Subspace Embedding and Linear Regression with Orlicz Norm

Alexandr Andoni [1] **Chengyu Lin** [1] **Ying Sheng** [1] **Peilin Zhong** [1] **Ruiqi Zhong** [1]

Abstract

We consider a generalization of the classic linear regression problem to the case when the loss is an Orlicz norm. An Orlicz norm is parameterized by a non-negative convex function $G : \mathbb{R}_+ \to \mathbb{R}_+$ with $G(0) = 0$: the Orlicz norm of a vector $x \in \mathbb{R}^n$ is defined as $\|x\|_G = \inf \{\alpha > 0 \mid \sum_{i=1}^n G(|x_i|/\alpha) \leq 1\}$. We consider the cases where the function $G(\cdot)$ grows subquadratically. Our main result is based on a new oblivious embedding which embeds the column space of a given matrix $A \in \mathbb{R}^{n\times d}$ with Orlicz norm into a lower dimensional space with ℓ_2 norm. Specifically, we show how to efficiently find an embedding matrix $S \in \mathbb{R}^{m\times n}, m < n$ such that $\forall x \in \mathbb{R}^d, \Omega(1/(d \log n)) \cdot \|Ax\|_G \leq \|SAx\|_2 \leq O(d^2 \log n) \cdot \|Ax\|_G$. By applying this subspace embedding technique, we show an approximation algorithm for the regression problem $\min_{x\in\mathbb{R}^d} \|Ax-b\|_G$, up to a $O(d \log^2 n)$ factor. As a further application of our techniques, we show how to also use them to improve on the algorithm for the ℓ_p low rank matrix approximation problem for $1 \leq p < 2$.

1. Introduction

Numerical linear algebra problems play a significant role in machine learning, data mining, and statistics. One of the most important such problems is the regression problem, see some recent advancements in, e.g., (Zhong et al., 2016; Bhatia et al., 2015; Jain & Tewari, 2015; Liu et al., 2014; Dhillon et al., 2013). In a linear regression problem, given a data matrix $A \in \mathbb{R}^{n\times d}$ with n data points $A^1, A^2, \cdots, A^n$ in $\mathbb{R}^d$ and the response vector $b \in \mathbb{R}^n$, the goal is to find a set of coefficients $x^* \in \mathbb{R}^d$ such that

$$x^* = \arg\min_{x\in\mathbb{R}^d} l(Ax - b), \tag{1}$$

where $l : \mathbb{R}^n \to \mathbb{R}_+$ is the loss function. When $l(y) = \|y\|_2^2 = \sum_{i=1}^n y_i^2$, then the problem is the classic least square regression problem (ℓ_2 regression). While there has been extensive research on efficient algorithms for solving ℓ_2 regression, it is not always a suitable loss function to use.

In many settings, alternative choices for a loss function lead to qualitatively better solutions x^*. For example, a popular such alternative is the least absolute deviation (ℓ_1) regression — with $l(y) = \|y\|_1 = \sum_{i=1}^n |y_i|$ — which leads to solutions that are more robust than those of ℓ_2 regression (see (Wikipedia; Gorard, 2005). In a nutshell, the ℓ_2 regression is suitable when the data contains Gaussian noise, whereas ℓ_1 — when the noise is Laplacian or sparse.

A further popular class of loss functions $l(\cdot)$ arises from *M-estimators*, defined as $l(y) = \sum_{i=1}^n M(y_i)$ where $M(\cdot)$ is an M-estimator function (see (Zhang, 1997) for a list of M-estimators). The benefit of (some) M-estimators is that they enjoy advantages of both ℓ_1 and ℓ_2 regression. For example, when $M(\cdot)$ is the Huber function (Huber et al., 1964), then the regression looks like ℓ_2 regression when y_i is small, and looks like ℓ_1 regression otherwise. However, these loss functions come with a downside: they depend on the scale, and rescaling the data may give a completely different solution!

Our contributions. We introduce a generic algorithmic technique for solving regression for an entire class of loss functions that includes the aforementioned examples, and in particular, a "scale-invariant" version of M-estimators. Specifically, our class consists of loss functions $l(y)$ that are Orlicz norms, defined as follows: given a non-negative convex function $G : \mathbb{R}_+ \to \mathbb{R}_+$ with $G(0) = 0$, for $x \in \mathbb{R}^n$, we can define $\|x\|_G$ to be an Orlicz norm with respect to $G(\cdot)$: $\|x\|_G \triangleq \inf \{\alpha > 0 \mid \sum_{i=1}^n G(|x_i|/\alpha) \leq 1\}$. Note that ℓ_p norm, for $p \in [1, \infty)$, is a special case of Orlicz norm with $G(x) = x^p$. Another important example is the following "scale-free" version of M-estimator. Taking $f(\cdot)$ to be a Huber function, i.e.

$$f(x) = \begin{cases} x^2/2 & |x| \leq \delta \\ \delta(|x| - \delta/2) & \text{otherwise} \end{cases}$$

for some constant δ, we take $G(x) = f(f^{-1}(1)x)$. Then the norm $\|x\|_G$ looks like ℓ_2 norm when x is flat, and looks like ℓ_1 norm when x is sparse. Figure 1 shows the unit norm ball of this kind of Orlicz norm.

[*]Equal contribution [1]Computer Science Department, Columbia University, New York City, NY 10027, U.S.A.. Correspondence to: Peilin Zhong <pz2225@columbia.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

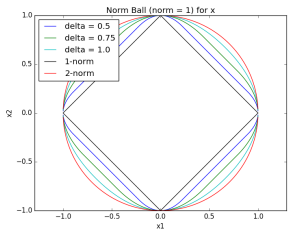

Figure 1. Unit norm balls of Orlicz norm induced by normalized Huber functions with different δ.

Our main result is a generic algorithm for solving any regression problem Eqn. (1) with any loss function that is a "nice" Orlicz norm; see Section 2 for a formal definition of "nice", and think of it as subquadratic for now.

Our main result employs the concept of *subspace embeddings*, which is a powerful tool for solving numerical linear algebra problems, and as such has many applications beyond regression. We say that a subspace embedding matrix $S \in \mathbb{R}^{m \times n}$ embeds the column space of $A \in \mathbb{R}^{n \times d}$ $(n > m)$ with u-norm into a subspace with v-norm, if $\forall x \in \mathbb{R}^d$, we have $\|Ax\|_u / \alpha \leq \|SAx\|_v \leq \beta \|Ax\|_u$ where $\alpha\beta$ is called distortion (approximation). A long line of work studied ℓ_2 regression problem based on ℓ_2 subspace embedding techniques; see, e.g., (Clarkson & Woodruff, 2009; 2013; Nelson & Nguyên, 2013). Furthermore, there are works on ℓ_p regression problem based on ℓ_p subspace embedding techniques (see, e.g. (Sohler & Woodruff, 2011; Meng & Mahoney, 2013; Clarkson et al., 2013; Woodruff & Zhang, 2013)), and similarly for M-estimators (Clarkson & Woodruff, 2015).

Our overall results are composed of four parts:

1. We develop the first subspace embedding method for all "nice" Orlicz norms. The embedding obtains a distortion factor polynomial in d, which was recently shown necessary (Wang & Woodruff, 2018).
2. Using the above subspace embedding, we obtain the first approximation algorithm for solving the linear regression problem with any "nice" Orlicz norm loss.
3. As a further illustration of the power of the subspace embedding method, we employ it towards improving on the best known result for another problem: ℓ_p low rank approximation for $1 \leq p < 2$ from (Song et al., 2017), which is the "ℓ_p-version of PCA".
4. Finally, we complement our theoretical results with experimental evaluation of our algorithms. Our experiments reveal that that the solution of regression under the Orlicz norm induced by Huber loss is much better than the solution given by regression under ℓ_1 or ℓ_2 norms, under natural noise distributions in practice. We also perform experiments for Orlicz regression with different Orlicz functions G and show their behavior under different noise settings, thus exhibiting the flexibility of our framework.

To the best of our knowledge, our algorithms are the first low distortion embedding and regression algorithms for general Orlicz norm. For the problem of low rank approximation under ℓ_p norm, $p \in [1, 2)$, our algorithms achieve simultaneously the best approximation and the best running time. In contrast, all the previous algorithms achieve either the best approximation, or the best running time, but not both at the same time.

Our algorithms for subspace embedding and regression are simple, and in particular are not iterative. In particular, for the subspace embedding, the embedding matrix S is generated independently of the data. In the regression problem, we multiply the input with the embedding matrix, and thus reduce to the ℓ_2 regression problem, for which we can use any of the known algorithm.

Technical discussion. Next we highlight some of our techniques used to obtain the theoretical results.

Subspace embedding. Our starting point is a technique introduced in (Andoni et al., 2017) for the Orlicz norms, which can be seen as an embedding that has guarantees for a *fixed vector* only. In contrast, our main challenge here is to obtain an embedding for *all vectors* $x \in \mathbb{R}^n$ in a certain d-dimensional subspace. Consider a random diagonal matrix $D \in \mathbb{R}^{n \times n}$ with each diagonal entry is a "generalized exponential" random variable, i.e., drawn from a distribution with cumulative distribution function $1 - e^{-G(x)}$. Then, for a fixed $x \in \mathbb{R}^d$, (Andoni et al., 2017) show that $\|D^{-1}Ax\|_\infty$ is not too small with high probability. We can combine this statement together with a net argument and the dilation bound on $\|D^{-1}Ax\|_G$, to argue that $\forall x \in \mathbb{R}^d$, $\|D^{-1}Ax\|_\infty$ is not too small.

The other direction is more challenging — to show that for a given matrix $A \in \mathbb{R}^{n \times d}$, and any fixed $x \in \mathbb{R}^d$, $\|D^{-1}Ax\|_G$ cannot be too large. Once we show this "dilation bound", we combine it with the well-conditioned basis argument (similar to (Dasgupta et al., 2009)), and prove that $\forall x \in \mathbb{R}^d$, $\|D^{-1}Ax\|_G$ cannot be too large. Overall, we have that $\forall x \in \mathbb{R}^d$, $\|D^{-1}Ax\|_G \leq O(d^2 \log n) \cdot \|Ax\|_G$, and $\|D^{-1}Ax\|_\infty \geq \Omega(1/(d \log n)) \cdot \|Ax\|_G$. Since ℓ_2 norm is sandwiched by $\|\cdot\|_G$ and ℓ_∞ norm, we have that $\forall x \in \mathbb{R}^d, \Omega(1/(d \log n)) \cdot \|Ax\|_G \leq \|D^{-1}Ax\|_2 \leq O(d^2 \log n) \cdot \|Ax\|_G$. Then, the remaining part is to use standard techniques (Woodruff & Zhang, 2013; Woodruff, 2014) to perform the ℓ_2 subspace embedding for the column space of $D^{-1}A$. See Theorem 16 for details.

The actual proof of the dilation bound is the most technically intricate result. Traditionally, since the p^{th} power of the ℓ_p norm is the sum of the p^{th} power of all the entries, it is easy to bound the expectation by using linearity of the expectation. However it is impossible to apply this analysis to Orlicz norm directly since Orlicz norm is not an "entry-wise" norm. Instead, we exploit a key observation that the

Orlicz norm of vectors which are on the unit ball can be seen as the sum of contribution of each coordinate. Thus, we propose a novel analysis for any fixed vector by analyzing the behavior of the normalized vector which is on the unit Orlicz norm ball. To extend the dilation bound for a fixed vector to all the vectors in a subspace, we generalize the definition of ℓ_p norm well-conditioned basis to the Orlicz norm case, and then show that the Auerbach basis provides a good basis for Orlicz norm. To the best of our knowledge, this is the first time Auerbach basis are used to analyze the dilation bound of an embedding method for a norm distinct from an ℓ_p norm. See Section 3 for details.

Regression with Orlicz norm. Here, given a matrix $A \in \mathbb{R}^{n\times d}$, a vector $b \in \mathbb{R}^n$, the goal is to solve Equation 1 with Orlicz norm loss function. We can now solve this problem directly using the subspace embedding from above, in particular by applyingit to the column space of $[A\ b]$. We obtain an $O(d^3 \log^2 n)$ approximation ratio, which we can further improve by observing that it actually suffices to have the dilation bound on $\|D^{-1}Ax^*\|_G$ only for the optimal solution x^* (as opposed to for an arbitrary x). Using this observation, we improve the approximation ratio to $O(d \log^2 n)$. See Theorem 18 for details. We evaluate the algorithm's performance and show that it performs well (even when n is not much larger than d). See Section 5.

ℓ_p low rank matrix approximation. The ℓ_p norm is a special case of the Orlicz norm $\|\cdot\|_G$, where $G(x) = x^p$. This connection allows us to consider the following problem: given $A \in \mathbb{R}^{n\times d}, n \geq d \geq k \geq 1$, find a rank-$k$ matrix $B \in \mathbb{R}^{n\times d}$ such that $\|A - B\|_p$ is minimized. Here we consider the case of $1 \leq p < 2$ and $k = \omega(\log n)$. The best known algorithm for this problem is from (Song et al., 2017), which uses the dense p-stable transform to achieves $k^2 \cdot \mathrm{poly}(\log n)$ approximation ratio. It has the downside that its runtime does not compare favorably to the golden standard of runtime linear in the sparsity of the input. To improve the runtime, one can apply the sparse p-stable transform and achieve input sparsity runtime, but that comes at the cost of an $\Omega(k^6)$ factor loss in the approximation ratio.

Using the above techniques, we develop an algorithm with best of both worlds: $k^2 \cdot \mathrm{poly}(\log n)$ approximation ratio and the input sparsity running time at the same time. In particular, the main inefficiency of the algorithm (Song et al., 2017) is the relaxation from ℓ_p norm to ℓ_2 norm, which incurs a further $\mathrm{poly}(k)$ approximation factor. In contrast, the embedding based on exponential random variables embeds ℓ_p norm to ℓ_2 norm directly, without further approximation loss. Our embedding also comes with its own pitfalls — as we now need to deal with mixed norms — thus requiring a new analysis of the overall algorithm. See Theorem 23 for details.

2. Notations and preliminaries

In this paper, we denote $\mathbb{R}_+$ to be the set of nonnegative reals. Define $[n] = \{1, 2, \cdots, n\}$. Given a matrix $A \in \mathbb{R}^{n\times d}$, $\forall i \in [n], j \in [d]$, A^i and A_j denotes the i^{th} row and the j^{th} column of A respectively. $\mathrm{nnz}(A)$ denotes the number of nonzero entries of A. The column space of $A \in \mathbb{R}^d$ is $\{y \mid \exists x \in \mathbb{R}^d, y = Ax\}$. $\forall p \neq 2$, $\|A\|_p \triangleq (\sum |A_{i,j}|^p)^{1/p}$, i.e. entrywise p-norm. $\|A\|_F$ defines the Frobenius norm of A, i.e. $(\sum A_{i,j}^2)^{1/2}$. $A^\dagger$ denotes the Moore-Penrose pseudoinverse of A. Given an invertible function $f(\cdot)$, let $f^{-1}(\cdot)$ be the inverse function of $f(\cdot)$. If $f(\cdot)$ is not invertible in $(-\infty, +\infty)$ but it is invertible in $[0, +\infty)$, then we denote $f^{-1}(\cdot)$ to be the inverse function of $f(\cdot)$ in domain $[0, +\infty)$. inf and sup denote the infimum and supremum respectively. $f'(x), f'_+(x), f'_-(x)$ denote the derivative, right derivative and left derivative of $f(x)$, respectively. Similarly, define $f''(x)$ for the second derivatives, and we define $f''_+(x) = \lim_{h\to 0^+}(f'(x+h) - f'_+(x))/h$. In the following, we give the definition of Orlicz norm.

Definition 1 (Orlicz norm) *For any nonzero monotone nondecreasing convex function $G : \mathbb{R}_+ \to \mathbb{R}_+$ with $G(0) = 0$. Define Orlicz norm $\|\cdot\|_G$ as: $\forall n \in \mathbb{Z}, n \geq 1, x \in \mathbb{R}^n, \|x\|_G = \inf\{\alpha > 0 \mid \sum_{i=1}^n G(|x_i|/\alpha) \leq 1\}$.*

For any function $G_1(\cdot)$ which is valid to define an Orlicz norm, we can always "simplify/normalize" the function to get another function G_2 such that computing $\|\cdot\|_{G_1}$ is equivalent to computing $\|\cdot\|_{G_2}$.

Fact 2 *Given a function $G_1 : \mathbb{R}_+ \to \mathbb{R}_+$ which can induce an Orlicz norm $\|\cdot\|_{G_1}$ (Definition 1), define function $G_2 : \mathbb{R}_+ \to \mathbb{R}_+$ as the following: $G_2(x) = \begin{cases} G_1(G_1^{-1}(1)x) & 0 \leq x \leq 1 \\ sx - (s-1) & x > 1 \end{cases}$ where $s = \sup\{(G_2(y) - G_2(x))/(y-x) \mid 0 \leq x \leq y \leq 1\}$. Then $\|\cdot\|_{G_2}$ is a valid Orlicz norm. Furthermore, $\forall n \in \mathbb{Z}, n \geq 1, x \in \mathbb{R}^n$, we have $\|x\|_{G_1} = \|x\|_{G_2}/G_1^{-1}(1)$.*

Thus, without loss of generality, in this paper we consider the Orlicz norm induced by function G which satisfies $G(1) = 1$, and $G(x)$ is a linear function for $x > 1$. In addition, we also require that $G(x)$ grows no faster than quadratically in x. Thus, we define the property $\mathcal{P}$ of a function $G : \mathbb{R} \to \mathbb{R}_+$ as the following: 1) G is a nonzero monotone nondecreasing convex function in $[0, \infty)$; 2) $G(0) = 0, G(1) = 1, \forall x \in \mathbb{R}, G(x) = G(-x)$; 3) $G(x)$ is a linear function for $x > 1$, i.e. $\exists s > 0, \forall x > 1, G(x) = sx + (1-s)$; 4) $\exists \delta_G > 0$ such that G is twice differentiable on interval $(0, \delta_G)$. Furthermore, $G'_+(0)$ and $G''_+(0)$ exist, and either $G'_+(0) > 0$ or $G''_+(0) > 0$; 5) $\exists C_G > 0, \forall 0 < x < y, G(y)/G(x) \leq C_G(y/x)^2$.

The condition 1 is required to define an Orlicz norm. The conditions 2,3 are required because we can always do the simplification/normalization (see Fact 2). The condition 4 is required for the smoothness of G. The condition 5 is due to the subquadratic growth condition. Subquadratic

Table 1. Some of M-estimators.

HUBER	$\begin{cases} x^2/2 & \lvert x\rvert \le c \\ c(\lvert x\rvert - c/2) & \lvert x\rvert > c \end{cases}$
$\ell_1 - \ell_2$	$2(\sqrt{1+x^2/2}-1)$
"FAIR"	$c^2\left(\lvert x\rvert/c - \log(1+\lvert x\rvert/c)\right)$

growth condition is necessary for sketching $\sum_{i=1}^n G(x_i)$ with sketch size sub-polynomial in the dimension n, as shown by (Braverman & Ostrovsky, 2010). For example, if $G(x) = x^p$ for some $p > 2$, then $\|\cdot\|_G$ is the same as $\|\cdot\|_p$. It is necessary to take $\Omega(n^{1-2/p})$ space to sketch ℓ_p norm in n-dimensional space. Condition 5 is also necessary for 2-concave property, (Kwapien & Schuett, 1985; Kwapie & Schtt, 1989) shows that $\|\cdot\|_G$ can be embedded into ℓ_1 space if and only if G is 2-concave. Although (Schtt, 1995) gives an explicit embedding to ℓ_1, it cannot be computed efficiently.

There are many potential choices of $G(\cdot)$ which satisfies property $\mathcal{P}$, the following are some examples: 1) $G(x) = |x|^p$ for some $1 \le p \le 2$. In this case $\|\cdot\|_G$ is exactly the ℓ_p norm $\|\cdot\|_p$; 2) $G(x)$ can be a normalized M-estimator function (see (Zhang, 1997)), i.e. define $f(x)$ to be one of the functions in Table 1. and let $G(x) = \begin{cases} f(f^{-1}(1)x) & |x| \le 1 \\ G'_-(1)|x| - (G'_-(1)-1) & |x| > 1 \end{cases}$.

The following presents some useful properties of function G with property $\mathcal{P}$. See Appendix for details of proofs of the following Lemmas.

Lemma 3 *Given a function $G(\cdot)$ with property $\mathcal{P}$, then $\forall 0 \le x \le 1, x^2/C_G \le G(x) \le x$.*

Lemma 4 *Given a function $G(\cdot)$ with property $\mathcal{P}$, then $\forall x \in \mathbb{R}^n, \|x\|_2/\sqrt{C_G} \le \|x\|_G \le \|x\|_1$.*

Lemma 5 *Given a function $G(\cdot)$ with property $\mathcal{P}$, then $\forall 0 < x < y$, we have $y/x \le G(y)/G(x)$.*

Lemma 6 *Given a function $G(\cdot)$ with property $\mathcal{P}$, there exist a constant $\alpha_G > 0$ which may depend on G, such that $\forall 0 \le a, b$, if $ab \le 1$, then $G(a)G(b) \le \alpha_G G(ab)$.*

3. Subspace embedding for Orlicz norm using exponential random variables

In this section, we develop the subspace embedding under the Orlicz norms which are induced by functions G with the property $\mathcal{P}$. We first show how to embed the subspace with $\|\cdot\|_G$ norm into a subspace with ℓ_2 norm, and then we use dimensionality reduction techniques for the ℓ_2 norm. Overall, we will prove Theorem 16 stated at the end of this section. Before discussing the details, we give formal definitions of subspace embedding.

Definition 7 (Subspace embedding for Orlicz norm) *Given a matrix $A \in \mathbb{R}^{n\times d}$, if $S \in \mathbb{R}^{m\times n}$ satisfies $\forall x \in \mathbb{R}^d, \|Ax\|_G/\alpha \le \|SAx\|_v \le \beta\|Ax\|_G$ where $\alpha, \beta \ge 1, \|\cdot\|_v$ is a norm (can still be $\|\cdot\|_G$), then we say S embeds the column space of A with Orlicz norm into the column space of SA with v-norm. The distortion is $\alpha\beta$.*

If the distortion and the v-norm are clear from the context, we just say S is a subspace embedding matrix for A.

Definition 8 (Subspace embedding for ℓ_2 norm) *Given a matrix $A \in \mathbb{R}^{n\times d}$, if $S \in \mathbb{R}^{m\times n}$ satisfies $\forall x \in \mathbb{R}^d, (1-\varepsilon)\|Ax\|_2^2 \le \|SAx\|_2^2 \le (1+\varepsilon)\|Ax\|_2^2$, then we say S is a subspace embedding of column space of A.*

There are many choices of ℓ_2 subspace embedding matrix A satisfying the above definition. Examples are: random sign JL matrix (Achlioptas, 2003; Clarkson & Woodruff, 2009), fast JL matrix (Ailon & Chazelle, 2009), and sparse embedding matrices (Clarkson & Woodruff, 2013; Meng & Mahoney, 2013; Nelson & Nguyên, 2013).

The main technical thrust is to embed $\|\cdot\|_G$ into ℓ_2 norm. As the embedding matrix, we use $S = \Pi D^{-1}$ where Π is one of the above ℓ_2 embedding matrices and D is a diagonal matrix of which diagonal entries are i.i.d. random variables draw from the distribution with CDF $1 - e^{-G(t)}$. Equivalently, each entry on the diagonal of D is $G^{-1}(u)$, where u is an i.i.d. sample from the standard exponential distribution, i.e. CDF is $1 - e^{-t}$. In Section 3.1, we will prove that $\forall x \in \mathbb{R}^d$, $\|D^{-1}Ax\|_G$will not be too large. In Section 3.2, we will show that $\forall x \in \mathbb{R}^d$, $\|D^{-1}Ax\|_\infty$ cannot be too small. Then due to Lemma 4, we know that $\|D^{-1}Ax\|_2$ is a good estimator to $\|Ax\|_G$. In Section 3.3, we show how to put all the ingredients together.

3.1. Dilation bound

We construct a randomized linear map $f : \mathbb{R}^n \to \mathbb{R}^n$: $(x_1, x_2, ..., x_n) \overset{f}{\mapsto} (x_1/u_1, x_2/u_2, ..., x_n/u_n)$ where each u_i is drawn from a distribution with CDF $1 - e^{-G(t)}$. Notice that for proving the dilation bound, we do not need to assume u_i are independent.

Theorem 9 *Given $x \in \mathbb{R}^n$, let $\|\cdot\|_G$ be an Orlicz norm induced by function $G(\cdot)$ which has property $\mathcal{P}$, and let $f(x) = (x_1/u_1, x_2/u_2, ..., x_n/u_n)$, where each u_i is drawn from a distribution with CDF $1 - e^{-G(t)}$. Then with probability at least $1 - \delta - O(1/n^{19})$, $\|f(x)\|_G \le O(\alpha_G \delta^{-1}\log(n))\|x\|_G$, where α_G is a constant may depend on function $G(\cdot)$.*

Proof sketch: By taking union bound, we have $\forall i \in [n], u_i \ge G^{-1}(1/n^{20})$ with high probability. Let $\alpha = \|x\|_G$. For $\gamma \ge 1$, we want to upper bound the probability that $\|f(x)\|_G \ge \gamma\alpha$. This is equivalent to upper bound the probability that $\|f(x)/(\gamma\alpha)\|_G \ge 1$. Notice that $\Pr(\|f(x)/(\gamma\alpha)\|_G \ge 1) = \Pr(\sum G(x_i/\alpha \cdot 1/(\gamma u_i)) \ge 1)$. By Markov inequality, it suffices to bound the expectation of $\sum G(x_i/\alpha \cdot 1/(\gamma u_i))$ conditioned on u_i are not too small. By lemma 6, $\sum G(x_i/\alpha \cdot 1/(\gamma u_i)) \le \alpha_G/\gamma \cdot \sum G(x_i/\alpha) \cdot 1/G(u_i)$. Because u_i is not too small, the conditional expectation of $1/G(u_i)$ is roughly $O(\log n)$. So the probability that $\|f(x)\|_G \ge \gamma\alpha$ is bounded by $O(\alpha_G \log n/\gamma)$, set $\gamma = O(\log n)\alpha_G/\delta$, we can complete the proof. See appendix for the details of the whole proof.

The final step is to use a well-conditioned basis; see details in appendix.We then obtain the following theorem.

Theorem 10 *Let $G(\cdot)$ be a function which has property $\mathcal{P}$. Given a matrix $A \in \mathbb{R}^{n\times m}$ with rank $d \leq n$, let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix of which each entry on the diagonal is drawn from a distribution with CDF $1 - e^{-G(t)}$. Then, with probability at least 0.99, $\forall x \in \mathbb{R}^m, \|D^{-1}Ax\|_G \leq O(\alpha_G d^2 \log n)\|Ax\|_G$, where $\alpha_G \geq 1$ is a constant which only depends on $G(\cdot)$.*

3.2. Contraction bound

As in Section 3.1, we construct a randomized linear map $f : \mathbb{R}^n \to \mathbb{R}^n$: $(x_1, x_2, ..., x_n) \overset{f}{\mapsto} (x_1/u_1, x_2/u_2, ..., x_n/u_n)$ where each u_i is an i.i.d. random variable drawn from a distribution with CDF $1-e^{-G(t)}$. Notice that the difference from proving the dilation bound is that we need u_i to be independent here. We use the following theorem:

Theorem 11 (Lemma 3.1 of (Andoni et al., 2017)) *Given $x \in \mathbb{R}^n$, let $\|\cdot\|_G$ be an Orlicz norm induced by function $G(\cdot)$ which has property $\mathcal{P}$, and let $f(x) = (x_1/u_1, x_2/u_2, ..., x_n/u_n)$, where each u_i is an i.i.d random variable drawn from a distribution with CDF $1 - e^{-G(t)}$. Then for $\alpha \geq 1$, with probability at least $1 - e^{-\alpha}$, $\|f(x)\|_\infty \geq \|x\|_G/\alpha$.*

By combining the result with the net argument (see appendix), and Theorems 11, 10, we get the following:

Theorem 12 *$G(\cdot)$ is a function with property $\mathcal{P}$. Given a matrix $A \in \mathbb{R}^{n\times m}$ with rank $d \leq n$, 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 $1 - e^{-G(t)}$. Then, with probability at least 0.98, $\forall x \in \mathbb{R}^m, \Omega(1/(\alpha'_G d \log n))\|Ax\|_G \leq \|D^{-1}Ax\|_\infty$, where $\alpha'_G \geq 1$ is a constant which only depends on $G(\cdot)$.*

Proof sketch: Set $\varepsilon = 1/\mathrm{poly}(nd)$, we can build an ε-net (see Appendix) N for the column space of A. By taking the union bound over all the net points, we have $\forall x \in N$, $\|D^{-1}x\|_\infty$ is not too small. Due to Theorem 10, we have $\forall x$ in the column space of A, $\|D^{-1}x\|_G$ is not too large. Now, for any unit vector y in the column space of A, we can find the closest point $x \in N$, and $\|x - y\|_2 \leq \varepsilon$. Since $\|D^{-1}y\|_\infty \geq \|D^{-1}x\|_\infty - \|D^{-1}(y-x)\|_\infty$, $\|D^{-1}x\|_\infty$ is not too small, and $\|D^{-1}(y-x)\|_\infty$ is not too large, we can get a lower bound for $\|D^{-1}y\|_\infty$. See appendix for details.

3.3. Putting it all together

We now combine Theorem 12, Theorem 10, and Lemma 4, to get the following theorem.

Theorem 13 *Let $G(\cdot)$ be a function which has property $\mathcal{P}$. Given a matrix $A \in \mathbb{R}^{n\times m}$ with rank $d \leq n$, 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 $1 - e^{-G(t)}$. Then, with probability at least 0.98, $\forall x \in \mathbb{R}^m, \Omega(1/(\alpha'_G d \log n))\|Ax\|_G \leq \|D^{-1}Ax\|_2 \leq O(\alpha''_G d^2 \log n)\|Ax\|_G$, where $\alpha''_G, \alpha'_G \geq 1$ are two constants which only depend on $G(\cdot)$.*

The above theorem successfully embeds $\|\cdot\|_G$ into ℓ_2 space. We now use ℓ_2 subspace embedding to reduce the dimension. The following two theorems provide efficient ℓ_2 subspace embeddings.

Theorem 14 ((Clarkson & Woodruff, 2013)) *Given matrix $A \in \mathbb{R}^{n\times m}$ with rank d. Let $t = \Theta(d^2/\varepsilon^2)$, $S = \Phi Y \in \mathbb{R}^{t\times n}$, where $Y \in \mathbb{R}^{n\times n}$ is a diagonal matrix with each diagonal entry independently uniformly chosen to be ± 1, $\Phi \in \mathbb{R}^{t\times n}$ is a binary matrix with $\Phi_{h(i),i} = 1, \forall i \in [n]$, and remaining entries 0. Here $h : [n] \to [t]$ is a random hashing function such that for each $i \in [n]$, $h(i)$ is uniformly distributed in $[t]$. Then with probability at least 0.99, $\forall x \in \mathbb{R}^m, (1-\varepsilon)\|Ax\|_2^2 \leq \|SAx\|_2^2 \leq (1+\varepsilon)\|Ax\|_2^2$. Furthermore, SA can be computed in $\mathrm{nnz}(A)$ time.*

Theorem 15 (See e.g. (Woodruff, 2014)) *Given matrix $A \in \mathbb{R}^{n\times m}$ with rank d. Let $t = \Theta(d/\varepsilon^2)$, $S \in \mathbb{R}^{t\times n}$ be a random matrix of i.i.d. standard Gaussian variables scaled by $1/\sqrt{t}$. Then with probability at least 0.99, $\forall x \in \mathbb{R}^m, (1-\varepsilon)\|Ax\|_2^2 \leq \|SAx\|_2^2 \leq (1+\varepsilon)\|Ax\|_2^2$.*

We conclude the full theorem for our subspace embedding:

Theorem 16 *Let $G(\cdot)$ be a function which has property $\mathcal{P}$. 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 $1 - e^{-G(t)}$. Let $\Pi_1 \in \mathbb{R}^{t_1\times n}$ be a sparse embedding matrix (see Theorem 14) and let $\Pi_2 \in \mathbb{R}^{t_2\times t_1}$ be a random Gaussian matrix (see Theorem 15) where $t_1 = \Omega(d^2), t_2 = \Omega(d)$. Then, with probability at least 0.9, $\forall x \in \mathbb{R}^d, \Omega(1/(\alpha'_G d \log n))\|Ax\|_G \leq \|\Pi_2\Pi_1 D^{-1}Ax\|_2 \leq O(\alpha''_G d^2 \log n)\|Ax\|_G$, where $\alpha''_G, \alpha'_G \geq 1$ are two constants which only depend on $G(\cdot)$. Furthermore, $\Pi_2\Pi_1 D^{-1}A$ can be computed in $\mathrm{nnz}(A) + \mathrm{poly}(d)$ time.*

4. Applications

In this section, we discuss regression problem with Orlicz norm error measure, and low rank approximation problem with ℓ_p norm, which is a special case of the Orlicz norms.

4.1. Linear regression under Orlicz norm

We first give the definition of regression problem with Orlicz norm.

Definition 17 *Function $G(\cdot)$ has property $\mathcal{P}$. Given $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$, the goal is to solve the following minimization problem $\min_{x\in\mathbb{R}^d} \|Ax - b\|_G$.*

Theorem 18 *Let $G(\cdot)$ have property $\mathcal{P}$. Given $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$, Algorithm 1 can output a solution $\hat{x} \in \mathbb{R}^d$ such that with probability at least 0.8, $\|A\hat{x} - b\|_G \leq O(\beta_G d \log^2 n) \min_{x\in\mathbb{R}^d} \|Ax - b\|_G$, where β_G is a constant which may depend on $G(\cdot)$. In addition, the running time of Algorithm 1 is $\mathrm{nnz}(A) + \mathrm{poly}(d)$.*

Algorithm 1 Linear regression with Orlicz norm $\|\cdot\|_G$

1: **Input:** $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$.
2: **Output:** $\hat{x}$.
3: Let $t_1 = \Theta(d^2), t_2 = \Theta(d)$.
4: Let $\Pi_1 \in \mathbb{R}^{t_1\times n}$ be a random sparse embedding matrix, $\Pi_2 \in \mathbb{R}^{t_2\times t_1}$ be a random gaussian matrix, and $D \in \mathbb{R}^{n\times n}$ be a random diagonal matrix with each diagonal entry independently drawn from distribution whose CDF is $1 - e^{-G(t)}$. (See Theorem 16.)
5: Compute $\hat{x} = (\Pi_2\Pi_1 D^{-1}A)^\dagger \Pi_2\Pi_1 D^{-1}b$.

Proof sketch: Let $S = \Pi_2\Pi_1 D^{-1}$ be the subspace embedding for column space of $[A\ b]$. Let $x^* = \arg\min_{x\in\mathbb{R}^d} \|Ax - b\|_G$. Due to Theorem 16, $\|A\hat{x} - b\|_G$ is bounded by $O(d\log n)\|S(A\hat{x} - b)\|_2 \leq O(d\log n)\|S(Ax^* - b)\|_2 \leq O(d\log n)\|D^{-1}(Ax^* - b)\|_2$. Due to Theorem 9, $\|D^{-1}(Ax^* - b)\|_2 \leq O(1)\|D^{-1}(Ax^* - b)\|_G \leq O(\log n)\|Ax^* - b\|_G$.

4.2. Regression with combined loss function

In this section, we want to point out that our technique can be used on solving regression problem with more general cost function. Recall that the goal is to solve the minimization problem $\min_{x\in\mathbb{R}^d} \|Ax - b\|_G$. Now, we consider there are multiple goals, and we want to minimize a linear combination of the costs. Now we give the definition of regression problem with combined cost function.

Definition 19 *Suppose function $G_1(\cdot), G_2(\cdot), ..., G_k(\cdot)$ satisfies property $\mathcal{P}$. Given $A_1 \in \mathbb{R}^{n_1\times d}, A_2 \in \mathbb{R}^{n_2\times d}, ..., A_k \in \mathbb{R}^{n_k\times d}, b_1 \in \mathbb{R}^{n_1}, b_2 \in \mathbb{R}^{n_2}, ..., b_k \in \mathbb{R}^{n_k}$, the goal is to solve the following minimization problem $\min_{x\in\mathbb{R}^d} \sum_{i=1}^k \|A_i x - b_i\|_{G_i}$.*

The idea of solving this problem is that we can embed every term into l_1 space, and then merge them into one term. By the standard technique, there is a way to embed l_2 space to l_1 space. We show the embedding as below. For the completeness, we put the proof of this lemma to the appendix.

Lemma 20 *Let $Q \in \mathbb{R}^{t\times n}$ be a random matrix with each entry drawn uniformly from i.i.d. $\mathcal{N}(0,1)$ Gaussian distribution. Let $B = (\sqrt{\pi/2}/t)\cdot Q$. If $t = \Omega(\epsilon^{-2}n\log(n\epsilon^{-1}))$, then with probability at least 0.98, $\forall x \in \mathbb{R}^n$, $\|Bx\|_1 \in ((1-\epsilon)\|x\|_2, (1+\epsilon)\|x\|_2)$.*

Theorem 21 *Let $k > 0$ be a constant, and $G_1(\cdot), G_2(\cdot), ..., G_k(\cdot)$ satisfy property $\mathcal{P}$. Given $A_1 \in \mathbb{R}^{n_1\times d}, A_2 \in \mathbb{R}^{n_2\times d}, ..., A_k \in \mathbb{R}^{n_k\times d}, b_1 \in \mathbb{R}^{n_1}, b_2 \in \mathbb{R}^{n_2}, ..., b_k \in \mathbb{R}^{n_k}$, Algorithm 2 can output a solution $\hat{x} \in \mathbb{R}^d$ such that with probability at least 0.7, $\sum_{i=1}^k \|A_i\hat{x} - b_i\|_{G_i} \leq O(\beta'_G d\log^2 n)\min_{x\in\mathbb{R}^d}\sum_{i=1}^k \|A_i x - b_i\|_{G_i}$, where β'_G is a constant which may depend on $G_1(\cdot), G_2(\cdot), ..., G_k(\cdot)$. In addition, the running time of Algorithm 2 is $\sum_{i=1}^k \mathrm{nnz}(A_i) + \mathrm{poly}(d)$.*

Algorithm 2 Linear regression with combined loss functions

1: **Input:** $A_1 \in \mathbb{R}^{n_1\times d}, A_2 \in \mathbb{R}^{n_2\times d}, ..., A_k \in \mathbb{R}^{n_k\times d}, b_1 \in \mathbb{R}^{n_1}, b_2 \in \mathbb{R}^{n_2}, ..., b_k \in \mathbb{R}^{n_k}$
2: **Output:** $\hat{x}$.
3: Let $t_1 = \Theta(d^2), t_2 = \Theta(d), t_3 = \Theta(t_2\log(t_2))$.
4: Let $\Pi_1^{(1)} \in \mathbb{R}^{t_1\times n_1}, \cdots \Pi_1^{(k)} \in \mathbb{R}^{t_1\times n_k}$ be k random sparse embedding matrices, $\Pi_2^{(1)}, \cdots, \Pi_2^{(k)} \in \mathbb{R}^{t_2\times t_1}$ be k random Gaussian matrices, and $D^{(1)} \in \mathbb{R}^{n_1\times n_1}, \cdots, D^{(k)} \in \mathbb{R}^{n_k\times n_k}$ be k random diagonal matrices where each diagonal entry of $D^{(i)}$ is independently drawn from distribution whose CDF is $1 - e^{-G_i(t)}$. (See Theorem 16.) Let $Q^{(1)}, \cdots, Q^{(k)} \in \mathbb{R}^{t_3\times t_2}$ be random matrices with each entry drawn uniformly from i.i.d. $\mathcal{N}(0,1)$ Gaussian distribution. $\forall i \in [k]$, let $B^{(i)} = (\sqrt{\pi/2}/t_3)\cdot Q^{(i)}$ (see Lemma 20.) Let $B \in \mathbb{R}^{kt_3\times kt_2}, \Pi_2 \in \mathbb{R}^{kt_2\times kt_1}, \Pi_1 \in \mathbb{R}^{kt_1\times\sum_{j=1}^k n_j}, D \in \mathbb{R}^{\sum_{j=1}^k n_j\times\sum_{j=1}^k n_j}$ be four block diagonal matrices such that $\forall i \in [k]$, the i^{th} block of B, Π_2, Π_1, D is $B^{(i)}, \Pi_2^{(i)}, \Pi_1^{(i)}, D^{(i)}$ respectively.
5: Let $A = [A_1^\top, A_2^\top, ..., A_k^\top]^\top, b = [b_1^\top, b_2^\top, ..., b_k^\top]^\top$ and $S = B\Pi_2\Pi_1 D^{-1}$.
6: Use classical method of solving l_1 regression to get $\hat{x} = \arg\min_{x\in\mathbb{R}^d} \|S(Ax - b)\|_1$.

Proof Sketch: Let $A = [A_1^\top, A_2^\top, ..., A_k^\top]^\top, b = [b_1^\top, b_2^\top, ..., b_k^\top]^\top$, and $S = B\Pi_2\Pi_1 D^{-1}$ be the subspace embedding for column space of $[A\ b]$. Let $S_i = B^{(i)}\Pi_2^{(i)}\Pi_1^{(i)}(D^{(i)})^{-1}$. Notice that $\forall x, \|S(Ax - b)\|_1 = \sum_{i=1}^k \|S_i(A_i x - b_i)\|_1$. Let $x^* = \arg\min_{x\in\mathbb{R}^d}\sum_{i=1}^k \|A_i x - b_i\|_{G_i}$. Due to Theorem 16 and Lemma 20, $\sum_{i=1}^k \|A_i\hat{x} - b_i\|_{G_i}$ is bounded by $O(d\log n)\sum_{i=1}^k \|S_i(A_i\hat{x} - b_i)\|_1 = O(d\log n)\|S(A\hat{x} - b)\|_1 \leq O(d\log n)\|S(Ax^* - b)\|_1 = \sum_{i=1}^k \|S_i(A_i x^* - b_i)\|_1$. Due to Theorem 16, $\sum_{i=1}^k \|S_i(A_i x^* - b_i)\|_1 \leq O(\log n)\sum_{i=1}^k \|A_i x^* - b_i\|_{G_i}$.

One application of the above Theorem is to solve the LASSO (Least Absolute Shrinkage Sector Operator) regression. In LASSO regression problem, the goal is to minimize $\|Ax - b\|_2^2 + \lambda\|x\|_1$, where λ is a parameter of regularizer. It is easy to show that it is equivalent to minimize $\|Ax - b\|_2 + \lambda'\|x\|_1$ for some other parameter λ'. When we look at $\|Ax - b\|_2 + \lambda'\|x\|_1$, we can set $A_1 = A, b_1 = b, A_2 = \lambda' I, b_2 = 0, G_1(\cdot) \equiv x^2, G_2(\cdot) \equiv x$, then we are able to apply Theorem 21 to give a good approximation. The merit of our algorithm is that it is very simple, and can be computed very fast.

4.3. ℓ_p norm low rank approximation using exponential random variables

We discuss a special case of Orlicz norm $\|\cdot\|_G$, ℓ_p norm, i.e. $G(x) \equiv x^p$ for $p \in [1,2]$. When rank parameter k is $\omega(\log n + \log d)$, by using exponential random variables, we can significantly improve the approximation ratio of input sparsity time algorithms shown by (Song et al., 2017). The

high level ideas combine the results of (Woodruff & Zhang, 2013; Song et al., 2017) and the dilation bound in Section 3. We define the problem in the following. See Appendix for the proof of Theorem 23.

Definition 22 *Let $p \in [1, 2]$. Given $A \in \mathbb{R}^{n\times d}, n \geq d, k \in \mathbb{Z}, 1 \leq k \leq \min(n, d)$, the goal is to solve the following minimization problem:* $\min_{U\in\mathbb{R}^{n\times k}, V\in\mathbb{R}^{k\times d}} \|UV - A\|_p^p$.

Algorithm 3 ℓ_p norm low rank approximation using exponential random variables.

1: **Input:** $A \in \mathbb{R}^{n\times d}, k \in \mathbb{Z}, \min(n, d) \geq k \geq 1$.
2: **Output:** $\hat{U} \in \mathbb{R}^{n\times k}, \hat{V} \in \mathbb{R}^{k\times d}$.
3: Let $t_1 = \Theta(k^2), t_2 = \Theta(k), t_3 = \Theta(k \log k)$.
4: Let $\Pi_1, S_1 \in \mathbb{R}^{t_1\times n}$ be two random sparse embedding matrices, $\Pi_2, S_2 \in \mathbb{R}^{t_2\times t_1}$ be two random gaussian matrices, and $D_1, D_2 \in \mathbb{R}^{n\times n}$ be two random diagonal matrices with each diagonal entry independently drawn from distribution whose CDF is $1 - e^{-t^p}$. (See Theorem 16.)
5: Let $T_2, R \in \mathbb{R}^{d\times t_3}$ be two random matrix, with i.i.d. entries drawn from standard p-stable distribution.
6: Let $S = S_2 S_1 D_1^{-1}, T_1 = \Pi_2 \Pi_1 D_2^{-1}$.
7: Solve $\hat{X}, \hat{Y} = \arg\min_{X\in\mathbb{R}^{t_2\times k}, Y\in\mathbb{R}^{k\times t_3}} \|T_1 ARXYSAT_2 - T_1 AT_2\|_F^2$.
8: $\hat{U} = AR\hat{X}, \hat{V} = \hat{Y}SA$.

Theorem 23 *Let $1 \leq p \leq 2$. Given $A \in \mathbb{R}^{n\times d}, n \geq d, k \in \mathbb{Z}, 1 \leq k \leq \min(n, d)$, with probability at least $2/3$, $\hat{U}, \hat{V}$ outputted by Algorithm 3 satisfies: $\|\hat{U}\hat{V} - A\|_p^p \leq \alpha \min_{U\in\mathbb{R}^{n\times k}, V\in\mathbb{R}^{k\times d}} \|UV - A\|_p^p$, where $\alpha = O(\min((k \log k)^{4-p} \log^{2p+2} n, (k \log k)^{4-2p} \log^{4+p} n))$. In addition, the running time of Algorithm 3 is $\mathrm{nnz}(A) + (n + d)\mathrm{poly}(k)$.*

5. Experiments

Implementation setups can be seen in appendix.

5.1. Orlicz Norm Linear Regression

In this section, we show that our algorithm i) has reasonable and predictable performance under different scenarios and ii) is flexible, general and easy to use. We perform 3 sets of experiments. The first is to compare its performance with the standard ℓ_1 and ℓ_2 regression under different noise assumptions and dimensions of the regression problem; the second is to compare the performance of Orlicz regression with different G under different noise assumptions; the third is to experiment with Orlicz function G that is different from standard ℓ_p and Huber function. We evaluate the performance of our Orlicz norm linear regression algorithm on simulated data.

Comparison with ℓ_1 and ℓ_2 regression We would like to see whether Orlicz norm linear regression leads to expected performance relative to ℓ_1 and ℓ_2 regression. We choose our Orlicz norm $\|\cdot\|_G$ to be induced by the normalized Huber function where the Huber function is defined as $f(x) = \begin{cases} x^2/2 & |x| \leq \delta \\ \delta \cdot (|x| - \delta/2) & \text{o.w.} \end{cases}$. We chose the parameter δ to be 0.75. Intuitively, it is between ℓ_1 and ℓ_2 norm (see Figure 1). In all the simulations, we generate matrix $A \in \mathbb{R}^{n\times d}$, ground truth $x^* \in \mathbb{R}^d$, and b to be Ax^* plus some particular noise. We evaluate the performance of each algorithm by the ℓ_2 distances between the output x and the ground truth x^*. In terms of algorithm details, since n, d are not too large in our simulation, we did not apply the ℓ_2 subspace embedding to reduce the dimension; we only use reciprocal exponential random diagonal embedding matrix to embed $\|\cdot\|_G$ to ℓ_2 norm (see Theorem 13)[1].

Table 2. Comparisons of different regressions in different noise and dimension settings; each entry is the error of ℓ_1, ℓ_2, Orlicz norm regression. As expected, ℓ_2/ℓ_1 regression lead to best performance under Sparse/Gaussian noise setting, and the performance of Orlicz norm regression lies in between.

	Gaussian	Sparse	Mixed
balance	211.2/**194.5**/197.3	**25.3**/30.7/30.0	37.9/37.8/**37.5**
overconstraint	25.3/**20.0**/24.9	**2e-9**/1.6/1.5	8.7/7.6/**7.5**

We experiment with two n, d combinations, i) $n = 200, d = 10$ ii) $n = 100, d = 75$, and 3 noise setting with i) Gaussian noise ii) sparse noise and iii) mixed noise (addition of i) and ii)), altogether $2 \times 3 = 6$ setting. The detail of data simulation can be seen in appendix. For each experiment we repeat 50 times and compute the mean. The results are shown in Table 2. Orlicz norm regression has better performance than ℓ_1 and ℓ_2 when the noise is mixed. When the noise is Gaussian or sparse, Orlicz norm regression works better than ℓ_1 and ℓ_2 respectively. We did not experiment with Huber loss regression, since if we rescale the data and make it small/large in absolute values, the Huber regression will degenerate into respectively ℓ_2/ℓ_1 regression (see Introduction). See appendix for results on approximation ratio.

Choice of δ for G as a normalized Huber function We compare the performance of Orlicz norm regression induced by G as normalized Huber loss function with different δ under different noise assumptions. We fix $n = 500, d = 30$ and generate A and x as in the first set of experiments (see appendix). The noise is a mixture of $N(0, 5)$ Gaussian noise and sparse noise on 1% entries with different scale of uniform noise from $[-s\|Ax^*\|_2, s\|Ax^*\|_2]$, where scale s is chosen from [0, 0.5, 1, 2]. Under each noise assumptions with different scale s, we compare the performance of Orlicz norm regression induced by G with δ from [0.05, 0.1, 0.2, 0.4, 1, 2]. We repeat each experiment 50 times and report the mean of the ℓ_2 distance between output x and the ground truth x^*. The result is shown in Figure 2. When the scale is 0/2, the noise is almost Gaussian/sparse and we expect ℓ_2/ℓ_1 norm and thus large/small δ to perform the best; anything scale lying in between these extremes will have an optimal δ in between. We observe the expected trend: as s increases, the performance is optimal with smaller δ.

[1] We use MATLAB's *linprog* to solve ℓ_1 regression.

Table 3. Orlicz regression with different choices of G, mean of the ℓ_2 distances between the output and the ground truth in 50 repetitions of experiments.

ℓ_1	$\ell_{1.5}$	ℓ_2	$G_{\delta=0.25}$	$G_{\delta=0.75}$	$G_{\ell_{1.5}}$
17.0	45.0	909.8	60.2	405.7	**14.7**

Beyond Huber function - A General Framework We explore a variant Orlicz function G and evaluate it under a particular setting; the evaluation criteria is the same as the first set of the experiments. The G is of the same form aforementioned, except that it now grows at the order of $x^{1.5}$ when x is small. We denote it by $G_{\ell_{1.5}}$, which is the normalization of function f, and f is defined as: $f(x) = \begin{cases} x^{1.5}/1.5 & x \le \delta \\ \delta^{0.5} \cdot (|x| - \delta/3) & \text{o.w.} \end{cases}$. We generate a 500×30 matrix A and the ground truth vector x^* in the same way as before, and then add $N(0, 5)$ Gaussian noises and 1 sparse outlier with scale $s = 100$. We find that the modified $G_{\ell_{1.5}}$ under this settings outperforms $\ell_1, \ell_2, \ell_{1.5}, G_{\delta=0.25}, G_{\delta=0.75}$ regression by a significant amount where $G_{\delta=0.25}, G_{\delta=0.75}$ are Orlicz norm induced by regular normalized Huber function with $\delta = 0.25, 0.75$ respectively. The results are shown in Table 3. This experiment demonstrates that our algorithm is i) flexible enough to combine the advantage of norm functions, ii) general for any function that satisfies the nice property, and iii) easy to experiment with different settings, as long as we can compute G and G^{-1}.

5.2. ℓ_1 low rank matrix approximation

In this section, we evaluate the performance of the ℓ_1 low rank matrix approximation algorithm. We mainly compare the ℓ_1 norm error of our algorithm with the error of (Song et al., 2017) and standard PCA. Inputs are a matrix $A \in \mathbb{R}^{n \times d}$ and a rank parameter k; the goal is to output a rank k matrix B such that $\|A - B\|_1$ is as small as possible. The details of implementations are in the appendix. For each input, we run the algorithm 50 times and pick the best solution.

Datasets. We first run experiment on synthetic data: we randomly choose two matrices $U \in \mathbb{R}^{2000 \times 5}, V \in \mathbb{R}^{5 \times 2000}$ with each entry drawn uniformly from $(0, 1)$ Then we randomly choose 100 entries of UV, and add random outliers uniformly drawn from $(-100, 100)$ on those entries, thus we can get a matrix $A \in \mathbb{R}^{2000 \times 2000}$. In our experiment, $\|A\|_1$ is about 5.0×10^6. Then, we run experiments on real datasets *diabetes* and *glass* in UCI repository(Bache & Lichman, 2013). The data matrix of *diabetes* has size 768×8, and the data matrix of *glass* has size 214×9. For each data matrix, we randomly add outliers on 1% number of entries.

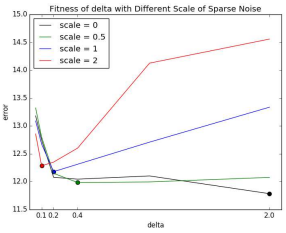

Figure 2. Performance of Orlicz regression with G induced by different δ under different scale of sparse noise. The larger the sparse noise, the smaller the δ that leads to the best performance, which makes the norm closer to ℓ_1

For each dataset, we evaluate the $\|A - B\|_1$. The result for the experiment on synthetic data is shown in Table 4, and the results for *diabetes* and *glass* are shown in Figure 3. The running time of algorithm in (Song et al., 2017) on *diabetes* and on *glass* are 5.69 and 11.97 seconds respectively, with ours being 3.18 and 3.74 seconds respectively. We also find that our algorithm consistently outperforms the other two alternatives (note that the y-coordinates are at log scale with base 10).

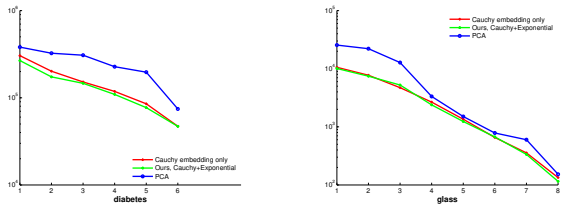

Figure 3. ℓ_1 norm error v.s. target rank.

Table 4. ℓ_1 rank-5 approximation on the synthetic data.

	Opt	PCA	(Song et al., 2017)	Ours
ℓ_1 loss ($\times 10^4$)	0.50	3.53	1.36	**1.04**

6. Conclusion and Future Work

In this paper we presented an efficient subspace embedding algorithm for orlicz norm and demonstrated its usefulness in regression/low rank approximation problem on synthetic and real datasets. Nevertheless, $O(d \log^2 n)$ is still a large theoretical approximation factor, and hence it is worth i) investigating whether the theoretical approximation ratio can be smaller if input are under some statistical distribution ii) calculating the actual approximation ratio with ground truth obtained by some slower but more accurate optimization algorithm. It is also worth examining whether our exponential embedding sketching method preserves the statistical properties of the regression error, since we assumed a different noise distribution from Gaussian/double-exponential as a starting point (Raskutti & Mahoney, 2014; Lopes et al., 2018).

Acknowledgements

Research supported in part by Simons Foundation (#491119 to Alexandr Andoni), NSF (CCF-1617955, CCF- 1740833), and Google Research Award.

References

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

Ailon, N. and Chazelle, B. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

Andoni, A., Nikolov, A., Razenshteyn, I., and Waingarten, E. Approximate near neighbors for general symmetric norms. In *Proceedings of the Forty-ninth annual ACM symposium on Theory of computing*, 2017.

Bache, K. and Lichman, M. Uci machine learning repository. *URL http://archive. ics. uci. edu/ml*, 901, 2013.

Bhatia, K., Jain, P., and Kar, P. Robust regression via hard thresholding. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 721–729. Curran Associates, Inc., 2015.

Braverman, V. and Ostrovsky, R. Zero-one frequency laws. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 281–290. ACM, 2010.

Clarkson, K. L. and Woodruff, D. P. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 205–214. ACM, 2009.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 81–90. ACM, 2013.

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

Clarkson, K. L., Drineas, P., Magdon-Ismail, M., Mahoney, M. W., Meng, X., and Woodruff, D. P. The fast cauchy transform and faster robust linear regression. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 466–477. Society for Industrial and Applied Mathematics, 2013.

Dasgupta, A., Drineas, P., Harb, B., Kumar, R., and Mahoney, M. W. Sampling algorithms and coresets for ℓ_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

Dhillon, P., Lu, Y., Foster, D. P., and Ungar, L. New subsampling algorithms for fast least squares regression. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 360–368. Curran Associates, Inc., 2013.

Gorard, S. Revisiting a 90-year-old debate: the advantages of the mean deviation. *British Journal of Educational Studies*, 53(4):417–430, 2005.

Huber, P. J. et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Jain, P. and Tewari, A. Alternating minimization for regression problems with vector-valued outputs. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1126–1134. Curran Associates, Inc., 2015.

Kwapien, S. and Schuett, C. Some combinatorial and probabilistic inequalities and their application to banach space theory. *Studia Mathematica*, 82:91–106, 1985.

Kwapie, S. and Schtt, C. Some combinatorial and probabilistic inequalities and their application to banach space theory ii. *Studia Mathematica*, 95(2):141–154, 1989.

Liu, H., Wang, L., and Zhao, T. Multivariate regression with calibration. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 127–135. Curran Associates, Inc., 2014.

Lopes, M. E., Wang, S., and Mahoney, M. W. Error estimation for randomized least-squares algorithms via the bootstrap. *arXiv preprint arXiv:1803.08021*, 2018.

Meng, X. and Mahoney, M. W. 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*, pp. 91–100. ACM, 2013.

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

Raskutti, G. and Mahoney, M. A statistical perspective on randomized sketching for ordinary least-squares. *arXiv preprint arXiv:1406.5986*, 2014.

Schtt, C. On the embedding of 2-concave orlicz spaces into l1. *Studia Mathematica*, 113(1):73–80, 1995.

Sohler, C. and Woodruff, D. P. Subspace embeddings for the ℓ_1-norm with applications. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 755–764. ACM, 2011.

Song, Z., Woodruff, D. P., and Zhong, P. 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.

Wang, R. and Woodruff, D. P. Tight bounds for ℓ_p oblivious subspace embeddings. *arXiv preprint arXiv:1801.04414*, 2018.

Wikipedia. https://en.wikipedia.org/wiki/Least_absolute_deviations.

Woodruff, D. and Zhang, Q. Subspace embeddings and ℓ_p regression using exponential random variables. In *Conference on Learning Theory*, pp. 546–567, 2013.

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

Zhang, Z. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and vision Computing*, 15(1):59–76, 1997.

Zhong, K., Jain, P., and Dhillon, I. S. Mixed linear regression with multiple components. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2190–2198. Curran Associates, Inc., 2016.