Appendix

A Related Works

Existing literature studied the robust regression with respect to Huber loss function [1, 2]. Such regression can be applied to solve many problems like the people counting problem [3]. To speed up the regression process, some dimensional reduction techniques can be used to reduce the number of observations [4], also faster algorithms have been proposed to address the robust regression with reasonable assumption [5]. Besides, different models of regression were explored, such as Gaussian process regression [6], active regression with adaptive Huber loss [7].

Recent years, there are lots of randomized sketching and embedding techniques developed for solving numerical linear algebra problems. There is a long line of works, e.g. [8, 9, 10] for ℓ_2 subspace embedding, and works, e.g. [11, 12, 13, 14] for ℓ_p subspace embedding. For more related works, we refer readers to the book [15]. Based on sketching/embedding techniques, there is a line of works studied ℓ_2 and ℓ_p regressions, e.g. [9, 16, 12, 10, 13]. [17] studied linear regression with M-estimator error measure. We refer to the survey [18] for more details.

Frobenius norm low rank matrix approximation problem is also known as PCA problem. This problem is well studied. The fastest algorithm is shown by [9]. For the entrywise ℓ_p norm low rank approximation problem, there is no known algorithm with theoretical guarantee until the work [19]. [19] works only for $1 \leq p \leq 2$. Recently, [20] gives algorithms for all $p \geq 1$. But either the running time is not in polynomial or the rank of the output is not exact k.

B Proof of Fact 2.

Proof. Notice that G_1 is a nonzero nondecreasing convex function on $\mathbb{R}_+$, thus $G_1^{-1}(1)$ exists, and G_2 is a nonzero nondecreasing function. In addition because $s = \sup\left\{\frac{1}{y-x}\left(G_2(y) - G_2(x)\right) \mid 0 \leq x \leq y \leq 1\right\}$, G_2 is also convex. Thus $\|\cdot\|_{G_2}$ is Orlicz norm. Let $x \in \mathbb{R}^n$. Notice that if $\alpha > 0$ satisfies $\sum_{i=1}^n G_1(|x_i|/\alpha) \leq 1$, then $\forall i \in [n], G_1(|x_i|/\alpha) \leq 1$. It means that $|x_i| \leq G_1^{-1}(1)\alpha$, thus $\sum_{i=1}^n G_2(|x_i|/(G_1^{-1}(1)\alpha)) = \sum_{i=1}^n G_1(|x_i|/\alpha) \leq 1$. Similarly if α satisfies $\sum_{i=1}^n G_2(|x_i|/\alpha) \leq 1$, then $\sum_{i=1}^n G_1(G_1^{-1}(1)|x_i|/\alpha) = \sum_{i=1}^n G_2(|x_i|/\alpha) \leq 1$. Therefore, $\|x\|_{G_1} = \|x\|_{G_2}/G_1^{-1}(1)$. □

C Proof of Lemma 3.

Due to convexity of G and $G(1) = 1, G(0) = 0$, $\forall x \in [0,1], G(x) \leq xG(1) + (1-x)G(0) = x$. Since $x \leq 1$, $G(1)/G(x) \leq C_G(1/x)^2$, we have $G(x) \geq x^2/C_G$.

D Proof of Lemma 4.

With out loss of generality, we can assume $\forall i \in [n], x_i \geq 0$. Let $x \in \mathbb{R}^n, \alpha = \|x\|_G$. We have $\sum_{i=1}^n G(x_i/\alpha) = 1$. If $x_i/\alpha \leq 1$, due to the convexity of G, $G(x_i/\alpha) \leq G(1)\cdot x_i/\alpha + G(0)\cdot(1 - x_i/\alpha) =$

$G(1) \cdot x_i/\alpha = x_i/\alpha$. If $x_i/\alpha > 1$, then $G(x_i/\alpha) > 1$ which contradicts to $\sum_{i=1}^n G(x_i/\alpha) = 1$. Thus, $\|x\|_G \leq \|x\|_1$.

$\|x/\alpha\|_2^2 = \sum_{i=1}^n (x_i/\alpha)^2 \leq \sum_{i=1}^n C_G G(x_i/\alpha) = C_G$. Then

$$\|x\|_2 \leq \sqrt{C_G}\alpha.$$

E Proof of Lemma 5.

Due to the convexity of $G(\cdot)$ and $G(0) = 0, \forall 0 < x < y$, we have $G(x) \leq G(y)x/y + G(0)(1 - x/y) = G(y)x/y$. Thus, $y/x \leq G(y)/G(x)$.

F Proof of Lemma 6.

It is easy to see that $\forall x > 0, G(x) \neq 0$, since otherwise for $y > x$, the condition $G(y)/G(x) < C_G(y/x)^2$ would be violated. Let $s = G'_+(1)$. There are several cases.

If $a \geq 1$ or $b \geq 1$. Without loss of generality, assume $a \geq 1$.$G(a)G(b)/G(ab) = (sa - (s - 1))G(b)/G(ab) \leq saG(b)/G(ab)$. since $ab \geq b$, $G(ab)/G(b) \geq a$. Therefore, $G(a)G(b)/G(ab) \leq saG(b)/G(ab) \leq s$.

If $a, b \leq 1, 0.5 \leq a \leq 1$ or $0.5 \leq b \leq 1$, we want to show $G(a)G(b)/G(ab)$ is still bounded. Without loss of generality, assume that $0.5 \leq a \leq 1$. Then $G(a)G(b)/G(ab) = G(a)\frac{G(b)}{G(ab)} \leq G(b)/G(ab) \leq C_G/a^2 \leq 4C_G$.

If $a, b \leq 0.5$ and $G'(0) > 0$. Let $G'(0) = c > 0$. Therefore, there is a constant δ_1 which may depend on G such that $\forall x \in (0, \delta_1)$, $|\frac{G(x)-G(0)}{x-0} - c| < \frac{c}{2}$. Therefore, $\forall x \in (0, \delta_1]$, $G(x) > \frac{c}{2}x$. Due to Lemma 5, $\forall x > \delta_1$, $G(x)/x > G(\delta_1)/\delta_1 > c/2$. Therefore, $\forall x, G(x) \geq \frac{c}{2}x$. Since $b \leq 0.5, ab \leq a \leq 1$. Since G is convex, $G(a) \leq \frac{1-a}{1-ab}G(ab) + \frac{a-ab}{1-ab}G(1) = \frac{1-a}{1-ab}G(ab) + \frac{a-ab}{1-ab}$. Therefore, $G(a) \leq G(ab) + (1-a)/(1-ab) \leq G(ab) + 2a$. Similarly, $G(b) \leq 2b + G(ab)$. Then we have $G(a)G(b) \leq (2b + G(ab))(2a + G(ab)) \Rightarrow G(a)G(b)/G(ab) \leq ab/G(ab) + (2a + 2b) + G(ab) \leq 2/c + 2 + 1 \leq 2/c + 3$.

If $a, b \leq 0.5, G'_+(0) = 0, G''_+(0) = c_2 > 0$. Let $\epsilon = c_2/4$. Since G is twice differentiable in $(0, \delta_G)$ and $G'_+(0), G''_+(0)$ exist, by Taylor's Theorem, there is a constant $\delta_2 > 0$ which may depend on G such that $|G(x) - (G(0) + G'_+(0)x + c_2x^2/2)| \leq \epsilon x^2$. Therefore, $\forall x \in (0, \delta_2), G(x) \geq c_2x^2/4, G(x) \leq c_2x^2$. Hence, $\forall a, b \in (0, \delta_2]$, $G(a)G(b)/G(ab) \leq \frac{G(a)G(b)}{c_2a^2b^2/4} \leq \frac{4}{c_2}\frac{G(a)}{a^2}\frac{G(b)}{b^2} \leq \frac{4}{c_2}c_2^2 = 4c_2$. Consider a or $b > \delta_2$. Without loss of generality, assume $a > \delta_2$. Similar to the previous argument, $G(a)G(b)/G(ab) \leq G(a)\frac{G(b)}{G(ab)} \leq G(b)/G(ab) \leq C_Gb^2/(ab)^2 \leq C_G/\delta_2^2$. Thus $G(a)G(b)/G(ab)$ is bounded by C_G/δ_2^2 in this case.

G Proof of Theorem 9

Without loss of generality, we assume $\forall i \in [n], x_i \geq 0$. Fix $i \in [n]$, we have

$$\Pr(u_i \geq G^{-1}(1/n^{20})) = e^{-G(G^{-1}(1/n^{20}))} \geq 1 - 1/n^{20}.$$

Define $\mathcal{E}$ to be the event that $\forall i \in [n], u_i \geq G^{-1}(1/n^{20})$. By taking union bound over n coordinates, $\mathcal{E}$ happens with probability at least $1 - 1/n^{19}$. Let $\alpha = \|x\|_G$. Then, for any $\gamma \geq 1$, we have

$$\begin{aligned}
&\Pr(\|f(x)\|_G \geq \gamma\alpha) \\
=&\Pr(\|f(x)\|_G \geq \gamma\alpha \mid \mathcal{E}) \Pr(\mathcal{E}) + \Pr(\|f(x)\|_G \geq \gamma\alpha \mid \neg\mathcal{E}) \Pr(\neg\mathcal{E}) \\
\leq&\Pr(\|f(x)\|_G \geq \gamma\alpha \mid \mathcal{E}) \Pr(\mathcal{E}) + \Pr(\neg\mathcal{E}) \\
=&\Pr(\|f(x)/(\gamma\alpha)\|_G \geq 1 \mid \mathcal{E}) \Pr(\mathcal{E}) + \Pr(\neg\mathcal{E}) \\
\leq&\mathbf{E}\left(\sum_{i=1}^{n} G\left(x_i/\alpha \cdot \frac{1}{\gamma u_i}\right) \mid \mathcal{E}\right) \Pr(\mathcal{E}) + 1/n^{19} \\
=&\sum_{i=1}^{n} \mathbf{E}\left(G\left(x_i/\alpha \cdot \frac{1}{\gamma u_i}\right) \mid \mathcal{E}\right) \Pr(\mathcal{E}) + 1/n^{19}.
\end{aligned}$$

Let $r = G^{-1}(1/n^{20})$. For a fixed $i \in [n]$,

$$\begin{aligned}
&\mathbf{E}\left(G\left(x_i/\alpha \cdot \frac{1}{\gamma u_i}\right) \mid \mathcal{E}\right) \Pr(\mathcal{E}) \\
=&\int_r^{\infty} G\left(\frac{x_i/\alpha}{u\gamma}\right) e^{-G(u)} G'(u) \mathrm{d}u \\
\leq&\frac{1}{\gamma} G(x_i/\alpha) \int_1^{\infty} e^{-G(u)} \mathrm{d}G + \frac{1}{\gamma} \int_r^1 G\left(\frac{x_i/\alpha}{u}\right) e^{-G(u)} \mathrm{d}G \\
\leq&\frac{1}{\gamma} G(x_i/\alpha) + \frac{1}{\gamma} \int_r^1 G\left(\frac{x_i/\alpha}{u}\right) e^{-G(u)} \mathrm{d}G \\
\leq&\frac{1}{\gamma} G(x_i/\alpha) + \frac{1}{\gamma} \alpha_G G(x_i/\alpha) \int_r^1 \frac{1}{G(u)} e^{-G(u)} \mathrm{d}G \\
\leq&O(\log n) \frac{\alpha_G}{\gamma} G(x_i/\alpha),
\end{aligned}$$

where α_G is a constant may depend on $G(\cdot)$. The first inequality follows by $G(x_i/\alpha \cdot 1/(\gamma u)) \leq 1/\gamma \cdot G(x_i/\alpha \cdot 1/u) + (1 - 1/\gamma) \cdot G(0) \leq G(x_i/\alpha \cdot 1/u)/\gamma \leq G(x_i/\alpha)/\gamma$. The second inequality follows by $\int_r^{\infty} e^{-x} \mathrm{d}x \leq 1$. The third inequality follows by Lemma 6. Since $x_i/\alpha \leq 1$, then there is an α_G such that $G(u)G(x_i/\alpha \cdot 1/u) \leq \alpha_G G(x_i/\alpha)$. The last inequality follows by $\int_{1/n^{20}}^{\infty} e^{-x}/x \mathrm{d}x = O(\log n)$.

Thus, we have

$$\sum_{i=1}^{n} \mathbf{E}\left(G\left(x_i/\alpha \cdot \frac{1}{\gamma u_i}\right) \mid \mathcal{E}\right) \Pr(\mathcal{E}) \leq O(\log n) \frac{\alpha_G}{\gamma} \sum_{i=1}^{n} G(x_i/\alpha) \leq O(\log n) \frac{\alpha_G}{\gamma}.$$

Then,

$$\Pr(\|f(x)\|_G \geq \gamma\alpha) \leq O(\log n) \frac{\alpha_G}{\gamma} + 1/n^{19}.$$

It is equivalent to

$$\Pr(\|f(x)\|_G \leq \gamma\alpha) \geq 1 - O(\log n) \frac{\alpha_G}{\gamma} - 1/n^{19}.$$

Set $\gamma = O(\log n) \frac{\alpha_G}{\delta}$, we complete the proof.

H Proof of Theorem 10

Similar to [21], we can define a well conditioned basis for Orlicz norm.

Definition H.1 (Well conditioned basis for Orlicz norm)**.** *Given a matrix $A \in \mathbb{R}^{n\times m}$ with rank d, let $U \in \mathbb{R}^{n\times d}$ be a matrix which has the same column space of A. If U satisfies 1. $\forall x \in \mathbb{R}^d$, $\|x\|_\infty \leq \beta \|Ux\|_G$, 2. $\sum_{i=1}^d \|U_i\|_G \leq \alpha$, then U is an (α, β, G)-well conditioned basis of A.*

Fortunately, the such good basis exists for Orlicz norm.

Theorem H.2 (See Connection to Auerbach basis in Section 3.1 of [21])**.** *Given a matrix $A \in \mathbb{R}^{n\times m}$ with rank d and norm $\|\cdot\|_G$, there exist a matrix $U \in \mathbb{R}^{n\times d}$ which is a $(d, 1, G)$ well conditioned basis of A.*

Proof of Theorem 10. Notice that $D^{-1}Ax$ is exactly the same as $f(Ax)$. There is a matrix $U \in \mathbb{R}^{n\times d}$ which is $(d, 1, G)$-well conditioned basis of A. Since $\forall x \in \mathbb{R}^m$, there is always a vector $y \in \mathbb{R}^d$ such that $Ax = Uy$, we only need to prove that with probability at least 0.99,

$$\forall x \in \mathbb{R}^d, \|D^{-1}Ux\|_G \leq O(\alpha_G d^2 \log n)\|Ux\|_G,$$

where D, α_G are the same as stated in the Theorem. According to Theorem 9, if we look at a fixed $i \in [d]$, then with probability at least $1 - 0.01/d$, $\|D^{-1}U_i\|_G \leq O(\alpha_G d \log(n))$. By taking union bound, with probability at least 0.99, $\forall i \in [d]$, $\|D^{-1}U_i\|_G \leq O(\alpha_G d \log(n))$. Now we have, for any $x \in \mathbb{R}^d$,

$$\begin{aligned}\|D^{-1}Ux\|_G &\leq \sum_{i=1}^d |x_i| \|D^{-1}U_i\|_G \leq \|x\|_\infty \sum_{i=1}^d \|D^{-1}U_i\|_G \leq O(\alpha_G d \log(n))\|x\|_\infty \sum_{i=1}^d \|U_i\|_G \\ &\leq O(\alpha_G d^2 \log(n))\|Ux\|_G.\end{aligned}$$

The first inequality follows by triangle inequality. The third inequality follows by $\forall i \in [d]$, $\|D^{-1}U_i\|_G \leq O(\alpha_G d \log(n))$. The forth inequality follows by $(d, 1, G)$-well conditioned basis. □

I Proof of Theorem 12

Now, in the following, we present the concept of ε-net.

Definition I.1 (ε-net for ℓ_2 norm)**.** *Given $A \in \mathbb{R}^{n\times m}$ with rank d, let S be the ℓ_2 unit ball in the column space of A, i.e. $S = \{y \mid \|y\|_2 = 1, \exists x \in \mathbb{R}^m, y = Ax\}$. Let $N \subset S$, if $\forall x \in S, \exists y \in N$ such that $\|x - y\|_2 \leq \varepsilon$, then we say N is an ε-net for S.*

The following theorem gives an upper bound of the size of ε-net.

Theorem I.2 (Lemma 2.2 of [15])**.** *Given $A \in \mathbb{R}^{n\times m}$ with rank d, let S be the ℓ_2 unit ball in the column space of A. There exist an ε-net N for S, such that $|N| \leq (1 + 4/\varepsilon)^d$.*

It suffices to prove $\forall x \in \mathbb{R}^m, \|Ax\|_2 = 1$ we have $\Omega(1/(\alpha'_G d \log n))\|Ax\|_G \leq \|D^{-1}Ax\|_\infty$. Let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix of which each entry on the diagonal is an i.i.d. random variable drawn from the distribution with CDF $1 - e^{-G(t)}$. Let $\alpha'_G \geq 1$ be a sufficiently large constant. Let S be the ℓ_2 unit ball in the column space of A. Let $t_1 = \Theta(\alpha'_G d \log n), t_2 = \Theta(\alpha_G d^2 \log n)$, where

α_G is the parameter stated in Theorem 10. Set $\varepsilon = O(1/(\sqrt{n}C_G t_1 t_2))$. There exist an ε-net N for S, and

$$|N| = e^{O(d(\log n + \log(C_G \alpha'_G \alpha_G)))}.$$

By taking union bound over the net points, according to Theorem 11, with probability at least 0.99,

$$\forall x \in N, \|D^{-1}x\|_\infty \geq \Omega(1/(\alpha'_G d \log n))\|x\|_G. \tag{1}$$

Also due to Theorem 10, with probability at least 0.99,

$$\forall x \in S, \|D^{-1}x\|_G \leq O(\alpha_G d^2 \log n)\|x\|_G. \tag{2}$$

By taking union bound, with probability at least 0.98, the above two events will happen. Then, in this case, consider a $y \in S$, let $x \in N$ such that $\|x - y\|_2 \leq \varepsilon$, let $z = x - y$, we have

$$\begin{aligned}
\|D^{-1}y\|_\infty &\geq \|D^{-1}x\|_\infty - \|D^{-1}z\|_\infty \\
&\geq \frac{1}{t_1}\|x\|_G - t_2\sqrt{C_G}\|z\|_G \\
&\geq \frac{1}{t_1}\|y\|_G - \frac{t_2}{t_1}\|z\|_G - t_2\sqrt{C_G}\|z\|_G \\
&\geq \frac{1}{t_1}\|y\|_G - 2t_2\sqrt{C_G}\|z\|_G \\
&\geq \frac{1}{t_1}\|y\|_G - O(\frac{2}{\sqrt{C_G}t_1}) \\
&\geq \Omega(1/t_1)\|y\|_G \\
&= \Omega(1/(\alpha'_G d \log n))\|y\|_G.
\end{aligned}$$

The first inequality follows by triangle inequality. The second inequality follows by Equation 1, Equation 2, and Lemma 4, i.e. $\|D^{-1}z\|_\infty \leq \|D^{-1}z\|_2 \leq \sqrt{C_G}\|D^{-1}z\|_G$. The third inequality follows by triangle inequality. The forth inequality follows by $t_1, C_G \geq 1$. The fifth inequality follows by Lemma 4: $\|z\|_G \leq \|z\|_1 \leq \sqrt{n}\|z\|_2 = \sqrt{n}\varepsilon = O(1/(C_G t_1))$. The sixth inequality follows by Lemma 4: $\|y\|_G \geq \frac{1}{\sqrt{C_G}}\|y\|_2 = 1/\sqrt{C_G}$.

J Proof of Theorem 13

Due to Theorem 12 and Theorem 10, with probability at least 0.98, $\forall x \in \mathbb{R}^m$,

$$\begin{aligned}
&\Omega(1/(\alpha'_G d \log n))\|Ax\|_G \leq \|D^{-1}Ax\|_\infty \leq \|D^{-1}Ax\|_2. \\
&\|D^{-1}Ax\|_2 \leq \sqrt{C_G}\|D^{-1}Ax\|_G \leq O(\sqrt{C_G}\alpha'_G d^2 \log n)\|Ax\|_G.
\end{aligned}$$

K Proof of Theorem 16

Due to Theorem 14 and Theorem 15, with probability at least 0.95, $\forall x$, $\|\Pi_2\Pi_1 D^{-1}Ax\|_2$ is a constant approximation to $\|\Pi_1 D^{-1}Ax\|_2$ and $\|\Pi_1 D^{-1}Ax\|_2$ is a constant approximation to $\|D^{-1}Ax\|_2$. Combining with Theorem 13, we complete the proof.

L Proof of Theorem 18

Let $x^* = \arg\min_{x\in\mathbb{R}^d} \|Ax-b\|_G$. Due to Theorem 9, with probability at least 0.99,

$$\|D^{-1}(Ax^*-b)\|_G \leq O(\alpha_G \log n)\|Ax^*-b\|_G. \tag{3}$$

Now let $A' = [A\ b] \in \mathbb{R}^{n\times(d+1)}$. Due to Theorem 16, with probability at least 0.9, we have

$$\forall x \in \mathbb{R}^{d+1}, \Omega(1/(\alpha'_G d\log n))\|A'x\|_G \leq \|\Pi_2\Pi_1 D^{-1}A'x\|_2. \tag{4}$$

Then,

$$\begin{aligned}
\|A\hat{x}-b\|_G &\leq O(\alpha'_G d\log n)\|\Pi_2\Pi_1 D^{-1}(A\hat{x}-b)\|_2 \\
&\leq O(\alpha'_G d\log n)\|\Pi_2\Pi_1 D^{-1}(Ax^*-b)\|_2 \\
&\leq O(\alpha'_G d\log n)\|D^{-1}(Ax^*-b)\|_2 \\
&\leq O(\alpha'_G \sqrt{C_G} d\log n)\|D^{-1}(Ax^*-b)\|_G \\
&\leq O(\alpha_G\alpha'_G \sqrt{C_G} d\log^2 n)\|Ax^*-b\|_G.
\end{aligned}$$

The first inequality follows by Equation 4. The second inequality follows by $\hat{x} = (\Pi_2\Pi_1 D^{-1}A)^\dagger \Pi_2\Pi_1 D^{-1}b$, which is the optimal solution for $\min_{x\in\mathbb{R}^d}\|\Pi_2\Pi_1 D^{-1}(Ax-b)\|_2$. The third inequality follows by Theorem 14 and Theorem 15. The forth inequality follows by Lemma 4. The last inequality follows by Equation 3. Let $\beta_G = \alpha'_G\sqrt{C_G}$, we complete the proof of the correctness of Algorithm 1.

For the running time, according to Theorem 16, computing $\Pi_2\Pi_1 D^{-1}A$ and $\Pi_2\Pi_1 D^{-1}b$ needs $\mathrm{nnz(A)} + \mathrm{poly}(d)$ time. Since $\Pi_2\Pi_1 D^{-1}A$ has size $\mathrm{poly}(d)$, computing $\hat{x} = (\Pi_2\Pi_1 D^{-1}A)^\dagger\Pi_2\Pi_1 D^{-1}b$ needs $\mathrm{poly}(d)$ running time. The total running time is $\mathrm{nnz(A)} + \mathrm{poly}(d)$.

M proof of Lemma 20

Before we prove the Lemma, we need following tools.

Lemma M.1 (Concentration bound for sum of half normal random variables)**.** *For any k i.i.d. random Gaussian variables $z_1, z_2, \cdots, z_k$, we have that*

$$\Pr\left(\frac{1}{k}\sum_{i=1}^k |z_i| \in \left((1-\varepsilon)\sqrt{2/\pi}, (1+\varepsilon)\sqrt{2/\pi}\right)\right) \geq 1 - e^{-\Omega(k\varepsilon^2)}.$$

Lemma M.2. *Let $G \in \mathbb{R}^{k\times m}$ be a random matrix with each entry drawn uniformly from i.i.d. $N(0,1)$ Gaussian distribution. With probability at least 0.99, $|G|_2 \leq 10\sqrt{km}$.*

Proof. Since $\mathbf{E}\left(\|G\|_F^2 = km\right)$, we have that $\Pr(\|G\|_F^2 \geq 100km) \leq 0.01$. Thus, with probability at least 0.99, we have $\|G\|_2 \leq \|G\|_F \leq 10\sqrt{km}$. □

Now, let us prove the lemma.

Proof of Lemma 20. Without loss of generality, we only need to prove $\forall x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, we have $\|Bx\|_1 \in (1-\varepsilon, 1+\varepsilon)$. Let set $S = \{v \mid v \in \mathbb{R}^n, \|v\|_2 = 1\}$. Due to Theorem I.2, we can find a set $G \subset S$ which satisfies that $\forall u \in S$ there exists $v \in G$ such that $\|u-v\|_2 \leq (\varepsilon/(1000n))^{10}$ and $|G| \leq (4000n/\varepsilon)^{20n}$. Let $k \geq c\varepsilon^{-2}n\ln(n/\varepsilon)$ where c is a sufficiently large constant. By Lemma M.1,

we have that for a fixed $v \in G$, with probability at least $1 - e^{-1000n \ln(4000n/\varepsilon)}$, $\|Bv\|_1 \in (1-\varepsilon, 1+\varepsilon)$. By taking union bound over all the points in G, we have

$$\Pr\left(\forall v \in G, \|Bv\|_1 \in (1-\varepsilon, 1+\varepsilon)\right) \geq 1 - e^{-980n \ln(4000n/\varepsilon)} \geq 0.99.$$

Now, consider $\forall x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, i.e. $x \in S$, we can find $v \in G$ such that $\|v - x\|_2 \leq (\varepsilon/(1000n))^{10}$, and let $u = v - x$. Then, conditioned on $\|B\|_2 \leq 10\sqrt{tn} \cdot \sqrt{\pi/2}/t$, we have

$$\begin{aligned}
\|Bx\|_1 &\in (\|Bv\|_1 - \|Bu\|_1, \|Bv\|_1 + \|Bu\|_1) \\
&\subseteq (1 - (\varepsilon + \sqrt{t}\|B\|_2\|u\|_2), 1 + (\varepsilon + \sqrt{t}\|B\|_2\|u\|_2)) \\
&\subseteq (1 - 2\varepsilon, 1 + 2\varepsilon)
\end{aligned}$$

where the first relation follows by triangle inequality, the second relation follows by $\|Bu\|_1 \leq \sqrt{t}\|Bu\|_2 \leq \sqrt{t}\|B\|_2\|u\|_2$, and the last relation follows by $\|u\|_2 \leq (\varepsilon/(1000n))^{10}, \|B\|_2 \leq 10\sqrt{tn} \cdot \sqrt{\pi/2}/t$.

According to Lemma M.2, we know that with probability at least 0.99, we have $\|B\|_2 \leq 10\sqrt{tn} \cdot \sqrt{\pi/2}/t$. By taking union bound, we have with probability at least 0.98, $\forall x \in S$, $\|Bx\|_1 \in (1 - 2\varepsilon, 1 + 2\varepsilon)$. By adjusting the ε, we complete the proof. □

N Proof of Theorem 21

Without loss of generality, we assume constant $k \leq 2$. Otherwise, we can always adjust constants in all the related theorems and lemmas to make larger k work. Let $x^* = \arg\min_{x \in \mathbb{R}^d} \sum_{i=1}^k \|A_i x - b_i\|_{G_i}$. By Theorem 9 and taking union bound, we have that with probability at least 0.98,

$$\forall i \in \{1, 2, \cdots, k\}, \|(D^{(i)})^{-1}(A_i x^* - b_i)\|_{G_i} \leq O(\alpha_{G_i} \log n)\|A_i x^* - b_i\|_{G_i}. \tag{5}$$

Now let $A_i' = [A_i \ b_i] \in \mathbb{R}^{n_i \times (d+1)}$. Due to Theorem 16 and union bound, with probability at least 0.8, we have

$$\forall x \in \mathbb{R}^{d+1}, \Omega(1/(\alpha_{G_i}' d \log n_i))\|A_i' x\|_{G_i} \leq \|\Pi_2^{(i)} \Pi_1^{(i)} (D^{(i)})^{-1} A_i' x\|_2. \tag{6}$$

Then,

$$\begin{aligned}
\sum_{i=1}^{k} \|A_i\hat{x} - b_i\|_{G_i} &\leq \sum_{i=1}^{k} O(\alpha'_{G_i} d \log n) \|\Pi_2^{(i)} \Pi_1^{(i)} (D^{(i)})^{-1} (A_i\hat{x} - b_i)\|_2 \\
&\leq \sum_{i=1}^{k} O(\alpha'_{G_i} d \log n) \|B^{(i)} \Pi_2^{(i)} \Pi_1^{(i)} (D^{(i)})^{-1} (A_i\hat{x} - b_i)\|_1 \\
&\leq O(\max_{i\in[k]} \alpha'_{G_i} d \log n) \|B\Pi_2\Pi_1 D^{-1}(A\hat{x} - b)\|_1 \\
&\leq O(\max_{i\in[k]} \alpha'_{G_i} d \log n) \|B\Pi_2\Pi_1 D^{-1}(Ax^* - b)\|_1 \\
&\leq O(\max_{i\in[k]} \alpha'_{G_i} d \log n) \sum_{i=1}^{k} \|B^{(i)} \Pi_2^{(i)} \Pi_1^{(i)} (D^{(i)})^{-1} (A_i x^* - b_i)\|_1 \\
&\leq O(\max_{i\in[k]} \alpha'_{G_i} d \log n) \sum_{i=1}^{k} \|\Pi_2^{(i)} \Pi_1^{(i)} (D^{(i)})^{-1} (A_i x^* - b_i)\|_2 \\
&\leq O(\max_{i\in[k]} \alpha'_{G_i} d \log n) \sum_{i=1}^{k} \|(D^{(i)})^{-1} (A_i x^* - b_i)\|_2 \\
&\leq O((\max_{i\in[k]} \sqrt{C_{G_i}})(\max_{i\in[k]} \alpha'_{G_i}) d \log n) \sum_{i=1}^{k} \|(D^{(i)})^{-1} (A_i x^* - b_i)\|_{G_i} \\
&\leq O((\max_{i\in[k]} \alpha_{G_i})(\max_{i\in[k]} \sqrt{C_{G_i}})(\max_{i\in[k]} \alpha'_{G_i}) d \log^2 n) \sum_{i=1}^{k} \|A_i x^* - b_i\|_{G_i}.
\end{aligned}$$

The first inequality follows by Equation 6. The second inequality follows by Lemma 20. The forth inequality follows by $\hat{x}$ is the optimal solution for $\min_{x\in\mathbb{R}^d} \|B\Pi_2\Pi_1 D^{-1}(Ax - b)\|_1$. The sixth inequality follows by Lemma 20. The seventh inequality follows by Theorem 14 and Theorem 15. The eighth inequality follows by Lemma 4. The last inequality follows by Equation 5. Let $\beta'_G = (\max_{i\in[k]} \alpha_{G_i})(\max_{i\in[k]} \sqrt{C_{G_i}})(\max_{i\in[k]} \alpha'_{G_i})$, we complete the proof of the correctness of Algorithm 2.

For the running time, according to Theorem 16, computing $\Pi_2\Pi_1 D^{-1}A$ and $\Pi_2\Pi_1 D^{-1}b$ needs $\sum_{i=1}^{k} \mathrm{nnz(A_i)} + \mathrm{poly}(d)$ time. Due to Lemma 20, the size of B is $\mathrm{poly}(d)$. To compute $B\Pi_2\Pi_1 D^{-1}A$ and $B\Pi_2\Pi_1 D^{-1}b$, we need additional $\mathrm{poly}(d)$ time. Since $B\Pi_2\Pi_1 D^{-1}A$ has size $\mathrm{poly}(d)$, computing the optimal solution of $\min_{x\in\mathbb{R}^d} \|B\Pi_2\Pi_1 D^{-1}(Ax - b)\|_1$ by using linear programming needs $\mathrm{poly}(d)$ running time. The total running time is $\sum_{i=1}^{k} \mathrm{nnz(A_i)} + \mathrm{poly}(d)$.

O Proof of Theorem 23

Before we prove the Theorem, we need to show following Lemmas.

Lemma O.1 ([19]). *Let $A \in \mathbb{R}^{n\times d}, R \in \mathbb{R}^{d\times t_3}, k$ be the same as in the Algorithm 3, then with probability at least* 0.9,

$$\min_{X\in\mathbb{R}^{t_3\times k}, Y\in\mathbb{R}^{k\times d}} \|ARXY - A\|_p^p \leq O((k\log k)^{1-p/2} \log n) \min_{U\in\mathbb{R}^{n\times k}, V\in\mathbb{R}^{k\times d}} \|UV - A\|_p^p.$$

Lemma O.2 ([13]). *Let $1 \leq p \leq 2$. Given a matrix $A \in \mathbb{R}^{n\times d}$, $d \leq n$, let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix of which each entry on the diagonal is an i.i.d. random variable drawn from the*

distribution with CDF $1-e^{-t^p}$. *Let* $\Pi_1 \in \mathbb{R}^{t_1 \times n}$ *be a sparse embedding matrix (see Theorem 18) and let* $\Pi_2 \in \mathbb{R}^{t_2 \times t_1}$ *be a random Gaussian matrix (see Theorem 19) where* $t_1 = \Omega(d^2), t_2 = \Omega(d)$. *Then, with probability at least* 0.9,

$$\forall x \in \mathbb{R}^d, \Omega(1/\min\{(d\log d)^{1/p}, (d\log d\log n)^{1/p-1/2}\})\|Ax\|_p \leq \|\Pi_2\Pi_1 D^{-1}Ax\|_2.$$

Lemma O.3. *Let* $A \in \mathbb{R}^{n\times d}, S \in \mathbb{R}^{t_2\times n}, R \in \mathbb{R}^{d\times t_3}, k$ *be the same as in the Algorithm 3, then with probability at least* 0.9,

$$\min_{X\in\mathbb{R}^{t_2\times k}, Y\in\mathbb{R}^{k\times t_2}} \|ARXYSA - A\|_p^p \leq \beta \min_{U\in\mathbb{R}^{n\times k}, V\in\mathbb{R}^{k\times d}} \|UV - A\|_p^p,$$

where $\beta = O(\min((k\log k)^{2-p/2}\log^{p+1} n, (k\log k)^{2-p}\log^{2+p/2} n))$.

Proof. Let $X^*, V^* = \arg\min_{X\in\mathbb{R}^{t_3\times k}, V\in\mathbb{R}^{k\times d}} \|ARXV - A\|_p^p$. Let $U^* = ARX^*, \tilde{V} = (SU^*)^\dagger SA$. Let $\gamma = \min\{k\log k, (k\log k\log n)^{1-p/2}\}$. We have

$$\begin{aligned}
\|U^*\tilde{V} - A\|_p^p &\leq 2^{p-1}\|U^*(\tilde{V} - V^*)\|_p^p + 2^{p-1}\|U^*V^* - A\|_p^p \\
&\leq O(\gamma)\sum_{i=1}^d \|SU^*(\tilde{V} - V^*)_i\|_2^p + 2^{p-1}\|U^*V^* - A\|_p^p \\
&\leq O(\gamma)\sum_{i=1}^d (\|S(U^*\tilde{V} - A)_i\|_2 + \|S(U^*V^* - A)_i\|_2)^p + 2^{p-1}\|U^*V^* - A\|_p^p \\
&\leq O(\gamma)\sum_{i=1}^d (2\|S(U^*V^* - A)_i\|_2)^p + 2^{p-1}\|U^*V^* - A\|_p^p \\
&\leq O(\gamma)\sum_{i=1}^d \|D_1^{-1}(U^*V^* - A)_i\|_2^p + 2^{p-1}\|U^*V^* - A\|_p^p \\
&\leq O(\gamma)\|D_1^{-1}(U^*V^* - A)\|_p^p + 2^{p-1}\|U^*V^* - A\|_p^p \\
&\leq O(\gamma)\log^p(nd)\|U^*V^* - A\|_p^p + 2^{p-1}\|U^*V^* - A\|_p^p \\
&= O(\gamma\log^p(n))\|U^*V^* - A\|_p^p.
\end{aligned}$$

The first inequality follows by convexity of x^p. The second inequality follows by Lemma O.2. The third inequality follows by triangle inequality. The forth inequality follows by $\tilde{V} = (SU^*)^\dagger SA$. The fifth inequality follows by Theorem 14 and Theorem 15. The sixth inequality follows by $p \leq 2$. The seventh inequality follows by Theorem 9.

Due to Lemma O.1, we have

$$\|U^*V^* - A\|_p^p \leq O((k\log k)^{1-p/2}\log n) \min_{U\in\mathbb{R}^{n\times k}, V\in\mathbb{R}^{k\times d}} \|UV - A\|_p^p.$$

Thus, we have

$$\begin{aligned}
&\min_{X,Y} \|ARXYSA - A\|_p^p \\
\leq &\|U^*\tilde{V} - A\|_p^p \\
\leq &O(\min((k\log k)^{2-p/2}\log^{p+1} n, (k\log k)^{2-p}\log^{2+p/2} n)) \min_{U,V} \|UV - A\|_p^p.
\end{aligned}$$

□

Lemma O.4 ([19]). *Let $A \in \mathbb{R}^{n\times d}, S \in \mathbb{R}^{t_2\times n}, R \in \mathbb{R}^{d\times t_3}, k, T_2 \in \mathbb{R}^{d\times t_3}$ be the same as in the Algorithm 3, then with probability at least 0.9, if for $\alpha \geq 1$, $\tilde{X}, \tilde{Y}$ satisfy*

$$\|AR\tilde{X}\tilde{Y}SAT_2 - AT_2\|_p^p \leq \alpha \min_{X,Y} \|ARXYSAT_2 - AT_2\|_p^p,$$

then

$$\|AR\tilde{X}\tilde{Y}SA - A\|_p^p \leq \alpha O(\log n) \min_{X,Y} \|ARXYSA - A\|_p^p.$$

Lemma O.5. *Let $A \in \mathbb{R}^{n\times d}, S \in \mathbb{R}^{t_2\times n}, R \in \mathbb{R}^{d\times t_2}, k, T_1 \in \mathbb{R}^{t_2\times n}, T_2 \in \mathbb{R}^{d\times t_3}$ be the same as in the Algorithm 3, then with probability at least 0.9, if for $\alpha \geq 1$*

$$\sum_{i=1}^{t_3} \|T_1(AR\tilde{X}\tilde{Y}SAT_2 - AT_2)_i\|_2^p \leq \alpha \min_{X,Y} \sum_{i=1}^{t_3} \|T_1(ARXYSAT_2 - AT_2)_i\|_2^p,$$

then

$$\|AR\tilde{X}\tilde{Y}SAT_2 - AT_2\|_p^p \leq \alpha\beta \min_{X,Y} \|ARXYSAT_2 - AT_2\|_p^p,$$

where $\beta = O(\min(k \log k \log^p n, (k \log k)^{1-p/2} \log^{1+p/2} n))$.

Proof. Let $X^*, Y^* = \arg\min_{X,Y} \sum_{i=1}^{t_3} \|T_1(ARXYSAT_2 - AT_2)_i\|_2^p$. Let $L = AR, N = SAT_2, M = AT_2$. Let $\gamma = \min\{k \log k, (k \log k \log n)^{1-p/2}\}$. Let $\tilde{H} = \tilde{X}\tilde{Y}$ and let $H^* = X^*Y^*$. We have

$$\begin{aligned}
&\|L\tilde{H}N - M\|_p^p \\
\leq& 2^{p-1}\|L\tilde{H}N - LH^*N\|_p^p + 2^{p-1}\|LH^*N - M\|_p^p \\
\leq& O(\gamma) \sum_{i=1}^{t_3} \|T_1(L\tilde{H}N - LH^*N)_i\|_2^p + 2^{p-1}\|LH^*N - M\|_p^p \\
\leq& O(\gamma) \sum_{i=1}^{t_3} (\|T_1(L\tilde{H}N - M)_i\|_2 + \|T_1(LH^*N - M)_i\|_2)^p + 2^{p-1}\|LH^*N - M\|_p^p \\
\leq& O(\gamma) \sum_{i=1}^{t_3} (\|T_1(L\tilde{H}N - M)_i\|_2^p + \|D_2^{-1}(LH^*N - M)_i\|_2^p) + 2^{p-1}\|LH^*N - M\|_p^p \\
\leq& O(\gamma) (\sum_{i=1}^{t_3} \|T_1(L\tilde{H}N - M)_i\|_2^p + \|D_2^{-1}(LH^*N - M)\|_p^p) + 2^{p-1}\|LH^*N - M\|_p^p \\
\leq& O(\gamma) (\alpha \sum_{i=1}^{t_3} \|T_1(LH^*N - M)_i\|_2^p + \|D_2^{-1}(LH^*N - M)\|_p^p) + 2^{p-1}\|LH^*N - M\|_p^p \\
\leq& O(\gamma) (\alpha \sum_{i=1}^{t_3} \|D_2^{-1}(LH^*N - M)_i\|_2^p + \|D_3^{-1}(LH^*N - M)\|_p^p) + 2^{p-1}\|LH^*N - M\|_p^p \\
\leq& O(\gamma)\alpha \|D_2^{-1}(LH^*N - M)\|_p^p + 2^{p-1}\|LH^*N - M\|_p^p \\
\leq& O(\gamma \log^p(n))\alpha \|LH^*N - M\|_p^p.
\end{aligned}$$

The first inequality follows by convexity of x^p. The second inequality follows by Lemma O.2. The third inequality follows by triangle inequality. The forth inequality follows by convexity of x^p, Theorem 14 and Theorem 15. The fifth inequality follows by $p \leq 2$. The sixth inequality follows by the property of $\tilde{X}, \tilde{Y}$. The seventh inequality follows by Theorem 14 and Theorem 15. The eighth inequality follows by $p \leq 2$. Then the ninth inequality follows by Theorem 9. □

Now let us prove Theorem:

Proof. Notice that $\hat{X}, \hat{Y} = \arg\min_{X \in \mathbb{R}^{t_2 \times k}, Y \in \mathbb{R}^{k \times t_3}} \|T_1 ARXYSAT_2 - T_1 AT_2\|_F^2$, we have

$$(\sum_{i=1}^{t_3} \|T_1(AR\hat{X}\hat{Y}SAT_2 - AT_2)_i\|_2^p)^{1/p} \leq O((k\log k)^{1/p-1/2})(\min_{X,Y} \sum_{i=1}^{t_3} \|T_1(ARXYSAT_2 - AT_2)_i\|_2^p)^{1/p}.$$

It means

$$(\sum_{i=1}^{t_3} \|T_1(AR\hat{X}\hat{Y}SAT_2 - AT_2)_i\|_2^p) \leq O((k\log k)^{1-p/2})(\min_{X,Y} \sum_{i=1}^{t_3} \|T_1(ARXYSAT_2 - AT_2)_i\|_2^p).$$

According to Lemma O.5, we have

$$\|AR\hat{X}\hat{Y}SAT_2 - AT_2\|_p^p \leq \beta_1 \min_{X,Y} \|ARXYSAT_2 - AT_2\|_p^p,$$

where $\beta_1 = O(\min((k\log k)^{2-p/2}\log^p n, (k\log k)^{2-p}\log^{1+p/2} n))$. Due to Lemma O.4, we have

$$\|AR\hat{X}\hat{Y}SA - A\|_p^p \leq O(\beta_1 \log n) \min_{X,Y} \|ARXYSA - A\|_p^p.$$

Then, according to Lemma O.3, we have

$$\|AR\hat{X}\hat{Y}SA - A\|_p^p \leq \beta_2 \min_{U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{k\times d}} \|UV - A\|_p^p,$$

where $\beta_2 = O(\min((k\log k)^{4-p}\log^{2p+2} n, (k\log k)^{4-2p}\log^{4+p} n))$. For the running time: $SA, T_1 A$ can be computed in $\mathrm{nnz}(A)$ time. Thus, total running time is $\mathrm{nnz}(A) + (n+d)\mathrm{poly}(k)$. □

P Implementation Setups

We implement all the algorithms in MATLAB. We ran experiments on a machine with 16G main memory and Intel Core i7-3720QM@2.60GHz CPU. The operating system is Ubuntu 14.04.5 LTS. All the experiments were in single threaded mode.

Q Data Simulation for Comparison with ℓ_1 and ℓ_2 Regression

We generate a matrix $A \in \mathbb{R}^{n\times d}, x^* \in \mathbb{R}^d$ as following: set each entry of the first $d+5$ rows of A as i.i.d. standard random Gaussian variable, each entry of x^* as i.i.d. standard random Gaussian variable. For $n \geq i \geq d+6$, we uniformly choose $p \in [d+5]$, and set $A^i = A^p, b_i = b_p$. We perform experiments under 3 different noise assumptions and 2 dimension combinations of N, d and in total $3 \times 2 = 6$ experiments. The 3 different noise assumptions are, respectively i) $N(0, 50)$ Gaussian noise with on all the entries of Ax^*; ii) sparse noise, where we randomly pick 3% number of entries of Ax^*, and add uniform random noise from $[-\|Ax^*\|_2, \|Ax^*\|_2]$ on each entry to get b; iii) mixed noise, which is $N(0, 5)$ Gaussian noise plus sparse noise. The 2 different dimension combinations are i) balance, where $n = 100 \approx d = 75$; ii) overconstraint, where $n = 200 \gg d = 10$.

R Experiments on Approximation Ratio

Here is a documentation of our preliminary experiments on calculating the actual approximation ratio for the experiment settings mentioned in **Section 5.1, Comparison with ℓ_1 and ℓ_2 regression**. The approximation ratio of interest is calculated as follows: $\frac{\|Ax'-b\|_G}{\|Ax^*-b\|_G}$, where x' is the output of our novel embedding based algorithm and x^* is the optimal solution. Since $\|\cdot\|_G$ is convex, we can formulate this problem as a convex optimization problem and use a vanilla gradient descent algorithm to calculate the optimal solution. We heuristically stop our gradient descent algorithm when the one step brings less than 10^{-7} improvement on the loss function and set the learning rate to be 0.001. Admittedly, we have not yet thoroughly and rigidly examined the convergence of the vanilla gradient descent algorithm (a direction of future work), and hence such calculation of approximation ratio is only a preliminary attempt.

Under the mixed noise setting, we varied different scale s of the uniform noise to be $0, 1, 2, 3$ and delta to be $0.1, 0.25, 0.5, 0.75$. With $n = 200, d = 10$, for each of these $4 * 4 = 16$ settings, we run the algorithm repeatedly for 50 times, and the worst approximation ratio is 1.06 among these 800 runs. Experimentally, it is far below the theoretical guarantee $d\log^2(n) \approx 584 \gg 1.06$, and the approximation ratio is robust among different noise settings. For $n = 100, d = 75$, due to time limit, we only run each of the 16 settings for 5 times, and the worst approximation ratio is 1.31.

S Implementation Detail for Low Rank Approximation

- For our algorithm, set $t_1 = 4k, t_2 = 8t_1$, set $S \in \mathbb{R}^{t_1\times n}, T_1 \in \mathbb{R}^{t_2\times n}$ to be two random cauchy matrices, and set $R \in \mathbb{R}^{d\times t_1}, T_2 \in \mathbb{R}^{d\times t_2}$ to be two embedding matrices with exponential random variables (see Theorem 16.) We solve the minimization problem $\min_{X,Y} \|T_1ARXYSAT_2 - T_1AT_2\|_F^2$, and set $B = ARXYSA$.
- For algorithm in [19], we set $t_1 = 4k, t_2 = 8t_1$. We set $S \in \mathbb{R}^{t_1\times n}, T_1 \in \mathbb{R}^{t_2\times n}, R \in \mathbb{R}^{d\times t_1}, T_2 \in \mathbb{R}^{d\times t_2}$ to be four random cauchy matrices. We solve the minimization problem $\min_{X,Y} \|T_1ARXYSAT_2 - T_1AT_2\|_F^2$, and set $B = ARXYSA$.
- For PCA, we project A onto the space spanned by top k singular vectors to get B.

References

[1] Olvi L Mangasarian and David R. Musicant. Robust linear and support vector regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):950–955, 2000.

[2] Art B Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

[3] Jacopo Cavazza and Vittorio Murino. People counting by huber loss regression.

[4] Leo N Geppert, Katja Ickstadt, Alexander Munteanu, Jens Quedenfeld, and Christian Sohler. Random projections for bayesian regression. *Statistics and Computing*, pages 1–23, 2015.

[5] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 721–729. Curran Associates, Inc., 2015.

[6] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[7] Jacopo Cavazza and Vittorio Murino. Active regression with adaptive huber loss. *arXiv preprint arXiv:1606.01568*, 2016.

[8] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[9] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2013.

[10] Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 117–126. IEEE, 2013.

[11] Christian Sohler and David P Woodruff. Subspace embeddings for the ℓ_1-norm with applications. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 755–764. ACM, 2011.

[12] Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 91–100. ACM, 2013.

[13] David Woodruff and Qin Zhang. Subspace embeddings and ℓ_p regression using exponential random variables. In *Conference on Learning Theory*, pages 546–567, 2013.

[14] Ruosong Wang and David P Woodruff. Tight bounds for ℓ_p oblivious subspace embeddings. *arXiv preprint arXiv:1801.04414*, 2018.

[15] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

[16] Petros Drineas, Michael W Mahoney, S Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

[17] Kenneth L Clarkson and David P Woodruff. Sketching for m-estimators: A unified approach to robust regression. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 921–939. Society for Industrial and Applied Mathematics, 2015.

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

[19] Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise ℓ_1-norm error. In *Proceedings of the 49th Annual Symposium on the Theory of Computing*. ACM, arXiv preprint arXiv:1611.00898, 2017.

[20] Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, and David P Woodruff. Algorithms for ℓ_p low rank approximation. *arXiv preprint arXiv:1705.06730*, 2017.

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