How Hard is Robust Mean Estimation?

Samuel B. Hopkins HOPKINS@BERKELEY.EDU
University of California, Berkeley
Jerry Li JERRL@MICROSOFT.COM
Microsoft Research

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Robust mean estimation is the problem of estimating the mean $\mu \in \mathbb{R}^d$ of a d-dimensional distribution D from a list of independent samples, an ε-fraction of which have been arbitrarily corrupted by a malicious adversary. Recent algorithmic progress has resulted in the first polynomial-time algorithms which achieve *dimension-independent* rates of error: for instance, if D has covariance I, in polynomial-time one may find $\hat{\mu}$ with $\|\mu - \hat{\mu}\| \leq O(\sqrt{\varepsilon})$. However, error rates achieved by current polynomial-time algorithms, while dimension-independent, are sub-optimal in many natural settings, such as when D is sub-Gaussian, or has bounded 4-th moments.

In this work we give worst-case complexity-theoretic evidence that improving on the error rates of current polynomial-time algorithms for robust mean estimation may be computationally intractable in natural settings. We show that several natural approaches to improving error rates of current polynomial-time robust mean estimation algorithms would imply efficient algorithms for the small-set expansion problem, refuting Raghavendra and Steurer's small-set expansion hypothesis (so long as $\mathsf{P} \neq \mathsf{NP}$). We also give the first direct reduction to the robust mean estimation problem, starting from a plausible but nonstandard variant of the small-set expansion problem.

Keywords: robust mean estimation, small-set expansion, complexity of learning, robust statistics, spectral graph theory

1. Introduction

Robust mean estimation is the following basic statistical problem: given a list of n samples $X_1, \ldots, X_n$ from some unknown probability distribution D on $\mathbb{R}^d$, an unknown ε-fraction of which have been arbitrarily corrupted by a malicious adversary, find a vector $\hat{\mu}$ such that $\|\hat{\mu} - \mathbb{E}_{X \sim \mathcal{D}} X\|$ is as small as possible, where (for this paper) $\| \cdot \|$ is the Euclidean norm.

Among other natural settings, robust mean estimation models estimation using data sets which contain outliers – due to random corruptions or malicious data poisoning – and, if D is assumed to lie in some class $\mathcal{C}$ of distributions, estimation when nature only produces data from a distribution which is ε-close to some distribution in $\mathcal{C}$ in statistical distance. It is the most elementary of many high-dimensional statistical estimation problems which become both statistically and computationally difficult in the presence of a small constant fraction of adversarial corruptions: robust covariance estimation, robust learning of hidden-variable models, and more.

Statisticians have studied estimation under adversarially-chosen corruptions since the 1960s, originally with the notion of "breakdown points" Anscombe (1960); Tukey (1960); Huber (1964); Tukey (1975). However, until recently, statistically-optimal rates of error when an ε-fraction of data is corrupted were out of reach for computationally efficient algorithms. For instance, if $X_1, \ldots, X_n$ are ε-corrupted samples from $\mathcal{N}(\mu, I)$, then the estimator which outputs the *Tukey median* of $X_1, \ldots, X_n$ with high probability achieves $\|\mu - \text{TukeyMedian}(X_1, \ldots, X_n)\| \leq O(\varepsilon)$, when $n \geq d/\varepsilon^2$ Tukey (1975). Unfortunately, the Tukey median is NP-hard to compute in high dimensions, at least for worst-case $X_1, \ldots, X_n$ Bernholt (2006).

Naive polynomial-time approaches, such as individually pruning X_i's at large distance to the rest of $X_1, \ldots, X_n$, suffer much worse rates of error: typically they lead to estimators $\hat{\mu}$ with $\|\mu - \hat{\mu}\| \leq O(\sqrt{\varepsilon d})$, even when the uncorrputed samples come from a nice distribution, such as a Gaussian as above. Notably, the rate of error for such estimators grows with the ambient dimension d.

Recently, the first dimension-independent error rates for robust mean estimation were achieved by Diakonikolas et al. (2016). Simultaneously and independently, Lai et al. (2016) achieved error for robust mean estimation scaling with the dimension as $O(\log d)$. These works sparked a great deal of activity in algorithm design for robust statistics, leading to new algorithms for robust mean estimation under sparsity assumptions, robust clustering and robust learning of mixture models, robust linear regression, and more (see Li (2018); Steinhardt (2018b) for surveys of recent work).

In spite of the substantial algorithmic success, current algorithms remain statistically sub-optimal in many settings, especially with respect to the dependence of the error rate $\|\mu - \hat{\mu}\|$ on ε. In this paper we are interested in the question:

> Do current polynomial-time algorithms for high-dimensional robust mean estimation achieve optimal error rates among all polynomial-time algorithms?

Contributions Our main contribution is a family of reductions from several variants of the *small-set expansion problem*, a close cousin of Khot's unique games problem, to robust mean estimation and related problems. These reductions show that (a) current approaches for improving error rates of existing algorithms for robust mean estimation under natural assumptions on D (such as bounded 4-th moments) would refute Raghavendra and Steurer's small-set expansion hypothesis, and (b) any efficient algorithm improving on the error rates of current algorithms for robust mean estimation under Steinhardt, Charikar, and Valiant's *resilience* assumption on D (see below) would refute a strengthened version of the small-set expansion hypothesis.

Our reductions employ tools from spectral graph theory. We reinterpret and strengthen ideas from Barak et al's proof that the $2 \to 4$ norm of a matrix is hard to approximate to any constant factor under the small set expansion hypothesis Barak et al. (2012a). Our reinterpretation results in a simple characterization of small sets of vectors in the spectral embedding of a small-set expander (see Section 3). This characterization leads to our main results. Along the way we dramatically simplify (and generalize) Barak et al. (2012a)'s proof of small-set expansion hardness of $2 \to q$ norms, which may be of independent interest.

Beating $\sqrt{\varepsilon}$: the complexity landscape We turn to a more quantitative discussion of our main question. In order for robust mean estimation to be information-theoretically solvable with non-trivial error guarantees (that is, solvable by any algorithm, irrespective of running time), some assumption must be made on the underlying distribution D. A common and mild assumption is that D has covariance $\Sigma \preceq I$. In this case, robust mean estimation is possible both information-theoretically and by polynomial-time algorithms with error rate $O(\sqrt{\varepsilon})$, and (up to constants) this is information-theoretically optimal.

Better scaling with ε is possible under stronger assumptions on D. For instance, if D has p-th moments bounded by a dimension-independent constant, then error rate $O(\varepsilon^{1-1/p})$ is information-theoretically achievable. Relatedly, Steinhardt, Charikar, and Valiant Steinhardt et al. (2017) introduced a weaker notion: (σ, ε)-*resilience*. A distribution D is (σ, ε)-resilient if every event of probability at least $(1 - \varepsilon)$ has conditional mean μ' with $\|\mu' - \mu\| \leq \sigma$, where μ is the mean of D. They show that ε-robust mean estimation is then possible with error $O(\sigma)$.

So far, no polynomial-time algorithm is known which achieves error better than $O(\sqrt{\varepsilon})$ under any resilience assumption, nor is a polynomial-time algorithm known which achieves error better than $O(\sqrt{\varepsilon})$ under a bounded p-th moments assumption. Thus, a second question which motivates this paper is:

> What structure in the distribution D of uncorrupted samples can be exploited by polynomial-time algorithms to perform robust mean estimation with error $\varepsilon^{1/2+\Omega(1)}$?

Of course *a priori* it could be that no polynomial-time algorithm has error better than $O(\sqrt{\varepsilon})$, but this is not the case. If D is Gaussian, then error $O(\varepsilon \log(1/\varepsilon))$ can be achieved in polynomial time Diakonikolas et al. (2016). And, if D has *certifiably* bounded p-th moments (a strengthening of p-th moment boundedness introduced independently by Hopkins and Li (2018) and by Kothari et al. (2018)), then error $O(\varepsilon^{1-1/p})$ is achievable in polynomial time. Furthermore, many natural distributions fall into the latter category: product distributions and strongly log-concave distributions, for example.

Thus, there is a nontrivial complexity landscape in robust mean estimation. Our results point to new points of hardness in this landscape. We show that *current approaches* to robust estimation under both resilience and moment-boundedness (which in particular also solve *certification* problems associated to moments and to resilience) would refute the small-set expansion hypothesis if they could be improved to error rate $\varepsilon^{1/2+\Omega(1)}$. And we show that *any efficient algorithm* achieving error $\varepsilon^{1/2+\Omega(1)}$ under a resilience assumption would refute a nonstandard version of the small-set expansion hypothesis (see Section B).

Complexity of learning under niceness assumptions Typically, results on computational complexity of learning take one of three forms: (1) reduction from an NP-hard problem, (2) reduction from a problem which is believed to be average-case hard, such as planted clique or learning parities with noise, or (3) unconditional lower bounds against of a restricted class of algorithms, such as statistical query (SQ) algorithms or particular hierarchies of convex programs.

Approach (1) is appealing because it can yield lower bounds which apply to all polynomial-time algorithms based on weak and well-tested assumptions like $P \neq NP$. Often, however, applications of approach (1) prove hardness of learning problems under input distributions which do not satisfy natural *niceness conditions* – assumptions like such as input data being drawn from a Gaussian distribution or from a distribution with bounded moments – because they rely on embedding gadgets in the input distribution. Algorithm designers often avoid such complexity results by assuming niceness conditions like these.

We follow approach (1) as well (subject to the small set expansion hypothesis), but our reductions produce nice input distributions, satisfying regularity conditions such as bounded moments or resilience; we therefore provide evidence from worst-case complexity that robust learning is hard *even under niceness assumptions*. The majority of technical work in this paper is devoted to showing that our reductions produce such nice distributions.

We note that approach (3), in the form of SQ lower bounds, has been investigated for the robust mean estimation problem – see Section 1.1.

Open problems This paper only begins the study of hardness of robust estimation problems based on worst-case complexity assumptions: there is a great deal left to do! We outline several open problems in Section C.

1.1. Related Work

Robust statistics The study of robust statistics, and specifically robust mean estimation, was initiated by seminal work of statisticians in the 60's and 70's Anscombe (1960); Tukey (1960); Huber (1964); Tukey (1975). However, it was not until recently that efficient algorithms were discovered for robust mean estimation in high dimensions which acheive nearly optimal error guarantees Diakonikolas et al. (2016); Lai et al. (2016); Diakonikolas et al. (2017b). The field has since experienced an explosion of algorithmic work. For a survey on more recent algorithmic results, see Li (2018); Steinhardt (2018b).

PAC learning lower bounds While there is a large literature on lower bounds for distributional learning problems either from average case assumptions or applying to a restricted classes of algorithms, there are only a handful of results we are aware of which base hardness of such problems on worst case hardness assumptions Kearns et al. (1994); Guruswami and Raghavendra (2009); Feldman et al. (2006); Applebaum et al. (2008); Regev (2009); Feldman and Kanade (2012); Bun and Zhandry (2016); Bubeck et al. (2018). Moreover, the lower bounds tend to be proved in a PAC learning sense, where the learning problem is *worst-case over distributions*. We consider a version of robust mean estimation which is worst-case over input distributions *belonging to a class of nice distributions, i.e. resilient distributions or those with bounded moments.* This amount of worst-case-ness allows us to base our results on worst-case hardness assumptions, but requires significant work in our reductions to produce such nice input distributions.

Computational lower bounds in robust statistics In the context of robust estimation, almost all known lower bounds were either against restricted classes of algorithms, notably statistical query algorithms Diakonikolas et al. (2017c, 2018, 2019), or against specific estimators Johnson and Preparata (1978); Bernholt (2006).

In particular, Diakonikolas et al. (2017c) proves an SQ lower bound in the setting of robust mean estimation for Gaussian distributions suggesting that the $O(\varepsilon)$ vs $O(\varepsilon \log 1/\varepsilon)$ gap between information-theoretically optimal error rates and those of known polynomial-time algorithms is likely inherent. For at least one of the problems we investigate – complexity of robust mean estimation under bounded moment assumptions – it would not be possible to prove an analogous SQ lower bound. This is because for every fixed p there is a simple (folklore) SQ *algorithm* which makes $\operatorname{poly}(d)$ statistical queries (with $1/\operatorname{poly}(d)$ tolerance) and robustly estimates the mean of a distribution with bounded p-th moments to (information-theoretically optimal) accuracy $O(\varepsilon^{1-1/p})$.[1]

[1] For simplicity we ignore the dependence of the number of statistical queries and tolerance on ε.

The only implementations of this algorithm we are aware of require $\exp(d)$ additional running time *but the SQ framework only allows for lower bounds on the number and tolerance of queries, not on the additional running time to process the answers to those queries*. We expect similar guarantees to be unachievable in polynomial time (especially in light of the present work), SQ lower bounds cannot evidence this. A different approach, such as the reduction-based arguments we pursue here, is required.

The only other work we are aware of giving a reduction from small-set expansion to prove complexity of a robust learning problem is Hardt and Moitra (2013), which gives lower bounds from small set expansion for the problem of identifying a low-dimensional subspace which contains a large fraction of a high-dimensional data set. While both their work and ours show reductions from the small-set expansion problem, the works otherwise diverge on a technical level – our reductions employ spectral graph theory, while theirs is largely combinatorial – and the results are incomparable. Furthermore, besides the constraint that the distribution of "good" samples lives on a low dimensional subspace, they enforce no additional niceness conditions. In particular, the distribution which results from their reduction is exponentially ill-conditioned. This stands again in contrast to the relative niceness of the distributions resulting from our reductions.

Klivans and Kothari Klivans and Kothari (2014) show hardness of robustly learning halfspaces with respect to Gaussian data; however, they start from an average-case hardness assumption (learning sparse parities with noise) rather than a worst-case one as we do here.

1.2. Results

The fundamental problem of study in this paper is robust mean estimation. At a high level, the question is as follows: given samples from a distribution D, a small fraction of which have been corrupted, estimate the mean of D as well as possible. There are several possible corruption models to consider. In this work, we will show lower bounds against the following (relatively weak) notion of corruption, which dates back to work of Huber in the 1960s Huber (1964):

Definition 1 (ε-contamination) *Let D be a distribution over $\mathbb{R}^d$. We say that that $X_1, \ldots, X_n$ is an ε-contaminated set of samples from D if the X_i are drawn i.i.d. from $(1-\varepsilon)D + \varepsilon N$, where N is an arbitrary, unknown distribution.*

This model is also known as *Huber's contamination model* in the robust statistics literature. The recent efficient algorithms Diakonikolas et al. (2016); Charikar et al. (2017); Steinhardt et al. (2017) actually work for slightly stronger notions of corruption. All of our lower bounds will be against learning from ε-contaminated samples, so in particular, they are also lower bounds against learning from corrupted samples as considered in these papers.

With these definitions, we can now formally state the robust mean estimation problem.

Problem 2 (Robust mean estimation) *Let D be a distribution with mean μ. Given δ-contaminated samples from D, output $\widehat{\mu}$ minimizing $\|\mu - \widehat{\mu}\|_2$ with high probability.*

We briefly note, as matter of notation, that in Problem 2 and the remainder of the paper, we will use δ (rather than ε, as is standard in robust statistics) to denote the fraction of corrupted samples. This will be helpful to stay notationally consistent with the literature on small-set expansion that we heavily rely on.

Without additional assumptions on D, Problem 2 is impossible: there is no way to distinguish between D and $\delta D + (1-\delta)N$, and since N can be arbitrary, the means of these two distributions can be arbitrarily far away. To make this problem statistically tractable, we must impose some conditions on D. In this paper we will focus on two previously considered conditions, namely, bounded moments and resilience.

1.2.1. Bounded moments

A canonical assumption in this area is that D has some number of bounded moments. For instance, arguably the most natural assumption is that D has bounded covariance. In this case, we have efficient algorithms matching the information theoretic lower bound:

Fact 1 (Diakonikolas et al. (2017a)) *Let $\mathcal{D}$ be the class of distributions over $\mathbb{R}^d$ distribution over $\mathbb{R}^d$ whose covariance have spectral norm at most* 1*, or equivalently, which have*

$$\mathbb{E}_{X \sim D} |\langle v, X \rangle - \langle v, \mathbb{E}_{X \sim D} X \rangle|^2 \leq 1 \;, \tag{1}$$

for all unit vectors v. There is a polynomial time algorithm, which for all small-enough $\delta > 0$ and all $D \in \mathcal{D}$, given a δ-contaminated set of samples from D of size $\mathrm{poly}(d, 1/\delta)$, outputs $\widehat{\mu}$ which with probability at least $9/10$, satisfies $\|\widehat{\mu} - \mathbb{E}_{X \sim D} \mu\| \leq O(\sqrt{\delta})$. Moreover, no estimator (efficient or not) can achieve $\|\widehat{\mu} - \mathbb{E}_{X \sim D} X\| \leq o(\sqrt{\delta})$ with probability greater than $1/10$ for all $X \in \mathcal{D}$.

Thus in this case, up to constants, there is no gap between the robustness of efficient and inefficient estimators. An obvious question is whether this can be strengthened by making additional structural assumptions on the data. For instance, what if we assume p bounded moments, for $p > 2$? Indeed, in this setting something stronger is possible, at least with exponential running time:

Fact 2 (folklore) *Let $p > 2$ and let $\mathcal{D}_p$ be the class of distributions over $\mathbb{R}^d$ whose p-th central moment is at most* 1*: that is, $D \in \mathcal{D}_p$ if and only if*

$$\mathbb{E}_{X \sim D} |\langle v, X \rangle - \langle v, \mathbb{E}_{X \sim D} X \rangle|^p \leq 1 \tag{2}$$

for all unit vectors v, There exists an exponential-time algorithm which for all $\delta > 0$ sufficiently small and all $D \in \mathcal{D}_p$, given a δ-contaminated set of samples from D of size $\mathrm{poly}(d, 1/\delta)$, outputs $\widehat{\mu}$ so that $\|\widehat{\mu} - \mathbb{E}_{X \sim D} X\| \leq O(\delta^{1-1/p})$ with probability at least $9/10$. Moreover, no estimator achieves error $\|\widehat{\mu} - \mathbb{E}_{X \sim D} X\| < o(\delta^{1-1/p})$ with probability at least $1/10$ over all of $\mathcal{D}_p$.

In particular, Fact 2 says that for $p > 2$, it is possible to outperform the guarantees of the algorithm in Fact 1 asymptotically as $\delta \to 0$. However, despite much work in the area, no efficient algorithms are known which achieve error better than $O(\delta^{1/2})$, i.e. the rate in the $p = 2$ case, unless even stronger assumptions are made. This leads to the question:

Question 3 *Is there some $p > 2$ and a polynomial-time algorithm which for all sufficiently-small $\delta > 0$ and all $D \in \mathcal{D}_p$ finds $\widehat{\mu}$ satisfying $\|\widehat{\mu} - \mathbb{E}_{X \sim D} X\| < o(\sqrt{\delta})$ with probability at least $9/10$ when given a δ-corrupted set of $\mathrm{poly}(d, 1/\delta)$ samples from D?*

Towards answering Question 3, we offer evidence that current techniques to algorithmically exploit moment boundedness cannot be extended to positively answer Question 3. The algorithms which achieve the guarantees in Fact 1 solve, as a subroutine, the problem of maximizing the left-hand side of (1) over all unit v. The algorithms of Hopkins and Li (2018); Kothari et al. (2018) which exploit p-th moment boundedness when the p-th moments satisfy additional structural assumptions analogously require subroutines which certify upper bounds on the left-hand side of (2).

A theorem of Barak et al. on hardness of computing the $2 \to q$ norm of a matrix already shows that this approach cannot be extended to $p \geq 4$ under only the assumptions specified in Question 3 without violating the small-set expansion hypothesis. In the following, D should be thought of as the uniform distribution over the vectors $a_1, \ldots, a_n$.

Theorem 4 (Barak et al. (2012a)) *If for any even $q \geq 2$ there is a polynomial-time algorithm which given $a_1, \ldots, a_n \in \mathbb{R}^d$ outputs a constant-factor approximation to $\max_{\|x\|=1} \frac{1}{n} \sum_{i=1}^n \langle a_i, x\rangle^q$, then there is a polynomial-time algorithm for small set expansion.*

In this work we strengthen Barak et al.'s result in several ways. Barak et al.'s result shows that for c, s with c/s arbitrarily large it is SSE-hard to distinguish a distribution with 4-th moment at least c from one with 4-th moment at most s. In statistical settings, however, it is natural to assume niceness of many moments. For instance: is it possible to distinguish a distribution D all of whose q-th moments for $q \leq 100$ have sub-Gaussian-type behavior (i.e. growing like $q^{q/2}$) from one whose 4-th moment is very large? An algorithm which could solve this decision problem seems likely to lead to an algorithm to improve on error $o(\sqrt{\delta})$, at least under the assumption that D has 100 sub-Gaussian moments.

We show that this apparently easier decision problem is still SSE hard. This requires modifying Barak et al.'s reduction so that in one case a distribution with sub-Gaussian moments is obtained; we do this by composing the reduction of Barak et al. with a smoothing/averaging step which we analyze via Rosenthal's moment inequality. The result addresses an open problem of Jacob Steinhardt Steinhardt (2018a). Additionally, we extend Barak et al.'s result to the case $p = 2 + \gamma$ for arbitrarily small γ, and we substantially simplify their proof.

Theorem 5 (Informal, see Theorem 24) *For any $p > 2$ and $q \in (2, p]$ and $c > s > s_0$ for some universal contant s_0, a polynomial time algorithm to distinguish the following two cases would yield a polynomial-time algorithm for the small-set expansion problem. Given $a_1, \ldots, a_n \in \mathbb{R}^d$, distinguish between: **yes:** there is a unit x that $\frac{1}{n} \sum_{i=1}^n |\langle a_i, x\rangle|^q > (cq)^{q/2}$, and **no:** for all unit x and $q \leq p$ it holds that $\frac{1}{n} \sum_{i=1}^n |\langle a_i, x\rangle|^q \leq (sq)^{q/2}$.*

1.2.2. Resilience

Another recently introduced assumption is that of *resilience:*

Definition 6 (Resilience, see Steinhardt (2018b)) *Let X be an $\mathbb{R}^d$-valued random variable with mean $\mathbb{E}\, X = \mu$. X is (σ, δ)*-resilient *in a norm $\| \cdot \|$ if for all events A with $\Pr A \geq 1 - \delta$, we have $\|\mathbb{E}\, X \mid A - \mu\| \leq \sigma$. Equivalently, X is (σ, δ)-resilient if for all events A with $\Pr A \leq \delta$, we have$\|\mathbb{E}\, X \mid A - \mu\| \leq \sigma \cdot \frac{1 - \Pr A}{\Pr A}$.*

In the remainder of the paper we will primarily consider the case where the norm $\| \cdot \|$ is the ℓ_2 norm in $\mathbb{R}^d$, since that is the setting in which our hardness results will apply. For the proof of equivalence, see Lemma 3 and Lemma 10 in Steinhardt et al. (2017).

It is not hard to show (see Corollary 32) that if D has second moments bounded by 1, then D is $(\sqrt{\delta}, \delta)$-resilient for all $\delta \leq 1/2$. Thus it might not be surprising that in this setting, we can achieve rates for robust mean estimation similar to those in Fact 1, at least inefficiently. However, there is already some asymptotic gap here between what is information-theoretically achievable and what is know to be achievable in polynomial time, since $(\sqrt{\delta}, \delta)$ resilience for a *fixed* δ is somewhat weaker than second moments bounded by 1.

Fact 3 (Steinhardt et al. (2017)) *Let $\mathcal{D}_\delta$ be the class of distributions over $\mathbb{R}^d$ which are $(\sqrt{\delta}, \delta)$-resilient. There exists an (exponential time) algorithm, which for all small-enough $\delta > 0$ and all $D \in \mathcal{D}_\delta$, given a δ-contaminated set of samples from D of size $\mathrm{poly}(d, 1/\delta)$, outputs $\widehat{\mu}$ which with probability at least $9/10$, satisfies $\|\widehat{\mu} - \mathbb{E}_{X\sim D} X\| \leq O(\sqrt{\delta})$. Furthermore, there is a polynomial-time algorithm which achieves $\|\widehat{\mu} - \mathbb{E}_{X\sim D} X\| \leq O(\sqrt{\delta \log(1/\delta)})$. Moreover, no estimator achieves error $\|\widehat{\mu} - \mathbb{E}_{X\sim D} X\| < c\sqrt{\delta}$ with probability at least $1/10$.*

A reasonable strengthening of this considers the condition that D is (σ, δ)-resilient for some $\sigma \ll \sqrt{\delta}$. The following basic fact about resilience shows that such assumptions suffice information-theoretically to achieve improved error rates.

Fact 4 (Steinhardt et al. (2017)) *There is an (inefficient) algorithm which given $\mathrm{poly}(d, 1/\delta)$ δ-contaminated samples from a (σ, δ)-resilient distribution D outputs $\widehat{\mu}$ such that with probability at least $9/10$ it holds that $\|\widehat{\mu} - \mathbb{E}_{X\sim D} X\| \leq O(\sigma)$.*

In particular, Fact 4 implies that if $\sigma \leq o(\sqrt{\delta})$, then it is information-theoretically possible to outperform even the exponential time algorithm from Fact 3. This leads to the question:

Question 7 *Is there a function $\sigma(\delta)$ and a polynomial-time algorithm which for all small-enough $\delta > 0$ given $\mathrm{poly}(d, 1/\delta)$ δ-contaminated samples from any (σ, δ)-resilient distribution D can find $\widehat{\mu}$ such that $\|\widehat{\mu} - \mathbb{E}_{X\sim D} X\| \leq o(\sqrt{\delta})$ with probability at least $9/10$?*

We prove two theorems suggesting a negative answer to Question 7. The first is in a similar spirit to Theorem 5. Existing algorithms (both efficient and inefficient) for robust mean estimation under resilience assumptions solve as a subroutine the problem of determining whether (the uniform distribution over) a set of samples is (σ, δ)-resilient. Thus, a potential route to design an algorithm for Question 7 is to improve existing guarantees for algorithms to check if a set of points is resilient. We show that such improvements would violate the small-set expansion hypothesis.

Theorem 8 (Informal, see Theorem 23) *For every sufficiently-small $s > 0$ there exists $\delta > 0$ such that an efficient algorithm for the following problem would yield an efficient algorithm for small set expansion: Given a set of points $a_1, \ldots, a_n \in \mathbb{R}^d$, distinguish between the cases **yes:** the uniform distribution on $\{a_1, \ldots, a_n\}$ is $(s\sqrt{\delta}, \delta)$ resilient, and **no:** it is not $(0.4\sqrt{\delta}, \delta)$-resilient.*

Our final theorem is the first in the literature to directly attack hardness for robust mean estimation via reduction from a worst-case complexity assumption, rather than reducing to related problems like certifying moment bounds or checking resilience as in Theorem 4, Theorem 5, and Theorem 8. We are able to show a negative answer to Question 7 under a strengthened small-set expansion hypothesis. Our strengthened version, which we call the *unique small-set expansion hypothesis* is as follows:

Hypothesis 9 (Unique Small-Set Expansion Hypothesis) *For every $\varepsilon > 0$ there exists $\delta > 0$ such that given a graph G, it is* NP*-hard to distinguish the following cases:* ***no*** *every set $S \subseteq [n]$ of δn vertices has expansion $\Phi_G(S) \geq 1 - \epsilon$, or* ***yes:*** *there exists a set $S \subseteq [n]$ of δn vertices in G such that $\Phi_G(S) \leq \epsilon$, and every other subset $T \subseteq [n]$ of δn vertices with $S \cap T = \varnothing$ has $\Phi_G(T) \geq 1 - \epsilon$.*

Here *unique* refers to the fact that in the **yes** case, the set S is the unique small nonexpanding set in G.[2] While we are not aware of this strengthening being considered previously in the literature, we also do not know any algorithmic techniques which could refute it. Hence we view the following theorem as at least a barrier to improving existing algorithms for robust mean estimation.

Theorem 10 (Informal, see Theorem 29) *If Question 7 has an affirmative answer then the Unique Small Set Expansion Hypothesis is false (or $P = NP$).*

It remains an interesting open problem to see if Theorem 10 can be strengthened to yield an algorithm for the (vanilla) small set expansion problem.

1.2.3. Spectral graph theory: Cheeger-style rounding for analytically sparse vectors

Our reductions involve spectral graph theory for small-set expanders, and one of our technical contributions is to substantially simplify current understanding of a simple structural question in spectral graph theory. This leads to the proofs of our main theorems, and answers an open question of Barak on simplification of the proof that the $2 \to 4$ norm is hard to approximate under the small-set expansion hypothesis (see Exercise 6.2 in Barak (2014)).

We review definitions formally in Section 2, but let us briefly recall some basics. For a regular n-node graph G and a set $S \subseteq [n]$, the expansion of S, denoted $\Phi_G(S)$, is the probability that a random walk initialized uniformly in S leaves it after one step. If we denote also by G the normalized adjacency matrix, then the expansion is $\Phi_G(S) = 1 - \langle 1_S, G1_S \rangle / |S|$. Of course, this makes sense only for indicator vectors 1_S of sets of vertices.

Cheeger's inequality extends the relationship between the quadratic form of G and expansion to other vectors. A consequence of Cheeger's inequality is the following fact:

Fact 5 (Consequence of Cheeger's inequality) *If v is any unit vector v where $\langle v, Gv \rangle \geq 1/2$ (and v is orthogonal to the all-1's vector), there is a level set S of the vector with $w_i = |v_i|$ with expansion $\Phi_G(S) \leq 0.99$.*

In the context of small-set expansion, it is important to detect the existence *small* sets of vertices – say, δn vertices for small constants δ – with expansion bounded away from 1. A key question is: what *analytical* properties of a vector v with $\langle v, Gv \rangle \geq 1/2$ give rise to a set of δn vertices S with expansion $\Phi_G(S) \leq 0.99$?

Barak et al. (2012a) showed that it is sufficient for v to be *analytically sparse*. In particular, they showed that if $\|v\|_4^4 \geq 1/\delta$ – that is, the 4-norm of v is similar to that of the (scaled) indicator vector of a set of size δn, then one may find a set of δn vertices in G with imperfect expansion. (Recall that sparse vectors, which are qualitatively similar to indicator vectors, have larger 4-norm than typical unit vectors.) One catch is that v must be completely contained in the span of eigenvectors of G of magnitude at least $1/2$, which is a stronger requirement than $\langle v, Gv \rangle \geq 1/2$.

[2] This use of "unique" should not be confused with Unique Games!

Theorem 11 (Consequence of Theorem 2.4 in Barak et al. (2012a)) *If there is v in the span of eigenvectors of G with eigenvalue at least $1/2$ such that $\|v\|_4^4 \geq 1/\delta n$, then G contains a set S of δn vertices having expansion $\Phi_G(S) \leq 1 - c$ for a universal $c > 0$. Furthermore, S may be found in polynomial time from G and v.*

While the vertex set S from this result can be found in polynomial time, Barak et al.'s procedure to find S from v is complex. In particular, it departs from the elegance of Cheeger's inequality that S can be taken to be a level set of v. Our tools give a simple proof of the following theorem, which we believe is novel – it directly characterizes the small set which can be recovered from v with large 4-norm in terms of level sets of v and the random walk on G.

Theorem 12 *If there is v in the span of eigenvectors of G with eigenvalue at least $1/2$ such that $\|v\|_4^4 \geq 1/\delta n$, then there is a level set S of the vector w defined by $w_i = |v_i|$ which has the following property. For some $t \leq O(\log n)$ there is level set of $G^t 1_S + G^{t+1} 1_S$ of size at most $O(\delta \log(1/\delta) \cdot n)$ having expansion $\Phi_G(S) \leq 1 - c$ for a universal $c > 0$. Here, 1_S is the $0/1$ indicator vector for the set S.*

Qualitatively, our theorem says that an analytically-sparse v in the high eigenspaces of G has a level set S such that if the random walk on G is initialized to the uniform distribution on S, eventually the random walk "discovers" a small cut of imperfect expansion. Thus, at the cost of a factor $\log(1/\delta)$ in the size of S as compared to the result of Barak et al., we recover some of the elegance of Cheeger's rounding procedure for turning v into a cut. We describe the proof of Theorem 12 in Appendix F.

2. Preliminaries

2.1. Spectral graph theory

Let $G = (V, E)$ be an n-node graph. We also denote by G the stochastic $n \times n$ random walk matrix associated to the graph G.

Definition 13 (Isotropic spectral embedding) *Let $\Pi_{1/2} \in \mathbb{R}^{n \times n}$ be the projector to the span of eigenvectors of G with eigenvalues at least $1/2$. Let A be a matrix such that $AA^\top = \Pi_{1/2}$. Without loss of generality, take the first column of A to be $\mathbf{1}/\sqrt{n}$, the (scaled) all-1s vector.*

Let $a_1, \ldots, a_n$ be the rows of A. We say that $(a_1, \ldots, a_n)$ is the spectral embedding of G, and if $b_i = \sqrt{n} a_i$ we say that $(b_1, \ldots, b_n)$ is the isotropic spectral embedding of G.

We will need the following basic facts; the proofs are elementary and omitted.

Fact 6 (Mean of a spectral embedding) *Let G be a graph and let $\Pi_{1/2}$ be the projector to the span of eigenvectors of G of eigenvalue at least $1/2$. Let $a_1, \ldots, a_n$ be the rows of the matrix A where $AA^\top = \Pi_{1/2}$; without loss of generality assume the first column of A is the vector $\frac{1}{\sqrt{n}} \cdot \mathbf{1}$. Then $\frac{1}{n} \sum a_i = (1/\sqrt{n}, 0, 0, \ldots, 0)$.*

Fact 7 $\mathbb{E}_{i \sim [n]} b_i b_i^\top = I$.

For any $S \subseteq V$, we denote by $\mathbf{1}_S \in \{0,1\}^n$ the $0/1$ indicator vector of S. For $v, w \in \mathbb{R}^n$ we often employ the usual Euclidean inner product $\langle v, w \rangle = \sum_{i \leq n} v_i w_i$.

If $S \subseteq V$ is a subset of vertices in G, its expansion is the probability that a random walk initialized inside S leaves S in one step: $\Phi_G(S) = 1 - \frac{1}{|S|} \cdot \langle \mathbf{1}_S, G \mathbf{1}_S \rangle$. We define the *expansion profile* of a graph S: for every $\delta > 0$, let $\Phi_G(\delta) = \inf_{|S| = \delta n} \Phi_G(S)$. We also let $\Phi_G^{\leq}(\delta) = \inf_{|S| \leq \delta n} \Phi_G(S)$ be a slightly modified version of expansion profile which takes into account all sets of size at most δn, rather than exactly δn.

A consequence of Lemma 31 is a local Cheeger inequality concerning the quadratic form $\langle f, G^2 f \rangle = \|Gf\|^2$ rather than $\langle f, Gf \rangle$. The proof is standard – see the appendix.[3]

Lemma 14 *Let G be an n-node regular graph, and let $\varepsilon, \delta, \gamma$ be so that $0 < \delta \leq \gamma$ and $\varepsilon > 0$. Let $f \in \mathbb{R}^n$ have nonnegative coordinates, and suppose that $\|Gf\|^2 \geq \varepsilon \|f\|^2$ and $\|f\|^2 \geq \frac{\gamma \|f\|_1^2}{\delta n}$. There is a level set S of the function $g = f + Gf$ with size at most δn and expansion $\Phi_G(S) \leq 1 - \Omega(\gamma \varepsilon^4)$.*

We will also require the following slight modification to Lemma 14, which states that in the special case of f being an indicator function for a subset, then we may additionally assume that the level set with poor expansion is additionally not too small. We are not aware of a black-box proof of Lemma 15 from Lemma 31, but our proof is a modification of the proof of Lemma 31 found in Steurer (2010b). For completeness we prove this lemma in the appendix.

Lemma 15 *There exist universal constants $0 < c < C$ such that the following holds. For every G an n-node regular graph and every small enough $\varepsilon, \delta, \eta$, if $S \subset [n]$ has $|S| = \delta n$ and $f = \mathbf{1}_S$ has $\|Gf\|^2 \geq \varepsilon \|f\|^2$, then there is a level set T of the function $g = (1-\eta) f + \eta G f$ with size $|T| \in \left[c \eta \varepsilon^2 \delta n, C \frac{\delta n}{\eta^2 \varepsilon^2} \right]$ and expansion $\Phi_G(T) \leq 1 - \Omega(\eta^2 \varepsilon^2)$. Moreover, if there is $R \subseteq [n]$ with $\Phi_G(R) \leq \varepsilon / 100$ and $|R| = \delta n$, and if $S \cap R = \varnothing$, then also T exists satisfying the previous properties and having $T \cap R = \varnothing$.*

2.2. Small-Set Expansion Hypotheses

Our reductions in this paper are from small-set expansion problems, which are conjectured to be computationally difficult to solve. At a high level, these assumptions say that it is hard to verify whether or not there exists a small set in a graph which does not expand well into the rest of the graph. There are two canonical versions of this Small-Set Expansion Hypothesis (SSEH) which the literature appears to consider interchangeable. However, for us it will be important to distinguish between the two. The first (and original) version of SSEH concerns $\Phi_G(\delta)$:

Hypothesis 16 ($=$-Small-Set Expansion Hypothesis ($\mathsf{SSEH}_=$) Raghavendra and Steurer (2010)) *For every constant $\epsilon > 0$ there is a small-enough $\delta > 0$ such that the following problem is NP-hard. Given a graph G, distinguish between $\Phi_G(\delta) \geq 1 - \epsilon$ and $\Phi_G(\delta) \leq \epsilon$.*

In particular, this statement is only about sets of size exactly δn. The second version of SSEH is essentially identical, except using $\Phi_G^{\leq}(\delta)$ instead of $\Phi_G(\delta)$.

[3] Theorem 2.1 in Steurer (2010a) is identical to Lemma 14 but is stated with the conclusion $\Phi_G(S) \leq 1 - \Omega(\varepsilon^2)$ rather than $\Phi_G(S) \leq 1 - \Omega(\varepsilon^4)$; however the only proof we are aware of appears to require the extra factor of $1/\varepsilon^2$. Generally ε is taken to be a tiny constant, so the difference is just one of constant factors.

Hypothesis 17 ($\leq$-Small-Set Expansion Hypothesis ($\mathsf{SSEH}_{\leq}$)) *For every constant $\epsilon > 0$ there is a small-enough $\delta > 0$ such that the following problem is NP-hard. Given a graph G, distinguish between $\Phi_G^{\leq}(\delta) \geq 1 - \epsilon$ and $\Phi_G^{\leq}(\delta) \leq \epsilon$.*

We are not aware of any equivalences or implications between these two (apparently very similar) problems. However, both versions of the problem have been widely used and called the "Small-Set Expansion Hypothesis" in the literature, see e.g. Barak et al. (2012a).

We remark that while these two problems are very similar, there do appear to be some subtle qualitative differences between them. In particular, in the context of this paper, $\mathsf{SSEH}_{=}$ (and variants thereof) implies hardness for problems related to resilience, whereas $\mathsf{SSEH}_{\leq}$ implies hardness for problems related to bounded moments. At a high level, this is because bounded moments is equivalent to resilience at every scale (see Corollary 32), and thus to control moments, we need to know what occurs at all sets of size at most δ, not just in a neighborhood around δ.

Acknowledgments

We thank Prasad Raghavendra for numerous helpful discussion regarding this paper, and in particular for the suggestion to use random walks to prove the upper bound in Section 3. We also thank David Steurer for answering emergency questions about small-set expansion.

References

Frank J Anscombe. Rejection of outliers. *Technometrics*, 2(2):123–146, 1960.

Benny Applebaum, Boaz Barak, and David Xiao. On basing lower-bounds for learning on worst-case assumptions. In *FOCS*, pages 211–220. IEEE Computer Society, 2008.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017.

Boaz Barak. Sum of squares upper bounds, lower bounds, and open questions. 2014. URL `https://www.boazbarak.org/sos/prev/files/all-notes.pdf`.

Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *STOC*, pages 307–326. ACM, 2012a.

Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. *CoRR*, abs/1205.4484, 2012b.

Thorsten Bernholt. Robust estimators are hard to compute. Technical report, Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in ?, 2006.

Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*, 2018.

Mark Bun and Mark Zhandry. Order-revealing encryption and the hardness of private learning. In *Theory of Cryptography Conference*, pages 176–206. Springer, 2016.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *STOC*, pages 47–60. ACM, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *FOCS*, pages 655–664. IEEE Computer Society, 2016.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 999–1008. PMLR, 2017a.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008, 2017b.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 73–84. IEEE, 2017c.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060. ACM, 2018.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

Vitaly Feldman and Varun Kanade. Computational bounds on statistical query learning. In *Conference on Learning Theory*, pages 16–1, 2012.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 563–574. IEEE, 2006.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM J. Comput.*, 39(2):742–765, 2009.

Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pages 354–375, 2013.

Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. ACM, 2018.

Peter J Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, 35 (1):73–101, 1964.

David S Johnson and Franco P Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

William B Johnson, Gideon Schechtman, and Joel Zinn. Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *The Annals of Probability*, pages 234–253, 1985.

Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 273–282. ACM, 1994.

Adam R. Klivans and Pravesh Kothari. Embedding hard learning problems into gaussian space. In *APPROX-RANDOM*, volume 28 of *LIPIcs*, pages 793–809. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2014.

Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, 2018.

Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *FOCS*, pages 665–674. IEEE Computer Society, 2016.

Jerry Li. *Principled Approaches to Robust Machine Learning and Beyond.* 2018.

Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *STOC*, pages 755–764. ACM, 2010.

Prasad Raghavendra, David Steurer, and Madhur Tulsiani. Reductions between expansion problems. In *IEEE Conference on Computational Complexity*, pages 64–73. IEEE Computer Society, 2012.

Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *J. ACM*, 56 (6):34:1–34:40, 2009.

Jacob Steinhardt. Talk at stoc 2018 workshop on computational phase transitions. 2018a.

Jacob Steinhardt. *Robust Learning: Information Theory and Algorithms.* Stanford University, Thesis, 2018b.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*, 2017.

David Steurer. Subexponential algorithms for d-to-1 two-prover games and for certifying almost perfect expansion. *Available at the authors website*, pages 2–1, 2010a.

David Steurer. *On the complexity of unique games and graph expansion*. Citeseer, 2010b.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

J.W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

Appendix A. Conditional Means of Small Sets in the Spectral Embedding

In this section we prove the following two key lemmas, which characterize the spectral embeddings of small sets of vertices in small-set expanders. They suggest the following perspective on embeddings of small-set expanders, which is at the heart of all our arguments: if G is a (δ, ϵ)-small-set expander, small sets of vectors in its spectral embedding cannot have average too far from the origin, while a small nonexpanding set in G embeds to a set of vectors whose average is far from the origin.

Slightly more formally, for every $S \subseteq [n]$ with $|S| \leq \delta n$, if $\Phi_G(S) \geq 1 - \varepsilon$, then

$$\left\| \tfrac{1}{|S|} \sum_{i \in S} b_i \right\| \approx \varepsilon^{\Omega(1)} / \sqrt{\delta} \,.$$

(At least, if S is not too small.) On the other hand, if $\Phi_G(S) \leq \varepsilon$, then

$$\left\| \tfrac{1}{|S|} \sum_{i \in S} b_i \right\| \approx 1/\sqrt{\delta} \gg \varepsilon^{\Omega(1)} / \sqrt{\delta} \,.$$

Now we make this formal. The first lemma shows that a small non-expanding set in a graph G has a spectral embedding far from the origin. It has been observed several times before (see e.g. Barak et al. (2012a)). We include the proof in the Appendix for completeness.

Lemma 18 *Suppose G is an n-node graph. Let $b_1, \ldots, b_n$ be the isotropic spectral embedding of G. Then, every $T \subseteq [n]$ satisfies*

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|^2 \geq \frac{n}{|T|} \cdot \left(\frac{1}{2} - \Phi_G(T) \right) \,.$$

The second lemma shows that if G is a small-set expander then every small set of vectors in its spectral embedding has mean near the origin. By correctly setting parameters, something qualitatively similar would follow as a corollary of Theorem 2.4 in Barak et al. (2012a), but our proof is much simpler than that route. We show that for such a set T, if $\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|$ were too large, then eventually the random walk on G, initialized to the uniform distribution on T, would find a small set with small expansion.

Lemma 19 *Let G be an n-node graph. Suppose ε, δ are such that $\Phi_G^{\leq}(\delta) \geq 1 - \varepsilon$, and $\varepsilon < \varepsilon_0$ for some universal constant $\varepsilon_0 > 0$. Let $b_1, \ldots, b_n$ be the isotropic spectral embedding of G. For every $T \subseteq [n]$ with $|T| \leq \delta n$,*

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\| \leq C' \exp \left(C \cdot \frac{\log(\delta n / |T|)}{\log(1/\varepsilon)} \right) \cdot \frac{\varepsilon^{1/10}}{\sqrt{\delta}} \,,$$

where $C', C > 0$ are universal constants.

Proof Up to scaling, $\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|$ is the magnitude of the projection of the uniform probability distribution $\mathbf{1}_T/|T|$ on T into the span of eigenvectors of G with eigenvalue at least $1/2$. We first argue that this magnitude is not affected by too much if we replace $\mathbf{1}_T/|T|$ with $G^t \cdot \mathbf{1}_T/|T|$, which is the probability distribution which results from running the random walk in G for t steps.

To see this, first express $\frac{1}{|T|} \sum_{i \in T} b_i$ in terms of the indicator vector $\mathbf{1}_T \in \{0,1\}^n$:

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|^2 = \frac{n}{|T|^2} \cdot \left\| A^\top \mathbf{1}_T \right\|^2 = \frac{n}{|T|^2} \sum_{i \,:\, \lambda_i \geq 1/2} \langle v_i, \mathbf{1}_T \rangle^2$$

where the columns of A are the eigenvectors v_i of G with eigenvalue λ_i at least $1/2$. For any $t \in \mathbb{N}$, note that

$$\left\| A^\top G^t \mathbf{1}_T \right\|^2 = \sum_{i \,:\, \lambda_i \geq 1/2} \langle G^t v_i, \mathbf{1}_T \rangle^2 = \sum_{i \,:\, \lambda_i \geq 1/2} \lambda_i^t \langle v_i, \mathbf{1}_T \rangle^2 \geq 2^{-t} \left\| A^\top \mathbf{1}_T \right\|^2 .$$

Our aim is to use the local Cheeger inequality to control $\left\| A^\top G^t \mathbf{1}_T \right\|^2$, which will be possible so long as the collision probability of $G^t \mathbf{1}_T$ is like that of the uniform distribution on a set of size at most δn. First, since $\Pi_{1/2} \preceq I$, we have $\left\| A^\top G^t \mathbf{1}_T \right\|^2 \leq \left\| G^t \mathbf{1}_T \right\|^2$.

By the local Cheeger inequality (Lemma 14) with $\gamma = \varepsilon^{0.1}$, there is a constant C such that for every t, either $\|G(G^t \mathbf{1}_T)\|^2 < C\varepsilon^{0.1} \|G^t \mathbf{1}_T\|^2$ or $\|G^t \mathbf{1}_T\|^2 < \varepsilon^{0.2} \|G^t \mathbf{1}_T\|_1^2 / \delta n = \varepsilon^{0.1} |T|^2/\delta n$. (Otherwise the assumption $\Phi_G(\delta) \geq 1 - \epsilon$ is violated.) The last equality follows because $\|\mathbf{1}_T\|_1 = |T|$ and G preserves 1-norms and nonnegativity of nonegative vectors.

Pick t to be the smallest integer such that the second alternative holds; i.e. $\|G^t \mathbf{1}_T\|^2 < \varepsilon^{0.2} |T|^2/\delta n$. (Such t must exist because for smaller t and small enough ε the norm $\|G^t \mathbf{1}_T\|^2$ strictly decreases in each step of the random walk.) Then putting together our previous bounds,

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|^2 \leq \frac{\varepsilon^{0.2} n}{|T|^2} \cdot 2^t \cdot \|G^t \mathbf{1}_T\|^2 \leq 2^t \cdot \frac{\varepsilon^{0.2}}{\delta} .$$

We just need to bound t, the smallest integer such that $\|G^t \mathbf{1}_T\|^2 < \varepsilon^{0.2} |T|^2/\delta n$. If ε is small enough, for every $t' < t$ we know that $\|G^{t'} \mathbf{1}_T\|$ is decreasing; in particular $\|G^{t'+1} \mathbf{1}_T\| < C\varepsilon^{0.1} \|G^{t'} \mathbf{1}_T\|$. Since $\|1_T\|^2 = |T|$, the number t just has to be large enough that $\varepsilon|T|/\delta n \geq (C\varepsilon^{0.1})^t$, which rearranges to $t \geq \frac{\log(\delta n/|T|)}{\log(C/\varepsilon^{0.1})}$. Putting it together, we find

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|^2 \leq \exp\left(C_1 \cdot \frac{\log(\delta n/|T|)}{\log 1/\varepsilon} \right) \cdot \frac{\varepsilon}{2\delta} \leq C_2 \exp\left(C_1 \cdot \frac{\log(\delta n/|T|)}{\log 1/\varepsilon} \right) \frac{\varepsilon^{0.2}}{\delta}$$

for some universal $C_1, C_2 \geq 0$. ■

Proving our main theorems from Lemma 18 and Lemma 19 We briefly describe how all our main results can be obtained using the preceding two lemmas and related ideas. To prove Theorem 8 on hardness of checking resilience of a set of points in $\mathbb{R}^d$, we take the set of points to be the spectral embedding of a graph G. Then if G is a small-set expander, one may see that no tail event in the uniform distribution over the embedding – that is, no small set of vectors – can deviate far from the origin, by Lemma 19.[4] On the other hand, if G has a small non-expanding set, resilience is immediately violated by applying Lemma 18.[5]

To prove Theorem 5, we again take the vectors $a_1, \dots, a_n$ in the theorem statement to be the embedding of a graph G. Lemma 19 leads to tail bounds for the uniform distribution over these vectors, which can then be translated into upper bounds on the moments of the distribution by Fact 9 in the case that G is a small-set expander. On the other hand, if G has a small non-expanding set then Lemma 18 can be leveraged to prove lower bounds on the p-th moments of the uniform distribution on $a_1, \dots, a_n$ for $p > 2$. We then combine this with an averaging argument to gain better control over even more moments of the distribution when the graph is a small-set expander, while arguing that this averaging does not decrease the p-th moment in the presence of a small non-expanding set.

The proof of Theorem 10 is similar, with one key difficulty. To arrive at the end of the reduction in the setting of robust mean estimation under resilience, there must be a set of adversarially corrupted points, but the remaining points must be resilient. This is where we critically leverage our strengthened small-set expansion hypothesis. We strengthen the hypothesis in the following way: we suppose that small-set expansion remains hard if in one case we are promised that G contains one small set S with $\Phi_G(S) \leq \varepsilon$ but for all other T with $|T| = \delta n$ and $T \cap S = \varnothing$ it holds that $\Phi_G(T) \geq 1 - \varepsilon$. The resulting control over deviations of small sets *in the embedding of* $[n] \setminus S$, via local Cheeger inequalities adapted to account for the presence of the set S, allows us to show that the embedding of $[n] \setminus S$ is resilient.

Appendix B. Hardness of Certifying Conditions for Robust Mean Estimation

In this section we show that it is SSE-hard to decide whether a set of points satisfy resilience or bounded moments beyond the $\sqrt{\delta}$ barrier. In particular, in this regime improved certification algorithms would likely lead to improved polynomial-time error rates for robust mean estimation under bounded moment or resilience assumptions.

Throughout this section, given an instance G of SSE, as in Section 3, we will let $\Pi_{1/2}$ be the projector to the span of eigenvectors of G having eigenvalue at least $1/2$, and we let $b_1, \dots, b_n$ be the isotropic spectral embedding of G.

B.1. Consequences of SSE

To prove our hardness from SSE, we will actually reduce from the following more quantitative problems, which are known to be polynomial-time equivalent to SSE.

[4] In reality, we must use a version of Lemma 19 which applies to Φ_G rather than $\Phi_G^{\leq}$ and takes only one step of the walk – this lemma is really just the local Cheeger inequality. See Section A.

[5] Actually, this is true only if the set has size $\Omega(\delta n)$, rather than perhaps having size, say, $\sqrt{n}$. This is why to prove hardness of resilience we need to start $\mathsf{SSEH}_=$.

B.1.1. Gap SSE

The first allows us to go from $\mathsf{SSEH}_=$ to assuming control over all sets of size in some constant size window around δn. In particular, consider the following variant of SSEH:

Hypothesis 20 (Gap =-Small-Set Expansion Hypothesis ($\mathsf{Gap\text{-}SSEH}_=$) Raghavendra et al. (2012)) *For all small-enough $\varepsilon > 0$ and $M \geq 1$, there exists a small-enough $\delta \leq 1/M$ so that the following problem is NP hard. Given a graph G on n vertices, distinguish between:*

yes: *There exists a non-expanding set $S \subseteq [n]$ with $|S| = \delta n$ an $\Phi_G(S) \leq \varepsilon$.*
no: *All sets $S \subseteq V$ with $|S| \in \left[\frac{\delta n}{M}, M\delta n\right]$ have $\Phi_G(S) \geq 1 - \varepsilon$.*

Then it is known that this problem is equivalent to $\mathsf{SSEH}_=$:

Proposition 21 (Raghavendra et al. (2012)) *$\mathsf{SSEH}_=$ holds if and only if $\mathsf{Gap\text{-}SSEH}_=$ holds.*

B.1.2. Quantitative SSE

It has been shown that in SSE, quantitative relationships between the parameters ε, δ may be taken. Specifically, Raghavendra et al. (2012) shows:

Proposition 22 (Raghavendra et al. (2012), Theorem 3.5) *For every sufficiently small $\delta, \varepsilon, \gamma > 0$ the following problem is NP-hard assuming $\mathsf{SSEH}_\leq$: Given a graph G, distinguish between:*

yes: *$\Phi_G^{\leq}(\delta) \leq \varepsilon$.*
no: *For all $\delta' \in [0,1]$ it holds that $\Phi_G^{\leq}(\delta') \geq 1 - (\delta')^{\Omega(\varepsilon)} - \gamma/\delta'$.*

B.2. Hardness of certifying resilience

In this section, we prove the following theorem.

Theorem 23 *Under $\mathsf{SSEH}_=$, for all sufficiently small constants $s > 0$ there exists $\delta(s) > 0$ such that given $S \subseteq \mathbb{R}^d$ it is NP-hard to distinguish between:*

yes: *the uniform distribution X on S is $(s\sqrt{\delta}, \delta)$-resilient.*
no: *there is an event A in the uniform distribution on S such that $\Pr(A) = \delta$ and $\|\mathbb{E}\, X \mid A - \mathbb{E}\, X\| > 0.4 \cdot \sqrt{\delta} \cdot \frac{1 - \Pr A}{\Pr A}$.*

Proof [Proof of Theorem 23] Let $\varepsilon > 0$ be sufficiently small. We start with an instance G of $\mathsf{Gap\text{-}SSEH}_=$ with parameter ε, $M \geq 1$ to be set later, and corresponding $\delta = \delta(\varepsilon, M)$. Our reduction is simple: we let the set S be $S = \{b_i\}_{i=1}^n$. Observe that an event E in the uniform distribution supported S directly corresponds to a subset $T \subseteq S$, and moreover $\Pr E = |T|/n$. We verify that an efficient algorithm certifying $(s\sqrt{\delta}, \delta)$-resilience of S, for some $s = s(\varepsilon)$, would solve SSE. (We will show $s = \Theta(\varepsilon^{1/8})$ suffices.)

There are two cases to check. Suppose there exists a set $T \subset [n]$ with $|T| = \delta n$, so that $\Phi_G(T) \leq \varepsilon$. Let A be the event associated to that set. Then by Lemma 18, we have

$$\begin{aligned} \|\mathbb{E}\, X|A\|^2 &= \left\| \frac{1}{T} \sum_{i \in T} b_i \right\|^2 \\ &\geq \frac{1/2 - \varepsilon}{\Pr A}, \end{aligned}$$

and so in particular since $\Pr A = \delta$, we have $\|\mathbb{E}\, X|A\| \geq \sqrt{1/2 - \varepsilon} \cdot \sqrt{\delta} \cdot \frac{1}{\Pr A} \geq \sqrt{\delta} \cdot \frac{1}{2 \Pr A}$, for ε sufficiently small. Since $\mathbb{E}\, X = e_1$ and therefore $\|\mathbb{E}\, X\| = 1 \ll \|\mathbb{E}\, X|A\|$ for δ small, in this case the resulting set S is in the **no** case for resilience.

We now check the other case. Suppose $\Phi_G(S) \geq 1 - \varepsilon$ for all S with $|S| \in [\delta n/M, M\delta n]$. We wish to verify that in this case the resulting distribution is in the **yes** case for resilience. First, observe that for any set $T \subseteq [n]$ with associated event E, we have the bound

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|^2 = n \cdot \left\| A^\top \frac{\mathbf{1}_T}{|T|} \right\|^2 \overset{(a)}{\leq} \frac{n}{|T|} \leq \frac{1}{\Pr E} \;,$$

where (a) follows since A has spectral norm at most 1. (Here A is the matrix such that $AA^\top = \Pi_{1/2}$.) Let r be a constant to be optimized later. If $\Pr E < r\delta$, then immediately $\|\mathbb{E}\, X|E\| \leq \frac{\sqrt{r\delta}}{\Pr E}$. On the other hand, if $\Pr E \in [r\delta, \delta]$, then if T is the associated set, we must have

$$\begin{aligned} \left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|^2 &= n \left\| A^\top \frac{\mathbf{1}_T}{|T|} \right\|^2 \\ &\overset{(a)}{\leq} 4n \left\| G \frac{\mathbf{1}_T}{|T|} \right\|^2 \\ &\overset{(b)}{\leq} \frac{4n\sqrt{\varepsilon}}{r} \left\| \frac{\mathbf{1}_T}{|T|} \right\|^2 \\ &= \frac{4\sqrt{\varepsilon}}{r \Pr E} \;, \end{aligned}$$

where (a) follows since $AA^\top = \Pi_{1/2} \preceq 4GG^\top$, and (b) follows since if we let $M = O\left(\frac{1}{r^2 \varepsilon^2}\right)$, then this follows from Lemma 15 with $\eta = r$ (as otherwise we would witness a set with size in $[\delta n/M, M\delta n]$ with $\Phi_G(S) < 1 - \varepsilon$). As a result, we have

$$\|\mathbb{E}\, X|E\| \leq \frac{2\varepsilon^{1/4}}{r^{1/2}\sqrt{\Pr E}} \leq \frac{2\varepsilon^{1/4}\sqrt{\delta}}{r^{1/2} \Pr E} \;.$$

Thus, if we let $r = \varepsilon^{1/4}$, then in all cases, we have

$$\|\mathbb{E}\, X|E\| \leq \frac{2\varepsilon^{1/8}\sqrt{\delta}}{\Pr E} \;.$$

Since again $\|\mathbb{E}\, X\| = 1$, this implies that for δ sufficiently small, S is $(s\sqrt{\delta}, \delta)$-resilient, for $s = \Theta(\varepsilon^{1/8})$. Thus we are in the **yes** case for resilience. Our choice of $s = \Theta(\varepsilon^{1/8})$ ensures that $\mathsf{SSEH}_=$ applies for all small-enough ε and hence for all small-enough s. This completes the proof of correctness of the reduction. ■

B.3. Hardness of certifying bounded moments

This section is dedicated to the proof of the following theorem:

Theorem 24 *Under $\mathsf{SSEH}_{\leq}$, there exists a constant $s > 0$ such that for any $q > 2$, $c > s$, $t \in (2, q]$, given $S \subseteq \mathbb{R}^d$ it is NP-hard to distinguish the cases:*

> ***yes:*** *the uniform distribution X on S satisfies*
>
> $$\sup_{\|v\|=1} |\langle v, X - \mathbb{E} X\rangle|^r \leq (sr)^{r/2} ,$$
>
> *for all $2 < r \leq q$.*
> ***no:*** *there exists a unit vector $v \in \mathbb{R}^d$ so that*
>
> $$\sup_{\|v\|=1} |\langle v, X - \mathbb{E} X\rangle|^t > (ct)^{t/2} .$$

We first show that the following intermediate problem is NP-hard under $\mathsf{SSEH}_{\leq}$:

Lemma 25 *There exists a universal constant $c \in [0, 1]$ such that for all $q > 2$ and all small-enough δ the following problem is NP-hard assuming $\mathsf{SSEH}_{\leq}$. Given a set S of n points in $\mathbb{R}^d$ so that the uniform distribution X over S is isotropic, distinguish between:*

> **yes:** *There exists an event E with probability $\Pr E \leq \delta$ so that $\|\mathbb{E} X|E - \mathbb{E} X\| \geq \frac{0.4}{\sqrt{\Pr E}}$. In particular, by Fact 9, this implies $\mathbb{E} |\langle v, X\rangle|^r > \frac{0.4^r}{\delta^{r/2-1}}$ for some unit vector v, and any $r > 2$.*
> **no:** $\mathbb{E} |\langle v, X\rangle|^r \leq \frac{\delta^{cr}}{\delta^{r/2-1}}$ *for all unit vectors v and all $r \in (2, q]$.*

We remark that this lemma (with different terminology) is very similar to the reduction presented in Barak et al. (2012b), in their proof that SSE implies hardness for certifying $2 \to 4$ norms of tensors. We give a proof here which simplifies and generalizes several key steps in their argument, and which gives us stronger guarantees which will be useful later.

The proof requires some bookkeeping, but the approach is simple. We will take S to be the isotropic spectral embedding of a graph G. The **yes** case is easy to establish For the **no** case, we first observe (Fact 9) that moment bounds of the type in Lemma 25 are essentially equivalent to large-deviation tail bounds – i.e. inequalities of the form $\Pr(\langle X, v\rangle > t) \leq p(t)$ for unit vectors v and various deviation magnitudes t. We obtain such deviation inequalities from Lemma 19, which shows that no small set of vectors in the spectral embedding of a small-set expander can deviate far from the origin.

Proof [Proof of Lemma 25] Fix $q > 2$. Let G be an instance of the problem given in Proposition 22 on n vertices with $\delta < 0.05$ and $\epsilon < 0.05$ sufficiently small that Proposition 22 applies, and $\gamma = \delta^{1+\varepsilon}$. Let b_i be the isotropic spectral embedding of G. Let $S = \{b_i\}_{i=1}^n$. We now verify that this set achieves the desired properties.

First, suppose that there exists a set $T \subseteq [n]$ with $|T| \leq \delta$ so that $\Phi_G(T) \leq \varepsilon$. Then, by Lemma 18, we have

$$\|\mathbb{E} X|E\|^2 \geq \frac{0.5 - \varepsilon}{\Pr E} ,$$

and so since $\|\mathbb{E} X\|^2 = 1 \ll \frac{0.45}{\Pr E}$, we have $\|\mathbb{E} X|E - \mathbb{E} X\| \geq \frac{0.4}{\sqrt{\Pr E}}$, for $\delta < 0.05$. Thus, in this case the set S belongs to the **yes** case of Proposition 22.

On the other hand, suppose that $\Phi_G^{\leq}(\delta') \geq 1 - (\delta')^{\Omega(\varepsilon)} - \gamma/\delta'$ for all $\delta' \in [0,1]$. Fix any $r \in (2, q]$. Our goal will be to use Fact 9, which for any $s > r$ supplies the following bound on $\mathbb{E}\, |\langle v, X\rangle - \mathbb{E}\langle v, X\rangle|^r$ for any unit v (by elementary integration):

$$\mathbb{E}\, |\langle v, X\rangle - \mathbb{E}\langle v, X\rangle|^r \leq \sup_E\, (2 \Pr E)^{r/s} \cdot |\, \mathbb{E}\langle X, v\rangle | E - \mathbb{E}\langle X, v\rangle|^r \cdot \frac{s}{s-r}\,,$$

where the supremum is over all events E.

By Cauchy-Schwarz, for any unit v and event E,

$$|\, \mathbb{E}\langle v, X\rangle | E - \mathbb{E}\langle v, X\rangle| \leq \|\, \mathbb{E}\, X | E - \mathbb{E}\, X\|\,. \tag{3}$$

So,

$$\mathbb{E}\, |\langle v, X\rangle - \mathbb{E}\langle v, X\rangle|^r \leq \sup_E\, (2 \Pr E)^{r/s} \cdot \|\, \mathbb{E}\, X | E - \mathbb{E}\, X\|^r \cdot \frac{s}{s-r}\,. \tag{4}$$

We choose $s = r \cdot \frac{\log(1/\delta)}{\log(1/\delta) - 1}$, so that $s/(s-r) = \log(1/\delta)$.

We will bound the supremum in (4) by separately considering two cases: $\Pr E \leq \delta/2$ and $\Pr E > \delta/2$. First, let E have $\Pr E \leq \delta/2$. By our choice of γ, we know that

$$\Phi_G(\delta) \geq 1 - \delta^{\Omega(\varepsilon)}\,.$$

Using this in conjunction with Lemma 19, we know that there exist universal constants $C, C' > 0$ so that

$$\|\, \mathbb{E}\, X | E\| \leq C' \left(\frac{\delta}{\Pr(E)}\right)^{C/\varepsilon \log(1/\delta)} \cdot \frac{\delta^{\Omega(\varepsilon)}}{\sqrt{\delta}}$$

and hence by triangle inequality

$$\|\, \mathbb{E}\, X | E - \mathbb{E}\, X\| \leq C' \left(\frac{\delta}{\Pr(E)}\right)^{C/\varepsilon \log(1/\delta)} \cdot \frac{\delta^{\Omega(\varepsilon)}}{\sqrt{\delta}} + 1$$

because $\|\, \mathbb{E}\, X\| = 1$. For small-enough δ, we have $C' \left(\frac{\delta}{\Pr(E)}\right)^{C/\varepsilon \log(1/\delta)} \cdot \frac{\delta^{\Omega(\varepsilon)}}{\sqrt{\delta}} \geq 1$ for all E, and hence

$$\|\, \mathbb{E}\, X | E - \mathbb{E}\, X\| \leq 2C' \left(\frac{\delta}{\Pr(E)}\right)^{C/\varepsilon \log(1/\delta)} \cdot \frac{\delta^{\Omega(\varepsilon)}}{\sqrt{\delta}}\,.$$

Returning to the expression from (4),

$$(2 \Pr E)^{r/s} \cdot \|\, \mathbb{E}\, X | E - \mathbb{E}\, X\|^r \cdot \frac{s}{s-r} \leq (2 \Pr E)^{r/s} \cdot \left(\frac{2C' \delta^{\Omega(\varepsilon)}}{\sqrt{\delta}}\right)^r \cdot \left(\frac{\delta}{\Pr(E)}\right)^{Cr/\varepsilon \log(1/\delta)} \cdot \log\frac{1}{\delta}\,.$$

By elementary algebra, using our choice of s and the bound $\Pr E \leq \delta/2$, so long as $\log(1/\delta) > Cr/\varepsilon + 1$, we have

$$(2 \Pr E)^{r/s} \left(\frac{\delta}{\Pr(E)}\right)^{Cr/\varepsilon \log(1/\delta)} \leq \delta^{1 - 1/\log(1/\delta)} \leq O(\delta)\,.$$

So all together we got

$$(2\Pr E)^{r/s} \cdot \|\mathbb{E}\, X|E - \mathbb{E}\, X\|^r \cdot \frac{s}{s-r} \leq C \cdot \frac{(2C'\delta)^{\Omega(\varepsilon)\cdot r}}{\delta^{r/2-1}} \cdot \log\frac{1}{\delta} \,.$$

for some (different) universal constant C. For some universal c_1, c_2 if we choose $\eta = c_1 \delta^{c_2 \varepsilon}$, then for every small-enough δ,

$$(2\Pr E)^{r/s} \cdot \|\mathbb{E}\, X|E - \mathbb{E}\, X\|^r \cdot \frac{s}{s-r} \leq \delta \cdot \left(\frac{\eta}{\delta}\right)^{r/2} \,.$$

We turn to the case of $\Pr E > \delta/2$. By hypothesis, $\Phi_G^{\leq}(\Pr E) \geq 1 - (\Pr E)^{\Omega(\varepsilon)}$. So by Lemma 19 applied with $\delta' = \Pr(E)$, we obtain that for some universal C,

$$\|\mathbb{E}\, X|E - \mathbb{E}\, X\| \leq C \cdot \left(\frac{(\delta')^{\Omega(\varepsilon)}}{\delta'}\right)^{1/2} \,,$$

(where we used $\|\mathbb{E}\, X\| = 1$ again).

Using this to bound (4) for events with $\Pr E > \delta/2$, and recalling our choice of s above, we have

$$\begin{aligned}(\Pr E)^{r/s} \cdot \|\mathbb{E}\, X|E - \mathbb{E}\, X\|^r \cdot \frac{s}{s-r} &\leq \Pr(E)^{1-1/\log(1/\delta)} \cdot C^r \cdot \left(\frac{(\delta')^{\Omega(\varepsilon)}}{\delta'}\right)^{r/2} \cdot \log(1/\delta) \\ &\leq \frac{O(\delta)^{\Omega(\varepsilon r)}}{\delta^{r/2-1}} \\ &\leq \delta \cdot \left(\frac{\eta}{\delta}\right)^{r/2}\end{aligned}$$

for small-enough δ and the choice of η above; the second simplification just uses $\Pr(E) \geq \delta/2$.

We conclude by (4) that for all $\delta < \delta_0(\varepsilon, r)$ it holds that

$$\mathbb{E}\,|\langle X - \mathbb{E}\, X, v\rangle|^r \leq \frac{\eta^{r/2}}{\delta^{r/2-1}} \,.$$

Thus picking $\delta < \min_{r \leq q} \delta_0(\varepsilon, r)$, we conclude that the set of vectors S is in the **no** case.

Finally, the distribution over S is as stated not isotropic, because the first coordinate of every vector is 1. Indeed, it is a standard fact that the distribution which is simply the uniform distribution over the vectors in S with the first coordinated removed is mean zero and isotropic. However, it is easily to check that the proof above goes through for S projected off of the first coordinate. Then the resulting distribution is indeed isotropic, and satisfies all the desired guarantees as in the Lemma. This completes the proof. ■

The second lemma we need to prove Theorem 24 is the following inequality for p-th moments of sums of independent random variables.

Fact 8 (Rosenthal's Theorem, see e.g. Johnson et al. (1985)) *Let $p \geq 2$, and let $X_1, \ldots, X_n$ be independent with $\mathbb{E}\, X_i = 0$ and $\mathbb{E}\,|X_i|^p < \infty$ for all $i = 1, \ldots, n$. Then*

$$\mathbb{E}\left|\sum X_i\right|^p \leq (C_1 p)^p \cdot \left(\sum \mathbb{E}\left[|X_i|^p\right]\right) + (C_2 p)^{p/2} \cdot \left(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]\right)^{p/2} \,,$$

for some universal constants C_1, C_2.

Proof [Proof of Theorem 24] Let S be an instance of the problem in Lemma 25 with parameters q and δ. We show how to construct a set S' over $n^{O(1/\delta)}$ points in time $n^{O(1/\delta)}$ so that a **yes** instance of the problem in Lemma 25 is mapped to a **yes** instance of the problem in Theorem 24, and similarly for **no** instances. Composing this reduction with Lemma 25 immediately yields Theorem 24.

To achieve this, we simply let S' be the set

$$S' = \left\{ \sqrt{\frac{\delta}{\alpha}} \sum_{i_1,\ldots,i_{\alpha/\delta}} X_i \, : \, i \in [|S|] \right\} ,$$

or equivalently, the uniform distribution D' over S' is the the sum of α/δ i.i.d. samples from the uniform distribution D over S, scaled by $\sqrt{\delta/\alpha}$. Here $\alpha \leq 1$ is a parameter depending only on q and c to be tuned later. Clearly $|S'| = n^{O(1/\delta)}$ and can be constructed in time $n^{O(1/\delta)}$ given a construction of S. We now check soundness and completeness.

Suppose S is an **yes** instance from Lemma 25. Then there exists an event E of D with $\Pr E \leq \delta$ and a unit vector v so that $|\mathbb{E}\langle v, X\rangle | E - \mathbb{E}\langle v, X\rangle| \geq \frac{0.4}{\sqrt{\Pr E}}$. The event E corresponds to some $T \subset S$ with $|S| \leq \delta |S|$. Let E' be the event in D' that at least one X_i in the sum belongs to S. By standard estimates, $\Pr[X' \in S'] = 1 - (1 - \Pr E)^{\alpha/\delta} = \Omega(\alpha \Pr E / \delta)$ for $\Pr E \leq \delta$. Moreover, since $\mathbb{E}_{X \sim D} X = 0$, we have that $\| \mathbb{E}_{X' \sim D'} X' | E' \| = \sqrt{\delta/\alpha} \cdot \| \mathbb{E}_{X \sim D} X | E \| \geq 0.4 \cdot \sqrt{\frac{\delta}{\alpha \Pr E}}$. Hence, using the contribution of the event E to the t-th moments to lower-bound them (Fact 9), for $t \in (2, q]$, there exists some unit vector v so that

$$\begin{aligned}
\mathbb{E} |\langle v, X\rangle|^t &\geq \left(\Pr E'\right) \cdot (0.4)^t \cdot \left(\frac{\delta}{\alpha \Pr E}\right)^{t/2} \\
&\geq \Omega\left(\frac{\alpha \Pr E}{\delta}\right) (0.4)^t \cdot \left(\frac{\delta}{\alpha \Pr E}\right)^{t/2} \\
&\geq \left(\frac{1}{\alpha}\right)^{\Omega(t)} \left(\frac{\delta}{\Pr E}\right)^{t/2-1} \\
&\geq \left(\frac{1}{\alpha}\right)^{\Omega(t)} \geq (ct)^{t/2} ,
\end{aligned}$$

for α chosen such that $c = \frac{1}{t}(1/\alpha)^{\Omega(1)}$. Hence S' is an instance of the **yes** case.

Suppose on the other hand that S is a **no** instance. Let v be an arbitrary unit vector. Let $X' \sim D'$, so that $X' = \sqrt{\delta/\alpha}\left(\sum_{i=1}^{\alpha/\delta} X_i\right)$ where $X_i \sim D$ are independent. Then, by Rosenthal's inequality (Fact 8) applied to the random variables $Z_i = \sqrt{\delta}\langle v, X_i\rangle$, we see that there are universal constants C_1, C_2 such that for any $r \in (2, q]$,

$$\begin{aligned}
\mathbb{E} \left|\langle v, X'\rangle\right|^r &\leq (C_1 r)^r \cdot \left((\delta/\alpha)^{r/2-1} \, \mathbb{E}_{X \sim E} |\langle v, X\rangle|^r\right) + (C_2 r)^{r/2} \\
&\leq (C_1 r)^r \cdot \delta^{\Omega(r)} \cdot (1/\alpha)^{r/2-1} + (C_2 r)^{r/2} ,
\end{aligned}$$

by Lemma 25. Using our previous choice for α, we see that if δ is small enough as a funcion of c, q then the second term dominates, and we get

$$\sup_{\|v\|=1} \mathbb{E} \left|\langle v, X'\rangle\right|^r \leq (sr)^{r/2} ,$$

for some universal constant $s = O(1)$. Thus in this case we are in the **no** case. This completes the proof. ■

Appendix C. Unique-SSE and Robust Estimation

In this section we prove Theorem 29 on hardness of robust estimation under USSEH.

Definition 26 (Almost-SSE) *Suppose G is an n-node graph. We say that G is an almost (ε, δ) small set expander if:*

- *there is $S \subseteq [n]$ with $|S| = \delta n$ and $\Phi_G(S) \leq \varepsilon$, and*
- *every $T \subseteq [n]$ with $|T| = \delta n$ and $T \cap S = \varnothing$ has $\Phi_G(T) \geq 1 - \varepsilon$.*

Hypothesis 27 (Unique Small-Set Expansion Hypothesis USSEH) *For every $\varepsilon > 0$ there is a small-enough $\delta > 0$ such that the following problem is NP-hard. Given an n-node graph G, distinguish between the cases: **yes:** G is an almost (ε, δ) small set expander, and **no:** $\Phi_G(\delta) \geq 1 - \epsilon$.*

Problem 28 (α, β-approximate robust mean estimation under resilience) ***Input:*** *$b_1, \ldots, b_n \in \mathbb{R}^d$ and $\delta > 0$, such that there exists $S \subseteq [n]$ with $|S| = (1 - \delta)n$ which is $(\alpha\sqrt{\delta}, \delta)$-resilient.* ***Output:*** *A vector $\hat{\mu} \in \mathbb{R}^n$ such that $\|\hat{\mu} - \mathbb{E}_{i \sim S} b_i\| \leq \beta\sqrt{\delta}$.*

Theorem 29 *Suppose USSEH. There is an absolute constant $\beta^* < 1$ such that if for any constant $\alpha < \beta^*$ Problem 28 has a polynomial-time algorithm then $\mathsf{P} = \mathsf{NP}$.*

Our main tool is the "moreover" clause in Lemma 15 which allows for G to be an almost (ε, δ) small set expander rather than a small set expander. This allows us to prove the following result characterizing the means of embeddings of small sets in G which do not overlap with the small non-expanding set.

Lemma 30 *Suppose that G is an n-node almost (ε, δ) small set expander for $\varepsilon < \varepsilon_0$, where $\varepsilon_0 > 0$ is a universal constant. Let $T \subseteq [n]$ have $|T| \leq \delta n$ and no intersection with the small non-expanding set in G. Let $b_1, \ldots, b_n$ be the isotropic spectral embedding of G. Then*

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\| \cdot \frac{|T|}{n} \leq 2\varepsilon^{0.05}\sqrt{\delta}\,.$$

Proof We proceed as in the proof of Theorem 23. By definition, $\frac{1}{|T|}\sum_{i \in T} b_i = A^\top \cdot \frac{\sqrt{n}}{|T|} 1_T$ where A has columns which are the eigenvectors of G with eigenvalue at least $1/2$. We will combine two bounds, one for $|T| \ll \delta n$ and one for $|T| \approx \delta n$.

Firstly, because $\|A\| \leq 1$, we have

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\|^2 = n \cdot \left\| A^\top \frac{1_T}{|T|} \right\|^2 \leq n \cdot \|1_T / |T|\|^2 = \frac{n}{|T|}\,.$$

Let $r = r(\varepsilon, \delta)$ be a constant to be chosen later. If $|T| \leq r\delta n$, then we find

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\| \cdot \frac{|T|}{n} \leq \sqrt{\frac{|T|}{n}} \leq \sqrt{r\delta}\,.$$

Now we address sets with sizes in the range $|T| \in [r\delta n, \delta n]$. Here we will use a local Cheeger inequality – Lemma 15. We are interested in

$$\frac{n}{|T|^2} \langle 1_T, \Pi_{1/2} 1_T \rangle \leq \frac{4n}{|T|^2} \|G 1_T\|^2\,.$$

Picking $\eta = \varepsilon^{0.1}$, if $\|G 1_T\|^2 \geq \varepsilon^{0.1} \|1_T\|^2$ then there is a set R of size in the range $|R| \in [c\varepsilon^{0.3} r\delta n, C\delta n / \varepsilon^{0.4}]$ for some universal constants c, C, with expansion $\Phi_G(R) \leq 1 - \Omega(\varepsilon^{0.4})$. Furthermore, $R \cap S = \varnothing$.

By subsampling at random or adding vertices as necessary, we find that there is a set R' of size δn which does not overlap S and has expansion $\Phi_G(R') \leq 1 - \Omega(r\varepsilon^{0.8})$. Choosing $r = \varepsilon^{0.1}$ and ε sufficiently small, this violates that G is an almost (ε, δ) small set expander. So it must be that $\|G 1_T\|^2 \leq \varepsilon^{0.1} \|1_T\|^2 \leq \varepsilon^{0.1} \delta n$. We therefore find that for $|T| \in [r\delta n, \delta n]$,

$$\left\| \frac{1}{|T|} \sum_{i \in T} b_i \right\| \cdot \frac{|T|}{n} \leq \frac{|T|}{n} \cdot \frac{2\sqrt{n}}{|T|} \cdot \varepsilon^{0.05} \sqrt{\delta n} = 2\sqrt{\varepsilon^{0.1} \delta}\,.$$

■

Proof [Proof of Theorem 29] We will analyze the following reduction from small-set expansion to robust mean estimation under resilience. Let β^* be a small-enough absolute constant. (We can choose it later). Let $\alpha < \beta^*$.

Given an n-node graph G and parameters $\varepsilon, \delta > 0$, let $b_1, \ldots, b_n$ be the isotropic spectral embedding of G. Let μ be the output of an oracle for Problem 28 with parameters $\alpha, \beta^*, \delta/2$ on input $b_1, \ldots, b_n$. Let $e_1 \in \mathbb{R}^d = (1, 0, 0, \ldots, 0)$ be the first standard basis vector. If $\|\mu - e_1\| > 2\beta^* \sqrt{\delta}$ then output **yes**. Otherwise output **no**.

We need to show that there exists $\varepsilon > 0$ such that for all $\delta > 0$ the following two statements hold:

Soundness: If $\Phi_G(\delta) > 1 - \varepsilon$ then $\|\mu - e_1\| \leq 2\beta^* \sqrt{\delta}$.

Completeness: If G is an almost (ε, δ) small set expander then $\|\mu - e_1\| > 2\beta^* \sqrt{\delta}$.

We address the statements in turn, beginning with soundness. By the proof of Theorem 23, if $\Phi_G(\delta) > 1 - \varepsilon$ then the uniform distribution on $\{b_1, \ldots, b_n\}$ is $(2\varepsilon^{1/8} \sqrt{\delta}, \delta)$-resilient.

Hence every subset of S of size $(1 - \delta/2)n$ is also $(4\varepsilon^{1/8} \sqrt{\delta}, \delta/2)$-resilient. Fix one such subset S. By Fact 6, we have $\mathbb{E}_{i \sim [n]} b_i = e_1$. Hence by resilience, $\| \mathbb{E}_{i \sim S} b_i - e_1 \| \leq 2\varepsilon^{1/8} \sqrt{\delta}$. By the guarantee of our robust mean estimation oracle, so long as $2\varepsilon^{1/8} \sqrt{\delta} \leq \alpha$ then $\|\mu - \mathbb{E}_{i \sim S} b_i\| \leq \beta^* \sqrt{\delta}$. By triangle inequality,

$$\|\mu - e_1\| \leq \|\mu - \mathbb{E}_{i \sim S} b_i\| + \| \mathbb{E}_{i \sim S} b_i - e_1 \| \leq (\beta^* + 2\varepsilon^{1/8}) \sqrt{\delta} \leq 2\beta^* \sqrt{\delta}$$

for small-enough $\epsilon = \epsilon(\alpha, \beta^*)$.

Now we move on to completeness. Let $S \subseteq [n]$ be the δn-size subset of vertices with $\Phi_G(S) \leq \varepsilon$. Let $v = \mathbb{E}_{i \sim S} b_i$ and let $w = \mathbb{E}_{i \notin S} b_i$. Since (by Fact 6) we have $\mathbb{E}_{i \sim [n]} b_i = e_1$, simple calculations show that

$$w = \frac{e_1 - \delta v}{1 - \delta} \,.$$

This rearranges to

$$e_1 - w = \frac{\delta v + \delta e_1}{1 - \delta} \,.$$

We first establish that the set $\{b_i\}_{i \notin S}$ is $(4\varepsilon^{0.05}\sqrt{\delta}, \delta/4)$-resilient. Let $R \subseteq [n] \setminus S$ have size at most $|R| \leq \delta n/2$. Then by Lemma 30, we have

$$\left\| \frac{1}{|R|} \sum_{i \in R} b_i \right\| \leq 2\varepsilon^{0.05}\sqrt{\delta} \cdot \frac{n}{|R|} \,.$$

Hence by triangle inequality we have

$$\left\| \frac{1}{|R|} \sum_{i \in R} b_i - w \right\| \leq 2\varepsilon^{0.05}\sqrt{\delta} \cdot \frac{n}{|R|} + \|w\|$$

and so finally

$$\left\| \frac{1}{|R|} \sum_{i \in R} b_i - w \right\| \cdot \frac{|R|}{n} \leq 2\varepsilon^{0.05}\sqrt{\delta} + \delta \cdot \|w\| \,.$$

It follows that $\{b_i\}_{i \notin S}$ is $(2\varepsilon^{0.05} + \delta\|w\|, \delta/4)$-resilient. By Lemma 30, $\|w\| \leq 2\varepsilon^{0.05}\sqrt{1/\delta}$. So ultimately, $\{b_i\}_{i \notin S}$ is $(4\varepsilon^{0.05}\sqrt{\delta}, \delta/4)$-resilient.

Therefore, we must have that $\|\mu - w\| \leq O(\varepsilon^{0.05}\sqrt{\delta})$. At the same time, by Lemma 18, we have $\|v\|^2 \geq 1/2\delta$, so $\|w - e_1\| \geq \Omega(\sqrt{\delta})$. So,

$$\|\mu - e_1\| = \|(\mu - w) + (w - e_1)\| \geq \|w - e_1\| - \|\mu - w\| \geq \Omega(\sqrt{\delta}) - O(\varepsilon^{0.05}\sqrt{\delta}) \,.$$

So, for sufficiently small β^* and ε, we find that for all δ, $\|\mu - e_1\| > 2\beta^*\sqrt{\delta}$. ■

Appendix D. Conclusion and Open Problems

In this paper we give evidence from worst case complexity assumptions that improving existing algorithms for robust mean estimation may be hard. These results are far from complete, however, and there are a number of very interesting open questions in this area.

The most natural question is whether or not we can show that improving current algorithms for robust mean estimation assuming bounded moments or resilience is impossible under SSEH. There are a number of interesting sub-questions:

- Can the uniqueness assumption be removed in the proof that USSEH implies improved robust mean estimation under resilience is NP-hard? As far as we are aware it could even be that SSEH and USSEH are equivalent – are they?.

- Does SSEH or a variant (such as USSEH) imply that improved robust mean estimation is hard under bounded moment assumptions? Our current techniques are unable to prove this for USSEH: they would require an analogue of Lemma 19 in the setting that G is a graph which contains a unique small non-expanding set. That lemma requires running a random walk on the graph G for about $\log n$ steps; we do not know how to ensure that such a random walk avoids entering the small non-expanding set (or, if it does, how to control its behavior across the nonexpanding cut).

Another interesting question is whether or not these techniques can be used to show hardness for other questions in robust estimation, such as list learning Charikar et al. (2017), or robust sparse mean estimation Balakrishnan et al. (2017). (Unlike for the main problems addressed in this paper, SQ lower bounds for these are already known Diakonikolas et al. (2017c, 2018).) We conjecture that the current spectral-based algorithms for these problems are optimal, even with additional assumptions on resilience or moments.

It is also interesting to ask whether SSEH-type assumptions can be avoided all together. In addition to showing that approximating the $2 \to 4$-norm is SSEH-hard, the authors of Barak et al. (2012a) also show it is NP-hard assuming the Exponential Time Hypothesis. That proof does not appear to easily adapt to our setting, however, because it is not clear the instance of the $2 \to 4$-norm problem it produces can be transformed into a distribution with sub-Gaussian moments as we require, nor can we easily control the kind of tail events we require to prove hardness of resilience. Nonetheless, it seems plausible that hardness for some robust estimation problem could be shown under assumptions weaker than SSEH.

Appendix E. Omitted Proofs from Section 2

E.1. Proofs of Local Cheeger Inequalities

Lemma 31 (Local Cheeger Inequality Steurer (2010b)) *For every $v \in \mathbb{R}^n$ there is a level set $S \subseteq V$ of the vector $w_i = v_i^2$ with $|S| \leq \delta n$ and expansion*

$$\Phi_G(S) \leq \frac{\sqrt{1 - \langle v, Gv\rangle^2/\|v\|^4}}{1 - \|v\|_1^2/\delta n \|v\|^2} \,.$$

Proof [Proof of Lemma 14] We follow the proof in Steurer (2010a), keeping track of a factor of $1/\varepsilon^2$ missing in that proof; at the end we apply a standard sub-sampling reduction used in e.g. Raghavendra et al. (2012).

First, dividing by $\|f\|_1$, we may assume that $\|f\|_1 = 1$; i.e. that f is a probability vector. We will apply Lemma 31 to the distribution $g = (f + Gf)/2$. Clearly $\|g\|_1 = 1$.

Since G is contractive in 2-norm, we have $\|g\|_2 \leq \|f\|_2$. But since f, Gf are nonnegative, also $\|g\|_2 \geq \|f\|_2/2$.

Finally, consider

$$\langle g, Gg\rangle = \langle f, Gf\rangle + 2\langle f, G^2 f\rangle + \langle f, G^3 f\rangle \geq 2\|Gf\|^2 \geq 2\varepsilon\|f\|^2$$

where we used that $\langle f, G^3 f\rangle, \langle f, Gf\rangle \geq 0$ by nonnegativity, and we used our hypothesis on $\|Gf\|^2$. Plugging these bounds into Lemma 31, we find that there is a level set S of g having size at most $\delta n/(\gamma\varepsilon^2)$ such that

$$\Phi_G(S) \leq \frac{\sqrt{1 - \langle g, Gg\rangle^2/\|g\|^4}}{1 - 20\varepsilon^2/(\delta n\|g\|^2)} \leq \frac{\sqrt{1 - 4\varepsilon^2}}{1 - 20\varepsilon^2} \leq 1 - \Omega(\varepsilon^2)\,.$$

Let T be a random subset of S of size δn. A simple computation shows that

$$\mathbb{E}_T(1 - \Phi_G(T)) \geq \gamma\varepsilon^2(1 - \Phi_G(S)) \geq \Omega(\gamma\varepsilon^4)\,.$$

as claimed. ■

Proof [Proof of Lemma 15] We begin by proving the statement prior to the "moreover," then we describe how the proof may be slightly altered in the case that G contains a small non-expanding set.

Our proof proceeds very similarly to the proof in Steurer (2010b). Let c, C be constants to be determined later. Let T_t be a random subset drawn from the following distribution: first, t is drawn uniformly from $[0, 1]$, then $T_t = \{i \in [n] : g_i^2 \geq t\}$. We first establish a number of properties of this distribution. Observe that since $|f_i| \leq 1$ for all i, then since G is a random walk matrix, $|Gf_i| \leq 1$ for all i as well, and so $g_i^2 \leq 1$ for all i. Therefore, by a simple calculation, we have that

$$\mathbb{E}_t\left[|T_t|\right] = \sum_{i \in [n]} g_i^2 = \|g\|^2 \,.$$

We also have that $\|g\|^2 \geq (1-\eta)^2 \|f\|^2 = (1-\eta)^2 \delta n$. Moreover, if $t \leq (1-\eta)^2$, $S \subseteq T_t$ and hence $|T_t| \geq \delta n$. Therefore

$$\Pr_t\left[|T_t| < \delta n\right] \leq 2\eta\,. \tag{5}$$

For any $U, V \subseteq [n]$, let $G(U,V) = \Pr_{(i,j)\sim G}[i \in U, j \in V]$ be the fraction of edges going from U to T, so that $\Phi_G(U) = nG(U, [n] \setminus U)/|U|$.

Then, by the same calculations as those done in Steurer (2010b), we still have the following three inequalities:

$$\mathbb{E}_t \left[|T_t|^2 \right] \leq \|g\|_1^2 \ , \tag{6}$$

$$\mathbb{E}_t \left[|T_t| \mathbf{1}_{|T_t| > C\delta n/(\eta\varepsilon)^2} \right] \leq \frac{\eta^2 \varepsilon^2}{C\delta n} \mathbb{E}_t \left[|T|^2 \right] \ , \tag{7}$$

$$n \cdot \mathbb{E}_t \, G(T_t, [n] \setminus T_t) \leq \|g\|^2 \sqrt{1 - \langle g, Gg \rangle^2 / \|g\|^4} \ . \tag{8}$$

We now specialize each of these three inequalities to our setting. Observe that f is nonnegative and satisfies $\|f\|_1 = |S| = \delta n$, and so because G is a random walk matrix, we have $\|g\|_1 = |S|$, and so (6) simply becomes

$$\mathbb{E}_t \left[|T_t|^2 \right] \leq (\delta n)^2 \ .$$

Plugging this bound into (7) yields that

$$\mathbb{E}_t \left[|T_t| \mathbf{1}_{|T_t| > C\delta n/(\eta\varepsilon)^2} \right] \leq \frac{\varepsilon^2 \eta^2}{C} \delta n \ . \tag{9}$$

Finally, observe that

$$\begin{aligned} \langle g, Gg \rangle &= (1-\eta)^2 \langle f, Gf \rangle + 2\eta(1-\eta) \|Gf\|^2 + \eta^2 \langle f, G^3 f \rangle \\ &\overset{(a)}{\geq} 2\eta(1-\eta) \|Gf\|^2 \\ &\overset{(b)}{\geq} 2\eta(1-\eta)\varepsilon \|f\|^2 \ , \end{aligned} \tag{10}$$

where (a) follows from the nonnegativity of f, and (b) follows from assumption. Moreover $\|g\|^2 \leq \|f\|^2 = \delta n$ since G is contractive in ℓ_2. Thus (8) simplifies in our setting to give

$$n \cdot \mathbb{E}_t \, G(T_t, [n] \setminus T_t) \leq \|g\|^2 \sqrt{1 - (2\eta(1-\eta)\varepsilon)^2} = \|g\|^2 (1 - \Omega(\eta^2 \varepsilon^2)) \ . \tag{11}$$

Now, let T^* be the level set of g^2 with size in the range $I = [c\eta\varepsilon^2 \delta n, C\delta n/(\eta\varepsilon)^2]$ with minimal $\Phi(T^*)$. Since $\Phi(T) = n \cdot \frac{G(U,[n]\setminus U)}{|U|}$, we have

$$\begin{aligned} \Phi(T^*) &\leq n \frac{\mathbb{E}_t \, G(T_v, [n] \setminus T_t)}{\mathbb{E}_t \, |T_t| \mathbf{1}_{|T_t| \in I}} \\ &\leq n \frac{\mathbb{E}_t \, G(T_v, [n] \setminus T_t)}{\mathbb{E}_t \, |T_t| \mathbf{1}_{|T_t| \in I}} \\ &= n \frac{\mathbb{E}_t \, G(T_v, [n] \setminus T_t)}{\mathbb{E}_t \, |T_t| - \mathbb{E}_t \, |T_t| \mathbf{1}_{|T_t| < c\eta\varepsilon^2 \delta n} - \mathbb{E}_t \, |T_t| \mathbf{1}_{|T_t| > C\delta n/(\eta\varepsilon)^2}} \\ &\overset{(a)}{\leq} n \frac{\mathbb{E}_t \, G(T_v, [n] \setminus T_t)}{\mathbb{E}_t \, |T_t| - c'\eta^2 \varepsilon^2 \delta n} \\ &\overset{(b)}{\leq} \frac{\|g\|^2 (1 - O(\eta^2 \varepsilon^2))}{\|g\|^2 (1 - 2\varepsilon^2 \eta^2)} \\ &\leq 1 - O(\eta^2 \varepsilon^2) \ , \end{aligned}$$

for an appropriate choice of c sufficiently small and C sufficiently large. Here (a) follows from (9), and since

$$\mathbb{E}_t\,|T_t|\mathbf{1}_{|T_t|<c\eta\varepsilon^2\delta n} \le c\eta\varepsilon^2\delta n\,\Pr\left[|T_t| < c\eta\varepsilon^2\delta n\right] \le c\eta^2\varepsilon^2\delta n \tag{12}$$

by (5), and (b) follows since $\|g\|_2^2 \ge (1-\eta)^2\|f\|^2 = (1-\eta)^2\delta n$. This completes the proof, except for the “moreover” statement, proved below. ■

Proof [Proof of Lemma 15, “moreover” part] Suppose that G contains a set R as described in the lemma statement. We describe how the preceeding proof may be altered to ensure that $T \cap R = \varnothing$.

The idea is to replace the function g with the function $g' = \Pi_{\overline{R}} g$, the projection of g to the coordinates outside R. The random thresholding procedure is applied to the coordinates of g' to produce the set T; because g' is supported off of R it holds that $T \cap R = \varnothing$ with probability 1.

We now verify that properties of g used above also apply to g'. Since f is supported off of R, (5) continues to hold. Equations (6), (7), (8) hold for any choice of g and hence in particular for g'.

Because $\|g'\|_1 \le \|g\|_1$, we obtain (9) when T is chosen according to the thresholding procedure on g.

We need to lower bound $\langle g', Gg'\rangle$ to obtain an analogue of (10). By expanding, we find

$$\langle g', Gg'\rangle = \langle g, Gg\rangle + 2\langle g'-g, Gg'\rangle + \langle g'-g, G(g'-g)\rangle\,.$$

Because $\Phi_G(R) \le \varepsilon/10$ and $|S| = |R| = \delta n$, we obtain that $\|g'-g\|_1 = \|\Pi_R g\|_1 \le \eta\varepsilon\delta n/10$. And because as noted before $|g_i| \le 1$ for all i, we have $\|g'\|_\infty \le 1$. So $|\langle g'-g, Gg'\rangle| \le \eta\varepsilon\delta n$; the same argument applies to $|\langle g'-g, G(g'-g)\rangle|$. And we proved above that $\langle g, Gg\rangle \ge 2\eta(1-\eta)\varepsilon\|f\|^2$. We may assume $\eta \le 1/2$, so it follows that $\langle g', Gg'\rangle \ge \eta(1-\eta)\varepsilon\|f\|^2$. Thus up to a factor of 2, we obtain the analogue of (10) for g' in place of g.

Since $\|g'\|^2 \le \|g\|^2$, we also obtain

$$n\,\mathbb{E}_t\,G(T_t, [n]\setminus T_t) \le \|g'\|^2\sqrt{1-(\eta(1-\eta)\varepsilon^2)} = \|g\|^2(1-\Omega(\eta^2\varepsilon^2))$$

as in (11).

Finally, since $\Pi_{\overline{R}} f = f$ (since f is supported off of R), it still holds that $\Pr[|T_t| < c\eta\varepsilon^2\delta n] \le \eta$ as in (12). The rest of the proof goes through unchanged. ■

E.2. Equivalence of Moments and Mean Shifts

We will repeatedly use the following elementary fact, which proves a near equivalence of moment bounds and mean shifts for $\mathbb{R}$-valued random variables.

Fact 9 *Let X be a $\mathbb{R}$-valued random variable with mean zero, and let $q \ge 1$. Then:*

- **Moment bounds implies bounded deviation** *Suppose $\mathbb{E}\,|X|^q$ is finite. Then for any event A, we have $|\mathbb{E}\,X\,|A| \le \left(\frac{\mathbb{E}\,|X|^q}{\Pr[A]}\right)^{1/q}$.*

- **Bounded deviation implies moment bounds** *For any p, let $C_p = \sup_A \Pr[A]\cdot|\mathbb{E}\,X\,|A|^p$. For every $p > q$, $\mathbb{E}\,|X|^q \le (2C_p)^{q/p}\cdot\frac{p}{p-q}$.*

Proof [Proof of Fact 9] We first prove the first implication. By Holder's inequality, we have

$$|\mathbb{E}\, X 1_A| \quad \leq (\mathbb{E}\, |X|^q)^{1/q} \Pr[A]^{1-1/q} \,,$$

and so

$$|\mathbb{E}\, X \,|A| = \frac{1}{\Pr[A]} \,|\mathbb{E}\, X 1_A| \leq \left(\frac{\mathbb{E}\, |X|^q}{\Pr[A]} \right)^{1/q} \quad ,$$

as claimed.

We now turn to the second implication. For any $t \geq 0$,

$$\Pr[|X| \geq t] = \Pr[X \geq t] + \Pr[X \leq -t] \leq \frac{C_p}{|\mathbb{E}\, X \mid X \geq t|^p} + \frac{C_p}{|\mathbb{E}\, X \mid X \leq -t|^p} \leq \frac{2C_p}{t^p} \,.$$

Recall that $\mathbb{E}\, |X|^q = \int_0^\infty \Pr[|X|^q \geq s] \, ds$. We will split this integral into two parts, because we know two different bounds on $\Pr[|X|^q \geq s]$. First of all, for any s we have $\Pr[|X|^q \geq s] \leq 1$ Second of all, when $s > (2C_p)^{q/p}$ a better bound is given by $\Pr[|X|^q \geq s] \leq 2C_p / s^{p/q} < 1$. So,

$$\mathbb{E}\, |X|^q = \int_0^\infty \Pr[|X|^q \geq s] \, ds \leq \int_0^{(2C_p)^{q/p}} 1 \, ds + \int_{(2C_p)^{q/p}}^\infty \frac{2C_p}{s^{p/q}} \, ds \,.$$

The first integral is just $(2C_p')^{q/p}$. The second is

$$\int_{(2C_p')^{q/p}}^\infty \frac{2C_p'}{s^{p/q}} \, ds = \frac{1}{\frac{p}{q} - 1} \cdot [(2C_p)^{q/p}]^{-p/q+1}$$

so long as $p > q$. (Otherwise the integral does not exist.)

Putting these together,

$$\mathbb{E}\, |X|^q \leq (2C_p)^{q/p} + \frac{1}{\frac{p}{q} - 1} \cdot 2C_p \cdot [(2C_p)^{q/p}]^{-p/q+1} = (2C_p)^{q/p} \cdot \left(1 + \frac{1}{(\frac{p}{q} - 1)} \right) \,.$$

Finally, note that $1 + 1/(\frac{p}{q} - 1) = \frac{p-q}{p-q} + \frac{q}{p-q} = \frac{p}{p-q}$, which finishes the proof. ■

As a simple corollary of this, we observe that moment bounds are equivalent to resilience "at every scale". For simplicity of exposition, we will state and prove the claim for ℓ_2 norm, however, the claim holds much more generally as well. This gives a novel characterization of resilience which may be of independent interest.

Corollary 32 *Let X be an $\mathbb{R}^d$-valued random variable with mean $\mathbb{E}\, X = \mu$, and let $q \geq 1$. Then:*

- **Moment bounds imply multi-scale resilience** *Suppose there exists a constant $C > 0$ so that $\mathbb{E} \langle v, X \rangle^q \leq C$ for all unit vectors v. Then, X is $(\frac{2C^{1/q}}{\delta^{1/q-1}}, \delta)$-resilient for all $\delta \leq 1/2$.*
- **Multi-scale resilience implies moment bounds** *Let $p > q$, and let C_p be so that X is $(\frac{C_p^{1/p}}{\delta^{1/p-1}}, \delta)$-resilient for all $\delta \leq 1/2$. Then,*

  $$\mathbb{E}\, |\langle v, X - \mu \rangle|^q \leq (2C_p)^{q/p} \frac{p}{p-q}$$

  for all unit vectors $v \in \mathbb{R}^d$.

Proof We first prove the first implication. Let v be an arbitrary unit vector, let $\delta \in (0, 1/2)$ and let A be an event with $\Pr[A] \le \delta$. Then, by our assumption and Fact 9, we know that

$$|\mathbb{E}\langle v, X\rangle | A - \langle v, \mu\rangle| \le \left(\frac{C}{\Pr A}\right)^{1/q} = \frac{C^{1/q}}{\Pr[A]^{1/q-1}} \cdot \frac{1}{\Pr A} \le \frac{C^{1/q}}{\delta^{1/q-1}} \cdot \frac{1}{\Pr A} \le \frac{2C^{1/q}}{\delta^{1/q-1}} \frac{1 - \Pr A}{\Pr A} .$$

Taking a supremum of this inequality over all unit vectors v immediately yields the desired bound.

We now prove the other direction. For any unit vector $v \in \mathbb{R}^d$, and any event A with $\Pr A \le 1/2$, and by our assumption of resilience (taking $\delta = \Pr A$), we have

$$\Pr[A] \cdot |\mathbb{E}\langle v, X\rangle | A - \langle v, \mu\rangle|^p \le \Pr[A] \cdot \|\mathbb{E}\, X | A - \mu\|^p \le C_p . \tag{13}$$

Moreover, for any event A with $\Pr A > 1/2$, we also have

$$\Pr[A] \cdot |\mathbb{E}\langle v, X\rangle | A - \langle v, \mu\rangle|^p = \Pr[A] \cdot \left(\frac{\Pr A^c}{\Pr A}\right)^{p-1} \Pr[A^c] \cdot |\mathbb{E}\langle v, X\rangle | A^c - \langle v, \mu\rangle|^p \le C_p ,$$

where the last inequality follows from (13). Thus, by Fact 9, we have

$$\mathbb{E}\, |\langle v, X - \mu\rangle|^q \le (2C_p)^{q/p} \frac{p}{p-q} ,$$

as claimed. ■

We briefly remark that to generalize this statement to more general norms $\|\cdot\|$, it suffices to have the moment bound be taken over all unit vectors over the dual norm $\|\cdot\|^*$. The proof is a fairly standard generalization of this argument and we omit the proof for simplicity of exposition.

Appendix F. Omitted Proofs from Section 3

Proof [Proof of Lemma 18] Let $B = \sqrt{n}A$. We start by expanding:

$$\left\|\frac{1}{|T|}\sum_{i \in T} b_i\right\|^2 = \frac{1}{|T|^2} \cdot \mathbf{1}_T^\top B B^\top \mathbf{1}_T^\top = \frac{n}{|T|^2} \mathbf{1}_T^\top \Pi_{1/2} \mathbf{1}_T .$$

Let $v_1, \ldots, v_n$ be the eigenvectors of G, with associated eigenvalues $\lambda_1, \ldots, \lambda_n$. Since G is stochastic, $|\lambda_i| \le 1$. So for any vector v we have

$$v^\top \Pi_{1/2} v = \sum_{i \,:\, \lambda_i \ge 1/2} \langle v, v_i\rangle^2 \ge \sum_{i=1}^{n} \lambda_i \langle v, v_i\rangle^2 - \frac{1}{2} \cdot \|v\|^2 = v^\top G v - \frac{1}{2} \cdot \|v\|^2 .$$

Putting it together,

$$\left\|\frac{1}{|T|}\sum_{i \in T} b_i\right\|^2 \ge \frac{n}{|T|^2} \cdot \left(\mathbf{1}_T^\top G \mathbf{1}_T - \frac{1}{2} \cdot |T|\right) = \frac{n}{|T|^2} \cdot |T| \cdot \left(\frac{1}{2} - \Phi_G(T)\right) .$$

■

Appendix G. Sketch of random-walk rounding for analytically sparse vectors

In this section we describe the proof of Theorem 12. The key is the following lemma, which says that if w is a vector in the high eigenspaces of G with $\|w\|_4^4 \geq 1/\delta n$, then there is a level set S of $|w|$ (applying the absolute value function coordinate-wise) containing $O(\delta n)$ coordinates such that $\|\frac{1}{|S|}\sum_{i \in S} b_i\|$ is large. The rest of the proof follows the same argument in Lemma 19, showing that as the random walk is run from initial distribution over S, it must eventually encounter a small cut with imperfect expansion or else it would violate the local Cheeger inequality.

Lemma 33 *Let G a graph with isotropic spectral embedding $b_i, \ldots, b_n$ and corresponding uniform distribution D. Suppose there exists a unit vector v so that $\mathbb{E}\langle v, X\rangle^4 \geq \frac{1}{\delta}$. Then, there exists $t > 0$ so that if we let $S = \{i : |\langle v, X\rangle| > t\}$, then $|S| \leq O(\delta n)$ and $|\mathbb{E}\langle v, X\rangle|E| \geq \Omega\left(\frac{1}{\delta^{1/2}\log^{1/4} 1/\delta}\right)$.*

Proof By Fact 9, we know that for all $p > 4$ there exists some event A so that

$$\Pr[A]^{4/p}\, |\mathbb{E}\langle v, X\rangle|A|^4 \cdot \frac{p}{p-4} \geq \frac{1}{\delta}\ . \tag{14}$$

Let A_p be the set which achieves the largest value for the LHS in (14). Without loss of generality, we may take A_p to be of the form $A_p = \{i : |\langle v, X\rangle| > t_p\}$ for some $t_p > 0$, since such sets maximize the mean shift in the direction v.

Because $\mathbb{E}\langle v, X\rangle^2 = 1$, by Fact 9, we must have $|\mathbb{E}\langle v, X\rangle|A_p| \leq \frac{1}{\sqrt{\Pr[A_p]}}$. Thus, $\Pr[A_p] \leq \delta^{p/(2p-4)}$, as otherwise we would have

$$\Pr[A_p]^{4/p} \cdot |\mathbb{E}\langle v, X\rangle|A_p|^4 \quad \leq \tfrac{1}{\Pr[A_p]^{2-4/p}} < \tfrac{1}{\delta}\ ,$$

which contradicts our choice of A_p. This implies that for all $p > 4$, we have

$$\delta^{4/(2p-4)}\, |\mathbb{E}\langle v, X\rangle|A_p|^4 \cdot \frac{p}{p-4} \geq \frac{1}{\delta}$$

For $p \leq 6$, if we let $q = p - 2$ we have that

$$\delta^{4/(2p-4)} \frac{p}{p-4} \geq \delta^{2/q} \frac{4}{q-2}\ ,$$

so optimizing over $q > 2$ and using Fact 10 (see below) yields that by choosing $q = 2\frac{\log 1/\delta}{\log 1/\delta - 1}$, we obtain that

$$|\mathbb{E}\langle v, X\rangle|A_{q+2}|^4 \geq \frac{O(1)}{\delta^2 \log 1/\delta}\ .$$

Finally, in this case, we have

$$\Pr A_{q+2} \leq \delta^{1+1/(2\log 1/\delta)} = O(\delta)\ .$$

This completes the proof of the lemma. ■

Fact 10 *Let $x \in (0,1)$, and let $r \geq 2$. Then we have*

$$\min_{s>r} x^{r/s} \frac{s}{s-r} \leq ex \log \frac{1}{x} \ ,$$

and the minimum is attained at $s = r \cdot \frac{\log 1/x}{\log 1/x - 1}$.

Proof [Proof of Fact 10] By monotonicity of logarithm, it suffices to find the minimizer of the function

$$f(s) = \frac{r}{s} \log x + \log s - \log(s-r) \ .$$

Taking derivatives, we find that

$$f'(s) = -\frac{r}{s^2} \log x + \frac{1}{s} - \frac{1}{s-r} \ .$$

Thus solving for $f'(s) = 0$, the minimizer of f must satisfy

$$\frac{r}{s^2} \log \frac{1}{x} = \frac{r}{s(s-r)} \ ,$$

or equivalently $s/(s-r) = \log 1/x$ and $r/s = 1 - 1/\log(1/x)$. Plugging these bounds into the original function yields the desired estimate. ■