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 $G : \mathbb{R} \mapsto \mathbb{R}^{\geq 0}$, with $\|y\|_G \equiv \sum_i G(y_i)$. The M-estimators generalize ℓ_p regression, for which $G(x) = |x|^p$. We first show that the *Huber* measure can be computed up to relative error ϵ in $O(\mathtt{nnz}(A)\log n + \mathrm{poly}(d(\log n)/\varepsilon))$ time, where $\mathtt{nnz}(A)$ denotes the number of non-zero entries of the matrix A. Huber is arguably the most widely used M-estimator, enjoying the robustness properties of ℓ_1 as well as the smoothness properties of ℓ_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 $O(\mathtt{nnz}(A) + \mathrm{poly}(d\log n))$ 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 descendent 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 ℓ_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 $\Theta(nmd)$ 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 $\mathtt{nnz}(A)$ 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 $O(\mathtt{nnz}(A))$ or $O(\mathtt{nnz}(A)\log n)$. This prominently includes regression problems on "tall and thin" matrices with $n \gg d$, both in the least-squares (ℓ_2) and robust (ℓ_1) cases. There are also recent recursive sampling-based algorithms for ℓ_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 $O(\mathtt{nnz}(A))$ time results were obtained for quantile regresion [42], by relating it to ℓ_1 regression. A natural question raised by these works is which families of penalty functions can be computed in $O(\mathtt{nnz}(A))$ or $O(\mathtt{nnz}(A)\log n)$ time.

M-estimators. Here we further extend the "$\mathtt{nnz}$" regime to general statistical M-*estimators*, specified by a measure function $G : \mathbb{R} \mapsto \mathbb{R}^{\geq 0}$, 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 ℓ_1 regression with the low variance of ℓ_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

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

combining an ℓ_2-like measure for small x with an ℓ_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 ℓ_1, while enjoying the same robustness properties with respect to outliers. Moreover, while some measures, such as ℓ_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 ℓ_1 and ℓ_2 properties. We obtain an algorithm yielding an ε-approximation with respect to the Huber norm of the residual; as stated in Theorem 2.1, the algorithm needs $O(\mathtt{nnz}(A)\log n) + \mathrm{poly}(d/\varepsilon)$ time (see, e.g., [31] for convex programming algorithms for solving Huber in $\mathrm{poly}(d/\varepsilon)$ time when the dimension is $\mathrm{poly}(d/\varepsilon)$).

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*[1] 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 $O(\mathtt{nnz}(A))$ time, and needs $O(\mathrm{poly}(d \log n))$ space; we show that it can be used in $O(\mathrm{poly}(d \log n))$ 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 $\alpha \in [1, 2]$ and $C_G > 0$ so that for all a, a' with $|a| \geq |a'| > 0$,

$$\left|\frac{a}{a'}\right|^\alpha \geq \frac{G(a)}{G(a')} \geq C_G \left|\frac{a}{a'}\right| \qquad (1.1)$$

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 $O(\mathtt{nnz}(A) + \mathrm{poly}(d \log n))$ 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 ℓ_p-norms (which are a special case of M-estimators). E.g., for the $L_1 - L_2$ estimator in which $G(x) = 2(\sqrt{1 + x^2/2} - 1)$, the Fair estimator in which $G(x) = c^2 \left[\frac{|x|}{c} - \log(1 + \frac{|x|}{c})\right]$, or the Huber estimator,

[1]Verbin and Zhang call the construction a *Rademacher sketch*; with apologies, we prefer our name, for this application.

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 $(A_{i:}, b_i)$, where a sampling matrix S' is obtained such that $\|S'(Ax-b)\|_{H,w}$ is a useful approximation to $\|Ax-b\|_H$. The sampling probabilities are based on a combination of the ℓ_1 leverage score vector $u \in \mathbb{R}^n$, and the ℓ_2 leverage score vector $u' \in \mathbb{R}^n$. The ℓ_1 vector u can be used to obtain good sampling probabilities for ℓ_1 regression, and similarly for u' and ℓ_2. Since the Huber measure has a mixed ℓ_1/ℓ_2 character, we are able to use a combination of ℓ_1 and ℓ_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, τ, and its ℓ_1 and ℓ_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 ℓ_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 we 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 $S_0, S_1, \ldots S_{h_{\max}}$, for a parameter $h_{\max}$, each comprising a block of rows of our sketching matrix:

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

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 ± 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 ℓ_2, implying here that the matrix S_0, which applies to all the coordinates of $y = Ax$, has the property that $\|S_0 Ax\|_2$ is a good estimator for $\|Ax\|_2$ for all $x \in \mathbb{R}^d$; in particular, each coordinate of $S_0 y$ is the magnitude of the ℓ_2 norm of the coordinates in the contributing group.

Why should our construction, based on ℓ_2 embeddings, be suitable for, e.g., ℓ_1, with $\|D(w)SAx\|_1$ 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 $S_{h_{\max}}$ for the smallest random subset $L_{h_{\max}}$ of $y = Ax$ to be sketched; we can think of $S_{h_{\max}} y$ as one coordinate of $y = Ax$, chosen uniformly at random and sign-flipped. The expectation of $\|S_{h_{\max}} y\|_1$ 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 $\|S_0 y\|_2$ is a good estimate. On the other hand, when y is the vector with all coordinates $1/n$, the variance of $\|S_{\log_b n} y\|_1$ is zero, while $\|S_0 y\|_2 \approx \|y\|_2$ 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, $\|SAx\|_{G,w}$ is not too much smaller than $\|Ax\|_G$;
- dilation, so with constant probability, $\|SAx\|_{G,w}$ is not too much bigger than $\|Ax\|_G$.

This asymmetry in probabilities means that some results are out of reach, but still allows approximation algorithms for $\min_x \|Ax-b\|_G$. (We blur the distinction between applying S to A for vectors $x \in \mathbb{R}^d$, and to $[A\ b]$ for vectors $[x\ -1]$.) If the optimum x_{OPT} for the original problem has $\|S(Ax_{\mathrm{OPT}} - b)\|_G$ 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 ε-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*") $L_{\hat{h}}$ for which $|W \cap L_{\hat{h}}|$ is small relative to the number of rows of $S_{\hat{h}}$ (each row of $S_{\hat{h}}$ 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 $\|Sy\|_{G,w}$, and therefore that $\|Sy\|_{G,w}$ will not be too small. Also, the level $L_{\hat{h}}$ is chosen such that the expected number of entries of the weight class is large enough that the random variable $|W \cap L_{\hat{h}}|$ 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 ℓ_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 $\|Sy\|_{G,w}$ will not be too big, we show that W will not contribute too much to levels other than the "Goldilocks" level $L_{\hat{h}}$: for $h < \hat{h}$, for which $|L_h \cap W|$ is expected to be large, the fact that members of $W \cap L_h$ 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 ℓ_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 $\|Sy\|_{G,w}$ 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 $\|S_h y\|$.

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 ϵ-Approximation for the Huber Measure.

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

$$H(a) \equiv \begin{cases} a^2/2\tau & \text{if } |a| \le \tau \\ |a| - \tau/2 & \text{otherwise.} \end{cases}$$

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 $O(\mathtt{nnz}(A)\log n) + \mathrm{poly}(d/\epsilon)$ time, given an $n \times d$ matrix A with $\mathtt{nnz}(A)$ 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 $\|Ax' - b\|_H \le (1+\varepsilon)\min_{x \in \mathbb{R}^d} \|Ax - b\|_H$.*

We will need to relate the Huber norm to the ℓ_1 and ℓ_2 norms. The following lemma is shown via a case analysis of the coordinates of the vector z.

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

$$\Theta(n^{-1/2})\min\{\|z\|_1, \|z\|_2^2/2\tau\} \le \|z\|_H \le \|z\|_1.$$

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

$$G(a) \equiv \begin{cases} a^2/2\tau & \text{if } |a| \le \tau \\ |a| & \text{otherwise.} \end{cases}$$

Then $\|z\|_H \le \|z\|_G \le 2\|z\|_H$, 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

$$|z_1| \le |z_2| \le \ldots \le |z_s| \le \tau \le |z_{s+1}| \le \ldots \le |z_n|.$$

(We may have $s = 0$, when all $|z_i| \ge \tau$, or $s = n$, when all $|z_i| \le \tau$.) 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 $\sqrt{\frac{L}{s}}$, one coordinate equal to U, and remaining coordinates equal to 0. Then,

$$\|w\|_G = s \cdot \frac{L}{s2\tau} + U = \frac{L}{2\tau} + U = \|z\|_G. \tag{2.2}$$

Moreover,

$$\|w\|_1 = U + s \cdot \frac{\sqrt{L}}{\sqrt{s}} = U + \sqrt{sL} \ge \|z\|_1, \tag{2.3}$$

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

$$\frac{\|w\|_2^2}{2\tau} = \frac{L}{2\tau} + \frac{U^2}{2\tau} \ge \frac{\|z\|_2^2}{2\tau}, \tag{2.4}$$

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

Combining (2.2), (2.3), and (2.4), in order to show $\|z\|_G = \Omega(n^{-1/2}) \min(\|z\|_1, \|z\|_2^2/2\tau)$ it suffices to show $\|w\|_G = \Omega(n^{-1/2}) \min(\|w\|_1, \|w\|_2^2/2\tau)$. By the above, this is equivalent to showing

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

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

$$U + \frac{L}{2\tau} = \Omega(n^{-1/2}) \cdot \min\left(U + \sqrt{nL},\ \frac{U^2}{2\tau} + \frac{L}{2\tau}\right). \tag{2.5}$$

Note that we can assume $U \ne 0$, 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 $U \ge \tau$. We break the analysis into cases:

Case: $\frac{U^2}{2\tau} + \frac{L}{2\tau} \le \frac{1}{4}(U + \sqrt{nL})$. 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 $\frac{L}{2\tau} \ge U$. 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 $U \ge \tau$, we have $\frac{U}{4} + \frac{U^2}{4\tau} + \frac{L}{2\tau} \le \frac{U}{4} + \frac{\sqrt{nL}}{4}$, which implies that $\frac{U^2}{\tau} \le \sqrt{nL}$. So we just need to show that $\frac{L}{2\tau} = \Omega(n^{-1/2})\frac{\sqrt{nL}}{2}$, or equivalently, $\sqrt{L} = \Omega(\tau)$. Since $\frac{L}{2\tau} \ge U \ge \tau$, we have $L = \Omega(\tau^2)$, as desired.

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

Case: $\frac{1}{4}(U+\sqrt{nL}) < \frac{U^2}{2\tau} + \frac{L}{2\tau}$. 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 $U \ge \frac{L}{2\tau}$, and so we need to show $U = \Omega(n^{-1/2})(U + \sqrt{nL})$, which is equivalent to showing $U = \Omega(\sqrt{L})$. Since $\frac{L}{2\tau} \le U$, we have $\sqrt{L} = O(\sqrt{U\tau}) = O(U)$, using that $U \ge \tau$. This completes this case.

Otherwise, we have $\frac{L}{2\tau} \ge U$ and need to show $\frac{L}{2\tau} = \Omega(n^{-1/2})(U + \sqrt{nL})$. We can assume $\sqrt{nL} \ge U$, otherwise this is immediate using $\frac{L}{2\tau} \ge U$, 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 $U + \sqrt{nL} = \Theta(\sqrt{nL})$ realizes the minimum, and so $\sqrt{nL} = O\left(\frac{U^2}{2\tau} + \frac{L}{2\tau}\right)$, and using that $U \le \frac{L}{2\tau}$, this implies $\sqrt{nL} = O\left(\frac{L}{2\tau} \cdot \frac{U}{2\tau} + \frac{L}{2\tau}\right)$. Since $U \ge \tau$, it follows that $\sqrt{nL} = O\left(\frac{LU}{\tau^2}\right)$. Now using that $U \le \sqrt{nL}$, 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 $w_i \ne 0$, then $w_i \ge 1$.

In our recursion we will have $\|w\|_\infty \le \mathrm{poly}(n)$ for some polynomial that depends on where we are in the recursion. These conditions imply that we can partition the positive coordinates of w into $O(\log n)$ groups P^j, for which $P^j = \{i \mid 2^{j-1} \le w_i < 2^j\}$.

Let A^j denote the restriction of the input matrix A to those rows i in the set P^j. For each j, let U^j be an (α, β)-well-conditioned basis for A^j with respect to ℓ_1, meaning U^j has the same column span as A^j, $\sum_{i \in P^j} |U_i^j|_1 = \alpha$, and for all x, $\|x\|_\infty \le \beta \|U^j x\|_1$ [10]. Here U_i^j is the i-th row of U^j. Let V^j be an approximately orthonormal basis for the column span of A^j, that is, $\sum_{i \in P^j} \|V_i^j\|_2^2 = O(d)$ and for all x, $\|V^j x\|_2 = (1 \pm 1/2)\|x\|_2$. Here V_i^j is the i-th row of V^j. Let $A^j = U^j J^j$ and $A^j = V^j K^j$, where J^j and K^j are $d \times d$ matrices.

For each j and $i \in P^j$, let $q_i^j = \frac{\|U_i^j\|_1}{\alpha}$ and let $r_i^j = \frac{\|V_i^j\|_2^2}{\sum_{i' \in P^j} \|V_{i'}^j\|_2^2}$. For $i \notin P^j$, let $q_i^j = 0$ and $r_i^j = 0$.

Set $s = C_0 \cdot n^{1/2} \max(\alpha \cdot \beta, d) \cdot d\varepsilon^{-2} \log(n/\varepsilon)$ for a sufficiently large constant $C_0 > 0$. Suppose we independently sample each row i of A with probability $p_i = \min(1, \Theta(s \cdot \sum_j (q_i^j + r_i^j)))$ (the fact that we choose $\Theta(s \cdot \sum_j (q_i^j + r_i^j))$ instead of $s \cdot \sum_j (q_i^j + r_i^j)$ in the definition of p_i will give us some flexibility in designing a fast algorithm, as we will see).

For $i \in [n]$, let $w'_i = 0$ if we do not sample row i, and otherwise $w'_i = w_i/p_i$. The expected number of non-zero elements of w' is $O(s \log n)$. This is because for each of the $O(\log n)$ possibilities of j, $\sum_i q_i^j + r_i^j = O(1)$. Note that if $w'_i \ne 0$, then $w'_i \ge 1$. Moreover, by a union bound over the n coordinates, with probability $1 - 1/n^C$ we have $\|w'\|_\infty \le n^{C+1}\|w\|_\infty$, since the probability that any i for which $p_i \le 1/n^{C+1}$ is sampled is at most $1/n^C$.

Theorem 2.2. *(Huber Embedding) With the notation defined above, for any fixed $x \in \mathbb{R}^d$,*

$$\begin{aligned}\Pr[(1-\varepsilon)\|Ax\|_{H,w} \le \|Ax\|_{H,w'} \le (1+\varepsilon)\|Ax\|_{H,w}] \\ \ge 1 - \exp(-C_2 d \log(n/\varepsilon)),\end{aligned}$$

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 $\mathbf{E}[X] = \sum_{i=1}^n p_i (w_i/p_i) H(A_i x) = \|Ax\|_{H,w}$.

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 $\mathbf{E}[X_i^2] < \infty$ and $X_i \ge 0$. Set $X = \sum_i X_i$ and let $\gamma > 0$. Then,*

$$\Pr[X \le E[X] - \gamma] \le \exp\left(\frac{-\gamma^2}{2\sum_i \mathbf{E}[X_i^2]}\right).$$

If $X_i - \mathbf{E}[X_i] \le \Delta$ for all i, then with $\sigma_i^2 = \mathbf{E}[X_i^2] - \mathbf{E}[X_i]^2$ we have

$$\Pr[X \ge \mathbf{E}[X] + \gamma] \le \exp\left(\frac{-\gamma^2}{2\sum_i \sigma_i^2 + 2\gamma\Delta/3}\right).$$

If for some i we have $p_i = 1$, then $\mathbf{E}[X_i] = X_i = w_i H(A_i x)$. It follows that such X_i do not contribute to the deviation of X from $\mathbf{E}[X_i]$, 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 $i \in P^j$. We use Lemma 2.1 to do the case analysis.

Case $|A_i x| \ge \tau$ and $\|A^j x\|_H = \Omega(n^{-1/2}\|A^j x\|_1)$. It follows that

$$\begin{aligned}\frac{H(A_i x)}{p_i} &= \frac{|A_i x| - \tau/2}{p_i} \le \frac{|A_i x|}{p_i} \le \frac{|A_i x|}{\Theta(s) q_i^j} = \frac{|A_i x|\alpha}{\Theta(s)\|U_i^j\|_1} \\ &\le \frac{\|U_i^j\|_1 \|J^j x\|_\infty \alpha}{\Theta(s)\|U_i^j\|_1} \le \frac{\alpha\beta\|A^j x\|_1}{\Theta(s)} \\ &\le \frac{\alpha\beta O(n^{1/2})\|A^j x\|_H}{s} = \frac{O(\|A^j x\|_H)}{C_0 \varepsilon^{-2} d \log(n/\varepsilon)}.\end{aligned}$$

Case $|A_i x| \ge \tau$ and $\|A^j x\|_H = \Omega(n^{-1/2}\|A^j x\|_2^2/(2\tau))$. We claim in this case that $\|A^j x\|_H = \Omega(n^{-1/2}\|A^j x\|_1)$ as well. Suppose not, so that $\frac{\|A^j x\|_1}{\|A^j x\|_H} = \omega(n^{1/2})$.

Let $S \subseteq [n]$ be the set of $\ell \in [n]$ for which $|(A^j x)_\ell| \ge \tau$. Then $\|(A^j x)_S\|_H \ge \|(A^j x)_S\|_1/2$. Hence,

$$\begin{aligned}\omega(n^{1/2}) &= \frac{\|A^j x\|_1}{\|A^j x\|_H} = \frac{\|(A^j x)_S\|_1 + \|(A^j x)_{[n]\setminus S}\|_1}{\|A^j x\|_H} \\ &\le 2 + \frac{\|(A^j x)_{[n]\setminus S}\|_1}{\|A^j x\|_H},\end{aligned}$$

so that $\|(A^j x)_{[n]\setminus S}\|_1 = \omega(n^{1/2})\|A^j x\|_H$.

Given a value of $\|(A^j x)_{[n]\setminus S}\|_1$, the value $\|(A^j x)_{[n]\setminus S}\|_2^2$ is minimized when all of the coordinates are equal:

$$\begin{aligned}\|(A^j x)_{[n]\setminus S}\|_H &\ge n \cdot \left(\frac{\|(A^j x)_{[n]\setminus S}\|_1}{n}\right)^2 /(2\tau) \\ &= \frac{\|(A^j x)_{[n]\setminus S}\|_1^2}{2\tau n}.\end{aligned}$$

Note also that $\|(A^j x)_S\|_H \ge \tau/2$ since there exists an i for which $|A_i x| \ge \tau$ given that we are in this case.

So in order for the condition that $\|(A^j x)_{[n]\setminus S}\|_1 = \omega(n^{1/2})\|A^j x\|_H$, it must be the case that

$$\|(A^j x)_{[n]\setminus S}\|_1 = \omega(n^{1/2}) \cdot \left(\tau + \frac{\|(A^j x)_{[n]\setminus S}\|_1^2}{2\tau n}\right).$$

The right hand side of this expression is minimized when $\tau^2 = \frac{\|(A^j x)_{[n]\setminus S}\|_1^2}{2n}$, which implies $\Theta(\tau^2 n) = \|(A^j x)_{[n]\setminus S}\|_1^2$, or equivalently, $\|(A^j x)_{[n]\setminus S}\|_1 = \Theta(\tau\sqrt{n})$. But then we have

$$\Theta(\tau\sqrt{n}) = \|(A^j x)_{[n]\setminus S}\|_1 = \omega(n^{1/2}) \cdot 2\tau,$$

which is a contradiction. Hence, $\|A^j x\|_H = \Omega(n^{-1/2}\|A^j x\|_1)$, and this case reduces to the first case.

Case $|A_i x| \leq \tau$ and $\|A^j x\|_H = \Omega(n^{-1/2}\|A^j x\|_1)$. It follows that

$$\frac{H(A_i x)}{p_i} = \frac{(A_i x)^2}{2\tau p_i} \leq \frac{\tau |A_i x|}{2\tau p_i} = \frac{|A_i x|}{2p_i},$$

using that $|A_i x| \leq \tau$. Now we have the same derivation as in the first case, up to a factor of 2.

Case $|A_i x| \leq \tau$ and $\|A^j x\|_H = \Omega(n^{-1/2}\|A^j x\|_2^2/(2\tau))$. It follows using the properties of V^j that

$$\begin{aligned}
\frac{H(A_i x)}{p_i} &= \frac{(A_i x)^2}{2\tau p_i} \leq \frac{(A_i x)^2}{\tau \Theta(s) r_i^j} \leq \frac{(A_i x)^2 O(d)}{\tau s \|V_i^j\|_2^2} \\
&\leq \frac{\|V_i^j\|_2^2 \|K^j x\|_2^2 O(d)}{\tau s \|V_i^j\|_2^2} \leq \frac{\|A^j x\|_2^2 O(d)}{\tau s} \\
&\leq \frac{O(d)\tau n^{1/2}\|A^j x\|_H}{\tau s} \\
&= \frac{O(\|A^j x\|_H)}{C_0 \varepsilon^{-2} d \log(n/\varepsilon)}.
\end{aligned}$$

Hence, in all cases, if $i \in P^j$ then

$$\frac{H(A_i x)}{p_i} \leq \frac{\|A^j x\|_H}{C_1 \varepsilon^{-2} d \log(n/\varepsilon)}$$

for an arbitrarily large constant $C_1 > 0$. Then,

$$X_i - \mathbf{E}[X_i] \leq X_i \leq \frac{w_i H(A_i x)}{p_i} \leq \frac{w_i \|A^j x\|_H}{C_1 \varepsilon^{-2} d \log(n/\varepsilon)}.$$

Moreover, using the notation of Fact (2.1),

$$\begin{aligned}
\sum_{i:p_i<1} \sigma_i^2 &\leq \sum_{i:p_i<1} E[X_i^2] \\
&= \sum_j \sum_{i:p_i<1,\ i\in P^j} w_i H(A_i x) \frac{w_i H(A_i^j x)}{p_i} \\
&\leq \sum_j \frac{\|A^j x\|_{H,w}}{C_1 \varepsilon^{-2} d \log(n/\varepsilon)} \cdot \sum_{i:p_i<1,\ i\in P^j} w_i H(A_i x) \\
&\leq \sum_j \frac{\|A^j x\|_{H,w}^2}{C_1 \varepsilon^{-2} d \log(n/\varepsilon)} \leq \frac{\|Ax\|_{H,w}^2}{C_1 \varepsilon^{-2} d \log(n/\varepsilon)}.
\end{aligned}$$

Setting $\gamma = \varepsilon \|Ax\|_{H,w}$, and applying Fact (2.1),

$$\begin{aligned}
&\Pr[\|Ax\|_{H,w'} \leq \|Ax\|_{H,w} - \gamma] \\
&\qquad \leq \exp\left(\frac{-\gamma^2 C_1 \varepsilon^{-2} d \log(n/\varepsilon)}{2(\|Ax\|_{H,w})^2}\right) \\
&\qquad \leq \exp(-C_2 d \log(n/\varepsilon)),
\end{aligned}$$

and also

$$\begin{aligned}
&\Pr[\|Ax\|_{H,w'} \geq \|Ax\|_{H,w} + \gamma] \\
&\quad \leq \exp\left(\frac{-\gamma^2}{2\frac{(\|Ax\|_{H,w})^2}{C_1\varepsilon^{-2} d\log(n/\varepsilon)} + \frac{2}{3}\gamma \frac{\|Ax\|_{H,w}}{C_1 \varepsilon^{-2} d \log(n/\varepsilon)}}\right) \\
&\quad \leq \exp(-C_2 d \log(n/\varepsilon)),
\end{aligned}$$

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 $C_G = 1$.*

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 $C \equiv a/a'$. 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 $(C - 1/2)/(1/2) = 2C - 1$. 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

$$\begin{aligned}
H(a)/H(a') &= (Ca' - \tau/2)/((a')^2/(2\tau)) \\
&= 2\tau C/a' - \tau^2/(a')^2.
\end{aligned}$$

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 $O(\texttt{nnz}(A))+\text{poly}(d/\varepsilon)$ 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 $c = \|y'\|_2/(1+\varepsilon)$. Then by Lemma 4.4 as we will see in our net argument in §4, applied to $w = 1^n$, if $y^* = Ax^* - b$, where $x^* = \text{argmin}_x \|Ax - b\|_H$, then $c \le \|y^*\|_2 \le \kappa c n^{3/2}$, where $\kappa > 0$ is a sufficiently large constant.

To apply Theorem 2.2 with $w = 1^n$ 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^1 and r_i^1, but only up to a constant factor since $p_i = \min(1, \Theta(s \cdot (q_i^1 + r_i^1)))$. Recall, $q_i^1 = \frac{\|U_i^1\|_1}{\alpha}$ and let $r_i = \frac{\|V_i^1\|_2^2}{\sum_{i=1}^d \|V_i^1\|_2^2}$, where U_i^1 and V_i^1 denote the i-th row of U^1 and V^1, respectively. Here U^1 is an (α, β)-well-conditioned basis for A with respect to ℓ_1, meaning U^1 has the same column span as A, $\sum_{i=1}^d |U_i^1|_1 = \alpha$, and for all x, $\|x\|_\infty \le \beta \|U^1 x\|_1$. By Lemma 49 and Theorem 50 of [9] (see also [34, 41]), the q_i^1 can be computed in $O(\texttt{nnz}(A) \log n) + \text{poly}(d/\varepsilon)$ time, for a U^1 with $\alpha, \beta \le \text{poly}(d)$. Similarly, by Theorem 29 of [9], in $O(\texttt{nnz}(A) \log n) + \text{poly}(d/\varepsilon)$ time we can compute the r_i, for a matrix V^1 for which $\sum_i \|V_i^1\|_2^2 = \Theta(d)$ and for all x, $\|V^1 x\| = (1 \pm 1/2)\|x\|_2$. These steps succeed with probability $1 - 1/\log^C n$ probability for arbitrarily large constant $C > 0$.

The vector w' in Theorem 2.2 can be computed in $O(n)$ time, and the expected number of non-zero entries of w' 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 $\texttt{nnz}(w') \le n^{1/2}(\log^2 n)\text{poly}(d/\varepsilon)$.

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 $\texttt{nnz}(A)$ time and T has $\text{poly}(d/\varepsilon)$ rows.

Now consider the regression problem $\min_x \|Ax - b\|_{H,w'}$ subject to the constraint $\|TAx - Tb\|_2 \le 2\kappa c n^{3/2}$. 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 $w_i' = w_i/p_i$ with probability p_i (and zero otherwise), by a union bound the probability that a p_i of a nonzero w_i' is less than $1/n^2$ is at most $n/n^2 = 1/n$, so with probability $1 - o(1)$, $\|w'\|_\infty \le n^2$, implying $\|Ax - b\|_{H,w'} \le n^2 \|Ax - b\|_H$ 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 2κ) 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 $\|TAx^* - Tb\|_2 \le 2\kappa c n^{3/2}$, then $\|Ax^* - b\|_{H,w} \le (1+\varepsilon) \min_x \|Ax - b\|_{H,w}$.

Thus, we have reduced the original regression problem to the regression problem $\min_x \|Ax - b\|_{H,w'}$ constrained by $\|TAx - Tb\|_2 \le 2\kappa c n^{3/2}$, 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 ℓ-th recursive step, $\ell \ge 2$, we are given the regression problem $\min_x \|Ax - b\|_{H,w_{\ell-1}}$ subject to the constraint $\|TAx - Tb\|_2 \le 2\kappa c n^{3/2+2\ell-2}$ (we use the same matrix T in all steps), and we reduce the problem to solving $\min_x \|Ax - b\|_{H,w_\ell}$ subject to the constraint $\|TAx - Tb\|_2 \le 2\kappa c n^{3/2+2\ell-2}$. We now describe the ℓ-th recursive step.

We inductively have that $\|w_{\ell-1}\|_\infty \le n^{2\ell-2}$. We first group the weights of $w_{\ell-1}$ into $O(\log n)$ groups P^j. For each group we compute U_i^j and V_i^j as above, thereby obtaining w_ℓ in $O(t_{\ell-1} \log n)$ expected time, where $t_{\ell-1}$ is the number of non-zero weights in $w_{\ell-1}$. The expected value of t_ℓ is $O(t_{\ell-1}^{1/2} \max(\alpha \cdot \beta, d) \cdot \varepsilon^{-2} d \log(n/\varepsilon) \log n)$. We can condition on $t_{\ell-1}$ being $O(n^{1/2^{\ell-1}} \text{poly}(d\varepsilon^{-1} \log n))$ as all events jointly succeed with probability $1 - o(1)$. We thus have $t_\ell = n^{1/2^\ell} \text{poly}(d\varepsilon^{-1} \log n)$ with probability $1 - o(1)$. We now consider the regression problem $\min_x \|Ax - Ab\|_{H,w_\ell}$ subject to the constraint $\|TAx - Tb\|_2 \le 2\kappa c n^{3/2} \|w_{\ell-1}\|_\infty \le 2\kappa c n^{3/2+2\ell-2}$. 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_ℓ and satisfies $\|w_\ell\|_\infty \le n^2 \|w_{\ell-1}\|_\infty \le n^{2\ell}$. We can thus apply Lemma 4.5 with S equal to the identity to conclude that with probability $1 - o(1)$, if $x^* = \text{argmin}_x \|Ax - b\|_{H,w_\ell}$ subject to the constraint $\|TAx^* - Tb\|_2 \le 2\kappa c n^{3/2+2\ell-2}$, then $\|Ax^* - b\|_{H,w_{\ell-1}} \le (1+\varepsilon) \min_x \|Ax - b\|_{H,w_{\ell-1}}$.

It follows that for ℓ a large enough constant, and by scaling ε by a constant factor, we will have that with probability $1 - o(1)$, if $x^* = \text{argmin}_x \|Ax - b\|_{H,w_\ell}$ subject to the constraint $\|TAx^* - Tb\|_2 \le 2\kappa c n^{3/2+2\ell-2}$, then $\|Ax^* - b\|_H \le (1+\varepsilon) \min_x \|Ax - b\|_H$. Moreover, $t_\ell \le n^{1/2^\ell} \text{poly}(d\varepsilon^{-1} \log n)$. 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_ℓ^C time for a fixed constant $C > 0$. Setting $2^\ell > C/2$ and assuming the $\mathrm{poly}(d\varepsilon^{-1}\log n)$ 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 $O(\mathtt{nnz}(A)\log n) + \mathrm{poly}(d/\varepsilon)$.

3 M-sketches for M-estimators.

Given a function $G : \mathbb{R} \mapsto \mathbb{R}^+$ with $G(a) = G(-a)$, and $G(0) = 0$, we can use the sketch of $z \in \mathbb{R}^n$ to estimate $\|z\|_G \equiv \sum_p G(z_p)$, 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 ℓ_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 $G^{-1}(\|tz\|_G) = |t|G^{-1}(\|z\|_G)$, 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 ℓ_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 $S \in \mathbb{R}^{Nh_{\max} \times n}$, where $h_{\max} \equiv \lfloor \log_b(n/m) \rfloor$, and our scaling vector $w \in \mathbb{R}^{Nh_{\max}}$. The entries of S are $S_{j,p} \leftarrow \Lambda_p$, and the entries of w are $w_j \leftarrow \beta b^{h_p}$, where $\beta \equiv (b - b^{-h_{\max}})/(b-1)$, $j \leftarrow g_p + Nh_p$, and

(3.6)

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

all independently. Let L_h be the multiset $\{z_p \mid h_p = h\}$, and $L_{h,i}$ the multiset $\{z_p \mid h_p = h, g_p = i\}$; that is, L_h is multiset of values at a given level, $L_{h,i}$ is the multiset of values in a bucket. We can write $\|Sz\|_{G,w}$ as $\sum_{i\in[N],h\in[0,h_{\max}]} \beta b^h G(\|L_{h,i}\|_\Lambda)$, where $\|L\|_\Lambda$ denotes $|\sum_{z_p \in L} \Lambda_p z_p|$.

(The function $\|\|_\Lambda$ is a semi-norm (if we map sets back to vectors), with $\|L\|_\Lambda \le \|L\|_1$, $\mathbf{E}_\Lambda[\|L\|_\Lambda^2] = \|L\|_2^2$, and all $(\mathbf{E}_\Lambda[\|L\|_\Lambda^k])^{1/k}$ 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 $\mathtt{OPT}_G \equiv \min_{x\in\mathbb{R}^d} \|Ax - b\|_G$. There is an algorithm that in $\mathtt{nnz}(A) + \mathrm{poly}(d\log n)$ time, with constant probability finds $\hat{x}$ such that $\|A\hat{x} - b\|_G \le O(1)\mathtt{OPT}_G$.*

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 z, and let Y denote the multiset comprising the coordinates of y. For $\hat{Z} \subset Z$, let $G(\hat{Z}) \subset Y$ denote $\{G(z_p) \mid z_p \in \hat{Z}\}$.

Let $\|Y\|_k$ denote $[\sum_{y\in Y} |y|^k]^{1/k}$, so $\|Y\|_1 = \|y\|_1$.

Hereafter multisets will just be called "sets".

Weight classes. For our analysis, fix $\gamma > 1$, and for integer $q \ge 1$, let W_q denote *weight class* $\{y_p \in Y \mid \gamma^{-q} \le y_p \le \gamma^{1-q}\}$.

We have $\beta b^h\, \mathbf{E}[\|G(L_h) \cap W_q\|_1] = \|W_q\|_1$.

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 $\|y'\|_1 \ge 1 - n(\varepsilon/n) = 1 - \varepsilon$. Thus for some purposes we can neglect W_q for $q > q_{\max} \equiv \log_\gamma(n/\varepsilon)$, up to error ϵ. Moreover, we can assume that $\|W_q\|_1 \ge \varepsilon/q_{\max}$, since the total contribution of weight classes of smaller total weight to $\|y\|_1$ is at most ϵ.

Let $h(q)$ denote $\lfloor \log_b(|W_q|/\beta m) \rfloor$ for $|W_q| \ge \beta m$, and zero otherwise, so that

$$m \le \mathbf{E}[|G(L_{h(q)}) \cap W_q|] \le bm$$

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 $h(q) \leq \lfloor \log_b(n/\beta m) \rfloor = h_{\max}$.

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 $N = bcm$.

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 $\log N \leq \varepsilon^2 m$. We will consider γ 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 $4 q_{\max} h_{\max} \exp(-\varepsilon^2 m/3) \leq C^{-\varepsilon^2 m}$ for a constant $C > 1$, the event $\mathcal{E}$ holds, that for all $q \leq q_{\max}$ 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 $t \equiv Z - \mathbf{E}[Z] = Z - \mu$ exceeds $\epsilon\mu$ is at most

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

Taking the exponential, and using a union bound over all events (including the event that $-t$ exceeds $\epsilon s/\beta b^h$) completes the first claim, with half the claimed failure probability, using Assumption 3.1 to shown 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 $\sum_p \mathbf{E}[X_p^2] \leq \sum_p \mathbf{E}[X_p] = \|W_q\|_1/\beta b^h$.

LEMMA 3.2. *For $h \in [h_{\max}]$, suppose $Q \subset \{q \mid h(q) = h, |W_q| \geq \beta m\}$, and $\hat{W} \subset Y$ contains $W_Q \equiv \cup_{q \in Q} W_q$. If $|G(L_h) \cap \hat{W}| \leq \varepsilon N$, 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 $\hat{W}$. 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 $\hat{W}$.

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 $\beta^{-1} b^{-h} (1 - |G(L_h) \cap \hat{W}|/N) \geq \beta^{-1} b^{-h} (1 - \varepsilon)$. We have $a_q \geq (1-\varepsilon)\beta^{-1} b^{-h} |W_q|$ in expectation.

To show that this holds with high probability, let $\hat{Z}_k \equiv \mathbf{E}[Z_s \mid Z_k]$. 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 $\hat{W}$.

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

$$\begin{aligned} \frac{\|W_q^*\|_1}{\|G(L_h) \cap W_q\|_1} &\geq \frac{(1-\varepsilon)X\gamma^{-q}}{(1-\varepsilon)X\gamma^{-q} + 2\varepsilon X \gamma^{1-q}} \\ &\geq 1 - 2\gamma\varepsilon/(1-\varepsilon). \end{aligned}$$

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

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

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

then with failure probability δ,

$$\max_{i \in [N]} \|G(L_{h,i}) \cap \bar{W}\|_1 \leq \frac{7}{6} T \log(N/\delta).$$

Proof. This directly follows from Lemma 2 of [9], (which follows directly from Bernstein's inequality), where t of that lemma is N, T is the same, $u_{s:n}$ is $\bar{W}$, r is $\|\bar{W}\|_1$, and δ_h is δ. The bound for N also uses $\|\bar{W}\|_2^2 \leq \|\bar{W}\|_\infty \|\bar{W}\|_1$.

3.4 Leverage Scores. The ℓ_2 *leverage scores* $u \in \mathbb{R}^n$ have $u_i \equiv \|U_{i:}\|_2^2$, where U is an orthogonal basis for the *columnspace* $C(A) \equiv \{Ax \mid x \in \mathbb{R}^d\}$. We will use the standard facts that these values satisfy $\|u\|_\infty \leq 1$ and $\|u\|_1 \leq d$, and for $y \in C(A)$ with $\|y\|_2 = 1$, $y_i^2 \leq u_i$ for $i \in [n]$.

We will condition on a likely event involving the top leverage scores. This lemma will be used to bound the effect of those W_q with $|W_q|$ small and weight γ^{-q} large.

LEMMA 3.4. *For $A \in \mathbb{R}^{n\times d}$, let $u \in \mathbb{R}^n$ denote the ℓ_2 leverage score vector of A. For N_1, N_2 with $N_2 \geq N_1$ and with $N_1 N_2 \leq \kappa N$, for $\kappa \in (0, 1/2)$, let Y_1 and Y_2 denote the sets of indices of the N_1 and N_2 largest coordinates of u, so that $Y_1 \subset Y_2$. Then with probability at least $1 - 2\kappa$, the event $\mathcal{E}_c$ 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 $\mathcal{E}_c$ 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 2κ, so with probability at least $1-2\kappa$, the number of collisions is less than $2\kappa < 1$, that is, zero.

We use the ℓ_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 3.5. *If u_p is the k'th largest ℓ_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 $z = Ux \in C(A)$,

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

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

$$\sum_{z_q\leq z_p}\frac{G(z_q)}{G(z_p)} \geq \sum_{z_q\leq z_p}\left|\frac{z_q}{z_p}\right|^\alpha \geq \sum_{z_q\leq z_p}\left|\frac{z_q}{z_p}\right|^2 \geq k/2d,$$

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

$$\sum_{z_q\geq z_p}\frac{G(z_q)}{G(z_p)} \geq C_G\sum_{z_q\geq z_p}\left|\frac{z_q}{z_p}\right| \geq C_G\left[\sum_{z_q\geq z_p}\left|\frac{z_q}{z_p}\right|^2\right]^{1/2} \geq C_G\sqrt{k/2d},$$

and the claimed inequality follows.

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

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

LEMMA 3.6. *For $v \in T \subset Z$,*

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

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

$$\frac{\|T-v\|_2}{|v|} \leq \varepsilon^{1/\alpha}, \tag{3.7}$$

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

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 $|\,|v| \pm \|T - v\|_\Lambda|$, so if $\|T\|_\Lambda \leq |v|$, then $\|T\|_\Lambda = |\,|v| - \|T - v\|_\Lambda|$. We have

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

proving the first claim. For the second claim, we have $|v'| < |v|$ for $v' \in T - v$, since $G(v') \leq \|T - v\|_G \leq \varepsilon G(v)$, 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,

$$\mathbf{E}_\Lambda[G(\|T\|_\Lambda)] \geq \mathbf{E}_\Lambda\left[\left(1 - \frac{\|T-v\|_\Lambda}{|v|}\right)^2\right]G(v) \geq \left(1 - 2\frac{\mathbf{E}_\Lambda[\|T-v\|_\Lambda]}{|v|}\right)G(v).$$

Using the Khintchine inequality and (3.7), we have

$$\frac{\mathbf{E}_\Lambda[\|T-v\|_\Lambda]}{|v|} \le \frac{C\|T-v\|_2}{|v|} \le C\varepsilon^{1/\alpha},$$

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 $h = 0 \dots h_{\max}$, let

$$\begin{aligned}
\hat{Q}_h &\equiv \{q \mid h(q) = h, |W_q| \ge \beta m\}\\
M_\ge &\equiv \log_\gamma(2(1+3\varepsilon)b/\varepsilon)\\
Q_h &\equiv \{q \in \hat{Q}_h \mid q \le M_\ge + \min_{q\in\hat{Q}_h} q\}\\
M_< &\equiv \log_\gamma(m/\varepsilon) = O(\log_\gamma(b/\varepsilon))\\
Q_< &\equiv \{q \mid |W_q| < \beta m, q \le M_<\}\\
Q^* &\equiv Q_< \cup [\cup_h Q_h].
\end{aligned} \tag{3.8}$$

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 $\mathcal{E}$ of Lemma 3.1, $\sum_{q\in Q^*}\|W_q\|_1 \ge 1-5\varepsilon$.*

Proof. The total weight of those weight classes with $|W_q| \le \beta m$ and $q > M_<$ is at most

$$\beta m \sum_{q>M_<} \gamma^{1-q} \le \beta m(\varepsilon/m)\gamma \sum_{g>0}\gamma^{-q} \le \varepsilon\beta/(1-1/\gamma) \le 4\varepsilon,$$

for $\gamma \ge 2$ and $\beta \le 2$.

For given $h > 0$, let $q_h^* \equiv \min_{q\in\hat{Q}_h} q$. The ratio of the total weight of classes in $\hat{Q}_h \setminus Q_h$ to $\|W_{q_h^*}\|_1$ is at most

$$\begin{aligned}
\frac{1}{(1-\varepsilon)\gamma^{-q_h^*}m}&\gamma^{-q_h^*-M_\ge}\sum_{q>0}(1+\varepsilon)bm\gamma^{1-q}\\
&= b\frac{\varepsilon}{2b(1+3\varepsilon)}\frac{1+\varepsilon}{1-\varepsilon}\sum_{q\ge 0}\gamma^{-q}\\
&= \frac{\varepsilon}{2(1+3\varepsilon)}\frac{1+\varepsilon}{1-\varepsilon}\frac{1}{1-1/\gamma}\\
&\le \varepsilon,
\end{aligned}$$

under the assumptions on γ and ε. So $\sum_h \|W_{\hat{Q}_h\setminus Q_h}\|_1 \le \sum_h \varepsilon\|W_{q_h^*}\|_1 \le \varepsilon$.

Putting together the bounds for the two cases, the total is at most 5ε, as claimed.

Lemma 3.8. *Assume that condition $\mathcal{E}$ of Lemma 3.1 holds, and that condition $\mathcal{E}_c$ of Lemma 3.4 holds for $N_1 = N_2 = O(C_G^{-2}\varepsilon^{-2}dm^2)$. Let $Q'_h \equiv \{q \mid q \le M'_h\}$, where $M'_h \equiv \log_\gamma(\beta b^{h+1}m^2q_{\max})$. Then there is $N = O(N_1^2 + m^2b\varepsilon^{-1}q_{\max})$ 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. *$|W_q^*| \ge (1-\varepsilon)\beta^{-1}b^{-h(q)}|W_q|$;*
2. *each $x \in W_q^*$ is in a bucket with no other member of W_{Q^*};*
3. *$\|W_q^*\|_1 \ge (1-4\gamma\varepsilon)\beta^{-1}b^{-h}\|W_q\|_1$.*
4. *for $q \in Q_h$, each $x \in W_q^*$ is in a bucket with no member of $W_{Q'_h}$;*

Proof. There is N_1 satisfying the given bound so that Lemma 3.5 implies that $y \notin Y_1$ must be smaller than $C_G^{-1}\sqrt{2d/N_1} \le \varepsilon/m$, and therefore not in W_q for $q \in Q_<$. Therefore $W_{Q_<} \subset Y_1$, and with the assumption of condition $\mathcal{E}_c$, no member of $W_{Q_<}$ is in the same bucket as any other member of that set. We will take $W_q^* \leftarrow W_q$ for $q \in Q_<$.

For each h, apply Lemma 3.2 to Q_h and with $\hat{W} \leftarrow W_{Q^*} \equiv W_{Q_<}\cup_{q\in Q_h} W_q$, so that, using condition $\mathcal{E}$,

$$\begin{aligned}
|G(L_h)\cap\hat{W}| &\le M_<\beta m + M_\ge(1+\varepsilon)bm\\
&= O(mb\log_\gamma(b/\varepsilon)).
\end{aligned}$$

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

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

For (4), let $\hat{W} \leftarrow W_{Q_h} \cup W_{Q'_h}$. Since $|W_{Q'_h}|\gamma^{-M'_h} \le \|y\|_1 \le 1$, so that $|W_{Q'_h}| \le \beta b^{h+1}m^2q_{\max}$, we have

$$\begin{aligned}
|G(L_h)\cap\hat{W}| &\le |G(L_h)\cap W_{Q_h}| + |G(L_h)\cap W_{Q'_h}|\\
&\le (1+\varepsilon)bmM_\ge + (1+\varepsilon)\beta^{-1}b^{-h}|W_{Q'_h}|\\
&\le O(bm^2q_{\max}),
\end{aligned}$$

using condition $\mathcal{E}$. Since $|G(L_h)\cap\hat{W}| \le \varepsilon N$ for large enough $N = O(m^2b\varepsilon^{-1}q_{\max})$, we can apply Lemma 3.2 to obtain (4).

Lemma 3.9. *Let $G : \mathbb{R} \mapsto \mathbb{R}^+$ as above. Assume that condition $\mathcal{E}$ of Lemma 3.1 holds, and Assumption 3.1, and that condition $\mathcal{E}_c$ of Lemma 3.4 holds for $N_1 = N_2 = O(C_G^{-2}\varepsilon^{-2}dm^2)$. There is $N = O(N_1^2 +$*

$\varepsilon^{-2}m^2 bq_{\max})$, *so that for* $h \in [h_{\max}]$ *and* $q \in Q_h$ *with* $\|W_q\|_1 \geq \varepsilon/q_{\max}$, *we have*

$$\sum_{y_p \in W_q^*} G(\|L(y_p)\|_\Lambda) \geq (1-\varepsilon^{1/\alpha})\|W_q\|_1$$

with failure probability at most $C^{-\varepsilon^2 m}$ *for fixed* $C > 1$.

Proof. For any $q \in Q_h$ we have

$$\begin{aligned}|W_q| &\leq (1+\varepsilon)\beta b^h\, \mathbf{E}[|G(L_h) \cap W_q|] \\ &\leq (1+\varepsilon)\beta b^h bm\end{aligned}$$

by condition $\mathcal{E}$ and the definition of $h(q) = h$; since

$$|W_q|\gamma^{1-q} \geq \|W_q\|_1 \geq \varepsilon/q_{\max},$$

using $\|W_q\|_1 \geq \varepsilon/q_{\max}$ from the lemma statement, we have for any $y_p \in W_q$,

(3.9)
$$y_p \geq \gamma^{-q} \geq (\varepsilon/q_{\max})/\gamma|W_q| \geq \varepsilon/b^{h+1}\gamma\beta m(1+\varepsilon)q_{\max}.$$

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 $\gamma/\beta b^{h+1}m^2 q_{\max}$, so if $\bar{W}$ comprises $G(L_h) \cap (Y \setminus W_{Q_h'})$, we have $\|\bar{W}\|_\infty \leq \gamma/\beta b^{h+1}m^2 q_{\max}$. Using condition $\mathcal{E}$, $\|\bar{W}\|_1 \leq (1+\varepsilon)b^{-h}$, using just the condition $\|Y\|_1 = 1$. Therefore the given N is larger than the $O(bm\varepsilon^{-2}q_{\max})$ 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 ε, obtaining $\mathbf{E}_\Lambda[G(\|L(v)\|_\Lambda)] \geq (1 - C'\varepsilon^{1/\alpha})G(v)$. (Here we fold a factor of $(2\gamma^2)^{1/\alpha}$ into C', recalling that we consider γ to be fixed.) Using this relation and condition $\mathcal{E}$, we have

$$\begin{aligned}\|W_q\|_1 &\leq \beta b^h \|W_q^*\|_1/(1-4\gamma\varepsilon) \qquad \text{from Lem 3.8.3} \\ &\leq \beta b^h \sum_{G(v)\in W_q^*} \frac{\mathbf{E}_\Lambda[G(\|L(v)\|_\Lambda)]}{(1-4\gamma\varepsilon)(1-C'\varepsilon^{1/\alpha})},\end{aligned}$$

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 $G(\|L(v)\|_\Lambda)$ to $G(v)$ via the first claim of Lemma 3.6, which implies $G(\|L(v)\|_\Lambda) \geq (1 - 2\|L(v) - v\|_\Lambda/|v|)G(v)$. Writing $V \equiv G^{-1}(W_q^*)$, we have

$$\begin{aligned}&\sum_{v\in V} G(\|L(v)\|_\Lambda) \\ &\quad \geq \sum_{\substack{v\in V \\ \|L(v)\|_\Lambda > |v|}} G(v) \quad + \sum_{\substack{v\in V \\ \|L(v)\|_\Lambda \leq |v|}} \left(1 - 2\frac{\|L(v)-v\|_\Lambda}{|v|}\right) G(v) \\ &\quad \geq \|W_q^*\|_1 - 2\sum_{\substack{v\in V \\ \|L(v)\|_\Lambda \leq |v|}} \frac{\|L(v)-v\|_\Lambda}{|v|}\gamma^{1-q}.\end{aligned}$$

It remains to upper bound the sum. Since $\|L(v)\|_\Lambda = ||v| \pm t|$, where $t \equiv \|L(v) - v\|_\Lambda$, if $\|L(v)\|_\Lambda \leq |v|$, then $t \leq 2|v|$.

Since

$$\mathbf{E}[t \mid t \leq 2|v|] \leq \mathbf{E}[t] \leq C_1\|L(v) - v\|_2 \leq C_1 C'\varepsilon^{1/\alpha}|v|,$$

using Khintchine's inequality and (3.7), and similarly $\mathbf{E}[t^2 | t \leq 2|v|] \leq C_2(C'\varepsilon^{1/\alpha})^2 v^2$, we can use Bernstein's inequality to bound

$$\begin{aligned}\sum_{\substack{v\in V \\ \|L(v)\|_\Lambda \leq |v|}} \frac{\|L(v)-v\|_\Lambda}{|v|}\gamma^{1-q} &\leq \sum_{\substack{v\in V \\ \|L(v)\|_\Lambda \leq |v|}} C_3\varepsilon^{1/\alpha}\gamma^{1-q} \\ &\leq C_4\varepsilon^{1/\alpha}\|W_q^*\|_1,\end{aligned}$$

with failure probability $\exp(-\varepsilon^2 m)$. Hence

$$\begin{aligned}&\sum_{v\in V} G(\|L(v))\|_\Lambda) \\ &\quad \geq \|W_q^*\|_1 - 2C_4\varepsilon^{1/\alpha}\|W_q^*\|_1 \\ &\quad = \|W_q^*\|_1(1 - 2C_4\varepsilon^{1/\alpha}) \\ &\quad \geq \beta^{-1}b^{-h}\|W_q\|_1(1-4\gamma\varepsilon)(1-2C_4\varepsilon^{1/\alpha}),\end{aligned}$$

using condition $\mathcal{E}$. Adjusting constants, the result follows.

LEMMA 3.10. *Assume that condition* $\mathcal{E}$ *of Lemma 3.1 holds, and Assumption 3.1, and* $\mathcal{E}_c$ *of Lemma 3.4 holds for large enough* $N_1 = O(C_G^{-2}\varepsilon^{-2}dm^2)$ *and* $N_2 = O(C_G^{-2}d(\varepsilon^{2\alpha}m^{4+\alpha} + \varepsilon^{4\alpha-4}m^{2+2\alpha}))$. *Then for* $q \in Q_<$,

$$\sum_{v\in G^{-1}(W_q)} \|G(L(v))\|_\Lambda \geq (1-\varepsilon^{1/\alpha})\|W_q\|_1$$

with failure probability at most $C^{-\varepsilon^2 m}$ *for a constant* $C > 1$.

Proof. For all $y_p \in W_{Q_<}$, we have

$$y_p \geq \frac{\varepsilon}{m}.$$

Let
$$\gamma' \equiv \min\{\frac{\varepsilon^{-\alpha}}{3\varepsilon^{\alpha} m^{2+\alpha/2}}, \frac{\varepsilon^{2-2\alpha}}{m^{1+\alpha}}\},$$
so that N_2 of the lemma statement is at least $2d/{\gamma'}^2 C_G^2$. Then condition $\mathcal{E}_c$ and Lemma 3.5 imply that every member of W_q is in a bucket with no entry other than itself larger than γ'.

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 $\bar{W} \equiv Y \setminus Y_2$, so that

$$\|\bar{W}\|_\infty \le \gamma' \le \frac{\varepsilon^{-\alpha}}{3m^{2+\alpha/2}} = \left(\frac{1}{\varepsilon^{1-1/\alpha}\sqrt{m}}\right)^\alpha \frac{\varepsilon}{3m}\frac{1}{\varepsilon^2 m}.$$

The result is that with large enough $N = O(m^{1+\alpha/2}\varepsilon^{-2+\alpha})$, and assuming $\log N \le \varepsilon^2 m$, so that $\log(N/\delta) \le 2\varepsilon^2 m$, we have for v with $G(v) = y_p \in W_q$,

$$\begin{aligned}\|G(L(v)) \cap \bar{W}\|_1 &\le \frac{7}{6}\left(\frac{1}{\varepsilon^{1-1/\alpha}\sqrt{m}}\right)^\alpha \frac{2\varepsilon}{3m} \\ &\le \left(\frac{1}{\varepsilon^{1-1/\alpha}\sqrt{m}}\right)^\alpha |y_p|,\end{aligned}$$

that is,

$$\frac{\|L(v) - v\|_G}{G(v)} \le \left(\frac{1}{\varepsilon^{1-1/\alpha}\sqrt{m}}\right)^\alpha,$$

so that from (3.7), we have

$$\frac{\|L(v) - v\|_2^2}{v^2} \le \frac{\varepsilon^{2/\alpha}}{\varepsilon^2 m}. \tag{3.10}$$

Since

$$\|\bar{W}\|_\infty \le \frac{\varepsilon^{2-2\alpha}}{m^{1+\alpha}} = \left(\frac{1}{m\varepsilon^{2-1/\alpha}}\right)^\alpha \frac{\varepsilon}{m},$$

we also have, for all $v' \in L(v) - v$, and using that $G(v) \ge \varepsilon/m$,

$$\left|\frac{v'}{v}\right| \le \left(\frac{G(v')}{G(v)}\right)^{1/\alpha} \le \frac{1}{m\varepsilon^{2-1/\alpha}}. \tag{3.11}$$

From (3.11), we have that the summands determining $\|L(v) - v\|_\Lambda$ have magnitude at most $|v|\varepsilon^{1/\alpha}/\varepsilon^2 m$. From (3.10), we have $\|L(v) - v\|_2^2$ is at most $v^2\varepsilon^{2/\alpha}/\varepsilon^2 m$. It follows from Bernstein's inequality that with failure probability $\exp(-\varepsilon^2 m)$, $\|L(v) - v\|_\Lambda \le \varepsilon^{1/\alpha}|v|$. Applying the first claim of Lemma 3.6, we have $G(\|L(v)\|_\Lambda) \ge (1 - 2\varepsilon^{1/\alpha})G(v)$, 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 $\mathcal{E}$ of Lemma 3.1 holds, and Assumption 3.1, and condition $\mathcal{E}_c$ of Lemma 3.4 holds for $N_1 = O(C_G^{-2}\varepsilon^{-2}dm^2)$ and $N_2 = O(C_G^{-2}d(\varepsilon^{2\alpha}m^{4+\alpha} + \varepsilon^{4\alpha-4}m^{2+2\alpha}))$, with $N = O(N_1 N_2 + \varepsilon^{-2}m^2 b q_{\max})$. Then $\|Sz\|_{G,w} \ge \|z\|_G(1 - \epsilon^{1/\alpha})$, with failure probability no more than $C^{-\varepsilon^2 m}$, for absolute $C > 1$.*

Proof. (We note that c, b, and 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/ε, which ultimately will require that n is not extremely large relative to d.)

Recalling Q^* from (3.8), let $Q^{**} \equiv \{q \mid q \in Q^*, \|W_q\|_1 \ge \varepsilon/q_{\max}\}$. Assuming conditions $\mathcal{E}$ and $\mathcal{E}_c$, we have, with probability $1 - C^{-\varepsilon^2 m}$,

$$\begin{aligned}\|Sz\|_{G,w} &= \sum_{h,i}\beta b^h G(\|L_{h,i}\|_\Lambda) && \text{Def.} \\ &\ge \sum_{q\in Q^{**}, v\in W_q^*}\beta b^{h(q)} G(\|L(v)\|_\Lambda) && \text{Lem 3.8} \\ &\ge \sum_{q\in Q^{**}}\beta b^{h(q)}(1 - \varepsilon^{1/\alpha})\|W_q^*\|_1 && \text{Lems 3.9, 3.10} \\ &\ge \sum_{q\in Q^{**}}(1 - \varepsilon^{1/\alpha})(1 - 4\gamma\varepsilon)\|W_q\|_1 && \text{Lem 3.8.}\end{aligned}$$

Using Lemma 3.7,

$$\sum_{q\in Q^{**}}\|W_q\|_1 \ge -q_{\max}(\varepsilon/q_{\max}) + \sum_{q\in Q^*}\|W_q\|_1 \ge 1 - 6\varepsilon.$$

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 $\mathbf{E}[\|Sz\|_{G,w}] = O(h_{\max})\|z\|_G$, which implies a tail bound via Markov's inequality; first, some lemmas.

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

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

$$\begin{aligned}
&\mathbf{E}[G(\|T\|_\Lambda)] \\
&\quad = \mathbf{E}[G(\|T\|_\Lambda) \mid \mathcal{L}\,]\,\mathbf{P}\{\mathcal{L}\} + \mathbf{E}[G(\|T\|_\Lambda) \mid \neg\mathcal{L}\,]\,\mathbf{P}\{\neg\mathcal{L}\} \\
&\quad \le \mathbf{E}\left[\frac{\|T\|_\Lambda^\alpha}{\|T\|_2^\alpha} G(\|T\|_2) \mid \mathcal{L}\right]\mathbf{P}\{\mathcal{L}\} + G(\|T\|_2) \\
&\quad = \mathbf{E}[\|T\|_\Lambda^\alpha \mid \mathcal{L}\,]\,\mathbf{P}\{\mathcal{L}\}\frac{G(\|T\|_2)}{\|T\|_2^\alpha} + G(\|T\|_2) \\
&\quad \le \mathbf{E}[\|T\|_\Lambda^\alpha]\frac{G(\|T\|_2)}{\|T\|_2^\alpha} + G(\|T\|_2) \\
&\quad \le CG(\|T\|_2),
\end{aligned}$$

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

LEMMA 3.12. *For $T \subset Z$, $G(\|T\|_2) \le \|T\|_G$, and so $\mathbf{E}_\Lambda[G(\|T\|_\Lambda)] \le C\|T\|_G$.*

Proof. Using the growth upper bound for G,

$$\begin{aligned}
\frac{\|T\|_G}{G(\|T\|_2)} &= \sum_{z_p \in T} \frac{G(z_p)}{G(\|T\|_2)} \\
&\ge \sum_{z_p \in T} \left|\frac{z_p}{\|T\|_2}\right|^\alpha \\
&\ge \sum_{z_p \in T} \left|\frac{z_p}{\|T\|_2}\right|^2 \\
&= 1.
\end{aligned}$$

The last claim follows from this and the previous lemma.

THEOREM 3.3. *Assuming condition $\mathcal{E}$ of Lemma 3.1, $\mathbf{E}_{g,\ell,\Lambda}[\|Sz\|_{G,w}] = O(h_{\max})\|z\|_G$.*

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

$$\begin{aligned}
&\mathbf{E}_{g,\ell,\Lambda}[\|Sz\|_{G,w}] \\
&\qquad = \sum_h \beta b^h \sum_i \mathbf{E}_\Lambda[G(\|L_{h,i}\|_\Lambda)] \\
&\qquad \le \sum_h \beta b^h \sum_i C(1+\varepsilon)\|y\|_1/\beta b^h \\
&\qquad = \sum_h \sum_i C(1+\varepsilon)\|y\|_1 = h_{\max} C(1+\varepsilon)\|z\|_G,
\end{aligned}$$

and the theorem follows, picking bounded ε.

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

$$M^* \equiv bmM_\ge + \beta m M_< = O(mb\log_\gamma(b/\varepsilon))$$

in Λ-semi-norm, recalling $M_\ge$ and $M_<$ defined in (3.8). That is,

$$\|Sz\|_{Gc,w} \equiv \sum_j \beta b^j \sum_{i \in [M^*]} G(\|L_{j,(i)}\|_\Lambda),$$

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

The proof of the bounded contraction of $\|Sz\|_{G,w}$, Theorem 3.2, only requires lower bounds on $\|G(L_{h,i})\|_\Lambda$ 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 $\|Sy\|_{G,w}$ uses only the largest such buckets in Λ-norm, the proven bounds on contraction continue to hold, and in particular $\|Sz\|_{Gc,w} \ge (1-\varepsilon)\|Sz\|_{G,w}$.

Moreover, the dilation of $\|\|_{Gc,w}$ is constant:

THEOREM 3.4. *There is $c = O(\log_\gamma(b/\varepsilon)(\log_b(n/m)))$ and $b \ge c$, recalling $N = mbc$, such that*

$$\mathbf{E}[\|Sz\|_{Gc,w}] \le C\|z\|_G$$

for a constant C.

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

$$\mathbf{E}[\sum_i G(\|L_{h,i}\|_\Lambda)] \le C\|L_h\|_G = C\|G(L_h)\|_1. \tag{3.12}$$

We will consider the contribution of each weight class separately. The contribution of W_q at $h = h(q)$ is $\beta b^h\|G(L_h) \cap W_q\|_1 \le \|W_q\|_1(1+\varepsilon)$, 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 $M^*|L_{h,i^*} \cap W_q|\gamma^{1-q}$, where L_{h,i^*} contains the largest number of members of W_q among the buckets on level h. When $|G(L_h) \cap W_q| \ge N\log N$, $|G(L_{h,i^*}) \cap W_q| \le 4|G(L_h) \cap W_q|/N$, with failure probability at most $1/N$. (This follows by applying Bernstein's inequality to the sum of random variables X_i, where $X_i = 1$ when the i'th element of W_q falls in a given bucket, and $X_i = 0$ otherwise, followed by a union bound over the buckets.) At level $h = h(q) - k$, $|G(L_h) \cap W_q| \ge b^k m(1-\varepsilon)$, using assumption $\mathcal{E}$ of Lemma 3.1, so to obtain $b^k m(1-\varepsilon) \ge N\log N$ it suffices that $k \ge 2 + 2\log_b c \ge \log_b(N\log(N)/m(1-\varepsilon))$, using

$N = bcm$, obtaining for those k a contribution for W_q is within a constant factor of

$$\beta b^h M^*(4\beta^{-1}b^{-h}|W_q|/N)\gamma^{1-q} \le \frac{O(\log_\gamma(b/\varepsilon))}{c}\|W_q\|_1,$$

using the bound on M^* given above. Adding this contribution to that for $k \le 2 + 2\log_b c$, we obtain an overall bound for W_q and $h < h(q)$ that is within a constant factor of $(1 + \log_b c + h_{\max}\frac{M^*}{N})\|W_q\|_1$, and therefore within a constant factor of $\|W_q\|_1$ under the given conditions on b and c.

For $h = h(q) + k$ for $k > 0$, the expected size of $G(L_h) \cap W_q$ is at most m/b^{k-1}; this quantity is also an upper bound for the probability that $G(L_h) \cap W_q$ is non-empty. Thus for the $q_{\max}$ non-negligible sets W_q, by a union bound the event $\mathcal{E}_s$ holds with failure probability δ, that all $W_q \cap L_{h(q)+k}$ will be empty for large enough $k = O(\log_b q_{\max} m/\delta)$. For each q and k, condition $\mathcal{E}$ implies that the contribution $\beta b^h \sum_i \|G(L_{h,i})\|_\Lambda \le (1 + \varepsilon)\|W_q\|_1$, and so the total contribution is $C\|W_q\|_1 \log_b q_{\max} m/\delta$, within a constant factor of $\|W_q\|_1$, under given conditions.

Note that if G is convex, then so is $\|Sz\|_{Gc,w}$, 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 $G^{-1}(\|\cdot\|_{Gc,w})$. If both conditions hold, then $G^{-1}(\|\cdot\|_G)$ is a norm, and so is $G^{-1}(\|\cdot\|_{Gc,w})$.

The Ky Fan k-norm of a vector $y \in \mathbb{R}^n$ is $\sum_{i\in[k]} |y_{(i)}|$, where $y_{(i)}$ denotes the i'th largest entry of y in magnitude. Thus the Ky Fan 1-norm of y is $\|y\|_\infty$, and the Ky Fan n-norm of y is $\|y\|_1$. 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: $\|z\|_{Gc,w}$ 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 ε-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 $|b| \le \varepsilon|a|$, $G(a + b) = (1 \pm C\varepsilon)G(a)$.*

Proof. First suppose that a and b have the same sign. Then by montonicity, $G(a) \le G(a + b)$. Moreover, by the growth condition,

$$\frac{G(a+b)}{G(a)} \le \left|\frac{a+b}{a}\right|^2 \le (1+\varepsilon)^2 \le 1 + 3\varepsilon,$$

and so $G(a + b) \le (1 + 3\varepsilon)G(a)$.

Now suppose a and b have the opposite sign. Then $G(a + b) \le G(a)$ by monotonicity, and again by the growth condition,

$$\frac{G(a)}{G(a+b)} \le \left|\frac{a}{a+b}\right|^2 \le (1+2\varepsilon)^2 \le 1 + 5\varepsilon,$$

and so $G(a + b) \ge G(a)/(1 + 5\varepsilon)$, and so $G(a + b) \ge (1 - 5\varepsilon)G(a)$.

LEMMA 4.2. *(Approximate Scale Invariance) For all a and $C \ge 1$, $G(Ca) \le C^2 G(a)$.*

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

LEMMA 4.3. *(Perturbation of the weighted M-Estimator) There is a constant $C' > 0$ for which for any e and any w, with $\|e\|_{G,w} \le \varepsilon^5\|y\|_{G,w}$,*

$$\|y + e\|_{G,w} = (1 \pm C'\varepsilon)\|y\|_{G,w}.$$

Proof. By Lemma 4.2, $G(\frac{1}{\varepsilon^2}e_i) \le \frac{1}{\varepsilon^4}G(e_i)$, and so $\|\frac{1}{\varepsilon^2}e\|_{G,w} \le \frac{1}{\varepsilon^4}\|e\|_{G,w} \le \varepsilon\|y\|_{G,w}$, where the final inequality follows by the assumption of the lemma.

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

Now consider an $i \in [n] \setminus S$. In this case $|y_i| \le \varepsilon(\frac{1}{\varepsilon^2}|e_i|)$. Using that G is monotonically non-decreasing and applying Lemma 4.1 again,

$$G(e_i + y_i) \le G(\frac{1}{\varepsilon^2}e_i + y_i) = (1 \pm C\varepsilon)G(e_i/\varepsilon^2),$$

so that

$$\begin{aligned}\sum_{i\in[n]\setminus S} w_i G(y_i + e_i) &\le (1 + C\varepsilon)\|e/\varepsilon^2\|_{G,w} \\ &\le (1 + C\varepsilon)\varepsilon\|y\|_{G,w}.\end{aligned}$$

Again using that G is monotonically non-decreasing, we note that

$$\begin{aligned}\sum_{i\in[n]\setminus S} w_i G(y_i) \le \sum_{i\in[n]\setminus S} w_i G(e_i/\varepsilon) &\le \|e/\varepsilon\|_{G,w} \\ &\le \|e/\varepsilon^2\|_{G,w} \le \varepsilon\|y\|_{G,w}.\end{aligned}$$

Hence,

$$\begin{aligned}&\|y + e\|_{G,w} \\ &= \sum_{i\in S} w_i G(y_i + e_i) + \sum_{i\in[n]\setminus S} w_i G(y_i + e_i) \\ &= (1 \pm C\varepsilon)\sum_{i\in S} w_i G(y_i) \pm (1 + C\varepsilon)\varepsilon\|y\|_{G,w} \\ &= (1 \pm O(\varepsilon))\sum_{i\in[n]} w_i G(y_i) \pm (2 + C\varepsilon)\varepsilon\|y\|_{G,w} \\ &= (1 \pm O(\varepsilon))\|y\|_{G,w}.\end{aligned}$$

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 $x^* = \mathrm{argmin}_x \|Ax - b\|_{G,w}$. Then $c \leq \|y^*\|_2 \leq \kappa c n^{3/2} \|w\|_\infty$, 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, $\|y\|_{G,w} \leq \|w\|_\infty \cdot n \cdot G(c)$ 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, $\|y^*\|_G \geq G(d/\sqrt{n})$. Since y^* is the minimizer for G with weight vector w, necessarily $G(d/\sqrt{n}) \leq \|w\|_\infty \cdot n \cdot G(c)$. If $d/\sqrt{n} \leq c$, the lemma follows. Otherwise, by the lower bound on the growth condition for G, $G(d/\sqrt{n}) \geq G(c) \cdot C_G d/(c\sqrt{n})$, and so $C_G d/(c\sqrt{n}) \leq \|w\|_\infty \cdot n$. Hence, $d \leq \|w\|_\infty n^{3/2} c / C_G$.

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 $|N| \leq (n/\varepsilon)^{C_N d}$ for which if both:*

1. *$\|S(Ax - b)\|_{G,w'} = (1 \pm \varepsilon)\|Ax - b\|_{G,w}$ 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 $\|Ax - b\|_2 \leq \kappa c n^{3/2} \|w\|_\infty$, for an arbitrary constant $\kappa > 0$, it holds that

$$\|S(Ax - b)\|_{G,w'} = (1 \pm \varepsilon)\|Ax - b\|_{G,w}.$$

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 weight vector w', then the following conclusion holds: for all x for which $\|Ax - b\|_2 \leq \kappa c n^{3/2}\|w\|_\infty$, 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_α 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 $\|y - e\|_2 \leq \frac{\varepsilon^5}{n^{2C_S+2}}\alpha$. It is well-known that there exists an N_α for which $|N_\alpha| \leq \left(\frac{3n^{2C_S+2}}{\varepsilon^5}\right)^{d+1}$ [32]. We define

$$N = N_c \cup N_{c(1+\varepsilon)} \cup N_{c(1+\varepsilon)^2} \cup \cdots \cup N_{\kappa c n^{3/2}\|w\|_\infty}.$$

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 $\|y\|_2 \leq \kappa c n^{3/2}\|w\|_\infty$. 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,

$$\begin{aligned}\|S(e-y)\|_{G,w'} &\leq n^{C_S}\|e-y\|_{G,w} \\ &\leq n^{C_S} \cdot \|w\|_\infty \cdot nG(\|e-y\|_2),\end{aligned}$$

using the fact that each coordinate of $e - y$ 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,

$$\begin{aligned}G(\|e-y\|_2) &\leq \frac{\|e-y\|_2}{C_G\|y\|_2}G(\|y\|_2) \\ &= O\left(\frac{\varepsilon^5}{n^{2C_s+2}}\right)G(\|y\|_2).\end{aligned}$$

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, $G(\|y\|_2) \leq nG(\|y\|_2/\sqrt{n})$. Combining these inequalities, we have

$$\begin{aligned}\|S(e-y)\|_{G,w'} &\leq n^{2C_s+1}G(\|e-y\|_2) \\ &= O\left(\frac{\varepsilon^5}{n}\right)G(\|y\|_2) \\ &= O(\varepsilon^5)\|y\|_{G,w}. \qquad (4.13)\end{aligned}$$

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

$$\|y\|_{G,w} = \|e + (y - e)\|_{G,w} = (1 \pm O(\varepsilon))\|e\|_{G,w}. \qquad (4.14)$$

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, $\|Se\|_{G,w'} = (1 \pm \varepsilon)\|e\|_{G,w}$ 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 $\|Sy\|_{G,w'} = (1 \pm O(\varepsilon))\|Se\|_{G,w'}$, and combining this with the assumption of the theorem and (4.14),

$$\|Sy\|_{G,w'} = (1 \pm O(\varepsilon))\|Se\|_{G,w'} = (1 \pm O(\varepsilon))\|e\|_{G,w} = (1 \pm O(\varepsilon))\|y\|_{G,w}.$$

For the second part of the lemma, suppose we only had that for all $e \in N$, $(1-\varepsilon)\|e\|_{G,w} \le \|Se\|_{G,w'}$. 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

$$\|Sy\|_{G,w'} = (1 \pm O(\varepsilon))\|Se\|_{G,w'}.$$

Using (4.14), we have

$$\|Sy\|_{G,w'} = (1 \pm O(\varepsilon))\|Se\|_{G,w'} \ge (1 - O(\varepsilon))\|e\|_{G,w} \ge (1 - O(\varepsilon))\|y\|_{G,w},$$

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 $N = O(C_G^{-2} d^2 m^{6+\alpha})$, putting $m = O(d \log n)$, and $\epsilon = 1/2$. 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 $\|w\|_\infty \le n^{C_s}$ for a constant $C_s > 0$ by definition of the sketch, since $h_{max} \le \log n$. 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, $G(\|L_{h,i}\|_\Lambda) \le n^{O(1)} \sum_{p \in L_{h,i}} G(p)$. By monotonicity of G, we have $G(\|L_{h,i}\|_\Lambda) \le G(\|L_{h,i}\|_1)$. By the growth condition on G, for $a \ge b$ we have

$$\frac{G(a+b)}{G(a)} \le \frac{(a+b)^2}{a^2} \le 2 + \frac{2b^2}{a^2} \le 2 + \frac{2G(b)}{G(a)},$$

and so $G(a+b) \le 2G(a) + 2G(b)$. Applying this inequality recursively $\lceil \log |L_{h,i}| \rceil$ times, we have $G(\|L_{h,i}\|_1) \le n \sum_{p \in L_{h,i}} |y_p|$, 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 $O(n)^{C_N d} C_1^{-m} < 1$, for $m = O(d \log n)$, assuming conditions $\mathcal{E}$ and $\mathcal{E}'_c$.

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

By making the totals of the failure probabilities for conditions $\mathcal{E}$ and $\mathcal{E}'_c$, 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 $\mathtt{nnz}(A)$ time and T has $\mathrm{poly}(d)$ rows.

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

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

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

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

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *J. ACM*, 54(2), 2007.

[3] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC*, pages 557–563, 2006.

[4] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):21:1–21:12, June 2013.

[5] S. Bernstein. *Theory of Probability*. Moscow, 1927.

[6] Jean Bourgain and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. *CoRR*, abs/1311.2542, 2013.

[7] C. Boutsidis and A. Gittens. Improved matrix algorithms via the Subsampled Randomized Hadamard Transform. *ArXiv e-prints*, March 2012.

[8] Vladimir Braverman and Rafail Ostrovsky. Zero-one frequency laws. In *STOC*, pages 281–290, 2010.

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

sity time. In *STOC*, 2013. Full version at `http://arxiv.org/abs/1207.6365`.

[10] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.

[11] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *STOC*, pages 341–350, 2010.

[12] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *SODA*, pages 1117–1126, 2006.

[13] Amit Deshpande and Santosh Vempala. Adaptive sampling and fast low-rank matrix approximation. In *APPROX-RANDOM*, pages 292–303, 2006.

[14] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006.

[15] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1):158–183, 2006.

[16] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM J. Comput.*, 36(1):184–206, 2006.

[17] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *CoRR*, abs/1109.3843, 2011.

[18] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *SODA*, pages 1127–1136, 2006.

[19] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-based methods. In *APPROX-RANDOM*, pages 316–326, 2006.

[20] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In *ESA*, pages 304–314, 2006.

[21] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *CoRR*, abs/0710.1435, 2007.

[22] Antoine Guitton and William W. Symes. Robust and stable velocity analysis using the Huber function, 1999.

[23] S. J. Haberman. Convexity and estimation. *Ann. Statist.*, 17:1631–1661, 1989.

[24] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 1964.

[25] Piotr Indyk and David P. Woodruff. Optimal approximations of the frequency moments of data streams. In *STOC*, pages 202–208, 2005.

[26] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, 1982.

[27] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. In *SODA*, pages 1195–1206, 2012.

[28] Ioannis Koutis, Gary L. Miller, and Richard Peng. Approaching optimality for solving SDD linear systems. In *FOCS*, pages 235–244, 2010.

[29] Ioannis Koutis, Gary L. Miller, and Richard Peng. A nearly-$m \log n$ time solver for SDD linear systems. In *FOCS*, pages 590–598, 2011.

[30] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[31] Olvi L. Mangasarian and David R. Musicant. Robust linear and support vector regression. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(9):950–955, 2000.

[32] Jiri Matousek. *Lectures on Discrete Geometry.* Springer, 2002.

[33] A. Maurer. A bound on the deviation probability for sums of non-negative random variables. *Journal of Inequalities in Pure and Applied Mathematics*, 4(1), 2003.

[34] X. Meng and M. W. Mahoney. Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression. *ArXiv e-prints*, October 2012.

[35] Gary L. Miller and Richard Peng. Iterative approaches to row sampling. *CoRR*, abs/1211.2713, 2012.

[36] Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. *CoRR*, abs/1211.1002, 2012.

[37] Wojciech Niemiro. Asymptotics for m-estimators defined by convex minimization. *Ann. Statist.*, 20(3):1514–1533, 1992.

[38] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006.

[39] D. A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *36th Annual ACM Symposium on Theory of Computing*, pages 81–90. ACM, 2004.

[40] Elad Verbin and Qin Zhang. Rademacher-sketch: A dimensionality-reducing embedding for sum-product norms, with an application to earth-mover distance. In *ICALP (1)*, pages 834–845, 2012.

[41] David P. Woodruff and Qin Zhang. Subspace embeddings and l_p-regression using exponential random variables. *CoRR*, abs/1305.5580, 2013.

[42] Jiyan Yang, Xiangrui Meng, and Michael W. Mahoney. Quantile regression for large-scale applications. In *ICML (3)*, pages 881–887, 2013.

[43] Zhengyou Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing Journal*, 15(1):59–76, 1997. `http://research.microsoft.com/en-us/um/people/zhang/INRIA/Publis/Tutorial-Estim/node24.html`.