Submitted to the Annals of Statistics
arXiv: 1206.6536

NEARLY OPTIMAL MINIMAX ESTIMATOR FOR HIGH DIMENSIONAL SPARSE LINEAR REGRESSION

By Li Zhang

Microsoft Research Silicon Valley

We present estimators for a well studied statistical estimation problem: the estimation for the linear regression model with soft sparsity constraints (ℓ_q constraint with $0 < q \leq 1$) in the high-dimensional setting. We first present a family of estimators, called *the projected nearest neighbor estimator* and show, by using results from Convex Geometry, that such estimator is within a logarithmic factor of the optimal for any design matrix. Then by utilizing a semi-definite programming technique developed in [41], we obtain an approximation algorithm for computing the minimax risk for any such estimation task and also a polynomial time nearly optimal estimator for the important case of ℓ_1 sparsity constraint. Such results were only known before for special cases, despite decades of studies on this problem. We also extend the method to the adaptive case when the parameter radius is unknown.

1. Introduction. In the classical estimation problem with linear regression model, one observes a noisy $\widetilde{y}$ of some $y \in \mathbb{R}^n$ where $y = X\theta$ for a given $n \times p$ matrix X (called the *design matrix*) and an unknown $\theta \in \mathbb{R}^p$ and wishes to estimate y or θ. Recently, there have been enormous interests in the high-dimensional setting which in addition assumes that the design matrix is high-dimensional, i.e. when $p \gg n$, and θ satisfies certain sparsity constraints. Such sparsity constraints can be "hard", when it bounds the number of non-zero components in θ, or "soft", when θ is assumed to belong to the unit ℓ_q ball for $0 < q \leq 1$. In the existing study, the focus has so far been on the condition needed for X such that certain (typically polynomial time) estimators are nearly optimal or achieve lowest possible error for the given parameters. The work along this line has been quite successful [19, 1, 2, 7, 16, 13, 17, 14, 5, 11, 4, 43, 44, 34, 18] and produced many characterization of X (typically Gaussian random matrix) for which a polynomial time nearly optimal estimator exists.

The main departure point of this study is that we consider the problem of designing nearly optimal estimator for *any* given design matrix X,

AMS 2000 subject classifications: Primary 62J05; secondary 62G20, 62C20

Keywords and phrases: minimax estimation, linear regression, sparsity constraint, compressed sensing, optimal minimax estimator, projected nearest neighbor estimator, orthogonal projection estimator, nearest neighbor estimator.

i.e. we make no assumption about X. As the main contribution of this paper, we present a family of estimators, which we call *the projected nearest neighbor estimator*(PNN), and show that for any design matrix X, there is a projected nearest neighbor estimator that is nearly optimal in terms of the prediction risk for the corresponding linear regression problem over soft sparsity constraints. As a consequence, we obtain a polynomial time algorithm to compute the approximate minimax risk for any such problem and a polynomial time estimator in the important case of $q = 1$. Our results represent the first provably nearly optimal estimators without any constraint on the design matrix for $0 < q \leq 1$. We also design an adaptive estimator for the case when the ℓ_1 radius is not given.

We believe that studying optimal estimator for arbitrary X is important for multiple reasons. First, in practice we often do not have control over the design matrix or even the distribution of the design matrix. The design matrix might be “ill”-conditioned such that no estimator can achieve good accuracy. On the other hand, the design matrix may have a structure, as is often the case in practice, rather than completely random. In this case, it is important to take advantage of such structure to obtain better accuracy. Secondly, while there have been many characterization (typically some isometry property on X) known for certain algorithms to work well, it is often difficult to tell if the required property holds for a given X. So most results assume that X come from Gaussian random matrix. Thirdly, relaxing the requirement about the design X calls for the development of new algorithms as well as new analysis tools. Indeed, to argue the optimality of our estimator, we have to utilize novel tools from Convex Geometry (the classical restricted invertibility result by Bourgain and Tzafriri [8]).

1.1. *Problem setup.* In the linear regression problem, one observes $\widetilde{y} = y + g \in \mathbb{R}^n$, where $y = X\theta$ for a given $n \times p$ matrix X and an unknown vector $\theta \in \ell_q(C)$ for $0 < q \leq 1$, where $\ell_q(C) = \{(\theta_1, \ldots, \theta_p) :\ (\sum_i |\theta_i|^q)^{1/q} \leq C\}$. In addition, the noise g is a random vector drawn from the multivariate Gaussian distribution with the covariance matrix $\sigma^2 I$. In this paper, we only consider the prediction estimation, i.e. on the estimation of y but not θ. We use the standard total squared loss[1] to measure the error of an estimation, i.e.

$$\mathrm{loss}(\widehat{y}, y) = \|\widehat{y} - y\|^2 = \sum_i (\widehat{y}_i - y_i)^2\,.$$

For an estimator $M : \mathbb{R}^n \to \mathbb{R}^n$, we define the expected error of M on an

[1] We use the total squared error instead of the common mean squared error purely for the brevity of notation.

input y and on Gaussian error as

$$\mathrm{err}_M(y,\sigma) = \mathbb{E}_{\widetilde{y}=y+g;g\sim\mathcal{G}(\sigma)} \operatorname{loss}(M(\widetilde{y}), y) = \mathbb{E}_{\widetilde{y}=y+g;g\sim\mathcal{G}(\sigma)} \left\| M(\widetilde{y}) - y \right\|^2 .$$

Following [20], for $K \subseteq \mathbb{R}^n$, the risk of M over K is defined as

$$R_M(K,\sigma) = \sup_{y\in K} \mathrm{err}_M(y,\sigma)\,. \tag{1}$$

Define the minimax risk, denote by $R^*(K,\sigma)$, as the minimum achieveable risk among all the possible estimators, i.e.

$$R^*(K,\sigma) = \inf_M R_M(K,\sigma)\,. \tag{2}$$

For the aforementioned linear model with sparsity constraint $\ell_q(C)$, we have $K = X\ell_q(C)$ for an $n \times p$ design matrix X. Clearly, the minimax risk R^* ranges between 0 and $n\sigma^2$ and depends on the structure of X. The main goal of this paper is to design an estimator M such that $R_M(X\ell_q(C),\sigma)$ is close to $R^*(X\ell_q(C),\sigma)$ for any given X. For our main results, we consider the case where the sparsity radius C is given. Since we will only consider the prediction risk, we can assume, by rescaling X, that $C = 1$. In what follows, we write ℓ_q for $\ell_q(1)$. In addition, we only consider the high dimensional case where $p \geq n$ because for $p < n$, we can apply a rotation to the design matrix so that the last $n - p$ rows are entirely 0. Since Gaussian noise is invariant under rotation, this does not affect the minimax risk, and the dimensions of the design matrix is effectively reduced to $p \times p$.

1.2. *Main contribution.* We present a family of estimators, called the *projected nearest neighbor estimator* (PNN), that can achieve nearly optimal risk for *any* design matrix X and any given $0 < q \leq 1$. The projected nearest neighbor estimator is a combination of two classic estimators: the *orthogonal projection estimator*, in which the estimation is obtained by projecting the observation $\widetilde{y}$ to a properly chosen subspace, and the *nearest neighbor estimator*, in which $\widetilde{y}$ is mapped to the closest point (in terms of ℓ_2 distance) on the ground truth set K. The projected nearest neighbor estimator is defined with respect to an orthogonal projection P. It is the summation of two components: one, similar to the orthogonal projection estimator, is the projection $P\widetilde{y}$ of $\widetilde{y}$ by P; the other, similar to the nearest neighbor estimator, is the nearest neighbor projection of $P^\perp\widetilde{y}$ on $P^\perp K$, where $P^\perp$ is the projection orthogonal to P. As the main contribution of this work, we show that for any X, $0 < q \leq 1$, and $\sigma > 0$, there always exists a projection P so that the corresponding projected nearest neighbor estimator for $K = X\ell_q$ is nearly minimax optimal. More precisely, we show that [2]

Theorem 1. *For any given $n \times p$ matrix X, $0 < q \leq 1$, and $\sigma \geq 0$, there exists a projected nearest neighbor estimator M such that*

$$R_M(X\ell_q, \sigma) = O(c_q(\log^{1-q/2} p) R^*(X\ell_q, \sigma))\,,$$

where $c_q = O(2^{\frac{1}{q}} \frac{1}{q} \ln \frac{2}{q})$ is a constant dependent on q only.

In the above theorem, the projection P is chosen in two steps: 1. for each $0 \leq k \leq n$, a k dimensional projection P_k is chosen to minimize $\max_i \|P^{\perp} x_i\|$ where x_i's are column vectors of X; 2. a proper k^* is chosen to minimize the risk among all the P_k's. Finding the projection in Step 1 turns out to be NP-hard. However, by using the semi-definite programming technique in [41], we can compute an approximately optimal projection and therefore an approximate minimax risk in polynomial time.

Theorem 2. *For any given $n \times p$ matrix X, $0 < q \leq 1$, and $\sigma \geq 0$, we can compute an $O(c_q \log p)$ approximation[3] of $R^*(X\ell_q, \sigma)$ in polynomial time. When $q = 1$, there is a randomized polynomial time estimator that is within $O(\log p)$ factor of the optimal.*

The above two results assume that the radius of ℓ_q ball is given. For $q = 1$, we can extend the estimator to the adaptive case when $\|\theta\|_1$ is unknown. Using the similar idea to the projected nearest neighbor estimator, we have that

Theorem 3. *There is a polynomial time adaptive estimator A such that for any given $n \times p$ matrix X, θ, and $\sigma > 0$,*

$$\mathrm{err}_A(X\theta, \sigma) = O(\log p \cdot R^*(X\ell_1(\|\theta\|_1), \sigma) + \sqrt{n \log n}\sigma^2)\,. \tag{3}$$

Notice that the first term of the above error is $O(\log p)$ factor within the oracle risk bound when $\|\theta\|_1$ is given. While we do not quite get the true oracle bound due to the presences of the additive term of $\sqrt{n \log n}\sigma^2$, the bound becomes a true (and non-trivial) oracle bound for a rather large range of $\|\theta\|_1$. See Remark 7 for a more detailed discussion.

[2] Throughout this paper, the O notation only hides some absolute constant, i.e. a constant independent of any of the parameters, such as $n, p, q, \sigma, X, \theta, y$.

[3] A quantity a is a c-approximation of $a^* \geq 0$, if $a^* \leq a \leq ca^*$.

1.3. *Intuition.* We provide some high level intuition of the projected nearest neighbor estimator. The orthogonal projection estimator, by projecting the observation to a chosen subspace, effectively identifies the "leading factors" in the ground truth set. It works well when K is "skewed". However by simple projection, it ignores the detailed local geometry of K. This makes it less effective when K has many constraints or has constraints involving many dimensions, e.g. when K satisfies sparse constraints. On the other hand, the nearest neighbor estimator, by projecting to the nearest neighbor, depends more on the local geometry of K. But it ignores the global geometry of K so it works well when the body is not skewed along any direction. In some sense, the projected nearest neighbor estimator achieves the optimality by taking both global and local geometry into account: it first identifies the skewed dimensions and then applies the nearest neighbor estimator to the "residual" space which is less biased.

It is long known that the nearest neighbor estimator may be far away from the optimal when there is strong correlation among column vectors of the design matrix X [21, 22, 45]. There have been many methods proposed to deal with this problem. The projection phase can be viewed as one way to remove the correlation such that the residual vectors are less biased. This might not be obvious as the projection only minimizes the maximum of ℓ_2 norm of the projection, a seemingly different quantity. However, in order for the the projected vectors to be all short, they necessarily "span" all the directions because otherwise we could "tilt" the projection to reduce the longest projection. This intuition can actually be made rigorous with the help of tools from Convex Geometry [8].

The technical analysis of the projected nearest neighbor estimator is inspired by two recent works, one is the analysis on the nearest neighbor estimator by Raskutti, Wainwright, and Yu [34]; the other is on the optimality of the orthogonal projection estimator by Javanmard and the author [26]. In [34], it is shown that if X satisfies a certain isometry property, then the nearest neighbor estimator is close to optimal. On the other hand, [26] shows that for symmetric convex bodies there always exists a projection such that the orthogonal projection estimator is close to optimal. At the very high level, we combine the analysis of these two results and show that there always exists a nearly optimal projection of X such that the bound in [34] is nearly optimal on the projected body.

While the main machinery in our analysis is similar to what is in [34] and [26], we need further insights for our problem. For the nearest neighbor analysis, we need a slightly different analysis than [34] to obtain an upper bound suit our purpose. This also allows our result hold for all ranges of p, n.

The lower bound is obtained by extending the techniques in [26] to the sets of the form $X\ell_q$ for $0 < q \leq 1$. The technique utilizes some classical results from Banach space geometry, first started by Bourgain and Tzafriri [8] and fully developed by Szarek, Talagrand, and Giannopoulous [36, 23].

Despite its somewhat involved analysis, the projected nearest neighbor estimator suggests a quite natural heuristic: project $K = X\ell_q$ to a subspace to make it more "round" before applying other estimators (in our case the nearest neighbor estimator). This approach is probably already being used in practice. As the main result in this paper, we prove that such heuristics can actually lead to a nearly optimal estimator. In addition, a nearly optimal projection can be found in polynomial time via semi-definite programming technique in [41].

For the adaptive estimator, we consider the case of $q = 1$. The well known Lasso [38] and Dantzig selector [14] can be viewed as the adaptive version of the nearest negibhor estimator. According to [5], these estimators can achieve an error bound dependent on $\|\theta\|_1$, which is the same as the oracle risk bound of PNN when the projection is taken as the identity projection. We can apply Lasso or Dantzig selector to the projection of X and to obtain the oracle risk bound of PNN under different projection dimensions. This way, we can obtain a set of estimations among which one achieves the true oracle risk bound! Unfortunately, we cannot reliably determine which one it is. By using ideas from hypothesis testing, we can only choose one within $O(\sqrt{n \log n}\sigma^2)$ error, which accounts for the additive bound in Theorem 3.

More concretely, in PNN, the optimal projection dimension is a staircase function of the parameter radius. So we try to "guess" $\|\theta\|_1$ at those critical values at which the optimal dimension changes value. The problem then reduces to a hypothesis testing problem on whether $y = X\theta$ belongs to some convex body. By using the statistics of $\|\widetilde{y} - \widehat{y}\|_2^2$, we can achieve the claimed bound. Our procedure is similar in spirit to the classical Lepski's recipe [28, 6] for converting a non-adaptive estimator to an adaptive one. But there is a significant difference as the PNN estimator is non-linear, and the projections at different dimensions lack a nested structure. As a result, our bound leaves an additive gap of $\sqrt{n \log n}\sigma^2$.

1.4. *Related work.* There are vast amounts of work on the minimax risk estimator. We refer to [30, 39, 27] for comprehensive surveys. Despite many studies on this subject, optimal or nearly optimal estimators are only known for special types of bodies.

One particularly interesting case is when the parameter space is sparse. It is long known that no linear estimator works well under such constraints (see for example [20]). Instead, one needs non-linear estimator such as the thresholding estimator to achieve nearly optimal risk. Recently, much attention has been paid to the (hard) sparsity constraint defined as the number of non-zero components, dubbed as ℓ_0 quantity, of a vector. This problem, called *compressive sensing* in the literature, is computationally infeasible in general so the study has focused on the condition under which nearly optimal polynomial time estimator exists [1, 2, 7, 16, 13, 14, 11, 4, 43, 44].

The case of $q = 1$ is closely related to Lasso [38], which is the nearest neighbor estimator for the case of $q = 1$ and later evolves to solving a regularized nearest neighbor problem with the ℓ_1 norm penalty. While Lasso has proved to be very effective, it is known that when the design matrix has strong correlation, the Lasso estimator may not produce a good estimation [21, 22]. Various methods have been proposed to remove the correlations [21, 22, 45] by using different penalty terms. The projected nearest neighbor estimator can also be viewed as a way to remove correlation. The difference is that our method can be shown to be close to the optimal solution for any design matrix X. In the projected nearest neighbor estimator, we choose the projection dimension that balance two error terms. Similar technique has appeared before. For example, in [3], the estimation is chosen among greedy approximations of the span of vectors of varying size, and the optimal choice is by balancing two error terms. In [10], the dimension is controlled by a stopping rule dependent on the noise structure. Despite these similarity, the optimality of the projected nearest neighbor estimator requires careful choice of the projection via solving a semi-definite program. It is unlikely that the greedy algorithm can achieve the same goal. On the other hand, the computational efficiency of the greedy algorithm makes it (or some variation) an attractive practical alternative to the more complex projection phase in this paper.

Many authors also consider (arguably more flexible and realistic) soft sparsity constraints in the form of $\theta \in \ell_q$ for $0 < q \leq 1$, the setting considered in this paper. In [19], asymptotically tight bounds are obtained for $X = I$, the identity matrix. A similar notion of roughness was studied in [29] in which soft-thresholding estimator is shown to be nearly optimal, again for $X = I$, but extended to more general noise and loss models. In [17], it is shown that there exists design matrices X which allow fairly accurate estimation when there is no noise. In [42], the authors presented several upper bounds, dependent on the design matrix X, on the loss of the Lasso and Dantzig selector methods when applied to soft sparsity constraints. They also show that the upper bound is nearly optimal for a family of X's. Then in [34], it is shown that the nearest neighbor estimator is nearly optimal if X

satisfies certain isometry property which holds for Gaussian random matrix X. In [18], it is shown that for Gaussian random matrix, the (polynomial time) ℓ_1 penalized least squares is nearly optimal. Despite all these studies, no nearly optimal estimator is known for general design matrix X. So our knowledge is limited to the case where X is a diagonal matrix or when X satisfies strong isometry properties. In [12], the authors showed a lower bound of the minimax risk on the estimation of θ for any design matrix and with the hard sparsity constraint, but it could be far away from the upper bound in general.

Among previous work, [34] is particularly relevant to our current work. In [34], the authors show, among many other results, an upper bound for the nearest neighbor estimator which depends on q and the radius of K. While this could be far away from the optimal, it turns out if we apply proper projection of K, the radius of the projection can be made so that the resulted bound is near optimal. For this we follow similar approach in [26], in which they show that the orthogonal projection estimator is nearly optimal for symmetric linear constraints. But we need to adapt the argument in [26] as $X\ell_q$ have exponentially many faces and can be non-convex.

As mentioned earlier, the transformation from non-adaptive estimator to the adaptive one is similar to Lepski's method [28, 6] but there are significant differences as our non-adaptive estimator does not quite satisfy the properties required by Lepski's method.

2. Preliminaries.

2.1. *Basic notations and definitions.*

For a vector $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and $q > 0$, denote by $\|x\|_q = (\sum_i |x_i|^q)^{1/q}$. When $p \geq 1$, $\|x\|_q$ is a norm. When $0 < q < 1$, $\|x\|_q$ is not a norm but it is quasi-convex as there is a constant c dependent on q such that for any x, y, $\|x+y\|_q \leq c(\|x\|_q + \|y\|_q)$. We use $\ell_q^p(r)$ to denote the p-dimensional q-ball with radius r, i.e.

$$\ell_q^p(r) = \{x \in \mathbb{R}^p \,:\, \|x\|_q \leq r\}\,.$$

We often drop p when the dimension is clear from the context. We use ℓ_q as a short hand for $\ell_q(1)$. For a set $K \subseteq \mathbb{R}^n$ containing the origin, define the q-radius of K as $\|K\|_q = \sup_{x \in K} \|x\|_q$. In all these notations, whenever q is omitted, it means $q = 2$.

We use $\mathcal{G}^n(\sigma)$ to denote the distribution of n-dimensional Gaussian random variable with covariance matrix $\sigma^2 I$. Again, we often drop n and σ when they are clear from the context.

As standard, $f = O(g)$ if there exists a constant $c > 0$ such that $f \leq c \cdot g$ and $f = \Omega(g)$ if there exists a constant $c > 0$ such that $f \geq c \cdot g$. Throughout this paper, high probability is understood as the probability of $1 - 1/n^2$.

2.2. *Minimax risk.* An estimator M is a map from $\mathbb{R}^n$ to $\mathbb{R}^n$: it takes a noisy observation $\widetilde{y} = y + g$ of an unknown vector $y \in \mathbb{R}^n$ and maps it to an estimation $\widehat{y} = M(\widetilde{y})$. Here we consider the noise drawn from $\mathcal{G}^n(\sigma)$. As described early, the risk $R_M(K, \sigma)$ of M is defined as the maximum expected error among y in K, i.e.

$$R_M(K, \sigma) = \sup_{y \in K} \mathbb{E}_{\widetilde{y} = y + g; g \sim \mathcal{G}(\sigma)} [\|M(\widetilde{y}) - y\|^2] \,.$$

The minimax risk of K is defined as the minimum achievable risk for K, i.e. $R^*(K, \sigma) = \inf_M R_M(K, \sigma)$. We state a well known lower bound on the minimax risk of Euclidean balls which we will use later.

Lemma 4. $R^*(\ell_2^n(r), \sigma) = \Omega(\min(n\sigma^2, r^2))$.

2.3. *Orthogonal projection estimator.* The orthogonal projection estimator T is a special type of linear estimator. It is defined with respect to some linear subspace. The estimation is simply by projecting the observation $\widetilde{y} \in \mathbb{R}^n$ to the subspace. Let $\mathcal{P}_k$ denotes all the k-dimensional linear subspaces in $\mathbb{R}^n$. For $P \in \mathcal{P}_k$, we also use P denote the orthogonal projection to P. The estimator T_P is then defined as $T_P(\widetilde{y}) = P\widetilde{y}$.

Since Gaussian random vector is invariant under the rotation, we have that $R_{T_P}(K, \sigma) = k\sigma^2 + \sup_{y \in K} \|y - Py\|^2 = k\sigma^2 + \sup_{y \in K} \|P^\perp y\|^2$, where $P^\perp$ denotes the $n - k$ dimensional subspace orthogonal to P. For $0 \leq k \leq n$, define *Kolmogorov width* (as in [31]) as

$$d_k(K) = \inf_{P \in \mathcal{P}_k} \sup_{y \in K} \|y - Py\| \,.$$

For ℓ_2 norm, this definition is equivalent to following more convenient form, which we will use through the paper.

$$d_k(K) = \inf_{P \in \mathcal{P}_k} \|P^\perp(K)\| = \inf_{P \in \mathcal{P}_{n-k}} \|P(K)\| \,.$$

Clearly, $d_k(K)$ is monotonically decreasing with k. Kolmogorov width determines the minimax risk of the orthogonal projection estimators [20]. Let R_T denote the minimum risk among all the orthogonal projection estimators.

Lemma 5. $R_T(K, \sigma) = \min_k(k\sigma^2 + d_k(K)^2)$.

The orthogonal projection estimator is long known to be nearly optimal for ellipsoids [32, 25] and more generally for quadratically convex and orthosymmetric objects [20]. However, it is also well known that the orthogonal projection estimator (actually any linear estimator) can be far away from optimal for the ℓ_1 ball and therefore does not work well for linear regression with sparsity constraints.

Lemma 6 ([20]).

$$R_T(\ell_1^n, 1/\sqrt{n}) = \Omega(\sqrt{n/\log n} R^*(\ell_1^n, 1/\sqrt{n}))\,.$$

2.4. *Nearest neighbor estimator.* The nearest neighbor estimator is another well known estimator. It maps an observation to the nearest point on K, i.e. $N_K(\widetilde{y}) = \operatorname{argmin}_{\widehat{y} \in K} \|\widehat{y} - \widetilde{y}\|$. The nearest neighbor estimator is a non-linear estimator and works well for "skinny" objects such as the ℓ_1 ball. However, we can construct an example (Section 6.1) to demonstrate it is far from optimal. Denote by $R_N(K, \sigma)$ the risk of the nearest neighbor estimator.

Lemma 7. *There exist ellipsoids $E_n \subset \mathbb{R}^n$ for $n = 1, 2, \ldots$ such that $R_N(E_n, 1) = \Omega(\sqrt{n} R^*(E_n, 1))$.*

3. Projected nearest neighbor estimator.

We now describe the projected nearest neighbor estimator, which is defined with respect to some low dimensional orthogonal projection. Given a k-dimensional subspace $P \in \mathcal{P}_k$, we define the projected nearest neighbor estimator H_P as follows. Let $P^\perp$ denote the $n-k$ dimensional subspace orthogonal to P. Recall that we also use Px, $P^\perp x$ to denote, respectively, the orthogonal projection to the space P and $P^\perp$. The estimator H_P is defined as

$$H_P(\widetilde{y}) = P\widetilde{y} + N_{P^\perp K}(P^\perp \widetilde{y})\,.$$

In other words, H_P consists of two components, one of which is the projection to the subspace P and the other the nearest neighbor of $P^\perp \widetilde{y}$ to $P^\perp K$. We use $R_H(K, \sigma) = \inf_q R_{H_P}(K, \sigma)$ to denote the minimum risk achievable by the projected nearest neighbor estimator for given K, σ.

When the projection is set as the identity projection, the corresponding PNN is the same as the nearest neighbor estimator. In addition, for the same projection, the projected nearest neighbor estimator outperforms the corresponding orthogonal projection estimator. So the projected nearest neighbor estimator subsumes both the nearest neighbor and the orthogonal projection estimators. In the following, we give an example to show the projected nearest neighbor estimator can outperform both the orthogonal projection and the nearest neighbor estimators by a large factor.

Example 8. Consider the ellipsoid defined as

$$E_{n,k} = \{x \,:\, \frac{1}{\sqrt{n}} \sum_{i=1}^{k} x_i^2 + \sum_{i=k+1}^{n} x_i^2 \le 1\} .$$

Let $K = E_{n^2,k} \times \ell_1^n(\sqrt{n})$ with $k \le n$. By the above discussion, we can see that for the orthogonal projection estimator $R_T(K,1) = \Theta(n)$, and for the nearest neighbor estimator $R_N(K,1) = \Theta(n)$, but $R_H(K,1) = O(\sqrt{n \log n})$ by setting P to be the k-dimensional projection spanned by the k long axes of $E_{n^2,k}$. This demonstrates a large gap between the projected nearest neighbor estimator and both the orthogonal projection and the nearest neighbor estimators.

To study the performance of the projected nearest neighbor estimator. We first need the following error bound for the nearest neighbor estimator from [34].

Proposition 9. *For $0 < q \le 1, K = X\ell_q^p$, the nearest neighbor estimator N has risk*

$$R_N(K,\sigma) = O(c_q \|K\|^q \sigma^{2-q} (\log p)^{1-q/2}) ,$$

where $c_q = O(2^{\frac{1}{q}} \frac{1}{q} \ln \frac{2}{q})$ is a constant dependent on q only.

The above bound is almost identical to Theorem 4(a) in [34]. We will present a slightly different proof which applies to wider combination of parameters. For clarity and completeness, we present the proof in Section 6.2. According to Proposition 9, the error is bounded by $\|K\|^q$. Hence, if we fix the dimension of the projection in a PNN estimator, in order to minimize the risk, we should seek the projection P that minimizes $\|PK\|$, i.e. realizes Kolmogorov width. By using this projection, we obtain the following upper bound of the projected nearest neighbor estimator.

Corollary 10. *For any $0 < q \le 1$ and any $K = X\ell_q^p$,*

$$R_H(K,\sigma) = O(\min_{0 \le k \le n} (k\sigma^2 + c_q d_k(K)^q \sigma^{2-q} (\log p)^{1-q/2})) , \tag{4}$$

where c_q is the same as in Proposition 9.

Proof. For any fixed k, the error consists of two terms: $O(k\sigma^2)$ for the projection, and $O(c_q d_k(K)^q \sigma^{2-q}(\log p)^{1-q/2})$ for the nearest neighbor estimation. The second term comes from Proposition 9 with $\|K\|$ replaced by $d_k(K)$ if we apply the projection that realizes $d_k(K)$. Clearly, we can choose k with the minimum bound. □

To show (4) is nearly optimal, we prove an almost matching lower bound in terms of the Kolmogorov width. This is the key technical contribution of the paper and relies on the classic restricted invertibility property developed by Bourgain and Tzafriri [8]. The proof of is in Section 6.3.

Theorem 11. *For $K = X\ell_q^p$,*

$$R^*(K,\sigma) = \Omega(\max_{0\le k\le n} \min(k\sigma^2, k^{1-2/q} d_k(K)^2))\,. \tag{5}$$

Theorem 1 follows readily from Corollary 10 and Theorem 11 by setting k to equalize two terms in (5). The details are in Section 6.4.

Remark 1. In the proof of Theorem 1, we choose k^* such that $d_k(X) \approx k^{1/q}\sigma$. When q goes to 0, then k^* goes to 1. Therefore, when q is close to 0, the projected nearest neighbor estimator becomes the ordinary nearest neighbor algorithm. As stated in Theorem 4(b) in [34], the risk of the nearest neighbor estimator is $O(s\log(p/s)\sigma^2)$ for $\theta \in \ell_0(s)$. On the other hand, if the rank of X is at least s, then $R^*(X\ell_0(s),\sigma) = \Omega(s\sigma^2)$. Hence the nearest neighbor estimator (and the projected nearest neighbor estimator) is $O(\log p)$ minimax for the hard sparsity constraint. This is consistent with the bound in Theorem 1 by letting $q \to 0$.

Remark 2. In the proof of Theorem 11, we actually showed that there exists a submatrix X' which consists of $k \le n$ columns of X such that the minimax risk of $X'\ell_q^k$ is close to that of $X\ell_q^p$. In some sense, this means that there is a hardest sub-problem which has at most n columns.

Remark 3. Our technique still leaves a gap of $(\log p)^{1-q/2}$. We do not know if this gap is inherent to the projected nearest neighbor estimator or due to the deficiency of the analysis. We note that the upperbound cannot be improved in general, as demonstrated by the example of ℓ_1 ball. There might be a chance to improve the lowerbound by a factor of $\sqrt{\log k}$ by more sophisticated techniques. But this is still insufficient to close the gap as k might be much smaller than p.

Remark 4. While PNN may sound similar to the technique of low dimension projection, there are significant differences. For example, when applying low dimension projection, we typically would like to preserve the original metric structure, and often a random projection suffices. In our case, however, we would like to make the projection as small as possible, and it requires more careful selection of the projection. Indeed, it is easy to show that a random projection would fail for our purpose.

4. Algorithms.

While the analysis of projected nearest neighbor estimators is somewhat involved, the resulted algorithm is quite straightforward. There are two separate parts in the projected nearest neighbor estimator. First, for given K and σ, compute the optimal projection P and k. Second, for any observation $\widetilde{y}$, apply the projection and then compute the nearest neighbor of $P^{\perp}\widetilde{y}$ to $P^{\perp}K$.

We will describe these two steps separately. For the first step, by the proof of Theorem 1, it suffices to compute $d_k(K)$. This problem is however NP-hard [9]. But since $K = X\ell_q^p$, $\|P^{\perp}K\|$ must be realized at one of p column vectors of X (see the proof of Lemma 19). Let $V = \{x_i \ : \ i = 1, \ldots, p\}$ be the p column vectors of X. Then computing $d_k(K)$ reduces to computing an $n-k$ dimensional projection P' such that $\max\{\|P'v\| \ : \ v \in V\}$ as small as possible. This problem has been studied in [41], and it is shown one can compute an $O(\sqrt{\log p})$ approximation by the semi-definite programming relaxation. The following proposition is the main result of [41].

Proposition 12. *For any $n \times p$ matrix X, $0 < q \leq 1$, and $0 \leq k \leq n$, we can compute in polynomial time an $O(\sqrt{\log p})$ approximation to $d_k(X\ell_q)$. In addition, we can compute an $n-k$ dimensional subspace P' in randomized polynomial time such that with high probability, $\|P'(X\ell_q)\| = O(\sqrt{\log p}\, d_k(X\ell_q))$.*

As for the second step, we need to compute the nearest neighbor on $K = X\ell_q^p$ for any given point. This can be done by convex programming for $q = 1$. Unfortunately, we do not know how to compute it efficiently for $q < 1$. So we can only claim polynomial time nearly optimal estimator for $K = X\ell_1$, as described in Algorithm 1. For description simplicity we have described the algorithm in which we try all $k = 1, 2, \ldots, n$. Since $d_k(K)$ is monotonically decreasing, the complexity can be reduced by using a binary search. Theorem 2 follows from the above discussion.

The following proof summarizes our above discussion.

Proof. **[Theorem 2]** By Proposition 12, we can compute an $O(\sqrt{\log p})$

Algorithm 1 Nearly optimal estimator for $X\ell_1$.

```
Input: design matrix $X$ and observation $\widetilde{y}$.
Output: $\widehat{y}$.
Let $x_1, \ldots, x_p$ be column vectors of $X$. Denote the set by $Y$;
for $k \in \{1, \ldots, p\}$ do
    Compute a projection $P_k$ such that $z_k = \|P_k Y\| = O(\sqrt{\log p}) d_k(K)$;
    Compute $r_k = k\sigma^2 + z_k \sigma \sqrt{\log p}$;
end for
Pick $k^* = \mathrm{argmin}_k\, r_k$, and let $P = P_{k^*}$ and $P^\perp$ be the subspace orthogonal to $P$;
Compute $\widehat{y}'$ as the nearest neighbor of $P^\perp \widetilde{y}$ to the convex hull of $\pm P^\perp x_1, \ldots, \pm P^\perp x_p$.
This can be done by using any polynomial time convex programming algorithm.
Set $\widehat{y} = P\widetilde{y} + \widehat{y}'$.
```

approximation d'_k of $d_k(X\ell_q)$. Using this approximation, we compute

$$R' = O\big(\min_{0 \le k \le n} (k\sigma^2 + c_q d'^q_k \sigma^{2-q} (\log p)^{1-q/2})\big) .$$

Since $d_k(K) \le d'_k \le c\sqrt{\log p}\, d_k(K)$ for some constant $c > 0$, we have that

$$R_H(K, \sigma) \le R' = c^q \log^{q/2} p R_H(K, \sigma) .$$

By Theorem 1, R_H is an $O((\log p)^{1-q/2})$ approximation of R^*, so R' is an $O((\log p)^{q/2} (\log p)^{1-q/2}) = O(\log p)$ approximation of R^*.

When $q = 1$, by Proposition 12, we can compute the nearly optimal projection P and use convex programming to compute the nearest neighbor of $P\widetilde{y}$ to $PX\ell_1$. The former can be done in randomized polynomial time and the latter in polynomial time. □

Remark 5. The first step of the algorithm uses the semi-definite programming relaxation to compute a nearly optimal projection of $X\ell_q$. While it has guaranteed approximation ratio, it can be time consuming. In practice, the projections on the principal subspaces of $X\ell_2$ might serve as a good heuristics.

Remark 6. We do not have a polynomial time estimator for $0 < q < 1$ because of the lack of a polynomial time algorithm for computing the nearest neighbor to the non-convex body of $K = X\ell_q$. While such nearest neighbor problem is hard, for our purpose an approximate nearest neighbor is sufficient. In addition, we only need to succeed in an average sense as $\widetilde{y} = y + g$ for $y \in K$ and g an i.i.d. Gaussian noise. It is interesting to know if there exists an efficient procedure in this particular setting. We note that this problem can be formulated under the framework of the smoothed analysis [35]. In both cases, we are interested in minimizing the expected performance of an algorithm (or an estimator) in the worst case.

5. Adaptive estimator when C is not given.

The projected nearest neighbor estimator in the last section is nearly minimax optimal once the sparsity radius is given. In this section, we extend the same idea to design an adaptive estimator to deal with the case when the sparsity radius is not known. Write $C = \|\theta\|_1$. Ideally, one would like to achieve some kind of oracle inequality with the error bound proportional to $R^*(X\ell_1(C), \sigma)$, i.e. the nearly optimal risk bound assuming C is available. We can only partially achieve this goal with an extra additive term of $\sqrt{n \log n}\sigma^2$. Here we will focus on the case of $q = 1$ for the simplicity of the exposition.

Again let $K = X\ell_1$. Intuitively, the adaptive estimator will search for the unknown C at some discrete values. In view of the upper bound in Corollary 10, we will only try those C's which equalize the two error terms in (4).

Define $C_k = k\sigma/d_k(K)$ for $k = 0, 1, \cdots, n/2$. C_k has the following properties:

1. $C_0 \leq C_1 \leq C_2 \leq \cdots$ is monotonically increasing, since d_k is non-increasing.
2. There is a constant $c > 0$, for $C \geq C_k$,

$$R^*(X\ell_1(C), \sigma) \geq ck\sigma^2 \,. \tag{6}$$

   This follows from Theorem 11.

Further we define P_k to be the $n - k$ dimensional projection that realizes $d_k(K)$, i.e. minimizes $\max_{1 \leq i \leq n} \|Px_i\|$ among all the $n - k$ dimensional projection. The adaptive estimator will estimate $\widetilde{y}_k = P_k\widetilde{y}$ against $P_kX\ell_1(C_k)$ using the nearest neighbor estimator, starting from $k = 0$. Suppose that the outcome is $\widehat{y}_k$. It is easy to show that among the n estimations $\widehat{y}_k$ for $k = 0, \ldots, n$, there is one that satisfies the true oracle risk bound, i.e. with high probability, there exists $0 \leq k \leq n$ such that

$$\|\widehat{y}_k - y\|^2 = O(\sqrt{\log p}R^*(X\ell_1(\|\theta_1\|), \sigma)) \,.$$

Unfortunately, we cannot determine reliably which one it is. Instead, we can only choose one which is within $O(\sqrt{n \log n}\sigma^2)$ error. This is by finding the minimum k such that $\|\widetilde{y}_k - \widehat{y}_k\|^2$ is not too large (defined precisely later). Algorithm 2 contains a formal description.

Now we will show that the estimator given in Algorithm 2 satisifies the bound stated in Theorem 3. The proof requires some properties on $\|\widehat{y}_k - \widetilde{y}_k\|^2$ as described in Lemma 13. Denote by $y_k = P_ky$ and $K_k = P_kX\ell_1(C_k)$. Let δ_k denote the ℓ_2 distance between y_k and K_k, i.e. $\delta_k = \min_{z \in K_k} \|y_k - z\|$.

Algorithm 2 Adaptive projected nearest neighbor estimator

```
**Input:** design matrix $X$ and observation $\widetilde{y}$.
**Output:** estimation $\widehat{y}$.
 1: for $k \in \{0, 1, \cdots, n/2\}$ do
 2:     Compute the $n-k$ dimensional projection $P_k$ that approximately minimizes $\|PX\|$;
 3:     Compute $\widetilde{y}_k = P_k\widetilde{y}$, $X_k = P_k X$, and $\Delta_k = \max_i P_k x_i$;
 4:     Set $C_k = k\sigma/\Delta_k$
 5:     Compute $\widehat{y}_k$ to be the nearest neighbor of $\widetilde{y}_k$ on $X_k\ell_1(C_k)$
 6:     if $\|\widehat{y}_k - \widetilde{y}_k\|^2 \leq (n-k)\sigma^2 + 2\sqrt{n\log n}\sigma^2$ then
 7:         Set $\widehat{y} = \widehat{y}_k + P_k^{\perp}\widetilde{y}$ and return;
 8:     end if
 9: end for
10: Set $\widehat{y} = \widetilde{y}$.
```

LEMMA 13. *There are constant $c_1, c_2 > 0$ such that the following holds with high probability*

1. *If $y_k \in K_k$, then*
$$\|\widehat{y}_k - \widetilde{y}_k\|^2 \leq (n-k)\sigma^2 + 2\sqrt{n\log n}\sigma^2 \,.$$
2. *If $\delta_k^2 \geq c_1(\sqrt{n\log n}\sigma^2 + k\sigma^2\log p)$, then*
$$\|\widehat{y}_k - \widetilde{y}_k\|^2 \geq (n-k)\sigma^2 + 2\sqrt{n\log n}\sigma^2 \,.$$
3. *If $\delta_k^2 \leq c_1(\sqrt{n\log n}\sigma^2 + k\sigma^2\log p)$, then*
$$\|\widehat{y}_k - y_k\|^2 \leq c_2(\sqrt{n\log n}\sigma^2 + k\sigma^2\log p) \,.$$

By Lemma 13.1 and 2, Step 6 in Algorithm 2 serves as a test for whether y_k is sufficiently separated from K_k. When $y_k \in K_k$, then the test is true with high probability, and the algorithm outputs $\widehat{y}$ and returns. But when the separation between y_k and K_k is large enough ($c_1(\sqrt{n\log n}\sigma^2 + k\sigma^2\log p)$), then Step 6 would test false with high probability. Theorem 3 follows from Lemma 13.

PROOF. [**Theorem 3**] If the test at Step 6 outputs false for some k, then by Lemma 13.1, $y_k \notin K_k$. Thus $y \notin X\ell_1(C_k)$, i.e. $C \geq C_k$. By (6), we have that $R^*(X\ell_1(C), \sigma) \geq ck\sigma^2$.

On the other hand, if Step 6 tests true for k, then by Lemma 13.2, $\delta_k^2 \leq c_1(\sqrt{n\log n}\sigma^2 + k\sigma^2\log p)$, and by Lemma 13.3, $\widehat{y}$ returned at Step 7 satisfies that
$$\|\widehat{y} - y\|^2 = \|\widehat{y}_k - y_k\|^2 + k\sigma^2 \leq c_2(\sqrt{n\log n}\sigma^2 + k\sigma^2\log p) + k\sigma^2 \,.$$

We distinguish three outcomes of Step 6.

- Step 6 tests true for $k = 0$. In this case,
$$\|\widehat{y} - y\|^2 \leq c_2\sqrt{n \log n}\sigma^2 \, .$$
- Step 6 test true for some $k > 0$ and therefore is false for $k - 1$. In this case
$$R^*(X\ell_1(C), \sigma) \geq c(k-1)\sigma^2 \, ,$$
and
$$\begin{aligned}\|\widehat{y} - y\|^2 &\leq c_2(\sqrt{n \log n}\sigma^2 + k\sigma^2 \log p) + k\sigma^2 \\ &= O(\sqrt{n \log n}\sigma^2 + R^*(X\ell_1(C), \sigma) \log p) \, .\end{aligned}$$
- Step 6 is never true so Step 10 is reached. In particular, the test is false for $k = n/2$ and hence $R^*(X\ell_1(C), \sigma) \geq c_1(n/2 - 1)\sigma^2$ but then $\|\widehat{y} - y\|^2 = O(n\sigma^2) = O(R^*(X\ell_1(C), \sigma))$.

In all the above cases, the bound in Theorem 3 holds. □

Remark 7. When $R^*(X\ell_1(\|\theta\|_1), \sigma) \geq \sqrt{n}\sigma^2$, the bound (3) in Theorem 3 becomes a true oracle risk bound (within $O(\log p)$ factor). In view of the proof of Theorem 1, this happens when $\|\theta\|_1 d_{\sqrt{n}}(X\ell_1) \geq \sqrt{n}\sigma$, i.e. when $\|\theta\|_1 \geq \sqrt{n}\sigma / d_{\sqrt{n}}(X\ell_1)$. In such case, the risk ranges between $\sqrt{n \log n}\sigma^2$ and $n\sigma^2$. So the bound (3) is nearly optimal and non-trivial for a rather large range of $\|\theta\|_1$.

Remark 8. It might be possible to apply the Lasso or Dantzig selector estimators to the projection $P_k X$ to obtain $\widehat{y}_k$ and then choose one $\widehat{y}_k$ similar to Algorithm 2. This would probably result in the same bound as in (3). We choose our current exposition because Lemma 13.2 relies on the fact that $\widehat{y}_k$ is the nearest neigbhor to $P_k\widetilde{y}$. It is not immediately clear whether it also holds for Lasso or Dantzig selector.

Remark 9. One may wonder if it is possible to get rid of $\sqrt{n \log n}\sigma^2$ factor and obtain a pure oracle inequality bound. If such a bound is possible, then when $C = 0$, the estimator needs to map all the observations to 0. Since it is impossible to distinguish 0 and a sphere with radius $n^{1/4}\sigma$, there might be a good reason for such an additive separation to be expected.

6. Proofs.

6.1. *Proof of Lemma 7 (bad example for the nearest neighbor estimator).* We will now construct a bad example for the nearest neighbor estimator. While it is well known that the nearest neighbor estimator can be non-optimal, we could not find a definitive reference for a large gap.In our example, we will demonstrate a large gap of $\sqrt{n}$. Consider the ellipsoid

$$E_n = \{y = (y_1, \ldots, y_n) : \sum_{i=1}^{n-1} y_i^2 + \frac{y_n^2}{\sqrt{n}} \leq 1\} .$$

Set $\sigma = 1$. The orthogonal projection estimator $M(\widetilde{y}) = (0, \ldots, 0, \widetilde{y}_n)$ has minimax error

$$M(\widetilde{y}) = \sum_{i=1}^{n-1} y_i^2 + \mathbb{E}[(\widetilde{y}_n - y_n)^2] \leq 2 . \tag{7}$$

On the other hand, we show that the nearest neighbor estimator has error $\Omega(\sqrt{n})$. For any $\widetilde{y} = (\widetilde{y}_1, \ldots, \widetilde{y}_n)$, by using Lagrangian multiplier, we have that the nearest point $\widehat{y}$ to $\widetilde{y}$ on E_n satisfies that $\widetilde{y}_i = (1+\lambda)\widehat{y}_i$ for $i = 1, \ldots, n-1$ and $\widetilde{y}_n = (1+\lambda/\sqrt{n})\widehat{y}_n$. Now, pick $y = (0, \ldots, 0, n^{1/4}) \in E_n$. Then with high probability $\sum_{i=1}^{n-1} \widetilde{y}_i^2 = \Omega(n)$. By

$$\sum_{i=1}^{n-1} \widetilde{y}_i^2 = (1+\lambda)^2 \sum_{i=1}^{n-1} \widehat{y}_n^2 \leq (1+\lambda)^2 ,$$

we have $\lambda = \Omega(\sqrt{n})$. But then $\widehat{y}_n \leq c\widetilde{y}_n \leq cn^{1/4}$ for some constant $c < 1$. Thus, with high probability $\|\widehat{y} - y\| = \Omega(n^{1/4})$. So the nearest neighbor estimator has error $\Omega(n^{1/2})$. Since the projection estimator achieves the risk of $O(1)$, we have constructed an example to show that the nearest neighbor estimator can be $\Omega(\sqrt{n})$ factor larger than the optimal.

6.2. *Proof of Proposition 9.* It is well known that the error of the nearest neighbor estimator is determined by the metric structure of K. For two bodies $K_1, K_2 \subseteq \mathbb{R}^n$, define the (dyadic) entropy number $e_k(K_1, K_2)$, for any $k \geq 0$, as the minimum ϵ such that K_1 can be covered by 2^k copies of ϵK_2. When K_2 is the unit ℓ_2 ball, we simply write it as $e_k(X_1)$.

For a random vector $g \in \mathcal{G} = \mathcal{G}^n(1)$ and any $y \in \mathbb{R}^n$, let g_y denote the random variable $g \cdot y \in \mathbb{R}$. The classical Dudley bound states that there is a constant $c > 0$ such that

$$\mathbb{E}_{g \sim \mathcal{G}}[\sup_{y \in K} |g_y|] \leq c \sum_{k=0}^{\infty} 2^{k/2} e_{2^k}(K)$$

We need a slight variation of the above bound where the summation is over k above some threshold. For $\delta \geq 0$, write

$$\begin{aligned} k(\delta) &= \lfloor \log(\min\{k : \ e_k(K) \leq \delta\}) \rfloor \,, \\ \gamma(K,\kappa) &= \sum_{k=\kappa}^{\infty} 2^{k/2} e_{2^k}(K) \,, \\ K(\delta) &= K \cap \ell_2^n(\delta) \,. \end{aligned}$$

With the above notations,

Lemma 14. *There is a constant $c > 0$, for any $t > 0$,*

$$\mathrm{Prob}_{g\sim\mathcal{G}}[\sup_{y \in K(\delta)} |g_y| \geq t\gamma(K, k(\delta))] \leq \exp(-ct^2 2^{k(\delta)}) \,.$$

Proof. By the standard chaining argument [37]. Clearly the result holds if we replace $e_k(K)$ with any upper bound of $e_k(K)$. □

Now we prove Proposition 9. Without loss of generality, we assume $\sigma = 1$. We apply the standard technique to bound the error of the nearest neighbor estimator by the supreme of Gaussian processes [40, 34]. The starting point is the well-known observation that for $\widehat{y} = N_K(\widetilde{y})$,

$$\|\widehat{y} - y\|^2 \leq 2(\widetilde{y} - y) \cdot (\widehat{y} - y) \,. \tag{8}$$

Since $\widehat{y}, y \in K = X\ell_q^p$ and by the quasi-convexity of ℓ_q^p for $0 < q \leq 1$, we have that $\widehat{y} - y \in c'K$ for $c' = 2^{\frac{1}{q}}$. Observe that $g = \widetilde{y} - y$ is a Gaussian random vector. We can bound $\|\widehat{y} - y\|$ through Dudley bound over ℓ_q^p ball as follows.

To apply Lemma 14, we need an estimate on the entropy number of $K = X\ell_q^p$. Write $\Delta = \|K\|$. The following is a consequence of [15, 24]. For completeness, we include the derivation in Appendix A.

Lemma 15.

$$e_{2k}(X\ell_q^p, \ell_2^n) = \begin{cases} O(\Delta) & k \leq \log p \\ O\left(\left(f_q \frac{\log(1+p/k)}{k}\right)^{1/q-1/2} \Delta\right) & \log p \leq k \leq p \\ O\left(2^{-2k/p}(f_q/p)^{1/q-1/2}\Delta\right) & k \geq p \,. \end{cases} \tag{9}$$

where $f_q = O(\frac{1}{q} \ln \frac{2}{q})$ is a constant dependent on q only.

Now the crucial lemma is

Lemma 16. *Suppose that $\Delta \le p^{1/q}(\log p)^{1/2}$ and $\Delta/p^{1/q-1/2} \le \delta \le \Delta$, for any constant $d > 0$, there exists $c(q,d) > 0$, dependent on q and d only, such that*

$$\begin{aligned} & \mathrm{Prob}_{g\sim\mathcal{G}}[\sup_{y\in K, \|y\|\le\delta} |g_y| \ge c(q,d)\Delta^{\frac{q}{2-q}}\delta^{\frac{2-2q}{2-q}}\sqrt{\log p}] \\ \le \quad & p^{-d(\frac{\Delta}{\delta})^{q/2}}. \end{aligned}$$

Proof. The proof is by applying Lemma 14 and 15. By Lemma 15, for

$$\Delta/p^{1/q-1/2} \le \delta \le \Delta\,,$$

we have,

$$k(\delta) = O((\Delta/\delta)^{\frac{2q}{2-q}}\log p) = O(p)\,.$$

Therefore,

$$\begin{aligned} \gamma(K, k(\delta)) &= \sum_{k=k(\delta)}^{\infty} 2^{k/2} e_{2^k}(K) \\ &= \sum_{k=k(\delta)}^{\log p} 2^{k/2} e_{2^k}(K) + \sum_{k=\log p}^{\infty} 2^{k/2} e_{2^k}(K)\,. \end{aligned}$$

By Lemma 15, it is easily seen that for both terms, the dominant term is the first term, i.e. when $k = k(\delta)$ and $k = \log p$, respectively. Plugging in $e_k(K)$ for these values, we have

$$\begin{aligned} \gamma(K, k(\delta)) &\le O\left(\delta\sqrt{(\Delta/\delta)^{\frac{2q}{2-q}}\log p}\right) + O\left(\sqrt{p}\Delta/p^{1/q-1/2}\right) \\ &\le O\left(\Delta^{\frac{q}{2-q}}\delta^{\frac{2-2q}{2-q}}\sqrt{\log p} + p^{1-1/q}\Delta\right)\,. \end{aligned}$$

It is easy to verify that with $\delta \ge \Delta/p^{1/q-1/2}$,

$$\Delta^{\frac{q}{2-q}}\delta^{\frac{2-2q}{2-q}}\sqrt{\log p} \ge c\Delta p^{1-1/q}\sqrt{\log p}\,,$$

for some constant $c' > 0$. So the first term dominates , that is

$$\gamma(K, k(\delta)) = O(\Delta^{\frac{q}{2-q}}\delta^{\frac{2-2q}{2-q}}\sqrt{\log p})\,.$$

The claim now follows from Lemma 14. □

With the above preparation, we are ready to prove Proposition 9.

Proof. **[Proposition 9]** We assume $\sigma = 1$. Recall $\Delta = \|K\|$. We can further assume

$$\sqrt{\log p} \leq \Delta \leq n^{1/q}/(\log p)^{(2-q)/2q} \,. \tag{10}$$

Otherwise the claim follows immediately by using the trivial bound of $O(\min(\Delta^2, n\sigma^2))$. Together with the assumption that $p = \Omega(n/\log n)$, the upper bound in (10) implies that

$$\Delta = O(p^{1/q}(\log p)^{1/2}) \,. \tag{11}$$

Write $\delta_0 = c\Delta^{q/2}(\log p)^{1/2-p/4}$ for some sufficiently large c such that $\delta_0 \geq \Delta/p^{1/q-1/2}$. This is possible as $\Delta = O(p^{1/q}(\log p)^{1/2})$. Hence, by applying Lemma 16, we have that for $\delta_0 \leq \delta \leq \Delta$ and any $d > 0$ there exists $c(q, d) > 0$ such that

$$\mathrm{Prob}_{g\sim\mathcal{G}}[\sup_{y\in K(\delta)} |g_y| \geq c(q,d)\Delta^{\frac{q}{2-q}}\delta^{\frac{2-2q}{2-q}}\sqrt{\log p}] \leq p^{-d(\frac{\Delta}{\delta})^{q/2}} \,.$$

Now denote by $\mathcal{E}$ the following event

$$\begin{aligned} &\exists y \;\; (\delta_0 \leq \|y\| \leq \Delta) \\ &\quad \wedge \left(|g_y| \geq t\sqrt{\log p}\Delta^{\frac{q}{2-q}}\|y\|^{\frac{2-2q}{2-q}}\right) \,. \end{aligned}$$

By the peeling argument we show that we can choose t, dependent on q only, such that $\mathrm{Prob}[\mathcal{E}] \leq p^{-4/q}$. Define

$$\overline{K}(\delta) = K(\delta) \setminus K(\delta/2) \,.$$

Clearly $\overline{K}(\delta) \subseteq K(\delta)$ and for any $y \in \overline{K}(\delta)$, $\|y\| \geq \delta/2$. By these we have

$$\mathrm{Prob}[\sup_{y\in\overline{K}(\delta)} |g_y| \geq t_q\sqrt{\log p}\Delta^{\frac{q}{2-q}}\|y\|^{q/2}] \leq p^{-d(\Delta/\delta)^{q/2}} \,.$$

Hence for any $d > 0$, there is $c(q, d) > 0$ such that

$$\begin{aligned} &\mathrm{Prob}[\mathcal{E}] \\ = \;& \mathrm{Prob}[\sup_{y\in K, \|y\|\geq\delta_0} |g_y| \geq c(q,d)\sqrt{\log p}\Delta^{\frac{q}{2-q}}\|y\|^{\frac{2-2q}{2-q}}] \\ \leq \;& \sum_{k=0}^{\log(\Delta/\delta_0)} \mathrm{Prob}[\sup_{y\in\overline{K}(2^k\delta_0)} |g_y| \geq c(q,d)\sqrt{\log p}\Delta^{\frac{q}{2-q}}\|y\|^{\frac{2-2q}{2-q}}] \\ \leq \;& \sum_{k=0}^{\log(\Delta/\delta_0)} p^{-d(\Delta/(2^k\delta_0))^{q/2}} \,. \end{aligned}$$

Now choosing $d = 4/q$ and setting $t_q = c(p, 4/q)$, we have that $\mathrm{Prob}[\mathcal{E}] = O(p^{-4/q})$. Let $z = \widehat{y} - y$. So for $\|z\| \geq \delta_0$, with probability $1 - O(p^{-4/q})$,

$$\|z\|^2 \leq 2|w \cdot z| \leq t_q \sqrt{\log p}\Delta^{\frac{q}{2-q}} \|z\|^{\frac{2-2q}{2-q}} .$$

That is

$$\|z\| = O(\Delta^{q/2}(\log p)^{1/2-q/4}) = O(\delta_0) .$$

Hence with probability $1 - O(p^{-4/q})$,

$$\|\widehat{y} - y\|^2 = O(\delta_0^2) = O(\Delta^q(\log p)^{1-q/2}) .$$

Since $\|\widehat{y} - y\| \leq 2\Delta \leq 2p^{1/q}$, we have that

$$\mathbb{E}[\|\widehat{y} - y\|^2] \leq \delta_0^2 + O(p^{-4/q} \cdot 2p^{2/q}) = O(\Delta^q(\log p)^{1-q/2}) .$$

For general $\sigma > 0$, we apply the standard scaling formula of $R_N(K, \sigma) = \sigma^2 R_N(K/\sigma, 1)$ and complete the proof of Proposition 9. The constant of $c_q = O(2^{\frac{1}{q}} \frac{1}{q} \ln \frac{2}{q})$ comes from multiplying c' and f_q in Lemma 15. □

6.3. *Proof of Theorem 11.* To establish the lower bound, we consider the largest Euclidean ball of various dimension contained in K. Intuitively, we show that if Kolmogorov width of K is large then it has to contain a large enough Euclidean ball, in terms of both radius and the dimension, which allows us to nearly match the upper bound. The crucial technical tool is the restricted invertibility result by Bourgain and Tzafriri [8] and developed by Szarek and Talagrand [36] and Giannopoulous [23].

Definition 17. For a set of vectors S, let $\mathrm{span}[S]$ denote the linear subspace spanned by S. A set $V = \{v_1, \ldots, v_s\}$ is called δ-wide if for any $1 \leq i \leq s$, $\mathrm{dist}(v_i, \mathrm{span}[V/\{v_i\}]) \geq \delta$, where $\mathrm{dist}(v, P)$ denotes the minimum distance between v and any vector in P.

The following proposition can be gleaned from work in [8, 36, 23]. See [26] (Proposition 5.2) for a proof.

Proposition 18. *For any δ-wide set $V = \{v_1, \ldots, v_s\}$, there exists $S \subseteq \{1, \ldots, s\}$ with $|S| \geq (1-\epsilon)s$ such that for any $\alpha = (\alpha_j)_{j\in S}$, $\|\sum_{j\in S} \alpha_j v_j\| \geq c\sqrt{\epsilon/s}\delta \sum_{j\in S} |\alpha_j|$, where c is an absolute constant.*

We make the following observation.

LEMMA 19. *Suppose that $K = X\ell_q^p$ and $X = (x_1, \ldots, x_p)$. Then for any $k > 0$, there exists $k+1$ vectors $V \subseteq \{x_1, \ldots, x_p\}$ such that V is $d_k(K)$ wide.*

PROOF. For a set of points $p_1, \ldots, p_s$ and $k \geq s-1$, let $\mathrm{vol}_k(p_1, \ldots, p_s)$ denote the k-volume of the convex hull of $p_1, \ldots, p_s$.

We find $k+1$ points $V = \{v_1, \ldots, v_{k+1}\}$ in K such that the $k+1$ volume of the simplex spanned by the origin O and $v_1, \ldots, v_{k+1}$ is the maximum, i.e.

$$V = \mathrm{argmax}_{y_1, \ldots, y_{k+1} \in K} \mathrm{vol}_{k+1}(O, y_1, \ldots, y_{k+1}) .$$

Since K is a compact set, $V \subseteq K$. We first show that V is $d_k(K)$ wide. Consider the k-dimensional subspace P spanned by $v_1, \ldots, v_k$. By the definition of d_k, we have $\sup_{y \in K} \|Py - y\| \geq d_k(K)$. Or equivalently

$$\sup_{y \in K} \mathrm{dist}(y, \mathrm{span}[\{v_1, \ldots, v_k\}] \geq d_k(K) . \tag{12}$$

On the other hand,

$$\begin{aligned} &\mathrm{vol}_{k+1}(O, v_1, \ldots, v_{k+1}) \\ &= \frac{1}{k+1} \mathrm{vol}_k(O, v_1, \ldots, v_k) \cdot \mathrm{dist}(v_{k+1}, \mathrm{span}[\{v_1, \ldots, v_k\}]) . \end{aligned} \tag{13}$$

By the maximality of $\mathrm{vol}_{k+1}(O, v_1, \ldots, v_{k+1})$ and (12) and (13), we have

$$\mathrm{dist}(v_{k+1}, \mathrm{span}(v_1, \ldots, v_k)) \geq d_k(K) .$$

Repeating this argument for each v_i in V, we have that V is $d_k(K)$-wide. In addition, for $K = X\ell_1$, K is the convex hull of $\pm x_1, \ldots, \pm x_p$. Hence for any projection P, $\mathrm{argmax}_{x \in K} \|Px\|$ has to be a vertex of K. That is $V \subseteq \{\pm x_i : 1 \leq i \leq p\}$. It is easy to see that V can be chosen such that $V \subseteq \{x_1, \ldots, x_p\}$. Since $X\ell_q \subseteq X\ell_1$ for $0 < q < 1$, $d_k(X\ell_q) \leq d_k(X\ell_1)$. This holds for any $0 < q \leq 1$. □

Using Proposition 18 and Lemma 19, we have that

LEMMA 20. *There exists constant $c > 0$ such that for any $K = X\ell_q^p$, $k > 0$, and $0 < \epsilon < 1$, there exists a linear sub-space P such that $P \cap K$ contains an $(1-\epsilon)k$ dimensional ℓ_2 ball with radius $\Omega(\sqrt{\epsilon(1-\epsilon)} k^{1/2-1/q} d_k(K))$.*

Proof. Clearly we can assume that $d_k(K) > 0$. Let V be the $d_k(K)$-wide set as in Lemma 19. Write $S_0 = \{i : \, x_i \in V\}$. By Proposition 18, let $S \subseteq S_0$ be such that $|S| \geq (1-\epsilon)|S_0|$ and for any $\{\alpha_j\}_{j \in S}$,

$$\| \sum_{i \in S} \alpha_i x_i \| \geq c\sqrt{\epsilon/|S_0|}\, d_k(K) \sum_{i \in S} |\alpha_i| \,.$$

According to reverse Hölder inequality, for $x \in \mathbb{R}^p$ and $0 < q \leq 1$, $\|x\|_1 \geq n^{1-1/q}\|x\|_q$. Hence, for any $\{\alpha_i\}$ such that $\sum_{i \in S} |\alpha_i|^q = 1$, $\sum_{i \in S} |\alpha_i| \geq |S|^{1-1/q}$. Thus if $\|\alpha\|_q = 1$, then

$$\begin{aligned} \| \sum_{i \in S} \alpha_i x_i \| &\geq c\sqrt{\epsilon/|S_0|}\, d_k(K) |S|^{1-1/q} \\ &\geq \;\; c\sqrt{\epsilon(1-\epsilon)}|S|^{1/2-1/q} d_k(K) \,. \qquad (14) \end{aligned}$$

Let P be the sub-space spanned by x_i for $i \in S$. Since $\{x_i\}_{i \in S_0}$ is $d_k(K) > 0$ wide, they are linearly independent. That is $K \cap P$ is fully ($|S|$) dimensional. On the other hand by (14) for any v on the boundary of $K \cap P$, we have that

$$\|v\| \geq c\sqrt{\epsilon(1-\epsilon)}|S|^{1/2-1/q} d_k(K) \,.$$

Hence, $K \cap P$ contains an $|S|$-dimensional ℓ_2 ball with radius

$$c\sqrt{\epsilon(1-\epsilon)}|S|^{1/2-1/q} d_k(K) \,.$$

The claim follows by $|S| \leq k$ and $1/2 - 1/q < 0$. □

By Lemma 4, $R^*(\ell_2^k(r), \sigma) = \Omega(\min(k\sigma^2, r^2))$. In addition, by definition of minimax risk, for any $K_1 \supseteq K_2$, $R^*(K_1, \sigma) \geq R^*(K_2, \sigma)$ (see for example [20]). Choosing $\epsilon = 1/2$, we have that for $K = X\ell_q^p$,

$$R^*(K, \sigma) = \Omega(\max_k \min(k\sigma^2, k^{1-2/q} d_k(K)^2)) \,.$$

6.4. *Proof of Theorem 1.*

Proof. Let

$$k^* = \operatorname{argmax}_k \min(d_k(K), k^{1/q}\sigma) \,.$$

When there is a tie, we pick k^* to be the smallest among the ties. Clearly $0 < k^* < n$ since $d_n(K) = 0$. When $k^* = 1$, it is easy to show the claim holds. For $1 < k^* < n$, we distinguish two cases.

Case 1. $d_{k^*}(K) \geq (k^*)^{1/q}\sigma$.

In this case we have that $d_{k^*+1}(K) \leq (k^*+1)^{1/q}\sigma$. Otherwise, we would have that

$$\begin{aligned}&\min(d_{k^*+1}(K), (k^*+1)^{1/q}\sigma)\\ &= (k^*+1)^{1/q}\sigma > (k^*)^{1/q}\sigma\\ &\geq d_{k^*}(K) \geq \min(d_{k^*}(K), (k^*)^{1/q}\sigma)\,.\end{aligned}$$

This contradicts with the maximality of k^*. Since $d_{k^*}(K) \geq (k^*)^{1/q}\sigma$, $k^{1-2/q}d_{k^*}(K)^2 \geq k^*\sigma^2$. We apply the lower bound in 5) and obtain that

$$R^*(K,\sigma) = \Omega(k^*\sigma^2)\,.$$

For the upper bound, by taking $k = k^*+1$ in (4), we have

$$\begin{aligned}&R_H(K,\sigma)\\ &= O((k^*+1)\sigma^2 + c_q d_{k^*+1}(K)^q\sigma^{2-q}(\log p)^{1-q/2})\\ &= O((k^*+1)\sigma^2 + c_q((k^*+1)^{1/q}\sigma)^q\sigma^{2-q}(\log p)^{1-q/2})\\ &= O((k^*+1)\sigma^2(\log p)^{1-q/2})\\ &= O(R^*(K,\sigma)(\log p)^{1-q/2})\,.\end{aligned}$$

Case 2. $d_{k^*}(K) < (k^*)^{1/q}\sigma$.

In this case $d_{k^*}(K) \geq (k^*-1)^{1/q}\sigma$. Otherwise, we would have that $d_{k^*}(K) < (k^*-1)^{1/q}\sigma$ and $d_{k^*}(K) < d_{k^*-1}(K)$. The latter is due to that we pick k^* the smallest k in case there is a tie. This would imply that

$$\begin{aligned}&\min(d_{k^*-1}(K), (k^*-1)^{1/q}\sigma)\\ &> d_{k^*}(K) \geq \min(d_{k^*}(K), (k^*)^{1/q}\sigma)\,.\end{aligned}$$

Again it contradicts with the maximality of k^*. Hence for the lower bound, we have that

$$\begin{aligned}R^*(K,\sigma) &= \Omega((k^*)^{1-2/q}d_{k^*}(K)^2)\\ &= \Omega((k^*)^{1-2/q}(k^*-1)^{2/q}\sigma^2)\\ &= \Omega((k^*)^2\sigma^2)\,, \quad \text{by } k^* > 1.\end{aligned}$$

Setting $k = k^*$ in (4), we have

$$\begin{aligned}&R_H(K,\sigma)\\ &= O(k^*\sigma^2 + c_q d_{k^*}(K)^q\sigma^{2-q}(\log p)^{1-q/2})\\ &= O(k^*\sigma^2 + c_q((k^*)^{1/q}\sigma)^q\sigma^{2-q}(\log p)^{1-q/2})\\ &= O(k^*\sigma^2(\log p)^{1-q/2})\\ &= O(R^*(K,\sigma)(\log p)^{1-q/2})\,.\end{aligned}$$

Therefore, for any $0 < q \leq 1$ and $p = \Omega(n/\log n)$, for $K = X\ell_q^p$ where X is an $n \times p$ matrix, we have that $R_H(K,\sigma) = O((\log p)^{1-q/2} R^*(K,\sigma))$. □

6.5. *Proof of Lemma 13.*

PROOF. In what follows, all the statements hold with high probability, say $1 - 1/n^2$.

1. Since $\widetilde{y}_k - y_k$ is $n-k$ dimensional Gaussian vector, by the property of χ^2-distribution,
$$\|\widetilde{y}_k - y_k\|^2 \leq (n-k)\sigma^2 + 2\sqrt{n\log n}\sigma^2 .$$
Since $\|\widetilde{y}_k - \widehat{y}_k\| \leq \|\widetilde{y}_k - y_k\|$, the statement follows immediately.
2. Let z denote the nearest neighbor of y_k on K_k. So $\|z - y_k\| = \delta_k$. Further,
$$(\widehat{y}_k - z)\cdot(y_k - z) \leq 0 . \tag{15}$$
Following the same analysis for the nearest neighbor estimator, we have
$$\begin{aligned}&\|\widehat{y}_k - z\|^2\\ \leq\ & 2(\widehat{y}_k - z)\cdot(\widetilde{y}_k - z)\\ =\ & 2(\widehat{y}_k - z)\cdot(\widetilde{y}_k - y_k) + 2(\widehat{y}_k - z)\cdot(y_k - z)\\ \leq\ & 2(\widehat{y}_k - z)\cdot(\widetilde{y}_k - y_k) \quad \text{by (15)}\\ \leq\ & c_1 C_k d_k \sigma\sqrt{\log p}\\ =\ & 4c_1 k\sigma^2\sqrt{\log p} .\end{aligned}$$
Hence
$$\begin{aligned}&\|\widehat{y}_k - \widetilde{y}_k\|^2\\ \geq\ & \|\widehat{y}_k - z\|^2 + \|z - \widetilde{y}_k\|^2 + 2(\widehat{y}_k - z)\cdot(z - \widetilde{y}_k)\\ \geq\ & \|\widetilde{y}_k - z\|^2 + 2(\widehat{y}_k - z)\cdot(z - y_k) + 2(\widehat{y}_k - z)\cdot(y_k - \widetilde{y}_k)\\ \geq\ & \|\widetilde{y}_k - z\|^2 + 2(\widehat{y}_k - z)\cdot(y_k - \widetilde{y}_k) \quad \text{by (15)}\\ \geq\ & \|\widetilde{y}_k - z\|^2 - 2|(\widehat{y}_k - z)\cdot(y_k - \widetilde{y}_k)| .\end{aligned}$$
We bound these two terms separately.
$$\begin{aligned}&\|\widetilde{y}_k - z\|^2\\ =\ & \|\widetilde{y}_k - y_k\|^2 + \|y_k - z\|^2 + 2(\widetilde{y}_k - y_k)\cdot(y_k - z)\\ \geq\ & (n-k)\sigma^2 - 2\sqrt{n\log n}\sigma^2 + \delta_k^2 - 4\delta_k\sigma\sqrt{\log p} .\end{aligned}$$

By the analysis for the nearest neighbor estimator, we have

$$2|(\widehat{y}_k - z) \cdot (y_k - \widetilde{y}_k)| \leq c_1 C_k d_k \sigma \sqrt{\log p} = c_1 k \sigma^2 \sqrt{\log p}\,.$$

Putting them together, we can take $\delta_k^2 = c_2(\sqrt{n \log n}\sigma^2 + k\sigma^2\sqrt{\log p})$ for some sufficiently large c_2 and obtain

$$\|\widehat{y}_k - \widetilde{y}_k\|^2 \geq (n-k)\sigma^2 + 2\sqrt{n \log n}\sigma^2\,.$$

3. If $\delta_k^2 \leq c_1(\sqrt{n \log n}\sigma^2 + k\sigma^2 \log p)$, then according to the above

$$\|\widehat{y}_k - z\|^2 \leq O(k\sigma^2 \log p) = O(\delta_k^2)\,.$$

Hence

$$\|\widehat{y}_k - y_k\| \leq \|\widehat{y}_k - z\| + \|z - y_k\| = O(\delta_k)\,.$$

□

Acknowledgment. The author would like to thank Adel Javanmard and Tong Zhang for useful discussions and anonymous reviewers and the associate editor for many useful suggestions, especially for suggesting the extension to the adaptive estimation and the references of [28, 6].

REFERENCES

[1] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117:467–493, 2000.

[2] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[3] A. R. Barron, A. Choen, W. Dahmen, and R. A. Devore. Approximation and learning by greedy algorithms. *Then Annals of Statistics*, 26(1):64–94, 2008.

[4] M. Bayati and A. Montanari. The LASSO risk for Gaussian matrices. *CoRR*, abs/1008.2581, 2010.

[5] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[6] L. Birgé. *An alternative point of view on Lepski's method*, volume 36, pages 113–133. Institute of Mathematical Statistics, 2001.

[7] L. Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10:1039–1051, 2004.

[8] J. Bourgain and L. Tzafriri. Invertibility of 'large' submatrices with applications to the geometry of banach spaces and harmonic analysis. *Israel Journal of Mathematics*, 57(2):137–224, 1987.

[9] A. Brieden. Geometric optimizatoin problems likely not contained in APX. *Discrete and Computational Geometry*, 28:201–209, 2002.

[10] T. T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.

[11] T. T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58(3):1300–1308, 2010.

[12] E. Candés and M. A. Davenport. How well can we estimate a sparse vector? http://arxiv.org/abs/1104.5246, 2011.

[13] E. Candés and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[14] E. Candés and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[15] B. Carl and A. Pajor. Gelfand numbers of operators with values in a Hilbert space. *Invent. Math.*, 94:479–504, 1988.

[16] D. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. `http://www-stat.stanford.edu/~donoho/Reports/2004/l1l0approx.pdf`, 2004.

[17] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):489–509, 2006.

[18] D. Donoho, I. Johnstone, A. Maleki, and A. Montanari. Compressed sensing over ℓ_p-balls: Minimax mean square error. *CoRR*, abs/1103.1943, 2011.

[19] D. Donoho and I. M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

[20] D. Donoho, R. Liu, and B. MacGibbon. Minimax risk over hyperrectangles, and implications. *The Annals of Staistics*, 18(3):1416–1437, 1990.

[21] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[22] J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

[23] A. Giannopoulous. A note on the Banach-Mazur distance to the cube. *Operator Theory*, 77:67–73, 1995.

[24] O. Guedon and A. E. Litvak. Euclidean projections of p-convex body. In *Geometric Aspects of Functional Analysis*, pages 971–988. Springer-Verlag, 2000.

[25] I. A. Ibragimov and R. Z. Hasminskii. *Statistical Estimation: Asymptotic Theory.* Springer, 1981.

[26] A. Javanmard and L. Zhang. The minimax risk of truncated series estimators for symetric convex bodies. http://arxiv.org/abs/1201.2462, 2012.

[27] I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. available at http://www-stat.stanford.edu/people/faculty/johnstone, 2011.

[28] O. V. Lepskii. Asymptotically minimax adaptive estimation i: upper bounds. optimally adaptive estimates. *Theory Probability and Its Applications*, 36(4):682–697, 1991.

[29] J.-M. Loubes and S. van de Geer. Adaptive estimation in regression, using soft thresholding type penalties. *Statistica Neerlandica*, 56:453–478, 2002.

[30] A. Nemirovski. *Topics in Non-parametric Statistics.* Lecture Notes in Mathematics. Springer, 1998.

[31] A. Pinkus. *n-Widths in Approximation Theory.* Springer-Verlag, 1984.

[32] M. S. Pinsker. Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Information Transmission*, 16:120–133, 1980.

[33] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge University Press, 1989.

[34] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

[35] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, 2004.

[36] S. J. Szarek and M. Talagrand. An isomorphic version of the Sauer-Shelah lemma and the Banach-Mazur distance to the cube. In *GAFA Seminar 87-88, Lecture Notes in Mathematics*, pages 105–112. Springer-Verlag, 1989.

[37] M. Talagrand. *The Generic Chaining.* Springer Monographs in Mathematics. Springer-Verlag, 2005.

[38] R. Tibshirani. Regression shrinkage and selection vias the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[39] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, 2009.

[40] S. van de Geer. *Emprical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambrdige University Press, 2000.

[41] K. R. Varadarajan, S. Venkatesh, Y. Ye, and J. Zhang. Approximating the radii of point sets. *SIAM J. Comput.*, 36(6):1764–1776, 2007.

[42] F. Ye and C.-H. Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Maching Learning Research*, 11:3519–3540, 2010.

[43] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[44] T. Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011.

[45] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

APPENDIX A: THE ENTROPY NUMBER OF $X\ell_q$

By Guedon and Litvak [24] (Theorem 6)

$$e_k(\ell_q^p, \ell_1^p) = \begin{cases} \Theta(1) & k \le \log p \\ \Theta\left(\left(f_q \frac{\log(1+p/k)}{k}\right)^{1/q-1}\right) & \log p \le k \le p \\ \Theta\left(2^{-k/p}(f_q/p)^{1/q-1}\right) & k \ge p\,. \end{cases} \tag{16}$$

where $f_q = O(\frac{1}{q}\ln\frac{2}{q})$ is a constant dependent on q only.

and by Carl and Pajor [15],

$$e_k(X\ell_1^p, \ell_2^n) = \begin{cases} O(\Delta) & k \le \log p \\ O\left(\left(\frac{\log(1+p/k)}{k}\right)^{1/2}\Delta\right) & \log p \le k \le p \\ O\left(2^{-k/p}(1/p)^{1/2}\Delta\right) & k \ge p\,. \end{cases} \tag{17}$$

From the definition of e_k, we have (see also [33])

$$e_{k_1+k_2}(K_1, K_3) \leq e_{k_1}(K_1, K_2) e_{k_2}(K_2, K_3)\,. \tag{18}$$

By (18), $e_{2k}(X\ell_q^p, \ell_2^n) \leq e_k(\ell_q^p, \ell_1^p) e_k(X\ell_1^p, \ell_2^n)$. So we have

$$e_{2k}(X\ell_q^p, \ell_2^n) = \begin{cases} O(\Delta) & k \leq \log p \\ O\left(\left(f_q \frac{\log(1+p/k)}{k}\right)^{1/q-1/2} \Delta\right) & \log p \leq k \leq p \\ O\left(2^{-2k/p}(f_q/p)^{1/q-1/2}\Delta\right) & k \geq p\,. \end{cases} \tag{19}$$

Microsoft Research
1065 La Avenida
Mountain View, CA 94043
USA
E-mail: lzha@microsoft.com