Efficient Point-to-Subspace Query in ℓ^1 with Application to Robust Face Recognition

Ju Sun, Yuqian Zhang, and John Wright

Department of Electrical Engineering, Columbia University, New York, USA
{jusun, yuqianzhang, johnwright}@ee.columbia.edu

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. We show in theory this problem can be solved with a simple two-stage algorithm: (1) random Cauchy projection of query and subspaces into low-dimensional spaces followed by efficient distance evaluation (ℓ^1 regression); (2) getting back to the high-dimensional space with very few candidates and performing exhaustive search. We present preliminary experiments on robust face recognition to corroborate our theory.

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

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 from low-level tasks such as signal acquistion 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. To 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 $\mathcal{O}_i$, the recognition problem becomes one of finding the nearest low-dimensional structure: $\min_i d(\mathbf{q}, \mathcal{M}_i)$, where $\mathbf{q} \in \mathbb{R}^D$ is the test image, and $d(\cdot, \cdot)$ is some metric.

This paradigm is broad enough to encompass very classical work in face recognition [1] and object instance recognition [2], as well as more recent developments [3–5]. In situations in which sufficient training data is available to accurately fit the $\mathcal{M}_i$, it can achieve high recognition rates [6]. 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 nusiance variations that may be encountered. For example, variations in illumination may be well-captured using low-dimensional *linear* models [7, 8], whereas variations in pose or alignment are highly nonlinear [9].

Second, *how should we measure the distance between* $\mathbf{q}$ *and* $\mathcal{M}_i$*?* Typically, one adopts a metric $d(\cdot,\cdot)$ on $\mathbb{R}^D$, and then sets $d(\mathbf{q},\mathcal{M}_i) = \min_{\mathbf{v}\in\mathcal{M}_i} d(\mathbf{q},\mathbf{v})$. Here, again, the appropriate choice metric d depends on our prior knowledge. For example, if the observation $\mathbf{q}$ is known to be perturbed by i.i.d. Gaussian noise, minimizing the ℓ^2 norm $d(\mathbf{q},\mathbf{v}) = \|\mathbf{q}-\mathbf{v}\|_2$ 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, specularities, the ℓ^1 norm is a more robust alternative [5].

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

$$\min_i \min_{\mathbf{v}\in\mathcal{M}_i} d(\mathbf{q},\mathbf{v}) \tag{1}$$

using computational resources that depend as gracefully as possible on the ambient dimension D (typically 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 [7, 8] (say, in near-frontal face recognition), and also form a basic building block for modeling and computing with more general, nonlinear sets [10, 11].

Our methodology pertains to distances $d(\mathbf{q},\mathbf{v})$ induced by the ℓ^p norm $\|\mathbf{q}-\mathbf{v}\|_p$, with $p \in (0,2]$. 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 occlusion, cast shadows, or specularities [5]. 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. Given n linear subspaces $\mathcal{S}_1,\ldots,\mathcal{S}_n$ of $\mathbb{R}^D$ of dimension r and a query point $\mathbf{q} \in \mathbb{R}^D$, determine the nearest $\mathcal{S}_i$ to $\mathbf{q}$ in ℓ^1 norm.

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

$$\min_{\mathbf{v}\in\mathcal{S}_i} \|\mathbf{q}-\mathbf{v}\|_1, \tag{2}$$

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 (2). For example, for interior point methods [12], we have $T_{\ell^1}(D,r) = O(D^{3.5})$.

There exist more scalable first-order methods [13–16], 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 [17] or the problem is otherwise well-structured [18]. Even in the best case, however, the aforementioned algorithms have complexity $\Omega(nD)$.[1] When both terms are large, this dependence is prohibitive: *Although Problem 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, 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 iid standard Cauchy random variables. That is to say, instead of solving (2), we solve

$$\min_{\mathbf{v} \in \mathcal{S}_i} \|\mathbf{P}\mathbf{q} - \mathbf{P}\mathbf{v}\|_1. \tag{3}$$

We prove that if the embedded dimension d is sufficiently large – say $d = \mathrm{poly}(r \log n)$, then with constant probability the model $\mathcal{S}_i$ obtained from (3) is the same as the one obtained from the original optimization (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 (e.g., [19]). These methods are numerically reliable, and can yield a speedup of several orders of magnitude over the naive approach (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 $\mathbb{R}^d$, these repeated trials are affordable.

The end result is a simple, practical algorithm that guarantees to maintain the good properties of ℓ^1 regression, with substantially improved computational complexity. We demonstrate this on model problems in subspace-based face and digit recognition (in supplementary material). 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 small databases and hard time constraints.

[1] 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.

Relationship to existing work. Problem 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 [20]. 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 [21].

Recently two groups have proposed approaches to tackling larger r. Basri et. al. [22] 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 et. al. give several random hash functions for the case when the $\mathcal{S}_i$ are hyperplanes [23]. 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 $c < 1$ [24].

The above algorithms exploit special properties of the ℓ^2 version of Problem 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.[2] 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 is the unique 1-stable distribution, a property which has been widely exploited in previous algorithmic work [20, 26, 27].

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. 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 non-negligble distortion [28]. For a single subspace, embedding results exist – most notably due to Soehler and Woodruff [27], but the distortion incurred is so large as to render them inapplicable to Problem 1. Nevertheless, several elegant technical ideas in the proof of [27] turn out to be useful for analyzing Problem 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 [29, 30]. These results give conditions under which it is possible to recover a vector $\mathbf{x}$ from grossly corrupted observations $\mathbf{q} = \mathbf{v} + \mathbf{e}$, with $\mathbf{v} \in \mathcal{S}$ and the sparse error $\mathbf{e}$ unknown. These results are quite strong: they imply exact recovery, even if

[2] Although it could be possible if we are willing to accept time and space complexity exponential in r or D, ala [25].

the error $\mathbf{e}$ has nonnegligible fractions of nonzero entries, of arbitrary size. For example, [29] proves that under technical conditions, ℓ^1 minimization

$$\min \|\mathbf{e}\|_1 \quad \text{s.t.} \quad \mathbf{q} - \mathbf{e} \in \mathcal{S} \tag{4}$$

exactly recovers $\mathbf{e}$. [30] presents similar theory for the case when $\mathcal{S}$ is union of subspaces solved by a variant of optimization in (4).

On the other hand, exact recovery may be stronger than what is needed for recognition. For recognition, as formulated in this work, we only need to 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 [5] 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 $\mathcal{S}_1 \dots \mathcal{S}_n$.

2 Our Algorithm and Main Results

The core of our algorithm is summarized as follows.

Input: n subspaces $\mathcal{S}_1, \cdots, \mathcal{S}_n$ 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 iid Cauchy RV's ($d \ll D$) and Compute the projections $\mathbf{P}\mathcal{S}_1$, $\cdots$, $\mathbf{P}\mathcal{S}_n$
Test: Compute the projection $\mathbf{Pq}$, and compute its ℓ^1 distance to each of $\mathbf{P}\mathcal{S}_i$

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 $\mathcal{S}_\star$:

Theorem 1. *Suppose we are given n linear subspaces $\{\mathcal{S}_1, \cdots, \mathcal{S}_n\}$ of dimension r in $\mathbb{R}^D$ and any query point $\mathbf{q}$, and that the ℓ^1 distances of $\mathbf{q}$ to each of $\{\mathcal{S}_1, \cdots, \mathcal{S}_n\}$ are $\xi_{1'} \leq \cdots \leq \xi_{n'}$ when arranged in ascending order, with $\xi_{2'}/\xi_{1'} \geq \eta > 1$. For any fixed $\alpha < 1 - 1/\eta$, there exists $d \sim O\left[(r \log n)^{1/\alpha}\right]$ (assuming $n > r$), if $\mathbf{P} \in \mathbb{R}^{d \times D}$ is iid Cauchy, we have*

$$\arg\min_{i \in [n]} d_{\ell^1}(\mathbf{Pq}, \mathbf{P}\mathcal{S}_i) = \arg\min_{i \in [n]} d_{\ell^1}(\mathbf{q}, \mathcal{S}_i) \tag{5}$$

with (nonzero) constant probability.

The condition in Theorem 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 $\eta = 1/2$ 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 Johnson-Lindenstrauss lemma for dimensionality reduction in ℓ^2, even though JL-syle embeddings are impossible for ℓ^1.

Before stating our overall algorithm, we suggest two additional practical implications of Theorem 1. First, Theorem 1 only guarantees success 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 $d < d' \ll D$; we save this for future work.

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, we may instead set d according to the gap between $\xi_{1'}$ and $\xi_{k'}$, for some $k' > 2$. With this choice, Theorem 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 naive approach.

3 A Sketch of the Analysis

In this section, we sketch the analysis leading to Theorem 1. The basic rationale for using Cauchy projection is that the Cauchy distribution is the unique *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 iid Cauchy entries, then the vector $\mathbf{Pv} \equiv_d \|\mathbf{v}\|_1 \times \mathbf{z}$, where $\mathbf{z}$ is again an iid Cauchy vector, and $\equiv_d$ denotes equality in distribution. 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 iid *half-Cauchy*, with probability density function

$$f_{\mathcal{HC}}(x) = \frac{2}{\pi} \frac{1}{1+x^2} \quad \text{if } x \geq 0, \tag{6}$$

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 the unambiguously 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 non-negligible 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)

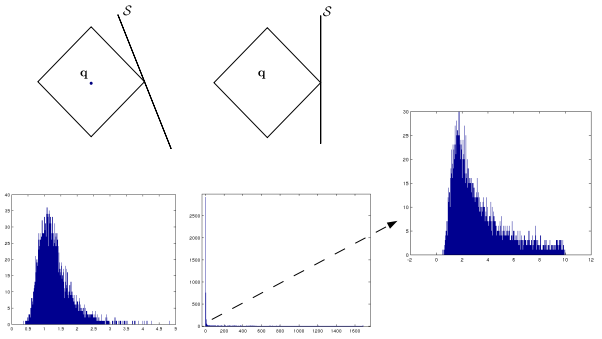

Fig. 1. Statistics of ℓ^1 distance ratios (after vs. before) by random projections over 10000 trials. The subspaces are randomly-oriented (1^{st} column) and axis-aligned (2^{nd} column), respectively. Here $r = 10$, $D = 10000$, $d = 35$, and $d_{\ell^1}(q, \mathcal{S}) = 1$.

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 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 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!). 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. Sec. 3.1 elaborates on property (i), while Sec. 3.2 describes the arguments needed to establish property (ii). Theorem 1 follows directly from the results in Secs. 3.1 and 3.2. This argument, as well as proofs of several routine or technical lemmas are deferred to the supplementary material.

3.1 Bounded expansion for the good subspace

Let $\mathbf{v}_\star \in \mathcal{S}_\star$ be a closest point to $\mathbf{q}$ in 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 always exists. After projection, $\mathbf{P}\mathbf{v}_\star$ might no longer be the closest point to $\mathbf{P}\mathbf{q}$. However, the distance $\|\mathbf{P}\mathbf{q}-\mathbf{P}\mathbf{v}_\star\|_1$ *does* upper bound the distance from $\mathbf{P}\mathbf{q}$ to $\mathbf{P}\mathcal{S}_\star$:

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

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:

Lemma 1. *There exists numerical constant $c \in (0,1)$ with the following property. If $\mathbf{w} \in \mathbb{R}^D$ be any fixed vector, and suppose that $\mathbf{P} \in \mathbb{R}^{d\times D}$ is a matrix with iid standard Cauchy entries. Then for any $\rho > 1$,*

$$\mathbb{P}\left[\|\mathbf{P}\mathbf{w}\|_1 > \rho\frac{2}{\pi}d\log d\,\|\mathbf{w}\|_1\right] \;<\; c+\frac{1-c}{\rho} \;<\; 1. \tag{7}$$

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:

$$\forall\, \mathbf{w} \in \mathcal{S}_i \oplus \mathrm{span}(\mathbf{q}), \quad \|\mathbf{P}\mathbf{w}\|_1 \;\ge\; \gamma\|\mathbf{w}\|_1. \tag{8}$$

Above, $\oplus$ denotes the direct sum of subspaces, so $\tilde{\mathcal{S}}_i = \mathcal{S}_i \oplus \mathrm{span}(\mathbf{q})$ 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 (8) holds, we have

$$\begin{aligned} d_{\ell^1}(\mathbf{P}\mathbf{q}, \mathbf{P}\mathcal{S}_i) \;&=\; \min_{\mathbf{v}\in\mathcal{S}_i} \|\mathbf{P}\mathbf{q}-\mathbf{P}\mathbf{v}\|_1 \quad\ge\quad \min_{\mathbf{v}\in\mathcal{S}_i} \|\mathbf{P}(\mathbf{q}-\mathbf{v})\|_1 \\ &\ge\; \min_{\mathbf{v}\in\mathcal{S}_i} \gamma\|\mathbf{q}-\mathbf{v}\|_1 \quad=\quad \gamma\, d_{\ell^1}(\mathbf{q}, \mathcal{S}_i), \end{aligned} \tag{9}$$

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

To show (8), we use a discretization argument. Let $\mathit{\Gamma}$ denote the intersection of the unit ℓ^1 "sphere" with the expanded subspace $\tilde{\mathcal{S}}_i$:

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

Recall that for any set $\mathit{\Gamma}$, an *ε-net* is a subset N_i such that for every $\mathbf{w} \in \mathit{\Gamma}$, $\|\mathbf{w}-\mathbf{w}'\|_1 \le \epsilon$ for some $\mathbf{w}' \in N$. Standard arguments (see [31]) show that for any $\epsilon > 0$, there exists an ϵ net N_i for $\mathit{\Gamma}$ of size at most $(3/\epsilon)^{d+1}$.

Consider the following two events:
- (ii.a) $\min_{\mathbf{w}'\in N} \|\mathbf{P}\mathbf{w}'\|_1 \ge \beta$, and
- (ii.b) For all $\mathbf{w} \in \tilde{\mathcal{S}}_i$, $\|\mathbf{P}\mathbf{w}\|_1 \;\le\; L\|\mathbf{w}\|_1$.

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

$$\|\mathbf{P}\mathbf{w}\|_1 \ge \|\mathbf{P}\mathbf{w}' + \mathbf{P}(\mathbf{w}-\mathbf{w}')\|_1 \;\ge\; \|\mathbf{P}\mathbf{w}'\|_1 - \|\mathbf{P}(\mathbf{w}-\mathbf{w}')\|_1 \;\ge\; \beta - L\epsilon. \tag{10}$$

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

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

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 2 (Concentration in Lower Tail). *Let $\mathbf{P} \in \mathbb{R}^{d\times D}$ be an iid Cauchy matrix. Then for any fixed vector $\mathbf{w} \in \mathbb{R}^D$ and $\alpha, \delta \in (0,1)$,*

$$\mathbb{P}\left[\|\mathbf{P}\mathbf{w}\|_1 < (1-\alpha)\,(1-\delta)\,\frac{2}{\pi} d\log d\,\|\mathbf{w}\|_1\right] \;<\; d^{1-\alpha}\exp\left(-\frac{\delta^2}{2\pi}d^{\alpha}\right). \tag{11}$$

This estimate gives the optimal power, d^α, in the exponent. The proof is straightforward, and is deferred to the supplementary material.

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

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

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

$$\mathbb{P}\left[\mathcal{E}_{\mathrm{net},i}\right] \;<\; |N_i|\, d^{1-\alpha}\exp\left(-\tfrac{\delta^2}{2\pi}d^{\alpha}\right). \tag{12}$$

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 [27]. 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 the supplementary material, and instead simply state the resulting bound:

Lemma 3. *Let $\mathbf{P} \in \mathbb{R}^{d\times D}$ be an iid Cauchy matrix, and $\mathcal{S}$ 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*

$$\mathbb{P}\left[L > t\,(r+1)\right] \le \frac{2d(r+1)}{\pi B} + \frac{2d(r+1)}{\pi t}\log\sqrt{1+B^2}. \tag{13}$$

The proof of Theorem 1 follows from Lemmas 1-3 above, by choosing appropriate values of the parameters B, t, δ and ϵ. We give the detailed calculation in the supplementary material.

4 Experiments

We present two experiments[3] to corroborate our theoretical result and demonstrate its particular relevance to subspace/manifold-based instance recognition.

[3] The second one on digit recognition is presented in the supplementary material.

4.1 Note on Implementation

Projection Matrices and Subspaces. Our main theorem is for any fixed set of subspaces and any fixed query point. Of course, if we fix $\mathbf{P}$ and consider many different $\mathbf{q}$, the success or failure will be dependent random variables. This suggests sampling a new matrix $\mathbf{P}$ for each test image, which would then require that we re-project each of the subspaces $\mathcal{S}_i$. In practice, it is more efficient to maintain a pool of k Cauchy projection matrices[4] $\mathbf{P}_j$ and store $\mathbf{P}_j\mathcal{S}_i$ for each i and j. During testing, we randomly sample a combination of N_{rep} (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 impressive performance. We fix $k = 20$ and normally set $N_{rep} = 3$ throughout.

Solvers for ℓ^1 Regression. We perform high-dimensional NS search in ℓ^1 (HDS) as baseline. Due to the large scale, we employ an Augmented Lagrange Method (ALM) numerical solver for the regression. All the other instances of ℓ^1 regression are in low dimensions and can be handled by interior point method (IPM) solvers. We will report typical running times, with the caveat that direct comparison may not be fair: the ALM solver is built for moderate accuracy with high scalability and subject to careful tuning of optimization parameters, while IPM solvers are meant for high accuracy. Despite this, our algorithm is often significantly faster.

4.2 Robust Face Recognition on Extended Yale B

Face images of one person taken with fixed pose and varying illumination are known to lie very close to a nine-dimensional linear subspace [8]. Because physical phenomena such as occlusions and specularities on faces may violate the linear model, we formulate the recognition problem as one of finding the closest subspace to $\mathbf{q}$ in ℓ^1 norm [5][5].

The Extended Yale B face dataset [7] (EYB, cropped version) 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 took half of the images for training (1205 in total) and the others for testing (1209 in total). 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

[4] The standard Cauchy projection matrix $\mathbf{P}$ generated as $\mathbf{A}./\mathbf{B}$, where both $\mathbf{A}$ and $\mathbf{B}$ are iid standard normal and "./" denotes elementwise matrix division.

[5] In other words, we formulate the problem as ℓ^1 nearest subspace (ℓ^1 NS) search. This is different from the idea of sparse representation in SRC [5] for face recognition. Since our focus here is not to propose a new or optimal face recognition algorithm (although 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 typical robust face recognition benchmarks.

based on the light source direction (*wrt.* the camera axis): images taken with either azimuth angle greater than 90° or elevation angle greater than 60° would be classified as extremely illuminated.

Recognition with Original Images. Fig. 2 presents the evolution of recognition rate on **Subset M** as the projection dimension (d) grows *with only one repetition of the projection* ($N_{rep} = 1$). Our experiment shows the HDS achieves

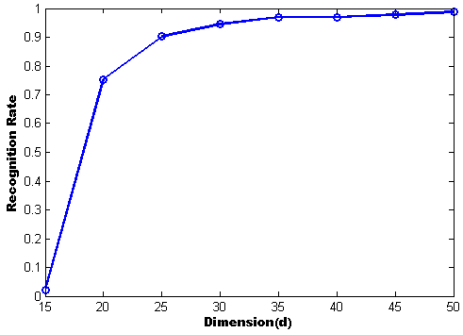

Fig. 2. Recognition rate versus projection dimension (d) *with one repetition* on **Subset M** face images of EYB. The recognition rate stays stable above 95% with $d \geq 25$. The high-dimensional NS in ℓ^1 achieves perfect (100%) recognition. Note the ambient dimension in this case is $D = 168 \times 192 = 32256$.

perfect recognition (100%) on this subset, implying recognition in this subset corresponds perfectly to NS search in ℓ^1. So Fig. 2 actually represents the evolution of "average" success probability for *one repetition* over the subset. Suppose the distance gap is significant such that $1/\alpha \to 1$, 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 3 repetitions and $d = 25$, the overall recognition rate is 99.56% (4 errors out of 909), nearly perfect. Fig. 3 presents the failing cases. They either contain significant artifacts

Fig. 3. Failing cases of our method on **Subset M** of EYB.

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 nearest subspaces is much less significant (one example shown in Fig. 4). 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, the HDS achieves 94.7% accuracy while our method achieves only 79.3% when $d = 25$ and $N_{back} = 5$ (N_{back} is the number of

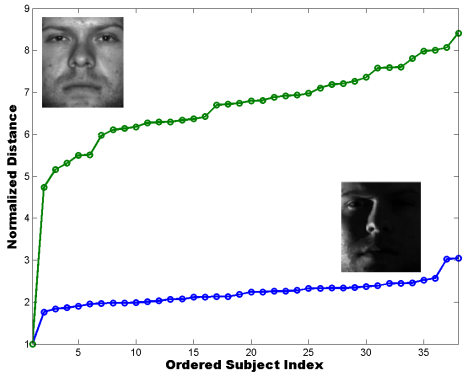

Fig. 4. Samples of moderately/extremely illuminated face images and their ℓ^1 distances to other subject subspaces. The subjects have been ordered in ascending order of ℓ^1 distance from the sample and the distances are normalized such that the first distance is 1. Note that for the moderately illuminated sample, a distance gap of about 4.8 is observed while this is only about 1.8 for the extremely illuminated sample.

back-research 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}.

	HDS	$d = 25$	$d = 50$	$d = 70$
$r = 15, N_{back} = 5$	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 ℓ^1 NS approach for recognition and particularly the capability of our method to preserve such property of ℓ^1, we corrupted each original test image with (1) randomly-distributed sparse corruptions, and (2) structured occlusions. For the first setting, we replaced, respectively, 5%, 10%, 15%, and 20% of randomly chosen pixels with iid uniform noise in $[0, 255]$[6]. For the second, the *lena* image of fixed size (i.e. depending on the desired percentage of occlusion) was randomly placed on each test image. Fig. 5 shows some typical samples of both cases, and also the effect of corruptions on distance gaps - corruptions significantly weaken the gaps. Therefore we set d to 50 and 70 in this experiment for comparison. Table 2 summarizes the recognition performances for each setting. Our method exhibits comparable level of performance with the HDS for corruptions less than 10% and observable performance lag beyond. This is a reasonable price to pay as we insist on working in low dimensions for efficiency.

Running Time. In our Matlab implementation, the typical time required for solving one instance of HDS is 8.3s (with ALM solver), and that for our method

[6] In other words, any valid pixel value for 8-bit gray-scaled image. Note also that our training is still half of all the samples as in last part, in contrast to the setting in [5], where only those moderately illuminated are considered.

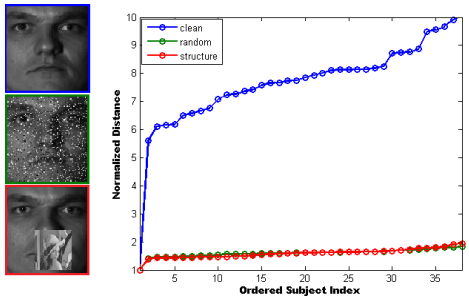

Fig. 5. Left: Sample of original images and the corrupted versions. In both corrupted images 20% of the pixels are contaminated. Right: Comparison of the ordered original ℓ^1 distances to other subspaces and that of after introducing the artificial corruptions. This distance gap is significantly suppressed due to the corruptions.

Table 2. Recognition Rate under Corruptions for all Test Samples on EYB. ($r = 15$)

Occlusion	Occluded Pixels	HDS	$d = 50$	$d = 70$
Random	5%	98.8%	96.2%	97.2%
	10%	98.6%	93.7%	95.2%
	15%	99.2%	89.2%	91.9%
	20%	99.2%	85.4%	87.8%
Structured	5%	98.7%	95.7%	96.7%
	10%	97.8%	91.3%	94.7%
	15%	95.9%	87.3%	91.6%
	20%	93.5%	82.7%	84.6%

is only about 1.2s (ℓ^1-magic interior point solver) which is mostly consumed by the back search in high dimensions. There is no observable difference in timing with or without the corruptions.

Acknowledgments. JS was supported by the Wei Family Private Foundation Fellowship.

References

1. Turk, M., Pentland, A.: Eigenfaces for recognition. In: CVPR. (1991)
2. Murase, H., Nayar, S.: Visual learning and recognition of 3D objects from appearance. IJCV **14**(1) (1995) 5–24
3. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. PAMI **23**(6) (2001) 681–685
4. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. PAMI **25**(9) (2003) 1063–1074
5. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. PAMI **31**(2) (2009) 210–227
6. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Towards a practical automatic face recognition system: Robust alignment and illumination by sparse representation. IEEE Trans. PAMI **34**(2) (2012) 372–386
7. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. PAMI **23**(6) (2001) 643–660

8. Basri, R., Jacobs, D.: Lambertian reflectance and linear subspaces. IEEE Trans. PAMI **25**(2) (2003) 218–233
9. Donoho, D., Grimes, C.: Image manifolds which are isometric to Euclidean space. J. of Math. Imag. and Vis. **23**(1) (2005) 5–24
10. Simard, P.Y., Cun, Y.A.L., Denker, J.S., Victorri, B.: Transformation invariance in pattern recognition - tangent distance and tangent propagation. In: Neural Networks: Tricks of the Trade, Springer (1998) 239–274
11. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290** (2000) 2323–2326
12. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
13. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics **32** (2004) 407–499
14. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. on Imag. Sci. **2**(1) (2009) 183–202
15. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for l1minimization with applications to compressed sensing. SIAM J. Imag. Sci **1**(1) 143–168
16. Yang, A., Ganesh, A., Ma, Y., Sastry, S.: Fast ℓ^1-minimization algorithms and an application in robust face recognition: A review. In: ICIP. (2010)
17. Donoho, D., Tsaig, Y.: Fast solution of ℓ^1-norm minimization problems when the solution may be sparse. IEEE Trans. IT **54**(11) (2008) 4789–4812
18. Agarwal, A., Negahban, S., Wainwright, M.: Fast global convergence of gradient methods for high-dimensional statistical recovery. In: NIPS. (2011)
19. Mattingley, J., Boyd, S.: CVXGEN: A code generator for embedded convex optimization. Optimization and Engineering **13**(1) (2012) 1–27
20. Datar, M., Indyk, P.: Locality-sensitive hashing scheme based on p-stable distributions. In: SCG, ACM Press (2004) 253–262
21. Andoni, A., Indyk, P., Krauthgamer, R., Nguyen, H.: Approximate line nearest neighbor in high dimensions. In: SODA. (2009)
22. Basri, R., Hassner, T., Zelnik-Manor, L.: Approximate nearest subspace search. IEEE Trans. PAMI **33**(2) (2011) 266–278
23. Jain, P., Vijayanarasimhan, S., Grauman, K.: Hashing hyperplane queries to near points with applications to large-scale active learning. In: NIPS. (2010)
24. Williams, R.: A new algorithm for optimal 2-constraint satisfaction and its implications. Theo. Comp. Sci. **348** (2005) 357–365
25. Magen, A., Zouzias, A.: Near optimal dimensionality reductions that preserve volumes. In: APPROX-RANDOM. (2008) 523–534
26. Li, P., Hastie, T., Church, K.: Nonlinear estimators and tail bounds for dimension reduction in ℓ^1 using cauchy random projections. JMLR **8** (2007) 2497–2532
27. Sohler, C., Woodruff, D.: Subspace embeddings for the ℓ_1-norm with applications. In: STOC. (2011)
28. Brinkman, B., Charikar, M.: On the impossibility of dimension reduction in ℓ^1. J. ACM **52** (2005) 766–788
29. Candés, E., Tao, T.: Decoding by linear programming. IEEE Trans. IT **51**(12) (2005) 4203–4215
30. Wright, J., Ma, Y.: Dense error correction via ℓ^1-minimization. IEEE Trans. IT **56**(7) (2010) 3540–3560
31. Ledoux, M.: The Concentration of Measure Phenomenon. AMS (2001)