Sparsifying Sums of Norms

Arun Jambulapati
Computer Science & Engineering
Univesity of Washington
Seattle, USA
jmblpati@alumni.stanford.edu

James R. Lee
Computer Science & Engineering
University of Washington
Seattle, USA
jrl@cs.washington.edu

Yang P. Liu
Mathematics
Stanford University
Palo Alto, USA
yangpliu@stanford.edu

Aaron Sidford
MS&E
Stanford University
Palo Alto, USA
sidford@stanford.edu

***Abstract*—For any norms $N_1,\dots,N_m$ on $\mathbb{R}^n$ and $N(x) := N_1(x)+\dots+N_m(x)$, we show there is a sparsified norm $\tilde{N}(x) = w_1N_1(x)+\dots+w_mN_m(x)$ such that $|N(x)-\tilde{N}(x)| \leqslant \varepsilon N(x)$ for all $x \in \mathbb{R}^n$, where $w_1,\dots,w_m$ 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 $\mathrm{poly}(n)$-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.[1]**

***Index Terms*—Randomness in Computing**

I. Introduction

Consider a collection $N_1,\dots,N_m : \mathbb{R}^n \to \mathbb{R}_+$ of semi-norms[2] on $\mathbb{R}^n$ and the semi-norm defined by

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

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 $\tilde{N}(x) := w_1N_1(x)+\dots+w_mN_m(x)$. 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

$$|N(x) - \tilde{N}(x)| \leqslant \varepsilon N(x)\,, \quad \forall x \in \mathbb{R}^n\,. \tag{I.1}$$

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

$$|N(x) - \tilde{N}(x)| \leqslant \varepsilon N(x)\,, \quad \forall x \in \{0,1\}^V\,, \tag{I.2}$$

where $x \in \{0,1\}^V$ naturally indexes cuts in G. A straightforward ℓ_1 variant of the discrete Cheeger inequality shows that (I.2) is equivalent to (I.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 semi-norm N is appropriately well-conditioned. Say that N is *(r,R)-rounded* if it holds that $r\|x\|_2 \leqslant N(x) \leqslant R\|x\|_2$ for all $x \in \ker(N)^\perp$, where $\ker(N) := \{x \in \mathbb{R}^n : N(x) = 0\}$.

Theorem I.1. *Consider a collection $N_1,\dots,N_m$ of semi-norms on $\mathbb{R}^n$ and $N(x) := N_1(x) + \dots + N_m(x)$. 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 semi-norm N is (r,R)-rounded, then weights realizing the approximation can be found in time $O(m(\log n)^{O(1)} + n^{O(1)})(\log(mR/r))^{O(1)}\mathcal{T}_{\mathrm{eval}}$ 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

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

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 $F(S) := f_1(S) + \dots + f_m(S)$. Given nonnegative weights $w_1,\dots,w_m$, define $\tilde{F}(S) := w_1f_1(S) + \dots + w_mf_m(S)$. 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

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

We thank anonymous reviewers for several helpful comments. 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.

[1] This paper is an extended abstract. The full paper can be accessed at https://arxiv.org/abs/2305.09049.

[2] 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 $x \neq 0$.

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 $\varepsilon > 0$, 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 III-B1). Therefore, Theorem I.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 I.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 $O(mn(\log n)^{O(1)} + \mathrm{poly}(n))\mathcal{T}_{\mathrm{eval}} \log^{O(1)}(mR)$, with high probability, assuming each f_i can be evaluated in time $\mathcal{T}_{\mathrm{eval}}$.

The deduction of Corollary I.2 from Theorem I.1 appears in Section III-B1.

Sums of higher powers. In the setting of graphs, *spectral sparsification* [ST11], a notion stronger than (I.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

$$\left|N(x)^2 - \tilde{N}(x)^2\right| \leqslant \varepsilon N(x)^2 , \quad \forall x \in \mathbb{R}^n . \tag{I.3}$$

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 $uv \in E$, (I.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

$$N_1(x) = |\langle a_1, x\rangle|, \ldots, N_m(x) = |\langle a_m, x\rangle|, \qquad a_1, \ldots, a_m \in \mathbb{R}^n . \tag{I.4}$$

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

$$\frac{N(x+y)^p + N(x-y)^p}{2} \leqslant N(x)^p + N(Sy)^p , \qquad x, y \in \mathbb{R}^n . \tag{I.5}$$

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 I.1.

Theorem I.3 (Sums of pth powers of uniformly smooth norms)**.** *Consider $p \geqslant 1$ and semi-norms $N_1, \ldots, N_m$ 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 semi-norm N is K-equivalent to a semi-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 II.3 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 $A_1, \ldots, A_m \in \mathbb{R}^{d\times n}$ and $\hat{N}_1, \ldots, \hat{N}_m$ 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 (I.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 (I.2) and (I.1)), the notions of approximation guaranteed by Theorem I.1 and Theorem I.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 I.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 $Z_1, \ldots, Z_m$ 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 I.3 shows the following: For any n-dimensional subspace $X \subseteq Z$ and $\varepsilon > 0$, there are indicies $i_1, \ldots, i_d \in \{1, \ldots, m\}$ 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 I.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 (I.3). In this setting, a sequence of works [SY19], [BST19], [KKTY21b], [KKTY21a], [JLS23], [Lee23] 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 I.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 I.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 I.3 can be applied to give nontrivial sparsification results in substantially more general settings, as the next example shows.

Example I.4 (Sparsification for matrix norms)**.** Consider a matrix generalization of this setting: $X \in \mathbb{R}^{d \times d}$, and matrices $S_1, \ldots, S_m$ with $S_i \in \mathbb{R}^{d_i \times d}$, and $T_1, \ldots, T_m$ 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 semi-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 [JLS23], [Lee23]: For every $\varepsilon > 0$, there is an $O(\varepsilon^{-2} \log(r) \cdot n \log n)$-sparse ε-approximation to N^2. In the full paper, we obtain the following generalization.

Theorem I.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 $2 \leqslant p_1, \ldots, p_m \leqslant p$. Suppose that $N_1, \ldots, N_m$ 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 $k_1, \ldots, k_m \leqslant r$, then each $\|A_i x\|_\infty$ is $O(1)$-equivalent to $\|A_i x\|_p$ for $p \asymp \log r$, and thus Theorem I.5 generalizes the aforementioned result for spectral hypergraph sparsifiers. One should note that, for any fixed $p \geqslant 2$, Theorem I.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 I.3. In the full version, 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 I.6 (General sums of pth powers)**.** *If $N_1, \ldots, N_m$ 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 I.3 directly for $K = \sqrt{n}, S = 1, p = 2$ results in a worse sparsity bound of $O(\varepsilon^{-2} n^2 \log(n/\varepsilon)^2 (\log n)^3)$.

II. 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 I.1. Our method for constructing sparsifiers is simply independent sampling: Consider a probability distribution $\rho = (\rho_1, \ldots, \rho_m) \in (0, 1]^m$

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:

$$\rho_i := \frac{\int_{\mathbb{R}^n} N_i(x)\, e^{-N(x)}\, dx}{\int_{\mathbb{R}^n} N(x)\, e^{-N(x)}\, dx}. \tag{II.1}$$

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, 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

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

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 the full paper 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 asserts that

$$\mathbb{E} \max_{x\in B_N} V_x \lesssim \int_0^\infty \sqrt{\log \mathcal{K}(B_N, d, r)}\, dr, \tag{II.3}$$

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

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

Then we can bound

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

Using this in (II.3) allows us to bound $\mathbb{E}\max_{x\in B_N} V_x$ by

$$M^{-1/2}\left(\max_{x\in B_N}\tilde{N}(x)\right)^{1/2}\int_0^\infty \sqrt{\log\mathcal{K}(B_N, (\mathcal{N}^\infty)^{1/2}, r)}\, dr$$
$$= M^{-1/2}\left(\max_{x\in B_N}\tilde{N}(x)\right)^{1/2}\int_0^{\sqrt{\kappa}} \sqrt{\log\mathcal{K}(B_N, \mathcal{N}^\infty, r^2)}\, dr, \tag{II.4}$$

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 (II.4), 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

$$\sqrt{\log\mathcal{K}(B_2^n, N, r)} \lesssim \frac{1}{r}\,\mathbb{E}\,[N(g)], \tag{II.5}$$

where g is a standard n-dimensional Gaussian.

An adaptation of the Pajor-Talagrand proof of (II.5) (see Lemma II.2) allows one to show that for any norms N and $\hat{N}$ on $\mathbb{R}^n$,

$$\log\mathcal{K}(B_N, \hat{N}, r) \lesssim \frac{1}{r}\,\mathbb{E}\left[\hat{N}(\boldsymbol{Z})\right], \tag{II.6}$$

where $\boldsymbol{Z}$ has density proportional to $e^{-N(x)}\,dx$. A closely related estimate was proved by Milman and Pajor [MP89]; see the remarks after (II.13). As this fact is simple and essential to our approach, we provide a proof here. We begin with a useful fact which bounds the measure of shifts of convex sets.

Lemma II.1 (Shift Lemma)**.** *Suppose N is a norm on $\mathbb{R}^n$. Define the probability measure μ on $\mathbb{R}^n$ by*

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

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

$$\mu(W+z) \geqslant \exp(-N(z))\,\mu(W)\,. \tag{II.7}$$

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

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

Now we bound

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

where the equality uses symmetry of W, the first inequality uses convexity of $\exp(x)$, and the second inequality uses the traingle inequality for N. □

Lemma II.2. *Let N and $\hat{N}$ be norms on $\mathbb{R}^n$. Define the probability measure μ on $\mathbb{R}^n$ so that*

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

Then for any $\varepsilon > 0$,

$$\log\left(\mathcal{K}(B_N,\hat{N},\varepsilon)/2\right) \leqslant \frac{2}{\varepsilon}\int \hat{N}(x)\,d\mu(x)\,.$$

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

$$\begin{aligned}1 \geqslant \mu\left(\bigcup_{j\in[M]} \lambda(x_j + B_{\hat{N}})\right) &= \sum_{j\in[M]} \mu\left(\lambda x_j + \lambda B_{\hat{N}}\right) \\ &\overset{(\text{II.7})}{\geqslant} \sum_{j\in[M]} e^{-\lambda N(x_j)}\mu(\lambda B_{\hat{N}}) \\ &\geqslant M e^{-\lambda}\mu(\lambda B_{\hat{N}})\,,\end{aligned}$$

where (II.7) used Lemma II.1 and the last inequality used $x_1,\ldots,x_M \in B_N$.

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

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

Combining with the preceding inequality yields the upper bound

$$(\log(M/2)) \leqslant \lambda\,. \qquad \square$$

Applying this with $\hat{N} = \mathcal{N}^\infty$ yields

$$\log\mathcal{K}(B_N,\mathcal{N}^\infty,r) \lesssim \frac{1}{r}\,\mathbb{E}\left[\mathcal{N}^\infty(\boldsymbol{Z})\right] = \frac{1}{r}\,\mathbb{E}\max_{j\in[M]}\frac{N_{i_j}(\boldsymbol{Z})}{\rho_{i_j}}\,. \tag{II.8}$$

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 II.3. *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 $t > 0$,*

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

This implies the following consequence.

Corollary II.4. *There is a constant $c > 0$ such that the following holds. Consider a semi-norm $\mathcal{N}$ on $\mathbb{R}^n$ and a random vector $\boldsymbol{Z}$ whose distribution is symmetric and log-concave. Then for any $t > 0$,*

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

Proof. 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 II.3 gives the desired result once we confirm the Lipschitz bound

$$\mathcal{N}(A^{1/2}x) \leqslant 2\,\mathbb{E}[\mathcal{N}(\boldsymbol{Z})]\cdot\|x\|_2\,. \tag{II.9}$$

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 standard facts on moments of log-concave variables: $\langle y, \boldsymbol{Z}\rangle$ is symmetric and log-concave. $\square$

With this in hand, a union bound gives

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

To make ρ a probability measure, we take $\rho_j := \mathbb{E}[N_j(\boldsymbol{Z})]/\mathbb{E}[N(\boldsymbol{Z})]$ for $j = 1,\dots,m$, and then (II.8) becomes

$$\log \mathcal{K}(B_N, \mathcal{N}^\infty, r) \lesssim \frac{1}{r}(\psi_n \log M)\,\mathbb{E}[N(\boldsymbol{Z})] = \frac{1}{r} n\psi_n \log(M)\,, \tag{II.10}$$

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 $e^{-N(x)}$. Thus we have

$$\int_{\sqrt{\kappa}/n^2}^{\sqrt{\kappa}} \sqrt{\log \mathcal{K}(B_N, \mathcal{N}^\infty, r^2)}\,dr \lesssim \left(n\psi_n \log M\right)^{1/2} \int_{\sqrt{\kappa}/n^2}^{\sqrt{\kappa}} \frac{1}{r}\,dr \lesssim \left(n\psi_n \log M\right)^{1/2} \log n\,. \tag{II.11}$$

Standard volume arguments in $\mathbb{R}^n$ 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 (II.11) into (II.4) 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 (II.10) 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 (II.2).

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 (II.1):

$$\rho_i := \frac{\int_{\mathbb{R}^n} N_i(x)^p\, e^{-N(x)^p}\,dx}{\int_{\mathbb{R}^n} N(x)^p\, e^{-N(x)^p}\,dx}\,. \tag{II.12}$$

Note that if $p = 2$ and one defines $N_i(x) := |(Ax)_i|$ for a matrix $A \in \mathbb{R}^{m\times n}$, then ρ_i are exactly the leverage scores of A (up to scaling by n). 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 (II.6) for p-uniformly smooth norms. It turns out however that 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 $\hat{N}$,

$$(\log \mathcal{K}(B_N, B_{\hat{N}}, r))^{1/p} \lesssim \frac{1}{r}\,\mathbb{E}[\hat{N}(\boldsymbol{Z})]\,. \tag{II.13}$$

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 I.6 for general norms, we must resort to a dimension-dependent version of (II.13). Moreover, we need to augment the sampling probabilities in (II.12) 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 I.5), in the full paper we formulate a generalization of ℓ_p Lewis weights, motivated by the construction of weights in [KKTY21a], [JLS23], [Lee23].

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

$$\max\{|\det(U)| : \alpha(U) \leqslant 1\}\,, \tag{II.14}$$

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

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

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

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

One can establish properties of the corresponding optimizer of (II.14), leading to the following generalization of the Lewis weights.

Definition II.5 (Block norm)**.** Consider any $p_1, \dots, p_m, q \in [1,\infty]$, 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 $\mathcal{N}(u) := \|(\mathcal{N}_1(u), \dots, \mathcal{N}_m(u))\|_q$.

Lemma II.6. *Consider $p_1, \dots, p_m \in [2,\infty]$ and $q \in [1,\infty)$. Let $\mathcal{N}_1, \dots, \mathcal{N}_m$ and $\mathcal{N}$ be as in Definition II.5. Fix $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^k)$ with $\operatorname{rank}(A) = n$, and denote for $j = 1, \dots, m$,*

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

Then there is a nonnegative diagonal matrix W such that for $U = (A^\top W A)^{-1/2}$, 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 $x \in \mathbb{R}^n$,

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

III. Algorithms

A. Computing the sampling weights via homotopy

We now 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 multiple settings (see, e.g., [MP12], [KLM+17], [JSS18], [AJSS19]).

o compute reasonable overestimates of the sampling weights $\{\rho_i\}$ from (II.1), one approach is to simply sample from the probability measure μ with density proportional to $e^{-N(x)}$, evaluate the N_i at the sample, and use a scaling of the average evaluation of N_i as the estimate of ρ_i. Sampling from a log-concave distribution, especially those induced by norms, is a well-studied task, and can be done in $n^{O(1)} \log^{O(1)}(nR/r)$ time if N is (r,R)-rounded; see [CV18], [JLLV21] and Theorem III.3, Corollary III.4. If a norm evaluation can be performed in time $\mathcal{T}_{\mathrm{eval}}$, this would naively require time $mn^{O(1)} \log^{O(1)}(nR/r)\mathcal{T}_{\mathrm{eval}}$, 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, for simplicity let us assume that N is a genuine norm that 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. To handle the case when N is a seminorm we modify this approach to instead obtain $\tilde{N}(x)$ such that $\tilde{N}(x) + \varepsilon r\|x\|_2$ is an ε-approximation to $N_{\varepsilon r}$ and argue that this suffices for $\tilde{N}$ to be an $O(\varepsilon)$-approximation of N.

B. Details and analysis

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

Theorem III.1 (Efficient sparsification)**.** *If N is (r,R)-rounded, then for any $\varepsilon \geqslant n^{-O(1)}$, there is an algorithm running in time $(m(\log n)^{O(1)} + n^{O(1)})(\log(mR/r))^{O(1)}\mathcal{T}_{\mathrm{eval}}$ 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 semi-norm on $\mathbb{R}^n$ that is K-equivalent to N, and let μ be the probability measure with density proportional to $e^{-\tilde{N}(x)}\,dx$.

Lemma III.2 (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 $\boldsymbol{X}_1, \dots, \boldsymbol{X}_k \in \mathbb{R}^n$ be independent samples from μ. Denote, for $i = 1, \dots, m$,

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

Since μ is log-concave, Corollary II.4 asserts there is a constant $c > 0$ such that

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

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

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

Thus (as discussed in the full paper) with high probability sampling proprotional to τ_i yields the desired sparse approximation. □

The preceding lemma shows that sampling from a distribution with $d\mu(x) \propto e^{-\tilde{N}(x)}$ suffices to efficiently sparsify a semi-norm N that is K-equivalent to $\tilde{N}$. A long line of work establishes algorithms that sample from a distribution that is close to uniform on any well-conditioned convex body $A \subseteq \mathbb{R}^n$, given only membership oracles to A. In the following statement, let B_2^n denote the Euclidean unit ball in $\mathbb{R}^n$.

Theorem III.3 ([JLLV21, Theorem 1.5], [CV18, Theorem 1.2])**.** *There is an algorithm that, given a convex body $A \subseteq \mathbb{R}^n$ satisfying $r \cdot B_2^n \subseteq A \subseteq R \cdot B_2^n$ and $\varepsilon > 0$, samples from a distribution that is within TV distance ε from the uniform measure on A using $O(n^3(\log \frac{nR}{\varepsilon r})^{O(1)})$ membership oracle queries to A, and $(n(\log \frac{nR}{\varepsilon r}))^{O(1)}$ additional time.*

When N is a norm, one obtains immediately an algorithm for sampling from the measure μ on $\mathbb{R}^n$ with density $d\mu(x) \propto e^{-N(x)}\,dx$ using evaluations of $N(x)$.

Corollary III.4. *There is an algorithm that, given an (r, R)-rounded norm N on $\mathbb{R}^n$ and $\varepsilon > 0$, samples from a distribution that is within TV distance ε from the measure μ with density proportional to $e^{-N(x)}\,dx$ using $O(n^3(\log \frac{nR}{\varepsilon r})^{O(1)})$ evaluations of $N(x)$, and $(n(\log \frac{nR}{\varepsilon r}))^{O(1)}$ additional time.*

Proof. Note that if $\boldsymbol{Z}$ has law μ, then the density of $N(\boldsymbol{Z})$ is proportional to $e^{-\lambda}\lambda^{n-1}$. Let $\boldsymbol{\lambda}$ be a sample from the latter distribution. The algorithm is as follows: Sample a point $\boldsymbol{X}$ from the uniform measure on B_N using Theorem III.3, and then output the point $\boldsymbol{\lambda}\boldsymbol{X}/N(\boldsymbol{X})$. □

Combining Lemma III.2 and Corollary III.4, 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 III.1.

Proof of Theorem III.1. Recall our assumption that $r\|x\|_2 \leqslant N(x) \leqslant R\|x\|_2$ for all $x \in \ker(N)^\perp$. 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 by sampling from $d\mu(x) \propto \exp(-R\|x\|_2)$ using Corollary III.4, we can use Lemma III.2 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. Note that $\tilde{N}_t$ is $(t/2, 4R)$-rounded. Thus, using Corollary III.4, 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 Corollary III.4 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 III.2 to obtain an $\tilde{O}(n)$-sparse 1/2-approximation to $N_{t/2}$.

After $O(\log(R/(\varepsilon r)))$ iterations, one obtains an $\tilde{O}(n)$-sparse 1/2-approximation to $N_{\varepsilon r}$. A final application of Lemma III.2 obtains an $O(n\varepsilon^{-2}\log(n/\varepsilon)(\log n)^{2.5})$-sparse ε-approximation to $N_{\varepsilon r}$. To conclude, note that for all $x \in \ker(N)^\perp$, $N_{\varepsilon r}$ is $(1+\varepsilon)$-equivalent to N. Moreover, in $N_{\varepsilon r}(x) = N(x)+\varepsilon r\|x\|_2$, only the summand $\varepsilon r\|x\|_2$ fails to vanish on $\ker(N)$. This can be removed from $N_{\varepsilon r}$ to obtain a $(1+2\varepsilon)$-approximation to N with the same sparsity. The result then follows by applying this procedure with a smaller value of ε. □

Remark III.5 (Algorithm for $1 < p \leqslant 2$)**.** We note that it is possible to extend Theorem III.1 to the setting of $1 < p \leqslant 2$ under a mild additional assumption. Specifically, we need to assume that each semi-norm N_i is itself K-equivalent to a p-uniformly smooth semi-norm $\mathcal{N}_i$ with constant $\mathcal{S}_p$, and that we have oracle access to $\mathcal{N}_i$.

For any weights $w_1, \dots, w_m \geqslant 0$, the semi-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 semi-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 semi-norms is also p-uniformly smooth quantiatively (see [Fig76]), it holds that N_w is K-equivalent to a semi-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 $1 \leqslant p \leqslant 2$.

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 I.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 III.1 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 II guarantees that $\mathbb{E}[w_1 + \cdots + w_m] = n$, so $\sum_{i=1}^m w_i \leqslant 2n$ 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 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. □

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. 7, 8

[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. 3

[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. 1, 4

[BLM89] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Math.*, 162(1-2):73–141, 1989. 3

[BSS12] Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012. 2, 3

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

[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. 3

[Che21] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture. *Geom. Funct. Anal.*, 31(1):34–61, 2021. 5

[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. 2

[CV18] Ben Cousins and Santosh S. Vempala. Gaussian cooling and $o^*(n^3)$ algorithms for volume and gaussian volume. *SIAM J. Comput.*, 47(3):1237–1273, 2018. 7, 8

[Fig76] T. Figiel. On the moduli of convexity and smoothness. *Studia Math.*, 56(2):121–155, 1976. 2, 8

[GM83] M. Gromov and V. D. Milman. A topological application of the isoperimetric inequality. *Amer. J. Math.*, 105(4):843–854, 1983. 5

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

[JLLV21] He Jia, Aditi Laddha, Yin Tat Lee, and Santosh S. Vempala. Reducing isotropy and volume to KLS: an $o^*(n^3\psi^2)$ volume algorithm. 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 961–974. ACM, 2021. 7, 8

[JLS23] Arun Jambulapati, Yang P. Liu, and Aaron Sidford. Chaining, group leverage score overestimates, and fast spectral hypergraph sparsification. In Barna Saha and Rocco A. Servedio, editors, *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pages 196–206. ACM, 2023. Full version at arXiv:2209.10539. 3, 6

[JLV22] Arun Jambulapati, Yin Tat Lee, and Santosh S Vempala. A slightly improved bound for the KLS constant. Available at arXiv: 2208.11644, 2022. 5

[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. 7, 8

[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. 3, 6

[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. 3

[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. 5

[Kla23] Bo'az Klartag. Logarithmic bounds for isoperimetry and slices of convex sets. Available at arxiv:2303.14938, 2023. 2, 5

[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. 7, 8

[Lee23] James R. Lee. Spectral hypergraph sparsification via chaining. In Barna Saha and Rocco A. Servedio, editors, *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pages 207–218. ACM, 2023. Full version at arXiv: 2209.04539. 3, 6

[Lew78] D. R. Lewis. Finite dimensional subspaces of L_p. *Studia Math.*, 63(2):207–212, 1978. 6

[Lew79] D. R. Lewis. Ellipsoids defined by Banach ideal norms. *Mathematika*, 26(1):18–29, 1979. 6

[Lov83] L. Lovász. Submodular functions and convexity. In *Mathematical programming: the state of the art (Bonn, 1982)*, pages 235–257. Springer, Berlin, 1983. 2

[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. 4

[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. 4, 6

[MP12] Gary L. Miller and Richard Peng. Iterative approaches to row sampling. *arXiv*, arxiv:1211.2713v1, 2012. 7, 8

[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. 4

[Rud99] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999. 2

[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

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011. 2, 4

[ST11] Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM J. Comput.*, 40(4):981–1025, 2011. 2

[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. 3

[Tal90] Michel Talagrand. Embedding subspaces of L_1 into l_1^N. *Proc. Amer. Math. Soc.*, 108(2):363–369, 1990. 3

[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. 3