is a centered subgaussian process with respect to the distance d. Then for some constants $c > 0, C > 1$ and any $\lambda > 0$,*

$$\mathbb{P}\left(\sup_{x \in T} |V_x| > C\left(\gamma_2(T, d) + \lambda \mathrm{diam}(T, d)\right)\right) \lesssim \exp\left(-c\lambda^2\right)\,.$$

In particular, if $Z = \sup_{x \in T} |V_x|$, then for any $\lambda > 0$,

$$\log \mathbb{E}[e^{\lambda Z}] \lesssim \lambda^2 \mathrm{diam}(T, d)^2 + \lambda \gamma_2(T, d)\,. \tag{2.6}$$

2.2 Sparsification via subgaussian processes

Consider $\varphi_1, \varphi_2, \ldots, \varphi_m : \mathbb{R}^n \to \mathbb{R}_+$, and define

$$F(x) := \sum_{j \in [m]} \varphi_j(x) .$$

Given a probability vector $\rho \in \mathbb{R}_+^m$, and an integer $M \geqslant 1$ and $\nu = (\nu_1, \ldots, \nu_M) \in [m]^M$, define the distance

$$d_{\rho,\nu}(x,y) := \left(\sum_{j \in [M]} \left(\frac{\varphi_{\nu_j}(x) - \varphi_{\nu_j}(y)}{\rho_{\nu_j} \cdot M} \right)^2 \right)^{1/2} . \tag{2.7}$$

and the function $\tilde{F}_{\rho,\nu} : \mathbb{R}^n \to \mathbb{R}_+$

$$\tilde{F}_{\rho,\nu}(x) := \frac{1}{M} \sum_{j \in [M]} \frac{\varphi_{\nu_j}(x)}{\rho_{\nu_j}} .$$

Denote $B_F := \{x \in \mathbb{R}^n : F(x) \leqslant 1\}$.

The next lemma employs a variant of a standard symmetrization argument to control $\mathbb{E} \max_{x \in B_F} |F(x) - \tilde{F}_{\rho,\boldsymbol{\nu}}(x)|$ using an associated subgaussian process (see, for example, [Tal14, Lem 9.1.11]). We also prove a version with a tail bound to show that our algorithms succeed with high probability.

Lemma 2.6. *Consider $M \geqslant 1$ and a probability vector $\rho \in \mathbb{R}_+^m$. Suppose that for some $\delta \in (0,1)$ and every $\nu \in [m]^M$, it holds that*

$$\gamma_2(B_F, d_{\rho,\nu}) \leqslant \delta \left(\max_{x \in B_F} \tilde{F}_{\rho,\nu}(x) \right)^{1/2} . \tag{2.8}$$

If $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M$ are sampled independently from ρ, then

$$\mathbb{E} \max_{x \in B_F} \left| F(x) - \tilde{F}_{\rho,\boldsymbol{\nu}}(x) \right| \lesssim \delta . \tag{2.9}$$

If it also holds that, for all $\nu \in [m]^M$,

$$\operatorname{diam}(B_F, d_{\rho,\nu}) \leqslant \hat{\delta} \left(\max_{x \in B_F} \tilde{F}_{\rho,\nu}(x) \right)^{1/2} , \tag{2.10}$$

then there is a universal constant $K > 0$ such that for all $0 \leqslant t \leqslant \frac{1}{2K\hat{\delta}}$,

$$\mathbb{P}\left(\max_{x \in B_F} \left| F(x) - \tilde{F}_{\rho,\boldsymbol{\nu}}(x) \right| > K(\delta + t\hat{\delta}) \right) \leqslant e^{-Kt^2/4}. \tag{2.11}$$

Proof. Note that $\mathbb{E}[\tilde{F}_{\rho,\boldsymbol{\nu}}(x)] = F(x)$ for every $x \in \mathbb{R}^n$. Thus for any convex function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$,

$$\mathop{\mathbb{E}}_{\boldsymbol{\nu}} \psi \left(\max_{F(x) \leqslant 1} \left| F(x) - \tilde{F}_{\rho,\boldsymbol{\nu}}(x) \right| \right) \leqslant \mathop{\mathbb{E}}_{\boldsymbol{\nu}, \tilde{\boldsymbol{\nu}}} \psi \left(\max_{F(x) \leqslant 1} \left| \tilde{F}_{\rho,\boldsymbol{\nu}}(x) - \tilde{F}_{\rho,\tilde{\boldsymbol{\nu}}}(x) \right| \right) , \tag{2.12}$$

where $\tilde{\boldsymbol{\nu}}$ is an independent copy of $\boldsymbol{\nu}$. The argument to ψ on the right-hand side can be written as

$$\max_{F(x)\leqslant 1}\left|\frac{1}{M}\sum_{j\in[M]}\left(\frac{\varphi_{\boldsymbol{\nu}_j}(x)}{\rho_{\boldsymbol{\nu}_j}}-\frac{\varphi_{\tilde{\boldsymbol{\nu}}_j}(x)}{\rho_{\tilde{\boldsymbol{\nu}}_j}}\right)\right|.$$

Since the distribution of $\varphi_{\boldsymbol{\nu}_j}(x)/\rho_{\boldsymbol{\nu}_j}-\varphi_{\tilde{\boldsymbol{\nu}}_j}(x)/\rho_{\tilde{\boldsymbol{\nu}}_j}$ is symmetric, we have

$$\sum_{j\in[M]}\frac{\varphi_{\boldsymbol{\nu}_j}(x)}{\rho_{\boldsymbol{\nu}_j}}-\sum_{j\in[M]}\frac{\varphi_{\tilde{\boldsymbol{\nu}}_j}(x)}{\rho_{\tilde{\boldsymbol{\nu}}_j}}\stackrel{\text{law}}{=}\sum_{j\in[M]}\varepsilon_j\cdot\left(\frac{\varphi_{\boldsymbol{\nu}_j}(x)}{\rho_{\boldsymbol{\nu}_j}}-\frac{\varphi_{\tilde{\boldsymbol{\nu}}_j}(x)}{\rho_{\tilde{\boldsymbol{\nu}}_j}}\right)$$

for any choice of signs $\varepsilon_1,\ldots,\varepsilon_M\in\{-1,1\}$. This yields the stochastic domination

$$\max_{F(x)\leqslant 1}\left|\frac{1}{M}\sum_{j=1}^{M}\frac{\varphi_{\boldsymbol{\nu}_j}(x)}{\rho_{\boldsymbol{\nu}_j}}-\frac{1}{M}\sum_{j=1}^{M}\frac{\varphi_{\tilde{\boldsymbol{\nu}}_j}(x)}{\rho_{\tilde{\boldsymbol{\nu}}_j}}\right|\preceq\max_{F(x)\leqslant 1}\left|\frac{1}{M}\sum_{j=1}^{M}\varepsilon_j\frac{\varphi_{\boldsymbol{\nu}_j}(x)}{\rho_{\boldsymbol{\nu}_j}}\right|+\max_{F(x)\leqslant 1}\left|\frac{1}{M}\sum_{j=1}^{M}\varepsilon_j\frac{\varphi_{\tilde{\boldsymbol{\nu}}_j}(x)}{\rho_{\tilde{\boldsymbol{\nu}}_j}}\right|. \tag{2.13}$$

Note that if we choose $\varepsilon_1,\ldots,\varepsilon_M\in\{-1,1\}$ to be uniformly random, then the quantity in the absolute value is a centered subgaussian process with respect to the distance $d_{\rho,\boldsymbol{\nu}}$ on B_F, so we are in position to apply Theorem 2.5.

Define $\mathcal{S}:=\mathbb{E}\max_{F(x)\leqslant 1}|F(x)-\tilde{F}_{\rho,\boldsymbol{\nu}}(x)|$ and apply (2.12) with $\psi(x)=x$ and (2.13) to obtain

$$\mathcal{S}\leqslant 2\mathop{\mathbb{E}}_{\boldsymbol{\nu}}\mathop{\mathbb{E}}_{\varepsilon}\max_{F(x)\leqslant 1}\left|\frac{1}{M}\sum_{j=1}^{M}\varepsilon_j\frac{\varphi_{\boldsymbol{\nu}_j}(x)}{\rho_{\boldsymbol{\nu}_j}}\right|\lesssim\mathop{\mathbb{E}}_{\boldsymbol{\nu}}[\gamma_2(B_F,d_{\rho,\boldsymbol{\nu}})],$$

where the final inequality is an application of Theorem 2.2. Now use (2.8) and concavity of the square root to bound

$$\mathbb{E}[\gamma_2(B_F,d_{\rho,\boldsymbol{\nu}})]\leqslant\delta\,\mathbb{E}\left[\left(\max_{F(x)\leqslant 1}\tilde{F}_{\rho,\boldsymbol{\nu}}(x)\right)^{1/2}\right]\leqslant\delta\left(\mathbb{E}\max_{F(x)\leqslant 1}\tilde{F}_{\rho,\boldsymbol{\nu}}(x)\right)^{1/2}.$$

The triangle inequality gives

$$\max_{F(x)\leqslant 1}\tilde{F}_{\rho,\boldsymbol{\nu}}(x)\leqslant 1+\max_{F(x)\leqslant 1}|F(x)-\tilde{F}_{\rho,\boldsymbol{\nu}}(x)|,$$

yielding the consequence

$$\mathcal{S}\leqslant 2\delta(1+\mathcal{S})^{1/2}.$$

Since $\delta<1$, this confirms (2.9).

Let us now verify (2.11). Fix $\lambda>0$ and define $\boldsymbol{Z}:=\max_{F(x)\leqslant 1}|F(x)-\tilde{F}_{\rho,\boldsymbol{\nu}}(x)|$. Applying (2.12) with $\psi(x)=e^{\lambda x}$ and (2.13), yields

$$\mathbb{E}[e^{\lambda\boldsymbol{Z}}]=\mathop{\mathbb{E}}_{\boldsymbol{\nu}}\exp\left(\lambda\max_{F(x)\leqslant 1}\left|F(x)-\tilde{F}_{\rho,\boldsymbol{\nu}}(x)\right|\right)\leqslant\mathop{\mathbb{E}}_{\boldsymbol{\nu}}\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_M}\exp\left(2\lambda\max_{F(x)\leqslant\lambda}\left|\frac{1}{M}\sum_{j\in[M]}\varepsilon_j\frac{\varphi_{\boldsymbol{\nu}_j}(x)}{\rho_{\boldsymbol{\nu}_j}}\right|\right)$$

$$\leqslant \mathop{\mathbb{E}}_{\boldsymbol{\nu}} \exp\left(C\left(\lambda^2 \mathrm{diam}(B_F, d_{\rho,\boldsymbol{\nu}})^2 + \lambda\gamma_2(B_F, d_{\rho,\boldsymbol{\nu}})\right)\right),$$

where the last inequality is an invocation of (2.6). Using (2.8) and (2.10), the latter quantity is bounded by

$$\mathop{\mathbb{E}}_{\boldsymbol{\nu}} \exp\left(C\left(\lambda^2\hat{\delta}^2\left(\max_{x\in B_F}\tilde{F}_{\rho,\boldsymbol{\nu}}(x)\right) + \lambda\delta\left(\max_{x\in B_F}\tilde{F}_{\rho,\boldsymbol{\nu}}(x)\right)^{1/2}\right)\right) \leqslant \mathbb{E}\exp\left(C\lambda^2\hat{\delta}^2(1+\boldsymbol{Z}) + C\lambda\delta\,(1+\boldsymbol{Z})^{1/2}\right).$$

Observe that for any $\alpha > 0$ and $z \geqslant 0$, we have $(1+z)^{1/2} \leqslant (1+\alpha)^{1/2} + \alpha^{-1/2} z$. Choose $\alpha := (4C\delta)^2$ so that

$$\mathbb{E}[e^{\lambda\boldsymbol{Z}}] \leqslant \exp\left(C\lambda^2\hat{\delta}^2 + C\lambda\delta(1+(4C\delta)^2)^{1/2}\right)\mathbb{E}\exp\left(\left(C\lambda^2\hat{\delta}^2 + \lambda/4\right)\boldsymbol{Z}\right).$$

Let us now assume that $t \leqslant 1/(2C\hat{\delta})$ and choose $\lambda := t/(2\hat{\delta})$. In this case, $C\lambda^2\hat{\delta}^2 \leqslant \lambda/4$, and therefore

$$\begin{aligned}\mathbb{E}[e^{\lambda\boldsymbol{Z}}] &\leqslant \exp\left(C\lambda^2\hat{\delta}^2 + C\lambda\delta(1+(4C\delta)^2)^{1/2}\right)\mathbb{E}\exp\left(\lambda\boldsymbol{Z}/2\right) \\ &\leqslant \exp\left(C\lambda^2\hat{\delta}^2 + C\lambda\delta(1+(4C\delta)^2)^{1/2}\right)(\mathbb{E}\, e^{\lambda\boldsymbol{Z}})^{1/2}.\end{aligned}$$

Taking logs and using $\delta \leqslant 1$ gives

$$\log\mathbb{E}[e^{\lambda\boldsymbol{Z}}] \leqslant K(\lambda^2\hat{\delta}^2 + \lambda\delta)$$

for some universal constant $K > 0$. Let us finally observe the standard consequence of Markov's inequality,

$$\log\mathbb{P}\left(\boldsymbol{Z} > K(\delta + t\hat{\delta})\right) \leqslant \log\mathbb{E}[e^{\lambda\boldsymbol{Z}}] - \lambda K(\delta + t\hat{\delta}) \leqslant K(\lambda^2\hat{\delta}^2 - t\lambda\hat{\delta}) = -Kt^2/4,$$

completing the proof. □

2.3 Concentration for Lipschitz functionals

We use the following standard concentration result for n-dimensional Gaussians (see, e.g., [LT11, (1.4)]).

Theorem 2.7. *Let $\boldsymbol{g}$ be a standard n-dimensional Gaussian. Then for every L-Lipschitz function $\varphi : \mathbb{R}^n \to \mathbb{R}$ and $t > 0$,*

$$\mathbb{P}\left(|\varphi(\boldsymbol{g}) - \mathbb{E}\,\varphi(\boldsymbol{g})| > t\right) \leqslant \exp\left(-t^2/(2L^2)\right).$$

We also use a classical moment inequality (see, e.g., [LW08, Rem. 5] for computation of the constant).

Theorem 2.8 ([BMP63])**.** *Suppose $\boldsymbol{X}$ is a real, symmetric, log-concave random variable. Then for every $p \geqslant q > 0$, it holds that*

$$(\mathbb{E}\,|\boldsymbol{X}|^p)^{1/p} \leqslant \frac{p}{q}(\mathbb{E}\,|\boldsymbol{X}|^q)^{1/q}.$$

Together with Theorem 1.7, this yields Corollary 1.8, as we now show.

Proof of Corollary 1.8. Define the covariance matrix $A := \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]$ and let $\boldsymbol{X} := A^{-1/2}\boldsymbol{Z}$. Then the law of $\boldsymbol{X}$ is log-concave and isotropic by construction. Thus Theorem 1.7 gives the desired result once we confirm the Lipschitz bound

$$\mathcal{N}(A^{1/2}x) \leqslant 2\,\mathbb{E}[\mathcal{N}(\boldsymbol{Z})] \cdot \|x\|_2 \,. \tag{2.14}$$

To this end, let $\mathcal{N}^*$ denote the dual norm to $\mathcal{N}$ and write

$$\begin{aligned}\mathcal{N}(A^{1/2}x) &= \sup_{\mathcal{N}^*(w)\leqslant 1} \langle w, A^{1/2}x\rangle \\ &= \sup_{\mathcal{N}^*(w)\leqslant 1} \langle A^{1/2}w, x\rangle \leqslant \|x\|_2 \sup_{\mathcal{N}^*(w)\leqslant 1} \|A^{1/2}w\|_2 \,.\end{aligned}$$

Then we have

$$\|A^{1/2}w\|_2 = \langle w, Aw\rangle^{1/2} = \left(\mathbb{E}[\langle w, \boldsymbol{Z}\rangle^2]\right)^{1/2} \leqslant 2\,\mathbb{E}[|\langle w, \boldsymbol{Z}\rangle|] \leqslant 2\mathcal{N}^*(w)\,\mathbb{E}[\mathcal{N}(\boldsymbol{Z})]\,,$$

where the penultimate inequality follows from Theorem 2.8, since $\langle y, \boldsymbol{Z}\rangle$ is a symmetric log-concave random variable. ◻

3 Sparsification for uniformly smooth norms

This section studies sparsification of uniformly smooth norms, and shows Theorem 1.1 and Theorem 1.3. We start by establishing a variant of the dual Sudakov lemma that will allow us to bound certain covering numbers.

3.1 Dual Sudakov lemmas for smooth norms

Lemma 3.1 (Shift lemma)**.** *Suppose N is a norm on $\mathbb{R}^n$ that is p-uniformly smooth with constant $\mathcal{S}_p$. Define the probability measure μ on $\mathbb{R}^n$ by*

$$d\mu(x) \propto \exp(-N(x)^p)\,.$$

Then for any symmetric convex body W and $z \in \mathbb{R}^n$,

$$\mu(W + z) \geqslant \exp(-\mathcal{S}_p^p N(z)^p)\,\mu(W)\,. \tag{3.1}$$

Proof. For any $z \in \mathbb{R}^n$, it holds that

$$\mu(W + z) = \frac{\int_W \exp(-N(x+z)^p)\,dx}{\int_W \exp(-N(x)^p)\,dx}\mu(W)\,.$$

Now we bound

$$\int_W \exp(-N(\sigma x + z)^p)\,dx = \int_W \mathop{\mathbb{E}}_{\sigma\in\{-1,1\}} \exp(-N(\sigma x + z)^p)\,dx$$

$$
\begin{aligned}
&\geqslant \int_W \exp\left(-\mathop{\mathbb{E}}_{\sigma\in\{-1,1\}} N(\sigma x+z)^p\right)\,dx\\
&\geqslant \int_W \exp\left(-(N(x)^p+\mathcal{S}_p^p N(z)^p)\right)\,dx\\
&= \exp(-\mathcal{S}_p^p N(z)^p)\int_W \exp(-N(x)^p)\,dx\,,
\end{aligned}
$$

where the equality uses symmetry of W, the first inequality uses convexity of $\exp(x)$, and the second inequality uses p-uniform smoothness of N (recall (1.5)). □

Lemma 3.2. *Let N and $\widehat{N}$ be norms on $\mathbb{R}^n$. Suppose that N is p-uniformly smooth with constant $\mathcal{S}_p$, and define the probability measure μ on $\mathbb{R}^n$ so that*

$$
d\mu(x) \propto \exp(-N(x)^p)\,.
$$

Then for any $\varepsilon > 0$,

$$
\left(\log\left(\mathcal{K}(B_N,\widehat{N},\varepsilon)/2\right)\right)^{1/p} \leqslant 2\frac{\mathcal{S}_p}{\varepsilon}\int \widehat{N}(x)\,d\mu(x)\,.
$$

Proof. By scaling $\widehat{N}$, we may assume that $\varepsilon = 1$. Suppose now that $x_1,\ldots,x_M \in B_N$ and $x_1 + B_{\widehat{N}},\ldots,x_M + B_{\widehat{N}}$ are pairwise disjoint. To establish an upper bound on M, let $\lambda > 0$ be a number we will choose later and write

$$
\begin{aligned}
1 \geqslant \mu\left(\bigcup_{j\in[M]}\lambda(x_j+B_{\widehat{N}})\right) &= \sum_{j\in[M]}\mu\left(\lambda x_j+\lambda B_{\widehat{N}}\right)\\
&\overset{(3.1)}{\geqslant} \sum_{j\in[M]} e^{-\lambda^p\mathcal{S}_p^p N(x_j)^p}\mu(\lambda B_{\widehat{N}}) \geqslant M e^{-\mathcal{S}_p^p\lambda^p}\mu(\lambda B_{\widehat{N}})\,,
\end{aligned}
$$

where (3.1) used Lemma 3.1 and the last inequality used $x_1,\ldots,x_M \in B_N$.

Now choose $\lambda := 2\int \widehat{N}(x)\,d\mu(x)$ so that Markov's inequality gives

$$
\mu(\lambda B_{\widehat{N}}) = \mu\left(\{x : \widehat{N}(x) \leqslant \lambda\}\right) \geqslant 1/2\,.
$$

Combining with the preceding inequality yields the upper bound

$$
\left(\log(M/2)\right)^{1/p} \leqslant \mathcal{S}_p\lambda\,.
$$

□

3.2 Entropy estimates

Here we establish our primary entropy estimate.

Definition 3.3. Consider $p \geqslant 1$ and denote $\hat{p} := \min(p,2)$. Let $\mathcal{N}$ be a norm on $\mathbb{R}^n$ that is $\hat{p}$-uniformly smooth with constant $\mathcal{S}$. Denote by μ the probability measure with $d\mu(x) \propto e^{-\mathcal{N}(x)^{\hat{p}}}$. Consider any norms $\mathcal{N}_1,\ldots,\mathcal{N}_M$ on $\mathbb{R}^n$, and define

$$
\mathsf{W}_j := \int \mathcal{N}_j(x)\,d\mu(x),\quad j\in[M]
$$

$$\mathsf{W}_{\max} := \max(\mathsf{W}_1, \dots, \mathsf{W}_M)$$
$$d(x,y) := \left(\sum_{j=1}^{M} (\mathcal{N}_j(x)^p - \mathcal{N}_j(y)^p)^2\right)^{1/2}$$
$$\mathcal{N}^\infty(x) := \max_{j\in[M]} \mathcal{N}_j(x)$$
$$\Lambda := \max_{x\in B_{\mathcal{N}}} \sum_{j=1}^{M} \mathcal{N}_j(x)^p \,.$$

We begin with some preliminary bounds valid for all $p \geqslant 1$. The next lemma will be useful in controlling the diameter of $(B_{\mathcal{N}}, d)$.

Lemma 3.4. *For any norm $\hat{N}$ on $\mathbb{R}^n$,*

$$\max_{x\in\mathcal{N}} \hat{N}(x) \leqslant 12\mathcal{S} \int \hat{N}(x)\, d\mu(x)\,.$$

Proof. Define $m := \int \hat{N}(x)\, d\mu(x)$. Markov's inequality gives $\mu(\lambda B_{\hat{N}}) \geqslant 3/4$, where $\lambda := 4m$. Now Lemma 3.1 (applied with $p = \hat{p}$) gives

$$\mu(\lambda B_{\hat{N}} + y) \geqslant \exp\left(-\mathcal{S}^{\hat{p}} \mathcal{N}(y)^{\hat{p}}\right) \mu(\lambda B_{\hat{N}}) \geqslant \frac{3}{4} \exp\left(-\mathcal{S}^{\hat{p}} \mathcal{N}(y)^{\hat{p}}\right)\,.$$

If $\mathcal{N}(y) \leqslant 3^{-1/\hat{p}}/\mathcal{S}$, this implies $\mu(\lambda B_{\hat{N}} + y), \mu(\lambda B_{\hat{N}} - y) > 1/2$. Thus $(\lambda B_{\hat{N}} + y) \cap (\lambda B_{\hat{N}} - y) \neq \emptyset$, and therefore some $z \in \lambda B_{\hat{N}}$ satisfies $z + y, z - y \in \lambda B_{\hat{N}}$. By convexity, we have $y \in \lambda B_{\hat{N}}$ as well, i.e., $\hat{N}(y) \leqslant \lambda$. Since this holds for any y satisfying $\mathcal{N}(y) \leqslant 3^{-1/\hat{p}}/\mathcal{S}$, the claim follows. □

Applying the preceding lemma with $\hat{N} = \mathcal{N}_j$ for each $j = 1, \dots, M$ gives the following.

Corollary 3.5. *It holds that*

$$\operatorname{diam}(B_{\mathcal{N}}, \mathcal{N}^\infty) \lesssim \mathcal{S} \max_{j\in[M]} \int \mathcal{N}_j(x)\, d\mu(x) = \mathcal{S}\mathsf{W}_{\max}\,.$$

We recall the following basic maximal inequality.

Fact 3.6. *If $\boldsymbol{X}_1, \dots, \boldsymbol{X}_M$ are nonnegative random variables satisfying $\mathbb{P}[\boldsymbol{X}_j \geqslant 1 + t] \leqslant C\exp(-t/\beta)$ for $t > 0$, $j \in [M]$ and some $C, \beta \geqslant 1$, then $\mathbb{E}[\max_{j\in[M]} \boldsymbol{X}_j] \lesssim \beta\left(1 + \log(CM)\right)$.*

Proof. A union bound gives $\mathbb{P}[\max_j \boldsymbol{X}_j \geqslant 1 + t] \leqslant CMe^{-\beta t}$, therefore for any $\theta > 1$,

$$\mathbb{E}[\max_j \boldsymbol{X}_j] = \int_0^\infty \mathbb{P}[\max_j \boldsymbol{X}_j \geqslant t] \leqslant \theta + CM \int_\theta^\infty e^{-\beta(t-1)}\, dt = \theta + C\beta M e^{\beta - \theta/\beta}\,.$$

Choosing $\theta := \beta(1 + 2\log(CM))$ gives $\mathbb{E}[\max_j \boldsymbol{X}_j] \lesssim \beta(1 + \log(CM))$. □

Lemma 3.7. *It holds that*

$$\int \mathcal{N}^\infty(x)\, d\mu(x) \lesssim \psi_n \log(M)\, \mathsf{W}_{\max}\,.$$

Proof. Suppose that $\boldsymbol{Z}$ has law μ, and define $\boldsymbol{X}_j := \mathcal{N}_j(\boldsymbol{Z})/\mathbb{E}[\mathcal{N}_j(\boldsymbol{Z})]$ for $j \in [M]$. Note that

$$\int \mathcal{N}^{\infty}(x)\, d\mu(x) = \mathbb{E}[\max_j \mathcal{N}_j(\boldsymbol{Z})] \leqslant \max_j \mathbb{E}[\mathcal{N}_j(\boldsymbol{Z})] \cdot \mathbb{E}[\max_j \boldsymbol{X}_j]\,.$$

Corollary 1.8 asserts that $\mathbb{P}[\boldsymbol{X}_j \geqslant t+1] \leqslant 2e^{-ct/\psi_n}$, and therefore Fact 3.6 gives $\mathbb{E}[\max_j \boldsymbol{X}_j] \lesssim \psi_n \log M$, establishing the first claimed inequality. ◻

For the remainder of this subsection, we restrict ourself to the range $p \in [1,2]$. We will control d by $\mathcal{N}^{\infty}$ using the next estimate.

Lemma 3.8. *For all $x, y \in B_{\mathcal{N}}$,*

$$d(x,y)^2 \leqslant 4\Lambda \left(\mathcal{N}^{\infty}(x-y)\right)^p\,. \tag{3.2}$$

Proof. Monotonicity of qth powers implies that for all $u, v \in \mathbb{R}$ we have $|u+v|^q \leqslant |u|^q + |v|^q$ for $q \in [0,1]$. Applying this with $u = a-b$ and $v = b$ gives $|a^q - b^q| \leqslant |a-b|^q$. Thus for real numbers $a, b \geqslant 0$ and $p \in [1,2]$, we have

$$|a^p - b^p| = |a^{p/2} - b^{p/2}||a^{p/2} + b^{p/2}| \leqslant |a-b|^{p/2}|a^{p/2} + b^{p/2}|\,.$$

Squaring both sides yields

$$|a^p - b^p|^2 \leqslant 2|a-b|^p(a^p + b^p)\,.$$

Applying this with $a = \mathcal{N}_j(x), b = \mathcal{N}_j(y)$ for each $j \in [M]$ we arrive at

$$d(x,y)^2 \leqslant 2\sum_{j=1}^{M} |\mathcal{N}_j(x) - \mathcal{N}_j(y)|^p(\mathcal{N}_j(x)^p + \mathcal{N}_j(y)^p) \leqslant 4\Lambda \max_{j\in[M]} |\mathcal{N}_j(x) - \mathcal{N}_j(y)|^p\,.$$

Finally, note that the triangle inequality gives $|\mathcal{N}_j(x) - \mathcal{N}_j(y)| \leqslant \mathcal{N}_j(x-y)$ for each $j \in [M]$, completing the proof. ◻

In conjunction with Corollary 3.5, this yields the following.

Corollary 3.9. *For $p \in [1,2]$, it holds that* $\mathrm{diam}(B_{\mathcal{N}}, d) \lesssim (\mathcal{S}\mathsf{W}_{\max})^{p/2}\sqrt{\Lambda}$.

We now prove our primary entropy estimate.

Lemma 3.10 (Entropy bound)**.** *It holds that*

$$\gamma_2(B_{\mathcal{N}}, d) \lesssim \left(\mathcal{S}\mathsf{W}_{\max}\psi_n \log M\right)^{p/2} \log(n)\sqrt{\Lambda}\,. \tag{3.3}$$

Proof. Since both sides of (3.3) scale linearly in the pth powers $\{\mathcal{N}_j^p\}$, we may assume that

$$\Lambda = \max_{x \in B_{\mathcal{N}}} \sum_{j=1}^{M} \mathcal{N}_j(x)^p = 1\,.$$

Thus Lemma 3.8 gives

$$d(x,y) \leqslant 2\mathcal{N}^{\infty}(x-y)^{p/2}\,, \qquad x, y \in B_{\mathcal{N}}\,. \tag{3.4}$$

Now, we have

$$\left(\log \mathcal{K}(B_{\mathcal{N}}, \mathcal{N}^\infty, \varepsilon)\right)^{1/p} \lesssim \frac{\mathcal{S}}{\varepsilon} \int \mathcal{N}^\infty(x)\, d\mu(x) \lesssim \frac{\mathcal{S}}{\varepsilon} \psi_n \log(M) \mathsf{W}_{\max}\,, \tag{3.5}$$

the first inequality uses Lemma 3.2 with $N = \mathcal{N}, \widehat{N} = \mathcal{N}^\infty$, and the second uses Lemma 3.7.

Define $Q := \psi_n \log(M)\mathsf{W}_{\max}$ and then using (3.4) we have

$$\sqrt{\log \mathcal{K}(B_{\mathcal{N}}, d, \varepsilon)} \lesssim \sqrt{\log \mathcal{K}(B_{\mathcal{N}}, \mathcal{N}^\infty, (\varepsilon/\sqrt{2})^{2/p})} \overset{(3.5)}{\lesssim} \frac{(\mathcal{S}Q)^{p/2}}{\varepsilon}\,,$$

which immediately yields

$$e_h(B_{\mathcal{N}}, d) \lesssim 2^{-h/2} (\mathcal{S}Q)^{p/2}\,, \qquad h \geqslant 0\,. \tag{3.6}$$

We then have

$$\begin{aligned} e_h(B_{\mathcal{N}}, d) \lesssim e_h(B_{\mathcal{N}}, \mathcal{N}^\infty)^{p/2} &\lesssim \operatorname{diam}(B_{\mathcal{N}}, \mathcal{N}^\infty)^{p/2}\, e_h(B_{\mathcal{N}^\infty}, \mathcal{N}^\infty)^{p/2} \\ &\lesssim \operatorname{diam}(B_{\mathcal{N}}, \mathcal{N}^\infty)^{p/2}\, 2^{-2^h p/2n}\,, , \end{aligned} \tag{3.7}$$

where the second inequality follows from $B_{\mathcal{N}} \subseteq \operatorname{diam}(B_{\mathcal{N}}, \mathcal{N}^\infty) \cdot B_{\mathcal{N}^\infty}$ and the final inequality is from Lemma 2.4.

Then using (3.6) and (3.7) in conjunction with the Dudley entropy bound (2.4) gives

$$\begin{aligned} \gamma_2(B_{\mathcal{N}}, d) &\lesssim \sum_{0 \leqslant h \leqslant 4 \log n} 2^{h/2} e_h(B_{\mathcal{N}}, d) + \operatorname{diam}(B_{\mathcal{N}}, \mathcal{N}^\infty)^{p/2} \sum_{h > 4 \log n} 2^{h/2} 2^{-2^h p/2n} \\ &\lesssim (\mathcal{S}Q)^{p/2} \log n + \operatorname{diam}(B_{\mathcal{N}}, \mathcal{N}^\infty)^{p/2}\,. \end{aligned}$$

In conjunction with Corollary 3.9, the proof is complete. □

3.2.1 Entropy for $p \geqslant 2$

We will continue working under the definition Definition 3.3, but now restrict ourselves to the regime $p \geqslant 2$, where $\hat{p} = 2$. We additionally define the quantity

$$\tilde{\Lambda} := \max_{x \in B_{\mathcal{N}}} \sum_{j=1}^{M} \mathcal{N}_j(x)^{2(p-1)}\,. \tag{3.8}$$

Lemma 3.11. *For all $x, y \in B_{\mathcal{N}}$, it holds that*

$$d(x,y) \leqslant p \mathcal{N}^\infty(x-y) \sqrt{2\tilde{\Lambda}}\,. \tag{3.9}$$

Proof. Note that for $p \geqslant 2$ and any $a, b \geqslant 0$, it holds that

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

Therefore

$$d(x,y) \leqslant p \max_{j \in [M]} \left(\left| \mathcal{N}_j(x) - \mathcal{N}_j(y) \right| \right) \left(\sum_{j=1}^{M} \mathcal{N}_j(x)^{2(p-1)} + \mathcal{N}_j(y)^{2(p-1)} \right)^{1/2}.$$ □

As in the previous section, we can use this in conjuction with Corollary 3.5 to bound the diameter, as $\mathrm{diam}(B_{\mathcal{N}}, d) \lesssim p\sqrt{\tilde{\Lambda}}\,\mathrm{diam}(B_{\mathcal{N}}, \mathcal{N}^{\infty})$.

Corollary 3.12. *For $p \geqslant 2$, it holds that $\mathrm{diam}(B_{\mathcal{N}}, d) \lesssim p\mathcal{S}\mathsf{W}_{\max}\sqrt{\tilde{\Lambda}}$.*

What follows is the analogous entropy bound.

Lemma 3.13 (Entropy bound)**.** *It holds that*

$$\gamma_2(B_{\mathcal{N}}, d) \lesssim p\left(\mathcal{S}\mathsf{W}_{\max}\psi_n \log(M)\log(n)\right)\sqrt{\tilde{\Lambda}}$$

Proof. Noting that both sides scale linearly in $\{\mathcal{N}_j^p\}$, we may assume that $\tilde{\Lambda} = 1$. Applying Lemma 3.11 then gives

$$d(x,y) \lesssim p\mathcal{N}^{\infty}(x-y)\,, \quad x, y \in B_{\mathcal{N}}\,. \tag{3.10}$$

Applying Lemma 3.2 with $N = \mathcal{N}, \hat{N} = \mathcal{N}^{\infty}$ yields

$$\sqrt{\log \mathcal{K}(B_{\mathcal{N}}, \mathcal{N}^{\infty}, \varepsilon)} \lesssim \frac{\mathcal{S}}{\varepsilon}\int \mathcal{N}^{\infty}(x)\,d\mu(x) \lesssim \frac{\mathcal{S}\psi_n \log M}{\varepsilon}\mathsf{W}_{\max}\,,$$

where the latter inequality is the first inequality in Lemma 3.7.

Thus defining $Q := \psi_n \log(M)\mathsf{W}_{\max}$, we have

$$e_h(B_{\mathcal{N}}, \mathcal{N}^{\infty}) \lesssim 2^{-h/2}\mathcal{S}Q\,, \quad h \geqslant 0\,. \tag{3.11}$$

Note that Corollary 3.12 gives

$$\mathrm{diam}(B_{\mathcal{N}}, \mathcal{N}^{\infty}) \lesssim \mathcal{S}\mathsf{W}_{\max}\,,$$

and therefore using Lemma 2.4,

$$e_h(B_{\mathcal{N}}, \mathcal{N}^{\infty}) \lesssim \mathcal{S}\mathsf{W}_{\max}e_h(B_{\mathcal{N}^{\infty}}, \mathcal{N}^{\infty}) \lesssim 2^{-2^h/n}\mathcal{S}\mathsf{W}_{\max}\,, \quad h \geqslant 0\,. \tag{3.12}$$

Using (3.11) and (3.12) in conjunction with the Dudley entropy bound (2.4) gives

$$\gamma_2(B_{\mathcal{N}}, \mathcal{N}^{\infty}) \lesssim \sum_{0 \leqslant h \leqslant 4\log n} 2^{h/2}e_h(B_{\mathcal{N}}, \mathcal{N}^{\infty}) + \mathcal{S}\mathsf{W}_{\max}\sum_{h > 4\log n} 2^{h/2}2^{-2^h/n} \lesssim \mathcal{S}Q\log n\,.$$

The proof is complete since $\gamma_2(B_{\mathcal{N}}, d) \lesssim p \cdot \gamma_2(B_{\mathcal{N}}, \mathcal{N}^{\infty})$ by (3.10). □

3.3 Sparsification

We now complement the preceding entropy estimates with control of our desired sampling probabilities, beginning with a simple fact.

Lemma 3.14. *For any norm N on $\mathbb{R}^n$ and $p \geqslant 1$, if μ is the probability measure with $d\mu(x) \propto e^{-N(x)^p}$, then*

$$\int N(x)^p\,d\mu(x) = \frac{n}{p}$$

Proof. Define $B(r) := \mathrm{vol}_n(rB_N) = r^n \,\mathrm{vol}_n(B_N)$, so $\frac{d}{dr}B(r) = nr^{n-1}\,\mathrm{vol}_n(B_N)$. Therefore,

$$\int_{\mathbb{R}^n} N(x)^p \, d\mu(x) = \frac{\int_{\mathbb{R}^n} N(x)^p e^{-N(x)^p} \, dx}{\int_{\mathbb{R}^n} e^{-N(x)^p} \, dx} = \frac{\int_0^\infty r^p e^{-r^p} \, dB(r)}{\int_0^\infty e^{-r^p} \, dB(r)} = \frac{\int_0^\infty e^{-r^p} r^{n-1+p} \, dr}{\int_0^\infty e^{-r^p} r^{n-1} \, dr} .$$

Make the subsitution $u = r^p$, yielding

$$\int_{\mathbb{R}^n} N(x)^p \, d\mu(x) = \frac{\int_0^\infty e^{-u} u^{n/p} \, du}{\int_0^\infty e^{-u} u^{n/p-1} \, du} = \frac{n}{p} ,$$

where the latter equality follows from integration by parts. □

Let us now fix $p \in [1,2]$, and consider norms $N_1, \ldots, N_m$ on $\mathbb{R}^n$. Define the norm $N(x)$ by $N(x)^p := N_1(x)^p + \cdots + N_m(x)^p$. Suppose that $\mathcal{N}$ is another norm on $\mathbb{R}^n$ that is p-uniformly smooth with constant $\mathcal{S}$, and that

$$\mathcal{N}(x) \leqslant N(x) \leqslant K\mathcal{N}(x) \qquad \forall x \in \mathbb{R}^n .$$

Let μ denote the probability measure whose density satisfies $d\mu(x) \propto e^{-\mathcal{N}(x)^p}$, and suppose that $\tau_1, \ldots, \tau_m \geqslant 0$ are numbers satisfying

$$\mathbb{E}[N_i(\mathbf{Z})^p] \leqslant \tau_i \leqslant 2\,\mathbb{E}[N_i(\mathbf{Z})^p] , \tag{3.13}$$

where $\mathbf{Z}$ has law μ. Define the probability vector $\rho \in \mathbb{R}_+^m$ by $\rho_i := \tau_i / \|\tau\|_1$ for $i = 1, \ldots, m$.

The next theorem establishes Theorem 1.3.

Theorem 3.15. *There is an explicit function $C : \mathbb{Z}_+ \to \mathbb{R}_+$ such that $C(n) \lesssim (K^2\mathcal{S}\psi_n)^p(\log n)^2$ for all $n \in \mathbb{Z}_+$, and so that for any $0 < \varepsilon < 1$ and $M \geqslant C(n)n(\log(n/\varepsilon))^p\varepsilon^{-2}$, the following holds. If $i_1, \ldots, i_M$ are indicies sampled independently from ρ, then with probability at least $1 - n^{-\Omega((\log n)^2)}$,*

$$\left| N(x)^p - \frac{1}{M}\sum_{j\in[M]} \frac{N_{i_j}(x)^p}{\rho_{i_j}} \right| \leqslant \varepsilon N(x)^p , \quad \forall x \in \mathbb{R}^n .$$

Proof. Note that, by (3.13),

$$\|\tau\|_1 \leqslant 2\sum_{j=1}^M \int N_j(x)^p \, d\mu(x) = 2\int N(x)^p \, d\mu(x) \leqslant 2K^p \int \mathcal{N}(x)^p \, d\mu(x) = 2K^p n/p, \tag{3.14}$$

where the last equality follows from Lemma 3.14.

Given $M \geqslant 1$ and $\nu \in [m]^M$, define the norms $\mathcal{N}_1, \ldots, \mathcal{N}_M$ by

$$\mathcal{N}_j(x)^p := \frac{N_{\nu_j}(x)^p}{M\rho_{\nu_j}} , \qquad j = 1, \ldots, M ,$$

and denote $\varphi_i(x) := N_i(x)^p$ for $i = 1, \ldots, m$ so that $d_{\rho,\nu}(x,y) = d(x,y)$, where $d_{\rho,\nu}$ is defined in (2.7), and d is from Definition 3.3.

We will apply Lemma 2.6 with $F := N$ and $\{\rho_i\}, \{\varphi_i\}$ defined as above. To do so, we require bounds on $\gamma_2(B_N, d_{\rho,\nu}) \leqslant \gamma_2(B_{\mathcal{N}}, d_{\rho,\nu})$ and $\mathrm{diam}(B_N, d_{\rho,\nu}) \leqslant \mathrm{diam}(B_{\mathcal{N}}, d_{\rho,\nu})$, where both inequalities follow because $B_N \subseteq B_{\mathcal{N}}$. Lemma 3.10 yields

$$\gamma_2(B_{\mathcal{N}}, d_{\rho,\nu}) = \gamma_2(B_{\mathcal{N}}, d) \lesssim \left(\mathcal{S}\mathsf{W}_{\max}\psi_n \log M\right)^{p/2} \log(n)\sqrt{\Lambda}, \tag{3.15}$$

Note that

$$\mathbb{E}[\mathcal{N}_j(\boldsymbol{Z})^p] = \frac{1}{M\rho_{\nu_j}} \mathbb{E}[N_{\nu_j}(\boldsymbol{Z})^p] \overset{(3.13)}{\leqslant} \frac{\tau_{\nu_j}}{M\rho_{\nu_j}} = \frac{\|\tau\|_1}{M}$$

for all $j = 1, \dots, M$, and therefore by monotonicity of pth moments,

$$\mathsf{W}_{\max} \leqslant \max_{j\in[M]} \left(\mathbb{E}[\mathcal{N}_j(\boldsymbol{Z})^p]\right)^{1/p} \leqslant \left(\frac{\|\tau\|_1}{M}\right)^{1/p}.$$

It also holds that

$$\Lambda = \max_{x\in B_{\mathcal{N}}} \sum_{j=1}^{M} \mathcal{N}_j(x)^p \leqslant K^p \max_{x\in B_N} \sum_{j=1}^{M} \mathcal{N}_j(x)^p = K^p \max_{x\in B_N} \tilde{F}_{\rho,\nu}(x).$$

Substituting these bounds into (3.15) gives

$$\begin{aligned}\gamma_2(B_N, d_{\rho,\nu}) &\lesssim M^{-1/2}\left(K^2\mathcal{S}\psi_n \log M\right)^{p/2} \log(n)\|\tau\|_1^{1/2}\sqrt{\max_{x\in B_N} \tilde{F}_{\rho,\nu}(x)} \\ &\overset{(3.14)}{\lesssim} \frac{\left(K^2\mathcal{S}\psi_n \log M\right)^{p/2}\sqrt{n}\log n}{M^{1/2}}\sqrt{\max_{x\in B_N} \tilde{F}_{\rho,\nu}(x)} \\ &\leqslant \delta\sqrt{\max_{x\in B_N} \tilde{F}_{\rho,\nu}(x)}\end{aligned}$$

for some choice of M sufficiently large and satisfying

$$M \lesssim n\delta^{-2}(\log n)^2(\log(n/\delta))^p(K^2\mathcal{S}\psi_n)^p.$$

In addition, Corollary 3.9 gives

$$\mathrm{diam}(B_{\mathcal{N}}, d_{\rho,\nu}) \lesssim (\mathcal{S}\mathsf{W}_{\max})^{p/2}\sqrt{\Lambda} \leqslant \left(K^p\mathcal{S}^p\frac{\|\tau\|_1}{M}\right)^{1/2}\left(\max_{x\in B_N} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}.$$

Using (3.14) and our choice of M, we have

$$\mathrm{diam}(B_{\mathcal{N}}, d_{\rho,\nu}) \leqslant \frac{C_0\delta}{(\log n)^{3/2}}\left(\max_{x\in B_N} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}$$

for some universal constant $C_0 > 0$.

From Lemma 2.6 (2.11), we conclude that for a universal constant $A > 0$ and any $0 \leqslant t \leqslant \frac{(\log n)^{3/2}}{2C_0 A\delta}$,

$$\mathbb{P}\left(\max_{x\in B_N}\left|N(x)^p - \frac{1}{M}\sum_{j=1}^{M}\frac{N_{i_j}(x)^p}{\rho_{i_j}}\right| > A\left(\delta + \frac{C_0 t\delta}{(\log n)^{3/2}}\right)\right) \leqslant e^{-At^2/4}. \tag{3.16}$$

For $t := \frac{(\log n)^{3/2}}{2C_0 A}$, this shows that with probability at least $1 - e^{-\Omega((\log n)^3)}$,

$$\left|N(x)^p - \frac{1}{M}\sum_{j=1}^{M}\frac{N_{i_j}(x)^p}{\rho_{i_j}}\right| \leqslant 2A\delta N(x)^p\,, \quad \forall x \in \mathbb{R}^n\,.$$

Choosing $\delta := \varepsilon/(2A)$ now yields the desired result. ◻

3.3.1 Sparsification for $p \geqslant 2$

We start by defining the sampling weights, analogous to Lemma 3.14.

Lemma 3.16. *For any norm N on $\mathbb{R}^n$ and $p \geqslant 1$, if μ is the probability measure with $d\mu(x) \propto e^{-N(x)^2}$, then*

$$\int N(x)^p \, d\mu(x) \leqslant \left(\frac{n+p}{2}\right)^{p/2}.$$

Proof. As in the proof of Lemma 3.14, write

$$\frac{\int N(x)^p e^{-N(x)^2}\,dx}{\int e^{-N(x)^2}\,dx} = \frac{\int_0^\infty r^p e^{-r^2} dB(r)}{\int_0^\infty e^{-r^2} dB(r)} = \frac{\int_0^\infty r^{p+n-1}e^{-r^2}\,dr}{\int_0^\infty r^{n-1}e^{-r^2}\,dr}\,.$$

Now make the substitution $u = r^2$, so the left-hand side is

$$\frac{\int_0^\infty u^{(p+n)/2-1}e^{-u}\,dr}{\int_0^\infty u^{n/2-1}e^{-u}\,dr} = \frac{\Gamma((p+n)/2-1)}{\Gamma((n/2)-1)}\,,$$

where we recall the definition of the Γ function: For real $t \geqslant 0$,

$$\Gamma(t) = \int_0^\infty e^{-u}u^t\,du\,.$$

Finally, note that $\Gamma(t+1) = t\Gamma(t)$ and $\Gamma(t+s) \leqslant t^s\Gamma(t)$ for all $0 < s < 1$ and $t \geqslant 0$ [Wen48], hence for $k := \lfloor p/2 \rfloor$,

$$\begin{aligned}\frac{\Gamma((p+n)/2-1)}{\Gamma((n/2)-1)} &= \left(\frac{n+p}{2}-2\right)\left(\frac{n+p}{2}-3\right)\cdots\left(\frac{n+p}{2}-(k+1)\right)\frac{\Gamma(\frac{n+p}{2}-k-1))}{\Gamma(\frac{n}{2}-1)}\\ &\leqslant \left(\frac{n+p}{2}-2\right)\left(\frac{n+p}{2}-3\right)\cdots\left(\frac{n+p}{2}-(k+1)\right)\left(\frac{n}{2}-1\right)^{p/2-k}\\ &\leqslant \left(\frac{n+p}{2}\right)^{p/2}.\end{aligned}$$

◻

Theorem 3.17. *Suppose $p > 2$ and $N_1, \ldots, N_m$ are norms on $\mathbb{R}^n$ such that the norm defined by $N(x)^p = N_1(x)^p + \cdots + N_m(x)^p$ is K-equivalent to a norm $\mathcal{N}$ that is 2-uniformly smooth with constant $\mathcal{S}$. There is a weight vector $w \in \mathbb{R}_+^m$ with*

$$|\mathrm{supp}(w)| \lesssim \frac{K^{2p}\mathcal{S}^p}{\varepsilon^2}\left(\frac{n+p}{2}\right)^{p/2}(\psi_n \log(n/\varepsilon)\log(n))^2$$

and such that

$$\left|N(x)^p - \sum_{i=1}^m w_i N_i(x)^p\right| \leqslant \varepsilon N(x)^p\,, \quad \forall x \in \mathbb{R}^n\,.$$

Proof. First, let us scale so that

$$\mathcal{N}(x) \leqslant N(x) \leqslant K\mathcal{N}(x)\,, \qquad \forall x \in \mathbb{R}^n\,.$$

Let μ denote the probability measure with $d\mu(x) \propto e^{-\mathcal{N}(x)^2}$, and define, for $i = 1, \ldots, m$,

$$\tau_i := \int N_i(x)^p \, d\mu(x)$$
$$\rho_i := \frac{\tau_i}{\|\tau\|_1}\,.$$

Note that Lemma 3.16 gives

$$\|\tau\|_1 \leqslant K^p \left(\frac{n+p}{2}\right)^{p/2}. \tag{3.17}$$

Given $M \geqslant 1$ and $\nu \in [m]^M$, define the norms $\mathcal{N}_1, \ldots, \mathcal{N}_M$ by

$$\mathcal{N}_j(x)^p := \frac{N_{\nu_j}(x)^p}{M\rho_{\nu_j}}\,,$$

and denote $\varphi_i(x) := N_i(x)^p$ for $i = 1, \ldots, m$ so that $d_{\rho,\nu}(x,y) = d(x,y)$, where $d_{\rho,\nu}$ is defined in (2.7), and d is from Definition 3.3.

We will apply Lemma 2.6 with $F := N$ and $\{\rho_i\}, \{\varphi_i\}$ defined as above. To do so, we require a bound on $\gamma_2(B_N, d_{\rho,\nu}) \leqslant \gamma_2(B_{\mathcal{N}}, d_{\rho,\nu})$. Lemma 3.13 yields

$$\gamma_2(B_{\mathcal{N}}, d_{\rho,\nu}) = \gamma_2(B_{\mathcal{N}}, d) \lesssim p\left(\mathcal{S}\mathsf{W}_{\max}\psi_n \log(M)\log(n)\right)\sqrt{\tilde{\Lambda}}\,. \tag{3.18}$$

Observe that

$$\mathsf{W}_{\max} = \max_{j\in[M]} \int \mathcal{N}_j(x)\, d\mu(x) \leqslant \max_{j\in[M]} \left(\int \mathcal{N}_j(x)^p \, d\mu(x)\right)^{1/p} \leqslant \left(\frac{\|\tau\|_1}{M}\right)^{1/p}. \tag{3.19}$$

Thus from Lemma 3.4 and monotonicity of pth moments, we see that for $j = 1, \ldots, M$,

$$\max_{x\in B_{\mathcal{N}}} \mathcal{N}_j(x)^p \lesssim \mathcal{S}^p \frac{\|\tau\|_1}{M}\,.$$

So for $x \in B_{\mathcal{N}}$, we can write

$$\sum_{j=1}^{M} \mathcal{N}_j(x)^{2(p-1)} \lesssim \left(\mathcal{S}^p \frac{\|\tau\|_1}{M}\right)^{(p-2)/p} \sum_{j=1}^{M} \mathcal{N}_j(x)^p \,.$$

Recalling the definitions of Λ (Definition 3.3) and $\tilde{\Lambda}$ (3.8), this gives

$$\sqrt{\tilde{\Lambda}} \leqslant \left(\mathcal{S}^p \frac{\|\tau\|_1}{M}\right)^{1/2-1/p} \sqrt{\Lambda}\,.$$

Note also that

$$\Lambda = \max_{x \in B_{\mathcal{N}}} \sum_{j=1}^{M} \mathcal{N}_j(x)^p \leqslant K^p \max_{x \in B_N} \sum_{j=1}^{M} \mathcal{N}_j(x)^p = K^p \max_{x \in B_N} \tilde{F}_{\rho,\nu}(x)\,.$$

Thus in conjunction with (3.17), (3.18), and (3.19), we have

$$\gamma_2(B_{\mathcal{N}}, d_{\rho,\nu}) \lesssim p \left(K^{2p} \mathcal{S}^p M^{-1} \left(\frac{n+p}{2}\right)^{p/2}\right)^{1/2} \left(\psi_n \log(M) \log(n)\right) \left(\max_{x \in B_N} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}.$$

So for every $\varepsilon \in (0,1)$, there is a choice of

$$M \lesssim \frac{K^{2p} \mathcal{S}^p p^2}{\varepsilon^2} \left(\frac{n+p}{2}\right)^{p/2} \left(\psi_n \log(n/\varepsilon) \log(n)\right)^2$$

such that

$$\gamma_2(B_{\mathcal{N}}, d_{\rho,\nu}) \lesssim \varepsilon \left(\max_{x \in B_N} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}.$$

Application of Lemma 2.6 gives

$$\mathop{\mathbb{E}}_{\boldsymbol{\nu}} \max_{x \in B_{\mathcal{N}}} \left| N(x)^p - \frac{1}{M} \sum_{j=1}^{M} \frac{N_{\boldsymbol{\nu}_j}(x)^p}{\rho_{\boldsymbol{\nu}_j}} \right| \lesssim \varepsilon\,. \qquad \square$$

3.4 Algorithms

We first present an efficient algorithm for sampling in the case $p = 1$. Consider norms $N_1, \ldots, N_m$ on $\mathbb{R}^n$ and suppose that each N_i can be evaluted in time $\mathcal{T}_{\text{eval}}$, and that $N(x) := N_1(x) + \cdots + N_m(x)$ is (r,R)-rounded for $0 < r \leqslant R$.

Theorem 3.18 (Efficient Recursive Sparsification)**.** *If N is (r,R)-rounded, then there is an algorithm running in time $(m(\log n)^{O(1)} + n^{O(1)})\mathcal{T}_{\text{eval}} \log(mR/r)$ that with high probability produces an $O(n\varepsilon^{-2} \log(n/\varepsilon)(\log n)^{2.5})$-sparse ε-approximation to N.*

Suppose now that $\tilde{N}$ is a norm on $\mathbb{R}^n$ that is K-equivalent to N, and let μ be the probability measure with density $\propto e^{-\tilde{N}(x)}$.

Lemma 3.19 (Sampling to sparsification)**.** *For $h \geqslant 1$, there is an algorithm that, given $O(h\psi_n \log(m+n))$ independent samples from μ and $\varepsilon > 0$, computes with probability at least $1-(m+n)^{-h}$, an s-sparse ε-approximation to N in time $O(m\psi_n \log(n+m)+s)\mathcal{T}_{\mathrm{eval}}$, where $s \leqslant O(K^2\varepsilon^{-2}n\varepsilon^{-2}\log(n/\varepsilon)(\log n)^{2.5})$.*

Proof. Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k \in \mathbb{R}^n$ be independent samples from μ. Denote, for $i=1,\ldots,m$,

$$\begin{aligned}\tau_i &:= \frac{3}{2}\frac{1}{k}\left(N_i(\boldsymbol{X}_1)+N_i(\boldsymbol{X}_2)+\cdots+N_i(\boldsymbol{X}_k)\right)\\ \sigma_i &:= \mathbb{E}[N_i(\boldsymbol{X}_1)]\,.\end{aligned}$$

Since μ is log-concave, Corollary 1.8 asserts there is a constant $c>0$ such that

$$\mathbb{P}\left(\left|N_i(x_j)-\sigma_i\right|>t\right) \leqslant 2\exp\left(-\frac{ct}{\psi_n\sigma_i}\right)$$

Consequently, for some $k \lesssim h\psi_n \log(m+n)$, it holds that

$$\mathbb{P}\left(\sigma_i \leqslant \tau_i \leqslant 2\sigma_i, i=1,\ldots,m\right) \geqslant 1-(m+n)^{-h}\,.$$

Thus with high probability, (3.13) it satisfied for $p=1$, and one obtains the desired sparse approximation using Theorem 3.15 with $p=1$. □

The preceding lemma shows that sampling from a distribution with $d\mu(x) \propto e^{-\tilde{N}(x)}$ suffices to efficiently sparsify a norm N that is K-equivalent to $\tilde{N}$. Also, there is a long line of work on sampling from logconcave distributions which shows that we can sample from a distribution close to μ in $n^{O(1)}$ evaluations of $\tilde{N}$.

Theorem 3.20 ([LV07, Theorem 2.1])**.** *There is an algorithm that produces a sample from a log-concave distribution μ with density $\propto e^{-f(x)}$ within total variation distance at most ε in $\tilde{O}(n^5)$ evaluations of $f(x)$ and $\tilde{O}(n^5)$ additional time.*

Combining Lemma 3.19 and Theorem 3.20, we see that if one can sample from the distribution induced by a sparsifier, then one can efficiently sparsify and if one can efficiently sparsify, then one can can perform the requisite sampling.

This chicken-and-egg problem has arisen for a variety of sparsification problems and there is a relatively simple and standard solution introduced in [MP12] that has been used in a range of settings; see, , e.g., [KLM+17, JSS18, AJSS19]).

Instead of simply sampling proportional to $e^{-N(x)}$ directly, we first sample proportional to the density $\exp(-(N(x)+t\|x\|_2))$, where t is chosen large enough that the sampling problem is trivial. This gives a sparsifier for $N(x)+t\|x\|_2$ which, in turn, can be used to efficiently sparsify $N(x)+t/2\|x\|_2$. Iterating allows us to establish Theorem 3.18.

Proof of Theorem 3.18. Recall our assumption that $r\|x\|_2 \leqslant N(x) \leqslant R\|x\|_2$ for all $x \in \mathbb{R}^n$. For $t \geqslant 0$, denote $N_t(x) := N(x)+t\|x\|_2$. Note that N_R is 2-equivalent to $R\|x\|_2$, and consequently we can use Lemma 3.19 to obtain an $\tilde{O}(n)$-sparse 1/2-approximation to N_R.

Now for any $t \in [\varepsilon r, R]$, suppose $\tilde{N}_t$ is an $\tilde{O}(n)$-sparse 1/2-approximation to N_t. Then using Theorem 3.20, we can compute a sample from the distribution with density $\propto e^{-\tilde{N}_t(x)}$ in time

$(n \log(R/r))^{O(1)}\mathcal{T}_{\text{eval}}$. We can ignore the total variation error in Theorem 3.20 as long as it is less than $m^{-O(1)}$, and charge it to the failure probability. Since $N_{t/2}$ is 2-equivalent to N_t, which is 2-equivalent to $\tilde{N}_t$, we can use Lemma 3.19 to obtain an $\tilde{O}(n)$-sparse $1/2$-approximation to $N_{t/2}$.

After $O(\log(R/r))$ iterations, one obtains an $\tilde{O}(n)$-sparse $1/2$-approximation to N_r, which is itself 2-equivalent to N. A final application of Lemma 3.19 obtains an $O(n\varepsilon^{-2}\log(n/\varepsilon)(\log n)^{2.5})$-sparse ε-approximation to N in the desired running time. □

Remark 3.21 (Algorithm for $1 < p \leqslant 2$)**.** We note that it is possible to extend Theorem 3.18 to the setting of $1 < p \leqslant 2$ under a mild additional assumption. Specifically, we need to assume that each norm N_i is itself K-equivalent to a p-uniformly smooth norm $\mathcal{N}_i$ with constant $\mathcal{S}_p$, and that we have oracle access to $\mathcal{N}_i$.

For any weights $w_1, \ldots, w_m \geqslant 0$, the norm $N_w(x) := (w_1 N_1(x)^p + \cdots + w_m N_m(x)^p)^{1/p}$ is then K-equivalent to the norm $\mathcal{N}_w(x) := (w_1\mathcal{N}_1(x)^p + \cdots + w_m\mathcal{N}_m(x)^p)^{1/p}$, where each $\mathcal{N}_i$ is p-uniformly smooth with constant $\mathcal{S}_p$. Since the ℓ_p sum of p-uniformly smooth norms is also p-uniformly smooth quantiatively (see [Fig76]), it holds that N_w is K-equivalent to a norm $\mathcal{N}_w$ that is p-uniformly smooth with constant $O(\mathcal{S}_p)$. One can then proceed along similar lines using the interpolants

$$N_t(x) := \left(N(x)^p + t\|x\|_2^p\right)^{1/p},$$

which are similarly K-equivalent to the p-uniformly smooth norm $\mathcal{N}_t(x) = \left(\mathcal{N}(x)^p + t\|x\|_2^p\right)^{1/p}$, since $\|\cdot\|_2$ is p-uniformly smooth with constant 1 for any $1 \leqslant p \leqslant 2$.

3.4.1 Sparsifying symmetric submodular functions

First recall that the Lovász extension $\bar{F}$ is a semi-norm. This follows because $\bar{F}$ can be expressed as

$$\bar{F}(x) = \int_{-\infty}^{\infty} F(\{i : x_i \leqslant t\})\, dt.$$

Note that the integral is finite because $F(\emptyset) = F(V) = 0$, and clearly $\bar{F}(cx) = c\bar{F}(x)$ for all $c > 0$. Also because F is symmetric we have $F(x) = \int_{-\infty}^{\infty} F(\{i : x_i \leqslant t\})\, dt = \int_{-\infty}^{\infty} F(\{i : x_i \geqslant t\})\, dt = F(-x)$. Finally, it is a standard fact that F is submodular if and only if $\bar{F}$ is convex. Thus, $\bar{F}$ is indeed a semi-norm.

Proof of Corollary 1.2. We assume that $\varepsilon \geqslant m^{-1/2}$, else the desired sparsity bound is trivial.

Let $\bar{f}_1, \ldots, \bar{f}_m$ denote the respective Lovász extensions of $f_1, \ldots, f_m$, and let $\bar{F}$ denote the Lovász extension of F. Define $\tilde{F}(x) := \bar{F}(x) + m^{-4}\|x\|_2$ and $\tilde{f}_i(x) := \bar{f}_i(x) + m^{-5}\|x\|_2$ so that $\tilde{F}(x) = \tilde{f}_1(x) + \cdots + \tilde{f}_m(x)$. Clearly each $\tilde{f}_i$ is $(m^{-5}, O(nR))$-rounded as $\tilde{f}_i(x) \leqslant 2\|x\|_\infty R \leqslant 2R\sqrt{n}\|x\|_2$. Thus Theorem 3.18 yields weights $w \in \mathbb{R}_+^m$ with the asserted sparsity bound and such that

$$\left|\tilde{F}(x) - \sum_{i=1}^{m} w_i \tilde{f}_i(x)\right| \leqslant \varepsilon\tilde{F}(x), \quad \forall x \in \mathbb{R}^n.$$

Additionally, the unbiased sampling scheme of Section 2.2 guarantees that $\mathbb{E}[w_1 + \cdots + w_m] = m$, so $\sum_{i=1}^m w_i \leqslant 2m$ with probability at least 1/2. Assuming this holds, let us argue that $\left|F(S) - \sum_{i\in[m]} w_i f_i(S)\right| \leqslant 2\varepsilon F(S)$ for all $S \subseteq V$. Indeed,

$$\left|F(S) - \sum_{i=1}^m w_i f_i(S)\right| \leqslant \varepsilon \tilde{F}(S) + \left(m + \sum_{i=1}^m w_i\right) m^{-5}\|x\|_2 \leqslant \varepsilon F(S) + m^{-3}\,.$$

This is at most $2\varepsilon F(S)$ if $F(S) \geqslant 1$, since we assumed that $\varepsilon \geqslant m^{-1/2}$.

If, on the other hand, $F(S) = 0$, then we conclude that all $f_i(S) = 0$ for all $i \in \mathrm{supp}(w)$. This is because the weights given by the independent sampling procedure (recall Section 2.2) are at least $1/M \geqslant 1/m$, and each function f_i is integer-valued. Thus $w_1 f_1(S) + \cdots + w_m f_m(S) = 0$ as well. □

4 Lewis weights

When working with a subspace of $L_p(\mu)$, it is often useful to perform a "change of density" in order to compare the L_p norm to some other norm (in the present setting, to an L_2 norm); see, e.g., [JS01]. A classical paper of Lewis [Lew78] describes a very useful change of density that has applications to sparsification problems like ℓ_p row sampling [CP15] and dimension reduction for linear subspaces of ℓ_p [BLM89, Tal95].

Let α be a norm on the space of linear operators $\mathbb{R}^n \to \mathbb{R}^n$. Following Lewis [Lew79], one can consider the corresponding optimization

$$\max\left\{\,|\det(U)| : \alpha(U) \leqslant 1\right\}. \tag{4.1}$$

As one example, suppose that $K \subseteq \mathbb{R}^n$ is a symmetric convex body and $\|x\|_K$ is the associated norm. Define the operator norm

$$\alpha(U) := \max_{\|x\|_2=1} \|Ux\|_K.$$

For an optimizer $U^\star$, one can check that $U^\star(B_2^n)$ is the John ellipsoid of K, i.e., a maximum volume ellipsoid such that $U^\star(B_2^n) \subseteq K$.

ℓ_p Lewis weights for a matrix A. Consider now a linear operator $A : \mathbb{R}^n \to \mathbb{R}^k$, and let $a_1 = A^\top e_1, \ldots, a_k = A^\top e_k$ be the rows of A. Denote $K := \mathrm{conv}(\pm a_1, \ldots, \pm a_k)$, let $\|\cdot\|_K^*$ be the dual norm, and define

$$\alpha(U) := \max_{\|x\|_2=1} \|U^\top x\|_K^* = \max_{\|x\|_2=1} \sup_{y\in K} \langle y, U^\top x\rangle = \max_{i\in[k]} \|Ua_i\|_2.$$

Generalizing further, define for any $1 \leqslant p < \infty$, the norm

$$\alpha(U) := \left(\sum_{i=1}^k \|Ua_i\|_2^p\right)^{1/p},$$

for which the preceding definition corresponds to the case $p = \infty$. We remark that this α coincides with the absolutely p-summing operator norm when U is considered as an operator $U : E \to F$

with $F := A(\mathbb{R}^n) \subseteq \ell_p^k$, and where E is $\mathbb{R}^n$ equipped with the Euclidean norm $x \mapsto \|(A^\top A)^{-1/2}x\|_2$. See, e.g., [DJT95] for background on absolutely summing operators.

For $1 \leqslant p \leqslant 2$, the optimality condition for (4.1) yields the existence of a nonnegative diagonal matrix W such that $\alpha((A^\top WA)^{-1/2}) \leqslant n^{1/p}$, and

$$\max_{i\in[k]} \frac{|\langle a_i, x\rangle|}{\|(A^\top WA)^{-1/2}a_i\|_2} \leqslant \|(A^\top WA)^{1/2}x\|_2 \leqslant \|Ax\|_p\,.$$

In particular, this bounds the contribution of every coordinate to the ℓ_p norm: Denoting $\alpha_i := \|(A^\top WA)^{-1/2}a_i\|_2$, we have

$$|\langle a_i, x\rangle| \leqslant \alpha_i\|Ax\|_p\,, \quad \forall x \in \mathbb{R}^n\,,$$

and $\alpha_1^p + \cdots + \alpha_k^p = \alpha((A^\top WA)^{-1/2})^p = n$.

Block ℓ_∞ weights. In order to construct spectral sparsifiers for hypergraphs, the authors of [KKTY21a] implicitly consider the following setting, couched in the language of effective resistances on graphs. Suppose $S_1 \cup \cdots \cup S_m = [k]$ forms a partition of the index set, and define

$$\alpha(U) := \left(\sum_{j=1}^m \max_{i\in S_j} \|Ua_i\|_2^2\right)^{1/2}.$$

Their construction was clarified and extended in [Lee22, JLS22]. We now present a substantial generalization that will be a useful tool in proving Theorem 1.5 and Theorem 1.6.

Definition 4.1 (Block norm)**.** Consider any $p_1, \ldots, p_m, q \in [1, \infty]$, and a partition $S_1 \cup \cdots \cup S_m = [k]$. For $p_j < \infty$, define

$$\mathcal{N}_j(u) := \left(\sum_{i\in S_j} |u_i|^{p_j}\right)^{1/p_j},$$

and for $p_j = \infty$, take $\mathcal{N}_j(u) := \max\{|u_i| : i \in S_j\}$. Define $\mathcal{N}(u) := \|(\mathcal{N}_1(u), \ldots, \mathcal{N}_m(u))\|_q$.

Throughout, we let $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^k)$ denotes the linear space of linear operators from $\mathbb{R}^n$ to $\mathbb{R}^k$, and denote $\mathcal{L}(\mathbb{R}^n) = \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. We prove the next lemma in Section 4.1.

Lemma 4.2. *Consider $p_1, \ldots, p_m \in [2, \infty]$ and $q \in [1, \infty)$. Let $\mathcal{N}_1, \ldots, \mathcal{N}_m$ and $\mathcal{N}$ be as in Definition 4.1. Fix $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^k)$ with $\operatorname{rank}(A) = n$, and denote for $j = 1, \ldots, m$,*

$$\alpha_j(U) := \mathcal{N}_j\big(\|UA^\top e_1\|_2, \ldots, \|UA^\top e_k\|_2\big)$$

Then there is a nonnegative diagonal matrix W such that for $U = (A^\top WA)^{-1/2}$, the following are true:

(1) It holds that

$$\alpha_1(U)^q + \cdots + \alpha_m(U)^q = \begin{cases} n & 1 \leqslant q \leqslant 2 \\ n^{q/2} & q \geqslant 2\,. \end{cases}$$

(2) For all $x \in \mathbb{R}^n$,

$$\mathcal{N}_j(Ax) \leqslant \alpha_j(U)\|U^{-1}x\|_2 \leqslant \alpha_j(U)\mathcal{N}(Ax)\,.$$

4.1 Block Lewis weights

For a norm $\mathcal{N}$ on $\mathbb{R}^k$ and $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^k)$, define the norm $\alpha_{\mathcal{N},A}$ on $\mathcal{L}(\mathbb{R}^n)$ by

$$\alpha_{\mathcal{N},A}(U) := \mathcal{N}\left(\|UA^\top e_1\|_2, \dots, \|UA^\top e_k\|_2\right) . \tag{4.2}$$

We consider the optimization (4.1). As observed in [SZ01], the analysis of (4.1) does not rely on duality in a fundamental way.

Lemma 4.3. *If $\mathcal{N}$ is continuously differentiable, then there is an invertible, self-adjoint $U \in \mathcal{L}(\mathbb{R}^n)$ such that $\alpha_{\mathcal{N},A}(U) = 1$, and*

$$U = (A^\top W A)^{-1/2} ,$$

where W is the diagonal matrix with

$$W_{ii} = \gamma \frac{\partial_{x_i} \mathcal{N}(\|UA^\top e_1\|_2, \dots, \|UA^\top e_k\|_2)}{\|UA^\top e_i\|_2} , \quad i = 1, \dots, k , \tag{4.3}$$

and

$$\gamma = n \left(\sum_{i=1}^{k} \|UA^\top e_i\|_2 \, \partial_{x_i} \mathcal{N}\left(\|UA^\top e_1\|_2, \dots, \|UA^\top e_k\|_2\right) \right)^{-1} . \tag{4.4}$$

Proof. Define the operator $B \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^k)$ by

$$B_{uv} := \partial_{U_{uv}} \alpha(U) = \sum_{i=1}^{k} \frac{\partial_{x_i} \mathcal{N}\left(\|UA^\top e_1\|_2, \dots, \|UA^\top e_k\|_2\right)}{\|UA^\top e_i\|_2} (UA^\top)_{ui} A_{iv} .$$

Note that

$$B = UA^\top (\gamma^{-1} W) A ,$$

where W is the diagonal matrix defined in (4.3).

Moreover, it holds that

$$\partial_{U_{uv}} \det(U) = \det(U) (U^{-1})_{vu} .$$

Hence if U is an optimal solution to (4.1) with $\alpha = \alpha_{\mathcal{N},A}$, then

$$UA^T W A = \gamma U^{-\top}$$

where $\gamma > 0$ is the Lagrange multiplier corresponding to the constraint $\alpha(U) \leqslant 1$. Replacing U by $(U^\top U)^{1/2}$, we may assume that U is self-adjoint.

To verify (4.4), use $U^{-2} = A^T W A = \sum_{i=1}^k W_{ii} A^\top e_i e_i^\top A$ to write

$$n = \operatorname{tr}(U^2 U^{-2}) = \operatorname{tr}(U^2 A^T W A) = \sum_{i=1}^{k} W_{ii} \operatorname{tr}(U^2 A^\top e_i e_i^\top A) = \sum_{i=1}^{k} W_{ii} \|UA^\top e_i\|_2^2 . \qquad \square$$

Lemma 4.4. *Suppose that $\mathcal{N}$ and $\mathcal{N}_1, \ldots, \mathcal{N}_m$ are as in Definition 4.1 with $p_j \in [2, \infty]$ for $j = 1, \ldots, m$, and $q \in [1, \infty)$. Then for any $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^k)$, there is a diagonal matrix W such that $U = (A^\top W A)^{-1/2}$ satisfies $\alpha_{\mathcal{N},A}(U)^q \leqslant 1$, and for $j \in [m]$ and $i \in S_j$,*

$$W_{ii} = \begin{cases} n u_i^{p_j - 2} \mathcal{N}_j(u)^{q - p_j} & p_j < \infty \\ n v_i \mathcal{N}_j(u)^{q-1} & p_j = \infty\,, \end{cases} \tag{4.5}$$

for some $v \in \mathbb{R}^k$ such that $\mathcal{N}_j^*(v) = \mathcal{N}_j(u)^{-1}$ when $p_j = \infty$.

Proof. Let $U \in \mathcal{L}(\mathbb{R}^n)$ be the map guaranteed by Lemma 4.3 applied with the $\mathcal{N}$ and A. (By a simple approximation argument, we may assume that $\mathcal{N}$ is C^1 for the cases where $q = 1$ or $p_j = \infty$.)

Note that for $j \in [m]$ and $i \in S_j$, if $p_j < \infty$, then

$$\partial_{u_i} \mathcal{N}(u_1, \ldots, u_k) = \operatorname{sign}(u_i) |u_i|^{p_j - 1} \mathcal{N}_j(u)^{q - p_j} \mathcal{N}(u)^{1-q}\,,$$

and otherwise, for $p_j = \infty$, we must consider the collection of subgradients $v \in \mathbb{R}^k$ with $\mathcal{N}_j^*(v) = \mathcal{N}_j(u)^{-1}$.

Defining $u_i := \|UA^\top e_i\|_2$ for $i = 1, \ldots, k$, this gives, for $j \in [m]$ and $i \in S_j$,

$$W_{ii} = \begin{cases} \gamma u_i^{p_j - 2} \mathcal{N}_j(u)^{q - p_j} & p_j < \infty \\ \gamma v_i \mathcal{N}_j(u)^{q-1} & p_j = \infty\,, \end{cases}$$

where we have used the fact that $\mathcal{N}(u) = \alpha_{\mathcal{N},A}(U) = 1$.

Now compute

$$\sum_{i=1}^{k} \|UA^\top e_i\|_2 \, \partial_{x_i} \mathcal{N}\big(\|UA^\top e_1\|_2, \ldots, \|UA^\top e_k\|_2\big) = \sum_{j=1}^{m} \mathcal{N}_j(u)^{q - p_j} \mathcal{N}_j(u)^{p_j} = 1\,.$$

From (4.4) we conclude that $\gamma = n$. □

Let us now use this to prove Lemma 4.2.

Proof of Lemma 4.2. Let $U = (A^\top W A)^{-1/2} \in \mathcal{L}(\mathbb{R}^n)$ be the operator guaranteed by Lemma 4.4 applied with norm $\mathcal{N}$ and $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^k)$. Denote $a_i := A^\top e_i$ for $i = 1, \ldots, k$ and the vector $u \in \mathbb{R}^k$ by $u_i = \|Ua_i\|_2$ for $i = 1, \ldots, k$. Recall that $\alpha_{\mathcal{N},A}(U)^q = \alpha_1(U)^q + \cdots + \alpha_m(U)^q = 1$.

Let us first perform the transformation $W \mapsto n^{-2/q} W$ and $U \mapsto n^{1/q} U$ so that $\alpha_1(U)^q + \cdots + \alpha_m(U)^q = n$ and, consulting (4.5), define, for $j \in [m]$ and $i \in S_j$,

$$w_i := W_{ii} = \begin{cases} u_i^{p_j - 2} \mathcal{N}_j(u)^{q - p_j} & p_j < \infty \\ v_i \mathcal{N}_j(u)^{q-1} & p_j = \infty\,, \end{cases}$$

where $v \in \mathbb{R}^k$ is such that $\mathcal{N}_j^*(v) = \mathcal{N}_j(u)^{-1}$ when $p_j = \infty$.

Now use $|\langle a_i, x\rangle| \leqslant \|Ua_i\|_2 \|U^{-1}x\|_2$ to bound, for $p_j < \infty$,

$$\mathcal{N}_j(Ax) = \left(\sum_{i \in S_j} |\langle a_i, x\rangle|^{p_j} \right)^{1/p_j} \leqslant \|U^{-1}x\|_2 \left(\sum_{i \in S_j} \|Ua_i\|_2^{p_j} \right)^{1/p_j}$$

$$= \|U^{-1}x\|_2\, \alpha_j(U) = \|U^{-1}x\|_2\, \mathcal{N}_j(u)\,, \tag{4.6}$$

verifying the first inequality in (2). Clearly the same argument applies for $p_j = \infty$ as well.

Let $w_{S_j} \in \mathbb{R}^{S_j}$ denote the vector with $(w_{S_j})_i = w_i$. We will abuse notation slightly and let $\|w_{S_j}\|_{p_j/(p_j-2)}$ denote $\|w_{S_j}\|_\infty$ when $p_j = 2$ and denote $\|w_{S_j}\|_1$ when $p_j = \infty$. For $2 < p_j < \infty$, we have

$$\|w_{S_j}\|_{p_j/(p_j-2)} = \left(\sum_{i \in S_j} u_i^{p_j}\right)^{(p_j-2)/p_j} \mathcal{N}_j(u)^{q-p_j} = \mathcal{N}_j(u)^{p_j-2}\mathcal{N}_j(u)^{q-p_j} = \mathcal{N}_j(u)^{q-2}\,. \tag{4.7}$$

This remains true in the other cases as well: When $p_j = 2$, $\|w_{S_j}\|_\infty = \mathcal{N}_j(u)^{q-2}$. When $p_j = \infty$, $\|w_{S_j}\|_1 = \mathcal{N}_j^*(v)\mathcal{N}_j(u)^{q-1} = \mathcal{N}_j(u)^{q-2}$.

Note that $\|U^{-\top}x\|_2^2 = \langle x, (U^\top U)^{-1}x\rangle$ and $(U^\top U)^{-1} = A^\top W A$. Using this and Hölder's inequality with exponents $p_j/(p_j-2)$ and $p_j/2$, write

$$\|U^{-\top}x\|_2^2 = \sum_{i=1}^k w_i |\langle a_i, x\rangle|^2 \leqslant \sum_{j=1}^m \|w_{S_j}\|_{p_j/(p_j-2)} \left(\sum_{i\in S_j} |\langle a_i, x\rangle|^{p_j}\right)^{2/p_j} = \sum_{j=1}^m \mathcal{N}_j(u)^{q-2}\mathcal{N}_j(Ax)^2\,, \tag{4.8}$$

where the last equality uses (4.7).

Case I: $1 \leqslant q \leqslant 2$.

To verify (1), note that

$$\sum_{j=1}^m \alpha_j(U)^q = \mathcal{N}(u)^q = n\,.$$

Now observe that (4.6) gives

$$\mathcal{N}_j(u)^{q-2}\mathcal{N}_j(Ax)^2 \leqslant \mathcal{N}_j(Ax)^q \|U^{-1}x\|_2^{2-q}\,.$$

In conjunction with (4.8), this yields

$$\|U^{-1}x\|_2^2 \leqslant \|U^{-1}x\|_2^{2-q} \sum_{j=1}^m \mathcal{N}_j(Ax)^q\,,$$

and simplifying yields

$$\|U^{-1}x\|_2 \leqslant \mathcal{N}(Ax)\,,$$

verifying the second inequality in (2).

Case II: $q > 2$.

Use Hölder's inequality with exponents $q/(q-2)$ and $q/2$ in (4.8) to bound

$$\|U^{-1}x\|_2^2 \leqslant \left(\sum_{j=1}^m \mathcal{N}_j(u)^q\right)^{(q-2)/q} \left(\sum_{j=1}^m \mathcal{N}_j(Ax)^q\right)^{2/q} \leqslant n^{1-2/q}\mathcal{N}(Ax)^2\,.$$

Now replacing U by $n^{1/2-1/q}U$ gives $\|U^{-1}x\|_2 \leqslant \mathcal{N}(Ax)$ and $\alpha_1(U)^q + \cdots + \alpha_m(U)^q = n^{q/2}$. □

4.1.1 Arbitrary norms

The $p_1 = \cdots = p_m = \infty$ case of Lemma 4.2 gives a consequence for *any* collection $N_1, \ldots, N_m$ of norms on $\mathbb{R}^n$ by embedding them into a subspace of ℓ_∞.

Lemma 4.5. *For any $1 \leqslant q < \infty$, there are numbers $\alpha_1, \ldots, \alpha_m \geqslant 0$ and a linear transformation $U : \mathbb{R}^n \to \mathbb{R}^n$ such that*

$$\alpha_1^q + \cdots + \alpha_m^q \leqslant \begin{cases} n & 1 \leqslant q \leqslant 2 \\ n^{q/2} & q \geqslant 2, \end{cases}$$

and for every $j \in [m]$ and $x \in \mathbb{R}^n$,

$$N_j(Ux) \leqslant \alpha_j \|x\|_2 \leqslant \alpha_j \left(\sum_{j=1}^m N_j(Ux)^q \right)^{1/q}.$$

Proof. Let N_j^* denote the dual norm to N_j. Then for every $\delta > 0$ there is a finite set $V_\delta \subseteq B_{N_j^*}$ such that

$$N_j(x) = \sup_{N_j^*(y) \leqslant 1} \langle x, y \rangle \geqslant (1-\delta) \max_{y \in V_\delta} \langle x, y \rangle = (1-\delta) \|A_j x\|_\infty$$

for some map $A_j : \mathbb{R}^n \to \mathbb{R}^{|V_\delta|}$.

Let $U_\delta : \mathbb{R}^n \to \mathbb{R}^n$ be the transformation guaranteed by Lemma 4.2 applied to the operator defined by $A(x) = A_1(x) \oplus \cdots \oplus A_m(x)$. This yields weights $\alpha_{1,\delta}, \ldots, \alpha_{m,\delta} \geqslant 0$ such that

$$\sum_{j=1}^m \alpha_{j,\delta}^q = \begin{cases} n & 1 \leqslant q \leqslant 2 \\ n^{q/2} & q \geqslant 2, \end{cases}$$

and for all $x \in \mathbb{R}^n$,

$$N_j(U_\delta x) \leqslant \alpha_{j,\delta} \|x\|_2 \leqslant (1-\delta)^{-1} \alpha_{j,\delta} N(U_\delta x),$$

where $N(x) := \left(\sum_{j=1}^m N_j(x)^q \right)^{1/q}$.

It holds that $N(U_\delta x) \leqslant \sqrt{n} \|x\|_2$ for all $x \in \mathbb{R}^n$, and therefore as long as N is a genuine norm on $\mathbb{R}^n$ (and not simply a semi-norm), the family $\{U_\delta : \delta > 0\}$ is contained in a compact subset of $\mathcal{L}(\mathbb{R}^n)$. In the case that N is not a genuine norm, one can simply restrict U_δ to the orthogonal complement of the subspace $\{x \in \mathbb{R}^n : N(x) = 0\}$ and argue there.

Define $s_\delta := (U_\delta, \alpha_{1,\delta}, \ldots, \alpha_{m,\delta})$, and then the preceding bounds similarly show that $\{s_\delta : \delta > 0\}$ lies in a compact subset of $\mathcal{L}(\mathbb{R}^n) \times \mathbb{R}^m$, and therefore contains an accumulation point $(U, \alpha_1, \ldots, \alpha_m)$ satisfying the desired conclusion. □

4.2 Sums of powers of general norms

Our goal now is to prove Theorem 1.6. For a general norm $N(x)$ on $\mathbb{R}^n$, directly applying Theorem 1.3 with the parameters $p = 2$, $K = \sqrt{n}$, $S = 1$ leads to the unimpressive bound of $O(\varepsilon^{-2} n^2 \log^5(n/\varepsilon))$. Towards improving this, we start by obtaining a (dimension-dependent) improvement on the shift lemma (Lemma 3.1).

Lemma 4.6. *Suppose N is a norm on $\mathbb{R}^n$ and $p \in [1,2]$. Let μ denote the probability measure μ on $\mathbb{R}^n$ satisfying*

$$d\mu(x) \propto \exp(-N(x)^p)\,.$$

Then for any measureable $W \subseteq \mathbb{R}^n$ and $z \in \mathbb{R}^n$, $\delta > 0$, we have

$$\mu(W+z) \geqslant e^{-\delta n/p} e^{-\left(\frac{1+\delta}{\delta}\right)^{p-1} N(z)^p} \mu(W)\,. \tag{4.9}$$

Proof. For any $\delta > 0$ and $x, z \in \mathbb{R}^n$, it holds that

$$N(x+z)^p \leqslant (N(x)+N(z))^p \leqslant (1+\delta)^{p-1} N(x)^p + \left(\tfrac{1+\delta}{\delta}\right)^{p-1} N(z)^p\,,$$

therefore

$$\begin{aligned}
\int_W \exp\left(-N(x+z)^p\right)\, dx &\geqslant e^{-\left(\frac{1+\delta}{\delta}\right)^{p-1} N(z)^p} \int_W \exp\left(-(1+\delta)^{p-1} N(x)^p\right)\, dx \\
&= (1+\delta)^{-(p-1)/p \cdot n}\, e^{-\left(\frac{1+\delta}{\delta}\right)^{p-1} N(z)^p} \int_W \exp(-N(x)^p)\, dx \\
&\geqslant e^{-\delta n/p} e^{-\left(\frac{1+\delta}{\delta}\right)^{p-1} N(z)^p} \int_W \exp(-N(x)^p)\, dx\,.
\end{aligned}$$

To complete the proof, observe that

$$\mu(W+z) = \frac{\int_W \exp\left(-N(x+z)^p\right)\, dx}{\int_W \exp\left(-N(x)^p\right)\, dx} \mu(W)\,. \qquad \square$$

Lemma 4.7. *Suppose N and $\widehat{N}$ are norms on $\mathbb{R}^n$. Denote the probability measure μ on $\mathbb{R}^n$ so that*

$$d\mu(x) \propto \exp(-N(x)^p)\,.$$

Then for any $\varepsilon > 0$,

$$\left(\log \mathcal{K}(B_N, \widehat{N}, \varepsilon)\right)^{1/2} \lesssim \left(\frac{\lambda}{\varepsilon}\right)^{p/2} + \sqrt{\frac{\lambda}{\varepsilon}} n^{\frac{1}{2}-\frac{1}{2p}}\,,$$

where

$$\lambda = \int \widehat{N}(x)\, d\mu(x)\,.$$

Proof. By scaling $\widehat{N}$, we may assume that $\varepsilon = 1$. Suppose now that $x_1, \ldots, x_M \in B_N$ and the balls $x_1 + B_{\widehat{N}}, \ldots, x_M + B_{\widehat{N}}$ are pairwise disjoint. To establish an upper bound on M, let $\lambda > 0$ be a number we will choose later and write

$$\begin{aligned}
1 \geqslant \mu\left(\bigcup_{j=1}^M \lambda(x_j + B_{\widehat{N}})\right) &= \sum_{j=1}^M \mu\left(\lambda x_j + \lambda B_{\widehat{N}}\right) \\
&\overset{(4.9)}{\geqslant} e^{-\delta n/p} \sum_{j=1}^M e^{-\lambda^p \left(\frac{1+\delta}{\delta}\right)^{p-1} N(x_j)^p} \mu(\lambda B_{\widehat{N}})
\end{aligned}$$

$$\geqslant M e^{-\delta n/p} e^{-\lambda^p \left(\frac{1+\delta}{\delta}\right)^{p-1}} \mu(\lambda B_{\widehat{N}}),$$

where the last inequality uses $x_1, \ldots, x_M \in B_N$. Now choose $\lambda := 2\int \widehat{N}(x)\, d\mu(x)$ so that Markov's inequality gives

$$\mu(\lambda B_{\widehat{N}}) = \mu\left(\{x : \widehat{N}(x) \leqslant \lambda\}\right) \geqslant 1/2,$$

yielding the upper bound

$$\log(M/2) \leqslant \frac{\delta n}{p} + \lambda^p \left(\tfrac{1+\delta}{\delta}\right)^{p-1}.$$

Choosing $\delta := \lambda/n^{1/p}$ and using $(1+1/\delta)^{p-1} \lesssim 1 + \delta^{-(p-1)}$ gives

$$\log(M/2) \lesssim \lambda n^{1-1/p} + \lambda^p. \qquad \square$$

4.2.1 Entropy estimate

We will work again in the setting of Definition 3.3.

Lemma 4.8. *Consider Definition 3.3 and additionally*

$$\kappa := \max_{x \in B_N} \mathcal{N}^\infty(x),$$
$$\lambda := \psi_n \log(M) \mathsf{W}_{\max}.$$

Then it holds that

$$\gamma_2(B_{\mathcal{N}}, d) \lesssim \left(\kappa^{\frac{p-1}{2}} \lambda^{\frac{1}{2}} n^{\frac{1}{2}-\frac{1}{2p}} + n^{-1}\kappa^{\frac{p}{2}} + \lambda^{\frac{p}{2}} \log n\right)\sqrt{\Lambda}. \tag{4.10}$$

Proof. Since both sides of (4.10) scale linearly in the values $\{\mathcal{N}_j^p\}$, we may assume that

$$\Lambda = \max_{x \in B_{\mathcal{N}}} \sum_{j=1}^M \mathcal{N}_j(x)^p = 1, \tag{4.11}$$

and then Lemma 3.8 gives

$$d(x,y) \leqslant 2\mathcal{N}^\infty(x-y)^{p/2}, \qquad x, y \in B_{\mathcal{N}}. \tag{4.12}$$

This yields the comparisons

$$\mathcal{K}(B_{\mathcal{N}}, d, r) \leqslant \mathcal{K}(B_{\mathcal{N}}, \mathcal{N}^\infty, (r/2)^{2/p}) \leqslant \mathcal{K}(B_{\mathcal{N}^\infty}, \mathcal{N}^\infty, \tfrac{(r/2)^{2/p}}{\kappa}) \leqslant \left(\frac{2\kappa}{(r/2)^{2/p}}\right)^n, \tag{4.13}$$

where the second inequality uses the definition of κ, and the final inequality follows from Lemma 2.4.

In particular, we have

$$\int_0^{\kappa^{p/2}/n^2} \sqrt{\log \mathcal{K}(B_{\mathcal{N}^\infty}, \mathcal{N}^\infty, (r/2)^{p/2})}\, dr \lesssim \sqrt{n} \int_0^{\kappa^{p/2}/n^2} \sqrt{\log \frac{\kappa}{r^{2/p}}}\, dr \lesssim n^{-3/2}\kappa^{p/2} \log n. \tag{4.14}$$

Lemma 4.7 with $N = \mathcal{N}, \widehat{N} = \mathcal{N}^\infty$ asserts that

$$(\log \mathcal{K}(B_{\mathcal{N}}, \mathcal{N}^\infty, (r/2)^{2/p})^{1/2} \lesssim \left(\frac{\lambda_0}{(r/2)^{2/p}}\right)^{p/2} + \sqrt{\frac{\lambda_0}{(r/2)^{2/p}}} n^{\frac{1}{2}-\frac{1}{2p}} \lesssim \frac{\lambda_0^{p/2}}{r} + \frac{\lambda_0^{1/2}}{r^{1/p}} n^{\frac{1}{2}-\frac{1}{2p}}, \tag{4.15}$$

where $\lambda_0 := \int \mathcal{N}^\infty(x)\, d\mu(x)$.

Dudley's entropy bound (2.5) in conjunction with (4.13) gives

$$\gamma_2(B_n, d) \lesssim \int_0^\infty \sqrt{\log \mathcal{K}(B_{\mathcal{N}}, d, r)}\, dr \lesssim \int_0^{(2\kappa)^{p/2}} \sqrt{\log \mathcal{K}(B_{\mathcal{N}}, \mathcal{N}^\infty, (r/2)^{2/p})}\, dr,$$

where we have used the fact that $\log \mathcal{K}(B_{\mathcal{N}}, \mathcal{N}^\infty, r) = 0$ for $r \geqslant \kappa$. Now using (4.14) and (4.15) yields

$$\begin{aligned}\gamma_2(B_{\mathcal{N}}, d) &\lesssim \frac{\kappa^{p/2}}{n} + \lambda_0^{p/2} \int_{\kappa^{p/2}/n^2}^{(2\kappa)^{p/2}} \frac{1}{r}\, dr + \lambda_0^{1/2} n^{\frac{1}{2}-\frac{1}{2p}} \int_{\kappa^{p/2}/n^2}^{(2\kappa)^{p/2}} \frac{1}{r^{1/p}}\, dr \\ &\lesssim \frac{\kappa^{p/2}}{n} + \lambda_0^{p/2} \log n + \lambda_0^{1/2} n^{\frac{1}{2}-\frac{1}{2p}} \kappa^{\frac{p-1}{2}}.\end{aligned}$$

To conclude, recall that Lemma 3.7 gives the estimate

$$\lambda_0 \lesssim \psi_n \log(M) \mathsf{W}_{\max} = \lambda. \qquad \square$$

4.2.2 Sparsification

We obtain our result for sparsifying sums of pth powers of arbitrary norms.

Theorem 4.9. *Suppose $1 \leqslant p \leqslant 2$ and $N_1, \ldots, N_m$ are norms on $\mathbb{R}^n$. Denote*

$$N(x) := (N_1(x)^p + \cdots + N_m(x)^p)^{1/p}.$$

Then there is a weight vector $w \in \mathbb{R}_+^m$ with

$$|\mathrm{supp}(w)| \lesssim \frac{n^{2-1/p} \psi_n \log(n/\varepsilon)}{\varepsilon^2} + \frac{n(\psi_n \log(n/\varepsilon))^p (\log n)^2}{\varepsilon^2},$$

and such that

$$\left| N(x)^p - \sum_{j=1}^m w_j N_j(x)^p \right| \leqslant \varepsilon N(x)^p, \qquad \forall x \in \mathbb{R}^n.$$

Proof of Theorem 4.9. Let $U : \mathbb{R}^n \to \mathbb{R}^n$ be the transformation guaranteed by Lemma 4.5 (with $p = q$). In particular, there are weights $\alpha_1, \ldots, \alpha_m \geqslant 0$ such that $\alpha_1^p + \cdots + \alpha_m^p \leqslant n$, and

$$N_j(x)^p \leqslant \alpha_j^p N(x)^p, \quad \forall x \in \mathbb{R}^n. \tag{4.16}$$

Let μ denote the measure with density $d\mu(x) \propto e^{-N(x)^p}$, and define for $i \in [m]$,

$$\tau_i := \int N_i(x)^p\, d\mu(x)$$

$$\rho_i := \frac{\tau_i + \alpha_i^p}{\sum_{i=1}^m (\tau_i + \alpha_i^p)}.$$

By Lemma 3.14, we have

$$\tau_1 + \cdots + \tau_m = \int N(x)^p \, d\mu(x) = \frac{n}{p}. \tag{4.17}$$

We will apply Lemma 4.8 with $\mathcal{N} = N$. Given $M \geqslant 1$ and $\nu \in [m]^M$, define

$$\begin{aligned} \mathcal{N}_j(x) &:= \frac{N_{\nu_j}(x)}{(M\rho_{\nu_j})^{1/p}}, \quad j = 1, \dots, M, \\ \varphi_i(x) &:= N_i(x)^p, \qquad i = 1, \dots, m, \end{aligned}$$

so that $d_{\rho,\nu}(x, y) = d(x, y)$, where d is the distance from (3.3). Observe that

$$\frac{N_i(x)^p}{\rho_i} \leqslant \sum_{i=1}^m (\tau_i + \alpha_i^p) \frac{N_i(x)^p}{\alpha_i^p} \overset{(4.16)}{\leqslant} 3nN(x)^p, \quad \forall x \in \mathbb{R}^n, i = 1, \dots, m,$$

and moreover

$$\int \mathcal{N}_j(x)^p \, d\mu(x) = \frac{1}{M\rho_{\nu_j}} \int N_{\nu_j}(x)^p \, d\mu(x) = \frac{\tau_{\nu_j}}{M\rho_{\nu_j}} \leqslant \frac{3n}{M}, \quad j = 1, \dots, M.$$

Therefore $\mathsf{W}_{\max} \leqslant (3n/M)^{1/p}$, and $\kappa \lesssim \left(\frac{n}{M}\right)^{1/p}$ and $\lambda \lesssim \left(\frac{n}{M}\right)^{1/p} \psi_n \log M$ in Lemma 4.8.

It follows that

$$\begin{aligned} \gamma_2(B_N, d) &\lesssim \left(\frac{n}{M}\right)^{1/2} \left(n^{\frac{1}{2} - \frac{1}{2p}} (\psi_n \log M)^{1/2} + (\psi_n \log M)^{p/2} \log n\right) \left(\max_{N(x) \leqslant 1} \sum_{j=1}^M \mathcal{N}_j(x)^p\right)^{1/2} \\ &\lesssim \varepsilon \left(\max_{N(x) \leqslant 1} \sum_{j=1}^M \mathcal{N}_j(x)^p\right)^{1/2}, \end{aligned}$$

for some choice of M satisfying

$$M \lesssim \frac{n^{2-1/p} \psi_n \log(n/\varepsilon)}{\varepsilon^2} + \frac{n(\psi_n \log(n/\varepsilon))^p (\log n)^2}{\varepsilon^2}.$$

With this, Lemma 2.6 yields

$$\mathbb{E}_{\boldsymbol{\nu}} \max_{x \in B_N} \left| N(x)^p - \frac{1}{M} \sum_{j \in [M]} \frac{N_{\boldsymbol{\nu}_j}(x)^p}{\rho_{\boldsymbol{\nu}_j}} \right| \lesssim \varepsilon.$$

Thus there exists a choice of $\nu \in [m]^M$ satisfying the bound, and the claim follows. □

4.3 Sums of squares of ℓ_p norms

Our goal now is to prove Theorem 1.5. We will require the following chaining estimate.

Theorem 4.10 ([Lee22, Lemma 2.12])**.** *Suppose $\mathcal{N}_1, \ldots, \mathcal{N}_M$ are norms on $\mathbb{R}^n$, and define*

$$d(x,y) := \left(\sum_{j=1}^{M} \left(\mathcal{N}_j(x)^2 - \mathcal{N}_j(y)^2\right)^2\right)^{1/2} \tag{4.18}$$

Then for any set $T \subseteq B_2^n$, it holds that

$$\gamma_2(T,d) \lesssim \sup_{x\in T}\left(\sum_{j=1}^{M}\mathcal{N}_j(x)^4\right)^{1/2} + \left(\kappa + \lambda\sqrt{\log n}\right)\sup_{x\in T}\left(\sum_{j=1}^{M}\mathcal{N}_j(x)^2\right)^{1/2},$$

where

$$\kappa := \mathbb{E}\max_{j\in[M]}\mathcal{N}_j(g)$$
$$\lambda := \max_{j\in[M]}\mathbb{E}\,\mathcal{N}_j(g),$$

and g is a standard n-dimensional Gaussian.

Proof of Theorem 1.5. By scaling the matrices $\{A_j\}$ suitably, we may assume that

$$\|A_j x\|_{p_j} \leqslant N_j(x) \leqslant K\|A_j x\|_{p_j}, \qquad \forall x \in \mathbb{R}^n, j = 1,\ldots,m. \tag{4.19}$$

Define $A : \mathbb{R}^n \to \mathbb{R}^k$ for $k := n_1 + \cdots + n_m$ by $A(x) = A_1(x) \oplus \cdots \oplus A_m(x)$ and let $S_1 \cup \cdots \cup S_m = [k]$ be the natural partition. Let $U \in \mathcal{L}(\mathbb{R}^n)$ be the transformation established in Lemma 4.2, and denote $\tau_j := \alpha_j(U)^2$ for $j = 1,\ldots,m$. Note that Lemma 4.2(1) ensures $\|\tau\|_1 \leqslant n$.

For $j \in [m]$, define

$$\rho_j := \tau_j/\|\tau\|_1,$$
$$\varphi_j(x) := N_j(x)^2.$$

Then using Lemma 2.6, our goal is to bound $\gamma_2(B_N, d_{\rho,\nu})$ for any $\nu \in [m]^M$.

To this end, define

$$\hat{N}_j(x) := \|A_j U x\|_{p_j}, \qquad j = 1,\ldots,m,$$
$$\hat{N}(x) := \left(\hat{N}_1(x)^2 + \cdots + \hat{N}_m(x)^2\right)^{1/2},$$

and then the second inequality in Lemma 4.2(2) implies

$$\|x\|_2 \leqslant \hat{N}(x),$$

hence $B_{\hat{N}} \subseteq B_2^n$, and $\hat{N}_j(x) \leqslant \tau_j\|x\|_2^2$ for all $x \in \mathbb{R}^n$, $j \in [m]$.

Fix $\nu \in [m]^M$ and let us finally define, for $i \in [M]$,

$$\mathcal{N}_i(x) := \frac{\hat{N}_{\nu_i}(x)}{\sqrt{\rho_{\nu_i}}}.$$

Let d be the corresponding distance given by (4.18). Then by construction, we have

$$d_{\rho,\nu}(x,y) \leqslant K^2 \frac{d(U^{-1}x, U^{-1}y)}{M}, \tag{4.20}$$

where the inequality uses (4.19).

Since $B_{\hat{N}} \subseteq B_2^n$, we can apply Theorem 4.10 with $T = B_{\hat{N}}$. Let us note first that, by Lemma 4.2(2) we have $\hat{N}_j(x)^2 \leqslant \tau_j \|x\|_2^2 \leqslant \tau_j \hat{N}(x)^2$. Therefore $\mathcal{N}_j(x)^2 \leqslant \|\tau\|_1 \hat{N}(x)^2$, and

$$\max_{\hat{N}(x)\leqslant 1} \sum_{j=1}^M \mathcal{N}_j(x)^4 \leqslant \|\tau\|_1 \max_{\hat{N}(x)\leqslant 1} \sum_{j=1}^M \mathcal{N}_j(x)^2. \tag{4.21}$$

Moreover, it holds that

$$\begin{aligned}
\mathbb{E}\, \hat{N}_j(g) \leqslant \left(\mathbb{E}\, \hat{N}_j(g)^{p_j}\right)^{1/p_j} &= \left(\sum_{i\in S_j} \mathbb{E}\, |\langle a_i, Ug\rangle|^{p_j}\right)^{1/p_j} \\
&= \left(\sum_{i\in S_j} \|Ua_i\|_2^{p_j}\, \mathbb{E}\, |g_1|^{p_j}\right)^{1/p_j} \\
&\lesssim \sqrt{p_j} \left(\sum_{i\in S_j} \|Ua_i\|_2^{p_j}\right)^{1/p_j} = \sqrt{p_j}\, \alpha_j(U) \leqslant \sqrt{p\tau_j},
\end{aligned}$$

where the first inequality uses the fact that $(\mathbb{E}[|g_1|^p])^{1/p} \lesssim \sqrt{p}$ for any $p \geqslant 1$, and the second uses $p_j \leqslant p$.

Recall that $\hat{N}_j(x)^2 \leqslant \tau_j \|x\|_2^2$, and therefore the Gaussian concentration inequality (Theorem 2.7) implies that

$$\mathbb{P}\left(\hat{N}_j(g) > \mathbb{E}\, \hat{N}_j(g) + t\sqrt{\tau_j}\right) \leqslant e^{-t^2/2}.$$

In particular, this gives

$$\mathbb{E} \max_{i\in[M]} \frac{\hat{N}_{\nu_i}(g)}{\sqrt{\rho_{\nu_i}}} \lesssim \sqrt{\|\tau\|_1 \log M}.$$

We conclude that

$$\begin{aligned}
\lambda &= \max_{j\in[M]} \mathbb{E}\, \mathcal{N}_j(g) \leqslant \sqrt{p\|\tau\|_1}, \\
\kappa &= \mathbb{E} \max_{j\in[M]} \mathcal{N}_j(g) \lesssim \sqrt{\|\tau\|_1 (p + \log M)}.
\end{aligned}$$

Combining these with (4.21) and $\|\tau\|_1 \leqslant n$, Theorem 4.10 yields

$$\gamma_2(B_{\hat{N}}, d) \lesssim \sqrt{n}\left(1 + \sqrt{p + \log M} + \sqrt{p \log n}\right)\left(\max_{\hat{N}(x)\leqslant 1} \sum_{j=1}^{M} \mathcal{N}_j(x)^2\right)^{1/2}. \tag{4.22}$$

Note that

$$\sum_{j=1}^{M} \mathcal{N}_j(x)^2 = \sum_{j=1}^{M} \frac{\hat{N}_{\nu_i}(x)^2}{\rho_{\nu_i}} \overset{(4.19)}{\leqslant} \sum_{j=1}^{M} \frac{N_{\nu_i}(U^{-1}x)^2}{\rho_{\nu_i}} = M\tilde{F}_{\rho,\nu}(U^{-1}x)\,.$$

And since (4.19) implies $N(U^{-1}x) \leqslant K\hat{N}(x)$, we have

$$\max_{\hat{N}(x)\leqslant 1} \sum_{j=1}^{M} \mathcal{N}_j(x)^2 \leqslant K^2 M \max_{N(U^{-1}x)\leqslant 1} \tilde{F}_{\rho,\nu}(U^{-1}x) \leqslant K^2 M \max_{N(x)\leqslant 1} \tilde{F}_{\rho,\nu}(x)\,.$$

Using this together with (4.22) gives

$$\gamma_2(B_{\hat{N}}, d) \lesssim KM^{1/2}\sqrt{n(p\log n + \log M)}\left(\max_{N(x)\leqslant 1} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}.$$

Combining this with (4.20), we conclude that

$$\gamma_2(B_N, d_{\rho,\nu}) \lesssim K^3 M^{-1/2}\sqrt{n(p\log n + \log M)}\left(\max_{N(x)\leqslant 1} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}.$$

Now choose $M \asymp \frac{K^6 pn \log(n/\varepsilon)}{\varepsilon^2}$ and apply Lemma 2.6 to obtain

$$\mathop{\mathbb{E}}_{\boldsymbol{\nu}} \max_{N(x)\leqslant 1} \left| N(x)^2 - \frac{1}{M}\sum_{j=1}^{M} \frac{N_{\boldsymbol{\nu}_j}(x)^2}{\rho_{\boldsymbol{\nu}_j}} \right| \leqslant \varepsilon\,,$$

completing the proof. □

Acknowledgments

James R. Lee is supported in part by NSF CCF-2007079 and a Simons Investigator Award. Yang P. Liu is supported by a Google Research Ph.D. Fellowship. Aaron Sidford is supported in part by a Microsoft Research Faculty Fellowship, NSF CCF-1844855, NSF CCF-1955039, a PayPal research award, and a Sloan Research Fellowship.

References

[AJSS19] AmirMahdi Ahmadinejad, Arun Jambulapati, Amin Saberi, and Aaron Sidford. Perron-Frobenius theory in nearly linear time: Positive eigenvectors, M-matrices, graph kernels, and other applications. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1387–1404. SIAM, 2019. 10, 26

[BCL94] Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Invent. Math.*, 115(3):463–482, 1994. 5

[BK96] András A. Benczúr and David R. Karger. Approximating s-t minimum cuts in (n^2) time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*, pages 47–55. ACM, New York, 1996. 2, 6

[BLM89] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Math.*, 162(1-2):73–141, 1989. 4, 28

[BMP63] Richard E. Barlow, Albert W. Marshall, and Frank Proschan. Properties of probability distributions with monotone hazard rate. *Ann. Math. Statist.*, 34:375–389, 1963. 14

[BSS12] Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012. 3, 5

[BSS14] Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM Rev.*, 56(2):315–334, 2014. 3

[BST19] Nikhil Bansal, Ola Svensson, and Luca Trevisan. New notions and constructions of sparsification for graphs and hypergraphs. In David Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 910–928. IEEE Computer Society, 2019. 5

[Che21] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture. *Geom. Funct. Anal.*, 31(1):34–61, 2021. 8

[CKP+17] Michael B. Cohen, Jonathan Kelner, John Peebles, Richard Peng, Anup B. Rao, Aaron Sidford, and Adrian Vladu. Almost-linear-time algorithms for Markov chains and new spectral primitives for directed graphs. In *STOC'17—Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 410–419. ACM, New York, 2017. 3

[CP15] Michael B. Cohen and Richard Peng. ℓ_p row sampling by Lewis weights. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*, pages 183–192. ACM, New York, 2015. 28

[DJT95] Joe Diestel, Hans Jarchow, and Andrew Tonge. *Absolutely summing operators*, volume 43 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995. 29

[Fig76] T. Figiel. On the moduli of convexity and smoothness. *Studia Math.*, 56(2):121–155, 1976. 4, 27

[GM83] M. Gromov and V. D. Milman. A topological application of the isoperimetric inequality. *Amer. J. Math.*, 105(4):843–854, 1983. 8

[Han56] Olof Hanner. On the uniform convexity of L^p and l^p. *Ark. Mat.*, 3:239–244, 1956. 5

[JLS22] A. Jambulapati, Y. Liu, and A. Sidford. Chaining, group leverage score overestimates, and fast spectral hypergraph sparsification. Available at arXiv:2209.10539. *To appear in* STOC'23, 2022. 5, 9, 29

[JLV22] Arun Jambulapati, Yin Tat Lee, and Santosh S Vempala. A slightly improved bound for the KLS constant. Available at arXiv: 2208.11644, 2022. 8

[JS01] William B. Johnson and Gideon Schechtman. Finite dimensional subspaces of L_p. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 837–870. North-Holland, Amsterdam, 2001. 28

[JSS18] Arun Jambulapati, Kirankumar Shiragur, and Aaron Sidford. Efficient structured matrix recovery and nearly-linear time algorithms for solving inverse symmetric m-matrices. *arXiv*, arxiv:1812.06295, 2018. 10, 26

[KKTY21a] Michael Kapralov, Robert Krauthgamer, Jakab Tardos, and Yuichi Yoshida. Spectral hypergraph sparsifiers of nearly linear size. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 1159–1170. IEEE, 2021. 5, 9, 29

[KKTY21b] Michael Kapralov, Robert Krauthgamer, Jakab Tardos, and Yuichi Yoshida. Towards tight bounds for spectral sparsification of hypergraphs. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 598–611. ACM, 2021. 5

[KL22] Bo'az Klartag and Joseph Lehec. Bourgain's slicing problem and KLS isoperimetry up to polylog. *Geom. Funct. Anal.*, 32(5):1134–1159, 2022. 8

[Kla23] Bo'az Klartag. Logarithmic bounds for isoperimetry and slices of convex sets. Available at arxiv:2303.14938, 2023. 4, 8

[KLM+17] Michael Kapralov, Yin Tat Lee, Cameron Musco, Christopher Musco, and Aaron Sidford. Single pass spectral sparsification in dynamic streams. *SIAM J. Comput.*, 46(1):456–477, 2017. 10, 26

[Lee22] James R Lee. Spectral hypergraph sparsification via chaining. Available at arXiv: 2209.04539. *To appear in* STOC'23., 2022. 5, 9, 29, 38

[Lew78] D. R. Lewis. Finite dimensional subspaces of L_p. *Studia Math.*, 63(2):207–212, 1978. 9, 28

[Lew79] D. R. Lewis. Ellipsoids defined by Banach ideal norms. *Mathematika*, 26(1):18–29, 1979. 9, 28

[Lov83] L. Lovász. Submodular functions and convexity. In *Mathematical programming: the state of the art (Bonn, 1982)*, pages 235–257. Springer, Berlin, 1983. 3

[LT11] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition. 7, 14

[LV07] László Lovász and Santosh S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007. 10, 26

[LW08] R. Latała and J. O. Wojtaszczyk. On the infimum convolution inequality. *Studia Math.*, 189(2):147–187, 2008. 14

[MP89] Vitali Milman and Alain Pajor. Cas limites dans des inégalités du type de Khinchine et applications géométriques. *C. R. Acad. Sci. Paris Sér. I Math.*, 308(4):91–96, 1989. 7, 9

[MP12] Gary L. Miller and Richard Peng. Iterative approaches to row sampling. *arXiv*, arxiv:1211.2713v1, 2012. 10, 26

[PTJ85] Alain Pajor and Nicole Tomczak-Jaegermann. Remarques sur les nombres d'entropie d'un opérateur et de son transposé. *C. R. Acad. Sci. Paris Sér. I Math.*, 301(15):743–746, 1985. 7

[Rud99] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999. 3

[RY22] Akbar Rafiey and Yuichi Yoshida. Sparsification of decomposable submodular functions. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10336–10344. AAAI Press, 2022. 2, 3

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011. 3, 6

[ST11] Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM J. Comput.*, 40(4):981–1025, 2011. 3

[SY19] Tasuku Soma and Yuichi Yoshida. Spectral sparsification of hypergraphs. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2570–2581. SIAM, 2019. 5

[SZ01] Gideon Schechtman and Artem Zvavitch. Embedding subspaces of L_p into l_p^N, $0 < p < 1$. *Math. Nachr.*, 227:133–142, 2001. 30

[Tal90] Michel Talagrand. Embedding subspaces of L_1 into l_1^N. *Proc. Amer. Math. Soc.*, 108(2):363–369, 1990. 4

[Tal95] M. Talagrand. Embedding subspaces of L_p in l_p^N. In *Geometric aspects of functional analysis (Israel, 1992–1994)*, volume 77 of *Oper. Theory Adv. Appl.*, pages 311–325. Birkhäuser, Basel, 1995. 4, 28

[Tal14] Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics*. Springer, Heidelberg, 2014. 11, 12

[Wen48] J. G. Wendel. Note on the gamma function. *Amer. Math. Monthly*, 55:563–564, 1948. 23