SIAM J. Imaging Sciences
Vol. 7, No. 4, pp. 2105–2138

Efficient Point-to-Subspace Query in ℓ^1 with Application to Robust Object Instance Recognition*

Ju Sun†, Yuqian Zhang†, and John Wright†

Abstract. Motivated by vision tasks such as robust face and object recognition, we consider the following general problem: given a collection of low-dimensional linear subspaces in a high-dimensional ambient (image) space, and a query point (image), efficiently determine the nearest subspace to the query in ℓ^1 distance. In contrast to the naive exhaustive search which entails large-scale linear programs, we show that the computational burden can be cut down significantly by a simple two-stage algorithm: (1) projecting the query and database subspaces into lower-dimensional space by random Cauchy matrix and solving small-scale distance evaluations (linear programs) in the projection space to locate the nearest candidates; (2) with few candidates upon independent repetition of (1), getting back to the high-dimensional space and performing exhaustive search. To preserve the identity of the nearest subspace with nontrivial probability, the projection dimension typically is a low-order polynomial of the subspace dimension multiplied by a logarithm of the number of the subspaces (Theorem 2.1). The reduced dimensionality and hence complexity render the proposed algorithm particularly relevant to vision applications such as robust face and object instance recognition that we investigate empirically.

Key words. ℓ^1 point-to-subspace distance, nearest subspace search, Cauchy projection, face recognition, subspace modeling

AMS subject classifications. 68U10, 68T45, 68W20, 68T10, 15B52

DOI. 10.1137/130936166

1. Introduction.

Although visual data reside in very high-dimensional spaces, they often exhibit much lower-dimensional intrinsic structure. Modeling and exploiting this low-dimensional structure is a central goal in computer vision, with impact on applications ranging from low-level tasks such as signal acquisition and denoising to higher-level tasks such as object detection and recognition.

In face and object recognition alone, many popular, effective techniques can be viewed as searching for the low-dimensional model which best matches the query (test) image (see, e.g., [25, 3]). With each object $\mathcal{O}$ of interest we may associate a low-dimensional subset $\mathcal{M} \subset \mathbb{R}^D$, which approximates the set of images of $\mathcal{O}$ that can be generated under different physical conditions—say, varying pose or illumination. Given n objects $\mathcal{O}_i$ and their corresponding approximation subsets $\mathcal{M}_i$, the recognition problem becomes one of finding the nearest low-

*Received by the editors September 9, 2013; accepted for publication (in revised form) July 28, 2014; published electronically November 6, 2014.
http://www.siam.org/journals/siims/7-4/93616.html
†Department of Electrical Engineering, Columbia University, New York, NY 10027 (jusun@ee.columbia.edu, yz2409@columbia.edu, johnwright@ee.columbia.edu). The research of the first author was supported by the Wei Family Private Foundation. The research of the second and third authors was partially supported by ONR N00014-13-1-0492 and Columbia University startup funding.

dimensional structure. To put it formally,

$$\arg\min_i d(\mathbf{q}, \mathcal{M}_i),$$

where $\mathbf{q} \in \mathbb{R}^D$ is the test image, and $d(\cdot, \cdot)$ is some prescribed point-to-set distance function.

This paradigm is broad enough to encompass very classical work in face recognition [37] and object instance recognition [32], as well as more recent developments [13, 7, 42]. In situations when sufficient training data are available to accurately fit the $\mathcal{M}_i$, it can achieve high recognition rates [39]. In applying it to a particular scenario, however, at least three critical questions must be answered.

First, *what is the most appropriate class of low-dimensional models $\mathcal{M}_i$?* The proper class of models may depend on the properties of the object $\mathcal{O}$, as well as the types of nuisance variations that may be encountered. For example, variations in illumination may be well captured using low-dimensional *linear* models [22, 5], whereas variations in pose or alignment are highly nonlinear [18].

Second, *how should we measure the distance $d(\mathbf{q}, \mathcal{M}_i)$ between $\mathbf{q}$ and $\mathcal{M}_i$?* Typically, one adopts a metric $\mathrm{dist}\,(\cdot, \cdot)$ on $\mathbb{R}^D$ and then sets

$$d(\mathbf{q}, \mathcal{M}_i) = \min_{\mathbf{v} \in \mathcal{M}_i} \mathrm{dist}\,(\mathbf{q}, \mathbf{v})\,.$$

Here, again, the appropriate metric $\mathrm{dist}\,(\cdot, \cdot)$ depends on our prior knowledge. For example, if the observation $\mathbf{q}$ is known to be perturbed by independent and identically distributed (i.i.d.) Gaussian noise, minimizing the metric induced by the ℓ^2 norm $\mathrm{dist}\,(\mathbf{q}, \mathbf{v}) = \|\mathbf{q} - \mathbf{v}\|_2$ yields a maximum likelihood estimator. However, in practice other norms may be more appropriate: for example, in situations where the data may have errors due to occlusions, shadows, or specularities, the ℓ^1 norm is a more robust alternative [42].

Finally, given an appropriate model and error distance, *how can we efficiently determine the nearest model to a given input query?* That is to say, we would like to solve

$$\arg\min_{i \in \{1,\ldots,n\}} \min_{\mathbf{v} \in \mathcal{M}_i} \mathrm{dist}\,(\mathbf{q}, \mathbf{v}) \tag{1.1}$$

using computational resources that depend as gracefully as possible on the ambient dimension D (typically the number of pixels in the image) and the number of models n. In practical applications, both of these quantities could be very large.

This paper. In this paper, we consider the case when the low-dimensional models $\mathcal{M}_i$ are *linear subspaces*. As mentioned above, subspace models are well justified for modeling illumination variations [22, 5] (say, in near-frontal face recognition) and also form a basic building block for modeling and computing with more general, nonlinear sets [35, 34].

Our methodology pertains to distances $\mathrm{dist}(\mathbf{q}, \mathbf{v})$ induced by the ℓ^p norm $\|\mathbf{q} - \mathbf{v}\|_p$, with

$p \in (0, 2]$.[1] We focus here on the ℓ^1 norm, $\|\mathbf{q} - \mathbf{v}\|_1 = \sum_i |q_i - v_i|$. The ℓ^1 norm is a natural and well-justified choice when the test image contains pixels that do not fit the model—say, due to moderate occlusions, cast shadows, or specularities [42]. For $p \in (0, 2]$, the ℓ^p norm with $p = 1$ strikes a unique compromise between computational tractability (convexity) and robustness to gross errors.

With this choice of models and distance, at recognition time we are left with the following computational task.

Problem 1.1. *Given n linear subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_n$ of dimension r and a query point $\mathbf{q}$, all in $\mathbb{R}^D$, determine the nearest $\mathcal{S}_i$ to $\mathbf{q}$ in the ℓ^1 norm.*

This problem has a straightforward solution: solve a sequence of n ℓ^1 regression problems,

$$\min_{\mathbf{v} \in \mathcal{S}_i} \|\mathbf{q} - \mathbf{v}\|_1, \tag{1.2}$$

and choose the i with the smallest optimal objective value. The total cost is $O(n \cdot T_{\ell^1}(D, r))$, where $T_{\ell^1}(D, r)$ is the time required to solve the linear program (1.2). For example, for interior point methods [8], we have $T_{\ell^1}(D, r) = O(D^{3.5})$.[2] There exist more scalable first-order methods [20, 6, 45, 43], which improve on the dependence on D at the expense of higher iteration complexity. The best known complexity guarantees for each of these methods are again superlinear in D, although linear runtimes may be achievable when the residual $\mathbf{q} - \mathbf{v}_\star$ is very sparse [19] or the problem is otherwise well structured [1]. Even in the best case, however, the aforementioned algorithms have complexity $\Omega(nD)$.[3] When both terms are large, this dependence is prohibitive: Although Problem 1.1 is simple to state and easy to solve in polynomial time, achieving real-time performance or scaling massive databases of objects appears to require a more careful study.

In this paper, we present a very simple, practical approach to Problem 1.1, with much improved computational complexity and reasonably strong theoretical guarantees. Rather than working directly in the high-dimensional space $\mathbb{R}^D$, we randomly embed the query $\mathbf{q}$ and subspaces $\mathcal{S}_i$ into $\mathbb{R}^d$, with $d \ll D$. The random embedding is given by a $d \times D$ matrix $\mathbf{P}$ whose entries are i.i.d. standard Cauchy random variables. That is to say, instead of solving (1.2), we solve

[1] Mathematically $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$ defines a valid norm only when $p \geq 1$, which in turn induces valid metric $\|\mathbf{x} - \mathbf{y}\|_p$. For $p \in (0, 1)$, though $\|\cdot\|_p$ is not a valid norm, one can verify that $\|\mathbf{x}\|_p^p = \sum_i |x_i|^p$ indeed also induces a valid metric, i.e., for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^D$, $\|\mathbf{x} - \mathbf{y}\|_p^p \geq 0$, $\|\mathbf{x} - \mathbf{y}\|_p^p = 0 \iff \mathbf{x} = \mathbf{y}$, $\|\mathbf{x} - \mathbf{y}\|_p^p = \|\mathbf{y} - \mathbf{x}\|_p^p$, and also the triangular inequality holds: $\|\mathbf{x} - \mathbf{z}\|_p^p \leq \|\mathbf{x} - \mathbf{y}\|_p^p + \|\mathbf{y} - \mathbf{z}\|_p^p$. These latter cases may turn out to be empirically interesting, as the ℓ^p "norm" for $p \in (0, 1)$ is actually a sharper proxy for the ℓ^0 counting norm (which is the main count for robustness to errors, as discussed in subsequent parts) than the ℓ^1 norm. Since stable distributions exist for all $\|\cdot\|_p$ ($p \in (0, 2]$), our current algorithm and analysis methodology is likely to extend to all $p \in (0, 2]$.

[2] We have suppressed the dependency on other factors, such as $\log \frac{1}{\varepsilon}$ (where ε denotes the target precision) and r to make things concise, because our main interest is mostly in the effect of D on the complexity. Lower order is possible for our specific case by some careful implementation; see, e.g., section 11.8.2 on page 617 of [8]. See also our discussion of running time in section 4.5.

[3] On a more technical level, when the $\mathcal{S}_i$ are fit to sample data, the aforementioned first-order methods may require tuning for optimal performance.

$$\min_{\mathbf{v}\in\mathcal{S}_i} \|\mathbf{Pq} - \mathbf{Pv}\|_1. \tag{1.3}$$

We prove that if the embedded dimension d is sufficiently large, say $d = \text{poly}(r \log n)$ (i.e., d is bounded by some polynomial of $r \log n$), then with constant probability the model $\mathcal{S}_i$ obtained from (1.3) is the same as the one obtained from the original optimization (1.2).

The required dimension d does not depend in any way on the ambient dimension D and is often significantly smaller, e.g., $d = 25$ vs. $D = 32,000$ for one typical example of face recognition. The resulting (small) ℓ^1 regression problems can be solved very efficiently using customized interior point solvers (see, e.g., [31]). These methods are numerically reliable and can yield a speedup of several folds over the standard approach relying on solving (1.2).

The price paid for this improved computational profile is a small increase in the probability of failure of the recognition algorithm, due to the use of a randomized embedding. Our theory quantifies how large d needs to be to render this probability of error under control. Repeated trials with independent projections $\mathbf{P}$ can then be used to make the probability of failure as small as desired. Because ℓ^1 regression is so much cheaper in the low-dimensional space $\mathbb{R}^d$ than in the original space $\mathbb{R}^D$ provided $d \ll D$, these repeated trials are affordable.

The end result is a simple, practical algorithm that guarantees maintaining the good properties of ℓ^1 regression, with substantially improved computational complexity. We demonstrate this on model problems in subspace-based face and object instance recognition. In addition to improved complexity in theory, we observe remarkable improvements on real data examples, suggesting that point-to-subspace query in ℓ^1 could become a practical strategy (or basic building block) for face and object recognition tasks involving large databases, or involving small databases under hard time constraints.

Relationship to existing work. Problem 1.1 is an example of a *subspace search* problem. For 0-dimensional affine subspaces in ℓ^2 (i.e., points), this problem coincides with the nearest neighbor problem. Its approximate version can be solved in time *sublinear* in n, the number of points, using randomized techniques such as locality sensitive hashing [16]. When the dimension r is larger than zero, the problem becomes significantly more challenging. For the case of $r = 1$, sublinear time algorithms exist, although they are more complicated [2].

Recently two groups have proposed approaches to tackling larger r. Basri, Hassner, and Zelnik-Manor [4] lift subspaces into a higher-dimensional vector space (identifying the subspace with its $D \times D$ orthoprojector) and then apply point-based near neighbor search. Jain, Vijayanarasimhan, and Grauman give several random hash functions for the case when the $\mathcal{S}_i$ are hyperplanes [26]. Both of these approaches pertain to ℓ^2 only. Both perform well on numerical examples, but have limitations in theory, as neither is known to yield an algorithm with provably sublinear complexity for all inputs. Results in theoretical computer science suggest that these limitations may be intrinsic to the problem: a sublinear time algorithm for approximate nearest hyperplane search would refute the strong version of the "exponential time hypothesis," which conjectures that general boolean satisfiability problems cannot be solved in time $O(2^{cn})$ for any $c < 1$ [40].

The above algorithms exploit special properties of the ℓ^2 version of Problem 1.1 and do not apply to its ℓ^1 variant. However, the ℓ^1 variant retains the aforementioned difficulties, suggesting that an algorithm for ℓ^1 near subspace search with sublinear dependence on n is

unlikely as well.[4] This motivates us to focus on ameliorating the dependence on D. Our approach is very simple and very natural: Cauchy projections are chosen because the Cauchy family is the unique ℓ^1-stable distribution, i.e., Cauchy projection of any given vector remains i.i.d. Cauchy (see (3.1) and Appendix A for details), a property which has been widely exploited in previous algorithmic work [16, 29, 36].

However, on a technical level, it is not obvious that Cauchy embedding should succeed for this problem. The Cauchy is a heavy tailed distribution, and because of this it does not yield embeddings that very tightly preserve distances between points, as in the Johnson–Lindenstrauss lemma[5] (JL lemma; see [27, 15]). In fact, for ℓ^1, there exist lower bounds showing that certain point sets in ℓ^1 cannot be embedded in significantly lower-dimensional spaces without incurring nonnegligible distortion [9].[6] For a single subspace, embedding results exist, most notably due to Sohler and Woodruff [36], but the distortion incurred is so large as to render them inapplicable to Problem 1.1. Nevertheless, several elegant technical ideas in the proof of [36] turn out to be useful for analyzing Problem 1.1 as well.

The problem studied here is also related to recent work on sparse modeling and sparse error correction. Indeed, one of the strongest technical motivations for using the ℓ^1 norm is its provable good performance in sparse error correction [11, 41]. These results give conditions under which it is possible to recover a vector $\mathbf{v}$ from grossly corrupted observation

$$\mathbf{q} = \mathbf{v} + \mathbf{e},$$

with $\mathbf{v} \in \mathcal{S}$ and the sparse error $\mathbf{e}$ unknown. These results are quite strong: they imply exact recovery, even if the error $\mathbf{e}$ has constant fractions of nonzero entries, of arbitrary magnitude. For example, [11] proves that under technical conditions, ℓ^1 minimization

$$\min \|\mathbf{e}\|_1 \quad \text{such that} \quad \mathbf{q} - \mathbf{e} \in \mathcal{S} \tag{1.4}$$

exactly recovers $\mathbf{e}$ when $\mathcal{S}$ is a linear subspace. Wright and Ma [41] present a similar theory for the case when $\mathcal{S}$ is a union of linear subspaces solved by a variant of optimization in (1.4).

On the other hand, exact recovery may be stronger than what is needed for recognition. For recognition, as formulated in this work, we need only know which subspace minimizes the distance $d(\mathbf{q}, \mathcal{S}_i)$—we do not need to precisely estimate the difference vector itself. The distinction is important: while [42] shows that significant dimensionality reduction is possible if there are no gross errors $\mathbf{e}$, when errors are present, the cardinality of the error vector gives a hard lower bound on the number of observations required for correct recovery. In contrast, for the simpler problem of finding the nearest model, it is possible to give an algorithm that uses very small d and is agnostic to the properties of $\mathbf{q}$ and $\mathcal{S}_1, \ldots, \mathcal{S}_n$.

[4] This could be possible if we are willing to accept time and space complexity exponential in r or D, a la [30].

[5] One version of the lemma (taken from [15]) states the following: for any $\varepsilon \in (0,1)$ and any $n \in \mathbb{N}$, let $k \in \mathbb{N}$ satisfy $k \geq 4\left(\varepsilon^2/2 - \varepsilon^3/3\right)^{-1} \log n$. Then, for any set $\mathcal{V}$ of n points in $\mathbb{R}^d$, there is a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $\mathbf{u}, \mathbf{v} \in \mathcal{V}$, $(1-\varepsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leq \|f(\mathbf{u}) - f(\mathbf{v})\|^2 \leq (1+\varepsilon)\|\mathbf{u}-\mathbf{v}\|^2$. Note in particular that k is independent of the ambient dimension d and depends on n only through its logarithm.

[6] In particular, it is shown in [9] that to keep the distortion within ε, it is necessary that the projection dimension be $n^{\Omega\left(1/\varepsilon^2\right)}$.

Solving the component regression problem in projected space is also reminiscent of research on approximate ℓ^1 regression (see, e.g., [36, 12]). The purpose in that line of work is to efficiently obtain an ε-approximate solution to a single ℓ^1 regression: any $\mathbf{x}$ such that

$$\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_{\ell^1} \leq (1+\varepsilon)\min_{\mathbf{z}} \|\mathbf{y}-\mathbf{A}\mathbf{z}\|_{\ell^1}.$$

Our purpose here is quite different: for a bunch of ℓ^1 regression problems, instead of being concerned with the quality of solving each individual problem, one need only ensure that the regression problem with the smallest objective value remains so after approximation. Moreover, state-of-the-art coreset-based approximation algorithms for ℓ^1 regression such as those in [14, 36, 12] depend heavily on obtaining some importance sampling measure (e.g., ℓ^1 leverage score of an ℓ^1 well-conditioned basis in [12]), which in turn depends on $\mathbf{A}$ and $\mathbf{y}$ simultaneously. In a database-query model that is common in recognition tasks, this complicated dependency directs a lot of computation to query time. By comparison, a considerable portion of computation (e.g., projection of the subspaces) in our framework can be performed during training, rendering the framework attractive when the recognition is under hard time constraints.

Notation. We define the most commonly used notations here. $d_{\ell^1}(\cdot,\cdot)$ is the ℓ^1 distance of a point to a subspace, i.e., $d_{\ell^1}(\mathbf{q},\mathcal{S}) = \min_{\mathbf{v}\in\mathcal{S}} \|\mathbf{q}-\mathbf{v}\|_{\ell^1}$. For any $k\in\mathbb{N}$, $[k]=\{1,\ldots,k\}$, and $\equiv_d$ denotes equality in distribution. Other notations will be defined within the text.

2. Our algorithm and main results.

The flow of our algorithm is summarized as follows. Our main theoretical result states that if d is chosen appropriately, with at least constant probability, the subspace $\mathcal{S}_{i_\star}$ selected will be the original closest subspace $\mathcal{S}_\star$.

Input: n subspaces $\mathcal{S}_1,\ldots,\mathcal{S}_n$ of dimension r and query $\mathbf{q}$
Output: Identity of the closest subspace $\mathcal{S}_\star$ to $\mathbf{q}$

Preprocessing: Generate $\mathbf{P}\in\mathbb{R}^{d\times D}$ with i.i.d. Cauchy random variables ($d \ll D$), and compute the projections $\mathbf{P}\mathcal{S}_1,\ldots,\mathbf{P}\mathcal{S}_n$. Repeat for independent repetitions of $\mathbf{P}$.

Candidate Search: Compute the projection $\mathbf{P}\mathbf{q}$, and compute its ℓ^1 distance to each of $\mathbf{P}\mathcal{S}_i$. Repeat for several versions of $\mathbf{P}$, and locate nearest candidates.

Refined Scanning: Scan the candidates in $\mathbb{R}^D$, and return $\mathcal{S}_\star$.

Theorem 2.1. *Suppose we are given n linear subspaces $\{\mathcal{S}_1,\ldots,\mathcal{S}_n\}$ of dimension r in $\mathbb{R}^D$ and any query point $\mathbf{q}$, and suppose $d_{\ell^1}(\mathbf{q},\mathcal{S}_1) \leq d_{\ell^1}(\mathbf{q},\mathcal{S}_i)/\eta$ for all $i\in[n]\setminus\{1\}$ and some $\eta>1$. Then, for any fixed $\alpha<1-1/\eta$, there exists $d=O[(r\log n)^{1/\alpha}]$ (assuming $n>r$) such that if $\mathbf{P}\in\mathbb{R}^{d\times D}$ is i.i.d. Cauchy, we have*

$$\arg\min_{i\in[n]} d_{\ell^1}(\mathbf{P}\mathbf{q},\mathbf{P}\mathcal{S}_i) = 1 \tag{2.1}$$

with (nonzero) constant probability.

The choice of the first subspace as the nearest is only for notational and expository convenience. Also we write $\arg\min_{i\in[n]} d_{\ell^1}(\mathbf{P}\mathbf{q},\mathbf{P}\mathbf{S}_i)=1$ to mean that the first subspace is the nearest *unambiguously*, i.e., the set of minimizers is a singleton (this comment applies to similar situations below). The condition in Theorem 2.1 depends on several factors. Perhaps the

most interesting is the relative gap η between the closest subspace distance and the second closest subspace distance. Notice that $\eta \in [1, \infty)$ and that the exponent $1/\alpha$ becomes large as η approaches one. This suggests that our dimensionality reduction will be most effective when the relative gap is nonnegligible. For example, when $\eta = 2$ the required dimension is proportional to r^2.

Notice also that d depends on the number of models n only through its logarithm. This rather weak dependence is a strong point, and, interestingly, mirrors the JL lemma for dimensionality reduction in ℓ^2, even though JL-style embeddings are impossible for ℓ^1.

Before stating our overall algorithm, we suggest two additional practical implications of Theorem 2.1. First, Theorem 2.1 guarantees success only with constant probability. This probability is easily amplified by taking T independent trials. Because the probability of failure drops exponentially in T, it usually suffices to keep T rather small. Each of these T trials generates one or more candidate subspaces $\mathbf{S}_i$. We can then perform ℓ^1 regression in $\mathbb{R}^D$ to determine which of these candidates is actually nearest to the query. Note that it may also be possible to perform this second step in $\mathbb{R}^{d'}$, where $d < d' \ll D$.

Second, the importance of the gap η suggests another means of controlling the resources demanded by the algorithm. Namely, if we have reason to believe that η will be especially small (i.e., approaching one), we may instead set d according to the gap between $\xi_{1'}$ and $\xi_{k'}$, for some $k' > 2$, where for any $i \in [n]$, $\xi_{i'}$ denotes the ℓ^1 distance of the query $\mathbf{q}$ to its ith nearest subspace. With this choice, Theorem 2.1 implies that with constant probability the desired subspace is amongst the $k' - 1$ nearest to the query. Again, all of these $k' - 1$ subspaces need to be retained for further examination. However, if $k' \ll n$, this is still a significant saving over the standard approach.

We complement our main result above with a result on the lower bound of the projecting dimension d, which basically says any randomized embedding that is oblivious to the query and subspaces has the target dimension dictated by $\log n, r$ and the reciprocal of $\log \eta_{\min}$, where $\eta_{\min}$ is a nominal relative distance gap (see below), in order to preserve the identity of the nearest subspace with nonnegligible probability.

Theorem 2.2. *Fix any $r, n \in \mathbb{N}, \eta_{\min} \in (1, \infty)$ and $\gamma \in (1/n, 1)$. Let $d \in \mathbb{N}$ satisfy the following: for all $D \geq r$, there exists a distribution μ over $\mathbb{R}^{d \times D}$ such that for all sets $\{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ of r-dimensional subspaces and point $\mathbf{q}$ in $\mathbb{R}^D$ with the property $d_{\ell^1}(\mathbf{q}, \mathcal{S}_1) \leq d_{\ell^1}(\mathbf{q}, \mathcal{S}_i)/\eta_{\min}$ for all $i \in [n]$, one has*

$$\mathbb{P}_{\mathbf{P} \sim \mu} \left[\arg\min_{i \in [n]} d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S}_i) = 1 \right] \geq \gamma. \tag{2.2}$$

Then $d \geq \max(C_1 \frac{1}{\log 3(\eta_{\min}+1)} \log \frac{1}{1-\gamma} \log n - C_2 \frac{r}{\log r}, r)$ for some numerical constants C_1, C_2.

We restrict the probability to be greater than $1/n$ to rule out any case worse than random guessing. The proof is provided in Appendix F. We note that there is a significant gap between the upper bound in Theorem 2.1 and the lower bound in Theorem 2.2. In particular, it is not clear whether $\eta_{\min}$ should enter the bound in its current form, which is extremely bad for small $\eta_{\min}$, or resemble our lower bound, which is significantly milder. Resolving these issues remains an open problem.

3. A sketch of the analysis. In this section, we sketch the analysis leading to Theorem 2.1. The basic rationale for using Cauchy projection is that the standard Cauchy is a *stable* distribution for the ℓ^1 norm: if $\mathbf{v} \in \mathbb{R}^D$ is any fixed vector, and $\mathbf{P} \in \mathbb{R}^{d \times D}$ is a matrix with i.i.d. Cauchy entries, then the vector

$$\mathbf{Pv} \equiv_d \|\mathbf{v}\|_1 \times \mathbf{z}, \tag{3.1}$$

where $\mathbf{z}$ is again an i.i.d. Cauchy vector. In fact, the Cauchy family is also the only stable distribution for the ℓ^1 norm (see Appendix A for more details). So, $\|\mathbf{Pv}\|_1 \equiv_d \|\mathbf{v}\|_1 \|\mathbf{z}\|_1 = \|\mathbf{v}\|_1 \sum_i |z_i|$. The random variables $|z_i|$ are i.i.d. *half-Cauchy*, with PDF

$$f_{\mathcal{HC}}(x) = \frac{2}{\pi} \frac{1}{1+x^2} \quad \text{if } x \geq 0, \tag{3.2}$$

and $f_{\mathcal{HC}}(x) = 0$ for $x < 0$.

In point-to-subspace query, we need to understand how $\mathbf{P}$ acts on many vectors $\mathbf{v}$ simultaneously, including the query $\mathbf{q}$ and all of the subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_n$. Here, we encounter a challenge: although the Cauchy is unambiguously the correct distribution for estimating ℓ^1 norms, it is rather ill-behaved: its mean and variance do not exist, and the sample averages $\frac{1}{n}\sum_i |z_i|$ do not obey the classical central limit theorem.

Figure 1 shows how this behavior affects the point-to-subspace distance $d_{\ell^1}(\mathbf{q}, \mathcal{S})$. The figure shows a histogram of the random variable $\psi = d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S})$, over randomly generated Cauchy matrices $\mathbf{P}$, for two different configurations of query $\mathbf{q}$ and subspace $\mathcal{S}$. Two properties are especially noteworthy. First, the upper tail of the distribution can be quite heavy: with nonnegligible probability, ψ may significantly exceed its median. On the other hand, the lower tail is much better behaved: with very high probability, ψ is not significantly smaller than its median. This inhomogeneous behavior (in particular, the heavy upper tail) precludes very tight distance-preserving embeddings using the Cauchy. However, our goal is *not* to find an embedding of the data, per se, but rather to find the nearest subspace, $\mathcal{S}_\star$, to the query. In

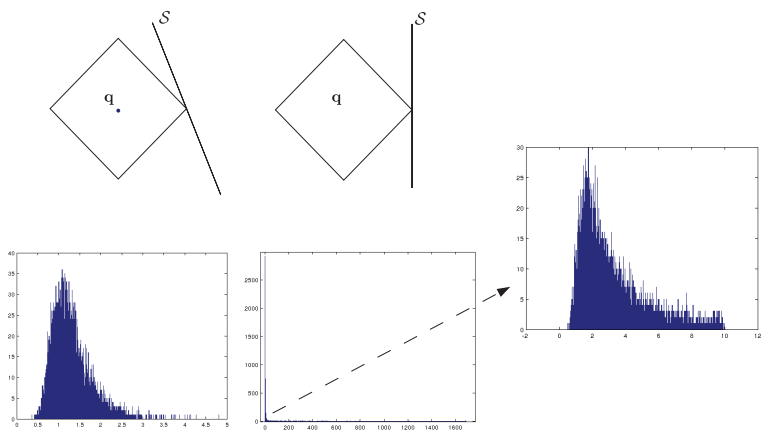

Figure 1. *Statistics of ℓ^1 distance ratios (after vs. before) by random projections over* 10000 *trials. The subspaces are randomly oriented (1st column) and axis-aligned (2nd column), respectively. Here $r = 10$, $D =$ 10000, $d = 35$, and $d_{\ell^1}(q, \mathcal{S}) = 1$.*

fact, for nearest subspace search, this inhomogeneous behavior is much less of an obstacle. To guarantee to find $\mathcal{S}_\star$, we need to ensure qualitatively that

(i) $\mathbf{P}$ does not increase the distance from $\mathbf{q}$ to $\mathcal{S}_\star$ too much, and
(ii) $\mathbf{P}$ does not shrink the distance from $\mathbf{q}$ to any of the other subspaces $\mathcal{S}_i$ too much.

The first property, (i), holds with constant probability: although the tail of ψ is heavy, with probability at least $1/2$, $\psi \leq \text{median}(\psi)$. For the second event, (ii), $\mathbf{P}$ needs to be well behaved on $n-1$ subspaces simultaneously. Notice, however, that for the bad subspaces $\mathcal{S}_i$, the lower tail in Figure 1 is most important. If projection happens to significantly increase the distance between $\mathbf{q}$ and $\mathcal{S}_i$, this will not cause an error (and may even help, in the sense that amplifying the distance to a "bad" subspace renders the chances of the "good" subspace being misdetected (hence failure) less likely). Since the lower tail is sharp, we *can* guarantee that if d is chosen correctly, $\mathbf{Pq}$ will not be significantly closer to any of the $\mathbf{P}\mathcal{S}_i$.

Below we describe some of the technical manipulations needed to carry this argument through rigorously and state key lemmas for each part. Section 3.1 elaborates on property (i), while section 3.2 describes the arguments needed to establish property (ii). Theorem 2.1 follows directly from the results in sections 3.1 and 3.2. This argument, as well as proofs of several routine or technical lemmas, are deferred to the appendices.

3.1. Bounded expansion for the good subspace. Let $\mathbf{v}_\star \in \mathcal{S}_\star$ be a closest point to $\mathbf{q}$ in the ℓ^1 norm before projection:

$$\mathbf{v}_\star \in \arg\min_{\mathbf{v}\in\mathcal{S}_\star} \|\mathbf{q}-\mathbf{v}\|_1.$$

Such a point $\mathbf{v}_\star$ may not be unique, but it always exists. After projection, $\mathbf{Pv}_\star$ might no longer be the closest point to $\mathbf{Pq}$. However, the distance $\|\mathbf{Pq}-\mathbf{Pv}_\star\|_1$ *does* upper bound the distance from $\mathbf{Pq}$ to $\mathbf{P}\mathcal{S}_\star$:

$$d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S}_\star) \;=\; \min_{\mathbf{h}\in\mathbf{P}\mathcal{S}_\star} \|\mathbf{Pq}-\mathbf{h}\|_1 \;\leq\; \|\mathbf{Pq}-\mathbf{Pv}_\star\|_1 \;=\; \|\mathbf{P}(\mathbf{q}-\mathbf{v}_\star)\|_1.$$

Hence, it is enough to show that $\mathbf{P}$ preserves the norm of the particular vector $\mathbf{w}=\mathbf{q}-\mathbf{v}_\star$. We use the following lemma for this purpose, the proof of which can be found in Appendix B.

Lemma 3.1. *There exists a numerical constant $c \in (0,1)$ with the following property. If $\mathbf{w} \in \mathbb{R}^D$ is any fixed vector, $2 \leq d \in \mathbb{N}$, and supposing that $\mathbf{P} \in \mathbb{R}^{d\times D}$ is a matrix with i.i.d. standard Cauchy entries, then*

$$\mathbb{P}\left[\|\mathbf{Pw}\|_1 \leq \frac{2}{\pi} d\log d\, \|\mathbf{w}\|_1\right] \;\geq\; c. \tag{3.3}$$

3.2. Bounded contraction for the bad subspaces. For the "bad" subspaces $\mathcal{S}_2,\ldots,\mathcal{S}_n$, our task is more complicated, since we have to show that under projection $\mathbf{P}$, *no* point in $\mathcal{S}_i$ comes close to $\mathbf{q}$. In fact, we will show something slightly stronger: for appropriate γ, with high probability the following holds for any i:

$$\forall\, \mathbf{w} \in \mathcal{S}_i \oplus \text{span}(\mathbf{q}), \quad \|\mathbf{Pw}\|_1 \;\geq\; \gamma\|\mathbf{w}\|_1. \tag{3.4}$$

Above, $\oplus$ denotes the direct sum of subspaces, so $\tilde{\mathcal{S}}_i = \mathcal{S}_i \oplus \operatorname{span}(\mathbf{q})$ is the linear span of $\mathcal{S}_i$ and the query together. Since for any $\mathbf{v} \in \mathcal{S}_i$, $\mathbf{q} - \mathbf{v} \in \tilde{\mathcal{S}}_i$, whenever (3.4) holds, we have

$$\begin{aligned} d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S}_i) &= \min_{\mathbf{v}\in\mathcal{S}_i} \|\mathbf{Pq} - \mathbf{Pv}\|_1 = \min_{\mathbf{v}\in\mathcal{S}_i} \|\mathbf{P}(\mathbf{q}-\mathbf{v})\|_1 \\ &\geq \min_{\mathbf{v}\in\mathcal{S}_i} \gamma\|\mathbf{q}-\mathbf{v}\|_1 = \gamma\, d_{\ell^1}(\mathbf{q}, \mathcal{S}_i), \end{aligned} \tag{3.5}$$

and the distance to any "bad" subspace $\mathcal{S}_i$ contracts by at most a factor of γ.

To show (3.4), we use a discretization argument. Let Γ denote the intersection of the unit ℓ^1 "sphere" with the expanded subspace $\tilde{\mathcal{S}}_i$:

$$\Gamma = \{\mathbf{w} \mid \|\mathbf{w}\|_1 = 1\} \cap \tilde{\mathcal{S}}_i.$$

Recall that for any set Γ, an *ε-net* is a subset $N_i \subset \Gamma$ such that for every $\mathbf{w} \in \Gamma$, $\|\mathbf{w} - \mathbf{w}'\|_1 \leq \epsilon$ for some $\mathbf{w}' \in N_i$. Standard arguments (see Lemma 3.18 on page 63 of [28]) show that for any $\epsilon > 0$, there exists an ϵ-net N_i for Γ of size at most $(3/\epsilon)^{r+1}$.

Consider the following two events:

(ii.a) $\min_{\mathbf{w}'\in N} \|\mathbf{Pw}'\|_1 \geq \beta$, and

(ii.b) for all $\mathbf{w} \in \tilde{\mathcal{S}}_i$, $\|\mathbf{Pw}\|_1 \leq L\|\mathbf{w}\|_1$.

When both hold, we have for any $\mathbf{w} \in \Gamma$ (with associated closest point $\mathbf{w}' \in N_i$)

$$\|\mathbf{Pw}\|_1 \geq \|\mathbf{Pw}' + \mathbf{P}(\mathbf{w}-\mathbf{w}')\|_1 \geq \|\mathbf{Pw}'\|_1 - \|\mathbf{P}(\mathbf{w}-\mathbf{w}')\|_1 \geq \beta - L\epsilon. \tag{3.6}$$

Moreover, since for any $\mathbf{w} \in \tilde{\mathcal{S}}_i$, $\mathbf{w}/\|\mathbf{w}\|_1 \in \Gamma$, we have that

$$\forall\, \mathbf{w} \in \tilde{\mathcal{S}}_i, \quad \|\mathbf{Pw}\|_1 \geq (\beta - L\epsilon)\|\mathbf{w}\|_1,$$

and we may set $\gamma = \beta - L\epsilon$. So, it is left to establish items (ii.a) and (ii.b) above.

Establishing (ii.a). We use the following tail bound.

Lemma 3.2 (concentration in lower tail). *Let $\mathbf{P} \in \mathbb{R}^{d\times D}$ be an i.i.d. Cauchy matrix. Then, for any fixed vector $\mathbf{w} \in \mathbb{R}^D$ and $\alpha, \delta \in (0,1)$,*

$$\mathbb{P}\left[\|\mathbf{Pw}\|_1 < (1-\alpha)(1-\delta)\frac{2}{\pi} d\log d\, \|\mathbf{w}\|_1\right] < d^{1-\alpha}\exp\left(-\frac{\delta^2}{2\pi}d^\alpha\right). \tag{3.7}$$

In hindsight, the exponent α in the power gives rise to the exponential factor in our bound for d in Theorem 2.1. Fortunately, we are able to establish a concrete lower bound on the probability, which shows that this estimate gives the optimal power. Detailed discussions and proofs are deferred to Appendix C.

This bound is sharp enough to allow us to simultaneously lower bound $\|\mathbf{Pw}'\|_1$ over all $\mathbf{w}' \in N_i$. Set

$$\beta_{\alpha,\delta} = (1-\alpha)(1-\delta)\tfrac{2}{\pi} d\log d,$$

and let $\mathcal{E}_{\text{net},i}$ denote the event that there exists $\mathbf{w}' \in N_i$ with $\|\mathbf{Pw}'\|_1 < \beta_{\alpha,\delta}\|\mathbf{w}'\|_1$:

$$\mathbb{P}[\mathcal{E}_{\text{net},i}] < |N_i|\, d^{1-\alpha}\exp\left(-\tfrac{\delta^2}{2\pi}d^\alpha\right). \tag{3.8}$$

Establishing (ii.b). In bounding the Lipschitz constant L in (ii.b), we have to cope with the heavy tails of the Cauchy, and simple arguments like the above argument for β are insufficient. Rather, we borrow an elegant argument of Sohler and Woodruff [36]. The rough idea is to work with a certain special basis for $\tilde{\mathcal{S}}_i$, which can be considered an ℓ^1 analogue of an orthonormal basis. Just as an orthonormal basis preserves the ℓ^2 norm, an ℓ^1 *well-conditioned basis* approximately preserves the ℓ^1 norm, up to distortion $(r+1)$. The argument then controls the action of $\mathbf{P}$ on the elements of this basis. Due to space limitations, we defer further discussion of this idea to Appendix D and instead simply state the resulting bound.

Lemma 3.3. *Let $\mathbf{P} \in \mathbb{R}^{d\times D}$ be an i.i.d. Cauchy matrix, and let $\mathcal{S}$ be a fixed subspace of dimension $r+1$. Set $L = \sup_{\mathbf{w}\in\mathcal{S}\setminus\{\mathbf{0}\}} \|\mathbf{P}\mathbf{w}\|_1/\|\mathbf{w}\|_1$. Then, for any $B > 0$, we have*

$$\mathbb{P}\left[L > t\,(r+1)\right] \le \frac{2d(r+1)}{\pi B} + \frac{2d}{\pi t}\log\sqrt{1+B^2}. \tag{3.9}$$

The proof of Theorem 2.1 follows from Lemmas 3.1–3.3 above by choosing appropriate values of the parameters B, t, δ, and ϵ. We give the detailed calculation in Appendix E.

Remark 3.4. *We do not allow $\eta = 1$ in Theorem 2.1, corresponding to ties in the nearest subspaces. In this special case, it seems natural that one instead ask the dimension reduction to preserve* any one of the nearest subspaces*; the problem actually becomes easier. To see this, one can fix one of the nearest subspaces as the "good" one, ignore the rest of the nearest, and treat all the rest as "bad" subspaces. Now the new relative distance gap $\eta_{\text{effective}} > 1$, and the number of distances we want to control becomes smaller than the number of subspaces present; hence the problem is actually* easier *as compared to a generic problem setting as in Theorem 2.1 with the same parameters (except for the slightly slacked target, as stated above).*

4. Experiments.

We present three experiments to corroborate our theoretical results and demonstrate their particular relevance to subspace-based robust instance recognition.

4.1. Note on implementation.

Projection matrices and subspaces. Theorem 2.1 is for any fixed set of subspaces and any fixed query point. Of course, if we fix the projection matrix $\mathbf{P}$ and consider many different query points, the success or failure of approximation to each query will be dependent. This suggests sampling a new matrix $\mathbf{P}$ for each new query, which would then require that we reproject each of the subspaces $\{\mathcal{S}_i\}$. In practice, it is more efficient to maintain a pool of k Cauchy projection matrices[7] $\{\mathbf{P}_j\}$ and store $\mathbf{P}_j\mathcal{S}_i$ for each i and j. During testing, we randomly sample a combination of N_{rep} ("rep" for repetition) matrices and corresponding projected subspaces and also apply these projections to the query. This sampling strategy from a finite pool does not generate independent projections for different query points, but it allows economic implementation and empirically still yields an impressive performance. We will specify the values for k and N_{rep} for different experiments.

Solvers for ℓ^1 regression. We perform high-dimensional nearest subspace (NS) search in ℓ^1 (HDL1) as a baseline. Considering the scale of ℓ^1 regression in this case, we employ an augmented Lagrange method (ALM) numerical solver [44] whenever the recognition performance

[7]The standard Cauchy projection matrix $\mathbf{P}$ can be generated as $\mathbf{A}./\mathbf{B}$, where both $\mathbf{A}$ and $\mathbf{B}$ are i.i.d. standard normals and "./" denotes elementwise matrix division.

is not noticeably affected (the case on extended Yale B below); otherwise we employ the more accurate interior point method (IPM) solvers [10] (for the synthesized experiment and ALOI). All the instances of ℓ^1 regression in the projected low dimensions are handled by IPM solvers.

4.2. Experiments with synthesized data. We independently generated $n = 100$ random subspaces in $\mathbb{R}^{10000}$ (i.e., $D = 10000$), each of which is 5-dimensional (i.e., $r = 5$). Each subspace is generated as the column span of a $D \times r$ i.i.d. standard normal matrix. We also prepared a pool of $k = 100$ Cauchy matrices of dimension $d \times D$, where d takes values in $\{10, 30, 50, 70, 90\}$.

To verify our theory (Theorem 2.1), we randomly picked one subspace and generated a sample $\mathbf{y} = \mathbf{Bx}$, where $\mathbf{B}$ is one orthonormal basis for the subspace and $\mathbf{x}$ contains i.i.d. standard normal entries. To induce a reasonable distance gap, and also simulate some sparse errors, we divided $\mathbf{y}$ by the magnitude of its largest entries and added errors that are uniformly distributed in $[-1, 1]$ to a θ-fraction of $\mathbf{y}$'s entries; i.e., we got $\hat{\mathbf{y}} = \mathbf{y} + \mathbf{e}_\theta$. We varied θ from 0.05 with 0.3, with 0.05 as step size. Growth in fraction of corruption diminishes the distance gap η, as evidenced from the legend of the left subfigure in Figure 2. To estimate the success probability of low-dimensional regression to retrieve the *nearest* (in principle not necessarily the originating) subspace, in each setting we exhausted our pool of projection matrices and obtained the empirical success rate. The left subfigure of Figure 2 reports the results. Note that here $r \log n \approx 23$; when the distance gap is not so small, say $\eta > 2$, $d = 30$ actually enjoys at least a 50% chance of preserving the NS. Also, to reasonably get the same level of success probability, small distance gaps evidently entail large projection dimensions.

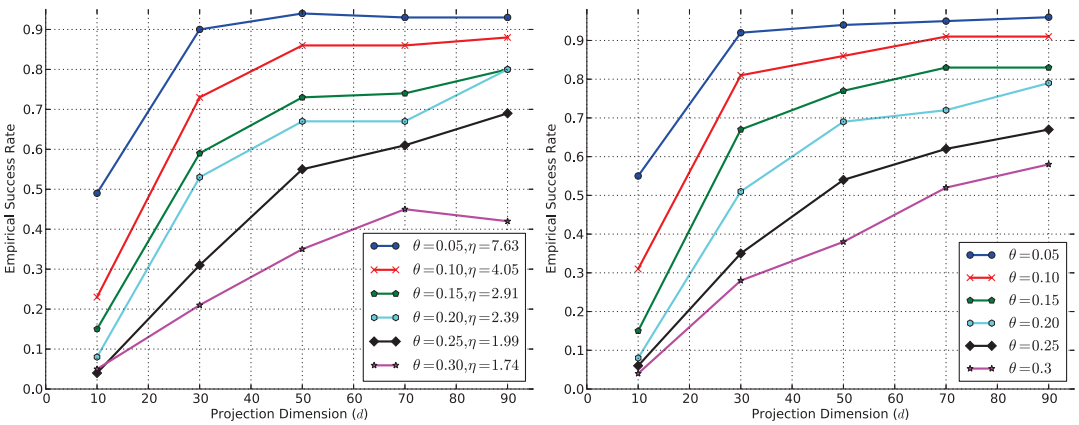

Figure 2. *Left: Probabilities of preserving the NSs by different projection dimensions for a fixed sample corrupted by different levels of additive errors. Right: Fraction of samples that still identify their NS after random projections of different dimensions.*

To emulate visual recognition scenarios such as we will do in the next experiments, we independently randomly generated 500 query points similar to $\hat{\mathbf{y}}$ and also varied θ similarly as above to induce different distance gaps. To keep things simple, for each query we randomly picked up one projection from the pool and omitted repetitions and refined scanning altogether. The success probability is now defined as the fraction of samples that successfully identify their respective NSs in randomly chosen low-dimensional space. The right subfigure

in Figure 2 gives such results. Again, even on this much trimmed version of our algorithm, $d = 30$ helps half of the samples find their NS when the corruption level is below 0.20!

4.3. Robust face recognition on extended Yale B. Under certain physical assumptions, images of one person taken with a fixed pose and varying illumination can be well approximated using a nine-dimensional linear subspace [5]. Because physical phenomena such as occlusions and specularities, as well as physical properties such as nonconvexity [46], may cause violation of the low-dimensional linear model, we formulate the recognition problem as one of finding the closest subspace to $\mathbf{q}$ in ℓ^1 norm [42].[8]

The cropped version of the extended Yale B (EYB) face dataset [22] contains cropped, well-aligned frontal face images (168×192) of 38 subjects under 64 illuminations (2,432 images in total, the 18 corrupted during acquisition not used here). For each subject, we randomly divided the images into two halves, leading to 1205 training images and 1209 test images. To better illustrate the behavior of our algorithm, we strategically divided the test set into two subsets: moderately illuminated (909, *Subset M*) and extremely illuminated (300, *Subset E*). The division is based on the light source direction (w.r.t. the camera axis): images taken with either azimuth angle greater than 90° or elevation angle greater than 60° would be classified as extremely illuminated.[9] Since all faces are supposed to be known, hence the closed-world assumption holds true in this setting.

Recognition with original images. Figure 3 presents the evolution of recognition rate on *Subset M* as the projection dimension (d) grows *with only one repetition of the projection* ($N_{rep} = 1$). We took the subspace dimension to be nine ($r = 9$) as conventional. Our experiment shows that the HDL1 achieves perfect recognition (100%) on this subset, implying that recognition in this subset corresponds perfectly to NS search in ℓ^1. So Figure 3 actually represents the evolution of "average" success probability for *one repetition* over the subset. Suppose the distance gap η is significant such that $1/\alpha \to 1$ (recall α is near $1 - 1/\eta$ in Theorem 2.1); our theorem suggests that one needs to set roughly $d = r \log n = 9 * \log 38 \approx 33$ to achieve a constant probability of success. Our result is consistent with this theoretical prediction, and the probability is already stable above 0.9 for $d \geq 25$. With three repetitions and $d = 25$, the overall recognition rate is 99.56% (four errors out of 909), nearly perfect. Figure 4 presents the failing cases. They either contain significant artifacts or approach the extremely illuminated cases, the failing mechanism and remedy of which are explained below.

For extremely illuminated face images, the ℓ^1 distance gap between the first and second NSs is much less significant (one example is shown in Figure 5). Our theory suggests d should be increased to compensate for the weak gap (because the exponent $1/\alpha$ becomes significant). Our experimental results confirm this prediction. Specifically, for $r = 15$ (we took this to be higher than 9 to account for the great variation due to extreme illuminations in this case), the HDL1 achieves 94.7% accuracy, while our method achieves only 79.3% when $d = 25$ and

[8] In other words, we formulate the problem as ℓ^1 NS search. This is different from the idea of sparse representation in SRC [42] for face recognition. Since our focus here is not to propose a new or optimal face recognition algorithm (although the ℓ^1 NS method happens to be new for the task), we prefer to save detailed discussions in this line for future work. Nevertheless, our preliminary results indeed suggest ℓ^1 NS is as competitive as SRC for the popular EYB face recognition benchmark we have used here.

[9] Note that this division does not closely match in any way the four subset divisions coming with the database, as described in [22].

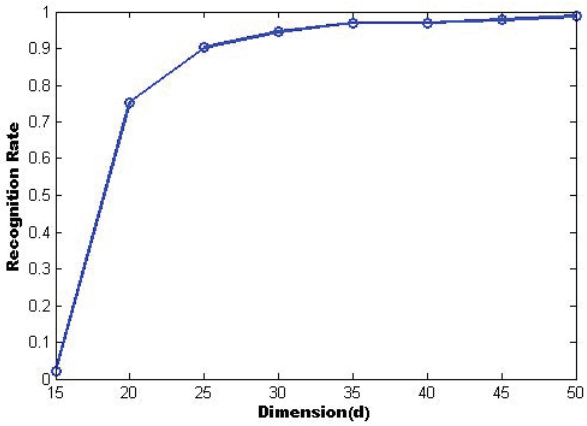

Figure 3. *Recognition rate vs. projection dimension (d)* with one repetition *on* Subset M *face images of EYB. The recognition rate stays stable above* 90% *with* $d \geq 25$. *The high-dimensional NS in* ℓ^1 *achieves perfect* (100%) *recognition. Note that the ambient dimension in this case is* $D = 168 \times 192 = 32256$.

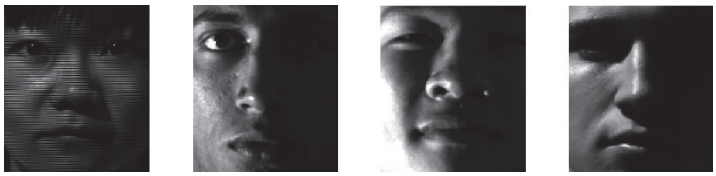

Figure 4. *Failing cases of our method on* Subset M *of EYB.*

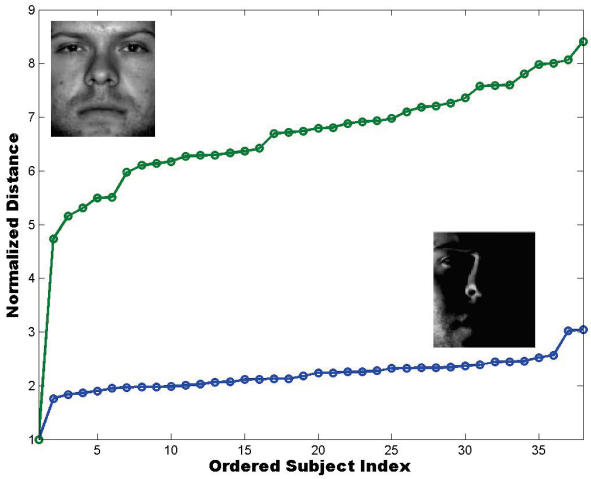

Figure 5. *Samples of moderately/extremely illuminated face images and their* ℓ^1 *distances to other subject subspaces. The subjects have been ordered in ascending order of* ℓ^1 *distance from the sample, and the distances are normalized such that the first distance is* 1. *Note that for the moderately illuminated sample, a distance gap of about* 4.8 *is observed, while this is only about* 1.8 *for the extremely illuminated sample.*

$N_{back} = 5$ (N_{back} is the number of back-research, i.e., "refined scanning" as in the algorithm description, in high dimensions). The recognition rate is boosted significantly when we increase

d or increase N_{back} (this is another way of amplifying the success probability), as evident from Table 1.

Table 1
Recognition rate (%) on Subset E *of EYB with varying d and N_{back}.*

	HDL1	$d = 25$	$d = 50$	$d = 70$
$r = 15, N_{back} = 5$	94.7	79.3	87.7	92.3
$r = 15, N_{back} = 10$	94.7	87.3	92.0	94.0

Recognition on artificially corrupted images. In order to illustrate the robustness of the ℓ^1 NS approach for recognition and particularly the capability of our method of preserving such property of ℓ^1, we emulated the robust recognition experiment on artificially corrupted images, as done in [42]. To be specific, Subset 1 and Subset 2, which comprise images taken under near-frontal illuminations, are used for training; and Subset 3 is used for testing.[10] We corrupted each original test image with (1) randomly distributed sparse corruptions and (2) structured occlusions. For the first setting, we replaced, respectively, 10% to 90% (with 10% resolution) of randomly chosen pixels of the test images with i.i.d. uniform integer values in $[0, 255]$.[11] For the second, the *mandril* image is scaled to, again, 10% to 90% (with 10% resolution) of the image size and imposed on the image with randomly chosen locations. Figure 6 shows some typical samples of both cases, and also the effect of corruptions on distance gaps—corruptions significantly weaken the gaps. In particular, the gap drops to 1 very rapidly as the corruption level increases, suggesting according to our theory that significant dimension reduction via projection is not likely beyond low corruption levels (say 20% from the plot).

To get a flavor of the level of approximation, we fix $k = 100$, $N_{rep} = 5$, $r = 9$, $N_{back} = 5$ and compare the HDL1 with our approximation scheme (dubbed LDL1) for $d = 100$, $d = 200$, and $d = 300$, respectively. To demonstrate the advantage of the ℓ^1 norm in terms of stability against corruptions, we also include comparison with the very natural ℓ^2 NS variant (dubbed HDL2).[12] Figure 7 summarizes the recognition performances for each setting. Our method exhibits a comparable level of performance with the HDL1 for corruptions less than or equal to 20% and observable performance lag beyond that level. This is a reasonable price to pay, as we insist on working in low dimensions for efficiency. In our current setting of the dimension, the performance of LDL1 (not HDL1) is even worse than HDL2 for the random corruption model, in particular when the corruption level is high. For the structured occlusion model, LDL1 is consistently better than HDL2. Increasing d is likely to improve the approximation accuracy further.

4.4. Object instance recognition. To investigate the applicability of our proposal for large-scale recognition tasks, we took a subset of the multipurpose Amsterdam Library of Object Images (ALOI) library [23].[13] This subset comprises images of 1000 toy-like objects with fixed poses, taken under 24 different illumination directions for each object, and hence includes 24 images per object. We randomly took 12 images of each object for training and the

[10] The subset division completely matches the division in [22].
[11] In other words, this means any valid pixel value for an 8-bit gray-scaled image.
[12] This is exactly the NS classifier that was compared to the SRC classifier in [42].
[13] Available online from http://aloi.science.uva.nl/.

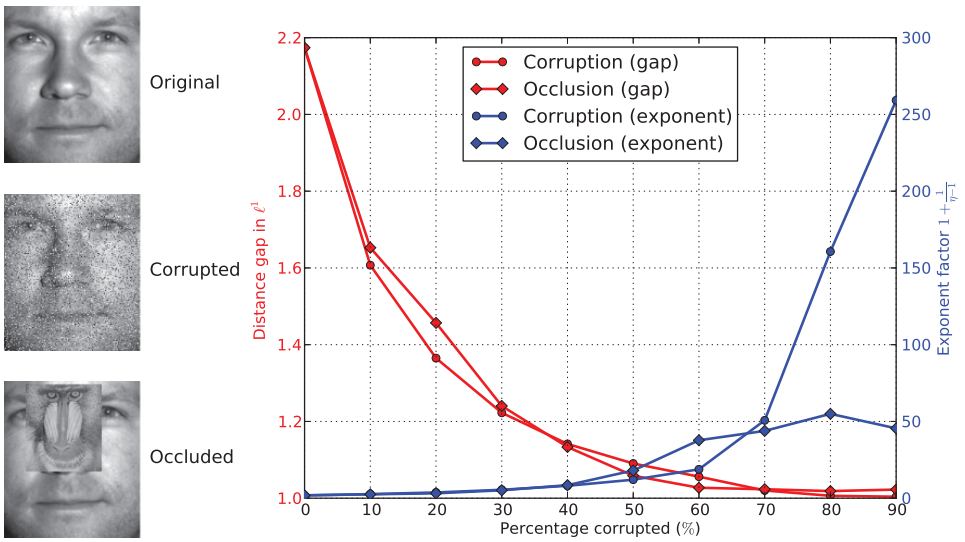

Figure 6. *Left: Sample of original images and the corrupted versions. In both corrupted images* 20% *of the pixels are contaminated. Right: Evolution of the distance gap due to corruptions and the corresponding exponents calculated as* $1+\frac{1}{\eta-1}$ *(in accordance with Theorem* 2.1*). The distance gap is estimated by taking one random example from each test subject.*

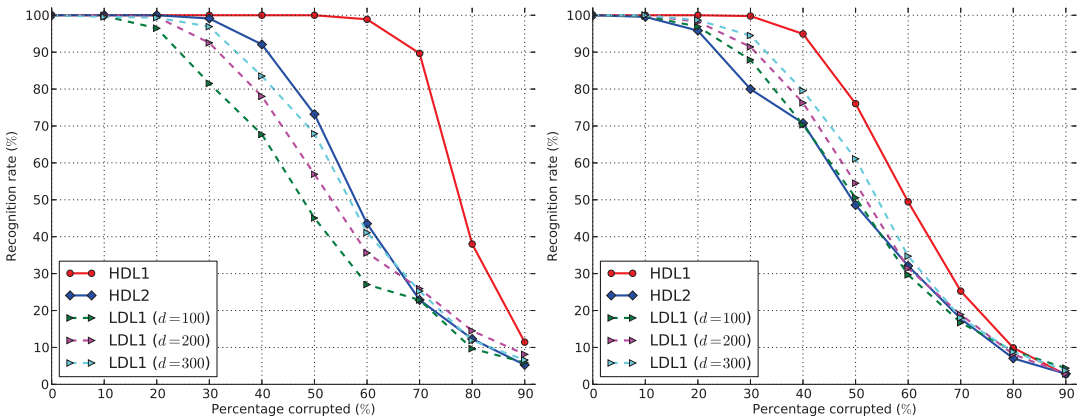

Figure 7. *Recognition rate under corruptions on EYB. (*$k = 100$, $N_{rep} = 3$, *and* $r = 9$ *throughout the experiments.) Left: under random corruptions. Right: under structured occlusion.*

rest for testing. Although these objects in general have nonconvex shapes and non-Lambertian reflectance property, we still approximate the collection of images of each object with a nine-dimensional subspace as proposed in [5]. This again turns the recognition problem naturally into a subspace search problem.

Again we are interested in robust recognition. We added random corruption of varying percentage ($10\% \sim 70\%$) to the test images, similar to the above for face images. We fixed $r = 9$, $k = 100$, $N_{rep} = 10$, $N_{back} = 30$. Table 2 compares the performance of HDL1 and HDL2 under image corruption.

The ℓ^1 NS method again exhibits impressive tolerance to these corruptions, as compared

Table 2

Recognition rate under corruptions for the selected (fixed pose but varying illumination conditions) ALOI subset ($r = 9$, $k = 100$, $N_{rep} = 10$, $N_{back} = 30$).

Corruption level (%)	0	10	20	30	40	50	60	70
HDL1 (%)	99.35	99.40	99.42	99.45	99.47	99.24	43.33	1.85
HDL2(%)	99.72	96.29	59.22	24.30	7.87	1.68	0.53	0.13
LDL1(%, $d = 200$)	99.41	99.10	89.54	66.74	42.62	—	—	—
Distance gap ($\tilde{\eta}$)	4.2858	1.3912	1.2074	1.1339	1.0833	1.0476	1.0117	—

to the ℓ^2 variant.[14] In particular, HDL1 tolerates corruptions up to 50% almost perfectly on the test set. By comparison, HDL2 fails badly for corruption level beyond 10%. Our approximation scheme, LDL1 with $d = 200$, turns out to be effective for corruptions lower than 20% (remains almost $\geq 90\%$ correct) and fails gradually beyond that. We did not try higher projection dimensions, as (1) the computational burden would expand rapidly, and (2) from the estimate in Figure 6, the exponent associated with the predicted dimensions by our theory would be significant for a distance gap lower than 1.2, leading to significant demand for large d.

4.5. Some results on running time. It is obvious that the running time of our algorithm is largely determined by how fast we can solve the ℓ^1 regression problem, i.e., $\min \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1$ for $\mathbf{A} \in \mathbb{R}^{\tilde{d} \times r}$, the cost of which will be denoted as $T_{\ell^1}(\tilde{d}, r)$. To be concrete, in our recognition tasks for object instance recognition, the straightforward exhaustive search in the high dimension $\mathbb{R}^D$ costs a total of $nT_{\ell^1}(D, r)$, whereas the two-level search algorithm we propose costs $nN_{rep}T_{\ell^1}(d, r) + N_{back}T_{\ell^1}(D, r)$ if we project onto a lower-dimensional $\mathbb{R}^d$ and repeat N_{rep} to boost the success probability and then select the best N_{back} for the refined scanning in the original space. So the proposed algorithm will be practically interesting when $T_{\ell^1}(d, r) \ll T_{\ell^1}(D, r)$.

We first experimented with simulated examples. We generate $\mathbf{A}$ as an orthonormal basis for an r-dimensional subspace in $\mathbb{R}^D$, where $D = 2^\rho$ and ρ varies from 4.5 to 15 with 0.5 step size, $r = 10$. For each D, $\mathbf{x}_0 \in \mathbb{R}^r$ is generated as i.i.d. Gaussians, and $y_0 = \mathbf{A}\mathbf{x}_0$. We then perform normalization and corruption addition, the same as we did in section 4.2, with the fraction of corruption θ taken from $\{0.2, 0.4, 0.6, 0.8, 1.0\}$. We take the ℓ^1 regression solver from ℓ^1 magic [10], which implements the customized IPM outlined in section 11.8.2 of [8]. Figure 8 plots the running time (in sec) vs. dimension ($\tilde{d}$), both in a based-2 logarithm. To make the comparison as fair as possible, we have turned on the `-singleCompThread` flag to ensure MATLAB is using only one thread for the simulation. It seems the running time scales approximately as $O(\tilde{d}^2)$. To see how that is relevant to our recognition problem, for $\theta = 0.2$, $T_{\ell^1}(256, 10) = 0.009s$, whereas $T_{\ell^1}(16384, 10) = 41.77s$. The running time differs by several orders of magnitude, giving our algorithm significant advantage!

[14] A systematic report of recognition results on ALOI is rare, with many only on a subset, say 300 objects, perhaps because of the significant scale. One exception is [21], which reports recognition performance under many different settings with state-of-the-art visual recognition schemes. Particularly relevant to our result here is that they evaluated recognition on the illumination subset we choose here with the biologically inspired HMAX model. With 25% of the data for training, they achieved 83.13% recognition rate.

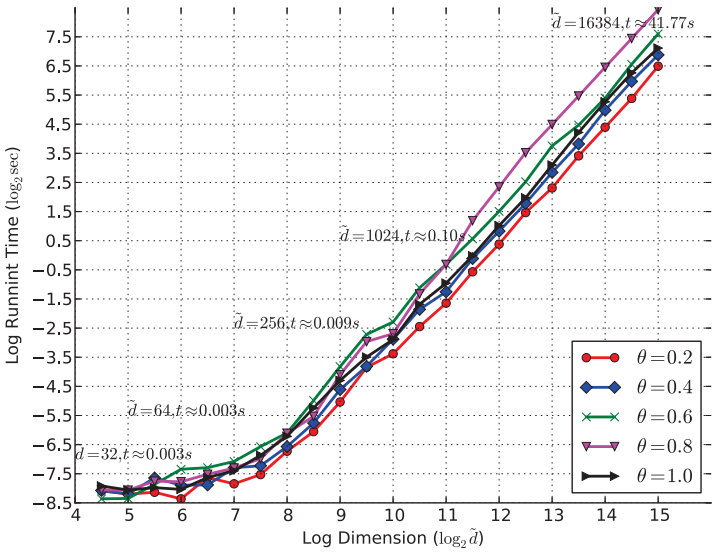

Figure 8. *(Log) running time vs. (log) dimension for the simulated ℓ^1 regression problems. It seems for the customized IPM solver that the complexity scales like $O(\tilde{d}^2)$. The computer runs* 64*bit Ubuntu* 13.10*, with Linux Kernel* 3.11.0-17 *and* 64 *bit MATLAB* 2012*b. The processor is Xeon E*56072.27*G, and the RAM is* 12*G. We did the simulation using only one thread by turning on the* `-singleCompThread` *flag for MATLAB.*

To illustrate what this means in practice, we take a random instance from the Yale B recognition task with 10% random corruptions and take $d = 100$. A previous experiment has confirmed this projection dimension works well for this case (see Figure 7). Again we take $N_{rep} = 5$ and $N_{back} = 5$ for the single-thread simulation; the high dimension exhaustive search costs 7496 secs, while the our two-level search algorithm needs only 467 secs,[15] over 16 times faster! The cost of our algorithm is largely dictated by N_{back} (empirically even smaller because of potential ties). In larger datasets, when N_{back} can be taken to be much smaller relative to n, the advantage could be more significant.

Appendix A. Notation and preliminaries. We present detailed proofs to our technical lemmas throughout the appendices. This part will provide some essential facts about stable distributions, in particular the Cauchy distribution. RV is short for random variable.

Definition A.1 (stable distributions [38, page 43]). *An RV Y is stable if and only if for arbitrary constants c_1 and c_2 there exist constants a and b such that*

$$c_1 Y_1 + c_2 Y_2 \equiv_d a + bY, \tag{A.1}$$

where $Y_1 \equiv_d Y_2 \equiv_d Y$. It is said to be strictly stable if and only if $c_1 Y_1 + c_2 Y_2 \equiv_d bY$ (i.e., one can take $a = 0$.).

Theorem A.2 (characteristic function of stable distributions [47, Theorem C.2]). *A nondegenerate distribution G is stable if and only if its characteristic function $\psi_G(t)$ satisfies*

$$\log \psi_G(t; \alpha, \beta, \gamma, \lambda) = \lambda \left(it\gamma - |t|^\alpha + it\omega_A(t, \alpha, \beta) \right), \tag{A.2}$$

[15] These daunting numbers can be significantly cut down by exploiting multicore/GPU programming. We have exploited multicore programming in our actual experiments over the recognition tasks.

where the real parameters $\alpha \in (0,2]$, $\beta \in [-1,1]$, $\gamma \in (-\infty,\infty)$, *and* $\lambda \in (0,\infty)$ *and*

$$\omega_A(t,\alpha,\beta) = \begin{cases} |t|^{\alpha-1}\beta\tan(\pi\alpha/2) & \text{if } \alpha \neq 1, \\ -\beta(2/\pi)\log|t| & \text{if } \alpha = 1. \end{cases} \tag{A.3}$$

We will use $G^A(x;\alpha,\beta,\gamma,\lambda)$ to denote the stable distribution with characteristic function $\psi_A(t;\alpha,\beta,\gamma,\lambda)$, following the convention in [38]. Also we write $G^A(x;\alpha,\beta)$ when $\gamma = 0$ and $\lambda = 1$, thinking of this setting as the canonical form.

Definition A.3 ((symmetric) ℓ^p-stable distributions). *An RV* X *is called symmetric* ℓ^p*-stable for some* $p \in (0,2]$ *if the characteristic function*

$$\psi_X(t) = \exp(-c|t|^p) \tag{A.4}$$

for some $c > 0$ *and for all* $t \in \mathbb{R}$. *Its distribution is called symmetric* ℓ^p*-stable distribution.*

By comparing the characteristic functions, it is clear that a symmetric ℓ^p-stable distribution is the stable distribution $G^A(x;p,0,0,c)$ for some $c > 0$. It is also obvious that ℓ^p stable distributions exist for all $p \in (0,2]$ by virtue of the existence of the stable distribution with the corresponding parameters.

Lemma A.4 (property of (symmetric) ℓ^p-stable distributions). *Consider i.i.d. RVs* $X_1,\ldots,X_n$ *obeying a symmetric* ℓ^p*-stable distribution. Then, for any real sequence* $\{c_i\}_{i\in[n]}$, *we have*

$$\sum_{i=1}^n c_i X_i \equiv_d \left(\sum_{i=1}^n |c_i|^p\right)^{1/p} X, \tag{A.5}$$

where X *has the same distribution as* X_i*'s.*

Proof. Assume the characteristic functions of X_i's are $\psi(t) = \exp(-c|t|^p)$ for some $c > 0$. Then

$$\psi_{\sum_{i=1}^n c_i X_i}(t) = \mathbb{E}\left[\exp\left(it\sum_{i=1}^n c_i X_i\right)\right] = \prod_{i=1}^n \mathbb{E}\left[(itc_i X_i)\right] \tag{A.6}$$

$$= \prod_{i=1}^n \exp(-c|c_i|^p|t|^p) = \exp\left(-c\sum_{i=1}^n |c_i|^p|t|^p\right) \tag{A.7}$$

$$= \mathbb{E}\left[\exp\left(it\left(\sum_{i=1}^n |c_i|^p\right)^{1/p} X\right)\right] = \psi_{\left(\sum_{i=1}^n |c_i|^p\right)^{1/p} X}(t), \tag{A.8}$$

completing the proof. ∎

We will henceforth omit the word "symmetric" for simplicity when considering ℓ^p-stable distributions. In fact, we will deal exclusively with the standard Cauchy RVs $X \sim \mathcal{C}(0,1)$ with PDF $p_{\mathcal{C}}(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ and the standard half-Cauchy RVs $X \sim \mathcal{HC}(0,1)$ with PDF

$$p_{\mathcal{HC}}(x) = \begin{cases} \frac{2}{\pi}\frac{1}{1+x^2}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{A.9}$$

One remarkable aspect of the standard Cauchy is that it is ℓ^1-stable. Furthermore, by inverting the characteristic function as stated in Definition A.3, one can see that all ℓ^1-stable distribution has to be standard Cauchy or its scaled version (controlled by c) [38]. These facts are fundamental to our subsequent analysis. In addition, the following two-sided bound for the upper tail of a half-Cauchy RV will also be useful.

Lemma A.5. *For $X \sim \mathcal{HC}(0,1)$, we have for all $t \geq 1$,*

$$\frac{1}{\pi}\frac{1}{t} \leq \mathbb{P}[X \geq t] \leq \frac{2}{\pi}\frac{1}{t}. \tag{A.10}$$

Proof. We have

$$\frac{1}{\pi}\frac{1}{t} = \frac{2}{\pi}\int_t^\infty \frac{1}{2x^2}\,dx \leq \mathbb{P}[X \geq t] = \frac{2}{\pi}\int_t^\infty \frac{1}{1+x^2}\,dx \tag{A.11}$$

$$\leq \frac{2}{\pi}\int_t^\infty \frac{1}{x^2}\,dx = \frac{2}{\pi}\frac{1}{t}. \tag{A.12}$$

In fact, the upper bound holds for any $t > 0$. ∎

For any matrix $\mathbf{A}$, we will use $\mathbf{A}_{i*}$ to denote its ith row and $\mathbf{A}_{*j}$ its jth column.

Appendix B. Proof of Lemma 3.1.

We first describe the behavior of the sum of i.i.d. half-Cauchys in the limit, based on the *generalized central limit theorem* (GCLT), which we record below for the sake of completeness.

Theorem B.1 (see [38, GCLT, page 62]). *Let $X_1, \ldots, X_n$ be i.i.d. RVs with the distribution function $F_X(x)$ satisfying the conditions*

$$1 - F_X(x) \sim cx^{-\mu}, \quad x \to \infty, \tag{B.1}$$

$$F_X(x) \sim d|x|^{-\mu}, \quad x \to -\infty, \tag{B.2}$$

with $\mu > 0$.[16] *Then there exist sequences a_n and $b_n > 0$ such that the distribution of the centered and normalized sum*

$$Z_n = \frac{1}{b_n}\left(\sum_{i=1}^n X_i - a_n\right) \tag{B.3}$$

weakly converges to the stable distribution with parameters

$$\alpha = \begin{cases} \mu, & \mu \leq 2, \\ 2, & \mu > 2, \end{cases} \qquad \beta = \frac{c-d}{c+d} \tag{B.4}$$

as $n \to \infty$: $F_{Z_n}(x) \Rightarrow G^A(x; \alpha, \beta)$. In particular, when $\mu = 1$, one can take

$$a_n = \beta(c+d)\, n \log n, \quad b_n = \frac{\pi}{2}(c+d)\, n. \tag{B.5}$$

[16] Note that there are obvious typographical errors in equations (2.5.17) and (2.5.18) in the original theorem statement. This can be seen from, e.g., Theorem 2 of section 35 of Chapter 7 in [24].

Lemma B.2. *Let $X_1, \ldots, X_n$ be i.i.d. half-Cauchy RVs. Consider the sequence*

$$Z_n = \left(\sum_{i=1}^{n} X_i - \frac{2}{\pi} n \log n \right) / n.$$

One has

$$F_{Z_n}(x) \Rightarrow G^A(x; 1, 1). \tag{B.6}$$

Proof. We proceed by determining the parameters μ, c, d, α, β and sequences a_n and b_n as appearing in the GCLT above. For any half-Cauchy RV X, we have

$$1 - F_X(x) = \frac{2}{\pi} \int_x^\infty \frac{1}{1+x^2}\, dx = \frac{2}{\pi} \left(\frac{\pi}{2} - \arctan x \right) = \frac{2}{\pi} \arctan \frac{1}{x}. \tag{B.7}$$

When $x \to \infty$, $\left|\frac{1}{x}\right| \leq 1$. We expand $\arctan \frac{1}{x}$ into an infinite series

$$1 - F_X(x) = \frac{2}{\pi} \arctan \frac{1}{x} = \frac{2}{\pi} \sum_{m=0}^{\infty} \frac{(-1)^m (1/x)^{2m+1}}{2m+1} \sim \frac{2}{\pi} \frac{1}{x} \text{ as } x \to \infty. \tag{B.8}$$

So we have $\mu = 1$, $c = \frac{2}{\pi}$. Since $F_X(x) = 0$ for any $x \leq 0$, $d = 0$. Hence we have

$$\alpha = \mu = 1, \ \beta = \frac{c+d}{c-d} = 1, \tag{B.9}$$

with the centering and normalizing sequences

$$a_n = \beta (c+d) n \log n = \frac{2}{\pi} n \log n, \ b_n = \frac{\pi}{2} (c+d) n = n. \tag{B.10}$$

Hence the sequence Z_n converges weakly to $G^A(x; 1, 1)$ in distribution. ∎

A plot[17] of $G^A(x; 1, 1)$ is included in Figure 9, which will be useful to the following proof.

Proof of Lemma 3.1. By ℓ^1 stability of Cauchy, we have

$$\|\mathbf{P}\mathbf{w}\|_1 = \left\| \sum_{i=1}^{d} \mathbf{P}_{i*}\mathbf{w} \right\|_1 \equiv_d \|\mathbf{w}\|_1 \left\| \sum_{i=1}^{d} \Psi_i \right\|_1 \equiv_d \|\mathbf{w}\|_1 \sum_{i=1}^{d} \Phi_i, \tag{B.11}$$

where $\Psi_1, \ldots, \Psi_d$ are i.i.d. Cauchy and $\Phi_1, \ldots, \Phi_d$ their corresponding half-Cauchys. So we are interested in the behavior of the sequence

$$p_d \doteq \mathbb{P}\left[\sum_{i=1}^{d} \Phi_i \leq \frac{2}{\pi} d \log d \right] = \mathbb{P}\left[\frac{\sum_{i=1}^{d} \Phi_i - \frac{2}{\pi} d \log d}{d} \leq 0 \right]. \tag{B.12}$$

[17] We use implementation available online from http://math.bu.edu/people/mveillet/html/alphastablepub.html. The convention used here (designated with subscript "ST") is almost identical to Zolotarev's form A (designated with subscript "A") in [38], with the following correspondences: $\alpha_{ST} = \alpha_A$, $\beta_{ST} = \beta_A$, $\gamma_{ST} = \lambda_A$, $\delta_{ST} = \gamma_A \lambda_A$.

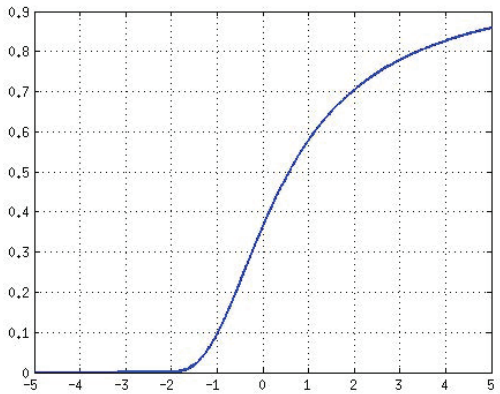

Figure 9. *Plot of the cumulative distribution function (CDF) of the stable distribution* $G^A(x;1,1)$.

Again we consider the sequence $S_d = \left(\sum_{i=1}^d \Phi_i - \frac{2}{\pi} d \log d\right)/d$. Since $S_d \Rightarrow G^A(x;1,1)$ as $d \to \infty$, and any stable distribution has a continuous CDF, we have for $x = 0$,

$$p_d = \mathbb{P}[S_d \le 0] \to G^A(0;1,1) \text{ as } d \to \infty. \tag{B.13}$$

So there exists $N \in \mathbb{N}$ such that for all $d > N$, $p_d \ge 0.3$, where we observe that the numerical value $G^A(0;1,1)$ is strictly greater than 0.3. So one can take the numerical constant c in the lemma as

$$c = \min(p_2, \dots, p_N, 0.3) \ge 0. \tag{B.14}$$

To see $c > 0$, note that for all $d \in \mathbb{N} \setminus \{1\}$,

$$p_d = \mathbb{P}\left[\sum_{i=1}^d \Phi_i \le \frac{2}{\pi} d \log d\right] \ge \mathbb{P}\left[\Phi_i \le \frac{2}{\pi} \log d \ \forall i \in [d]\right] \tag{B.15}$$

$$= \left(\mathbb{P}\left[\Phi_1 \le \frac{2}{\pi} \log d\right]\right)^d = \left[\frac{2}{\pi} \arctan\left(\frac{2}{\pi} \log d\right)\right]^d > 0. \tag{B.16}$$

Hence we complete the proof. ■

Appendix C. Proof of Lemma 3.2.

We will use $\mathbf{1}_{\text{conditional}}$ as the indicator function that assumes either 1 (when the conditional is asserted) or 0 (otherwise).

Proof of Lemma 3.2. Similar to the above it is enough to bound $\sum_{i=1}^d \Phi_i$. For the integer grid $1 < 2 < \cdots < k$, we have

$$\Phi_i \ge \mathbf{1}_{\Phi_i \ge 1} + \mathbf{1}_{\Phi_i \ge 2} + \cdots + \mathbf{1}_{\Phi_i \ge k} \tag{C.1}$$

and hence

$$\sum_{i=1}^d \Phi_i \ge \sum_{j=1}^k \sum_{i=1}^d \mathbf{1}_{\Phi_i \ge j}. \tag{C.2}$$

Notice that $\vartheta_j \doteq \sum_{i=1}^d \mathbf{1}_{\Phi_i \geq j}$ is the sum of d independent Bernoulli RVs with rate $\mathbb{P}[\Phi_1 \geq j]$ and hence $\mathbb{E}[\vartheta_j] = d\mathbb{P}[\Phi_1 \geq j]$. An application of the Chernoff bound gives us

$$\mathbb{P}\left[\vartheta_j < (1-\delta)\, d\mathbb{P}[\Phi_1 \geq j]\right] \leq \exp\left(-\frac{\delta^2 d\mathbb{P}[\Phi_1 \geq j]}{2}\right). \tag{C.3}$$

Now suppose that $\vartheta_j \geq (1-\delta)\, d\mathbb{P}[\Phi_1 \geq j]$ for all $j \in [k]$; in this case, we would have

$$\sum_{i=1}^d \Phi_i \geq \sum_{j=1}^k \sum_{i=1}^d \mathbf{1}_{\Phi_i \geq j} = \sum_{j=1}^k \vartheta_j \quad \text{(by (C.2) and the definition of } \vartheta_k) \tag{C.4}$$

$$\geq d(1-\delta) \sum_{j=1}^k \mathbb{P}[\Phi_1 \geq j] \quad \text{(by our assumption above)} \tag{C.5}$$

$$= d(1-\delta)\frac{2}{\pi} \sum_{j=1}^k \int_j^\infty \frac{1}{1+x^2}\, dx \quad (\Phi_1 \text{ is half-Cauchy}) \tag{C.6}$$

$$= d(1-\delta)\frac{2}{\pi} \sum_{j=1}^k \arctan(1/j) \tag{C.7}$$

$$\geq d(1-\delta)\frac{2}{\pi} \log(k+1) \quad \text{(by Lemma C.1 below).} \tag{C.8}$$

Hence

$$\mathbb{P}\left[\sum_{i=1}^d \Phi_i < (1-\delta)\, d\frac{2}{\pi}\log(k+1)\right] \leq \mathbb{P}\left[\exists\, j \in [k], \vartheta_j < (1-\delta)\, d\mathbb{P}[\Phi_1 \geq j]\right] \tag{C.9}$$

$$\leq \sum_{j=1}^k \exp\left(-\frac{\delta^2 d\mathbb{P}[\Phi_1 \geq j]}{2}\right) \quad \text{(union bound)} \tag{C.10}$$

$$\leq k \exp\left(-\frac{\delta^2 d}{2\pi k}\right). \tag{C.11}$$

It is always true that

$$\mathbb{P}\left[\sum_{i=1}^d \Phi_i < (1-\delta)\, d\frac{2}{\pi}\log d^{1-\alpha}\right] \leq \mathbb{P}\left[\sum_{i=1}^d \Phi_i \leq (1-\delta)\, d\frac{2}{\pi}\log\left(\lfloor d^{1-\alpha}\rfloor + 1\right)\right]. \tag{C.12}$$

Now by setting $k = \lfloor d^{1-\alpha}\rfloor \geq 1$ for the above bound we derived, we have

$$\mathbb{P}\left[\sum_{i=1}^d \Phi_i \leq (1-\delta)\, d\frac{2}{\pi}\log\left(\lfloor d^{1-\alpha}\rfloor + 1\right)\right] \tag{C.13}$$

$$\leq \lfloor d^{1-\alpha}\rfloor \exp\left(-\frac{\delta^2 d}{2\pi}\frac{1}{\lfloor d^{1-\alpha}\rfloor}\right) \tag{C.14}$$

$$\leq d^{1-\alpha} \exp\left(-\frac{\delta^2 d}{2\pi}\frac{1}{d^{1-\alpha}}\right), \tag{C.15}$$

which leads to the result we have claimed. ∎

Lemma C.1. *For all* $k \in \mathbb{N}$, $\sum_{j=1}^{k} \arctan(1/j) \geq \log(k+1)$.

Proof. It is true for $k = 1$ as $\pi/4 > \log(2)$. Now suppose the claim holds for $k-1$, i.e., $\sum_{j=1}^{k-1} \arctan(1/j) \geq \log(k)$; we need to show that it holds for k. It suffices to show that $\arctan(1/k) \geq \log(1+1/k)$. This follows from the fact that $\arctan x \geq \log(1+x)$ for $x \in [0,1]$. ∎

We next show that in some sense the bound we obtained in Lemma 3.2 above cannot be significantly improved.

Lemma C.2. *For any* $d \in \mathbb{N}$ *and any* $\beta \in (0,1)$ *such that* $d^\beta \geq 2$, *if* $\Phi_1, \ldots, \Phi_d$ *are i.i.d. half-Cauchy, then*

$$\mathbb{P}\left[\sum_{i=1}^{d} \Phi_i \leq \frac{2}{\pi}\beta d \log d + O(d)\right] \geq \frac{\exp\left(-Cd^{1-\beta}\right)}{1+\log d}, \tag{C.16}$$

where C *is some numerical constant.*

Proof. Let $k = d^\beta$. Note that when $\Phi_i \leq k$, we have

$$\Phi_i \leq \mathbf{1}_{\Phi_i \geq 0} + \mathbf{1}_{\Phi_i \geq 1} + \cdots + \mathbf{1}_{\Phi_i \geq k}. \tag{C.17}$$

We again define $\vartheta_j \doteq \sum_{i=1}^{d} \mathbf{1}_{\Phi_i \geq j}$ and $p_j \doteq \mathbb{P}[\Phi_1 \geq j]$; then we have

$$\mathbb{P}[\vartheta_k = 0] = (1-p_k)^d \geq \exp(2dp_k \log(1/2)) \geq \exp\left(-Cd^{1-\beta}\right), \tag{C.18}$$

where the second inequality above follows from the fact that $\log(1-y) \geq 2y\log(1/2)$ for $y \in [0, 1/2]$. Moreover, we note that

$$\mathbb{E}[\vartheta_j \mid \vartheta_k = 0] \leq \mathbb{E}[\vartheta_j] = dp_j, \tag{C.19}$$

so we have

$$\mathbb{P}\left[\sum_{i=1}^{d} \Phi_i > t \mid \vartheta_k = 0\right] \tag{C.20}$$

$$\leq \mathbb{P}\left[\sum_{i=1}^{d}\sum_{j=0}^{k} \mathbf{1}_{\Phi_i \geq k} > t \mid \vartheta_k = 0\right] \quad (\vartheta_k = 0 \text{ implies } \Phi_i \leq k \text{ for all } i, \text{ and by (C.17)}) \tag{C.21}$$

$$\leq \mathbb{P}\left[d + \sum_{j=1}^{k} \vartheta_j > t \mid \vartheta_k = 0\right] \quad (\text{exchange summation order and substitute into } \vartheta_k) \tag{C.22}$$

$$\leq \frac{d + \sum_{j=1}^{k} \mathbb{E}[\vartheta_j \mid \vartheta_k = 0]}{t} \quad (\text{by Markov inequality and linearity of expectation}) \tag{C.23}$$

$$\leq \frac{d + d\sum_{j=1}^{k} p_j}{t} \quad (\text{by (C.19)}) \tag{C.24}$$

$$\leq \frac{d + d/2 + 2d/\pi \int_1^k x^{-1}\, dx}{t} \quad (\text{substitute } p_j \text{ and upper bound finite sum by integral}) \tag{C.25}$$

$$= \frac{\frac{2}{\pi}\beta d \log d + \frac{3}{2}d}{t}. \tag{C.26}$$

We set

$$t = \left(1 + \frac{1}{\log d}\right)\left(\frac{2}{\pi}\beta d \log d + \frac{3}{2}d\right) = \frac{2}{\pi}\beta d \log d + O(d). \tag{C.27}$$

Then we have

$$\mathbb{P}\left[\sum_{i=1}^{d} \Phi_i \le t\right] \ge \mathbb{P}[\vartheta_k = 0]\,\mathbb{P}\left[\sum_{i=1}^{d} \Phi_i \le t \mid \vartheta_k = 0\right] \tag{C.28}$$

$$\ge \exp\left(-Cd^{1-\beta}\right)\left(1 - \frac{\log d}{1 + \log d}\right), \tag{C.29}$$

yielding the result. ■

Appendix D. Proof of Lemma 3.3. We will need the definition of well-conditioned basis and some existence lemma to proceed.

Definition D.1 (well-conditioned basis for subspaces [14]). *Let $\mathcal{S}$ be an r-dimensional linear subspace in $\mathbb{R}^D$. For $p \in [1, \infty)$, let $\|\cdot\|_q$ be the dual norm of $\|\cdot\|_p$. Then a matrix $\mathbf{U} \in \mathbb{R}^{D\times r}$ is an (α, β, p)-well-conditioned basis for $\mathcal{S}$ if the following hold: (1) columns of $\mathbf{U}$ are linearly independent; (2) $\|\mathbf{U}\|_p \le \alpha$; and (3) for all $\mathbf{z} \in \mathbb{R}^r$, $\|\mathbf{z}\|_q \le \beta\|\mathbf{U}\mathbf{z}\|_p$. $\mathbf{U}$ is said to be a p-well-conditioned basis for $\mathcal{S}$ if α and β are $r^{\mathcal{O}(1)}$ (i.e., polynomial in r) and independent of D.*

The next lemma asserts the existence of a 1-well-conditioned basis for any r-dimensional subspaces, justified by the existence of the Auerbach basis.

Lemma D.2 (existence of 1-well-conditioned basis [14]). *For any linear subspace $\mathcal{S}$ of dimension r, there exists an $(r, 1, 1)$-well-conditioned basis.*

Proof of Lemma 3.3. Fix a 1-well-conditioned basis $\mathbf{A}$ for $\mathcal{S}$. Suppose that

$$\sum_{j=1}^{r+1} \|\mathbf{P}\mathbf{A}_{*j}\|_1 \le t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1. \tag{D.1}$$

Since any vector $\mathbf{w} \in \mathcal{S}$ can be written as $\mathbf{w} = \mathbf{A}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^{r+1}$,

$$\|\mathbf{P}\mathbf{w}\|_1 = \|\mathbf{P}\mathbf{A}\mathbf{x}\|_1 = \left\|\mathbf{P}\sum_{j=1}^{r+1} \mathbf{A}_{*j}x_j\right\|_1 \le \sum_{j=1}^{r+1} |x_j|\,\|\mathbf{P}\mathbf{A}_{*j}\|_1 \tag{D.2}$$

$$\le \|\mathbf{x}\|_\infty \sum_{j=1}^{r+1} \|\mathbf{P}\mathbf{A}_{*j}\|_1 \le \|\mathbf{x}\|_\infty\, t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1 \tag{D.3}$$

$$\le \|\mathbf{A}\mathbf{x}\|_1\, t\,(r+1) = t\,(r+1)\,\|\mathbf{w}\|_1, \tag{D.4}$$

where the last inequality follows from the definition of a 1-well-conditioned basis. Hence, whenever (D.1) holds, $L \le t(r+1)$, and so

$$\mathbb{P}[L > t(r+1)] \le \mathbb{P}\left[\sum_{j=1}^{r+1} \|\mathbf{P}\mathbf{A}_{*j}\|_1 \ge t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1\right]. \tag{D.5}$$

We finish by upper bounding the probability on the right-hand side, which we define as ϖ. For all $i \in [d]$, $j \in [r+1]$, let $\Psi_{i,j} = |\mathbf{P}_{i*}\mathbf{A}_{*j}| \,/\, \|\mathbf{A}_{*j}\|_1$. Obviously, $\Psi_{i,j}$'s are all half-Cauchy RVs and also $\Psi_{i,j}$'s indexed by the same j are independent. Now

$$\varpi = \mathbb{P}\left[\sum_{j=1}^{r+1} \|\mathbf{P}\mathbf{A}_{*j}\|_1 \geq t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1\right] = \mathbb{P}\left[\sum_{j=1}^{r+1} \left(\|\mathbf{A}_{*j}\|_1 \sum_{i=1}^{d} \Psi_{i,j}\right) \geq t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1\right]. \tag{D.6}$$

Next we partition the probability space and relax a bit to obtain

$$\begin{aligned}
\varpi = \mathbb{P}&\left[\sum_{j=1}^{r+1} \left(\|\mathbf{A}_{*j}\|_1 \sum_{i=1}^{d} \Psi_{i,j}\right) \geq t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1 \;\middle|\; \exists\, \Psi_{i,j} > B\right] \mathbb{P}\left[\exists\, \Psi_{i,j} > B\right] \\
&+ \mathbb{P}\left[\sum_{j=1}^{r+1} \left(\|\mathbf{A}_{*j}\|_1 \sum_{i=1}^{d} \Psi_{i,j}\right) \geq t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1 \;\middle|\; \Psi_{i,j} \leq B\ \forall\, i,j\right] \mathbb{P}\left[\Psi_{i,j} \leq B\ \forall\, i,j\right] \qquad \text{(D.7)} \\
&\leq \mathbb{P}\left[\exists\, \Psi_{i,j} > B\right] \\
&+ \mathbb{P}\left[\sum_{j=1}^{r+1} \left(\|\mathbf{A}_{*j}\|_1 \sum_{i=1}^{d} \Psi_{i,j}\right) \geq t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1 \;\middle|\; \Psi_{i,j} \leq B\ \forall\, i,j\right] \mathbb{P}\left[\Psi_{i,j} \leq B\ \forall\, i,j\right]. \qquad \text{(D.8)}
\end{aligned}$$

Applying union bound to the first term and Markov inequality to the conditional probability in the second term, we have

$$\begin{aligned}
\varpi &\leq \frac{2d\,(r+1)}{\pi B} + \frac{\sum_{j=1}^{r+1} \left(\|\mathbf{A}_{*j}\|_1 \,\mathbb{E}\left[\sum_{i=1}^{d} \Psi_{i,j} \mid \Psi_{i,j} \leq B, \forall\, i,j\right]\right)}{t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1} \mathbb{P}\left[\Psi_{i,j} \leq B\ \forall\, i,j\right] \qquad \text{(D.9)} \\
&= \frac{2d\,(r+1)}{\pi B} + \frac{\left(\sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1\right) \mathbb{E}\left[\sum_{i=1}^{d} \Psi_{i,j} \mid \Psi_{i,j} \leq B, \forall\, i,j\right]}{t \sum_{j=1}^{r+1} \|\mathbf{A}_{*j}\|_1} \mathbb{P}\left[\Psi_{i,j} \leq B\ \forall\, i,j\right] \qquad \text{(D.10)} \\
&= \frac{2d\,(r+1)}{\pi B} + \frac{d\,\mathbb{E}\left[\Psi_{1,1} \mid \Psi_{1,1} \leq B\right]}{t} \mathbb{P}\left[\Psi_{i,j} \leq B\ \forall\, i,j\right], \qquad \text{(D.11)}
\end{aligned}$$

where in the last step we take $j = 1$ with loss of generality, as $\Psi_{i,j}$'s are i.i.d. half-Cauchy for any fixed j. We now define a new RV $\Psi_{1,1}^B$ as

$$\Psi_{1,1}^B = \begin{cases} \Psi_{1,1}, & \Psi_{1,1} \leq B, \\ 0, & \Psi_{1,1} > B \end{cases} \tag{D.12}$$

and note the fact that $\mathbb{E}\left[\Psi_{1,1} \mid \Psi_{1,1} \leq B\right] = \mathbb{E}\left[\Psi_{1,1}^B\right] / \mathbb{P}\left[\Psi_{1,1} \leq B\right]$; hence

$$\varpi \leq \frac{2d\,(r+1)}{\pi B} + \frac{d\,\mathbb{E}\left[\Psi_{1,1}^B\right]}{t\mathbb{P}\left[\Psi_{1,1} \leq B\right]} \mathbb{P}\left[\Psi_{i,j} \leq B\ \forall\, i,j\right] \leq \frac{2d\,(r+1)}{\pi B} + \frac{d\,\mathbb{E}\left[\Psi_{1,1}^B\right]}{t}, \tag{D.13}$$

where we have used the fact that $\mathbb{P}\left[\Psi_{i,j} \leq B\ \forall\, i,j\right] \leq \mathbb{P}\left[\Psi_{1,1} \leq B\right]$. We arrive at the claimed results by substituting the expectation

$$\mathbb{E}\left[\Psi_{1,1}^B\right] = \frac{2}{\pi} \int_0^B \frac{x}{1+x^2}\, dx = \frac{1}{\pi} \log\left(1+x^2\right)\Big|_0^B = \frac{1}{\pi} \log\left(1+B^2\right). \tag{D.14}$$

This completes the proof. ∎

Appendix E. Summing up: Proof of Theorem 2.1.

Proof of Theorem 2.1. By Lemma 3.1, with probability at least c,

$$d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S}_1) \;\le\; \left(\frac{2}{\pi} d \log d\right) d_{\ell^1}(\mathbf{q}, \mathcal{S}_1). \tag{E.1}$$

We apply Lemmas 3.2 and 3.3 to obtain a probabilistic lower bound on $d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S}_j)$ for each $j = 2, \dots, n$. As above, let $\tilde{\mathcal{S}}_j = \mathcal{S}_j \oplus \{\mathbf{q}\}$ denote the direct sum of $\mathcal{S}_j$ and the query point. Let N_j denote an ε-net for the intersection of $\tilde{\mathcal{S}}_j$ with the ℓ^1 ball, with size at most $(3/\varepsilon)^{r+1}$. Standard arguments guarantee the existence of such a net.

Applying Lemma 3.2 to each of the N_j, we obtain that

$$\|\mathbf{Pw}\|_1 \;\ge\; (1-\alpha)(1-\delta)\frac{2}{\pi} d \log d \tag{E.2}$$

for every $\mathbf{w} \in N_j$ and every $j \in \{2, \dots, n\}$, simultaneously, with probability at least

$$1 - (n-1)\left(\frac{3}{\varepsilon}\right)^{r+1} d^{1-\alpha} \exp\left(-\frac{\delta^2}{2\pi} d^{\alpha}\right). \tag{E.3}$$

At the same time, applying Lemma 3.3 to each $\tilde{S}_j$, we obtain that

$$\|\mathbf{Pw}\|_1 \;\le\; t(r+1)\,\|\mathbf{w}\|_1 \tag{E.4}$$

simultaneously for every $\mathbf{w} \in \tilde{\mathcal{S}}_j$, for each $j \in \{2, \dots, n\}$, with probability at least

$$1 - \frac{2d\,(r+1)\,(n-1)}{\pi B} - \frac{2d\,(n-1)}{\pi t} \log\sqrt{1+B^2}. \tag{E.5}$$

Here, $B > 0$ can be chosen freely to obtain the tightest possible bound on the probability of failure. For notational convenience, write

$$\xi = \frac{t(r+1)\varepsilon}{\frac{2}{\pi} d \log d}, \tag{E.6}$$

and notice that on the intersection of the good events introduced above, for every $\mathbf{h} \in \tilde{\mathcal{S}}_j$ with $\|\mathbf{h}\|_1 \le \varepsilon$,

$$\|\mathbf{Ph}\|_1 \;\le\; \left(\frac{2}{\pi} d \log d\right) \xi. \tag{E.7}$$

Consider an arbitrary $\mathbf{w} \in \mathcal{S}_j$. We can write

$$\frac{\mathbf{q} - \mathbf{w}}{\|\mathbf{q} - \mathbf{w}\|_1} = \mathbf{z} + \mathbf{h}, \tag{E.8}$$

with $\mathbf{z} \in N_j$, $\mathbf{h} \in \tilde{\mathcal{S}}_j$, and $\|\mathbf{h}\|_1 \le \varepsilon$. Applying $\mathbf{P}$ to both sides and using the triangle inequality, we obtain that

$$\begin{aligned} \|\mathbf{Pq} - \mathbf{Pw}\|_1 &\ge (\|\mathbf{Pz}\|_1 - \|\mathbf{Ph}\|_1)\, \|\mathbf{q} - \mathbf{w}\|_1 \\ &\ge \left(\frac{2}{\pi} d \log d\right)\, ((1-\alpha)(1-\delta) - \xi)\, \|\mathbf{q} - \mathbf{w}\|_1 . \end{aligned} \tag{E.9}$$

Hence, on the intersection of the good events introduced above, for each $j = 2, \dots, n$,

$$\begin{aligned} d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S}_j) &\ge \left(\frac{2}{\pi} d \log d\right)\, ((1-\alpha)(1-\delta) - \xi)\, d_{\ell^1}(\mathbf{q}, \mathcal{S}_j) \\ &\ge \left(\frac{2}{\pi} d \log d\right)\, ((1-\alpha)(1-\delta) - \xi)\, \eta\, d_{\ell^1}(\mathbf{q}, \mathcal{S}_1) \\ &\ge ((1-\alpha)(1-\delta) - \xi)\, \eta\, d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S}_1). \end{aligned} \tag{E.10}$$

So, as long as

$$(1-\alpha)(1-\delta) - \xi > 1/\eta, \tag{E.11}$$

the algorithm will succeed, except on an event of probability at most

$$\begin{aligned} \phi \;\doteq\; 1 - c + (n-1)\left(\frac{3}{\varepsilon}\right)^{r+1} d^{1-\alpha} \exp\left(-\frac{\delta^2}{2\pi} d^\alpha\right) \\ + \frac{2d\,(r+1)\,(n-1)}{\pi B} + \frac{2d\,(n-1)}{\pi t} \log\sqrt{1+B^2}. \end{aligned} \tag{E.12}$$

Our remaining task is to show that with the specified choice of d, (E.11) is satisfied, and the failure probability ϕ in (E.12) is bounded away from one by a constant.

Set $\zeta = 1 - \frac{1}{\eta} - \alpha$. By assumption, $\zeta > 0$. We will set $\delta = \zeta/3$ and ensure that $\xi \le \zeta/3$, which will imply that

$$(1-\alpha)(1-\delta) - \xi \;\ge\; 1 - \alpha - 2\zeta/3 \;>\; 1/\eta, \tag{E.13}$$

ensuring that (E.11) is satisfied. We choose

$$B = \frac{4}{c}\left(\frac{2d\,(r+1)\,(n-1)}{\pi}\right), \quad t = \frac{4}{c}\left(\frac{2}{\pi} d\,(n-1)\right) \cdot 2 \cdot 4 \cdot \log\left[\max\left(\frac{8}{c\pi}, d, r+1, n-1\right)\right].$$

These choices ensure that the quantity $2d(r+1)(n-1)/\pi B$ in (E.12) is at most $c/4$. Moreover, using that $B \ge 16/\pi \ge (1+\sqrt{5})/2$ and the crude bound $\log\sqrt{1+B^2} \le 2\log B$ for all $B \ge (1+\sqrt{5})/2$, we can show that the final term in (E.12) is at most $c/4$, giving

$$\begin{aligned} \phi &\le 1 - \frac{c}{2} + (n-1)\left(\frac{3}{\varepsilon}\right)^{r+1} d^{1-\alpha} \exp\left(-\frac{\delta^2}{2\pi} d^\alpha\right) \\ &= 1 - \frac{c}{2} + \exp\left(-\frac{\zeta^2}{18\pi} d^\alpha + (1-\alpha)\log d + (r+1)\log(3/\varepsilon) + \log(n-1)\right). \end{aligned} \tag{E.14}$$

It remains to choose ε and bound the exponential term above. We set

$$\varepsilon \;=\; \left(\frac{2}{\pi} d \log d\right) \frac{\zeta}{3} \frac{1}{t(r+1)}. \tag{E.15}$$

This ensures that $\xi = \frac{t(r+1)\varepsilon}{(2/\pi) d \log d} \leq \frac{\zeta}{3}$, as promised. Plugging in for t, we obtain

$$\varepsilon \;\geq\; \frac{C_1\, c\, \zeta\, \log d}{(n-1)(r+1)\log\left[\max\left(d, r+1, n-1\right)\right]}, \tag{E.16}$$

where C_1 is a numerical constant. Using the assumption that $n > r$, we can simplify this bound to

$$\varepsilon \;\geq\; \frac{C_2\, c\, \zeta}{n^2\, \log n}, \tag{E.17}$$

with C_2 numerical. The exponential term in (E.12) is then at most

$$\exp\left(-C_3 \zeta^2 d^{\alpha} + (1-\alpha)\log d + C_4 r \log\left(\frac{n}{c\zeta}\right)\right). \tag{E.18}$$

To ensure that this term is bounded by $c/4$, and hence the probability of failure is bounded away from one by a constant, it suffices to ensure that

$$d \;\geq\; C_5 \left(\frac{\log d + r \log\left(\frac{n}{c\zeta}\right) + \log\left(\frac{4}{c}\right)}{\zeta^2}\right)^{1/\alpha}. \tag{E.19}$$

■

Appendix F. Proof of Theorem 2.2. From a high level, our proof proceeds by exploiting the approximate subspace search to solve the sparse recovery problem. Invoking some known lower bounds for the sparse recovery problem, we arrive at the bound as stated in Theorem 2.2. We first record/show some useful results.

Proposition F.1 (number of measurements for stable sparse recovery [17, Theorem 5.2]). *For any constant $C \geq 1$, if any distribution μ over $\mathbb{R}^{m\times t}$ and any algorithm $\mathscr{A}$ obey that for all $\mathbf{x} \in \mathbb{R}^t$ and $\mathbf{A} \sim \mu$, $\hat{\mathbf{x}} = \mathscr{A}(\mathbf{A}\mathbf{x})$ and*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_1 \leq C \min_{\|\mathbf{x}'\|_0 \leq k} \|\mathbf{x} - \mathbf{x}'\|_1 \tag{F.1}$$

with probability at least $p > 3/4$, we must have $m \geq C_1 \frac{1}{2+2\log(2C+3)} k \log t/k$ for some numerical constant C_1.

The dependency of m on the approximation factor C is directly extracted from the proof of Theorem 5.2 in [17].

Proposition F.2 (approximate subset query [33, Theorem 3.1]). *There is a randomized sparse binary matrix $\mathbf{A}$ with $O\left(\frac{c}{\varepsilon}k\right)$ rows and recovery algorithm $\mathscr{A}$ such that for all $\mathbf{x} \in \mathbb{R}^t$ and $\mathcal{S} \subset [t]$ with $|\mathcal{S}| = k$, $\mathbf{x}' = \mathscr{A}(\mathbf{A}\mathbf{x}, \mathcal{S}) \in \mathbb{R}^t$ has $\operatorname{supp}(\mathbf{x}') \subset \mathcal{S}$ and*

$$\|\mathbf{x}' - \mathbf{x}_{\mathcal{S}}\|_1 \leq \varepsilon \|\mathbf{x} - \mathbf{x}_{\mathcal{S}}\|_1 \tag{F.2}$$

with probability at least $1 - 1/k^c$.[18]

Proof of Theorem 2.2. Consider the following distribution μ (on $\mathbf{A}$) and algorithm $\mathscr{A}$ for the k-sparse recovery problem as defined in Proposition F.1.

- μ is a distribution on $\mathbf{A} = \left[\begin{smallmatrix}\mathbf{A}_C \\ \mathbf{A}_B\end{smallmatrix}\right]$, where $\mathbf{A}_C$ comprises ℓ blocks of projection matrices, $\mathbf{A}_C^1, \ldots, \mathbf{A}_C^\ell \in \mathbb{R}^{m \times t}$, from the same distribution ν, stacked vertically, and $\mathbf{A}_B \in \mathbb{R}^{m' \times t}$ is a randomized sparse binary matrix with $m' = O\left(\frac{c\ell}{\varepsilon} k\right)$ rows from a distribution that verifies Proposition F.2. The distribution ν and parameters m, ℓ, c, ε, and C are specified below.
- For any $\mathbf{x} \in \mathbb{R}^t$, $\mathscr{A}$ comprises two steps, given $\mathbf{A}\mathbf{x}$:
  1. Identifying a subset of coordinates of $\mathbf{x}$ that probably contains large (in magnitude) elements. Suppose we target detecting the support of the largest k elements of $\mathbf{x}$. This is equivalent to identifying the nearest, out of the $\binom{t}{k}$ k-dimensional canonical subspaces [spanned by any k of the t canonical basis vectors (i.e., $\mathbf{e}_1, \ldots, \mathbf{e}_t$)], to $\mathbf{x}$ in the sense of ℓ^1 point-to-subspace distance.
     Let ν and m be a distribution-projection dimension pair that satisfies the hypothesis of Theorem 2.2 with the parameter tuples $\left(k, \binom{t}{k}, \eta_{\min}, \gamma\right)$. In particular, this means that if the canonical subspaces and $\mathbf{x}$ obey the gap condition dictated by $\eta_{\min}$, given $\mathbf{A}_C^i \mathbf{x}$, for all $i \in [\ell]$, we can identify the k significant supports as desired with probability at least γ. This is not true for all $\mathbf{x}$, however. Instead, w.l.o.g. assuming the first canonical subspace is the nearest, consider the following "partitioning"[19] of canonical subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_{\binom{t}{k}}$:
     - $\{1\}$
     - $\mathcal{I} \doteq \left\{\kappa \in \left[\binom{t}{k}\right] \setminus \{1\} : d_{\ell^1}(\mathbf{x}, \mathcal{S}_\kappa) < \eta_{\min} d_{\ell^1}(\mathbf{x}, \mathcal{S}_1)\right\}$
     - $\mathcal{J} \doteq \left\{\kappa \in \left[\binom{t}{k}\right] \setminus \{1\} : d_{\ell^1}(\mathbf{x}, \mathcal{S}_\kappa) \geq \eta_{\min} d_{\ell^1}(\mathbf{x}, \mathcal{S}_1)\right\}$

     Then $\mathcal{I} = \emptyset$ corresponds to cases when distance gap $\eta_{\min}$ is obeyed, so for all $i \in [\ell]$

     (F.3)
     $$\mathbb{P}\left[\operatorname*{arg\,min}_{\kappa \in \left[\binom{t}{k}\right]} d_{\ell^1}\left(\mathbf{A}_C^i \mathbf{x}, \mathbf{A}_C^i \mathcal{S}_\kappa\right) \in \{1\} \cup \mathcal{I}\right] = \mathbb{P}\left[\operatorname*{arg\,min}_{\kappa \in \left[\binom{t}{k}\right]} d_{\ell^1}\left(\mathbf{A}_C^i \mathbf{x}, \mathbf{A}_C^i \mathcal{S}_\kappa\right) = 1\right] \geq \gamma.$$

     If $\mathcal{J} = \emptyset$, for all $i \in [\ell]$,

     (F.4)
     $$1 = \mathbb{P}\left[\operatorname*{arg\,min}_{\kappa \in \left[\binom{t}{k}\right]} d_{\ell^1}\left(\mathbf{A}_C^i \mathbf{x}, \mathbf{A}_C^i \mathcal{S}_\kappa\right) \in \{1\} \cup \mathcal{I}\right] \geq \gamma.$$

     When $\mathcal{I} \neq \emptyset$ and $\mathcal{J} \neq \emptyset$, we consider in addition a spurious set $\mathcal{I}'$ with $|\mathcal{I}'| = |\mathcal{I}|$,

[18] For any vector $\mathbf{x}$, $\mathbf{x}_\Omega$ is a vector of the same length of $\mathbf{x}$, with coordinates in Ω^c set to 0; $\mathbf{x}_{\bar{\Omega}}$ is a restriction of $\mathbf{x}$ to its subvector indexed by Ω. The same convention applies to the matrix setting in a similar manner.

[19] The division may not be partitioning in a strictly mathematical sense, since $\mathcal{I}$ or $\mathcal{J}$ may be empty.

which consists of random duplicates of subspaces in $\mathcal{J}$. So in this case

$$\mathbb{P}\left[\underset{\kappa\in\left[\binom{t}{k}\right]}{\arg\min}\, d_{\ell^1}\left(\mathbf{A}_C^i\mathbf{x}, \mathbf{A}_C^i\mathcal{S}_\kappa\right) \in \{1\}\cup\mathcal{I}\right] \tag{F.5}$$

$$\geq \mathbb{P}\left[\underset{\kappa\in\{1\}\cup\mathcal{J}}{\arg\min}\, d_{\ell^1}\left(\mathbf{A}_C^i\mathbf{x}, \mathbf{A}_C^i\mathcal{S}_\kappa\right) = 1\right] \tag{F.6}$$

$$\geq \mathbb{P}\left[\underset{\kappa\in\{1\}\cup\mathcal{J}\cup\mathcal{I}'}{\arg\min}\, d_{\ell^1}\left(\mathbf{A}_C^i\mathbf{x}, \mathbf{A}_C^i\mathcal{S}_\kappa\right) = 1\right] \geq \gamma. \tag{F.7}$$

So in any case $\mathbf{A}_C^i$, $i\in[\ell]$ is enough to guarantee a constant probability of success γ, in order to identify one subspace that is within $\eta_{\min}$ of the best in terms of distance to $\mathbf{x}$. Denote the corresponding supports identified by the ℓ independent runs by Ω_i; for all $i\in[\ell]$ and $\Pi \doteq \cup_{i=1}^{\ell}\Omega_i$, we have

$$\mathbb{P}\left[\exists\, \mathcal{S}\subset\Pi : |\mathcal{S}| = k, \|\mathbf{x}_{\mathcal{S}^c}\|_1 \leq \eta_{\min}\min_{|\mathcal{T}|=k}\|\mathbf{x}_{\mathcal{T}^c}\|_1\right] \geq 1-(1-\gamma)^\ell. \tag{F.8}$$

We choose

$$\ell = -\log 5/\log(1-\gamma) \tag{F.9}$$

to make this probability at least $4/5$.

2. Estimating the value of $\mathbf{x}$ on the support from step 1. We denote $k' = |\Pi| \leq k\ell$. Given Π, by Proposition F.2, we can obtain an $\hat{\mathbf{x}}$ with $\mathbf{A}_B\mathbf{x}$ that obeys $\operatorname{supp}(\hat{\mathbf{x}})\subset\Pi$, and

$$\|\hat{\mathbf{x}}-\mathbf{x}_\Pi\|_1 \leq \varepsilon\|\mathbf{x}-\mathbf{x}_\Pi\|_1 \tag{F.10}$$

with probability at least $15/16$, provided

$$k^c \geq 16 \Longrightarrow c \geq \log 16/\log k. \tag{F.11}$$

Putting together the above constructions, with probability at least $4/5\times 15/16 = 3/4$, $\hat{\mathbf{x}}$ above satisfies

$$\|\hat{\mathbf{x}}-\mathbf{x}\|_1 = \|\hat{\mathbf{x}}-\mathbf{x}_\Pi\|_1 + \|\mathbf{x}_{\Pi^c}\|_1 \leq \varepsilon\|\mathbf{x}-\mathbf{x}_\Pi\|_1 + \|\mathbf{x}_{\Pi^c}\|_1 \tag{F.12}$$

$$\leq (1+\varepsilon)\|\mathbf{x}_{\Pi^c}\|_1 \leq (1+\varepsilon)\,\eta_{\min}\min_{\|\mathbf{x}'\|_0\leq k}\left\|\mathbf{x}-\mathbf{x}'\right\|_1. \tag{F.13}$$

Hence this $(\mu,\mathscr{A})$ pair respects the hypothesis in Proposition F.1, and so $\mathbf{A}$ must have at least $C_1 k\log(t/k)/[2+2\log(2(1+\varepsilon)\eta_{\min}+3)]$ rows for some constant C_1, or each $\mathbf{A}_C^i$ must have

$$\frac{1}{\ell}\left[\frac{C_1 k\log(t/k)}{2+2\log(2(1+\varepsilon)\eta_{\min}+3)} - C_2\frac{c\ell k}{\varepsilon}\right]$$

$$= C_1'\frac{1}{2+2\log(2(1+\varepsilon)\eta_{\min}+3)}\log\frac{1}{1-\gamma}k\log(t/k) - C_2'\frac{k}{\log k}\frac{1}{\varepsilon} \tag{F.14}$$

rows for some constants C_2, C_1', and C_2'. Note that we have $n = \binom{t}{k}$ subspaces in each subspace search problem, and hence by taking $\varepsilon = 1/2$ (corresponding to requiring $C = 1.5\eta_{\min}$ approximation for the k-sparse recovery problem we started with) we have $d \geq C_3 \frac{1}{\log 3(\eta_{\min}+1)} \log \frac{1}{1-\gamma} \log n - C_4 \frac{k}{\log k}$ for some numerical constants C_3, C_4, or translating to the parameter of Theorem 2.2,

$$d \geq C_3 \frac{1}{\log 3\,(\eta_{\min}+1)} \log \frac{1}{1-\gamma} \log n - C_4 \frac{r}{\log r}. \tag{F.15}$$

On the other hand, consider the $\binom{D}{r}$ canonical subspaces $\left\{\mathcal{S}_1, \ldots, \mathcal{S}_{\binom{D}{r}}\right\}$ spanned by any r subset of the canonical basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_D\}$. Let $\mathbf{0} \neq \mathbf{q} \in \mathcal{S}$, where $\mathcal{S}$ is another r-dimensional subspace and $\mathcal{S} \neq \mathcal{S}_i$ for all $i \in [\binom{D}{r}]$ and, moreover, $\mathbf{q} \notin \mathcal{S}_i$ for all i. Note that in this case $t = 1$ and $\eta = \infty$. For any projection matrix $\mathbf{P} \in \mathbb{R}^{d \times D}$, $\mathbf{Pq}$ is either $\mathbf{0}$ or spans a 1-dimensional subspace.

- To identify the original subspace unambiguously with nontrivial probability (i.e., better than random guessing in any case of ties), $\mathbf{Pq}$ cannot be zero, as for all i, $\mathbf{P}\mathcal{S}_i$ is again a subspace.
- When $\mathbf{Pq} \neq \mathbf{0}$, a necessary condition for unambiguous identifiability is $\mathbf{Pq} \notin \mathbf{P}\mathcal{S}_i$ for all i, or

$$\mathbf{Pq} \neq \mathbf{P}_{\mathcal{S}_i}\mathbf{y} \quad \forall \mathbf{y} \in \mathbb{R}^r, \forall i \in \left[\binom{D}{r}\right], \tag{F.16}$$

  where $\mathbf{P}_{\mathcal{S}_i}$ is the submatrix indexed by the canonical basis vectors associated with the subspace $\mathcal{S}_i$. Equivalently,

$$\mathbf{P}_{\mathcal{S}_i^c}\mathbf{q}_{\mathcal{S}_i^c} \neq \mathbf{P}_{\mathcal{S}_i}\mathbf{y} \quad \forall \mathbf{y} \in \mathbb{R}^r, \forall i \in \left[\binom{D}{r}\right]. \tag{F.17}$$

If $m \leq r$, then by rank argument, there exists $i \in [\binom{D}{r}]$ such that $\operatorname{span}(\mathbf{P}_{\mathcal{S}_i}) = \operatorname{span}(\mathbf{P})$, and hence $\mathbf{P}_{\mathcal{S}_i^c}\mathbf{q}_{\mathcal{S}_i^c} \in \operatorname{span}(\mathbf{P}_{\mathcal{S}_i})$, or there exists $\mathbf{y} \in \mathbb{R}^r$, such that $\mathbf{P}_{\mathcal{S}_i^c}\mathbf{q}_{\mathcal{S}_i^c} = \mathbf{P}_{\mathcal{S}_i}\mathbf{y}$, contradicting (F.17). So we must have $d \geq r$. ∎

REFERENCES

[1] A. Agarwal, S. Negahban, and M. Wainwright, *Fast global convergence of gradient methods for high-dimensional statistical recovery*, Ann. Statist., 40 (2012), pp. 2452–2482.

[2] A. Andoni, P. Indyk, R. Krauthgamer, and H. L. Nguyen, *Approximate line nearest neighbor in high dimensions*, in Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, 2009, pp. 293–301.

[3] R. Basri, T. Hassner, and L. Zelnik-Manor, *Approximate nearest subspace search with applications to pattern recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[4] R. Basri, T. Hassner, and L. Zelnik-Manor, *Approximate nearest subspace search*, IEEE Trans. Pattern Anal. Mach. Intell., 33 (2011), pp. 266–278.

[5] R. Basri and D. Jacobs, *Lambertian reflectance and linear subspaces*, IEEE Trans. Pattern Anal. Mach. Intell., 25 (2003), pp. 218–233.

[6] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.
[7] V. Blanz and T. Vetter, *Face recognition based on fitting a 3D morphable model*, IEEE Trans. Pattern Anal. Mach. Intell., 25 (2003), pp. 1063–1074.
[8] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[9] B. Brinkman and M. Charikar, *On the impossibility of dimension reduction in ℓ^1*, J. ACM, 52 (2005), pp. 766–788.
[10] E. Candès and J. Romberg, *ℓ^1-Magic: Recovery of Sparse Signals via Convex Programming*, http://users.ece.gatech.edu/justin/l1magic/downloads/l1magic.pdf (2005).
[11] E. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.
[12] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff, *The fast Cauchy transform and faster robust linear regression*, in Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, 2013, pp. 466–477.
[13] T. Cootes, G. Edwards, and C. Taylor, *Active appearance models*, IEEE Trans. Pattern Anal. Mach. Intell., 23 (2001), pp. 681–685.
[14] A. Dasgupta, P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney, *Sampling algorithms and coresets for ℓ_p regression*, in Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2008, pp. 932–941.
[15] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures Algorithms, 22 (2003), pp. 60–65.
[16] M. Datar and P. Indyk, *Locality-sensitive hashing scheme based on p-stable distributions*, in Proceedings of the Twentieth Annual ACM Symposium on Computational Geometry, 2004, pp. 253–262.
[17] K. Do Ba, P. Indyk, E. Price, and D. P. Woodruff, *Lower bounds for sparse recovery*, in Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, 2010, pp. 1190–1197.
[18] D. Donoho and C. Grimes, *Image manifolds which are isometric to Euclidean space*, J. Math. Imaging Vision, 23 (2005), pp. 5–24.
[19] D. Donoho and Y. Tsaig, *Fast solution of ℓ^1-norm minimization problems when the solution may be sparse*, IEEE Trans. Inform. Theory, 54 (2008), pp. 4789–4812.
[20] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist., 32 (2004), pp. 407–499.
[21] L. Elazary and L. Itti, *A Bayesian model for efficient visual search and recognition*, Vis. Res., 50 (2010), pp. 1338–1352.
[22] A. Georghiades, P. Belhumeur, and D. Kriegman, *From few to many: Illumination cone models for face recognition under variable lighting and pose*, IEEE Trans. Pattern Anal. Mach. Intell., 23 (2001), pp. 643–660.
[23] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, *The Amsterdam library of object images*, Int. J. Comput. Vis., 61 (2005), pp. 103–112.
[24] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison–Wesley, Reading, MA, 1968.
[25] J. Ho, M. Yang, J. Lim, K. C. Lee, and D. Kriegman, *Clustering appearances of objects under varying illumination conditions*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 11–18.
[26] P. Jain, S. Vijayanarasimhan, and K. Grauman, *Hashing hyperplane queries to near points with applications to large-scale active learning*, IEEE Trans. Pattern Anal. Mach. Intell., 36 (2014), pp. 276–288.
[27] W. B Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, in Conference in Modern Analysis and Probability, Contemp. Math. 26, AMS, Providence, RI, 1984, pp. 189–206.
[28] M. Ledoux, *The Concentration of Measure Phenomenon*, AMS, Providence, RI, 2001.
[29] P. Li, T. Hastie, and K Church, *Nonlinear estimators and tail bounds for dimension reduction in ℓ_1 using Cauchy random projections*, J. Mach. Learn. Res., 8 (2007), pp. 2497–2532.
[30] A. Magen and A. Zouzias, *Near optimal dimensionality reductions that preserve volumes*, in Approxi-

mation, Randomization and Combinatorial Optimization, Springer, Berlin, 2008, pp. 523–534.
[31] J. Mattingley and S. Boyd, *CVXGEN: A code generator for embedded convex optimization*, Optim. Eng., 13 (2012), pp. 1–27.
[32] H. Murase and S. Nayar, *Visual learning and recognition of* 3*D objects from appearance*, Int. J. Comput. Vis., 14 (1995), pp. 5–24.
[33] E. Price, *Efficient sketches for the set query problem*, in Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, 2011, pp. 41–56.
[34] S. Roweis and L. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science, 290 (2000), pp. 2323–2326.
[35] P. Simard, Y. Le Cun, J. Denker, and B. Victorri, *Transformation invariance in pattern recognition - tangent distance and tangent propagation*, in Neural Networks: Tricks of the Trade, Springer, Berlin, 1998, pp. 239–274.
[36] C. Sohler and D. P. Woodruff, *Subspace embeddings for the* ℓ_1*-norm with applications*, in Proceedings of the 43rd Annual ACM Symposium on Theory of Computing, 2011, pp. 755–764.
[37] M. Turk and A. Pentland, *Eigenfaces for recognition*, J. Cognit. Neurosci., 3 (1991), pp. 71–86.
[38] V. V. Uchaikin and V. M. Zolotarev, *Chance and Stability: Stable Distributions and Their Applications*, Vol. 3, VSP, Utrecht, The Netherlands, 1999.
[39] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, *Towards a practical automatic face recognition system: Robust alignment and illumination by sparse representation*, IEEE Trans. Pattern Anal. Mach. Intell., 34 (2012), pp. 372–386.
[40] R. Williams, *A new algorithm for optimal* 2*-constraint satisfaction and its implications*, Theoret. Comput. Sci., 348 (2005), pp. 357–365.
[41] J. Wright and Y. Ma, *Dense error correction via* ℓ^1*-minimization*, IEEE Trans. Inform. Theory, 56 (2010), pp. 3540–3560.
[42] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, *Robust face recognition via sparse representation*, IEEE Trans. Pattern Anal. Mach. Intell., 31 (2009), pp. 210–227.
[43] A. Yang, A. Ganesh, Y. Ma, and S. Sastry, *Fast* ℓ^1*-minimization algorithms and an application in robust face recognition: A review*, in Proceedings of the International Conference on Image Processing, 2010.
[44] A. Y. Yang, Z. Zhou, A. Balasubramanian, S. Sastry, and Y. Ma, *Fast* ℓ^1*-minimization algorithms for robust face recognition*, IEEE Trans. Image Process., 22 (2013), pp. 3234–3246.
[45] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for* ℓ_1*-minimization with applications to compressed sensing*, SIAM J. Imaging Sci., 1 (2008), pp. 143–168.
[46] Y. Zhang, C. Mu, H. Kuo, and J. Wright, *Towards guaranteed illumination models for nonconvex objects*, in Proceedings of the International Conference on Computer Vision, 2013.
[47] V.M. Zolotarev, *One-Dimensional Stable Distributions*, Transl. Math. Monogr. 65, AMS, Providence, RI, 1986.