Supplementary Material for "Dimensionality Reduction for Tukey Regression"

A. Preliminaries

For two real numbers a and b, we use the notation $a = (1 \pm \varepsilon)b$ if $a \in [(1-\varepsilon)b, (1+\varepsilon)b]$.

We use $\|\cdot\|_p$ to denote the ℓ_p norm of a vector, and $\|\cdot\|_{p,w}$ to denote the weighted ℓ_p norm, i.e.,

$$\|y\|_{p,w} = \left(\sum_{i=1}^{n} w_i |y_i|^p\right)^{1/p}.$$

For a vector $y \in \mathbb{R}^n$, a weight vector $w \in \mathbb{R}^n$ whose entries are all non-negative and a loss function $M : \mathbb{R} \to \mathbb{R}^+$ that satisfies Assumption 1, $\|y\|_{M,w}$ is defined to be

$$\|y\|_{M,w} = \sum_{i=1}^{n} w_i \cdot M(y_i).$$

We also define $\|y\|_M$ to be

$$\|y\|_M = \sum_{i=1}^{n} M(y_i).$$

For a vector $y \in \mathbb{R}^n$ and a real number $\tau \geq 0$, we define H_y to be the set $H_y = \{i \in [n] \mid |y_i| > \tau\}$, and L_y to be the set $L_y = \{i \in [n] \mid |y_i| \leq \tau\}$.

A.1. Tail Inequalities

Lemma A.1 (Bernstein's inequality)**.** *Suppose $X_1, X_2, \ldots, X_n$ are independent random variables taking values in $[-b, b]$. Let $X = \sum_{i=1}^{n} X_i$ and $\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i]$ be the variance of X. For any $t > 0$ we have*

$$\Pr[|X - \mathrm{E}[X]| > t] \leq 2\exp\left(-\frac{t^2}{2\,\mathrm{Var}[X] + 2bt/3}\right).$$

A.2. Facts Regarding the Loss Function

Lemma A.2. *Under Assumption 1, there is a constant $C > 0$ that depends only on p, for which for any a, b with $|b| \leq \varepsilon|a|$, we have $M(a+b) = (1 \pm C\varepsilon)M(a)$.*

Proof. Without loss of generality we assume $a > 0$. When $b \geq 0$, by Assumption 1.3, we have

$$M(a) \leq M(a+b) \leq (1+\varepsilon)^p \cdot M(a) \leq (1 + C\varepsilon)M(a).$$

When $b < 0$, we have

$$M(a) \geq M(a+b) \geq \left(\frac{a}{a+b}\right)^p M(a) \geq (1 - C\varepsilon)M(a).$$

□

Lemma A.3. *Under Assumption 1, there is a constant $C' > 0$ that depends only on p, for which for any $e, y \in \mathbb{R}^n$ and any weight vector w with $\|e\|_{M,w} \leq \varepsilon^{2p+1}\|y\|_{M,w}$,*

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

Proof. Clearly, by Assumption 1.3,

$$\|e/\varepsilon^2\|_{M,w} \leq \varepsilon^{-2p}\|e\|_{M,w} \leq \varepsilon\|y\|_{M,w}.$$

Let $S = \{i \in n \mid |e_i| \le \varepsilon |y_i|\}$. By Lemma A.2, for all $i \in S$ we have $M(y_i + e_i) = (1 \pm C\varepsilon)M(y_i)$. For all $i \in [n] \setminus S$, we have $|e_i| > \varepsilon|y_i|$. For sufficiently small ε, by Assumption 1.2 and Lemma A.2,

$$M(e_i + y_i) \le M(e_i/\varepsilon^2 + y_i) \le (1 + C\varepsilon)M(e_i/\varepsilon^2),$$

which implies

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

Furthermore,

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

Thus,

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

□

A.3. Facts Regarding Lewis Weights

In this section we recall some facts regarding leverage scores and Lewis weights.

Definition A.1. Given a matrix $A \in \mathbb{R}^{n \times d}$. The *leverage score* of a row $A_{i,*}$ is defined to be

$$\tau_i(A) = A_{i,*}(A^T A)^\dagger (A_{i,*})^T.$$

Definition A.2 ((Cohen & Peng, 2015))**.** For a matrix $A \in \mathbb{R}^{n \times d}$, its ℓ_p *Lewis weights* $\{u_i\}_{i=1}^n$ are the *unique weights* such that for each $i \in [n]$ we have

$$u_i = \tau_i(U^{1/2-1/p}A).$$

Here τ_i is the leverage score of the i-th row of a matrix and U is the diagonal matrix formed by putting the elements of u on the diagonal.

Theorem A.4 ((Cohen & Peng, 2015))**.** *There is an algorithm that receives a matrix $A \in \mathbb{R}^{n \times d}$ and outputs $\{\hat{u}\}_{i=1}^n$ such that*

$$u_i \le \hat{u}_i \le 2u_i,$$

where $\{u_i\}_{i=1}^n$ are the ℓ_p Lewis weights of A. Furthermore, the algorithm runs in $\widetilde{O}(\mathrm{nnz}(A) + d^{p/2+O(1)})$ time.

Theorem A.5 (Lewis's change of density (Lewis, 1978), see also (Wojtaszczyk, 1996, p. 113))**.** *Given a matrix $A \in \mathbb{R}^{n \times d}$ and $p \ge 1$, there exists a basis matrix $H \in \mathbb{R}^{n \times d}$ of the column space of A, such that if we define a weight vector $\overline{u} \in \mathbb{R}^n$ where $\overline{u}_i = \|H_{i,*}\|_2$, then the following hold:*

1. $\|\overline{u}\|_p^p \le d$*;*
2. $\overline{U}^{p/2-1}H$ *is an orthonormal matrix.*

Here $\overline{U}$ is the diagonal matrix formed by putting the elements of $\overline{u}$ on the diagonal.

Lemma A.6 (See, e.g., (Wojtaszczyk, 1996, p. 115))**.** *Given a matrix $A \in \mathbb{R}^{n \times d}$, for the basis matrix H and the weight vector $\overline{u}$ defined in Theorem A.5, for all $x \in \mathbb{R}^d$ we have*

$$\|\overline{U}^{p/2-1}Hx\|_2 \le \|Hx\|_p \le d^{1/p-1/2}\|\overline{U}^{p/2-1}Hx\|_2$$

when $1 \le p \le 2$*, and*

$$\|Hx\|_p \le \|\overline{U}^{p/2-1}Hx\|_2 \le d^{1/2-1/p}\|Hx\|_p$$

when $p \ge 2$*.*

Since $\overline{U}^{p/2-1}H$ *is an orthonormal matrix, for all* $x \in \mathbb{R}^d$ *we have*

$$\|x\|_2 \le \|Hx\|_p \le d^{1/p-1/2}\|x\|_2$$

when $1 \le p \le 2$*, and*

$$\|Hx\|_p \le \|x\|_2 \le d^{1/2-1/p}\|Hx\|_p$$

when $p \ge 2$*.*

Lemma A.7. *Given a matrix* $A \in \mathbb{R}^{n\times d}$ *and* $p \ge 1$*, the weight vector* u *defined in Definition A.2 and the weight vector* $\overline{u}$ *defined in Theorem A.5 satisfies*

$$u_i = \overline{u}_i^p.$$

Proof. We show that substituting $u_i = \overline{u}_i^p$ will satisfy

$$u_i = \tau_i(U^{1/2-1/p}A),$$

and thus the theorem follows by the uniqueness of Lewis weights.

Since leverage scores are invariant under change of basis (see, e.g., (Woodruff, 2014, p. 30)), we have

$$\tau_i(U^{1/2-1/p}A) = \tau_i(U^{1/2-1/p}H),$$

where H is the basis matrix defined in Theorem A.5. Substituting $u_i = \overline{u}_i^p$ we have

$$\tau_i(U^{1/2-1/p}A) = \tau_i(\overline{U}^{p/2-1}H).$$

However, since $\overline{U}^{p/2-1}H$ is an orthonormal matrix, and the leverage scores of an orthonormal matrix are just squared ℓ_2 norm of rows (see, e.g., (Woodruff, 2014, p. 29)), we have

$$\tau_i(U^{1/2-1/p}A) = \left(\overline{u}_i^{p/2-1}\|H_{i,*}\|_2\right)^2 = \overline{u}_i^p.$$

□

Lemma A.8. *Given a matrix* $A \in \mathbb{R}^{n\times d}$ *and* $p \ge 1$*, for all* $y \in \mathbf{im}(A)$ *and* $i \in [n]$*, we have*

$$|y_i|^p \le d^{\max\{0,p/2-1\}}u_i \cdot \|y\|_p^p.$$

Here $\{u_i\}_{i=1}^n$ *are the* ℓ_p *Lewis weights defined in Definition A.2.*

Proof. For all $y \in \mathbf{im}(A)$, we can write $y = Hx$ for some vector $x \in \mathbb{R}$ and the basis matrix H in Theorem A.5. By the Cauchy-Schwarz inequality,

$$|y_i|^p = |\langle x, H_{i,*}\rangle|^p \le \|x\|_2^p \cdot \|H_{i,*}\|_2^p,$$

which implies

$$|y_i|^p \le d^{\max\{0,p/2-1\}} \cdot \|y\|_p^p \cdot \|H_{i,*}\|_2^p$$

by Lemma A.6, which again implies

$$|y_i|^p \le d^{\max\{0,p/2-1\}}u_i \cdot \|y\|_p^p$$

since $\overline{u}_i = \|H_{i,*}\|_2$ and $u_i = \overline{u}_i^p$ by Lemma A.7. □

Lemma A.9. *Under Assumption 1, given a matrix $A \in \mathbb{R}^{n\times d}$, $\delta_{\mathsf{lewis}} \in (0,1)$, and a weight vector $w \in \mathbb{R}^n$ such that (i) $w_i \geq 1$ for all $i \in [n]$ and (ii) $\max_{i\in[n]} w_i \leq 2\min_{i\in[n]} w_i$. Let $w' \in \mathbb{R}^n$ be another weight vector which is defined to be*

$$w'_i = \begin{cases} w_i/p_i & \text{with probability } p_i \\ 0 & \text{with probability } 1-p_i \end{cases}$$

and p_i satisfies

$$p_i \geq \min\{1, \Theta(U_M/L_M \cdot d^{\max\{0,p/2-1\}} u_i \cdot \log(1/\delta_{\mathsf{lewis}})/\varepsilon^2)\},$$

then for any fixed vectors $x \in \mathbb{R}^d$ such that $\|Ax\|_\infty \leq \tau$, with probability at least $1-\delta_{\mathsf{lewis}}$ we have

$$\|Ax\|_{M,w} = (1\pm\varepsilon)\|Ax\|_{M,w'}.$$

Proof. Without loss of generality we assume $1 \leq w_i \leq 2$ for all $i \in [n]$. Let $y = Ax$. We use the random variable Z_i to denote

$$Z_i = w'_i M(y_i).$$

Clearly $\mathrm{E}[Z_i] = w_i M(y_i)$, which implies

$$\mathrm{E}[\|y\|_{M,w'}] = \|y\|_{M,w}.$$

Furthermore, $Z_i \leq 2M(y_i)/p_i$. Since $\|y\|_\infty \leq \tau$ and $L_M|y_i|^p \leq M(y_i) \leq U_M|y_i|^p$ when $|y_i| \leq \tau$, by Lemma A.8 we have

$$Z_i \leq 2U_M|y_i|^p/p_i \leq \Theta(L_M \cdot \|y\|_p^p \cdot \varepsilon^2/\log(1/\delta_{\mathsf{lewis}})) \leq \Theta(\|y\|_{M,w} \cdot \varepsilon^2/\log(1/\delta_{\mathsf{lewis}})).$$

Moreover, $\mathrm{E}[Z_i^2] \leq O((M(y_i))^2/p_i)$, which implies

$$\sum_{i=1}^n \mathrm{E}[Z_i^2] \leq O\left(\sum_{i=1}^n (M(y_i))^2/p_i\right).$$

By Hölder's inequality,

$$\sum_{i=1}^n \mathrm{E}[Z_i^2] \leq O(\|y\|_M) \cdot \max_{i\in[n]} M(y_i)/p_i \leq O(\|y\|_{M,w}^2 \cdot \varepsilon^2/\log(1/\delta_{\mathsf{lewis}})).$$

Furthermore, since

$$\mathrm{Var}\left[\sum_{i=1}^n Z_i\right] = \sum_{i=1}^n \mathrm{Var}[Z_i] \leq \sum_{i=1}^n \mathrm{E}[Z_i^2],$$

Bernstein's inequality in Lemma A.1 implies

$$\Pr\left[|\|y\|_{M,w'} - \|y\|_{M,w}| > t\right] \leq \exp\left(-\Theta\left(\frac{t^2}{\|y\|_{M,w} \cdot \varepsilon^2/\log(1/\delta_{\mathsf{lewis}}) \cdot t + \|y\|_{M,w}^2 \cdot \varepsilon^2/\log(1/\delta_{\mathsf{lewis}})}\right)\right).$$

Taking $t = \varepsilon \cdot \|y\|_{M,w}$ implies the desired result. □

Theorem A.10. *Given a matrix $A \in \mathbb{R}^{n\times d}$, $\delta_{\mathsf{subspace}} \in (0,1)$, and a weight vector $w \in \mathbb{R}^n$ such that (i) $w_i \geq 1$ for all $i \in [n]$ and (ii) $\max_{i\in[n]} w_i \leq 2\min_{i\in[n]} w_i$. Let $w' \in \mathbb{R}^n$ be another weight vector which is defined to be*

$$w'_i = \begin{cases} w_i/p_i & \text{with probability } p_i \\ 0 & \text{with probability } 1-p_i \end{cases}$$

and p_i satisfies

$$p_i \geq \min\{1, \Theta(d^{\max\{0,p/2-1\}} u_i \cdot (d\log(1/\varepsilon) + \log(1/\delta_{\mathsf{subspace}}))/\varepsilon^2)\},$$

then with probability at least $1-\delta_{\mathsf{subspace}}$, for all vectors $x \in \mathbb{R}^d$, we have

$$\|Ax\|_{p,w}^p = (1\pm\varepsilon)\|Ax\|_{p,w'}^p.$$

Proof. Let $\mathcal{N}$ be an ε-net for $\{Ax \mid \|Ax\|_{p,w} = 1\}$. Standard facts (see, e.g., (Woodruff, 2014, p. 48)) imply that $\log|\mathcal{N}| \le O(d\log(1/\varepsilon))$. Now we invoke Lemma A.9 with $\delta_{\mathsf{lewis}} = \delta_{\mathsf{subspace}}/|\mathcal{N}|$. Notice that $f(x) = |x|^p$ is also a loss function that satisfies Assumption 1, with $L_M = U_M = 1$ and $\tau = \infty$. Thus, if p_i satisfies

$$p_i \ge \Theta(d^{\max\{0,p/2-1\}} u_i \cdot (d\log(1/\varepsilon) + \log(1/\delta_{\mathsf{subspace}}))/\varepsilon^2),$$

then with probability $1 - \delta_{\mathsf{subspace}}$, simultaneously for all $x \in \mathcal{N}$ we have

$$\|Ax\|_{p,w}^p = (1 \pm \varepsilon)\|Ax\|_{p,w'}^p.$$

Now we can invoke the standard successive approximation argument (see, e.g., (Woodruff, 2014, p. 47)) to show that with probability $1 - \delta_{\mathsf{subspace}}$, simultaneously for all $x \in \mathbb{R}^d$ we have

$$\|Ax\|_{p,w}^p = (1 \pm O(\varepsilon))\|Ax\|_{p,w'}^p.$$

Adjusting constants implies the desired result. □

B. Finding Heavy Coordinates

B.1. A Polynomial Time Algorithm

1. Let $J = \emptyset$.
2. Repeat the following for α times:
   (a) Calculate $\{u_i\}_{i \in [n] \setminus J}$, which are the ℓ_p Lewis weights of the matrix $A_{[n]\setminus J,*}$.
   (b) For each $i \in [n] \setminus J$, if
   $$d^{\max\{0,p/2-1\}} u_i \ge \frac{1}{2\alpha},$$
   then add i into J.

Figure 6. Algorithm for finding the set J.

Theorem B.1. *For a given matrix $A \in \mathbb{R}^{n\times d}$, $\tau \ge 0$ and $p \ge 1$, the algorithm in Figure 6 returns a set of indices $J \subseteq [n]$ with size $|J| \le O(d^{\max\{p/2,1\}} \cdot \alpha^2)$, such that for all $y \in \mathbf{im}(A)$, if y satisfies (i) $\|y_{L_y}\|_p^p \le \alpha \cdot \tau^p$ and (ii) $|H_y| \le \alpha$, then $H_y \subseteq J$.*

Proof. Consider a fixed vector $y \in \mathbf{im}(A)$ that satisfies (i) $\|y_{L_y}\|_p^p \le \alpha \cdot \tau^p$ and (ii) $|H_y| \le \alpha$. For ease of notation, we assume $|y_1| \ge |y_2| \ge \cdots \ge |y_n|$. Of course, this order is unknown and is not used by our algorithm. Under this assumption, $H_y = \{1, 2, \ldots, |H_y|\}$.

We prove $H_y \subseteq J$ by induction. For any $i < |H_y|$, suppose $[i] \subseteq J$ and $i+1 \notin J$ after the i-th repetition of Step 2, we show that we will add $i+1$ into J in the $(i+1)$-th repetition of Step 2. Since, $[i] \subseteq J$ and $|y_1| \ge |y_2| \ge \cdots \ge |y_n|$,

$$\|y_{[n]\setminus J}\|_p^p \le \|y_{L_y}\|_p^p + \alpha|y_{i+1}|^p \le \alpha\tau^p + \alpha|y_{i+1}|^p.$$

Since $i+1 \in H_y$, we must have $|y_{i+1}| \ge \tau$, which implies

$$\frac{|y_{i+1}|^p}{\|y_{[n]\setminus J}\|_p^p} \ge \frac{1}{2\alpha}.$$

By Lemma A.8, this implies

$$d^{\max\{0,p/2-1\}} u_{i+1} \ge \frac{1}{2\alpha},$$

1. Let $|J| = O(d^{\max\{p/2,1\}} \cdot \alpha^2)$ as in Corollary B.2.
2. Repeat the following for $O(\log(|J|/\delta_{\mathsf{struct}}))$ times:
   (a) Randomly partition $[n]$ into $\Gamma_1, \Gamma_2, \ldots, \Gamma_\alpha$.
   (b) For each $j \in [\alpha]$, use the algorithm in Theorem A.2 to obtain weights $\{\hat{u}_i\}_{i \in \Gamma_j}$ such that $u_i \le \hat{u}_i \le 2u_i$, where $\{u_i\}_{i \in \Gamma_j}$ are the ℓ_p Lewis weights of the matrix $A_{\Gamma_j,*}$.
   (c) For each $j \in [\alpha]$, for each $i \in \Gamma_j$, if
   $$d^{\max\{0,p/2-1\}}\hat{u}_i \ge \frac{1}{6},$$
   then add i to I.

Figure 7. Algorithm for finding the set I.

where u_{i+1} is the ℓ_p Lewis weight of the row $A_{i+1,*}$ in $A_{[n]\setminus J,*}$, in which case we will add $i+1$ into J. Thus, $H_y \subseteq J$ since $|H_y| \le \alpha$.

Now we analyze the size of J. For the algorithm in Figure 6, we repeat the whole procedure α times. Each time, an index i will be added into I if and only if

$$d^{\max\{0,p/2-1\}}u_i \ge \frac{1}{2\alpha}.$$

However, since

$$\sum_{i \in [n]\setminus J} u_i = \sum_{i \in [n]\setminus J} \overline{u}_i^p \le d$$

by Theorem A.5, there are at most $O(d^{\max\{p/2,1\}} \cdot \alpha)$ such indices i. Thus, the total size of J is upper bounded by $O(d^{\max\{p/2,1\}} \cdot \alpha^2)$. □

The above algorithm also implies the following existential result.

Corollary B.2. *For a given matrix $A \in \mathbb{R}^{n \times d}$, $\tau \ge 0$ and $p \ge 1$, there exists a set of indices $J \subseteq [n]$ with size $|J| \le O(d^{\max\{p/2,1\}} \cdot \alpha^2)$, such that for all $y \in \mathbf{im}(A)$, if y satisfies (i) $\|y_{L_y}\|_p^p \le \alpha \cdot \tau^p$ and (ii) $|H_y| \le \alpha$, then $H_y \subseteq J$.*

B.2. An Input-sparsity Time Algorithm

To find a set of heavy coordinates, the algorithm in Theorem B.1 runs in polynomial time. In this section we present an algorithm for finding heavy coordinates that runs in input-sparsity time. The algorithm is described in Figure 7.

Theorem B.3. *For a given matrix $A \in \mathbb{R}^{n \times d}$, $\tau \ge 0$, $\delta_{\mathsf{struct}} \in (0,1)$, and $p \ge 1$, the algorithm in Figure 7 returns a set of indices $I \subseteq [n]$ with size $|I| \le \widetilde{O}(d^{\max\{p/2,1\}}\alpha \cdot \log(1/\delta_{\mathsf{struct}}))$, such that with probability at least $1 - \delta_{\mathsf{struct}}$, simultaneously for all $y \in \mathbf{im}(A)$, if y satisfies (i) $\|y_{L_y}\|_p^p \le \alpha \cdot \tau^p$ and (ii) $|H_y| \le \alpha$, then $H_y \subseteq I$. Furthermore, the algorithm runs in $\widetilde{O}\left(\left(\mathrm{nnz}(A) + d^{p/2+O(1)} \cdot \alpha\right) \cdot \log(1/\delta_{\mathsf{struct}})\right)$ time.*

Proof. Let J be the set with size $|J| \le O(d^{\max\{p/2,1\}} \cdot \alpha^2)$ whose existence is proved in Corollary B.2. For all $y \in \mathbf{im}(A)$, if y satisfies (i) $\|y_{L_y}\|_p^p \le \alpha \cdot \tau^p$ and (ii) $|H_y| \le \alpha$, then $H_y \subseteq J$. We only consider those $c \in J$ for which there exists $y \in \mathbf{im}(A)$ such that (i) $\|y_{L_y}\|_p^p \le \alpha \cdot \tau^p$, (ii) $|H_y| \le \alpha$ and (iii) $c \in H_y$, since we can remove other c from J and the properties of J still hold. For such $c \in H_y$ and the corresponding $y \in \mathbf{im}(A)$, suppose for some $j \in [\alpha]$ we have $c \in \Gamma_j$. Since $|H_y| \le \alpha$, with probability $(1 - 1/\alpha)^{|H_y|-1} \ge 1/e$, we have $\Gamma_j \cap H_y = \{c\}$. Furthermore, $\mathrm{E}[\|y_{L_y \cap \Gamma_j}\|_p^p] = \|y_{L_y}\|_p^p/\alpha \le \tau^p$. By Markov's inequality, with probability at least 0.8, we have $\|y_{L_y \cap \Gamma_j}\|_p^p \le 5\tau^p$. Thus, by a union bound, with probability at least $1/e - 0.2 > 0.1$, we have $\|y_{L_y \cap \Gamma_j}\|_p^p \le 5\tau^p$ and $\Gamma_j \cap H_y = \{c\}$. By repeating $O(\log(|J|/\delta_{\mathsf{struct}}))$ times, the success probability is at least $1 - \delta_{\mathsf{struct}}/|J|$. Applying a union bound over all $c \in J$, with probability $1 - \delta_{\mathsf{struct}}$, the stated conditions hold for all $c \in J$. We condition on this event in the rest of the proof.

Consider any $c \in J$ and $y \in \mathbf{im}(A)$ with the properties stated above. Since $|y_c| \geq \tau$, we have

$$\frac{|y_c|^p}{\|y_{\Gamma_j}\|_p^p} \geq \frac{|y_c|^p}{\|y_{\Gamma_j \cap L_y}\|_p^p + |y_c|^p} \geq \frac{1}{6}.$$

By Lemma A.8, we must have

$$d^{\max\{0,p/2-1\}} u_c \geq \frac{1}{6},$$

where u_c is the ℓ_p Lewis weight of the row $A_{c,*}$ in the matrix $A_{\Gamma_j,*}$, which also implies

$$d^{\max\{0,p/2-1\}} \hat{u}_c \geq \frac{1}{6}$$

since $\hat{u}_c \geq u_c$, in which case we will add c to I.

Now we analyze the size of I. For each $j \in [\alpha]$, we have

$$\sum_{i \in \Gamma_j} \hat{u}_i \leq 2 \sum_{i \in \Gamma_j} u_i = 2 \sum_{i \in \Gamma_j} \overline{u}_i^p \leq 2d$$

by Theorem A.5. For each $j \in [\alpha]$, there are at most $O(d^{\max\{p/2,1\}})$ indices i which satisfy

$$d^{\max\{0,p/2-1\}} \hat{u}_i \geq \frac{1}{6},$$

which implies we will add at most $O\left(\alpha \cdot d^{\max\{p/2,1\}}\right)$ elements into I during each repetition. The bound on the size of I follows since there are only $O(\log(|J|/\delta_{\mathsf{struct}})) = O(\log d + \log \alpha + \log(1/\delta_{\mathsf{struct}}))$ repetitions.

For the running time of the algorithm, since we invoke the algorithm in Theorem A.4 for $O(\log(|J|/\delta_{\mathsf{struct}}))$ times, and each time we estimate the ℓ_p Lewis weights of $A_{\Gamma_1,*}, A_{\Gamma_2,*}, \ldots, A_{\Gamma_{|\alpha|},*}$, which implies the running time for each repetition is upper bounded by

$$\sum_{j=1}^{|\alpha|} \widetilde{O}\left(\mathrm{nnz}(A_{\Gamma_j,*}) + d^{p/2+O(1)}\right) = \widetilde{O}\left(\mathrm{nnz}(A) + d^{p/2+O(1)} \cdot \alpha\right).$$

The bound on the running time follows since we repeat for $O(\log(|J|/\delta_{\mathsf{struct}}))$ times. □

The above algorithm and the probabilisitic method also imply the following existential result.

Corollary B.4. *For a given matrix $A \in \mathbb{R}^{n \times d}$, $\tau \geq 0$ and $p \geq 1$, there exists a set of indices $I \subseteq [n]$ with size $|I| \leq \widetilde{O}(d^{\max\{p/2,1\}} \cdot \alpha)$, such that for all $y \in \mathbf{im}(A)$, if y satisfies (i) $\|y_{L_y}\|_p^p \leq \alpha \cdot \tau^p$ and (ii) $|H_y| \leq \alpha$, then $H_y \subseteq I$.*

C. The Net Argument

C.1. Bounding the Norm

We will generally assume that for product Ax, the x involved is in $\mathbf{im}(A^\top)$, which is the orthogonal complement of the nullspace of A; any nullspace component of x would not affect Ax or SAx, and so can be neglected for our purposes.

Lemma C.1. *When the entries of A are integral, for any nonempty $\mathcal{S} \subset [n]$, $\|A^+_{\mathcal{S},*}\|_2 \leq \|A\|_2^d \mathrm{CP}(A)\sqrt{d}$, and under also Assumption 2.2, $\|A^+_{\mathcal{S},*}\|_2 \leq n^{O(d^2)}$.*

Proof. When $\mathcal{S}$ is a nonempty proper subset of $[n]$, then since $\|A_{\mathcal{S},*}\|_2 \leq \|A\|_2$ and $\mathrm{CP}(A_{\mathcal{S},*}) \leq \mathrm{CP}(A)$, we have that if $\|A^+_{\mathcal{S},*}\|_2 \leq \|A_{\mathcal{S},*}\|_2^d \mathrm{CP}(A_{\mathcal{S},*})\sqrt{d}$, then the lemma follows. So we can assume $S = [n]$.

First suppose A has full column rank, so that $A^\top A$ is invertible. For any $y \in \mathbb{R}^n$, $A^+ y$ is the unique solution x^* of $A^\top A x = A^\top y$. Applying Cramer's rule, the entries of x^* have the form $x_i = \frac{\det B_i}{\det A^\top A}$, where B_i is the same as $A^\top A$, except that the i'th column of B_i is $A^\top y$. The integrality of A implies $|\det A^\top A| \geq 1$; using that together with Hadamard's determinant inequality and the definition of the spectral norm, we have $\|x^*\|_2 \le \|A\|_2^d \mathrm{CP}(A)\|y\|_2\sqrt{d}$. Since this holds for any y, we have $\|A^+\|_2 \le \|A\|_2^d \mathrm{CP}(A)\sqrt{d}$ as claimed.

Now suppose A has rank $k < d$. Then there is $\mathcal{T} \subset [d]$ of size k whose members are indices of a set of k linearly independent columns of A. Moreover, if $x^* = A^+ y$ is a solution to $\min_x \|Ax - y\|_2$, then there is another solution where the entries with indices in $[d] \setminus \mathcal{T}$ are zero, since a given column not in $\mathcal{T}$ is a linear combination of columns in $\mathcal{T}$. That is, the solution to $\min_{x\in\mathbb{R}^k} \|A_{*,\mathcal{T}}x - y\|_2$ can be mapped directly to a solution x^* in $\mathbb{R}^k$ with the same Euclidean norm. Since $A_{*,\mathcal{T}}$ has full column rank, the analysis above implies that

$$\|x^*\|_2 \le \|A_{*,\mathcal{T}}\|_2^k \mathrm{CP}(A_{*,\mathcal{T}})\|y\|_2\sqrt{k} \le \|A\|_2^d \mathrm{CP}(A)\|y\|_2\sqrt{d},$$

so the bound on $\|A^+\|_2$ holds also when A has less than full rank.

The last statement of the lemma follows directly, using the definitions of $\|A\|_2$, $\mathrm{CP}(A)$, and Assumption 2.2. □

Lemma C.2. *If A has integral entries, and if Assumptions 1, 2.2, 2.3 hold, then Assumption 2.1 holds.*

Proof. Let $x_M^{C_1}$ be a C_1-approximate solution of $\min_x \|Ax - b\|_M$, which Assumption 2.1 requires to have bounded Euclidean norm. Let $\hat{M}(a) \equiv \min\{\tau^p, |a|^p\}$, so that Assumptions 1.4 and 1.5 imply that $L_M\hat{M}(a) \le M(a) \le U_M\hat{M}(a)$ for all a. Letting $x_M^* \equiv \operatorname{argmin}_x \|Ax - b\|_M$, and similarly defining $x_{\hat{M}}^*$, this condition implies that

$$\begin{aligned} \|Ax_M^{C_1} - b\|_{\hat{M}} &\le \frac{1}{L_M}\|Ax_M^{C_1} - b\|_M \\ &\le \frac{C_1}{L_M}\|Ax_M^* - b\|_M \\ &\le C_2\|Ax_M^* - b\|_{\hat{M}} \\ &\le C_2\|Ax_{\hat{M}}^* - b\|_{\hat{M}}, \end{aligned} \tag{3}$$

where $C_2 \equiv C_1 U_M/L_M$.

Let $\mathcal{S}$ denote the set of indices at which $|A_{i,*}x_M^{C_1} - b_i| \le \tau$. If $\mathcal{S}$ is empty, then $x_M^{C_1}$ can be assumed to be zero.

Similarly to our general assumption that $x_M^{C_1} \in \mathbf{im}(A^\top)$, we can assume that $x_M^{C_1} \in \mathbf{im}(A_{\mathcal{S},*}^\top)$, since any component of $x_M^{C_1}$ in the nullspace of $A_{\mathcal{S},*}$ can be removed without changing $A_{\mathcal{S},*}x_M^{C_1}$, and without increasing the $n - |S|$ contributions of τ^p from the remaining summands in $\|Ax_M^{C_1} - b\|_M$. (Here we used Assumption 1.5 that $M(a) = \tau^p$ for $|a| \ge \tau$.)

From $x_M^{C_1} \in \mathbf{im}(A^\top)$ it follows that $\|x_M^{C_1}\|_2 = \|A_{\mathcal{S},*}^+ A_{\mathcal{S},*}x_M^{C_1}\|_2 \le \|A_{\mathcal{S},*}^+\|_2 \|A_{\mathcal{S},*}x_M^{C_1}\|_2$, and since

$$\begin{aligned} \|A_{\mathcal{S},*}x_M^{C_1}\|_2 &\le \sqrt{n}\|A_{\mathcal{S},*}x_M^{C_1}\|_p \\ &\le \sqrt{n}(\|A_{\mathcal{S},*}x_M^{C_1} - b_S\|_p + \|b_S\|_p) \\ &\le C_2\sqrt{n}(\|Ax_{\hat{M}}^* - b\|_{\hat{M}}^{1/p} + \|b_S\|_p) \qquad \text{(by (3))} \\ &\le 2C_2\sqrt{n}\|b\|_p, \end{aligned}$$

we have $\|x_M^{C_1}\|_2 \le \|A_{\mathcal{S},*}^+\|_2 \|A_{\mathcal{S},*}x_M^{C_1}\|_2 \le \|A_{\mathcal{S},*}^+\|_2 2C_2\sqrt{n}\|b\|_p$, and so from Lemma C.1 and Assumption 2.2, the bound on $\|x_M^{C_1}\|_2$ of Assumption 2.1 follows. □

C.2. Net Constructions

Lemma C.3. *Under the given assumptions, for U as in Assumption 2.1, there exists a set $\mathcal{N}_\varepsilon \subseteq \mathbf{im}([A\ b])$ with size $|\mathcal{N}_\varepsilon| \le n^{O(d^3)} \cdot (1/\varepsilon)^{O(d)}$, such that for any x satisfying $\|x\|_2 \le U$, there exists $y' \in \mathcal{N}_\varepsilon$ such that*

$$\|(Ax - b) - y'\|_M \le \varepsilon^p.$$

Proof. Let $\hat{M}(a) \equiv \min\{\tau^p, |a|^p\}$. Assume for now that $\varepsilon \le \tau/2$, so that if $\|Ax\|_{\hat{M}} \le \varepsilon^p$, then every entry of Ax is no more than τ in magnitude, and so $\|Ax\|_{\hat{M}} = \|Ax\|_p^p$.

Let

$$B_\varepsilon \equiv \{Ax - b \mid \|Ax - b\|_{\hat{M}} \le \varepsilon^p\} = \{Ax - b \mid \|Ax - b\|_p \le \varepsilon\}$$

and

$$B_U \equiv \{Ax - b \mid \|x\|_2 \le U\} \subseteq \{Ax - b \mid \|Ax - b\|_p \le \sqrt{n} \cdot (\|A\|_2 U + \|b\|_2)\}.$$

From the scale invariance of the ℓ_p norm, and the volume in at-most d dimensions, $\mathrm{Vol}(B_\varepsilon) \ge (\varepsilon/(\sqrt{n} \cdot (\|A\|_2 U + \|b\|_2)))^d \,\mathrm{Vol}(B_U)$, so that at most $(\sqrt{n} \cdot (\|A\|_2 U + \|b\|_2)/\varepsilon)^d$ translates of B_ε can be packed into B_U without intersecting. Thus the set $\mathcal{N}_\varepsilon$ of centers of such a maximal packing of translates is an ε^p-cover of B_U, that is, for any point $y \in B_U$, there is some $y' \in \mathcal{N}$ such that $\|y' - y\|_p \le \varepsilon$, so that $\|y' - y\|_{\hat{M}} \le \varepsilon^p$.

If $\varepsilon > \tau/2$, we just note that a $(\tau/2)^p$-cover is also an ε^p-cover, and so there is an ε^p-cover of size $(\sqrt{n} \cdot (\|A\|_2 U + \|b\|_2)/\min\{\tau/2, \varepsilon\})^d$.

Plugging in the bounds for U from Assumption 2.1, and for τ, $\|b\|_2$, and $\|A\|_2 \le \max_{i\in[d]} \|A_{*,i}\|_2$ from Assumptions 2.2 and 2.3, the cardinality bound of the lemma follows.

This argument is readily adapted to more general $\|\cdot\|_M$, by noticing that $\|y - y'\|_M \le U_M \cdot \|y - y'\|_{\hat{M}}$ using Assumption 1.4 and adjusting constants. □

Lemma C.4. *Under the given assumptions, there exists a set $\mathcal{M}_\varepsilon^{\alpha,\beta} \subseteq \mathbf{im}([A\ b])$ with size $|\mathcal{M}_\varepsilon^{\alpha,\beta}| \le O\left(\frac{\beta/\alpha}{\varepsilon}\right) \cdot n^{O(d^2)} \cdot (1/\varepsilon)^{O(d)}$, such that for any x satisfying $\alpha \le \|Ax - b\|_p \le \beta \le \tau$, there exists $y' \in \mathcal{M}_\varepsilon^{\alpha,\beta}$ such that*

$$\|(Ax - b) - y'\|_M \le \varepsilon^p \cdot \|Ax - b\|_M.$$

Proof. We assume $\varepsilon \le \tau$, since otherwise we can take ε to be τ. By standard constructions (see, e.g., (Woodruff, 2014, p. 48)), there exists a set $\mathcal{M}_\gamma \subseteq \mathbf{im}([A\ b])$ with size $|\mathcal{M}_\gamma| \le (1/\varepsilon)^{O(d)}$, such that for any $y = Ax - b$ with $\|y\|_p = \gamma$, there exists $y' \in \mathcal{M}_\gamma$ such that $\|y - y'\|_p \le \gamma \cdot \varepsilon$.

Let $\mathcal{M}_\varepsilon^{\alpha,\beta}$ be

$$\mathcal{M}_\varepsilon^{\alpha,\beta} = \mathcal{M}_\alpha \cup \mathcal{M}_{(1+\varepsilon)\alpha} \cup \mathcal{M}_{(1+\varepsilon)^2\alpha} \cup \cdots \cup \mathcal{M}_\beta.$$

Clearly, by Assumption 2,

$$|\mathcal{M}_\varepsilon^{\alpha,\beta}| \le \log_{1+\varepsilon}(\beta/\alpha) \cdot n^{O(d^2)} \cdot (1/\varepsilon)^{O(d)} \le O\left(\frac{\beta/\alpha}{\varepsilon}\right) \cdot n^{O(d^2)} \cdot (1/\varepsilon)^{O(d)}.$$

Now we show that $\mathcal{M}_\varepsilon^{\alpha,\beta}$ satisfies the desired properties. For any $x \in \mathbb{R}^d$ such that $y = Ax - b$ satisfies $\alpha \le \|y\|_p \le \beta \le \tau$, we must have $|y_i| \le \tau$ for all entries of y. By normalization, there exists $\hat{y}$ such that $\|y - \hat{y}\|_p \le \varepsilon \cdot \|y\|_p$ and $\|\hat{y}\|_p = (1+\varepsilon)^i \cdot \alpha$ for some $i \in \mathbb{N}$. Furthermore, by the property of $\mathcal{M}_{(1+\varepsilon)^i\alpha}$, there exists $y' \in \mathcal{M}_{(1+\varepsilon)^i\alpha} \subseteq \mathcal{M}_\varepsilon^{\alpha,\beta}$ such that $\|\hat{y} - y'\|_p \le \varepsilon \cdot \|y'\|_p \le 2\varepsilon \cdot \|y\|_p$. Thus, by triangle inequality, we have $\|y - y'\|_p \le 3\varepsilon\|y\|_p$. For sufficiently small ε, since $\|y\|_p \le \tau$, we also have $\|y - y'\|_p \le \tau$, which implies $\|y - y'\|_\infty \le \tau$. Thus, using Assumption 1.4, we have

$$\|y - y'\|_M \le U_M \|y - y'\|_p^p \le U_M \cdot (3\varepsilon)^p \cdot \|y\|_p^p \le U_M/L_M (3\varepsilon)^p \|y\|_M.$$

Adjusting constants implies the desired properties. □

C.3. The Net Argument

Theorem C.5. *For any $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, given a matrix $S \in \mathbb{R}^{r\times n}$ and a weight vector $w \in \mathbb{R}^n$ such that $w_i \ge 0$ for all $i \in [n]$. Let $c = \min_x \|Ax - b\|_p$. If there exist $U_O, U_A, L_A, L_N \le \mathrm{poly}(n)$ such that*

1. $\|S(Ax_M^* - b)\|_{M,w} \le U_O \|Ax_M^* - b\|_M$*, where* $x_M^* = \mathrm{argmin}_x \|Ax - b\|_M$*;*
2. $L_A \|Ax - b\|_M \le \|S(Ax - b)\|_{M,w} \le U_A \|Ax - b\|_M$ *for all* $x \in \mathbb{R}^d$*;*
3. $\|Sy\|_{M,w} \ge L_N \|y\|_M$ *for all* $y \in \mathcal{N}_{\mathrm{poly}(\varepsilon\cdot\tau/n)} \cup \mathcal{M}^{c,c\cdot\mathrm{poly}(n)}_{\mathrm{poly}(\varepsilon/n)}$,

then, any C-approximate solution of $\min_x \|S(Ax-b)\|_{M,w}$ *with* $C \le \mathrm{poly}(n)$ *is a* $C \cdot (1+O(\varepsilon)) \cdot U_O/L_N$*-approximate solution of* $\min_x \|Ax-b\|_M$*. Here* $\mathcal{N}_{\mathrm{poly}(\varepsilon\cdot\tau/n)}$ *and* $\mathcal{M}^{c,c\cdot\mathrm{poly}(n)}_{\mathrm{poly}(\varepsilon/n)}$ *are as defined in Lemma C.3 and Lemma C.4, respectively.*

Proof. We distinguish two cases in the proof.

Case 1: $(C \cdot U_M \cdot U_A/(L_M \cdot L_A)) \cdot c^p \le \tau^p$**.** In this case, we prove that any C-approximate solution $x^C_{S,M,w}$ of $\min_x \|S(Ax-b)\|_{M,w}$ satisfies $c \le \|Ax^C_{S,M,w}-b\|_p \le (C \cdot U_M \cdot U_A/(L_M \cdot L_A))^{1/p} \cdot c \le \tau$. Let $x^*_p = \mathrm{argmin}_x \|Ax-b\|_p$, we have

$$\begin{aligned}
&\|Ax^C_{S,M,w}-b\|_M \\
\le&\|S(Ax^C_{S,M,w}-b)\|_{M,w}/L_A \\
\le&C \cdot \|S(Ax^*_p-b)\|_{M,w}/L_A \\
\le&C \cdot \|Ax^*_p-b\|_M \cdot U_A/L_A \\
\le&C \cdot \|Ax^*_p-b\|_p^p \cdot (U_M \cdot U_A)/L_A \\
=&C \cdot c^p \cdot (U_M \cdot U_A)/L_A.
\end{aligned}$$

Since $L_M \le 1$, this implies $\|Ax^C_{S,M,w}-b\|_M \le \tau^p$, which implies $\|Ax^C_{S,M,w}-b\|_\infty \le \tau$. Thus, $\|Ax^C_{S,M,w}-b\|_p^p \le \|Ax^C_{S,M,w}-b\|_M/L_M \le (C \cdot U_M \cdot U_A/(L_M \cdot L_A)) \cdot c^p$, which implies $\|Ax^C_{S,M,w}-b\|_p \le (C \cdot U_M \cdot U_A/(L_M \cdot L_A))^{1/p} \cdot c$. Moreover, by the definition of c we have $\|Ax^C_{S,M,w}-b\|_p \ge c$.

Since $(C \cdot U_M \cdot U_A/(L_M \cdot L_A))^{1/p} \le \mathrm{poly}(n)$, by Lemma C.4, there exists $y' \in \mathcal{M}^{c,c\cdot\mathrm{poly}(n)}_{\mathrm{poly}(\varepsilon/n)}$ such that $\|(Ax^C_{S,M,w}-b)-y'\|_M \le \mathrm{poly}(\varepsilon/n) \cdot \|Ax^C_{S,M,w}-b\|_M$. Notice that

$$\|S(Ax^C_{S,M,w}-b)\|_{M,w} = \|Sy' + S((Ax^C_{S,M,w}-b)-y')\|_{M,w}.$$

For Sy', since $y' \in \mathcal{M}^{c,c\cdot\mathrm{poly}(n)}_{\mathrm{poly}(\varepsilon/n)}$, we have

$$\|Sy'\|_{M,w} \ge L_N\|y'\|_M = L_N\|Ax^C_{S,M,w}-b+(y'-(Ax^C_{S,M,w}-b))\|_M.$$

Since $\|y'-(Ax^C_{S,M,w}-b)\|_M \le \mathrm{poly}(\varepsilon/n) \cdot \|Ax^C_{S,M,w}-b\|_M$, by Lemma A.3, we have $\|Ax^C_{S,M,w}-b+(y'-(Ax^C_{S,M,w}-b))\|_M \ge (1-\varepsilon)\|Ax^C_{S,M,w}-b\|_M$, which implies $\|Sy'\|_{M,w} \ge L_N(1-\varepsilon)\|Ax^C_{S,M,w}-b\|_M$. On the other hand, $\|S((Ax^C_{S,M,w}-b)-y')\|_{M,w} \le U_A\|(Ax^C_{S,M,w}-b)-y'\|_M \le \mathrm{poly}(\varepsilon/n) \cdot \|Ax^C_{S,M,w}-b\|_M$. Again by Lemma A.3, we have $\|S(Ax^C_{S,M,w}-b)\|_{M,w} \ge (1-\varepsilon)\|Sy'\|_{M,w} \ge L_N(1-O(\varepsilon))\|Ax^C_{S,M,w}-b\|_M$. Furthermore, since $x^C_{S,M,w}$ is a C-approximate solution of $\min_x \|S(Ax-b)\|_{M,w}$, we must have

$$\begin{aligned}
\|Ax^C_{S,M,w}-b\|_M &\le (1+O(\varepsilon))/L_N \cdot \|S(Ax^C_{S,M,w}-b)\|_{M,w} \\
&\le C \cdot (1+O(\varepsilon))/L_N \cdot \|S(Ax^*_M-b)\|_{M,w} \\
&\le C \cdot (1+O(\varepsilon)) \cdot U_O/L_N \cdot \|Ax^*_M-b\|_M.
\end{aligned}$$

Case 2: $(C \cdot U_M \cdot U_A/(L_M \cdot L_A)) \cdot c^p \ge \tau^p$**.** In this case, we first prove that any C-approximate solution $x^C_{S,M,w}$ of $\min_x \|S(Ax-b)\|_{M,w}$ is a $\mathrm{poly}(n)$-approximate solution of $\min_x \|Ax-b\|_M$. By Assumption 2.1, this implies all C-approximate solution $x^C_{S,M,w}$ of $\min_x \|S(Ax-b)\|_{M,w}$ satisfies $\|x^C_{S,M,w}\|_2 \le U$.

Consider any C-approximate solution $x^C_{S,M,w}$ of $\min_x \|S(Ax-b)\|_{M,w}$, we have

$$\begin{aligned}
\|Ax^C_{S,M,w}-b\|_M \le \|S(Ax^C_{S,M,w}-b)\|_{M,w}/L_A &\le C \cdot \|S(Ax^*_M-b)\|_{M,w}/L_A \\
&\le C \cdot U_A/L_A \cdot \|Ax^*_M-b\|_M \le \mathrm{poly}(n) \cdot \|Ax^*_M-b\|_M.
\end{aligned}$$

We further show that $\|Ax-b\|_M \ge \tau^p/\mathrm{poly}(n)$ for all $x \in \mathbb{R}^d$. If $\|Ax-b\|_\infty \ge \tau$, then the statement clearly holds. Otherwise, $\|Ax-b\|_M \ge L_M \cdot \|Ax-b\|_p^p \ge L_M c^p \ge L_M^2 L_A/(C \cdot U_M \cdot U_A) \cdot \tau^p \ge \tau^p/\mathrm{poly}(n)$. Thus, for any C-approximate solution $x^C_{S,M,w}$ of $\min_x \|S(Ax-b)\|_{M,w}$, there exists $y' \in \mathcal{N}_{\mathrm{poly}(\varepsilon\cdot\tau/n)}$ such that

$$\|y'-(Ax^C_{S,M,w}-b)\|_M \le \mathrm{poly}(\varepsilon \cdot \tau/n) \le \mathrm{poly}(\varepsilon/n) \cdot \|Ax^C_{S,M,w}-b\|_M.$$

The rest of the proof is exactly the same as that of Case 1. $\square$

D. A Row Sampling Algorithm for Tukey Loss Functions

In this section we present the row sampling algorithm. The row sampling algorithm proceeds in a recursive manner. We describe a single recursive step in Section D.1 and the overall algorithm in Section D.2.

D.1. One Recursive Step

The goal of this section is to design one recursive step of the row sampling algorithm. For a weight vector $w \in \mathbb{R}^n$, the recursive step outputs a sparser weight vector $w' \in \mathbb{R}^n$ such that for any set $\mathcal{N} \subseteq \mathbf{im}(A)$ with size $|\mathcal{N}|$, with probability at least $1 - \delta_{\mathsf{o}}$, simultaneously for all $y \in \mathcal{N}$,

$$\|y\|_{M,w'} = (1 \pm \varepsilon)\|y\|_{M,w}.$$

We maintain that if $w_i \neq 0$, then $w_i \geq 1$ and $\|w\|_\infty \leq n^2$ as an invariant in the recursion. These conditions imply that we can partition the positive coordinates of w into $2\log n$ groups P_j, for which $P_j = \{i \mid 2^{j-1} \leq w_i < 2^j\}$.

Now we define one recursive step of our sampling procedure. We split the matrix A into $A_{P_1,*}, A_{P_2,*}, \ldots, A_{P_{2\log n},*}$, and deal with each of them separately. For each $1 \leq j \leq 2\log n$, we invoke the algorithm in Theorem B.3 to identify a set I_j for the matrix $A_{P_j,*}$, for some parameter α and δ_{struct} to be determined. For each $1 \leq j \leq 2\log n$, we also use the algorithm in Theorem A.4 to calculate $\{\hat{u}_i\}_{i \in P_j}$ such that $u_i \leq \hat{u}_i \leq 2u_i$ where $\{u_i\}_{i\in P_j}$ are the ℓ_p Lewis weights of the matrix $A_{P_j,*}$.

Now for each $i \in P_j$, we define its sampling probability p_i to be

$$p_i = \begin{cases} 1 & i \in I_j \\ \min\{1, 1/2 + \Theta(d^{\max\{0,p/2-1\}}\hat{u}_i \cdot Y)\} & i \notin I_j \end{cases},$$

where $Y \equiv d\log(1/\varepsilon) + \log(\log n/\delta_{\mathsf{o}}) + U_M/L_M \log(|\mathcal{N}| \cdot \log n/\delta_{\mathsf{o}})/\varepsilon^2$.

For each $i \in [n]$, we set $w'_i = 0$ with probability $1 - p_i$, and set $w'_i = w_i/p_i$ with probabliity p_i. The finishes the definition of one step of the sampling procedure.

Let

$$F \equiv \sum_{1\leq j\leq 2\log n} |I_j| + \sum_{1\leq j \leq 2\log n} \sum_{i \in P_j \setminus I_j} \Theta(d^{\max\{0,p/2-1\}}\hat{u}_i \cdot Y).$$

Our first lemma shows that with probability at least $1 - \delta_{\mathsf{o}}$, the number of non-zero entries in w' is at most $\frac{2}{3}\|w\|_0$, provided $\|w\|_0$ is large enough.

Lemma D.1. *When $\|w\|_0 \geq 10F$, with probability at least $1 - \delta_{\mathsf{o}}$,*

$$\|w'\|_0 \leq \frac{2}{3}\|w\|_0.$$

Proof. Notice that

$$\mathrm{E}[\|w'\|_0] \leq \|w\|_0/2 + F.$$

By Bernstein's inequality in Lemma A.1, since $F \geq \Omega(\log(1/\delta_{\mathsf{o}}))$, with probability at least $1 - \exp(-\Omega(\|w\|_0)) \geq 1 - \exp(-\Omega(F)) \geq 1 - \delta_{\mathsf{o}}$, we have

$$\|w'\|_0 \leq \|w\|_0/2 + F + \|w\|_0/10 \leq \frac{2}{3}\|w\|_0.$$

□

Our second lemma shows that $\|w'\|_\infty$ is upper bounded by $2\|w\|_\infty$.

Lemma D.2. $\|w'\|_\infty \leq 2\|w\|_\infty$.

Proof. Since $p_i \geq 1/2$ for all $i \in [n]$, we have $\|w'\|_\infty \leq 2\|w\|_\infty$. □

We show that for sufficiently large constant C, if we set

$$\alpha = C \cdot U_M / L_M \cdot \log(|\mathcal{N}| \cdot \log n / \delta_{\mathsf{o}}) / \varepsilon^2$$

and $\delta_{\mathsf{struct}} = \delta_{\mathsf{o}} / (4 \log n)$, then with probability at least $1 - \delta_{\mathsf{o}}$, simultaneously for all $y \in \mathcal{N}$ we have

$$\|y\|_{M,w'} = (1 \pm \varepsilon) \|y\|_{M,w}.$$

By Theorem B.3 and Theorem A.5, since

$$\sum_{1 \le j \le 2 \log n} \sum_{i \in P_j \setminus I_j} \hat{u}_i \le O(d \log n),$$

this also implies

$$F = \widetilde{O}(d^{\max\{1,p/2\}} \log n \cdot (\log(|\mathcal{N}|/\delta_{\mathsf{o}}) \cdot \log(1/\delta_{\mathsf{o}}) + d)/\varepsilon^2).$$

Furthermore, for each $1 \le j \le 2 \log n$, we invoke the algorithm in Theorem A.4 and the algorithm in Theorem B.3 on $A_{P_1,*}, A_{P_2,*}, \ldots, A_{P_{2 \log n},*}$, and thus the running time of each recursive step is thus upper bounded by

$$\widetilde{O}((\mathrm{nnz}(A) + d^{p/2+O(1)} \cdot \alpha) \cdot \log(1/\delta_{\mathsf{struct}})) = \widetilde{O}((\mathrm{nnz}(A) + d^{p/2+O(1)} \cdot \log(|\mathcal{N}|/\delta_{\mathsf{o}}) \cdot /\varepsilon^2) \cdot \log(1/\delta_{\mathsf{o}})).$$

Now we consider a fixed vector $y \in \mathbf{im}(A)$. We use the following two lemmas in our analysis.

Lemma D.3. *With probability* $1 - \delta_{\mathsf{o}}/O(|\mathcal{N}| \cdot \log n)$*, the following holds:*

- *If* $\|y_{H_y \cap P_j}\|_{M,w} \ge C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\mathsf{o}})/\varepsilon^2$*, then*

$$\|y_{H_y \cap P_j}\|_{M,w'} = (1 \pm \varepsilon/2) \|y_{H_y \cap P_j}\|_{M,w};$$

- *If* $\|y_{H_y \cap P_j}\|_{M,w} < C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\mathsf{o}})/\varepsilon^2$*, then*

$$|\|y_{H_y \cap P_j}\|_{M,w'} - \|y_{H_y \cap P_j}\|_{M,w}| \le C \cdot U_M \cdot \tau^p \cdot 2^{j-2} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\mathsf{o}})/\varepsilon.$$

Proof. For each $i \in H_y \cap P_j$, we use Z_i to denote the random variable

$$Z_i = \begin{cases} w_i M(y_i)/p_i & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}.$$

Since $Z_i = w'_i M(y_i)$, we have

$$\|y_{H_y \cap P_j}\|_{M,w'} = \sum_{i \in H_y \cap P_j} Z_i.$$

It is clear that $Z_i \le 2^{j+1} \cdot U_M \cdot \tau^p$ since $p_i \ge 1/2$ and $w_i \le 2^j$, $\mathrm{E}[Z_i] = w_i M(y_i)$ and $\mathrm{E}[Z_i^2] = w_i^2 (M(y_i))^2/p_i$. By Hölder's inequality,

$$\sum_{i \in H_y \cap P_j} \mathrm{E}[Z_i^2] \le 2^{j+1} \cdot \|y_{H_y \cap P_j}\|_{M,w} \cdot U_M \cdot \tau^p.$$

Thus by Bernstein's inequality in Lemma A.1, we have

$$\Pr\left[\left|\sum_{i \in H_y \cap P_j} Z_i - \|y_{H_y \cap P_j}\|_{M,w}\right| \ge t\right] \le 2 \exp\left(-\frac{t^2}{2^{j+2} \cdot U_M \cdot \tau^p \cdot t/3 + 2^{j+2} \cdot \|y_{H_y \cap P_j}\|_{M,w} \cdot U_M \cdot \tau^p}\right).$$

When

$$\|y_{H_y \cap P_j}\|_{M,w} \ge C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\mathsf{o}})/\varepsilon^2,$$

we take

$$t = \varepsilon/2 \cdot \|y_{H_y \cap P_j}\|_{M,w} \ge C \cdot U_M \cdot \tau^p \cdot 2^{j-2} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\mathsf{o}})/\varepsilon.$$

By taking C to be some sufficiently large constant, with probability at least $1 - \delta_{\circ}/O(|\mathcal{N}| \cdot \log n)$,

$$\|y_{H_y \cap P_j}\|_{M,w'} = (1 \pm \varepsilon/2)\|y_{H_y \cap P_j}\|_{M,w}.$$

When

$$\|y_{H_y \cap P_j}\|_{M,w} < C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2,$$

we take

$$t = C \cdot U_M \cdot \tau^p \cdot 2^{j-2} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon.$$

By taking C to be some sufficiently large constant, with probability at least $1 - \delta_{\circ}/O(|\mathcal{N}| \cdot \log n)$,

$$|\|y_{H_y \cap P_j}\|_{M,w'} - \|y_{H_y \cap P_j}\|_{M,w}| \le C \cdot U_M \cdot \tau^p \cdot 2^{j-2} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon.$$

□

The proof of the following lemma is exactly the same as Lemma D.3.

Lemma D.4. *With probability $1 - \delta_{\circ}/O(|\mathcal{N}| \cdot \log n)$, the following holds:*

- *If $\|y_{L_y \cap P_j}\|_{M,w} \ge C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2$, then*
$$\|y_{L_y \cap P_j}\|_{M,w'} = (1 \pm \varepsilon/2)\|y_{L_y \cap P_j}\|_{M,w};$$
- *If $\|y_{L_y \cap P_j}\|_{M,w} < C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2$, then*
$$|\|y_{L_y \cap P_j}\|_{M,w'} - \|y_{L_y \cap P_j}\|_{M,w}| \le C \cdot U_M \cdot \tau^p \cdot 2^{j-2} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon.$$

Now we use Lemma D.3 and Lemma D.4 to analyze the sampling procedure.

Lemma D.5. *If we set $\alpha = C \cdot U_M/L_M \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2$, $\delta_{\mathsf{struct}} = \delta_{\circ}/(4\log n)$, then for each $1 \le j \le 2\log n$, with probability at least $1 - \delta_{\circ}/(2\log n)$, simultaneously for all $y \in \mathcal{N}$,*

$$\|y_{P_j}\|_{M,w'} = (1 \pm \varepsilon)\|y_{P_j}\|_{M,w}.$$

Applying a union bound over all $1 \le j \le 2\log n$, with probability at least $1 - \delta_{\circ}$, simultaneously for all $y \in \mathcal{N}$,

$$\|y\|_{M,w'} = (1 \pm \varepsilon)\|y\|_{M,w}.$$

Proof. By Theorem B.3, for each $1 \le j \le 2\log n$, with probability $1 - \delta_{\circ}/(4\log n)$, simultaneously for all $y \in \mathcal{N} \subseteq \mathbf{im}(A)$, if y satisfies (i) $\|y_{L_y \cap P_j}\|_p^p \le \alpha \cdot \tau^p$ and (ii) $|H_y \cap P_j| \le \alpha$, then we have $H_y \cap P_j \subseteq I_j$. We condition on this event in the remaining part of the proof.

Now we consider a fixed $y \in \mathcal{N}$. We show that $\|y_{P_j}\|_{M,w'} = (1 \pm \varepsilon)\|y_{P_j}\|_{M,w}$ with probability at least $1 - \delta_{\circ}/O(|\mathcal{N}| \cdot \log n)$. The desired bound follows by applying a union bound over all $y \in \mathcal{N}$.

We distinguish four cases in our analysis. We use T to denote a fixed threshold

$$T = C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2.$$

Case (i): $\|y_{H_y \cap P_j}\|_{M,w} < T$ and $\|y_{L_y \cap P_j}\|_{M,w} < T$. Since $\|y_{H_y \cap P_j}\|_{M,w} < T$, we must have

$$|H_y \cap P_j| < C \cdot U_M/L_M \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2 = \alpha.$$

Furthermore, we also have

$$\|y_{L_y \cap P_j}\|_p^p < C \cdot U_M/L_M \cdot \tau^p \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2 = \alpha \cdot \tau^p.$$

By Lemma A.9, with probability at least $1 - \delta_{\circ}/O(|\mathcal{N}| \cdot \log n)$, we have

$$\|y_{P_j \setminus I_j}\|_{M,w'} = (1 \pm \varepsilon)\|y_{P_j \setminus I_j}\|_{M,w},$$

since $H_y \cap P_j \subseteq I_j$. Moreover, $\|y_{I_j}\|_{M,w} = \|y_{I_j}\|_{M,w'}$ since $w_i = w'_i$ for all $i \in I_j$. Thus, we have $\|y_{P_j}\|_{M,w'} = (1 \pm \varepsilon)\|y_{P_j}\|_{M,w}$.

Case (ii): $\|y_{H_y \cap P_j}\|_{M,w} \geq T$ and $\|y_{L_y \cap P_j}\|_{M,w} \geq T$. By Lemma D.3 and Lemma D.4, with probability at least $1 - \delta_{\circ}/O(|\mathcal{N}| \cdot \log n)$,

$$\|y_{H_y \cap P_j}\|_{M,w'} = (1 \pm \varepsilon/2)\|y_{H_y \cap P_j}\|_{M,w}$$

and

$$\|y_{L_y \cap P_j}\|_{M,w'} = (1 \pm \varepsilon/2)\|y_{L_y \cap P_j}\|_{M,w},$$

which implies

$$\|y_{P_j}\|_{M,w'} = (1 \pm \varepsilon/2)\|y_{P_j}\|_{M,w}.$$

Case (iii): $\|y_{H_y \cap P_j}\|_{M,w} \geq T$ and $\|y_{L_y \cap P_j}\|_{M,w} < T$. By Lemma D.3 and Lemma D.4, with probability at least $1 - \delta_{\circ}/O(|\mathcal{N}| \cdot \log n)$,

$$\|y_{H_y \cap P_j}\|_{M,w'} = (1 \pm \varepsilon/2)\|y_{H_y \cap P_j}\|_{M,w}$$

and

$$\left|\|y_{L_y \cap P_j}\|_{M,w'} - \|y_{L_y \cap P_j}\|_{M,w}\right| \leq C \cdot U_M \cdot \tau^p \cdot 2^{j-2} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon.$$

Since

$$\|y_{P_j}\|_{M,w} \geq \|y_{H_y \cap P_j}\|_{M,w} \geq T \geq C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2,$$

we have

$$\left|\|y_{L_y \cap P_j}\|_{M,w'} - \|y_{L_y \cap P_j}\|_{M,w}\right| \leq \varepsilon/2 \cdot \|y_{P_j}\|_{M,w},$$

which implies

$$\|y_{P_j}\|_{M,w'} = (1 \pm \varepsilon)\|y_{P_j}\|_{M,w}.$$

Case (iv): $\|y_{H_y \cap P_j}\|_{M,w} < T$ and $\|y_{L_y \cap P_j}\|_{M,w} \geq T$. By Lemma D.3 and Lemma D.4, with probability at least $1 - \delta_{\circ}/O(|\mathcal{N}| \cdot \log n)$,

$$\|y_{L_y \cap P_j}\|_{M,w'} = (1 \pm \varepsilon/2)\|y_{L_y \cap P_j}\|_{M,w}$$

and

$$\left|\|y_{H_y \cap P_j}\|_{M,w'} - \|y_{H_y \cap P_j}\|_{M,w}\right| \leq C \cdot U_M \cdot \tau^p \cdot 2^{j-2} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon.$$

Since

$$\|y_{P_j}\|_{M,w} \geq \|y_{L_y \cap P_j}\|_{M,w} \geq T \geq C \cdot U_M \cdot \tau^p \cdot 2^{j-1} \cdot \log(|\mathcal{N}| \cdot \log n/\delta_{\circ})/\varepsilon^2,$$

we have

$$\left|\|y_{H_y \cap P_j}\|_{M,w'} - \|y_{H_y \cap P_j}\|_{M,w}\right| \leq \varepsilon/2 \cdot \|y_{P_j}\|_{M,w},$$

which implies

$$\|y_{P_j}\|_{M,w'} = (1 \pm \varepsilon)\|y_{P_j}\|_{M,w}.$$

□

Now we show that with probability $1 - \delta_{\circ}$, simultaneously for all $x \in \mathbb{R}^d$, $\|Ax\|_{p,w'}^p = (1 \pm \varepsilon)\|Ax\|_{p,w}^p$.

Lemma D.6. *For any $1 \leq j \leq 2\log n$, with with probability at least $1 - \delta_{\circ}/(2\log n)$, simultaneously for all $y = Ax$,*

$$\|y_{P_j}\|_{p,w'}^p = (1 \pm \varepsilon)\|y_{P_j}\|_{p,w}^p.$$

Applying a union bound over all $1 \leq j \leq 2\log n$, this implies with probability at least $1 - \delta_{\circ}$,

$$\|y\|_{p,w'}^p = (1 \pm \varepsilon)\|y\|_{p,w}^p.$$

Proof. For any fixed $1 \leq j \leq 2\log n$, by Theorem A.10, if we take $\delta_{\mathsf{subspace}} = \delta_{\circ}/(2\log n)$, with probability at least $1 - \delta_{\circ}/(2\log n)$, simultaneously for all $y = Ax$, we have

$$\|y_{P_j \setminus I_j}\|_{p,w'}^p = (1 \pm \varepsilon)\|y_{P_j \setminus I_j}\|_{p,w}^p.$$

Moreover, $\|y_{I_j}\|_{p,w}^p = \|y_{I_j}\|_{p,w'}^p$ since $w_i = w_i'$ for all $i \in I_j$. Thus, we have $\|y_{P_j}\|_{p,w'}^p = (1 \pm \varepsilon)\|y_{P_j}\|_{p,w}^p$. □

D.2. The Recursive Algorithm

We start by setting $w = 1^n$. In each recursive step, we use the sampling procedure defined in Section D.1 to obtain w', by setting $\delta_\circ = \delta/O(\log n)$ and $\varepsilon = \varepsilon'/O(\log n)$ for some $\varepsilon' > 0$. By Lemma D.1, for each recursive step, with probability at least $1 - \delta/(10\log n)$, we have $\|w'\|_0 \le 2/3\|w\|_0$. We repeat the recursive step until $\|w\|_0 \le 10F$.

By applying a union bound over all recursive steps, with probability $1 - \delta/10$, the recursive depth is at most $\log_{3/2} n$. By Lemma D.2, this also implies with probability $1 - \delta/10$, during the whole recursive algorithm, the weight vector w always satisfies $\|w\|_\infty \le 2^{\log_{1.5} n} \le n^2$. If we use w_{final} to denote the final weight vector, then we have

$$\|w_{\mathsf{final}}\|_0 \le 10F = \widetilde{O}(d^{\max\{1,p/2\}} \log n \cdot (\log(|\mathcal{N}|/\delta_\circ) \cdot \log(1/\delta_\circ) + d)/\varepsilon^2).$$

By Lemma D.5, and a union bound over all the $\log_{1.5} n$ recursive depths, with probability $1 - \delta$, simultaneously for all $y \in \mathcal{N}$, we have

$$\|Ax\|_{M,w_{\mathsf{final}}} = (1 \pm O(\varepsilon \cdot \log n))\|Ax\|_M = (1 \pm O(\varepsilon'))\|Ax\|_M.$$

Moreover, by Lemma D.6 and a union bound over all the $\log_{1.5} n$ recursive depths, with probability $1-\delta/10$, simultaneously for all $y = Ax$ we have

$$\|Ax\|_{p,w_{\mathsf{final}}}^p = (1 \pm O(\varepsilon \cdot \log n))\|Ax\|_{p,w}^p = (1 \pm O(\varepsilon'))\|Ax\|_{p,w}^p.$$

We further show that conditioned on this event, simultaneously for all $x \in \mathbb{R}^d$,

$$\|Ax\|_{M,w_{\mathsf{final}}} \ge \frac{L_M}{U_M \cdot n} \cdot \|Ax\|_M.$$

Consider a fixed vector $x \in \mathbb{R}^d$, if there exists a coordinate $i \in H_{Ax}$ such that $w_i > 0$, since $w_i \ge 1$ if $w_i > 0$, we must have

$$\|Ax\|_{M,w_{\mathsf{final}}} \ge w_i M((Ax)_i) \ge M((Ax)_i) \ge L_M \cdot \tau^p.$$

On the other hand,

$$\|Ax\|_M \le n \cdot U_M \cdot \tau^p,$$

which implies

$$\|Ax\|_{M,w_{\mathsf{final}}} \ge \frac{L_M}{U_M \cdot n} \cdot \|Ax\|_M.$$

Otherwise, $i \in L_{Ax}$ for all $i \in [n]$, which implies

$$\|Ax\|_{M,w_{\mathsf{final}}} \ge L_M \cdot \|Ax\|_{p,w_{\mathsf{final}}}^p \ge (1 - O(\varepsilon'))L_M\|Ax\|_{p,w}^p \ge \frac{(1 - O(\varepsilon'))L_M}{U_M}\|Ax\|_M.$$

Finally, since each recursive step runs in $\widetilde{O}((\mathrm{nnz}(A) + d^{p/2+O(1)} \cdot \log(|\mathcal{N}|/\delta) \cdot /\varepsilon^2) \cdot \log(1/\delta))$ time, and the number of recursive steps is upper bounded by $\log_{1.5} n$ with probability $1 - \delta/10$, the total running time is also upper bounded $\widetilde{O}((\mathrm{nnz}(A) + d^{p/2+O(1)} \cdot \log(|\mathcal{N}|/\delta) \cdot /\varepsilon^2) \cdot \log(1/\delta))$ with probability $1 - \delta/10$.

The following lemma can be proved by applying a union bound over all observations above, changing ε' to ε and changing A to $[A\ b]$.

Lemma D.7. *The algorithm outputs a vector $w_{\mathsf{final}} \in \mathbb{R}^n$, such that for any set $\mathcal{N} \subseteq \mathbf{im}([A\ b])$ with size $|\mathcal{N}|$, with probability $1 - \delta$, the algorithm runs in $\widetilde{O}((\mathrm{nnz}(A) + d^{p/2+O(1)} \cdot \log(|\mathcal{N}|/\delta) \cdot /\varepsilon^2) \cdot \log(1/\delta))$ time and the following holds:*

1. $\|w_{\mathsf{final}}\|_0 \le \widetilde{O}(d^{\max\{1,p/2\}} \log^3 n \cdot (\log(|\mathcal{N}|/\delta) \cdot \log(1/\delta) + d)/\varepsilon^2)$;
2. $\|w_{\mathsf{final}}\|_\infty \le n^2$;
3. *For all* $x \in \mathbb{R}^d$, $\|Ax - b\|_{M,w_{\mathsf{final}}} \ge \frac{L_M}{U_M \cdot n} \cdot \|Ax - b\|_M$.
4. *For all* $x \in \mathcal{N}$, $\|Ax - b\|_{M,w_{\mathsf{final}}} = (1 \pm \varepsilon)\|Ax - b\|_M$.

Combining Lemma D.7 with the net argument in Theorem C.5, we have the following theorem.

Theorem D.8. *By setting $|\mathcal{N}| = n^{O(d^3)} \cdot (1/\varepsilon)^{O(d)}$, the algorithm outputs a vector $w_{\mathsf{final}} \in \mathbb{R}^n$, such that with probability $1-\delta$, the algorithm runs in $\widetilde{O}((\mathrm{nnz}(A) + d^{p/2+O(1)}/\varepsilon^2 \cdot \log(1/\delta)) \cdot \log(1/\delta))$ time, $\|w_{\mathsf{final}}\|_0 \le \widetilde{O}(d^{p/2+O(1)} \log^4 n \cdot \log^2(1/\delta)/\varepsilon^2)$ and any C-approximate solution of $\min_x \|Ax-b\|_{M,w_{\mathsf{final}}}$ with $C \le \mathrm{poly}(n)$ is a $C \cdot (1+\varepsilon)$-approximate solution of $\min_x \|Ax-b\|_M$.*

Proof. Lemma D.7 implies that $U_O = 1+\varepsilon$, $L_N = 1-\varepsilon$, $L_A = \frac{L_M}{U_M \cdot n}$ and $U_A \le \|w_{\mathsf{final}}\|_\infty \le n^2$. Adjusting constants and applying Theorem C.5 imply the desired result. □

E. The M-sketch

In this section we give an oblivious sketch for Tukey loss functions. Throughout this section we assume $1 \le p \le 2$ in Assumption 1.

For convenience and to set up notation, we first describe the construction.

The sketch. Each coordinate z_p of a vector z to be sketched 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 is $S \in \mathbb{R}^{Nh_{\max} \times n}$, where $h_{\max} \equiv \lfloor \log_b(n/m) \rfloor$. Our weight vector $w \in \mathbb{R}^{Nh_{\max}}$ has entries $w_{i+1} \leftarrow \beta b^h$, for $i \in [Nh, N(h+1))$ and integer $h = 0, 1, \ldots, h_{\max}$, and $\beta \equiv (b - b^{-h_{\max}})/(b-1)$. Our sketch is reminiscent of sketches in the data stream literature, where we hash into buckets at multiple levels of subsampling (Indyk & Woodruff, 2005; Verbin & Zhang, 2012). However, the estimation performed in the sketch space needs to be the same as in the original space, which necessitates a new analysis.

The entries of S are $S_{j,p} \leftarrow \Lambda_p$, where $p \in [n]$ and $j \leftarrow g_p + Nh_p$ and

$$\begin{aligned} &\Lambda_p \leftarrow \pm 1 \text{ with equal probability} \\ &g_p \in [N] \text{ chosen with equal probability} \\ &h_p \leftarrow h \text{ with probability } 1/\beta b^h \text{ for integer } h \in [0, h_{\max}], \end{aligned} \tag{4}$$

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 $\|Sz\|_{M,w}$ as $\sum_{h \in [0,h_{\max}], i \in [N]} \beta b^h M(\|L_{h,i}\|_\Lambda)$, where $\|L\|_\Lambda$ denotes $|\sum_{z_p \in L} \Lambda_p z_p|$.

E.1. Accuracy Bounds for Sketching One Vector

We will show that our sketching construction has the property that for a given vector $z \in \mathbb{R}^n$, with high probability, $\|Sz\|_{M,w}$ is not too much smaller than $\|z\|_M$. We assume that $\|z\|_M = 1$, for notational convenience.

Define $y \in \mathbb{R}^n$ by $y_p = M(z_p)$, so that $\|y\|_1 = \|z\|_M = 1$. 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 $M(\hat{Z}) \subset Y$ denote $\{M(z_p) \mid z_p \in \hat{Z}\}$. Let $\|Y\|_k$ denote $\left(\sum_{y \in Y} |y|^k\right)^{1/k}$, so $\|Y\|_1 = \|y\|_1$. Hereafter multisets will just be called "sets".

Weight classes. Fix a value $\gamma > 1$, and for integer $q \ge 1$, let W_q denote the multiset comprising *weight class* $\{y_p \in Y \mid \gamma^{-q} \le y_p \le \gamma^{1-q}\}$. We have $\beta b^h \,\mathrm{E}[\|M(L_h) \cap W_q\|_1] = \|W_q\|_1$. 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 $y' \in \mathbb{R}^n$ with $y'_i \leftarrow y_i$ when $y_i > \varepsilon/n$, and $y'_i \leftarrow 0$ otherwise. Then $\|y'\|_1 \ge 1 - n(\varepsilon/n) = 1 - \varepsilon$. 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 contribution to $\|y\|_1$ of weight classes W_q of smaller total weight, added up for $q \le q_{\max}$, 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

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

for all W_q except those with $|W_q| < \beta m$, for which the lower bound does not hold.

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

E.2. Contraction Bounds

Here we will show that $\|Sz\|_{M,w}$ is not too much smaller than $\|z\|_M$. We will need some weak conditions among the parameters. Recall that $N = bcm$.

Assumption 3. *We will assume $b \ge m$, $b > c$, $m = \Omega(\log\log(n/\varepsilon))$, $\log b = \Omega(\log\log(n/\varepsilon))$, $\gamma \ge 2 \ge \beta$, an error parameter $\varepsilon \in [1/10, 1/3]$, and $\log N \le \varepsilon^2 m$. We will consider γ to be fixed throughout, that is, not dependent on the other parameters.*

We need lemmas that allow lower bounds on the contributions of the weight classes. First, some notation. For $h = 0, 1, \ldots, h_{\max}$, let

$$\begin{aligned}
M_< &\equiv \log_\gamma(m/\varepsilon) = O(\log_\gamma(b/\varepsilon))\\
Q_< &\equiv \{q \mid |W_q| < \beta m, q \le M_<\}\\
\hat{Q}_h &\equiv \{q \mid h(q) = h, |W_q| \ge \beta m\}\\
M_\ge &\equiv \log_\gamma(2(1+3\varepsilon)b/\varepsilon)\\
Q_h &\equiv \{q \in \hat{Q}_h \mid q \le M_\ge + \min_{q \in \hat{Q}_h} q\}\\
Q^* &\equiv Q_< \cup [\cup_h Q_h].
\end{aligned} \tag{5}$$

Here $Q_<$ is the set of indices of weight classes that have relatively few members, but contain relatively large weights. $\hat{Q}_h$ 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 E.1. *If $N \ge \max\{O(|M_<|dm^3\varepsilon), \widetilde{O}(d^2m^2/\varepsilon^2)\}$, then with constant probability, for all $z \in \mathbf{im}(A)$ and all $q \in Q_<$, the following event $\mathcal{E}_v$ holds: there are sets $W_q^* \subset W_q$, with $|W_q^*| \ge (1-\varepsilon)|W_q|$, such that for all $y \in W_q^*$,*

1. *they are isolated: they are the sole members of $W_{Q_<}$ in their bucket;*
2. *their buckets are low-weight: the set L of other entries in bucket containing $y \in W_q^*$ has $\|L\|_1 \le 1/\varepsilon^2 m^3$.*

Proof. Without loss of generality we assume $h(q)$ are the same for all $q \in M_<$, since otherwise we can deal with each $h(q)$ separately.

Let $\alpha = m/(L_M \cdot \varepsilon)$. By Lemma B.4, there exists a set $I \subseteq [n]$ with size $|I| = \widetilde{O}(d \cdot \alpha) = \widetilde{O}(d \cdot m/\varepsilon)$ such that for any $z \in \mathbf{im}(A)$, if z satisfies (i) $\|z_{L_z}\|_p^p \le \alpha \cdot \tau^p$ and (ii) $|H_z| \le \alpha$, then $H_z \subseteq I$. Let $\{u\}_{i \in [n]\setminus I}$ be the ℓ_p Lewis weights of $A_{[n]\setminus I,*}$ and let $J \subseteq [n] \setminus I$ be the set of indices of the $d \cdot m/\varepsilon \cdot U_M/L_M$ largest coordinates of u. Thus, $|J| \le O(d \cdot m/\varepsilon)$. Since J contains the $d \cdot m/\varepsilon \cdot U_M/L_M$ largest coordinates of u and

$$\sum_{i \in [n]\setminus I} u_i = \sum_{i \in [n]\setminus I} \overline{u}_i^p \le d$$

by Theorem A.5, for each $i \in [n] \setminus (I \cup J)$, we have $u_i \le d/(d \cdot m/\varepsilon \cdot U_M/L_M) \le \varepsilon/m \cdot L_M/U_M$.

If $\tau^p < \|z\|_M \cdot \varepsilon/m$, by Assumption 1.2, we have $M(z_i) \le \tau^p < \|z\|_M \cdot \varepsilon/m$ for all $i \in [n]$. In this case, we have $W_{Q<} = \emptyset$. Thus we assume $\tau^p \ge \|z\|_M \cdot \varepsilon/m$ in the remaining part of the analysis.

Since $\|z\|_M \ge |H_z| \cdot \tau^p$, we have $|H_z| \le m/\varepsilon$. Furthermore, by Assumption 1.4, $\|z_{L_z}\|_p^p \le \|z_{L_z}\|_M/L_M \le \|z\|_M/L_M \le \tau^p \cdot m/(L_M \cdot \varepsilon)$. Thus by setting $\alpha = m/(L_M \cdot \varepsilon)$ we have $H_z \subseteq I$. For each $i \in [n] \setminus I$, we have $|z_i| \le \tau$. By Lemma

A.8 and Assumption 1.4, for each $i \in [n] \setminus I$, $M(z_i) \le |z_i|^p / L_M \le u_i \cdot \|z_{[n]\setminus I}\|_p^p / L_M \le u_i \cdot \|z_{[n]\setminus I}\|_M \cdot U_M / L_M < u_i \cdot \|z\|_M \cdot U_M / L_M$. Thus for each entry $i \in [n] \setminus (I \cup J)$, we have $M(z_i) < \varepsilon / m \cdot \|z\|_M$.

Thus, the indices of all members of $W_{Q_<}$ are in $I \cup J$. By setting $N \ge |I \cup J|^2 / \kappa = \widetilde{O}(d^2 m^2 / \varepsilon^2)/\kappa$, the expected number of total collisions in $I \cup J$ is $|I \cup J|^2 / N \le \kappa$. Thus, by Markov's inequality, with probability $1 - 2\kappa$, the total number of collisions is upper bounded by $1/2$, i.e., there is no collision. This implies the first condition.

For the second condition, we use $\{u_i\}_{i \in [n]\setminus(I\cup J)}$ to denote the ℓ_p Lewis weights of $A_{i\in[n]\setminus(I\cup J),*}$. Consider a fixed $q \in M_<$. By the first condition, all elements in W_q are the sole members of $W_{Q_<}$ in their buckets. For each bucket we define $B_{h,i}$ to be the multiset $\{u_p \mid h_p = h, g_p = i, p \in [n] \setminus (I \cup J)\}$. By setting $N \ge \frac{U_M \cdot |M_<| \cdot d m^3 \varepsilon}{L_M \cdot \kappa}$, for each $y \in W_q$, $\mathrm{E}[\|B_{h,i}\|_1] \le d/N \le \frac{L_M}{U_M} \cdot \frac{1}{\varepsilon^2 m^3} \cdot \frac{\varepsilon \cdot \kappa}{|M_<|}$ where $L_{h,i}$ is the bucket that contains y. This is simply because $\sum_{i \in N} B_{h,i} \le \sum_{i \in [n]\setminus(I\cup J)} u_i \le d$ by Theorem A.5. We say a bucket is *good* if $\|B_{h,i}\|_1 \le \frac{L_M}{U_M} \cdot \frac{1}{\varepsilon^2 m^3}$. Notice that for $y \in W_q$, if y is in a good bucket $B_{h,i}$, then the set L of other entries in that bucket satisfies

$$\begin{aligned}
\|L\|_1 &= \sum_{y \in L} y \\
&= \sum_{p \in [n]\setminus(I\cup J) \mid h_p = h, g_p = i} M(z_p) \\
&\le \sum_{p \in [n]\setminus(I\cup J) \mid h_p = h, g_p = i} U_M \cdot |z_p|^p && \text{(Assumption 1.4)} \\
&\le \sum_{p \in [n]\setminus(I\cup J) \mid h_p = h, g_p = i} U_M \cdot u_p \cdot \|z_{[n]\setminus(I\cup J)}\|_p^p && \text{(Lemma A.8)} \\
&\le \sum_{p \in [n]\setminus(I\cup J) \mid h_p = h, g_p = i} U_M / L_M \cdot u_p \cdot \|z_{[n]\setminus(I\cup J)}\|_M && \text{(Assumption 1.4)} \\
&\le \|B_{h,i}\|_1 \cdot U_M / L_M \cdot \|z\|_M \\
&\le \frac{1}{\varepsilon^2 m^3} \cdot \|z\|_M.
\end{aligned}$$

Thus, it suffices to show that at least $(1-\varepsilon)|W_q|$ buckets associated with $y \in W_q$ are good.

By Markov's inequality, for each $y \in W_q$, with probability $1 - \varepsilon \cdot \kappa / |M_<|$, the bucket that contains y is good. Thus, for the $|W_q|$ buckets associated with $y \in W_q$, the expected number of good buckets is at least $(1 - \varepsilon \cdot \kappa / M_<)|W_q|$. Again, by Markov's inequality, with probability at least $1 - \kappa / |M_<|$, at least $(1-\varepsilon)|W_q|$ buckets associated with $y \in W_q$ are good, and we just take these $(1-\varepsilon)|W_q|$ good buckets to be W_q^*. By applying a union bound over all $q \in M_<$, the second condition holds with probability at least $1 - \kappa$. The lemma follows by applying a union bound over the two conditions and setting κ to be a small constant.

□

Lemma E.2 (Lemma 3.8 of (Clarkson & Woodruff, 2015b)). *Let $Q'_h \equiv \{q \mid q \le M'_h\}$, where $M'_h \equiv \log_\gamma(\beta b^{h+1} m^2 q_{\max})$. Then for large enough $N = O(m^2 b \varepsilon^{-1} q_{\max})$, with probability at least $1 - C^{-\varepsilon^2 m}$ for a constant $C > 1$, for each $q \in \cup_h Q_h$, 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. *each $x \in W_q^*$ is in a bucket with no member of $W_{Q'_h}$.*

For $v \in T \subset Z$, let $T - v$ denote $T \setminus \{v\}$.

Lemma E.3 (Lemma 3.6 of (Clarkson & Woodruff, 2015b)). *For $v \in T \subset Z$,*

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

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

$$\frac{\|T - v\|_2}{|v|} \leq \varepsilon^{1/2}, \tag{6}$$

and for a constant C, $\mathrm{E}_\Lambda[M(\|T\|_\Lambda)] \geq (1 - C\varepsilon^{1/2})M(v)$.

Lemma E.4 (Lemma 3.9 of (Clarkson & Woodruff, 2015b)). *Assume Assumption 3. There is $N = O(\varepsilon^{-2}m^2bq_{\max})$, so that for all $0 \leq h \leq h_{\max}$ and $q \in Q_h$ with $\|W_q\|_1 \geq \varepsilon/q_{\max}$, we have*

$$\sum_{y_p \in W_q^*} M(\|L(y_p)\|_\Lambda) \geq (1 - \varepsilon^{1/2})\|W_q\|_1$$

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

Lemma E.5. *Assume that $\mathcal{E}_v$ of Lemma E.1 holds, and Assumption 3. Then for $q \in Q_<$,*

$$\sum_{y_p \in W_q^*} M(\|L(y_p)\|_\Lambda) \geq (1 - \varepsilon^{1/2})\|W_q\|_1$$

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

Proof. Let $v \equiv z_p$ where $y_p = M(z_p)$, let $L(v)$ denote the $\{z_{p'} \mid M(z_{p'}) \in L\}$. Condition $\mathcal{E}_v$ and $M(v) \geq \varepsilon/m$ imply that

$$\|L(v) - v\|_2^2 \leq \|L\|_1 \leq 1/\varepsilon^2 m^3 < M(v)/\varepsilon m,$$

so that using (6) we have

$$\frac{\|L(v) - v\|_2^2}{|v|^2} \leq \frac{\|L(v) - v\|_M}{M(v)} \leq \frac{1}{\varepsilon m}. \tag{7}$$

Since $\|L\|_\infty \leq \|L\|_1$, we also have, for all $v' \in L(v) - v$, and using again $M(v) \geq \varepsilon/m$,

$$\left|\frac{v'}{v}\right| \leq \left(\frac{M(v')}{M(v)}\right)^{1/2} \leq \frac{1}{m\varepsilon^{3/2}}. \tag{8}$$

From (8), we have that the summands determining $\|L(v) - v\|_\Lambda$ have magnitude at most $|v|\varepsilon^{1/2}/\varepsilon^2 m$. From (7), we have $\|L(v) - v\|_2^2$ is at most $v^2\varepsilon/\varepsilon^2 m$. It follows from Bernstein's inequality that with failure probability $\exp(-\varepsilon^2 m)$, $\|L(v) - v\|_\Lambda \leq \varepsilon^{1/2}|v|$. Applying the first claim of Lemma E.3, we have $M(\|L(v)\|_\Lambda) \geq (1 - 2\varepsilon^{1/2})M(v)$, for all $v \in M^{-1}(W_q^*)$ with failure probability $\beta m M_< \exp(-\varepsilon^2 m)$. Summing over W_q^*, we have

$$\sum_{v \in M^{-1}(W_q*)} M(\|L(v)\|_\Lambda) \geq (1 - \varepsilon^{1/2})\|W_q^*\|_1 \geq (1 - 2\varepsilon\gamma)(1 - \varepsilon^{1/2})\|W_q\|_1.$$

This implies the bound, using Assumption 3, after adjusting constants. □

The above lemmas imply that overall, with high probability, the sketching-based estimate of $\|z\|_M$ of a single given vector z is very likely to not much smaller than $\|z\|_M$, as stated next.

Theorem E.6 (Theorem 3.2 of (Clarkson & Woodruff, 2015b)). *Assume Assumption 3, and condition $\mathcal{E}_v$ of Lemma E.1. Then $\|Sz\|_{M,w} \geq \|z\|_M(1 - \varepsilon^{1/2})$, with failure probability no more than $C^{-\varepsilon^2 m}$, for an absolute constant $C > 1$.*

E.3. A "Clipped" Version

For a vector z, we use $\|Sz\|_{Mc,w}$ to denote a "clipped" version of $\|Sz\|_{M,w}$, in which we ignore small buckets and use a subset of the coordinates of Sz as follows: $\|Sz\|_{Mc,w}$ is obtained by adding in only those buckets in level h that are among the top

$$M^* \equiv bmM_{\geq} + \beta mM_{<}$$

in $\|L_{h,i}\|_\Lambda$, recalling $M_{\geq}$ and $M_{<}$ defined in (5). Formally, we define $\|Sz\|_{Mc,w}$ to be

$$\|Sz\|_{Mc,w} = \sum_{h\in[0,h_{\max}],i\in[M^*]} \beta b^h M(\|L_{h,(i)}\|_\Lambda),$$

where $L_{h,(i)}$ denotes the level h bucket with the i-th largest $\|L_{h,i}\|_\Lambda$ among all the level h buckets.

The proof of the contraction bound of $\|Sz\|_{M,w}$ in Theorem E.6 requires only lower bounds on $M(\|L_{h,i}\|_\Lambda)$ for those at most M^* buckets on level h. Thus, the proven contraction bounds continue to hold for $\|Sz\|_{Mc,w}$, and in particular $\|Sz\|_{Mc,w} \geq (1-\varepsilon)\|Sz\|_{M,w}$.

E.4. Dilation Bounds

We use two prior bounds of (Clarkson & Woodruff, 2015b) on dilation; the first shows that the dilation is at most $O(\log n)$ in expectation, while the second shows that the "clipped" version gives $O(1)$ dilation with constant probability. Note that we need only expectations, since we need the dilation bound to hold only for the optimal solution as in Theorem C.5.

Theorem E.7 (Theorem 3.3 of (Clarkson & Woodruff, 2015b))**.** $\mathrm{E}[\|Sz\|_{M,w}] = O(h_{\max})\|z\|_M$.

Better dilation is achieved by using the "clipped" version $\|Sz\|_{Mc,w}$, as described in (Clarkson & Woodruff, 2015b).

Theorem E.8 (Theorem 3.4 of (Clarkson & Woodruff, 2015b))**.** *There is $c = O(\log_\gamma(b/\varepsilon)(\log_b(n/m)))$ and $b \geq c$, recalling $N = mbc$, such that*

$$\mathrm{E}[\|Sz\|_{Mc,w}] \leq C\|z\|_M$$

for a constant C.

E.5. Regression Theorem

Lemma E.9. *There is $N = O(d^2 h_{\max})$, so that with constant probability, simultaneously for all $x \in \mathbb{R}^d$,*

$$0.9/(n \cdot U_M/L_M)\|Ax-b\|_M \leq \|S(Ax-b)\|_{M,w} \leq U_M/L_M \cdot n^2 \cdot \|Ax-b\|_M.$$

Proof. For the upper bound,

$$\|Sz\|_{M,w} = \sum_{h\in[0,h_{\max}],i\in[N]} \beta b^h M(\|L_{h,i}\|_\Lambda).$$

The weights βb^h are less than n, and

$$\begin{aligned}
&M(\|L_{h,i}\|_\Lambda)\\
\leq &M(\|L_{h,i}\|_1)\\
\leq &M(n^{1-1/p}\|L_{h,i}\|_p) && \text{(Assumption 1.2)}\\
\leq &U_M \cdot n^{p-1}\|L_{h,i}\|_p^p && \text{(Assumption 1.4)}\\
\leq &U_M/L_M \cdot n \cdot \sum_{z_p\in L_{h,i}} M(z_p). && \text{(Assumption 1.4)}
\end{aligned}$$

Since any given z_p contributes once to $\|Sz\|_{M,w}$, $\|Sz\|_{M,w} \leq U_M/L_M \cdot n^2 \cdot \|z\|_M$.

For the lower bound, notice that

$$\|Sz\|_{2,w}^2 = \sum_{h\in[0,h_{\max}],i\in[N]} \beta b^h \|L_{h,i}\|_\Lambda^2.$$

For each $h \in [0, h_{\max}]$, since $N = O(d^2 h_{\max})$, with probability at least $1 - 1/(10h_{\max})$, simultaneously for all $z \in \mathbf{im}(A)$ we have

$$\sum_{i \in [N]} \|L_{h,i}\|_\Lambda^2 = (1 \pm 0.1) \sum_{z_p \in L_h} z_p^2,$$

since the summation on the left-hand side can be equivalently viewed as applying CountSketch (Clarkson & Woodruff, 2013; Nelson & Nguyen, 2012; Meng & Mahoney, 2012) on L_h. Thus, by applying union bound over all $h \in [0, h_{\max}]$, we have

$$\|Sz\|_{2,w}^2 = \sum_{h \in [0,h_{\max}], i \in [N]} \beta b^h \|L_{h,i}\|_\Lambda^2 \geq 0.9\|z\|_2^2. \tag{9}$$

If there exists some $i \in H_{Sz}$, since $w_i \geq 1$ for all i, we have

$$\|Sz\|_{M,w} \geq w_i M((Sz)_i) \geq M((Sz)_i) \geq \tau^p.$$

On the other hand,

$$\|z\|_M \leq n \cdot U_M \cdot \tau^p,$$

which implies

$$\|Sz\|_{M,w} \geq \|z\|_M / (n \cdot U_M).$$

If $H_{Sz} = \emptyset$, then

$$\begin{aligned}
&\|Sz\|_{M,w} \\
\geq& \sum_i w_i |(Sz)_i|^p \cdot L_M && \text{(Assumption 1.4)} \\
=& \|Sz\|_{p,w}^p \cdot L_M \\
\geq& \|Sz\|_{2,w}^p \cdot L_M && (p \leq 2) \\
\geq& 0.9\|z\|_2^p \cdot L_M && ((9)) \\
\geq& 0.9\|z\|_p^p \cdot L_M / n \\
\geq& 0.9\|z\|_M / (n \cdot U_M / L_M). && \text{(Assumption 1.4)}
\end{aligned}$$

□

The following theorem states that M-sketches can be used for Tukey regression, under the conditions described above.

Theorem E.10. *Under Assumption 1 and Assumption 2, there is an algorithm running in $O(\mathrm{nnz}(A))$ time, that with constant probability creates a sketched regression problem $\min_x \|S(Ax - b)\|_{M,w}$ where SA and Sb have $\mathrm{poly}(d \log n)$ rows, and any C-approximate solution $\tilde{x}$ of $\min_x \|S(Ax - b)\|_{M,w}$ with $C \leq \mathrm{poly}(n)$ satisfies*

$$\|A\tilde{x} - b\|_M \leq O(C \cdot \log_d n) \min_{x \in \mathbb{R}^d} \|Ax - b\|_M.$$

Moreover, any C-approximate solution $\hat{x}$ of $\min_x \|S(Ax - b)\|_{Mc,w}$ with $C \leq \mathrm{poly}(n)$ satisfies

$$\|A\hat{x} - b\|_M \leq O(C) \min_{x \in \mathbb{R}^d} \|Ax - b\|_M.$$

Proof. We set S to be an M-sketch matrix with large enough $N = \mathrm{poly}(d \log n)$. We note that, up to the trivial scaling by β, SA satisfies Assumption 2 if A does. We also set $m = O(d^3 \log n)$, and $\varepsilon = 1/10$. We apply Theorem C.5 to prove the desired result.

The given N is large enough for Theorem E.6 and Lemma E.9 to apply, obtaining a contraction bound with failure probability C_1^{-m}. By Theorem E.6, since the needed contraction bound holds for all members of $\mathcal{N}_{\mathrm{poly}(\varepsilon \cdot \tau/n)} \cup \mathcal{M}_{\mathrm{poly}(\varepsilon/n)}^{c, c \cdot \mathrm{poly}(n)}$, with failure probability $n^{O(d^3)} C_1^{-m} < 1$, for $m = O(d^3 \log n)$, assuming the condition $\mathcal{E}_v$.

Thus, by Theorem E.7, we have $U_O \leq O(\log_d n)$. By Lemma E.9, $L_A = 0.9/(n \cdot U_M/L_M)$ and $U_A = U_M/L_M \cdot n^2$. By Theorem E.6, $L_N = 1 - \varepsilon^{1/2} = \Omega(1)$. Thus, by Theorem C.5 we have

$$\|A\tilde{x} - b\|_M \leq O(C \cdot \log_d n) \min_{x \in \mathbb{R}^d} \|Ax - b\|_M.$$

A similar argument holds for C-approximate solution $\hat{x}$ of $\min_x \|S(Ax - b)\|_{Mc,w}$. □

F. Hardness Results and Provable Algorithms for Tukey Regression

F.1. Hardness Results

In this section, we prove hardness results for Tukey regression based on the *Exponential Time Hypothesis* (Impagliazzo & Paturi, 2001). We first state the hypothesis.

Conjecture 1 (Exponential Time Hypothesis (Impagliazzo & Paturi, 2001))**.** *For some constant $\delta > 0$, no algorithm can solve* 3-SAT *on n variables and $m = O(n)$ clauses correctly with probability at least $2/3$ in $O(2^{\delta n})$ time.*

Using Dinur's PCP Theorem (Dinur, 2007), Hypothesis 1 implies a hardness result for MAX-3SAT.

Theorem F.1 ((Dinur, 2007))**.** *Under Hypothesis 1, for some constant $\varepsilon > 0$ and $c > 0$, no algorithm can, given a* 3-SAT *formula on n variables and $m = O(n)$ clauses, distinguish between the following cases correctly with probability at least $2/3$ in $2^{n/\log^c n}$ time:*

- *There is an assignment that satisfies all clauses in ϕ;*
- *Any assignment can satisfy at most $(1 - \varepsilon)m$ clauses in ϕ.*

We make the following assumptions on the loss function $M : \mathbb{R} \to \mathbb{R}^+$. Notice that the following assumptions are more general than those in Assumption 1.

Assumption 4. *There exist real numbers $\tau \geq 0$ and $C > 0$ such that*

1. *$M(x) = C$ for all $|x| \geq \tau$.*
2. *$0 \leq M(x) \leq C$ for all $|x| \leq \tau$.*
3. *$M(0) = 0$.*

Now we give an reduction that transforms a 3-SAT formula ϕ with d variables and $m = O(d)$ clauses to a Tukey regression instance

$$\min_x \|Ax - b\|_M,$$

such that $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ with $n = O(d)$, and all entries in A are in $\{0, +1, -1\}$ and all entries in b are in $\{\pm k\tau \mid k \in \mathbb{N}, k \leq O(1)\}$. Furthermore, there are at most three non-zero entries in each row of A.

For each variable v_i in the formula ϕ, there is a variable x_i in the Tukey regression that corresponds to v_i. For each variable v_i, if v_i appears in Γ_i clauses in ϕ, we add $2\Gamma_i$ rows into $[A\ b]$. These $2\Gamma_i$ rows are chosen such that when calculating $\|Ax - b\|_M$, there are Γ_i terms of the form $M(x_i)$, and another Γ_i terms of the form $M(x_i - 10\tau)$. This can be done by taking the i-th entry of the corresponding row of A to be 1 and taking the corresponding entry of b to be either 0 or 10τ. Since $\sum_{i=1}^{d} \Gamma_i = 3m$ in a 3-SAT formula ϕ, we have added $6m = O(d)$ rows into $[A\ b]$. We call these rows Part I of $[A\ b]$.

Now for each clause $\mathcal{C} \in \phi$, we add three rows into $[A\ b]$. Suppose the three variables in $\mathcal{C}$ are v_i, v_j and v_k. The first row is chosen such that when calculating $\|Ax - b\|_M$, there is a term of the form $M(a + b + c - 10\tau)$, where $a = x_i$ if there is a positive literal that corresponds to v_i in $\mathcal{C}$ and $a = 10\tau - x_i$ if there is a negative literal that corresponds to v_i in $\mathcal{C}$. Similarly, $b = x_j$ if there is a positive literal that corresponds to v_j in $\mathcal{C}$ and $b = 10\tau - x_j$ if there is a negative literal that corresponds to v_j in $\mathcal{C}$. The same holds for c, x_k, and v_k. The second and the third row are designed such that when calculating $\|Ax - b\|_M$, there is a term of the form $M(a + b + c - 20\tau)$ and another term of the form $M(a + b + c - 30\tau)$.

Clearly, this can also be done while satisfying the constraint that all entries in A are in $\{0, +1, -1\}$ and all entries in b are in $\{\pm k\tau \mid k \in \mathbb{N}, k \le O(1)\}$. We have added $3m$ rows into $[A\ b]$. We call these rows Part II of $[A\ b]$.

This finishes our construction, with $6m + 3m = O(d)$ rows in total. It also satisfies all the restrictions mentioned above.

Now we show that when ϕ is satisfiable, if we are given any solution $\overline{x}$ such that

$$\|A\overline{x} - b\|_M \le (1 + \eta) \min_x \|Ax - b\|_M,$$

then we can find an assignment to ϕ that satisfies at least $(1 - 5\eta)m$ clauses.

We first show that when ϕ is satisfiable, the regression instance we constructed satisfies

$$\min_x \|Ax - b\|_M \le 5C \cdot m.$$

We show this by explicitly constructing a vector x. For each variable v_i in ϕ, if $v_i = 0$ in the satisfiable assignment, then we set x_i to be 0. Otherwise, we set x_i to be 10τ. For each variable v_i, since $x_i \in \{0, 10\tau\}$, for all the $2\Gamma_i$ rows added for it, there will be Γ_i rows contributing 0 when calculating $\|Ax - b\|_M$, and another Γ_i rows contributing C when calculating $\|Ax - b\|_M$. The total contribution from this part will be $3C \cdot m$. For each clause $\mathcal{C} \in \phi$, for the three rows added for it, there will be one row contributing 0 when calculating $\|Ax - b\|_M$, and another two rows contributing C when calculating $\|Ax - b\|_M$. This is by construction of $[A\ b]$ and by the fact that $\mathcal{C}$ is satisfied. Notice that $M(a + b + c - 10\tau) = 0$ if only one literal in $\mathcal{C}$ is satisfied, $M(a + b + c - 20\tau) = 0$ if two literals are satisfied, and $M(a + b + c - 30\tau) = 0$ if all three literals in $\mathcal{C}$ are satisfied. Thus, we must have $\min_x \|Ax - b\|_M \le 5C \cdot m$, which implies $\|A\overline{x} - b\|_M \le (1 + \eta)5C \cdot m$.

We first show that we can assume each $\overline{x}_i$ satisfies $\overline{x}_i \in [-\tau, \tau]$ or $\overline{x}_i \in [9\tau, 11\tau]$. This is because we can set $\overline{x}_i = 0$ otherwise without increasing $\|A\overline{x} - b\|_M$, as we will show immediately. For any $\overline{x}_i$ that is not in the two ranges mentioned above, its contribution to $\|A\overline{x} - b\|_M$ in Part I is at least $C \cdot 2\Gamma_i$. However, by setting $\overline{x}_i = 0$, its contribution to $\|A\overline{x} - b\|_M$ in Part I will be at most $C \cdot \Gamma_i$. Thus, by setting $\overline{x}_i = 0$ the total contribution to $\|A\overline{x} - b\|_M$ in Part I has been decreased by at least $C \cdot \Gamma_i$. Now we consider Part II of the rows in $[A\ b]$. The contribution to $\|A\overline{x} - b\|_M$ of all rows in $[A\ b]$ created for clauses that do not contain v_i will not be affected after changing $\overline{x}_i$ to be 0. For the $3\Gamma_i$ rows in $[A\ b]$ created for clauses that contain v_i, their contribution to $\|A\overline{x} - b\|_M$ is lower bounded by $C \cdot 2\Gamma_i$ and upper bounded by $C \cdot 3\Gamma_i$. The lower bound follows since for any three real numbers a, b and c, at least two elements in $\{a+b+c-10\tau, a+b+c-20\tau, a+b+c-30\tau\}$ have absolute value at least τ, and $M(x) = C$ for all $|x| \ge \tau$. Thus, by setting $\overline{x}_i = 0$ the total contribution to $\|A\overline{x} - b\|_M$ in Part II will be increased by at most $C \cdot \Gamma_i$, which implies we can set $\overline{x}_i = 0$ without increasing $\|A\overline{x} - b\|_M$.

Now we show how to construct an assignment to the 3-SAT formula ϕ which satisfies at least $(1 - 5\eta)m$ clauses, using a vector $\overline{x} \in \mathbb{R}^d$ which satisfies (i) $\|A\overline{x} - b\|_M \le (1 + \eta)5C \cdot m$ and (ii) $\overline{x}_i \in [-\tau, \tau]$ or $\overline{x}_i \in [9\tau, 11\tau]$ for all $\overline{x}_i$. We set $v_i = 0$ if $\overline{x}_i \in [-\tau, \tau]$ and set $v_i = 1$ if $\overline{x}_i \in [9\tau, 11\tau]$. To count the number of clauses satisfied by the assignment, we show that for each clause $\mathcal{C} \in \phi$, $\mathcal{C}$ is satisfied whenever $a + b + c \ge 7\tau$. Recall that $a = x_i$ if there is a positive literal that corresponds to v_i in $\mathcal{C}$ and $a = 10\tau - x_i$ if there is a negative literal that corresponds to v_i in $\mathcal{C}$. Similarly, $b = x_j$ if there is a positive literal that corresponds to v_j in $\mathcal{C}$ and $b = 10\tau - x_j$ if there is a negative literal that corresponds to v_j in $\mathcal{C}$. The same holds for c, x_k, and v_k. Since a, b and c are all in the range $[-\tau, \tau]$ or in the range $[9\tau, 11\tau]$, whenever $a + b + c \ge 7\tau$, we must have $a \ge 9\tau$, $b \ge 9\tau$ or $c \ge 9\tau$, in which case clause $\mathcal{C}$ will be satisfied. Thus, at least $(1 - 5\eta)m$ clauses will be satisfied, since otherwise $\|A\overline{x} - b\|_M$ will be larger than $3C \cdot m + 2C \cdot m + 5\eta C \cdot m = (1 + \eta)5C \cdot m$. Here the first term $3C \cdot m$ corresponds to the contribution from Part I, since any $\overline{x}_i$ must satisfy $|\overline{x}_i| \ge \tau$ or $|\overline{x}_i - 10\tau| \ge \tau$. The second and the third term $2C \cdot m + 5\eta C \cdot m$ corresponds to the contribution from Part II when at least $5\eta m$ clauses are not satisfied.

Our reduction implies the following theorem.

Theorem F.2. *Suppose there is an algorithm that runs in $T(d)$ time and succeeds with probability $2/3$ for Tukey regression with approximation ratio $1 + \eta$ when the loss function M satisfies Assumption 4 and the input data satisfies the following restrictions:*

1. *$A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ with $n = O(d)$.*
2. *All entries in A are in $\{0, +1, -1\}$ and all entries in b are in $\{\pm k\tau \mid k \in \mathbb{N}, k \le O(1)\}$.*
3. *There are at most three non-zero entries in each row of A.*

Then, there exists an algorithm that runs in $T(d)$ time for a 3-SAT formula on d variables and $m = O(d)$ clauses which distinguishes between the following cases correctly with probability at least $2/3$:

- *There is an assignment that satisfies all clauses in ϕ.*
- *Any assignment can satisfy at most $(1 - 5\eta)m$ clauses in ϕ.*

Combining Theorem F.1 and Theorem F.2 with the Hypothesis 1, we have the following corollary.

Corollary F.3. *Under Hypothesis 1, for some constant $\eta > 0$ and $C > 0$, no algorithm can solve Tukey regression with approximation ratio $1 + \eta$ and success probability $2/3$, and runs in $2^{d/\log^C d}$ time, when the loss function M satisfies Assumption 4 and the input data satisfies the following restrictions:*

1. *$A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ with $n = O(d)$.*
2. *All entries in A are in $\{0, +1, -1\}$ and all entries in b are in $\{\pm k\tau \mid k \in \mathbb{N}, k \le O(1)\}$.*
3. *There are at most three non-zero entries in each row of A.*

F.2. Provable Algorithms

In this section, we use the polynomial system verifier to develop provable algorithms for Tukey regression.

Theorem F.4 ((Renegar, 1992; Basu et al., 1996))**.** *Given a real polynomial system $P(x_1, x_2, \cdots, x_d)$ with d variables and n polynomial constraints $\{f_i(x_1, x_2, \cdots, x_d)\Delta_i 0\}_{i=1}^n$, where Δ_i is any of the "standard relations": $\{>, \ge, =, \ne, \le, <\}$, let D denote the maximum degree of all the polynomial constraints and let H denote the maximum bitsize of the coefficients of all the polynomial constraints. Then there exists an algorithm that runs in*

$$(Dn)^{O(d)} \operatorname{poly}(H)$$

time that can determine if there exists a solution to the polynomial system P.

Besides Assumption 1, we further assume that the loss function $M(x)$ can be approximated by a polynomial $P(x)$ with degree D, when $|x| \le \tau$. Formally, we assume there exist two constants $L_P \le 1 \le U_P$ such that when $|x| \le \tau$, we have

$$L_P P(|x|) \le M(|x|) \le U_P P(|x|).$$

Indeed, Assumption 1 already implies we can take $P(x) = x^p$, with $L_P = L_M$ and $U_P = U_M$ when p is an integer. However, for some loss function (e.g., the one defined in (1)), one can find a better polynomial to approximate the loss function. Since the approximation ratio of our algorithm depends on U_P/L_P, for those loss functions we can get an algorithm with better approximation ratio. We also assume Assumption 2 and all entries in A and b are integers.

We first show that under Assumption 2 and the assumption that all entries in A and b are integers, either $\|Ax - b\|_M = 0$ for some $x \in \mathbb{R}^d$, or $\|Ax - b\|_M \ge 1/2^{\operatorname{poly}(nd)}$ for all $x \in \mathbb{R}^d$.

Lemma F.5. *Suppose all entries in A and b are integers, under Assumption 1 and Assumption 2, either $\|Ax - b\|_M = 0$ for some $x \in \mathbb{R}^d$, or $\|Ax - b\|_M \ge 1/2^{\operatorname{poly}(nd)}$ for all $x \in \mathbb{R}^d$.*

Proof. We show that either there exists $x \in \mathbb{R}^d$ such that $Ax = b$, or $\|Ax - b\|_2 \ge 1/2^{\operatorname{poly}(nd)}$ for all $x \in \mathbb{R}^d$. Notice that $\|Ax - b\|_2 \ge 1/2^{\operatorname{poly}(nd)}$ implies $\|Ax - b\|_\infty \ge 1/2^{\operatorname{poly}(nd)}/\sqrt{n}$, and thus the claimed bound follows from Assumption 1.

Without loss of generality we assume A is non-singular. By the normal equation, we know $x^* = (A^TA)^{-1}(A^Tb)$ is an optimal solution to $\min_x \|Ax - b\|_2$. By Cramer's rule, all entries in x^* are either 0 or have absolute value at least $1/2^{\operatorname{poly}(nd)}$. This directly implies either $Ax^* - b = 0$ or $\|Ax^* - b\|_2 \ge 1/2^{\operatorname{poly}(nd)}$. □

Lemma F.5 implies that either $\|Ax - b\|_M = 0$ for some $x \in \mathbb{R}^d$, or $\|Ax - b\|_M \ge 1/2^{\operatorname{poly}(nd)}$ for all $x \in \mathbb{R}^d$. The former case can be solved by simply solving the linear system $Ax = b$. Thus we assume $\|Ax - b\|_M \ge 1/2^{\operatorname{poly}(nd)}$ for all $x \in \mathbb{R}^d$ in the rest part of this section.

To solve the Tukey regression problem $\min_x \|Ax-b\|_M$, we apply a binary search to find the optimal solution value OPT. Since $1/2^{\mathrm{poly}(nd)} \le \mathrm{OPT} \le n \cdot \tau^p \le 2^{\mathrm{poly}(nd)}$ by Assumption 1 and Assumption 2, the binary search makes at most $\log(2^{\mathrm{poly}(nd)}/\varepsilon) = \mathrm{poly}(nd) + \log(1/\varepsilon)$ guesses to the value of OPT to find a $(1+\varepsilon)$-approximate solution.

For each guess λ, we need to decide whether there exists $x \in \mathbb{R}^d$ such that $\|Ax-b\|_M \le \lambda$ or not. We use the polynomial system verifier in Theorem F.4 to solve this problem. We first enumerate a set of coordinates $S \subseteq [n]$, which are the coordinates with $|(Ax^*-b)_i| \ge \tau$, where $x^* = \mathrm{argmin}_x \|Ax-b\|_M$, and then solve the following decision problem:

$$\begin{aligned}&\sum_{i\in[n]\setminus S} P(\sigma_i(Ax-b)_i) + |S|\cdot\tau^p \le \lambda\\ \text{s.t } &\sigma_i^2 = 1, \forall i \in [n]\setminus S\\ &0 \le \sigma_i(Ax-b)_i \le \tau, \forall i \in [n]\setminus S.\end{aligned}$$

Clearly, $\sigma_i(Ax-b)_i = |(Ax-b)_i|$, and thus $L_P P(\sigma_i(Ax-b)_i) \le M((Ax-b)_i) \le U_P P(\sigma_i(Ax-b)_i)$. Thus by Assumption 1, for all $x \in \mathbb{R}^d$ and $S \subseteq [n]$,

$$L_P\|Ax-b\|_M \le \sum_{i\in[n]\setminus S} P(\sigma_i(Ax-b)_i) + |S|\cdot\tau^p.$$

Moreover,

$$\sum_{i\in[n]\setminus S} P(\sigma_i(Ax^*-b)_i) + |S|\cdot\tau^p \le U_P\|Ax^*-b\|_M$$

when $S = \{i \in [n] \mid |(Ax^*-b)_i| \ge \tau\}$, which implies the binary search will return a $((1+\varepsilon)\cdot U_P/L_P)$-approximate solution.

Now we analyze the running time of the algorithm. We make at most $\mathrm{poly}(nd) + \log(1/\varepsilon)$ guesses to the value of OPT. For each guess, we enumerate a set of coordinates S, which takes $O(2^n)$ time. For each set $S \subseteq [n]$, we need to solve the decision problem mentioned above, which has $n+d$ variables and $O(n)$ polynomial constraints with degree at most D. By Theorem F.4 this decision problem can be solved in $(nD)^{O(n)}$ time. Thus, the overall time complexity is upper bounded by $(nD)^{O(n)} \cdot \log(1/\varepsilon)$.

Notice that we can apply the row sampling algorithm in Theorem D.8 to reduce the size of the problem before applying this algorithm. This reduces the running time from $(nD)^{O(n)} \cdot \log(1/\varepsilon) = 2^{O(n\cdot(\log n + \log D))} \cdot \log(1/\varepsilon)$ to $2^{\widetilde{O}(\log D \cdot d^{p/2}\,\mathrm{poly}(d\log n)/\varepsilon^2)}$. Formally, we have the following theorem.

Theorem F.6. *Under Assumption 1 and 2, and suppose all entries in A and b are integers, and there exists a polynomial $P(x)$ with degree D and two constants $L_P \le 1 \le U_M$ such that when $|x| \le \tau$, we have*

$$L_P P(|x|) \le M(|x|) \le U_P P(|x|).$$

Then there exists an algorithm that returns a $((1+\varepsilon)\cdot U_P/L_P)$-approximate solution to $\min_x \|Ax-b\|_M$ and runs in $2^{\widetilde{O}(\log D \cdot d^{p/2}\,\mathrm{poly}(d\log n)/\varepsilon^2)}$ time.

Corollary F.7. *Under Assumption 2, and suppose all entries in A and b are integers, for the loss function M defined in (1) there exists an algorithm that returns a $(1+\varepsilon)$-approximate solution to $\min_x \|Ax-b\|_M$ and runs in $2^{\widetilde{O}(\mathrm{poly}(d\log n)/\varepsilon^2)}$ time.*