Sketching for M-Estimators: A Unified Approach to Robust Regression

Kenneth L. Clarkson* David P. Woodruff†

Abstract

We give algorithms for the M-estimators \( \min_x \|Ax - b\|_G \), where \( A \in \mathbb{R}^{n \times d} \) and \( b \in \mathbb{R}^n \), and \( \|y\|_G \) for \( y \in \mathbb{R}^n \) is specified by a cost function \( G : \mathbb{R}^n \mapsto \mathbb{R}_+^n \), with \( \|y\|_G \equiv \sum_i G(y_i) \). The M-estimators generalize \( \ell_p \) regression, for which \( G(x) = |x|^p \). We first show that the Huber measure can be computed up to relative error \( \epsilon \) in \( O(\text{nnz}(A) \log n + \text{poly}(d \log n) / \epsilon) \) time, where \( \text{nnz}(A) \) denotes the number of non-zero entries of the matrix \( A \). Huber is arguably the most widely used \( \ell_p \) estimator, enjoying the robustness properties of \( \ell_1 \) as well as the smoothness properties of \( \ell_2 \).

We next develop algorithms for general M-estimators. We analyze the M-sketch, which is a variation of a sketch introduced by Verbin and Zhang in the context of estimating the earthmover distance. We show that the M-sketch can be used much more generally for sketching any M-estimator provided \( G \) has growth that is at least linear and at most quadratic. Using the M-sketch we solve the M-estimation problem in \( O(\text{nnz}(A) + \text{poly}(d \log n)) \) time for any such \( G \) that is convex, making a single pass over the matrix and finding a solution whose residual error is within a constant factor of optimal, with high probability.

1 Introduction

In recent years there have been significant advances in randomized techniques for solving numerical linear algebra problems, including the solution of diagonally dominant systems [28, 29, 39], low-rank approximation[2, 9, 15, 12, 13, 34, 36, 38], overconstrained regression [9, 21, 34, 36, 38], and computation of leverage scores [9, 17, 34, 36]. There are many other references; please see for example the survey by Mahoney [30]. Much of this work involves the tool of sketching, which in generality is a descendnet of random projection methods as described by Johnson and Lindenstrauss[1, 4, 3, 11, 26, 27], and also of sampling methods [10, 14, 15, 16, 18, 19, 20]. Given a problem involving \( A \in \mathbb{R}^{n \times d} \), a sketching matrix \( S \in \mathbb{R}^{t \times n} \) with \( t \ll n \) is used to reduce to a similar problem involving the smaller matrix \( SA \), with the key property that with high likelihood with respect to the randomized choice of \( S \), a solution for \( SA \) is a good solution for \( A \). More generally, data derived using \( SA \) is used to efficiently solve the problem for \( A \). In cases where no further processing of \( A \) is needed, a streaming algorithm often results, since a single pass over \( A \) suffices to compute \( SA \).

An important property of many of these sketching constructions is that \( S \) is a subspace embedding, meaning that for all \( x \in \mathbb{R}^d \), \( \|SAx\| \approx \|Ax\| \). (Here the vector norm is generally \( \ell_p \) for some \( p \).) For the regression problem of minimizing \( \|Ax - b\| \) with respect to \( x \in \mathbb{R}^d \), for inputs \( A \in \mathbb{R}^{n \times d} \) and \( b \in \mathbb{R}^n \), a minor extension of the embedding condition implies \( S \) preserves the norm of the residual vector \( Ax - b \), that is \( \|S(Ax - b)\| \approx \|Ax - b\| \), so that a vector \( x \) that makes \( \|S(Ax - b)\| \) small will also make \( \|Ax - b\| \) small.

A significant bottleneck for these methods is the computation of \( SA \), taking \( \Theta(nd) \) time with straightforward matrix multiplication. There has been work showing that fast transform methods can be incorporated into the construction of \( S \) and its application to \( A \), leading to a sketching time of \( O(nd \log n) [3, 4, 7, 38] \).

Recently it was shown that there are useful sketching matrices \( S \) such that \( SA \) can be computed in time linear in the number \( \text{nnz}(A) \) of non-zeros of \( A \) [6, 9, 34, 36]. With such sketching matrices, various problems can be solved with a running time whose leading term is \( O(\text{nnz}(A)) \) or \( O(\text{nnz}(A) \log n) \). This prominently includes regression problems on “tall and thin” matrices with \( n \gg d \), both in the least-squares (\( \ell_2 \)) and robust (\( \ell_1 \)) cases. There are also recent recursive sampling-based algorithms for \( \ell_p \) regression [35], as well as sketching-based algorithms for \( p \in [1, 2] \) [34] and \( p > 2 \) [41], though the latter requires sketches whose size grows polynomially with \( n \). Similar \( O(\text{nnz}(A)) \) time results were obtained for quantile regression [42], by relating it to \( \ell_1 \) regression. A natural question raised by these works is which families of penalty functions can be computed in \( O(\text{nnz}(A)) \) or \( O(\text{nnz}(A) \log n) \) time.

M-estimators. Here we further extend the “nnz” regime to general statistical M-estimators, specified by a measure function \( G : \mathbb{R}^n \mapsto \mathbb{R}_+^n \), where \( G(x) = G(-x) \), \( G(0) = 0 \), and \( G \) is non-decreasing in \( |x| \). The result is a new “norm” \( \|y\|_G \equiv \sum_{i \in [n]} G(y_i) \). (In general these functions \( \|\|_G \) are not true norms, but we will sometimes refer to them as norms anyway.) An M-estimator is a solution to \( \min_x \|Ax - b\|_G \). For appropriate \( G \), M-estimators can combine the insensitivity to outliers of \( \ell_1 \) regression with the low variance of \( \ell_2 \) regression.

*IBM Research - Almaden
†IBM Research - Almaden
The Huber norm. The Huber norm [24], for example, is specified by a parameter $\tau > 0$, and its measure function $H$ is given by

$$H(a) = \begin{cases} a^2/2\tau & \text{if } |a| \leq \tau \\ |a| - \tau/2 & \text{otherwise,} \end{cases}$$

combining an $\ell_2$-like measure for small $x$ with an $\ell_1$-like measure for large $x$.

The Huber norm is of particular interest, because it is popular and “recommended for almost all situations” [43], because it is “most robust” in a certain sense [24], and because it has the useful computational and statistical properties implied by the convexity and smoothness of its defining function, as mentioned above. The smoothness makes it differentiable at all points, which can lead to computational savings over $\ell_1$, while enjoying the same robustness properties with respect to outliers. Moreover, while some measures, such as $\ell_1$, treat small residuals “as seriously” as large residuals, it is often more appropriate to have robust treatment of large residuals and Gaussian treatment of small residuals [22].

We give in §2 a sampling scheme for the Huber norm based on a combination of Huber’s $\ell_1$ and $\ell_2$ properties. We obtain an algorithm yielding an $\epsilon$-approximation with respect to the Huber norm of the residual; as stated in Theorem 2.1, the algorithm needs $O(\text{nnz}(A)\log n) + \text{poly}(d/\epsilon)$ time (see, e.g., [31] for convex programming algorithms for solving Huber in $\text{poly}(d/\epsilon)$ time when the dimension is poly($d/\epsilon$)).

$M$-sketches for $M$-estimators. We also show that the sketching construction of Verbin and Zhang [40], which they applied to the earthmover distance, can also be applied to sketching for general $M$-estimators. This construction, which we call the $M$-sketch\footnote{Verbin and Zhang call the construction a Rademacher sketch; with apologies, we prefer our name, for this application.} is constructed independently of the $G$ function specifying the $M$-estimator, and so the same sketch can be used for all $G$. That is, one can first sketch the input, in one pass, and decide later on the particular choice of penalty function $G$. That is, the entire algorithm for the problem $\min_x \|Ax - b\|_G$ is to compute $S \cdot A$ and $S \cdot b$, for a simple sketching matrix $S$ described below, and then solve the regression problem $\min_x \|SAx - Sb\|_{G,w}$, where $\|\|_{G,w}$ is defined as follows.

**Definition 1.1.** For dimension $m$ and non-negative weights $w_1, \ldots, w_m$, define the weighted $G$-measure of a vector $y \in \mathbb{R}^m$, denoted $\|y\|_{G,w}$, to be $\sum_{i \in [m]} w_i G(y_i)$. We refer to $w$ as the weight vector.

Notice that $\|y\|_G$ equals $\|y\|_{G,w}$ when $w_i = 1$ for all $i$. If the $G$ function is convex, then using the non-negativity of $w$, it follows that $\|y\|_{G,w}$ is a convex function of $y$.

The sketch $SA$ (and $Sb$) can be computed in $O(\text{nnz}(A))$ time, and needs $O(\text{poly}(d\log n))$ space; we show that it can be used in $O(\text{poly}(d\log n))$ time to find approximate $M$-estimators, that with constant probability have a cost within a constant factor of optimal. The success probability can be amplified by independent repetition and choosing the best solution found among the repetitions.

**Condition on $G$.** For our results we need some additional conditions on the function $G$ beyond symmetry and monotonicity: that it grows no faster than quadratically in $x$, and no slower than linearly. Formally: there is $a \in [1,2]$ and $C_G > 0$ so that for all $a, a'$ with $|a| \geq |a'| > 0$,

$$a^n \geq C_G \frac{|a|}{|a'|}$$

The subquadratic growth condition is necessary for a sketch with a sketching dimension sub-polynomial in $n$ to exist, as shown by Braverman and Ostrovsky [8]. Also, subquadratic growth is appropriate for robust regression, to reduce the effect of large values in the residual $Ax - b$, relative to their effect in least-squares. Almost all proposed $M$-estimators satisfy these conditions [43].

The latter linear lower bound on the growth of $G$ holds for all convex $G$, and many popular $M$-estimators have convex $G$ [43]. Moreover, the convexity of $G$ implies the convexity of $\|\|_G$, which is needed for computing a solution to the minimization problem in polynomial time. Convexity also implies significant properties for the statistical interpretation of the results, such as consistency and asymptotic normality [23, 37].

However, we do not require $G$ to be convex for our sketching results, and indeed some $M$-estimators are not convex; here we simply reduce a large non-convex problem, $\min_x \|Ax - b\|_G$, to a smaller non-convex problem $\min_x \|S(Ax - b)\|_{G,w}$ of a similar kind. The linear growth lower bound does imply that we are unable to apply sketching to some proposed $M$-estimators; the “Tukey” estimator, for example, whose $G$ function is constant for large argument values, is not included in our results. However, we can get close, in the sense that at the cost of more computation, we can handle $G$ functions that grow arbitrarily slowly.

Not only do we obtain optimal $O(\text{nnz}(A) + \text{poly}(d\log n))$ time approximation algorithms for these $M$-estimators, our sketch is the first to non-trivially reduce the dimension of any of these estimators other than the $\ell_p$-norms (which are a special case of $M$-estimators). E.g., for the $L_1 - L_2$ estimator in which $G(x) = 2(\sqrt{1 + x^2} - 1)$, the Fair estimator in which $G(x) = c^2 \left( \frac{|x|}{c} - \log(1 + \frac{|x|}{c}) \right)$, or the Huber estimator,
no dimensionality reduction for the regression problem was known.

1.1 Techniques.

Huber algorithm. Our algorithm for the Huber estimator, §2, involves importance sampling of the \((A_1, b_1)\), where a sampling matrix \(S'\) is obtained such that \(\|S'(Ax - b)\|_H\) is a useful approximation to \(\|Ax - b\|_H\). The sampling probabilities are based on a combination of the \(\ell_1\) leverage score vector \(u \in \mathbb{R}^n\), and the \(\ell_2\) leverage score vector \(u' \in \mathbb{R}^n\). The \(\ell_1\) vector \(u\) can be used to obtain good sampling probabilities for \(\ell_1\) regression, and similarly for \(u'\) and \(\ell_2\). Since the Huber measure has a mixed character, we are able to use a combination of \(\ell_1\) and \(\ell_2\) scores to obtain good sampling probabilities for Huber. A key observation we use is Lemma 2.1, which roughly bounds the Huber norm of a vector in terms of \(n\), \(\tau\), and its \(\ell_1\) and \(\ell_2\) norms, and leads to a recursive sampling algorithm. Several difficulties arise, most notably that the Huber norm is not scale-invariant, that is, for small arguments it scales quadratically with its input while for large arguments it scales linearly. This complicates the sampling, as well as simple aspects such as net arguments typically used for \(\ell_p\)-regression, which relied on scale-invariance.

The \(M\)-sketch construction. Our sketch, a variant of that of Verbin and Zhang [40], is given formally as (3.6) in §3.1. It can be seen as a form of sub-sampling and finding heavy hitters, techniques common in data streams [25]; however, most analyses are aware of concerning such data structures, with the exception of that of Verbin and Zhang for earthmover distance, require a median operation in the sketch space and thus do not preserve convexity. This is the first time such sketches have been considered and shown to work in the context of regression.

We describe here a variant construction, comprising a sequence of sketching matrices \(S_0, S_1, \ldots, S_{h_{\text{max}}}\), for a parameter \(h_{\text{max}}\), each comprising a block of rows of our sketching matrix:

\[
S \equiv \begin{bmatrix}
S_0 \\
S_1 \\
S_2 \\
\vdots \\
S_{h_{\text{max}}}
\end{bmatrix}.
\]

When applied to vector \(y \in \mathbb{R}^n\), each \(S_h\) ignores all but a subset \(L_h\) of \(n/b^h\) entries of \(y\), where \(b > 1\) is a parameter, and where those entries are chosen uniformly at random. (That is, \(S_h\) can be factored as \(S_h' S''_h\), where \(S''_h \in \mathbb{R}^{n/b^h \times n}\) samples row \(i\) of \(A\) by having column \(i\) with a single 1 entry, and the rest zero, and \(S'_h\) has only \(n/b^h\) nonzero entries.)

Each \(S_h\) implements a particular sketching scheme called COUNT-SKETCH on its random subset. COUNT-SKETCH splits the coordinates of \(y\) into groups (“buckets”) at random, and adds together each group after multiplying each coordinate by a random \(\pm 1\); each such sum constitutes a coordinate of \(S_h y\).

COUNT-SKETCH was recently [9, 34, 36] shown to be a good subspace embedding for \(\ell_2\), implying here that the matrix \(S_h\), which applies to all the coordinates of \(y = Ax\), has the property that \(\|S_h^0 Ax\|_2\) is a good estimator of \(\|Ax\|_2\) for all \(x \in \mathbb{R}^d\); in particular, each coordinate of \(S_h y\) is the magnitude of the \(\ell_2\) norm of the coordinates in the contributing group.

Why should our construction, based on \(\ell_2\) embeddings, be suitable for, e.g., \(\ell_1\), with \(\|D(w)S Ax\|_1\) an estimate of \(\|Ax\|_1\)? Why should the \(M\)-sketch be effective for that norm? Here \(D(w)\) is an appropriate diagonal matrix of weights \(w\). An intuition comes from considering the matrix \(S_{h_{\text{max}}}\) for the smallest random subset \(L_{h_{\text{max}}}\) of \(y = Ax\) to be sketched: we can think of \(S_{h_{\text{max}}} y\) as one coordinate of \(y = Ax\), chosen uniformly at random and sign-flipped. The expectation of \(\|S_{h_{\text{max}}} y\|_1\) is \(\sum_{i \in [n]} \|y_i\|_2 / n = \|y\|_1 / n\); with appropriate scaling from \(D(w)\), that smallest random subset yields an estimate of \(\|y\|_1 = \|Ax\|_1\). (This scaling is where the values \(w\) are needed.) The variance of this estimate is too high to be useful, especially when the norm of \(y\) is concentrated in one coordinate, say \(y_1 = 1\), and all other coordinates zero. For such a \(y\), however, \(\|y\|_2 = \|y\|_1\), so the base level estimator \(\|S_{h_{\text{max}}} y\|_2\) is a good estimate. On the other hand, when \(y\) is the vector with all coordinates \(1/n\), the variance of \(\|S_{h_{\text{max}}} y\|_1\) is zero, while \(\|S_{h_{\text{max}}} y\|_2 \approx \|y\|_2\) is quite inaccurate as an estimator of \(\|y\|_1\). So in these extreme cases, the extreme ends of the \(M\)-sketch are effective. The intermediate matrices \(S_h\) of the \(M\)-sketch help with less extreme cases of \(y\)-vectors.

Analysis techniques. While helpful to the intuition, the above observations are not used to prove the results here. The general structure of our arguments is to show that, conditioned on several constant probability events, for a fixed \(x \in \mathbb{R}^d\) there are bounds on:

- contraction, so with high probability, \(\|S A x\|_{G, w}\) is not too much smaller than \(\|A x\|_G\);
- dilation, so with constant probability, \(\|S A x\|_{G, w}\) is not too much bigger than \(\|A x\|_G\).

This asymmetry in probabilities means that some results are out of reach, but still allows approximation algorithms for \(\min_x \|A x - b\|_G\). (We blur the distinction between applying \(S\) to \(A\) for vectors \(x \in \mathbb{R}^d\), and to \(A b\) for vectors \([x, \mathbb{b}]\).) If the optimum \(x_{\text{OPT}}\) for the original problem has \(\|S (A x_{\text{OPT}} - b)\|_G\) that is not too large, then it will be a not-too-large solution for the
sketched problem \( \min_x \|S(Ax - b)\|_{G,w} \). If contraction bounds hold with high probability for a fixed vector \( Ax \), and a weak dilation bound holds for every \( Ax \), then an argument using a metric-space \( \varepsilon \)-net shows that the contraction bounds hold for all \( x \); thus, there will be no \( x \) that gives a good, small \( \|S(Ax - b)\|_{G,w} \) and bad, large \( \|Ax - b\|_G \).

The contraction and dilation bounds are shown on a fixed vector \( y \in \mathbb{R}^n \) by splitting up the coordinates of \( y \) into groups ("weight classes") with the members of a weight class having roughly equal magnitude. (For \( y = SAx \), it will convenient to consider weight classes based on the values \( G(y_i) \), not \( |y_i| \) itself; for this section we won’t dwell on this distinction: assume here \( G(a) = |a| \).) A weight class \( W \) is then analyzed with respect to its cardinality: there will be some random subset ("level") \( L_h \) for which \( |W \cap L_h| \) is small relative to the number of rows of \( S_h \) (each row of \( S_h \) corresponds to a bucket, as an implementation of COUNT-SKETCH), and therefore the members of \( W \) are spread out from each other, in separate buckets. This implies that each member of \( W \) makes its own independent contribution to \( \|S y\|_{G,w} \), and therefore that \( \|S y\|_{G,w} \) will not be too small. Also, the level \( L_h \) is chosen such that the expected number of entries of the weight class is large enough that the random variable \( |W \cap L_h| \) is concentrated around its mean with exponentially small failure probability in \( d \), and so this contribution from \( W \) is well-behaved enough to union bound over a net.

The above argument works when the weight class \( W \) has many members, i.e., at least \( d \) coordinates in order to achieve concentration. For those \( W \) without many members which still contribute significantly to \( \|y\|_G \), we need to ensure that as we range over \( y \) in the subspace, these weight classes only ever involve a small fixed set of coordinates. We show this by relating the \( G \) function to the function \( f(x) = x^2 \), and arguing that these weight classes only involve coordinates with a large \( \ell_2 \) leverage score; thus the number of such coordinates is small and they can be handled separately once for the entire subspace by conditioning on a constant probability event.

To show that \( \|S y\|_{G,w} \) will not be too big, we show that \( W \) will not contribute too much to levels other than the "Goldilocks" level \( L_h \); for \( h < \hat{h} \), for which \( |L_{\hat{h}} \cap W| \) is expected to be large, the fact that members of \( W \cap L_{\hat{h}} \) will be crowded together in a single bucket implies they will cancel each other out, roughly speaking; or more precisely, the fact that the COUNT-SKETCH buckets have an expectation that is the \( \ell_2 \) norm of the bucket entries implies that if a bucket contains a large number of entries from one weight class, those entries will make a lower contribution to the estimate \( \|S y\|_{G,w} \) than they did for \( L_{\hat{h}} \). For \( h \) a bit bigger than \( \hat{h} \), \( W \cap L_h \) will likely be empty, and \( W \) will make no contribution to \( \|S_{\hat{h}} y\| \).

This argument does not work when the function \( G \) has near quadratic growth, and would result in an \( O(\log n) \) dilation. By modifying the estimator we can achieve an \( O(1) \) dilation by ignoring small buckets, and adding only those buckets in a level \( h \) that are among the top ones in value. Note that if \( G \) is convex, then so is this "clipped" version, since at each level we are applying a Ky Fan norm. The distinction of taking the top number of buckets versus those buckets whose value is sufficiently large seems important here, since only the former results in a convex program.

1.2 Outline. We give our algorithm for the Huber \( M \)-estimator in §2.

Next we give some definitions and basic lemmas related to \( M \)-sketches, that for a given vector \( y \), under appropriate assumptions \( S \) does not contract \( y \) too much (§3.5). We also show it does not dilate it too much (§3.6). In §3.6.2, we sharpen the dilation result by changing slightly the way we use the sketches, improving the dilation bound while preserving the contraction bound.

2 \( \varepsilon \)-Approximation for the Huber Measure.

Here we consider specifically the Huber measure: for parameter \( \tau > 0 \), and \( a \in \mathbb{R} \), the Huber function

\[
H(a) = \begin{cases} 
\frac{a^2}{2\tau} & \text{if } |a| \leq \tau \\
|a| - \tau^2 / 2 & \text{otherwise.}
\end{cases}
\]

The Huber "norm" is \( \|z\|_H = \sum_p H(z_p) \).

The main theorem of this section, proven in §2.1:

**THEOREM 2.1.** (Input Sparsity Time Huber Regression) In \( O(\text{nnz}(A) \log n) + \text{poly}(d/\varepsilon) \) time, given an \( n \times d \) matrix \( A \) with \( \text{nnz}(A) \) non-zero entries and \( n \times 1 \) vector \( b \), with probability at least 4/5, one can find an \( x' \in \mathbb{R}^d \) for which \( \|Ax' - b\|_H \leq (1 + \varepsilon) \min_{x \in \mathbb{R}^d} \|Ax - b\|_H \).

We will need to relate the Huber norm to the \( \ell_1 \) and \( \ell_2 \) norms. The following lemma is shown via a case analysis of the coordinates of the vector \( z \).

**LEMMA 2.1.** (Huber Inequality) For \( z \in \mathbb{R}^n \),

\[
\Theta(n^{-1/2}) \min \left\{ \|z\|_1, \|z\|_2^2/2\tau \right\} \leq \|z\|_H \leq \|z\|_1.
\]

**Proof.** For the upper bound, we note that \( H(a) \leq |a| \), whether \( |a| \leq \tau \) or otherwise, and therefore \( \|z\|_H \leq \sum_p H(z_p) \leq \sum_{p} \|z_p\| = \|z\|_1 \). We now prove the lower bound. We consider a modified Huber measure \( \|z\|_G \) given a parameter \( \tau > 0 \) in which

\[
G(a) = \begin{cases} 
\frac{a^2}{2\tau} & \text{if } |a| \leq \tau \\
|a| & \text{otherwise.}
\end{cases}
\]
Then $\|z\|_H \leq \|z\|_G \leq 2\|z\|_H$, and so it suffices to prove the lower bound for $\|z\|_G$.

By permuting coordinates, which does not affect the inequality we are proving, there is an $s$ for which

$$|z_1| \leq |z_2| \leq \cdots \leq |z_s| \leq \tau \leq |z_{s+1}| \leq \cdots \leq |z_n|.$$  

(We may have $s = 0$, when all $|z_i| \geq \tau$, or $s = n$, when all $|z_i| \leq \tau$.) Let $U = \sum_{j=s+1}^{n} |z_j|$ and $L = \sum_{j=1}^{s} |z_j|^2$. Consider the $n$-dimensional vector $w$ with $s$ coordinates equal to $\frac{L}{\sqrt{s}}$, one coordinate equal to $U$, and remaining coordinates equal to 0. Then,

$$\|w\|_G = s \cdot \frac{L}{s^2 \tau} + U = \frac{L}{2\tau} + U = \|z\|_G.$$  

Moreover,

$$\|w\|_1 = U + s \cdot \frac{\sqrt{L}}{\sqrt{s}} = U + \sqrt{sL} \geq \|z\|_1,$$  

since subject to a 2-norm constraint $L$, the 1-norm is maximized when all $s$ coordinates are equal. Also,

$$\frac{\|w\|^2}{2\tau} = \frac{L}{2\tau} + \frac{U^2}{2\tau} \geq \frac{\|z\|^2}{2\tau},$$  

since subject to a 1-norm constraint $U$, the 2-norm is maximized when a single non-zero coordinate.

Combining (2.2), (2.3), and (2.4), in order to show $\|z\|_G = \Omega(n^{-1/2})\min(\|z\|_1, \|z\|^2/2\tau)$ it suffices to show $\|w\|_G = \Omega(n^{-1/2})\min(\|w\|_1, \|w\|^2/2\tau)$. By the above, this is equivalent to showing

$$U + \frac{L}{2\tau} = \Omega(n^{-1/2}) \cdot \min \left( U + \sqrt{nL}, \frac{U^2}{2\tau} + \frac{L}{2\tau} \right),$$  

which since $s \leq n$, is implied by showing

$$U + \frac{L}{2\tau} = \Omega(n^{-1/2}) \cdot \min \left( U + \sqrt{nL}, \frac{U^2}{2\tau} + \frac{L}{2\tau} \right).$$  

Note that we can assume $U \neq 0$, as otherwise the inequality is equivalent to showing $\frac{L}{2\tau} = \Omega(n^{-1/2}) \cdot \min \left( \sqrt{nL}, \frac{L}{2\tau} \right)$. This holds since $\frac{L}{2\tau} = \Omega(n^{-1/2}) \frac{L}{2\tau}$. So we can assume $U > 0$, and by definition of $U$, this implies that $U \geq \tau$. We break the analysis into cases:

**Case:** $\frac{U^2}{2\tau} + \frac{L}{2\tau} \leq \frac{1}{4}(U + \sqrt{nL})$. What we need to show in this case to prove (2.5) is $U + \frac{L}{2\tau} = \Omega(n^{-1/2})(\frac{U^2}{2\tau} + \frac{L}{2\tau})$.

Suppose first that $\frac{L}{2\tau} \geq U$. Then what we need to show in this case is that $\frac{L}{2\tau} = \Omega(n^{-1/2})(\frac{U^2}{2\tau} + \frac{L}{2\tau})$. Since $\frac{L}{2\tau}$ appears on both the left and right hand sides, this follows from showing that $\frac{L}{2\tau} = \Omega(n^{-1/2})(\frac{U^2}{2\tau})$. Using the definition of this case, and that $U \geq \tau$, we have $U^2/4 + \frac{L}{2\tau} \leq U^2/4 + \frac{\sqrt{nL}}{4}$, which implies that $U^2 \leq 4L = \frac{\sqrt{nL}}{\tau}$, or equivalently, $\sqrt{nL} = \Omega(\tau)$. Since $\frac{L}{2\tau} \geq U \geq \tau$, we have $L = \Omega(\tau^2)$, as desired.

Otherwise, we have $U \geq \frac{L}{2\tau}$ and to prove (2.5) we need to show $U = \Omega(n^{-1/2}) \left( \frac{U^2}{2\tau} + \frac{L}{2\tau} \right)$. We can assume $U^2/2\tau \geq \frac{L}{2\tau}$, otherwise this is immediate from the fact that $U \geq \frac{L}{2\tau}$, and so we need to show $U = \Omega(n^{-1/2})\frac{U^2}{2\tau}$, or equivalently, $\frac{U}{\sqrt{\tau}} = O(\sqrt{n})$. Now use the fact that $\frac{U^2}{2\tau} + \frac{L}{2\tau} = \Theta(U^2/\tau)$ realizes the minimum given the case that we are in, and so $\frac{U^2}{2\tau} = O(U + \sqrt{nL})$, or equivalently, $\frac{U}{\sqrt{\tau}} = O(1 + \sqrt{nL})$. Since as mentioned it holds that $\frac{U^2}{2\tau} \geq \frac{L}{2\tau}$, we have $U^2 \geq L$, and so $\frac{U^2}{2\tau} \leq \sqrt{n}$. It follows that $\frac{U}{\sqrt{\tau}} = O(\sqrt{n})$, which is what we needed to show.

**Case:** $\frac{1}{4}\left(U + \sqrt{nL}\right) < \frac{U^2}{2\tau} + \frac{L}{2\tau}$. What we need to show in this case to prove (2.5) is $U + \frac{L}{2\tau} = \Omega(n^{-1/2})(U + \sqrt{nL})$.

Suppose first that $U \geq \frac{L}{2\tau}$, and so we need to show $U = \Omega(n^{-1/2})(U + \sqrt{nL})$, which is equivalent to showing $U = \Omega(\sqrt{nL})$. Since $\frac{L}{2\tau} \leq U$, we have $\sqrt{nL} = O(\sqrt{U^2/\tau}) = O(U)$, using that $U \geq \tau$. This completes this case.

Otherwise, we have $\frac{L}{2\tau} \geq U$ and need to show $\frac{L}{2\tau} = \Omega(n^{-1/2})(U + \sqrt{nL})$. We can assume $\sqrt{nL} \geq U$, otherwise this is immediate using $\frac{L}{2\tau} \geq U$, and so we need to show $\frac{L}{2\tau} = \Omega(n^{-1/2})\sqrt{nL} = \Omega(\sqrt{L})$, or equivalently, $L = \Omega(\tau^2)$. Now we use the fact that $U + \sqrt{nL} = \Theta(\sqrt{U^2/\tau})$ realizes the minimum, and so $\sqrt{nL} = O\left(\frac{U^2}{2\tau} + \frac{L}{2\tau}\right)$, and using that $U \leq \frac{L}{2\tau}$, this implies $\sqrt{nL} = O\left(\frac{U^2}{2\tau} + \frac{L}{2\tau}\right)$. Since $U \geq \tau$, it follows that $\sqrt{nL} = O\left(\frac{U^2}{2\tau}\right)$. Now using that $U \leq \sqrt{nL}$, this implies that $L = \Omega(\tau^2)$, which is what we needed to show.

This completes the proof.

Suppose we want to solve the Huber regression problem $\min_{x \in \mathbb{R}^d} \|Ax - b\|_H$, where $A$ is an $n \times d$ matrix and $b$ an $n \times 1$ column vector. We will do so by a recursive argument, and for that we will need to solve $\min_{x \in \mathbb{R}^d} \|Ax - b\|_{H,w}$ for various weight vectors $w$. Note that $\|Ax - b\|_{H,w}$ is a non-negative linear combination of convex functions of $x$, and hence is convex. We develop a lemma for this more general problem, given $w$. We maintain that if $w_i \neq 0$, then $w_i \geq 1$.

In our recursion we will have $\|w\|_{\infty} \leq \operatorname{poly}(n)$ for some polynomial that depends on where we are in the
Theorem 2.2. (Huber Embedding) With the notation defined above, for any fixed $x \in \mathbb{R}^d$, 
\[
\Pr[(1 - \varepsilon)\|Ax\|_{H, w} \leq \|Ax\|_{H, w} \leq (1 + \varepsilon)\|Ax\|_{H, w}] \geq 1 - \exp(-C_2 d\log(n/\varepsilon)),
\]
for an arbitrarily large constant $C_2 > 0$.

**Proof.** Fix a vector $x$ and define the non-negative random variable $X_i = w_i \cdot H(A_i x)$. For $X = \sum_{i=1}^n X_i$, we have $E[X] = \sum_{i=1}^n p_i (w_i/p_i) H(A_i x) = \|Ax\|_{H, w}$.

We will use the following version of the Bernstein inequality.

**Fact 2.1. ([33, 5])** Let $\{X_i\}_{i=1}^n$ be independent random variables with $E[X_i^2] < \infty$ and $X_i \geq 0$. Set $X = \sum_{i=1}^n X_i$ and let $\gamma > 0$. Then,
\[
\Pr[X \leq E[X] - \gamma] \leq \exp \left( \frac{-\gamma^2}{2 \sum_i E[X_i^2]} \right).
\]

If $X_i - E[X_i] \leq \Delta$ for all $i$, then with $\sigma_i^2 = E[X_i^2] - E[X_i]^2$ we have
\[
\Pr[X \geq E[X] + \gamma] \leq \exp \left( \frac{-\gamma^2}{2 \sum_i \sigma_i^2 + 2 \gamma \Delta / 3} \right).
\]

If for some $i$ we have $p_i = 1$, then $E[X_i] = w_i H(A_i x)$. It follows that such $X_i$ do not contribute to the deviation of $X$ from $E[X_i]$, and therefore we can apply Fact 2.1 only to those $X_i$ for which $p_i < 1$.

In order to apply Fact 2.1, we first bound $H(A_i x)/p_i$, for the case when $p_i < 1$, by a case analysis. Suppose $i \in P_i$. We use Lemma 2.1 to do the case analysis.

**Case** $(A_i x) \geq \tau$ and $\|A_i x\|_H = \Omega(n^{-1/2}\|A_i x\|_H)$. It follows that
\[
\frac{H(A_i x)}{p_i} \leq \frac{|A_i x| - \tau/2}{p_i} \leq \frac{|A_i x|}{p_i} \leq \frac{|A_i x|}{\Theta(s) q_i^2} = \frac{|A_i x|}{\Theta(s) \|U_i^j\|_1} \leq \frac{\|U_i^j\|_1 \|J_j x\|_\infty}{\Theta(s) \|U_i^j\|_1} \leq \frac{\alpha \beta \|A_j x\|_1}{\Theta(s)} \leq \frac{\alpha \beta O(n^{1/2}) \|A_j x\|_H}{s} = \frac{O(\|A_j x\|_H)}{C_0 \varepsilon^{-2} \log(n/\varepsilon)}.
\]

Case $(A_i x) \geq \tau$ and $\|A_i x\|_H = \Omega(n^{-1/2}\|A_i x\|_H)$ as well. Suppose not, so that $\|A_i x\|_H = \omega(n^{1/2})$.

Let $S \subseteq [n]$ be the set of $i \in [n]$ for which $|(A_i x)_{S}| \geq \tau$. Then $\|A_j x\|_H \geq \|A_j x\|_H |S| \geq \|A_j x\|_H / 2$. Hence,
\[
\omega(n^{1/2}) = \|A_j x\|_H \geq \|A_j x\|_H |S| \geq \|A_j x\|_H / 2 + \|A_j x\|_H / 2,
\]
so that $\|A_j x\|_H / 2 = \omega(n^{1/2})$. The value $\|A_j x\|_H / 2$ is minimized when all of the coordinates are equal:
\[
\|A_j x\|_H / 2 = \omega(n^{1/2}) = \omega(n^{1/2})\|A_j x\|_H / 2.
\]

Given a value of $\|A_j x\|_H / 2$, the value $\|A_j x\|_H / 2$ is minimized when all of the coordinates are equal:
\[
\|A_j x\|_H / 2 = \omega(n^{1/2}) / 2.
\]

Note also that $\|A_j x\|_H / 2 \geq \tau/2$ since there exists an $i$ for which $|A_i x| \geq \tau$ given that we are in this case.

So in order for the condition that $\|A_j x\|_H / 2 = \omega(n^{1/2})$, it must be the case that
\[
\|A_j x\|_H / 2 = \omega(n^{1/2}) \cdot \left( \tau + \frac{\|A_j x\|_H / 2}{2\tau n} \right).
\]
The right hand side of this expression is minimized when
\[ \tau^2 = \frac{\| (A'x)_n \|_2^2}{2n}, \]
which implies \( \Theta(\tau^2 n) = \| (A'x) \|_n \), or equivalently, \( \| (A'x) \|_1 = \Theta(\tau \sqrt{n}) \). But then we have
\[
\Theta(\tau \sqrt{n}) = \| (A'x) \|_n \|_1 = \Omega(n^{1/2}) \cdot 2\tau,
\]
which is a contradiction. Hence, \( \| A'x \|_H = \Omega(n^{-1/2} \| A'x \|_1) \), and this case reduces to the first case.

**Case** \( |A_i| \leq \tau \) and \( \| A'x \|_H = \Omega(n^{-1/2} \| A'x \|_1) \).

It follows that
\[
\frac{H(A_i)}{p_i} = \frac{(A_i)_n^2}{\tau |A_i|} \leq \frac{(A_i)_n^2}{\tau |A_i|} = \frac{|A_i|}{2p_i},
\]
using that \( |A_i| \leq \tau \). Now we have the same derivation as in the first case, up to a factor of 2.

**Case** \( |A_i| \leq \tau \) and \( \| A'x \|_H = \Omega(n^{-1/2} \| A'x \|_1^2/(2\tau)) \).

It follows using the properties of \( V_j \) that
\[
\frac{H(A_i)}{p_i} = \frac{(A_i)_n^2}{2\tau p_i} \leq \frac{(A_i)_n^2}{\tau s} \tau |V_j| \| K_jx \|_2 O(d) \leq \frac{|A'x|}{\tau s} \tau |V_j| \| K_jx \|_2 O(d) \leq \frac{O(d) \tau n^{1/2} \| A'x \|_H}{\tau s} \leq \frac{O(\| A'x \|_H)}{C_0 \varepsilon^{-2} \log(n/\varepsilon)}.\]

Hence, in all cases, if \( i \in P^j \) then
\[
\frac{H(A_i)}{p_i} \leq \frac{\| A'x \|_H}{C_1 \varepsilon^{-2} \log(n/\varepsilon)} \] for an arbitrarily large constant \( C_1 > 0 \). Then,
\[
X_i - E[X_i] \leq X_i \leq \frac{w_i H(A_i)}{p_i} \leq \frac{w_i \| A'x \|_H}{C_1 \varepsilon^{-2} \log(n/\varepsilon)}.
\]

Moreover, using the notation of Fact (2.1),
\[
\sum_{i : p_i < 1} \sigma_i^2 \leq \sum_{i : p_i < 1} E[X_i^2] = \sum_j \sum_{i : p_i < 1} w_i H(A_i) \frac{w_i H(A_i)}{p_i} \leq \sum_j \frac{\| A'x \|_{H,w}}{C_1 \varepsilon^{-2} \log(n/\varepsilon)} \sum_{i : p_i < 1} w_i H(A_i) x \leq \sum_j \frac{\| A'x \|_{H,w}^2}{C_1 \varepsilon^{-2} \log(n/\varepsilon)} \leq \frac{\| A'x \|_{H,w}^2}{C_1 \varepsilon^{-2} \log(n/\varepsilon)}.\]

Setting \( \gamma = \varepsilon \| A'x \|_{H,w} \), and applying Fact (2.1),
\[
\Pr[\| A'x \|_{H,w'} \leq \| A'x \|_{H,w} - \gamma] \leq \exp \left( -\gamma^2 C_1 \varepsilon^{-2} \log(n/\varepsilon) \right) \leq \exp(-C_2 d \log(n/\varepsilon)),\]

and also
\[
\Pr[\| A'x \|_{H,w'} \geq \| A'x \|_{H,w} + \gamma] \leq \exp \left( -\frac{-\gamma^2}{2 C_1 \varepsilon^{-2} \log(n/\varepsilon)} \right) \leq \exp(-C_2 d \log(n/\varepsilon)),\]

where \( C_2 > 0 \) is a constant that can be made arbitrarily large by choosing \( C_0 > 0 \) arbitrarily large.

We now combine Theorem 2.2 with a net argument for the Huber measure. We will use those arguments in Section 4. To do so, we need the following lemma.

**Lemma 2.2. (Huber Growth Condition)** The function \( H(a) \) satisfies the growth condition \( (1.1) \) with \( \alpha = 2 \) and \( C_0 = 1 \).

**Proof.** We prove this by a case analysis. We can assume \( a \) and \( a' \) are positive since the inequality only depends on the absolute value of these quantities. For notational convenience, let \( C = a/a' \). If \( a = a' \), the lemma is immediate, so assume \( C > 1 \).

First suppose \( a' \geq \tau \). Then \( H(a)/H(a') = (Ca' - \tau/2)/(a' - \tau/2) \), which is maximized when \( a' = \tau \), yielding \( (C - 1)/2 \leq 2C - 1 \). Since \( 2C - 1 \leq C^2 \) for \( C \geq 1 \), the left inequality of (1.1) holds. Conversely, \( H(a)/H(a') \) is at least \( C \), and so the right inequality of (1.1) holds.

Next suppose \( a \geq \tau \) and \( a' < \tau \). Then
\[
H(a)/H(a') = (Ca' - \tau/2)/(a')^2/2\tau) = 2\tau C/a' - \tau^2/(a')^2.
\]

Then
\[
\frac{d[H(a)/H(a')]}{da'} = -2\tau C/a'^2 + 2\tau^2/(a')^3,
\]
and setting this equal to 0 we find that \( a' = \tau/C \) maximizes \( H(a)/H(a') \). In this case \( H(a)/H(a') = C^2 \), and so the left inequality of (1.1) holds. Since \( a \geq \tau, \tau > a' \geq \tau/C \), and since \( \frac{d[H(a)/H(a')]}{da'} < 0 \) for \( a' \in (\tau/C, \tau] \), \( H(a)/H(a') \) is minimized when \( a' = \tau \), in which case it equals \( 2C - 1 \). Since \( 2C - 1 \geq C \) for \( C \geq 1 \), the right inequality of (1.1) holds.
Finally, suppose \( a < \tau \). In this case
\[
H(a)/H(a') = a^2/(a')^2 = C^2,
\]
and the left inequality of (1.1) holds, and the right inequality holds as well.

2.1 Proof of Theorem 2.1, Huber algorithm running time.

Proof. We first solve the least squares regression problem \( \min_x \|Ax-b\|_2 \) in \( O(\text{nnz}(A)) + \text{poly}(d/\varepsilon) \) time using [9] up to a factor of \( 1 + \varepsilon \). This step succeeds with probability \( 1-o(1) \). Suppose \( y^* = Ax^* - b \) realizes this minimum. Let \( c = \|y^*\|_1/(1+\varepsilon) \). Then by Lemma 4.4 as we will see in our net argument in §4, applied to \( w = 1^a \), if \( y^* = Ax^* - b \), where \( x^* = \text{argmin}_x \|Ax-b\|_H \), then \( c \leq \|y^*\|_1 \leq 2\kappa n^{3/2} \), where \( \kappa > 0 \) is a sufficiently large constant.

To apply Theorem 2.2 with \( w = 1^a \) first note that all weights \( w_i \) are in the same group \( P_1 \). We then need to be able to compute the sampling probabilities \( q_i^t \) and \( r_i \), but only up to a constant factor since \( p_i^t = \min(1, \Theta(s \cdot (q_i^t + r_i))) \). Recall, \( q_i^t = \|v_i^t/h\|_t \) and let \( r_i = \sum \|v_i^t/h\|_t \), where \( U_1^t \) and \( V_1^t \) denote the \( i \)-th row of \( U_1^t \) and \( V_1^t \), respectively. Here \( U_1^t \) is an \((\alpha, \beta)\)-well-conditioned basis for \( A \) with respect to \( \ell_1 \), meaning \( U_1^t \) has the same column span as \( A \), \( \sum \|U_i^t\|_1 = \alpha \), and for all \( x \), \( \|x\|_\infty \leq \beta \|U_1^t x\|_1 \). By Lemma 49 and Theorem 50 of [9] (see also [34, 41]), the \( q_i^t \) can be computed in \( O(\text{nnz}(A) \log n) + \text{poly}(d/\varepsilon) \) time, for a \( U_1^t \) with \( \alpha, \beta \leq \text{poly}(d) \). Similarly, by Theorem 29 of [9], in \( O(\text{nnz}(A) \log n) + \text{poly}(d/\varepsilon) \) time we can compute the \( r_i \) for a matrix \( V_1^t \) for which \( \sum \|v_i^t/h\|_t = \Theta(d) \) and for all \( x \), \( \|V_1^t x\|_1 \leq (1+1/2)\|x\|_2 \). These steps succeed with probability \( 1-1/\log^2 n \) for arbitrarily large constant \( C > 0 \).

The vector \( w^t \) in Theorem 2.2 can be computed in \( O(n) \) time, and the expected number of non-zero entries of \( w^t \) is \( O(s \log n) = O(n^{1/2} \max(\alpha \cdot \beta, d) \cdot \varepsilon^{-2} d \log(n/\varepsilon) \log n) = n^{1/2} (\log^2 n) \text{poly}(d/\varepsilon) \), and so with probability \( 1-o(1) \), we will have \( \text{nnz}(w^t) \leq n^{1/2} (\log^2 n) \text{poly}(d/\varepsilon) \).

Let \( T \) be the sparse subspace embedding of \( [9] \), so that with probability \( 1-o(1) \), \( \|TAx\|_2 = (1 \pm \varepsilon) \|Ax\|_2 \) for all \( x \) and \( TA \) can be computed in \text{nnz}(A) \) time and \( H \) has \text{poly}(d/\varepsilon) \) rows.

Now consider the regression problem \( \min_x \|Ax-b\|_{H,w} \) subject to the constraint \( \|TAx-Tb\|_2 \leq 2\kappa n^{3/2} \). This 2-norm constraint is needed to ensure that we satisfy the conditions needed to apply Lemma 4.5 in our net argument in §4. By a union bound, Theorem 2.2 holds simultaneously for all points in a net \( N \) of size \((n/\varepsilon)^{O(d)} \). This step succeeds with probability \( 1-o(1) \). Moreover, since \( w'_t = w_t/p_i \) with probability \( p_i \) (and zero otherwise), by a union bound the probability that a \( p_i \) of a nonzero \( w'_t \) is less than \( 1/n^2 \) is at most \( n/n^2 = 1/n \), so with probability \( 1-o(1) \), \( \|Ax-b\|_{H,w} \leq n^{2} \|Ax-b\|_H \) for all \( x \).

Hence, we can apply Lemma 4.5 with \( S \) equal to the identity and our choice of \( w' \) (together with the input constant \( 2r \)) to conclude, by a union bound that with probability \( 1-o(1) \), if \( x^* = \text{argmin}_x \|Ax-b\|_{H,w} \) subject to the constraint \( \|TAx-Tb\|_2 \leq 2\kappa n^{3/2} \), then \( \|Ax^*-b\|_{H,w} \leq (1+\varepsilon) \|Ax^*-b\|_H \).

Thus, we have reduced the original regression problem to the regression problem \( \min_x \|Ax-b\|_{H,w} \) constrained by \( \|TAx-Tb\|_2 \leq 2\kappa n^{3/2} \), where \( w' \) has \( n^{1/2} \log^2 n \cdot \text{poly}(d/\varepsilon) \) non-zero entries. We now repeat this procedure recursively \( O(1) \) times. Let \( w_0 = 1^{n} \) and \( w_1 = w' \). In the \( \ell \)-th recursive step, \( \ell \geq 2 \), we are given the regression problem \( \min_x \|Ax-b\|_{H,w_{\ell-1}} \) subject to the constraint \( \|TAx-Tb\|_2 \leq 2\kappa n^{3/2} \). We now describe the \( \ell \)-th recursive step.

We inductively have that \( \|w_{\ell-1}\|_\infty \leq n^{2\ell-2} \). We first group the weights of \( w_{\ell-1} \) into \( O(\log n) \) groups \( P_j \). For each group we compute \( U_j^t \) and \( V_j^t \) as above, thereby obtaining \( w_j \) in \( O(t_0 \log n) \) expected time, where \( t_0 = \ell \) is the number of non-zero weights in \( w_{\ell-1} \). The expected value of \( t_0 = O(t_0 \log(n/\varepsilon) \log n) \). We can condition on \( t_0 \) being \( O(n^{1/2} \log d \log(n/\varepsilon) \log n) \) as events jointly succeed with probability \( 1-o(1) \). We thus have \( t_0 = n^{2/2} \log(n/\varepsilon) \log n) \) with probability \( 1-o(1) \). We now consider the regression problem \( \min_x \|Ax-Ab\|_{H,w_{\ell-1}} \) subject to the constraint \( \|TAx-Tb\|_2 \leq 2\kappa n^{3/2} \). By a union bound, Theorem 2.2 holds simultaneously for all points in a net \( N \) of size \((n/\varepsilon)^{O(d)} \), this step succeeding with probability \( 1-o(1) \). Moreover, the \( w' \) in Theorem 2.2 is equal to \( w \) and satisfies \( \|w\|_\infty \leq n^{2\ell-2} \|w_{\ell-1}\|_\infty \leq n^{-2\ell} \). We can thus apply Lemma 4.5 with \( S \) equal to the identity to conclude that with probability \( 1-o(1) \), if \( x^* = \text{argmin}_x \|Ax-b\|_{H,w_{\ell-1}} \) subject to the constraint \( \|TAx-Tb\|_2 \leq 2\kappa n^{3/2} \), then \( \|Ax^*-b\|_{H,w_{\ell-1}} \leq \min_{x} \|Ax-b\|_{H,w_{\ell-1}} \).

It follows that for \( \ell \) a large enough constant, and by scaling \( \varepsilon \) by a constant factor, we will have that with probability \( 1-o(1) \), if \( x^* = \text{argmin}_x \|Ax-b\|_{H,w_{\ell-1}} \) subject to the constraint \( \|TAx-Tb\|_2 \leq 2\kappa n^{3/2} \), then \( \|Ax^*-b\|_H \leq (1+\varepsilon) \min_{x} \|Ax-b\|_H \). Moreover, \( t_0 \leq n^{2\ell} \log(n/\varepsilon) \log n) \). This resulting problem is that of minimizing a convex function subject to a convex constraint and can be solved using the ellipsoid method.
in $t^C$ time for a fixed constant $C > 0$. Setting $2^C > C/2$ and assuming the poly$(d c^{-1} \log n)$ factor is at most $n^{1/2}$ gives us a running time of $O(n)$ to solve this last recursive step of the problem. The overall running time of the recursion is dominated by the time to compute the $U^j$ and $V^j$ in the different recursive levels, which itself is dominated by the top-most level of recursion. This gives an overall running time of $O(n z(A) \log n) + \text{poly}(d/\varepsilon)$.

3 $M$-sketches for $M$-estimators.

Given a function $G : \mathbb{R} \to \mathbb{R}^+$ with $G(a) = G(-a)$, and $G(0) = 0$, we can use the sketch of $z \in \mathbb{R}^n$ to estimate $\|z\|_G \equiv \sum_i G(z_i)$, assuming $G$ is monotone and satisfies the growth upper and lower bounds of (1.1).

Perhaps a more consistent notation would define the measure based on $G$ as $G^{-1}(\|z\|_G)$, by analogy with the $\ell_p$ norms. Moreover, $\|z\|_G$ does not in general satisfy the properties of a norm. However, if $G$ is convex, then $\|y\|_G$ is a convex function of $z$, and if also $G^{-1}(\|z\|_G)$ is scale-invariant, so that $G^{-1}(t |z|_G) = t |G^{-1}(\|z\|_G)|$, then $G^{-1}(\|z\|_G)$ is a norm.

The sketch. We use an extension of COUNT-SKETCH, which has been shown to be effective for subspace embeddings [9, 36, 34]. In that method, for a vector $z \in \mathbb{R}^n$, each coordinate $z_p$ is mapped via a hash function from $[n]$ to one of $N$ hash buckets, written as $g_p \in [N]$ for $p \in [n]$; a coordinate is generated for bucket $g \in [N]$ as $\sum_{g_p = g} \Lambda_p z_p$, where $\Lambda_p = \pm 1$ is chosen independently at random with equal probability for $+1$ and $-1$. The resulting $N$-vector has approximately the same $\ell_2$ norm as $z$.

Here we employ also sampling of the coordinates, as done in the context of estimating earthmover distance in [40], where each coordinate $z_p$ is mapped to a level $h_p$, and the number of coordinates mapped to level $h$ is exponentially small in $h$: for an integer branching factor $b > 1$, we expect the number of coordinates at level $h$ to be about a $b^{-h}$ fraction of the coordinates. The number of buckets at a given level is $N = bcm$, where integers $m, c > 1$ are parameters to be determined later.

Our sketching matrix implementing this approach is $S \in \mathbb{R}^{N h_{\max} \times n}$, where $h_{\max} \equiv \lceil \log_b (n/m) \rceil$, and our scaling vector $w \in \mathbb{R}^{N h_{\max}}$. The entries of $S$ are $S_{i,p} \leftarrow \Lambda_p$, and the entries of $w$ are $w_j \leftarrow \beta^{b h_p}$, where $\beta \equiv (b - b^{-h_{\max}})/(b - 1)$, $j \leftrightarrow g_p + N h_p$, and

$$
\Lambda_p \leftarrow \pm 1 \text{ with equal probability} \\
g_p \in [N] \text{ chosen with equal probability} \\
h_p \leftarrow h \text{ with probability } 1/\beta b^h \text{ for int } h \in [0, h_{\max}].
$$

all independently. Let $L_h$ be the multiset of values at a given level, $L_{h,i}$ is the multiset of values in a bucket. We can write $\|S z\|_{G,w}$ as $\sum_{i \in [N], h \in [0, h_{\max}]} \beta^h G(\|L_{h,i}\|_A)$, where $\|L\|_A$ denotes $|\sum_{z \in L} \Lambda_p z_p|$.

(The function $\|\cdot\|_A$ is a semi-norm (if we map sets back to vectors), with $\|L\|_A \leq \|L\|_1$, $E_{\Lambda} \|L\|_A^2 = \|L\|_2^2$, and all $(E_{\Lambda} \|L\|_2^k)^{1/k}$ within constant factors of $\|L\|_2$, by Khintchine’s inequality.)

Regression theorem. Our main theorem of this section states that $M$-sketches can be used for regression.

Theorem 3.1. (Input Sparsity Time Regression for $G$-functions) Let $\text{OPT}_G \equiv \min_{x \in \mathbb{R}^d} \|A x - b\|_G$. There is an algorithm that in $n z(A) + \text{poly}(d \log n)$ time, with constant probability finds $\hat{x}$ such that $\|A \hat{x} - b\|_G \leq O(1) \text{OPT}_G$.

The proof is deferred to §4.1; it requires a net argument, Lemma 4.5; the contraction bound Theorem 3.2 from §3.5; and from §3.6, a clipped variant Theorem 3.4 of the dilation bound Theorem 3.3. First, various definitions, assumptions, and lemmas will be given.

3.1 Preliminary Definitions and Lemmas for $M$-estimators. We will analyze the behavior of sketching on $z \in \mathbb{R}^n$. We assume that $\|z\|_G = 1$; this is for convenience of notation only, the same argument would apply to any particular value of $\|z\|_G$ (we do not assume scale-invariance of $G$).

Define $y \in \mathbb{R}^d$ by $y_p = G(z_p)$, so that $\|y\|_1 = \|z\|_G = 1$. A large part of our analysis will be related to $y$, although $y$ does not appear in the sketch. Let $Z$ denote the multiset comprising the coordinates of $y$, and let $Y$ denote the multiset comprising the coordinates of $y$. For $\tilde{Z} \subset Z$, let $G(\tilde{Z}) \subset Y$ denote $\{G(z_p) \mid z_p \in \tilde{Z}\}$. Let $\|Y\|_k \equiv \{\sum_{y \in Y} \|y\|_k^k\}$, so $\|Y\|_1 = \|y\|_1$. Hereafter multiset will just be called “sets”.

Weight classes. For our analysis, fix $\gamma > 1$, and for integer $q \geq 1$, let $W_q$ denote weight class $\{y_p \in Y \mid y^{-q} \leq y_p \leq \gamma y^{-q}\}$. We have $\beta b^h E[\|G(L_h) \cap W_q\|_1] = \|W_q\|_1$. For a set of integers $Q$, let $W_Q$ denote $\cup_{q \in Q} W_q$.

Defining $q_{\max}$ and $h(q)$. For given $\varepsilon > 0$, consider $y' \in \mathbb{R}^d$ with $y'_i \leftarrow y_i$ when $y_i > \varepsilon / n$, and $y'_i \leftarrow 0$ otherwise. Then $\|y'\|_1 \geq 1 - n(\varepsilon / n) = 1 - \varepsilon$. Thus for some purposes we can neglect $W_q$ for $q > q_{\max} \equiv \log_b (n/\varepsilon)$, up to error $\varepsilon$. Moreover, we can assume that $\|W_q\|_1 \geq \varepsilon / q_{\max}$, since the total contribution of weight classes of smaller total weight to $\|y\|_1$ is at most $\varepsilon$.

Let $h(q)$ denote $\lceil \log_b (|W_q|/\beta m) \rceil$ for $|W_q| \geq \beta m$, and zero otherwise, so that

$$m \leq E[\|G(L_{h(q)}) \cap W_q\|]$$
for all $W_q$ except those with $|W_q| < \beta m$, for which the lower bound does not hold.

Since $|W_q| \leq n$ for all $q$, we have $h(q) \leq \lfloor \log_b(n/\beta m) \rfloor = h_{\text{max}}$.

### 3.2 Assumptions About the Parameters.

There are many minor assumptions about the relations between various numerical parameters; some of them are collected here for convenience of reference. Recall that $N = bcn$.

**Assumption 3.1.** We will assume $b \geq m$, $b > c$, $m = \Omega(\log \log(n/\varepsilon))$, $\log b = \Omega(\log \log(n/\varepsilon))$, $\gamma \geq 2 \geq \beta$, an error parameter $\varepsilon \in (0, 1/3)$, and $\log N \leq \varepsilon^4 m$. We will consider $\gamma$ to be fixed throughout, that is, not dependent on the other parameters.

### 3.3 Distribution into Buckets.

The entries of $y$ are well-distributed into the buckets, as the following lemmas describe.

**Lemma 3.1.** For $\varepsilon \leq 1$, with failure probability at most $4q_{\text{max}}h_{\text{max}}\exp(-\varepsilon^2 m/3) \leq C^{-\varepsilon m}$ for a constant $C > 1$, the event $\mathcal{E}$ holds, that for all $q \leq q_{\text{max}}$ with $|W_q| \geq \beta m$, and all $h \leq h(q)$, that

$$|G(L_h) \cap W_q| = \beta^{-1} b^{-h}|W_q|(1 \pm \varepsilon),$$

and

$$||G(L_h) \cap W_q||_1 = \beta^{-1} b^{-h}||W_q||_1(1 \pm \varepsilon).$$

Here $a = b(1 \pm \varepsilon)$ means that $|a - b| \leq \varepsilon |b|$.

We will hereafter generally assume that $\mathcal{E}$ holds.

**Proof.** Let $s \equiv |W_q|$. When $s \geq \beta m$ and $h \leq h(q)$, in expectation $|G(L_h) \cap W_q|$ is equal to $\mu \equiv s/\beta b^h \geq m$, and $||G(L_h) \cap W_q||_1 \geq ||W_q||_1/\beta b^h$. We need that with high probability, deviations from these bounds are small.

Applying Bernstein’s inequality to the random variable $Z$ with binomial $B(s, 1/\beta b^h)$ distribution, the logarithm of the probability that $t \equiv Z - \mathbb{E}[Z] = Z - \mu$ exceeds $\varepsilon \mu$ is at most

$$-\frac{(\varepsilon \mu)^2}{\mu + (\varepsilon \mu)/3} \leq -\varepsilon^2 \mu/3 \leq -\varepsilon^2 m/3.$$

Taking the exponential, and using a union bound over all events (including the event that $t$ exceeds $\varepsilon \mu/\beta b^h$) completes the first claim, with half the claimed failure probability, using Assumption 3.1 to show that the claimed $C$ exists. For the second claim, there is a similar argument for the random variables $X_p$, which are equal to $y_p$ when $h_p = h$ and $y_p \in W_q$, and zero otherwise. Here $\sum_p \mathbb{E}[X_p^2] \leq \sum_p \mathbb{E}[X_p] = ||W_q||_1/\beta b^h$.

**Lemma 3.2.** For $h \in [h_{\text{max}}]$, suppose $Q \subset \{q \mid h(q) = h, |W_q| \geq \beta m\}$, and $W \subset Y$ contains $W_q \equiv \cup_{q \in Q} W_q$. If $|G(L_h) \cap W| \leq \varepsilon N$, then with failure probability at most $2|Q| \exp(-\varepsilon^2 m/3)$, each $W_q$ has $W_q^* \subset G(L_h) \cap W_q$ with $|W_q^*| \geq (1 - \varepsilon)\beta^{-1} b^{-h}|W_q|$, and where each entry of $W_q^*$ is in a bucket with no other element of $W$. Also if condition $\mathcal{E}$ of Lemma 3.1 holds, then

$$||W_q^*||_1 \geq (1 - 4\gamma \varepsilon)\beta^{-1} b^{-h}||W_q||_1.$$

**Proof.** We will show that for $q \in Q$, with high probability it will hold that $a_q \geq (1 - \varepsilon)\beta^{-1} b^{-h}|W_q|$, where $a_q$ is the number of buckets $G(L_h,i)$, over $i \in [N]$, containing a member of $W_q$, and no other members of $W$.

Consider each $q \in Q$ in turn, and the members of $W_q$ in turn, for $k = 1, 2, \ldots s \equiv |W_q|$, and let $Z_k$ denote the number of bins occupied by the first $k$ members of $W_q$. The probability that $Z_{k+1} > Z_k$ is at least $\beta^{-1} b^{-h}(1 - |G(L_h) \cap W_q|/N) \geq \beta^{-1} b^{-h}(1 - \varepsilon)$. We have $a_q \geq (1 - \varepsilon)\beta^{-1} b^{-h}|W_q|$ in expectation.

To show that this holds with high probability, let $\hat{Z}_k = \mathbb{E}[Z_k | Z_k]$. Then $\hat{Z}_1, \hat{Z}_2, \ldots$ is a Martingale with increments bounded by $1$, and with the second moment of each increment at most $\beta^{-1} b^{-h}$. Applying Freedman’s inequality gives a concentration for $a_q$, similar to the above application of Bernstein’s inequality, yielding a failure probability $2\exp(-\varepsilon^2 m/3)$.

Applying a union bound over all $|Q|$ yields that with probability at least $1 - 2|Q| \exp(-\varepsilon^2 m/3)$, for each $W_q$ there is $W_q^*$ of size at least $(1 - \varepsilon)\beta^{-1} b^{-h}|W_q|$ such that each member of $W_q^*$ is in a bucket containing no other member of $W$.

For the last claim, we compare the at least $(1 - \varepsilon)X$ entries of $W_q^*$, where $X \equiv \beta^{-1} b^{-h}|W_q|$, with the at most $(1 + \varepsilon)X - |W_q^*|$ entries of $G(L_h) \cap W_q$ not in $W_q^*$, using condition $\mathcal{E}$; we have

$$||W_q^*||_1 \geq (1 - \varepsilon)X\gamma^{-q} - (1 - \varepsilon)X\gamma^{-q} + 2\varepsilon X\gamma^{1-q} - \varepsilon^2 N \geq 1 - 2\gamma \varepsilon/(1 - \varepsilon).$$

Using condition $\mathcal{E}$ again to make the comparison with $||W_q||_1$, the claim follows.

**Lemma 3.3.** For $h \in [h_{\text{max}}]$, $\bar{W} \subset G(L_h)$, $T \geq ||\bar{W}||_\infty$, and $\delta \in (0, 1)$, if

$$N \geq \frac{6||\bar{W}||_1}{T \log(N/\delta)},$$

then with failure probability $\delta$,

$$\max_{i \in [N]} ||G(L_h,i) \cap \bar{W}||_1 \leq \frac{7}{6} T \log(N/\delta).$$
Leverage scores. This lemma will be used to bound the growth of $G$ that we need the linear lower bound of $\kappa$, the event $\Lambda = \{ v \}$, so with probability at least $1 - \kappa$, the event $E_\kappa$ holds that $S$ sends each member of $Y_1$ into a bucket containing no other member of $Y_2$.

We will hereafter generally assume that $E_\kappa$ holds.

Proof. For each member of $Y_2$, the expected number of members of $Y_1$ colliding with it, that is, in the same bucket with it, is $N_1/N$. The expected number of such collisions is therefore at most $N_1N_2/N < \kappa$. The probability that the number of collisions is at least twice its mean is at most $2\kappa$, so with probability at least $1 - 2\kappa$, the number of collisions is less than $2\kappa < 1$, that is, zero.

We use the $\ell_2$ leverage scores to bound the coordinates of $G(z)$; this is the one place in proving contraction bounds that we need the linear lower bound of (1.1) on the growth of $G$.

**Lemma 3.4.** If $u_p$ is the $k$th largest $\ell_2$ leverage score, then for $z \in C(A)$, $G(z_p) \leq \sqrt{2d/k\|z\|_G/C_G}$.

Here $C_G$ is the growth parameter from (1.1).

Proof. We have $u_p \leq d/k$, since $\sum_i u_i = d$. For $z = Ux \in C(A)$,

$$z_p^2 \leq (U_{pp}, x)^2 \leq \|U_{pp}\|^2 \|x\|^2 = u_p \|z\|^2 \leq (d/k) \|z\|^2.$$  

That is, $\sum_{z_i \geq z_p} z_i^2 / z_p^2 \geq k/d$. Suppose $\sum_{z_i \geq z_p} z_i^2 / z_p^2 \geq k/2d$. Then

$$\sum_{z_i \geq z_p} \frac{G(z_i)}{G(z_p)} \geq \sum_{z_i \geq z_p} \frac{z_i^2}{z_p^2} \geq \sum_{z_i \geq z_p} \left(\frac{z_i}{z_p}\right)^2 \geq k/2d,$$

and the claimed inequality follows. Otherwise, $\sum_{z_i \geq z_p} z_i^2 / z_p^2 \geq k/2d$, which implies

$$\sum_{z_i \geq z_p} \frac{G(z_i)}{G(z_p)} \geq C_G \sum_{z_i \geq z_p} \frac{z_i^2}{z_p} \geq C_G \left(\sum_{z_i \geq z_p} \frac{z_i}{z_p}\right)^2 \geq C_G \sqrt{k/2d},$$

and the claimed inequality follows.

**3.5 Contraction bounds.** Here we will show that $\|S\|_{G,w}$ is not too much smaller than $\|z\|_G$.

**3.5.1 Estimating $\|z\|_G$ using $S\bar{z}$.** For $v \in T \subset \mathcal{Z}$, let $T - v$ denote $T \setminus \{v\}$.

**Lemma 3.6.** For $v \in T \subset \mathcal{Z}$,

$$G(||T||_\Lambda) \geq \left(1 - \frac{\|T - v\|_\Lambda}{\|v\|}\right)^2 G(v),$$

and if $G(v) \geq \varepsilon^{-1} \|T - v\|_G$, then

$$\frac{\|T - v\|_2}{\|v\|} \leq \varepsilon^{1/\alpha},$$

and for a constant $C$, $E_\Lambda[G(||T||_\Lambda)] \geq (1 - C \varepsilon^{1/\alpha})G(v)$.

Proof. For the first claim, if $||T||_\Lambda \geq \|v\|$, then the claim is immediate since $G$ is non-decreasing. Otherwise, note that $||T||_\Lambda$ has the form $\|v\| - \|T - v\|_\Lambda$, so if $||T||_\Lambda \leq \|v\|$, then $||T||_\Lambda = \|v\| - ||T - v||_\Lambda$. We have

$$\frac{G(||T||_\Lambda)}{G(v)} \geq \left(\frac{||T||_\Lambda}{\|v\|}\right)^\alpha \geq \left(\frac{||T - v||_\Lambda}{\|v\|}\right)^2 = \left(1 - \frac{\|T - v\|_{\Lambda}}{\|v\|}\right)^2,$$

proving the first claim. For the second claim, we have $|v'| < \|v\|$ for $v' \in T - v$, since $G(v') \leq \|T - v\|_{\Lambda} \leq \varepsilon G(v)$, and $G$ is non-decreasing in $\|v\|$. Therefore

$$\varepsilon \geq \frac{\|T - v\|_G}{G(v)} = \sum_{v' \in T - v} \frac{G(v')}{G(v)} \geq \sum_{v' \in T - v} \left(\frac{|v'|}{|v|}\right)^\alpha \geq \sum_{v' \in T - v} \left(\frac{|v'|}{|v|}\right)^2,$$

and so (3.7) follows. For the third claim, we have from the first claim,

$$E_\Lambda[G(||T||_\Lambda)] \geq E_\Lambda\left[\left(1 - \frac{||T - v||_\Lambda}{\|v\|}\right)^2\right] G(v) \geq \left(1 - 2 \frac{E_\Lambda[||T - v||_\Lambda]}{\|v\|}\right) G(v).$$
Using the Khintchine inequality and (3.7), we have
\[
\frac{E\|T - v\|_4}{|v|} \leq C\frac{\|T - v\|_2}{|v|} \leq C\varepsilon^{1/\alpha},
\]
for a constant \(C\), so the claim follows, after adjusting constants.

We will need a lemma that will allow bounds on the contributions of the weight classes. First, some notation. For \(h = 0 \ldots h_{\text{max}}\), let
\[
\begin{align*}
\hat{Q}_h &\equiv \{ q \mid h(q) = h, |W_q| \geq \beta m \} \\
M_\geq &\equiv \log_\gamma(2(1 + 3\varepsilon)b/\varepsilon) \\
Q_h &\equiv \{ q \in \hat{Q}_h \mid q \leq M_\geq + \min q \} \\
M_\leq &\equiv \log_\gamma(m/\varepsilon) = O(\log_\gamma(b/\varepsilon)) \\
Q_\leq &\equiv \{ q \mid |W_q| < \beta m, q \leq M_\leq \} \\
Q^* &\equiv Q_\leq \cup \{hQ_h\}.
\end{align*}
\]
(3.8)

Here \(\hat{Q}_h\) gives the indices of \(W_q\) that are “large” and have \(h\) as the level at which between \(m\) and \(bm\) members of \(W_q\) are expected in \(L_h\). The set \(Q_h\) cuts out the weight classes that can be regarded as negligible at level \(h\).

**Lemma 3.7.** Using Assumption 3.1 and assuming condition \(E\) of Lemma 3.1, \(\sum_{q \in Q^*} \|W_q\|_1 \geq 1 - 5\varepsilon\).

**Proof.** The total weight of those weight classes with \(|W_q| \leq \beta m\) and \(q > M_\leq\) is at most
\[
\beta m \sum_{q > M_\leq} \gamma^{1-q} \leq \beta m (\varepsilon/m) \gamma \sum_{q > 0} \gamma^{-q} \leq \varepsilon (1 - 1/\gamma) \leq 4\varepsilon
\]
for \(\gamma \geq 2\) and \(\beta \leq 2\).

For given \(h > 0\), let \(q_h^* \equiv \min_{q \in \hat{Q}_h} q\). The ratio of the total weight of classes in \(Q_h \setminus \hat{Q}_h\) to \(\|W_{q_h^*}\|_1\) is at most
\[
\frac{1}{(1 - \varepsilon)^{\gamma - q_h^* M_\leq}} \sum_{q > 0} (1 + \varepsilon)b m \gamma^{-q} = \frac{\varepsilon}{2b(1 + 3\varepsilon)} \frac{1 + \varepsilon}{1 - \varepsilon},
\]
under the assumptions on \(\gamma\) and \(\varepsilon\). So \(\sum_{h} \|W_{Q_h \setminus \hat{Q}_h}\|_1 \leq \varepsilon \|W_{Q_h^*}\|_1 \leq \varepsilon\).

Putting together the bounds for the two cases, the total is at most \(5\varepsilon\), as claimed.

**Lemma 3.8.** Assume that condition \(E\) of Lemma 3.1 holds, and that condition \(E_c\) of Lemma 3.4 holds for \(N_1 = N_2 = O(C^2 \varepsilon^{-2} dm^2)\). Let \(Q_h \equiv \{ q \mid q \leq M'_h \},\) where \(M'_h \equiv \log_\gamma(b^{h+1} m^2 q_{\text{max}})\). Then there is \(N = O(N_1^2 + m^2 \varepsilon^{-1} q_{\text{max}})\) so that with probability at least \(1 - C^{-\varepsilon^2 m}\) for a constant \(C > 1\), for each \(q \in Q^*\), there is \(W_q^* \subset L_{h(q)} \cap W_q\) such that:

1. \(\|W_q^*\| \geq (1 - \varepsilon)\beta^{-1} b^{-h(q)}|W_q|\);
2. each \(x \in W_q^*\) is in a bucket with no other member of \(W_{Q_q^*}\);
3. \(\|W_q^*\|_1 \geq (1 - 4\gamma \varepsilon)\beta^{-1} b^{-h} \|W_q\|_1\);
4. for \(q \in Q_h\), each \(x \in W_q^*\) is in a bucket with no member of \(W_{Q_q^*}\).

**Proof.** There is \(N_1\) satisfying the given bound so that Lemma 3.5 implies that \(y \notin Y_1\) must be smaller than \(C_1 \sqrt{2d/N_1} \leq \varepsilon/m\), and therefore not in \(W_q\) for \(q \in Q_\leq\). Therefore \(W_{Q_\leq} \subset Y_1\), and with the assumption of condition \(E_c\), no member of \(W_{Q_\leq}\) is in the same bucket as any other member of that set. We will take \(W_q^* \leftarrow W_q\) for \(q \in Q_\leq\).

For each \(h\), apply Lemma 3.2 to \(Q_h\) and with \(W \leftarrow W_{Q_h} \equiv W_{Q_\leq} \cup q \in \hat{Q}_h\), so that, using condition \(E\),
\[
\|G(L_h) \cap \hat{W}\| \leq M_\leq \beta m + M_\geq (1 + \varepsilon)b m = O(m \log \log (b/\varepsilon)).
\]

To apply Lemma 3.2, we need \(N > \varepsilon^{-1} \|G(L_h) \cap \hat{W}\|\), and large enough \(N\) in \(O(m b \varepsilon^{-1} \log_\gamma(b/\varepsilon))\) suffices for this. We have (1) and (2), with failure probability \(2 M_\leq \exp(-\varepsilon^2 m)\).

Condition (3) follows either trivially, for \(q \in Q_\leq\), or from Lemma 3.2.

For (4), let \(W \leftarrow W_{Q_h} \cup W_{Q_h^*}\). Since \(|W_{Q_h^*}| \gamma^{-M_\leq} \leq \|y\|_1 \leq 1\), so that \(|W_{Q_h^*}| \leq b^{h+1} m^2 q_{\text{max}}\), we have
\[
\|G(L_h) \cap \hat{W}\| \leq \|G(L_h) \cap W_{Q_h}\| + \|G(L_h) \cap W_{Q_h^*}\| \leq (1 + \varepsilon)b m M_\geq + (1 + \varepsilon)\beta^{-1} b^{-h} \|W_{Q_h^*}\| \leq O(b m^2 q_{\text{max}}),
\]
using condition \(E\). Since \(\|G(L_h) \cap \hat{W}\| \leq \varepsilon N\) for large enough \(N = O(m^2 b \varepsilon^{-1} q_{\text{max}})\), we can apply Lemma 3.2 to obtain (4).

**Lemma 3.9.** Let \(G : \mathbb{R} \mapsto \mathbb{R}^+\) as above. Assume that condition \(E\) of Lemma 3.1 holds, and Assumption 3.1, and that condition \(E_c\) of Lemma 3.4 holds for \(N_1 = N_2 = O(C^{-2} \varepsilon^{-2} dm^2)\). There is \(N = O(N_1^2 + \cdots)\).
\[ \varepsilon^{-2} m^2 b q_{\text{max}} \], so that for \( h \in [h_{\text{max}}] \) and \( q \in Q_h \) with \( \|W_q\|_1 \geq \varepsilon / q_{\text{max}} \), we have
\[
\sum_{y_p \in W_q^*} G(\|L(y_p)\|_1) \geq (1 - \varepsilon^{1/\alpha})\|W_q\|_1
\]
with failure probability at most \( C^{-\varepsilon^2 m} \) for fixed \( C > 1 \).

**Proof.** For any \( q \in Q_h \) we have
\[
\|W_q\| \leq (1 + \varepsilon) \beta b^n \mathbf{E}[\|G(L_h) \cap W_q\|] \\
\leq (1 + \varepsilon) \beta b^n m
\]
by condition \( \mathcal{E} \) and the definition of \( h(q) = h \); since
\[
\|W_q\| \geq \|W_q\|_1 \geq \varepsilon / q_{\text{max}},
\]
using \( \|W_q\|_1 \geq \varepsilon / q_{\text{max}} \) from the lemma statement, we have for any \( y_p \in W_q \),
\[ (3.9) \]
\[
y_p \geq \gamma^{-q} \geq (\varepsilon / q_{\text{max}}) / \gamma \|W_q\| \geq \varepsilon / b^{h+1} \gamma \beta m (1 + \varepsilon) q_{\text{max}}.
\]

Condition 4 of Lemma 3.8 holds, since \( N_1, N_2, \) and \( N \) are large enough, and so we have that no bucket containing \( y_p \in W_q^* \) contains an entry larger than \( \gamma / \beta b^{h+1} m^2 q_{\text{max}} \), so if \( W \) comprises \( G(L_h) \cap (Y \setminus W q_{\text{max}}) \), we have \( \|W\| \leq \gamma / \beta b^{h+1} m^2 q_{\text{max}} \). Using condition \( \mathcal{E} \), \( \|W\|_1 \leq (1 + \varepsilon) b^{-h} \), using just the condition \( \|Y\|_1 = 1 \). Therefore the given \( N \) is larger than the \( O(b n \varepsilon^{-2} q_{\text{max}}) \) needed for Lemma 3.3 to apply, with \( \delta = \exp(-\varepsilon^2 m) \). This with \( (3.9) \) yields that for each \( y_p \in W_q^* \), the remaining entries in its bucket \( L \) have \( \|L - y_p\|_1 \leq 2 \gamma^{-2} \varepsilon |y_p| \), with failure probability \( \exp(-\varepsilon^2 m) \).

For each such isolated \( y_p \) we consider the corresponding \( z_p \) (denoted by \( v \) hereafter), and let \( L(v) \) denote the set of \( z \) values in the bucket containing \( v \). We apply Lemma 3.6 to \( v \) with \( L(v) \) taking the role of \( T \), and \( 2 \gamma^{-2} \varepsilon \) taking the role of \( \varepsilon \), obtaining \( \mathbf{E}_A[G(\|L(v)\|_1)] \geq (1 - C^* \varepsilon^{1/\alpha}) G(v) \). (Here we fold a factor of \((2\gamma^2)^{1/2} \) into \( C^* \), recalling that we consider \( \gamma \) to be fixed.) Using this relation and condition \( \mathcal{E} \), we have
\[
\|W_q\| \leq \beta b^n \sum_{G(v) \in W_q^*} \mathbf{E}_A[G(\|L(v)\|_1)] / (1 - 4 \gamma \varepsilon) \cdot (1 - C^* \varepsilon^{1/\alpha}),
\]
so the claim of the lemma follows, in expectation, after adjusting constants, and conditioned on events of failure probability \( C^{-\varepsilon^2 m} \) for constant \( C \).

To show the tail estimate, we relate each \( G(\|L(v)\|_1) \) to \( G(v) \) via the first claim of Lemma 3.6, which implies \( G(\|L(v)\|_1) \geq (1 - \varepsilon^{1/\alpha}) \mathbf{E}_A[G(\|L(v)\|_1)] \) with failure probability at most \( C^{-\varepsilon^2 m} \) for a constant \( C > 1 \).

**Proof.** For all \( y_p \in W_q < v \), we have
\[
y_p \geq \frac{\varepsilon}{m}.
\]
Let
\[ \gamma' \equiv \min\{ \frac{\varepsilon^{-\alpha}}{5\varepsilon^2 m^{2+\alpha/2}}, \frac{\varepsilon^{2-2\alpha}}{m^{1+\alpha}} \}, \]
so that \( N_2 \) of the lemma statement is at least \( 2d/\gamma'^2 C_2^2 \). Then condition \( E_\varepsilon \) and Lemma 3.5 imply that every member of \( W_q \) is in a bucket with no entry other than itself larger than \( \gamma' \).

Assume for the moment that all \( h_p = 0 \), that is, all values are mapped to level 0. We apply Lemma 3.3 to \( h = 0 \), with \( \delta \equiv \exp(-\varepsilon^2 m) \) and with \( \tilde{W} = Y \setminus Y_2 \), so that
\[ \|\tilde{W}\|_\infty \leq \gamma' \leq \frac{\varepsilon^{-\alpha}}{3m^{2+\alpha/2}} = \left( \frac{1}{\varepsilon^{1-1/\alpha} \sqrt{m}} \right)^\alpha \frac{1}{3m \varepsilon^2 m}. \]
The result is that with large enough \( N = O(m^{1+\alpha/2} \varepsilon^{-2+\alpha}) \), and assuming \( \log N \leq \varepsilon^2 m \), so that \( \log(N/\delta) \leq 2\varepsilon^2 m \), we have for \( v \) with \( G(v) = y_p \in W_q \),
\[ \|G(L(v)) \cap \tilde{W}\|_1 \leq \frac{7}{6} \left( \frac{1}{\varepsilon^{1-1/\alpha} \sqrt{m}} \right)^\alpha \frac{2\varepsilon}{3m} \leq \left( \frac{1}{\varepsilon^{1-1/\alpha} \sqrt{m}} \right)^\alpha \|y_p\|, \]
that is,
\[ \frac{\|L(v) - v\|_{G(v)}}{\|G(v)\|_2} \leq \left( \frac{1}{\varepsilon^{1-1/\alpha} \sqrt{m}} \right)^\alpha, \]
so that from (3.7), we have
\[ \frac{\|L(v) - v\|_{G(v)}}{v^2} \leq \frac{\varepsilon^{2/\alpha}}{\varepsilon^2 m}. \]

Since
\[ \|\tilde{W}\|_\infty \leq \frac{\varepsilon^{2-2\alpha}}{m^{1+\alpha}} = \left( \frac{1}{m \varepsilon^{2-1/\alpha}} \right)^\alpha \frac{\varepsilon}{m}, \]
we also have, for all \( v' \in L(v) - v \), and using that \( G(v) \geq \varepsilon/m \),
\[ \|v' - v\|_{G(v)} \leq \left( \frac{G(v')}{G(v)} \right)^{1/\alpha} \leq \frac{1}{m^2 \varepsilon^{2-1/\alpha}}. \]

From (3.11), we have that the summands determining \( \|L(v) - v\|_\Lambda \) have magnitude at most \( \varepsilon/\varepsilon^{1/\alpha} \varepsilon^2 m \). From (3.10), we have \( \|L(v) - v\|^2 \) is at most \( \varepsilon^2/\varepsilon^2 m \). It follows from Bernstein’s inequality with failure probability \( \exp(-\varepsilon^2 m) \), \( \|L(v) - v\|_\Lambda \leq \varepsilon^{1/\alpha} |v| \). Applying the first claim of Lemma 3.6, we have \( G(||L(v)||_\Lambda) \geq (1 - 2e^{1/\alpha}) G(v) \), for all \( v \in G^{-1}(W_q^*) \), with failure probability \( |W_q| \exp(-\varepsilon^2 m) \). This implies the bound after adjusting constants.

We can remove the assumption that all \( h_p = 0 \), because the bound on \( \|L(v) - v\|_\Lambda \) also holds when splitting up into levels.

Combining these lemmas, we have the following contraction bound.

**Theorem 3.2.** Assume condition \( E \) of Lemma 3.1 holds, and Assumption 3.1, and condition \( E_\varepsilon \) of Lemma 3.4 holds for \( N_1 = O(C_2^2 \varepsilon^{-2} d m^2) \) and \( N_2 = O((C_2^2 d(\varepsilon^2 m^{4+\alpha} + \varepsilon^{4-4\alpha} d^2 m^2)) \), with \( N = O(N_1 N_2 + \varepsilon^{-2} m^2 b_{\max}) \). Then \( \|Sz\|_{G,w} \geq \|z\|_G (1 - \varepsilon^{1/\alpha}) \), with failure probability no more than \( C^{-\varepsilon^2 m} \), for absolute \( C > 1 \).

**Proof.** (We note that \( c, b, m \) can be chosen such that the relations among these quantities and also \( N = cbm \) satisfy Assumption 3.1, up to the weak relations among \( m, b, \) and \( n/\varepsilon \), which ultimately will require that \( n \) is not extremely large relative to \( d \).)

Recalling \( Q^* \) from (3.8), let \( Q^* \equiv \{ q \mid q \in Q^*, \|W_q\| \geq \varepsilon/q_{\max} \} \). Assuming conditions \( E \) and \( E_\varepsilon \), we have, with probability \( 1 - C^{-\varepsilon^2 m} \),
\[ \|Sz\|_{G,w} = \sum_{h,i} \beta b^h G(||L_h,v||_\Lambda) \]
\[ \geq \sum_{q \in Q^*, v \in W_q^*} \beta b^h(q)(1 - \varepsilon^{1/\alpha}) ||W_q^*||_1 \] \[ \geq \sum_{q \in Q^*} (1 - \varepsilon^{1/\alpha})(1 - 4\varepsilon) ||W_q||_1 \]
\[ \geq \sum_{q \in Q^*} \varepsilon^{-1} - q_{\max}(\varepsilon/q_{\max}) + \sum_{q \in Q^*} ||W_q||_1 \geq 1 - 6\varepsilon. \]

Using Lemma 3.7,
\[ \sum_{q \in Q^*} ||W_q||_1 \geq -q_{\max}(\varepsilon/q_{\max}) + \sum_{q \in Q^*} ||W_q||_1 \geq 1 - 6\varepsilon. \]

Adjusting constants gives the result.

**3.6 Dilation bounds.** We prove two bounds for dilation, where the first gives a dilation that is at most a log factor, and the second gives a constant factor by using a different way to estimate distance based on the sketch.

**3.6.1 Bound for \( ||Sz||_{G,w} \).** Our first bound for dilation is \( E[||Sz||_{G,w}] = O(\max) ||z||_G \), which implies a tail bound via Markov’s inequality; first, some lemmas.

**Lemma 3.11.** For \( T \subset Z \), \( E_{\Lambda}[G(||T||_\Lambda)] \leq CG(||T||_2) \), for an absolute constant \( C \).

**Proof.** Let \( \mathcal{L} \) denote the event that \( ||T||_\Lambda \geq ||T||_2 \). Here
the expectation is with respect to $\Lambda$ only:

\[
E[G(||T||_\Lambda)] = E\left[G(||T||_\Lambda) \mid \mathcal{L}\right] P(\mathcal{L}) + E\left[G(||T||_\Lambda) \mid \mathcal{L}^c\right] P(\mathcal{L}^c)
\]

\[
\leq E\left[\frac{\|T\|^2}{\|T\|^2} G(||T||_\Lambda) \mid \mathcal{L}\right] P(\mathcal{L}) + G(||T||_\Lambda)
\]

\[
= E[\|T\|^2_\Lambda \mid \mathcal{L}] P(\mathcal{L}) G(||T||_\Lambda) + G(||T||_\Lambda)
\]

\[
\leq E[\|T\|^2_\Lambda \mid \mathcal{L}] P(\mathcal{L}) G(||T||_\Lambda) + G(||T||_\Lambda)
\]

\[
\leq C G(||T||_2),
\]

for a constant $C$, where the last inequality uses Khintchine.

**Lemma 3.12.** For $T \subset Z$, $G(||T||_2) \leq ||T||_G$, and so $E[\Lambda G(||T||_\Lambda)] \leq C ||T||_G$.

**Proof.** Using the growth upper bound for $G$,

\[
\frac{||T||_G}{G(||T||_2)} = \sum_{z_p \in \mathcal{P}} \frac{G(z_p)}{G(||T||_2)}
\]

\[
\geq \sum_{z_p \in \mathcal{P}} \left| \frac{z_p}{||T||_2} \right|^\alpha
\]

\[
\geq \sum_{z_p \in \mathcal{P}} \left| \frac{z_p}{||T||_2} \right|^2
\]

\[
= 1.
\]

The last claim follows from this and the previous lemma.

**Theorem 3.3. Assuming condition $E$ of Lemma 3.1,**

\[
E_{g,t,\Lambda}[||Sz||_{G,c,w}] = O(\max h) ||z||_G.
\]

**Proof.** Note that for each level $h$, \( \sum_{i} E[\Lambda G(||L_{h,i}||_\Lambda)] \leq C ||L_h||_G \), applying the previous lemma. Since $||L_h||_G = ||G(L_h)||_1 = (1 + \varepsilon)||y||_1/\beta b$ under assumption $E$, we have

\[
E_{g,t,\Lambda}[||Sz||_{G,c,w}]
\]

\[
= \sum_{h} \beta b^h \sum_{i} E[\Lambda G(||L_{h,i}||_\Lambda)]
\]

\[
\leq \sum_{h} \beta b^h \sum_{i} C(1 + \varepsilon)||y||_1/\beta b^h
\]

\[
= \sum_{h} \sum_{i} C(1 + \varepsilon)||y||_1 = \max h C(1 + \varepsilon)||z||_G,
\]

and the theorem follows, picking bounded $\varepsilon$.

**3.6.2 Bound for a “clipped” version.** We can achieve a better dilation than $O(\max h) = O(\log(\varepsilon n/d))$ by ignoring small buckets, using a subset of the coordinates of $Sz$, as follows: for a given sketch, our new estimate $||Sz||_{G,c,w}$ of $||z||_G$ is obtained by adding in only those buckets in level $h$ that are among the top

\[
M^* = bmM_\geq + bmM_\leq = O(mb \log \gamma (b/\varepsilon))
\]

in $\Lambda$-semi-norm, recalling $M_\geq$ and $M_\leq$ defined in (3.8). That is,

\[
||Sz||_{G,c,w} = \sum_{h} \beta b^h \sum_{i \in \{M^*\}} G(||L_{j,(i)}||_\Lambda),
\]

where $L_{j,(i)}$ denotes the level $j$ bucket with the $i$'th largest $\Lambda$-semi-norm among the level $j$ buckets.

The proof of the bounded contraction of $||Sz||_{G,c,w}$, Theorem 3.2, only requires lower bounds on $G(L_{h,k})||_\Lambda$ for those at most $M^*$ buckets on level $h$ containing some member of $W^*_q$ for $q \in Q^*$, for the $W^*_q$ defined in Lemma 3.8. Thus if the estimate of $||Sz||_{G,c,w}$ uses only the largest such buckets in $\Lambda$-norm, the proven bounds on contraction continue to hold, and in particular $||Sz||_{G,c,w} \geq (1 - \varepsilon)||Sz||_{G,c,w'}$.

Moreover, the dilation of $||Sz||_{G,c,w}$ is constant:

**Theorem 3.4.** There is $c = O(\log \gamma (b/\varepsilon)(\log(\gamma(\gamma/m)))$ and $b \geq c$, recalling $N = mbc$, such that

\[
E[||Sz||_{G,c,w}] \leq C ||z||_G
\]

for a constant $C$.

**Proof.** From Lemma 3.12, the contribution of level $h$ satisfies

\[
E[\sum_{i} G(||L_{h,i}||_\Lambda)] \leq C ||L_h||_G = C ||G(L_h)||_1.
\]

We will consider the contribution of each weight class separately. The contribution of $W_q$ at $h = h(q)$ is $\beta b^h ||G(L_h) \cap W_q||_1 \leq \|W_q\|_1(1 + \varepsilon)$, if all entries of $W_q$ land among the top $M^*$ buckets; otherwise the contribution will be smaller.

The expected contribution of $W_q$ at $h = h(q) - k$ for $k > 0$ is at most $M^*|L_{h,i} \cap W_q|\gamma^{1-k}$, where $L_{h,i} \cap$ contains the largest number of members of $W_q$ among the buckets on level $h$. When $|G(L_{h,i}) \cap W_q| \geq N \log N$, $|G(L_{h,i}) \cap W_q| \leq 4|G(L_{h,i}) \cap W_q|/N$, with failure probability at most $1/N$. (This follows by applying Bernstein’s inequality to the sum of random variables $X_k$, where $X_k = 1$ when the $i$'th element of $W_q$ falls in a given bucket, and $X_k = 0$ otherwise, followed by a union bound over the buckets.) At level $h = h(q) - k$, $|G(L_{h,i}) \cap W_q| \geq \beta b^h (1 - \varepsilon)$, using assumption $E$ of Lemma 3.1, so to obtain $b^h (1 - \varepsilon) \geq N \log N$ it suffices that $k \geq 2 + 2 \log b \geq \log(N \log(N)/m(1 - \varepsilon))$, using
\( N = bc m \), obtaining for those \( k \) a contribution for \( W_q \) is within a constant factor of
\[
\beta b^M (4^{b^{-1}} b^{-h}|W_q|/N)^{1-q} \leq \frac{O(\log_2_b (b/\varepsilon))}{c} \| W_q \|_1,
\]
using the bound on \( M^* \) given above. Adding this contribution to that for \( k \leq 2 + 2 \log_2_b c \), we obtain an overall bound for \( W_q \) and \( h < h(q) \) that is within a constant factor of \( (1 + \log_2_b c + h_{\max}(M^*_q)) \| W_q \|_1 \), and therefore within a constant factor of \( \| W_q \|_1 \) under the given conditions on \( b \) and \( c \).

For \( h = h(q) \) for \( k \) non-empty. Thus for the \( q_{\max} \) non-negligible sets \( W_q \), by a union bound the event \( \mathcal{E}_q \) holds with failure probability \( \delta \), that all \( W_q \cap L_{h(q)+k} \) will be empty for large enough \( k = O(\log_2_b q_{\max} m/\delta) \). For each \( q \) and \( k \), condition \( \mathcal{E}_q \) implies that the contribution \( \beta b^M \sum_k |G(L_{h+1})| \| A \|_\Lambda \leq (1 + \varepsilon) \| W_q \|_1 \), and so the total contribution is \( C \| W_q \|_1 \log_2_b q_{\max} m/\delta \), within a constant factor of \( \| W_q \|_1 \) under given conditions.

Note that if \( G \) is convex, then so is \( \| S \varepsilon \|_{G,c,w} \), since at each level we are applying a Ky Fan norm, discussed below; also, if \( G^{-1}(\| \cdot \|_G) \) is scale-invariant, then so is \( G^{-1}(\| \cdot \|_{G,c,w}) \). If both conditions hold, then \( G^{-1}(\| \cdot ||_G) \) is a norm, and so is \( G^{-1}(\| \cdot ||_{G,c,w}) \).

The Ky Fan \( k \)-norm of a vector \( y \in \mathbb{R}^n \) is
\[
\sum_{i \in [n]} |y(i)|^k, \text{ where } y(i) \text{ denotes the } i^{th} \text{ largest entry of } y \text{ in magnitude. Thus the Ky Fan 1-norm of } y \text{ is } \| y \|_\infty, \text{ and the Ky Fan } n \text{-norm of } y \text{ is } \| y \|_1. \text{ The matrix version of the norm arises by application to the vector of singular values.}

A disadvantage of this approach is that some smoothness is sacrificed: \( \| z \|_{G,c,w} \) is not a smooth function, even if \( G \) is; while this does not affect the fact that the minimization problem in the sketch space is polynomial time, it could affect the concrete polynomial time complexity, which we leave a subject for future work.

4 Net Argument.

We prove a general \( \varepsilon \)-net argument for M-estimators satisfying our growth condition (1.1).

We need a few lemmas to develop the net argument.

**Lemma 4.1.** (Bounded Derivative) There is a constant \( C > 0 \) for which for any \( a, b \) with \( |b| \leq \varepsilon |a| \), \( G(a + b) = (1 + C|\varepsilon|) G(a) \).

**Proof.** First suppose that \( a \) and \( b \) have the same sign. Then by mononicity, \( G(a) \leq G(a + b) \). Moreover, by the growth condition,
\[
\frac{G(a + b) - G(a)}{G(a)} \leq \left| \frac{a + b}{a} \right|^2 \leq (1 + \varepsilon)^2 \leq 1 + 3\varepsilon,
\]
and so \( G(a + b) \leq (1 + 3\varepsilon) G(a) \).

Now suppose \( a \) and \( b \) have the opposite sign. Then \( G(a + b) \leq G(a) \) by mononicity, and again by the growth condition,
\[
\frac{G(a)}{G(a + b)} \leq \left| \frac{a}{a + b} \right|^2 \leq (1 + 2\varepsilon)^2 \leq 1 + 5\varepsilon,
\]
and so \( G(a + b) \geq G(a)/(1 + 5\varepsilon) \), and so \( G(a + b) \geq (1 - 5\varepsilon) G(a) \).

**Lemma 4.2.** (Approximate Scale Invariance) For all \( a \) and \( C \geq 1 \), \( G(Ca) \leq C^2 G(a) \).

**Proof.** By the growth condition, \( G(Ca)/G(a) \leq C^2 \).

**Lemma 4.3.** (Perturbation of the weighted M-Estimator) There is a constant \( C' > 0 \) for which for any \( e \) and \( w \), with \( \| e \|_{G,c,w} \leq \varepsilon \| y \|_{G,c,w} \),
\[
\| y + e \|_{G,c,w} = (1 + C'\varepsilon) \| y \|_{G,c,w}.
\]

**Proof.** By Lemma 4.2, \( G(e_i) \leq \frac{C'}{C} G(e_i) \), and so \( \| e \|_{G,c,w} \leq \| y \|_{G,c,w} \), where the final inequality follows by the assumption of the lemma.

Now let \( S \subseteq [n] \) denote those columns \( i \) for which \( |e_i| \leq \varepsilon |y_i| \). By Lemma 4.1, \( G(y_i + e_i) = (1 + C\varepsilon) G(y_i) \).

Now consider an \( i \in [n] \setminus S \). In this case \( |y_i| \leq \varepsilon |e_i| \). Using that \( G \) is monotonically non-decreasing and applying Lemma 4.1 again,
\[
G(e_i + y_i) \leq G(\frac{1}{\varepsilon} e_i + y_i) = (1 + C\varepsilon) G(e_i/\varepsilon^2),
\]
so that
\[
\sum_{i \in [n] \setminus S} w_i G(y_i + e_i) \leq (1 + C\varepsilon) \| e \|_{G,c,w} \leq (1 + C\varepsilon) \| y \|_{G,c,w}.
\]

Again using that \( G \) is monotonically non-decreasing, we note that
\[
\sum_{i \in [n] \setminus S} w_i G(y_i) \leq \sum_{i \in [n] \setminus S} w_i G(e_i/\varepsilon) \leq \| e \|_{G,c,w} \leq \varepsilon \| y \|_{G,c,w}.
\]

Hence,
\[
\| y + e \|_{G,c,w} \leq \left( 1 + C\varepsilon \right) \sum_{i \in S} w_i G(y_i + e_i) + \sum_{i \in [n] \setminus S} w_i G(y_i + e_i) \leq (1 + O(\varepsilon)) \sum_{i \in S} w_i G(y_i) \leq \| y \|_{G,c,w}.
\]
This completes the proof.
LEMMA 4.4. (Relation of weighted M-Estimator to 2-Norm) Suppose \( w_i \geq 1 \) for all \( i \). Given an \( n \times d \) matrix \( A \), an \( n \times 1 \) column vector \( b \), let \( c = \min x \|Ax - b\|_2 \) (note the norm is the 2-norm here). Let \( y^* = Ax^* - b \), where \( x^* = \arg\min_{x} \|Ax - b\|_{G,w} \). Then \( c \leq \|y^*\|_2 \leq \kappa cn^{3/2}\|w\|_\infty \), where \( \kappa > 0 \) is a sufficiently large constant.

Proof. If \( c = 0 \), then there exists an \( x \) for which \( Ax = b \). In this case, since \( G(0) = 0 \), it follows that \( \|y^*\|_2 = 0 \). Now suppose \( c > 0 \) and let \( y \) be a vector of the form \( Ax - b \) of minimal 2-norm. Since \( \|y\|_2 = c \), each coordinate of \( y \) is at most \( c \). Hence, \( \|y\|_{G,w} \leq \|w\|_\infty \cdot G(c) \) by monotonicity of \( G \).

Now consider the 2-norm of \( y^* \), and let \( d = \|y^*\|_2 \).

By definition, \( d \geq c \). Moreover, there exists a coordinate of \( y^* \) of absolute value at least \( d/\sqrt{n} \). Hence, by monotonicity of \( G \) and using that \( w_i \geq 1 \) for all \( i \), \( \|y^*\|_2 \geq G(d/\sqrt{n}) \). Since \( y^* \) is the minimizer for \( G \) with weight vector \( w \), necessarily \( G(d/\sqrt{n}) \leq \|w\|_\infty \cdot G(c) \).

By lower bound on the growth condition for \( G \), \( G(d/\sqrt{n}) \geq \kappa c d/ \sqrt{n} \), and so \( c d/ \sqrt{n} \leq \|w\|_\infty \cdot n \cdot G(c) \).

Hence, \( d \geq \|w\|_\infty n^{3/2} c/G(c) \).

LEMMA 4.5. (Net for weighted M-Estimators) Let \( c = \min x \|Ax - b\|_2 \). For any constant \( C_S > 0 \) there is a constant \( C_N > 0 \) and a set \( N \subseteq \{Ax - b \mid x \in \mathbb{R}^d \} \) with \( |N| \leq (n/\varepsilon)^{C_N} \) for which both:

1. \( \|S(Ax - b)\|_{G,w} \leq (1 + \varepsilon)\|Ax - b\|_{G,w} \) holds for all \( Ax - b \in N \) and \( S \) is a matrix for which \( \|S(Ax - b)\|_{G,w} \leq n^{C_S}\|Ax - b\|_{G,w} \) for all \( x \) for an appropriate \( w' \),
2. \( \|w\|_\infty \leq n^{C_S} \) and \( w_i \geq 1 \) for all \( i \),

then for all \( x \) for which \( \|Ax - b\|_2 \leq \kappa cn^{3/2}\|w\|_\infty \), for an arbitrary constant \( \kappa > 0 \), it holds that

\[
\|S(Ax - b)\|_{G,w} \geq (1 - \varepsilon)\|Ax - b\|_{G,w}.
\]

Moreover, if the first condition is relaxed to state only that \( (1 - \varepsilon)\|Ax - b\|_{G,w} \leq \|S(Ax - b)\|_{G,w} \) for all \( Ax - b \in N \) and \( S \) is a matrix for which \( \|S(Ax - b)\|_{G,w} \leq n^{C_S}\|Ax - b\|_{G,w} \) for all \( x \) for an appropriate \( w' \), then the following conclusion holds: for all \( x \) for which \( \|Ax - b\|_2 \leq \kappa cn^{3/2}\|w\|_\infty \), for an arbitrary constant \( \kappa > 0 \), it holds that

\[
(1 - \varepsilon)\|Ax - b\|_{G,w} \leq \|S(Ax - b)\|_{G,w}.
\]

Proof. Let \( L \) be the subspace of \( \mathbb{R}^n \) of dimension at most \( d + 1 \) spanned by the columns of \( A \) together with \( b \). Let \( N_\alpha \) be a finite subset of \( \{z \mid z \in L \text{ and } \|z\|_2 = \alpha \} \) for which for any point \( y \) with \( \|y\|_2 = \alpha \), there exists a point \( e \in N_\alpha \) for which \( \|y - e\|_2 \leq \frac{\varepsilon^2}{n\|w\|_\infty} \). It is well-known that there exists an \( N_\alpha \) for which \( |N_\alpha| \leq \left( \frac{3n^{2C_S+2}}{\varepsilon^5} \right)^{d+1} \).

We define

\[
N = N_c \cup N_c(1+\varepsilon) \cup N_c(1+\varepsilon)^2 \cup \cdots \cup N_{kcn^{3/2}\|w\|_\infty}.
\]

Then

\[
|N| = O(\log(1+\varepsilon)\kappa n^{3/2}\|w\|_\infty) \left( \frac{3n^{2C_S+2}}{\varepsilon^5} \right)^{d+1} \leq \left( \frac{n}{\varepsilon} \right)^{C_N d}.
\]

where \( C_N > 0 \) is a large enough constant.

Now consider any \( x \in \mathbb{R}^d \) for which \( y = Ax - b \) satisfies \( \|y\|_2 \leq \kappa cn^{3/2}\|w\|_\infty \). By construction of \( N \), there exists an \( e \in N \) for which \( \|e - y\|_2 = O(\varepsilon^5/n^{2C_S+2})\|y\|_2 \). Then,

\[
\|S(e - y)\|_{G,w} \leq n^{C_S} \|e - y\|_{G,w} \leq n^{C_S} \cdot \|w\|_\infty \cdot nG(\|e - y\|_2),
\]

using the fact that each coordinate of \( e \) is at most \( \|e - y\|_2 \) in magnitude and that \( G \) is monotonically non-decreasing. By the lower bound on the growth condition on \( G \), \( G(\|e - y\|_2) \leq \left( \frac{\varepsilon^5}{n} \right) G(\|y\|_2) \).

Note that \( \|y\|_{G,w} \geq G(\|y\|_2/\sqrt{n}) \) by monotonicity and using that \( w_i \geq 1 \) for all \( i \). Furthermore, by the growth condition on \( G \), \( G(\|y\|_2/\sqrt{n}) \leq nG(\|y\|_2/\sqrt{n}) \). Combining these inequalities, we have

\[
\|S(e - y)\|_{G,w} \leq n^{2C_S+1} G(\|e - y\|_2) = O\left( \frac{\varepsilon^5}{n} \right) G(\|y\|_2) = O(\varepsilon^5) \|y\|_{G,w}.
\]

Note that the argument thus far was true for any \( S \) and \( w' \) for which \( \|S(Ax - b)\|_{G,w} \leq n^{C_S}\|Ax - b\|_{G,w} \) for all \( x \), and so in particular holds for \( S \) being the identity and \( w'_i = 1 \) for all \( i \in [n] \). So in particular we have \( \|e - y\|_{G,w} = O(\varepsilon^5) \|y\|_{G,w} \). Applying Lemma 4.3, it follows that

\[
\|y\|_{G,w} = \|e + (y - e)\|_{G,w} = (1 + O(\varepsilon)) \|e\|_{G,w}.
\]

Now we use the assumption of the theorem that for all \( e \in N \) with a particular choice of \( S \) and \( w' \) one has \( (1 - \varepsilon)\|e\|_{G,w} \leq \|Se\|_{G,w} \leq (1 + \varepsilon)\|e\|_{G,w} \). Then \( \|Sy\|_{G,w} = \|Se + S(y - e)\|_{G,w} \). Now, \( \|Se\|_{G,w} = (1 + \varepsilon)\|e\|_{G,w} \) by the assumption of the theorem, whereas \( \|S(y - e)\|_{G,w} = O(\varepsilon^5) \|y\|_{G,w} = O(\varepsilon^5) \|e\|_{G,w} \) by combining (4.13) and (4.14). So we can apply Lemma 4.3
to conclude that \( \|Sy\|_{G,w} = (1 + O(\varepsilon))\|Se\|_{G,w'} \), and combining this with the assumption of the theorem and (4.14),
\[
\|Sy\|_{G,w'} = (1 + O(\varepsilon))\|Se\|_{G,w'} = (1 + O(\varepsilon))\|e\|_{G,w'} = (1 + O(\varepsilon))\|y\|_{G,w}.
\]
For the second part of the lemma, suppose we only had that for all \( e \in N \), \((1 - \varepsilon)\|e\|_{G,w} \leq \|Se\|_{G,w'}\). We still have \( \|S(y - e)\|_{G,w'} = O(\varepsilon^5)\|e\|_{G,w'} \), and so we can still apply Lemma 4.3 to conclude that
\[
\|S(y - e)\|_{G,w'} = (1 + O(\varepsilon))\|Se\|_{G,w'}.
\]
Using (4.14), we have
\[
\|Sy\|_{G,w'} = (1 + O(\varepsilon))\|Se\|_{G,w'} \geq (1 - O(\varepsilon))\|y\|_{G,w'},
\]
which completes the proof.

### 4.1 Proof of Theorem 3.1

Using Lemma 4.5, and previous results on contraction and dilation, we can now prove Theorem 3.1.

**Proof.** The first algorithm: compute \( SA \) and \( Sb \), for \( S \) an \( M \)-sketch matrix with large enough \( N = O(C_G^{-2}d^2m^6n) \), putting \( m = O(d \log n) \), and \( \varepsilon = 1/2 \). This \( N \) is large enough for Theorem 3.2 to apply, obtaining a contraction bound with failure probability \( C_1^{-m} \).

To apply Lemma 4.5, we need to ensure the assumptions of the lemma are satisfied. For the second condition, note that indeed \( \|w\|_\infty \leq n^{C_s} \) for a constant \( C_s > 0 \) by definition of the sketch, since \( h_{max} \leq \log n \). For the first condition, because the second condition holds, it now suffices to bound \( \|Sy\|_{G,w} \) for an arbitrary vector \( y \). For this it suffices to show that for each level \( h \) and bucket \( i \), \( G(\|L_{h,i}\|) \leq n^{O(1)} \sum_{p \in L_{h,i}} G(p) \). By monotonicity of \( G \), we have \( G(\|L_{h,i}\|) \leq G(\|L_{h,i}\|_1) \). By the growth condition on \( G \), for \( a \geq b \) we have
\[
\frac{G(a + b)}{G(a)} \leq \frac{(a + b)^2}{a^2} \leq 2 + \frac{2b^2}{a^2} \leq 2 + \frac{2G(b)}{G(a)},
\]
and so \( G(a + b) \leq 2G(a) + 2G(b) \). Applying this inequality recursively \( \log \|L_{h,i}\| \) times, we have \( G(\|L_{h,i}\|_1) \leq n \sum_{p \in L_{h,i}} |y_p| \), which is what we needed to show (where with some abuse of notation, we use the definition \( y_p = G(z_p) \) given in §3.1).

Hence, we can apply Lemma 4.5, and by Theorem 3.2, the needed contraction bound holds for all members of the net \( N \) of Lemma 4.5, with failure probability \( O(n)^{C_s d C_1^{-m}} < 1 \), for \( m = O(d \log n) \), assuming conditions \( \mathcal{E} \) and \( \mathcal{E}' \).

For \( x_G \) minimizing \( \|Ax - b\|_G \), apply Theorem 3.3 to \( x_G \) and \( S \), so that with constant probability, \( \|S(Ax_G - b)\|_{G,w} = O(\log d \log n)\|Ax_G - b\|_G = O(\log d \log n)\|FT_G. \)

By making the totals of the failure probabilities for conditions \( \mathcal{E} \) and \( \mathcal{E}' \), for the contraction bound, and the dilation bound less than one, the overall failure probability is less than one. (Here we note that all such failure probabilities can be made less than 1/5, even if described as fixed.)

Let \( T \) be the sparse subspace embedding of [9], so that with probability \( 1 - o(1) \), \( \|TAx\|_2 = O(1)\|Ax\|_2 \) for all \( x \) and \( TA \) can be computed in \( nnz(A) \) time and \( T \) has \( \text{poly}(d) \) rows.

Find \( x_0 \) minimizing \( \|T(Ax - b)\|_2 \), and let \( c \equiv \|Ax_0 - b\|_2 \).

Now find \( \tilde{x} \) minimizing \( \|S(Ax - b)\|_{G,w} \), subject to \( \|T(Ax - b)\|_2 \leq \kappa c^{3/2} \), using the ellipsoid method, in \( \text{poly}(d \log n) \) time. Now Lemma 4.5 applies, implying that \( \tilde{x} \) satisfies the claim of the theorem.

A similar argument holds for \( \tilde{x} \), by minimizing \( \|S(Ax - b)\|_{G_c,w} \).

**Acknowledgements.** We acknowledge the support of the XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323. We thank the referees for their helpful comments.

### References


[9] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input spar-