Journal of Machine Learning Research 15 (2014) 1675-1711 Submitted 2/13; Revised 4/14; Published 5/14

Adaptive Minimax Regression Estimation over Sparse ℓ_q-Hulls

Zhan Wang ZWANG@FULCRM.COM
Fulcrum Analytics
Fairfield, CT 06824, USA

Sandra Paterlini SANDRA.PATERLINI@EBS.EDU
Department of Finance & Accounting
EBS Universität für Wirtschaft und Recht
Gustav-Stresemann-Ring 3
65189 Wiesbaden, Germany

Fuchang Gao FUCHANG@UIDAHO.EDU
Department of Mathematics
University of Idaho
Moscow, ID 83844, USA

Yuhong Yang YYANG@STAT.UMN.EDU
School of Statistics
University of Minnesota
313 Ford Hall
224 Church Street
Minneapolis, MN 55455, USA

Editor: Bin Yu

Abstract

Given a dictionary of M_n predictors, in a random design regression setting with n observations, we construct estimators that target the best performance among all the linear combinations of the predictors under a sparse ℓ_q-norm ($0 \leq q \leq 1$) constraint on the linear coefficients. Besides identifying the optimal rates of convergence, our universal aggregation strategies by model mixing achieve the optimal rates simultaneously over the full range of $0 \leq q \leq 1$ for any M_n and without knowledge of the ℓ_q-norm of the best linear coefficients to represent the regression function.

To allow model misspecification, our upper bound results are obtained in a framework of aggregation of estimates. A striking feature is that no specific relationship among the predictors is needed to achieve the upper rates of convergence (hence permitting basically arbitrary correlations between the predictors). Therefore, whatever the true regression function (assumed to be uniformly bounded), our estimators automatically exploit any sparse representation of the regression function (if any), to the best extent possible within the ℓ_q-constrained linear combinations for any $0 \leq q \leq 1$.

A sparse approximation result in the ℓ_q-hulls turns out to be crucial to adaptively achieve minimax rate optimal aggregation. It precisely characterizes the number of terms needed to achieve a prescribed accuracy of approximation to the best linear combination in an ℓ_q-hull for $0 \leq q \leq 1$. It offers the insight that the minimax rate of ℓ_q-aggregation is basically determined by an effective model size, which is a sparsity index that depends on

©2014 Zhan Wang, Sandra Paterlini, Fuchang Gao and Yuhong Yang.

q, M_n, n, and the ℓ_q-norm bound in an easily interpretable way based on a classical model selection theory that deals with a large number of models.

Keywords: high-dimensional sparse learning, minimax rate of convergence, model selection, optimal aggregation, sparse ℓ_q-constraint

1. Introduction

Learning a high-dimensional function has become a central research topic in machine learning. In this paper, we intend to provide a theoretical understanding on how well one can adaptively estimate a regression function by sparse linear combinations of a number of predictors based on i.i.d observations.

1.1 Motivation

Sparse modeling has become a popular area of research to handle high-dimensional linear regression learning. One notable approach is to exploit the assumption that the "true" linear coefficients have a bounded ℓ_q-norm (or simply q-norm) for some $0 \leq q \leq 1$, which implies that the parameter space is necessarily sparse in a proper sense. A major recent theoretical advancement is made by Raskutti, Wainwright, and Yu (2012), who derive minimax rates of convergence both for estimating the regression function and for estimating the parameter vector, which spells out how the sparsity parameter q (in conjunction with the number of predictors, say, M_n) affects the intrinsic capability of estimation for the ℓ_q-hulls. The results confirm that even if M_n is much larger than the sample size n, relatively fast rates of convergence in learning the linear regression function are possible. In a Gaussian sequence model framework, Donoho and Johnstone (1994) identify precisely how the ℓ_q-constraint ($q > 0$) on the mean vector affects estimation accuracy under the ℓ_p-loss ($p \geq 1$).

In this paper, differently from the fixed design setting in Raskutti et al. (2012); Negahban et al. (2012), under a random design, we examine the issue of minimax optimal estimation of a linear regression function in the ℓ_q-hulls for $0 \leq q \leq 1$. Besides confirming the same role of q on determining the minimax rate of convergence for estimation of the regression function also for the random design, we prove that the minimax rate can be adaptively achieved without *any* knowledge of q or the ℓ_q-radius of the linear coefficients. The adaptation results show that in high-dimensional linear regression learning, theoretically speaking, should the regression function happen to depend on just a few predictors (i.e., hard sparsity or $q = 0$) or only a small number of coefficients really matter (i.e., soft sparsity or $0 < q \leq 1$), the true sparsity nature is automatically exploited, leading to whatever the optimal rate of convergence for the situation. No restriction is imposed on M_n. To our knowledge, this is the most general result on minimax learning in the ℓ_q-hulls for $0 \leq q \leq 1$.

In reality, obviously, the soft or hard sparsity is only an approximation that hopefully captures the nature of the target function. To deal with possible model misspecification (i.e., the sparsity assumption may or may not be suitable for the data), our upper bound results on regression estimation will be given in a framework that permits the regression function to be outside of the ℓ_q-hulls. The risk bounds show that whichever soft or hard sparse representation of the true regression function by linear combination of the predictors

best describes the truth, the corresponding optimal performance (in rate) is automatically achieved (with some additional conditions for deriving matching minimax lower bounds).

Our aim of simultaneous adaptive estimation over the ℓ_q-hulls for all $0 \leq q \leq 1$ and positive ℓ_q-radius, especially with possible model misspecification, requires a deeper understanding of sparse approximation of functions in the ℓ_q-hulls than what is available in the literature. As a solution, we provide a sharp sparse approximation error bound for ℓ_q-hulls with $0 \leq q \leq 1$, which may also be relevant for studying other linear representation based high-dimensional sparse learning methods.

The aforementioned flexible approach to optimal estimation that allows model misspecification is done in the framework of aggregation of estimates. Besides the aspect of not assuming the true target function to have any specific relationship to the predictors/the initial estimates to be aggregated, the theory of aggregation emphasizes that the predictors/the initial estimates are basically arbitrary in their association with each other. With this characteristic in sight, the minimax rate of aggregation is properly determined by finding a specific set of initial estimates with known relationship (e.g., independence) under which an existing upper bound can be shown to be un-improvable up to a constant factor. In contrast, for minimax optimal regression, one works with whatever (hopefully weak) assumptions imposed on the predictors and tries to achieve the minimax rate of convergence for the function class of interest. With the above, the problem of aggregation of estimates is closely related to the usual regression estimation: A risk upper bound on aggregation of estimates readily gives a risk upper bound for regression estimation, but one has to derive minimax lower bounds for the specific regression learning settings. In this work, we will first give results on aggregation of estimates (where most work is on the upper bounds under minimal assumptions on the initial estimates) and then present results on minimax regression in ℓ_q-hulls (where most work is on deriving lower rates of convergence). The focus is on random design and additions results on fixed design are in Wang et al. (2011).

1.2 Aggregation of Estimates

The idea of sharing strengths of different learning procedures by combining them instead of choosing a single one has led to fruitful and exciting research results in statistics and machine learning. The theoretical advances have centered on optimal risk or loss bounds that require almost no assumption on the behaviors of the individual estimators to be aggregated. See, e.g., Yang (1996, 2000a); Catoni (1997, 2004); Juditsky and Nemirovski (2000); Nemirovski (2000); Yang (2004); Tsybakov (2003); Leung and Barron (2006) for early representative work (the reader is referred to Cesa-Bianchi and Lugosi 2006 for interesting results and references from an individual sequence perspective). While there are many different ways that one can envision to combine the advantages of the candidate procedures, the combining methods can be put into two main categories: those intended for *combining for adaptation*, which aim at combining the procedures to perform adaptively as well as the best candidate procedure no matter what the truth is, and those for *combining for improvement*, which aim at improving over the performance of all the candidate procedures in certain ways. Whatever the goal is, for the purpose of estimating the regression function, we expect to pay a price: the risk of the combined procedure is typically larger than the target risk. The

difference between the two risks (or a proper upper bound on the difference) is henceforth called *risk regret* of the combining method.

The research attention is often focused on one but the main step in the process of combining procedures, namely, *aggregation of estimates*, wherein one has already obtained estimates by all the candidate procedures (based on initial data, most likely from data splitting or previous studies; some exceptions are in e.g., Leung and Barron 2006; Dalalyan and Salmon 2012), and is trying to aggregate these estimates into a single one based on data that are independent of the initial data. The performance of the aggregated estimator (conditional on the initial estimates) plays the most important role in determining the total risk of the whole combined procedure, although proportion of the initial data and the later one certainly also influences the overall performance. In this work, we will mainly focus on the aggregation step.

It is now well-understood that given a collection of procedures, one only needs to pay a relatively small price for aggregation for adaptation (Yang 2000b; Catoni 2004; Tsybakov 2003). In contrast, aggregation for improvement under a convex constraint or ℓ_1-constraint on coefficients is associated with a higher risk regret (as shown in Juditsky and Nemirovski 2000; Nemirovski 2000; Yang 2004; Tsybakov 2003). Several other directions of aggregation for improvement, defined via proper constraints imposed on the ℓ_0-norm alone or in conjunction with the ℓ_1-norm of the linear coefficients, have also been studied, including linear aggregation (no constraint, Tsybakov 2003), aggregation to achieve the best performance of a linear combination of no more than a given number of initial estimates (Bunea et al. 2007) and also under an additional constraint on the ℓ_1-norm of these coefficients (Lounici 2007). Interestingly, combining for adaptation plays a fundamental role in combining for improvement: it serves as an effective tool in constructing multi-directional (or universal) aggregation methods, that is methods which simultaneously achieve the best performance in multiple specific directions of aggregation for improvement. This strategy was taken in, e.g., Yang (2004), Tsybakov (2003), Bunea et al. (2007), Rigollet and Tsybakov (2010), and Dalalyan and Tsybakov (2012b).

The goal of our work on aggregation is to propose aggregation methods that achieve the performance (in risk with/without a multiplying factor), up to a multiple of the optimal risk regret as defined in Tsybakov (2003), of the best linear combination of the initial estimates under the constraint that the ℓ_q-norm ($0 \le q \le 1$) of the linear coefficients is no larger than some positive number t_n (henceforth the *ℓ_q-constraint*). We call this type of aggregation *ℓ_q-aggregation*. It turns out that the optimal rate is simply determined by an *effective model size* m_*, which roughly means that only m_* terms are really needed for effective estimation. We strive to achieve the optimal ℓ_q-aggregation simultaneously for all q ($0 \le q \le 1$) and t_n ($t_n > 0$).

It is useful to note that the ℓ_q-aggregation provides a general framework: our proposed strategies enable one to reach the optimal bounds automatically and simultaneously for the major state-of-art aggregation strategies and more, as will be seen.

1.3 Plan of the Paper

The paper is organized as follows. In Section 2, we introduce notation and some preliminaries of the estimators and aggregation algorithms that will be used in our strategies for

learning. In Section 3, we derive optimal rates of ℓ_q-aggregation and show that our methods achieve multi-directional aggregation. In Section 4, we derive the minimax rate for linear regression with ℓ_q-constrained coefficients. A discussion is then reported in Section 5. Finally, we report in Appendix A the derivation of metric entropy and approximation error bounds for $\ell_{q,t_n}^{M_n}$-hulls, while Appendix B provides an insight from the sparse approximation bound based on classical model selection theory. The proofs of the results in Sections 3 and 4 are then provided in the Appendix C.

2. Preliminaries

Consider the regression problem where a dictionary of M_n prediction functions ($M_n \geq 2$ unless stated otherwise) are given as initial estimates of the unknown true regression function. The goal is to construct a linearly combined estimator using these estimates to pursue the performance of the best (possibly constrained) linear combinations. A learning strategy with two building blocks will be considered. First, we construct candidate estimators from subsets of the given estimates. Second, we aggregate the candidate estimators using aggregation algorithms to obtain the final estimator.

2.1 Notation and Definition

Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be n ($n \geq 2$) i.i.d. observations where $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})$, $1 \leq i \leq n$, take values in $\mathcal{X}$ with a probability distribution P_X. We assume the regression model

$$Y_i = f_0(\mathbf{X}_i) + \varepsilon_i, \qquad i = 1, \ldots n, \tag{1}$$

where f_0 is the unknown true regression function to be estimated. The random errors ε_i, $1 \leq i \leq n$, are independent of each other and of $\mathbf{X}_i$, and have the probability density function $h(x)$ (with respect to the Lebesgue measure or a general measure μ) such that $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = \sigma^2 < \infty$. The quality of estimating f_0 by using the estimator $\hat{f}$ is measured by the squared L_2 risk (with respect to P_X)

$$R(\hat{f}; f_0; n) = E\|\hat{f} - f_0\|^2 = E\left(\int (\hat{f} - f_0)^2 dP_X\right),$$

where, as in the rest of the paper, $\|\cdot\|$ denotes the L_2-norm with respect to the distribution of P_X.

Let $F_n = \{f_1, f_2, \ldots, f_{M_n}\}$ be a dictionary of M_n initial estimates of f_0. In this paper, unless stated otherwise, $\|f_j\| \leq 1$, $1 \leq j \leq M_n$. The condition is satisfied, possibly after a scaling, if the f_j's are uniformly bounded between known constants, and it may require additional assumptions on the distribution of P_X to check its validity for a general case. Consider the constrained linear combinations of the estimates $\mathcal{F} = \left\{f_\theta = \sum_{j=1}^{M_n} \theta_j f_j : \theta \in \Theta_n, f_j \in F_n\right\}$, where Θ_n is a subset of $\mathbb{R}^{M_n}$. Let

$$d^2(f_0; \mathcal{F}) = \inf_{f_\theta \in \mathcal{F}} \|f_\theta - f_0\|^2$$

denote the smallest approximation error to f_0 over a function class $\mathcal{F}$.

The problem of constructing an estimator $\hat{f}$ that pursues the best performance in $\mathcal{F}$ is called *aggregation of estimates*. We consider aggregation of estimates with sparsity constraints on θ. For any $\theta = (\theta_1, \ldots, \theta_{M_n})'$, define the ℓ_0-norm and the ℓ_q-norm $(0 < q \leq 1)$ by

$$\|\theta\|_0 = \sum_{j=1}^{M_n} I(\theta_j \neq 0), \text{ and } \|\theta\|_q = \left(\sum_{j=1}^{M_n} |\theta_j|^q \right)^{1/q},$$

where $I(\cdot)$ is the indicator function. Note that for $0 < q < 1$, $\|\cdot\|_q$ is not a norm but a quasinorm, and for $q = 0$, $\|\cdot\|_0$ is not even a quasinorm. However, we choose to refer them as norms for ease of exposition. For any $0 \leq q \leq 1$ and $t_n > 0$, define the ℓ_q-ball

$$B_q(t_n; M_n) = \left\{ \theta = (\theta_1, \theta_2, \ldots, \theta_{M_n})' : \|\theta\|_q \leq t_n \right\}.$$

When $q = 0$, t_n is understood to be an integer between 1 and M_n, and sometimes denoted by k_n to be distinguished from t_n when $q > 0$. Define the $\ell_{q,t_n}^{M_n}$*-hull* of F_n to be the class of linear combinations of functions in F_n with the ℓ_q-constraint

$$\mathcal{F}_q(t_n) = \mathcal{F}_q(t_n; M_n; F_n) = \left\{ f_\theta = \sum_{j=1}^{M_n} \theta_j f_j : \theta \in B_q(t_n; M_n), f_j \in F_n \right\}, 0 \leq q \leq 1, t_n > 0.$$

One of our goals is to propose an estimator $\hat{f}_{F_n} = \sum_{j=1}^{M_n} \hat{\theta}_j f_j$ such that its risk is upper bounded by a multiple of the smallest risk over the class $\mathcal{F}_q(t_n)$ plus a small risk regret term

$$R(\hat{f}_{F_n}; f_0; n) \leq C \inf_{f_\theta \in \mathcal{F}_q(t_n)} \|f_\theta - f_0\|^2 + REG_q(t_n; M_n),$$

where C is a constant that does not depend on f_0, n, and M_n, or $C = 1$ for some estimators. In various situations (e.g., adaptive estimation with data splitting as in Yang 2000a), the initial estimates can be made such that $\inf_{f_\theta \in \mathcal{F}_q(t_n)} \|f_\theta - f_0\|^2$ approaches zero as $n \to \infty$ in a proper manner. Thus, results with $C > 1$ (especially under heavy tailed random errors) are also of interest.

2.2 Two Starting Estimators

A key step of our strategy is the construction of candidate estimators using subsets of the initial estimates. The T- and AC-estimators, described below, were chosen because of the relatively mild assumptions for them to work with respect to the squared L_2 risk (each of them gives cleaner results in different aspects). Under the data generating model (1) and i.i.d. observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, suppose we are given m terms $\{g_1, \ldots, g_m\}$ (i.e., m functions of the original explanatory variables) as regressors.

When working on the minimax upper bounds in random design settings, we always make the following assumption on the true regression function.

Assumption BD: There exists a known constant $L > 0$ such that $\|f_0\|_\infty \leq L < \infty$.

To our knowledge, Assumption BD is typically assumed in the literature of aggregation of estimates. A recent work of Birgé (2014) successfully removes this limitation in a density estimation framework.

(T-estimator) Birgé (2006, 2004) constructed the T-estimator and derived its L_2 risk bounds under the Gaussian regression setting. The following proposition is a simple consequence of Theorem 3 of Birgé (2004). Suppose
T1. The error distribution $h(\cdot)$ is normal;
T2. $0 < \sigma < \infty$ is known.

Proposition 1 *Suppose Assumptions* BD *and* T1, T2 *hold. We can construct a T-estimator* $\hat{f}^{(T)}$ *such that*

$$E\|\hat{f}^{(T)} - f_0\|^2 \le C_{L,\sigma}\left(\inf_{\vartheta\in\mathbb{R}^m}\left\|\sum_{j=1}^m \vartheta_j g_j - f_0\right\|^2 + \frac{m}{n}\right),$$

where $C_{L,\sigma}$ *is a constant depending only on* L *and* σ*.*

(AC-estimator) For our purpose, consider the class of linear combinations with the ℓ_1-constraint $\mathcal{G}_s = \{g = \sum_{j=1}^m \vartheta_j g_j : \|\vartheta\|_1 \le s\}$ for some $s > 0$. Audibert and Catoni proposed a sophisticated AC-estimator $\hat{f}_s^{(AC)}$ (Audibert and Catoni 2010, page 25). The following proposition is a direct result from Theorem 4.1 in Audibert and Catoni (2010) under the following conditions.
AC1. There exists a constant $H > 0$ such that $\sup_{g,g'\in\mathcal{G}_s,\mathbf{x}\in\mathcal{X}} |g(\mathbf{x}) - g'(\mathbf{x})| = H < \infty$.
AC2. There exists a constant $\sigma' > 0$ such that $\sup_{\mathbf{x}\in\mathcal{X}} E\left((Y - g_s^*(\mathbf{X}))^2|\mathbf{X} = \mathbf{x}\right) \le (\sigma')^2 < \infty$, where $g_s^* = \inf_{g\in\mathcal{G}_s}\|g - f_0\|^2$.

Proposition 2 *Suppose Assumptions* AC1 *and* AC2 *hold. For any* $s > 0$*, we can construct an AC-estimator* $\hat{f}_s^{(AC)}$ *(that may depend on* H *and* σ'*) such that*

$$E\|\hat{f}_s^{(AC)} - f_0\|^2 \le \inf_{g\in\mathcal{G}_s}\|g - f_0\|^2 + c\left(2\sigma' + H\right)^2\frac{m}{n},$$

where c *is a pure constant.*

The risk bound for the AC-estimator improves over that for the T-estimator in terms of i) reducing the multiplying constant in front of the optimal approximation error to the best possible; ii) relaxing the normality assumption on the errors. But this is achieved at the expense of restricting the ℓ_1-norm of the linear coefficients in approximation. Note also that under the assumption $\|f_0\|_\infty \le L$, we can always enforce the estimators $\hat{f}^{(T)}$ and $\hat{f}_s^{(AC)}$ to be in the range of $[-L, L]$ with the same risk bounds in the propositions.

2.3 Two Aggregation Algorithms for Adaptation

Suppose N estimates $\check{f}_1, \ldots, \check{f}_N$ are obtained from N candidate procedures based on some initial data. Two aggregation algorithms, the ARM algorithm (Adaptive Regression by Mixing, Yang 2001) and Catoni's algorithm (Catoni 2004, 1999), can be used to construct the final estimator $\hat{f}$ by aggregating the candidate estimates $\check{f}_1, \ldots, \check{f}_N$ based on n additional i.i.d. observations $(\mathbf{X}_i, Y_i)_{i=1}^n$. The ARM algorithm requires knowing the form of the error distribution but it allows heavy tail cases. In contrast, Catoni's algorithm does not assume any functional form of the error distribution, but demands exponential decay of the tail probability.

(The ARM algorithm) Suppose
Y1. There exist two known constants $\underline{\sigma}$ and $\overline{\sigma}$ such that $0 < \underline{\sigma} \le \sigma \le \overline{\sigma} < \infty$;
Y2. The error density has the form $h(x) = h_0(x/\sigma)/\sigma$, where h_0 is known and has mean zero, variance 1, and a finite fourth moment. In addition, for each pair of constants $R_0 > 0$ and $0 < S_0 < 1$, there exists a constant B_{S_0,R_0} (depending on S_0 and R_0) such that for all $|R| < R_0$ and $S_0 \le S \le S_0^{-1}$,

$$\int h_0(x) \log \frac{h_0(x)}{S^{-1} h_0((x-R)/S)} dx \le B_{S_0,R_0}((1-S)^2 + R^2).$$

The condition Y2 can be shown to hold for Gaussian, Laplace and Student's t (with at least 3 degrees of freedom) distributions. We can construct an estimator $\hat{f}^Y$ which aggregates $\check{f}_1, \ldots, \check{f}_N$ by the ARM algorithm with prior probabilities π_k ($\sum_{k=1}^N \pi_k = 1$) on the procedures.

Proposition 3 *(Yang 2004, Proposition 1) Suppose Assumptions* BD *and* Y1, Y2 *hold, and* $\|\check{f}_k\|_\infty \le L < \infty$ *with probability* 1, $1 \le k \le N$. *The estimator* $\hat{f}^Y$ *by the ARM algorithm has the risk*

$$R(\hat{f}^Y; f_0; n) \le C_Y \inf_{1 \le k \le N} \left(\|\check{f}_k - f_0\|^2 + \frac{\sigma^2}{n} \left(1 + \log \frac{1}{\pi_k}\right)\right),$$

where C_Y *is a constant that depends on* $\underline{\sigma}, \overline{\sigma}, L$, *and also* h *(through the fourth moment of the random error and* B_{S_0,R_0} *with* $S_0 = \underline{\sigma}/\overline{\sigma}, R_0 = L$*).*

(Catoni's algorithm) Suppose for some positive constant $\alpha < \infty$, there exist known constants $U_\alpha, V_\alpha < \infty$ such that
C1. $E(\exp(\alpha|\varepsilon_i|)) \le U_\alpha$;
C2. $\frac{E(\varepsilon_i^2 \exp(\alpha|\varepsilon_i|))}{E(\exp(\alpha|\varepsilon_i|))} \le V_\alpha$.
Let $\lambda_C = \min\{\frac{\alpha}{2L}, (U_\alpha(17L^2 + 3.4V_\alpha))^{-1}\}$ and π_k be the prior for $\check{f}_k$, $1 \le k \le N$.

Proposition 4 *(Catoni 2004, Theorem 3.6.1) Suppose Assumptions* BD *and* C1, C2 *hold, and* $\|\check{f}_k\|_\infty \le L < \infty$, $1 \le k \le N$. *The estimator* $\hat{f}^C$ *that aggregates* $\check{f}_1, \ldots, \check{f}_N$ *by Catoni's algorithm has the risk*

$$R(\hat{f}^C; f_0; n) \le \inf_{1 \le k \le N} \left(\|\check{f}_k - f_0\|^2 + \frac{2}{n\lambda_C} \log \frac{1}{\pi_k} \right).$$

Juditsky et al. (2008) (p. 2200) give a similar result under simplified conditions.

3. ℓ_q-Aggregation of Estimates

Consider the setup from Section 2.1. We focus on the problem of aggregating the estimates in F_n to pursue the best performance in $\mathcal{F}_q(t_n)$ for $0 \le q \le 1$, $t_n > 0$, which we call *ℓ_q-aggregation of estimates*. To be more precise, when needed, it will be called $\ell_q(t_n)$-aggregation, and for the special case of $q = 0$, we call it $\ell_0(k_n)$-aggregation for $1 \le k_n \le M_n$.

3.1 The Strategy

For each $1 \le m \le M_n \wedge n$ and each subset model $J_m \subset \{1, 2, \ldots, M_n\}$ of size m, define $\mathcal{F}_{J_m} = \{\sum_{j\in J_m} \theta_j f_j : \theta_j \in \mathbb{R}, j \in J_m\}$. Let $\mathcal{F}^L_{J_m,s} = \{f_\theta = \sum_{j\in J_m} \theta_j f_j : \|\theta\|_1 \le s, \|f_\theta\|_\infty \le L\}$ $(s = 1, 2, ...)$ be the class of ℓ_1-constrained linear combinations in F_n with a sup-norm bound on f_θ. Our strategy is as follows.

Step I. Divide the data into two parts: $Z^{(1)} = (\mathbf{X}_i, Y_i)_{i=1}^{n_1}$ and $Z^{(2)} = (\mathbf{X}_i, Y_i)_{i=n_1+1}^{n}$.

Step II. Based on data $Z^{(1)}$, obtain a T-estimator for each function class $\mathcal{F}_{J_m}$, or obtain an AC-estimator for each function class $\mathcal{F}^L_{J_m,s}$ with $s \in \mathbb{N}$.

Step III. Based on data $Z^{(2)}$, combine all estimators obtained in step II and the null model ($f \equiv 0$) using Catoni's or the ARM algorithm. Let p_0 be a small positive number in $(0, 1)$. In all, we have to combine $\sum_{m=1}^{M_n\wedge n} \binom{M_n}{m}$ T-estimators with the weight $\pi_{J_m} = (1 - p_0)\left((M_n \wedge n)\binom{M_n}{m}\right)^{-1}$ and the null model with the weight $\pi_0 = p_0$, or combine countably many AC-estimators with the weight $\pi_{J_m,s} = (1 - p_0)\left((1 + s)^2(M_n \wedge n)\binom{M_n}{m}\right)^{-1}$ and the null model with the weight $\pi_0 = p_0$. (Note that sub-probabilities on the models do not affect the validity of the risk bounds to be given.)

For simplicity of exposition, from now on and when relevant, we assume n is even and choose $n_1 = n/2$ in our strategy. However, similar results hold for other values of n and n_1.

We use the expression "**E-G** strategy" for ease of presentation where $\mathbf{E} = \mathbf{T}$ or $\mathbf{AC}$ represents the estimators constructed in Step II, and $\mathbf{G} = \mathbf{C}$ or $\mathbf{Y}$ stands for the aggregation algorithm used in Step III. By our construction, Assumption AC1 is automatically satisfied: for each J_m, $H_{J_m,s} = \sup_{f,f'\in\mathcal{F}^L_{J_m,s},\mathbf{x}\in\mathcal{X}} |f(\mathbf{x}) - f'(\mathbf{x})| \le 2L$. Assumption AC2 is met with $(\sigma')^2 = \sigma^2 + 4L^2$.

We assume the following conditions are satisfied for each strategy, respectively.

$$\begin{aligned} \mathrm{A}_{\mathbf{T-C}} \text{ and } \mathrm{A}_{\mathbf{T-Y}} &: \text{BD, T1, T2.} \\ \mathrm{A}_{\mathbf{AC-C}} &: \text{BD, C1, C2.} \\ \mathrm{A}_{\mathbf{AC-Y}} &: \text{BD, Y1, Y2.} \end{aligned}$$

Given that T1, T2 are stronger than C1, C2 and Y1, Y2, it is enough to require their satisfaction in $\mathrm{A}_{\mathbf{T-C}}$ and $\mathrm{A}_{\mathbf{T-Y}}$.

3.2 Minimax Rates for ℓ_q-Aggregation of Estimates

Consider general M_n, t_n and $0 < q \leq 1$. The *ideal model size* (in order) that balances the approximation error and the estimation error under the ℓ_q-constraint over $1 \leq m \leq M_n \wedge n$ is

$$m^* = m^*(q, t_n) = \left\lceil 2\left(nt_n^2\tau\right)^{q/2} \right\rceil \wedge M_n \wedge n,$$

where $\tau = \sigma^{-2}$ is the precision parameter. The *effective model size* (in order) that yields the optimal rate of convergence, as will be shown, is

$$m_* = m_*(q, t_n) = \begin{cases} m^* & \text{if } m^* = M_n \wedge n, \\ \left\lceil \frac{m^*}{\left(1+\log \frac{M_n}{m^*}\right)^{q/2}} \right\rceil & \text{otherwise.} \end{cases}$$

See Appendix B for an explanation on why m_* is expected to yield the minimax rate of convergence. Let $\mathcal{F}_q^L(t_n) = \mathcal{F}_q(t_n) \cap \{f : \|f\|_\infty \leq L\}$ for $0 \leq q \leq 1$, and define

$$m_*^{\mathcal{F}} = \begin{cases} m_*(q, t_n) & \text{for case 1: } \mathcal{F} = \mathcal{F}_q(t_n), 0 < q \leq 1, \\ k_n \wedge n & \text{for case 2: } \mathcal{F} = \mathcal{F}_0(k_n), \\ m_*(q, t_n) \wedge k_n & \text{for case 3: } \mathcal{F} = \mathcal{F}_q(t_n) \cap \mathcal{F}_0(k_n), 0 < q \leq 1. \end{cases}$$

Note that in the third case, we are simply taking the smaller one between the effective model sizes from the soft sparsity constraint (ℓ_q-constraint with $0 < q \leq 1$) and the hard sparsity one (ℓ_0-constraint), and this smaller size determines the final sparsity. Define

$$REG(m_*^{\mathcal{F}}) = \sigma^2 \left(1 \wedge \frac{m_*^{\mathcal{F}} \cdot \left(1 + \log\left(\frac{M_n}{m_*^{\mathcal{F}}}\right)\right)}{n} \right),$$

which will be shown to be typically the optimal rate of the risk regret for ℓ_q-aggregation.

For case 3, we intend to achieve the best performance of linear combinations when both ℓ_0- and ℓ_q-constraints are imposed on the linear coefficients, which results in ℓ_q-aggregation using just a subset of the initial estimates and is called $\ell_0 \cap \ell_q$-aggregation. For the special case of $q = 1$, this $\ell_0 \cap \ell_1$-aggregation is studied in Yang (2004) (page 36) for multi-directional aggregation and in Lounici (2007) (called D-convex aggregation) more formally, giving also lower bounds. Our results below not only handle $q < 1$ but also close a gap of a logarithmic factor in upper and lower bounds in Lounici (2007).

For ease of presentation, we may use the same symbol (e.g., C) to denote possibly different constants of the same nature.

Theorem 5 *Suppose* $\mathrm{A}_{\mathbf{E-G}}$ *holds for the* ***E-G*** *strategy respectively. Our estimator* $\hat{f}_{F_n}$ *simultaneously has the following properties.*

(i) For ***T-*** *strategies, for* $\mathcal{F} = \mathcal{F}_q(t_n)$ *with* $0 < q \leq 1$, *or* $\mathcal{F} = \mathcal{F}_0(k_n)$, *or* $\mathcal{F} = \mathcal{F}_q(t_n) \cap \mathcal{F}_0(k_n)$ *with* $0 < q \leq 1$, *we have*

$$R(\hat{f}_{F_n}; f_0; n) \leq \left[C_0 d^2(f_0; \mathcal{F}) + C_1 REG(m_*^{\mathcal{F}})\right] \wedge \left[C_0 \left(\|f_0\|^2 \vee \frac{C_2\sigma^2}{n}\right)\right].$$

(ii) For ***AC-*** *strategies, for $\mathcal{F} = \mathcal{F}_q(t_n)$ with $0 < q \le 1$, or $\mathcal{F} = \mathcal{F}_0(k_n)$, or $\mathcal{F} = \mathcal{F}_q(t_n) \cap \mathcal{F}_0(k_n)$ with $0 < q \le 1$, we have*

$$R(\hat{f}_{F_n}; f_0; n) \le C_1 REG(m_*^{\mathcal{F}}) + C_0 \begin{cases} d^2(f_0; \mathcal{F}_q^L(t_n)) + \frac{C_2\sigma^2 \log(1+t_n)}{n} & \text{for case 1,} \\ \inf_{s\ge 1} \left(d^2(f_0; \mathcal{F}_1(s) \cap \mathcal{F}_0^L(k_n)) + \frac{C_2\sigma^2 \log(1+s)}{n} \right) & \text{for case 2,} \\ d^2(f_0; \mathcal{F}_q^L(t_n) \cap \mathcal{F}_0^L(k_n)) + \frac{C_2\sigma^2 \log(1+t_n)}{n} & \text{for case 3.} \end{cases}$$

Also, $R(\hat{f}_{F_n}; f_0; n) \le C_0 \left(\|f_0\|^2 \vee \frac{C_2\sigma^2}{n} \right)$.

For all these cases, C_0, C_1, and C_2 do not depend on n, f_0, t_n, q, k_n, M_n. These constants may depend on L, p_0, σ^2 or $\overline{\sigma}^2/\underline{\sigma}^2$, α, U_α, V_α when relevant. An exception is that $C_0 = 1$ for the ***AC-C*** *strategy.*

Remark 6 *For case 2, the boundedness assumption of $\|f_j\| \le 1$, $1 \le j \le M_n$ is not necessary.*

Remark 7 *If the true function f_0 happens to have a small L_2-norm such that $\|f_0\|^2 \vee \frac{\sigma^2}{n}$ is of a smaller order than $REG(m_*^{\mathcal{F}})$, then its inclusion in the risk bounds may improve the rate of convergence.*

Discussion of the bounds. Note that an extra term of $\log(1 + t_n)/n$ is present in the upper bounds of the estimator obtained by **AC-** strategies. For case 1, let us focus on the high-dimensional situation of M_n between order n and order $e^{O(n)}$. When t_n is no larger (in order) than $\sigma n^{1/q-1/2}$, the extra price $\log(1 + t_n)/n$ does not damage the rate of convergence if $\frac{\log(1+t_n)}{t_n^q}$ is no larger in order than $\frac{n^{q/2}}{\sigma^q}(\log M_n)^{1-q/2}$, which does hold as $n \to \infty$ when q is fixed. When t_n is at least of order $\sigma n^{1/q-1/2}$, $REG(m_*^{\mathcal{F}})$ is of order 1, and from Proposition 15 in the Appendix, it can be seen that under the conditions of the theorem, the risks of the AC- strategies are also of order 1. For case 2, the extra term in Theorem 5 is harmless in rate if for some $s \le e^{cn} \wedge e^{ck_n(1+\log(M_n/k_n))}$ for some constant $c > 0$, the ℓ_1-norm constraint does not enlarge the approximation error order.

Comparison to the existing literature. When $q = 1$, our theorem covers some important previous aggregation results. With $t_n = 1$, Juditsky and Nemirovski (2000) obtained the optimal result for large M_n; Yang (2004) gave upper bounds for all M_n, but the rate is slightly sub-optimal (by a logarithmic factor) when $M_n = O(\sqrt{n})$ and with a factor larger than 1 in front of the approximation error; Tsybakov (2003) derived the optimal rates for both large and small M_n (and also for linear aggregation) but under the assumption that the joint distribution of $\{f_j(\mathbf{X}), j = 1, ..., M_n\}$ is known. For the case $M_n = O(\sqrt{n})$, Audibert and Catoni (2010) have improved over Yang (2004) and Tsybakov (2003) by giving an optimal risk bound. Thus in the special case of $q = 1$, our result overcomes the aforementioned shortcomings by integrating the previous understandings together, with the additional generality of t_n. In the direction of adaptive aggregation, Dalalyan and Tsybakov (2012b) give risk bounds, up to a logarithmic factor, suitable for $q = 0, 1$. Our result here closes the logarithmic factor gap and also handles q between 0 and

1. Note also that Rigollet and Tsybakov (2010) obtain adaptive optimal ℓ_q-aggregation for $q = 0, 1$ under a fixed design Gaussian regression setting.

Next, we establish lower bounds for the aggregation problems that match (up to a multiplicative constant) the upper bounds above, which then implies that the estimators by our strategies are indeed minimax adaptive for ℓ_q-aggregation of estimates (with respect to both q and the ℓ_q-radius). Let $f_1, \ldots, f_{M_n}$ be an orthonormal basis with respect to the distribution of $\mathbf{X}$. Since the earlier upper bounds are obtained under the assumption that the true regression function f_0 satisfies $\|f_0\|_\infty \le L$ for some known (possibly large) constant $L > 0$, for our lower bound result below, this assumption will also be considered. For the last result in part (iii) below under the sup-norm constraint on f_0, the functions $f_1, \ldots, f_{M_n}$ are specially constructed on $[0, 1]$ and P_X is the uniform distribution on $[0, 1]$. See the proof for details.

In order to give minimax lower bounds without any norm assumption on f_0, let $\widetilde{m}_*^{\mathcal{F}}$ be defined the same as $m_*^{\mathcal{F}}$ except that the ceiling of n is removed. Define

$$\overline{REG}(\widetilde{m}_*^{\mathcal{F}}) = \frac{\sigma^2 \widetilde{m}_*^{\mathcal{F}} \cdot \left(1 + \log\left(\frac{M_n}{\widetilde{m}_*^{\mathcal{F}}}\right)\right)}{n} \wedge \begin{cases} t_n^2 & \text{for cases 1 and 3,} \\ \infty & \text{for case 2,} \end{cases}$$

$$\underline{REG}(m_*^{\mathcal{F}}) = REG(m_*^{\mathcal{F}}) \wedge \begin{cases} t_n^2 & \text{for cases 1 and 3,} \\ \infty & \text{for case 2.} \end{cases}$$

Theorem 8 *Suppose the noise ε follows a normal distribution with mean 0 and variance $\sigma^2 > 0$.*

(i) *For any aggregated estimator $\hat{f}_{F_n}$ based on an orthonormal dictionary $F_n = \{f_1, \ldots, f_{M_n}\}$, for $\mathcal{F} = \mathcal{F}_q(t_n)$, or $\mathcal{F} = \mathcal{F}_0(k_n)$, or $\mathcal{F} = \mathcal{F}_q(t_n) \cap \mathcal{F}_0(k_n)$ with $0 < q \le 1$, one can find a regression function f_0 (that may depend on $\mathcal{F}$) such that*

$$R(\hat{f}_{F_n}; f_0; n) - d^2(f_0; \mathcal{F}) \ge C \cdot \overline{REG}(\widetilde{m}_*^{\mathcal{F}}),$$

where C may depend on q (and only q) for cases 1 and 3 and is an absolute constant for case 2.

(ii) *Under the additional assumption that $\|f_0\| \le L$ for a known $L > 0$, the above lower bound becomes $C' \cdot \underline{REG}(m_*^{\mathcal{F}})$ for the three cases, where C' may depend on q and L for cases 1 and 3 and on L for case 2.*

(iii) *With the additional knowledge $\|f_0\|_\infty \le L$ for a known $L > 0$, the lower bound $C'' \cdot \underline{REG}(m_*^{\mathcal{F}})$ also holds for the following situations: 1) for $\mathcal{F} = \mathcal{F}_q(t_n)$ with $0 < q \le 1$, if $\sup_{f_\theta \in \mathcal{F}_q(t_n)} \|f_\theta\|_\infty \le L$; 2) for $\mathcal{F} = \mathcal{F}_0(k_n)$, if $\sup_{1 \le j \le M_n} \|f_j\|_\infty \le L < \infty$ and $\frac{k_n^2}{n}(1 + \log \frac{M_n}{k_n})$ are bounded above; 3) for $\mathcal{F} = \mathcal{F}_0(k_n)$, if $M_n / \left(1 + \log \frac{M_n}{k_n}\right) \le bn$ for some constant $b > 0$ and the orthonormal basis is specially chosen.*

Remark 9 *Consider the interesting high-dimensional case of M_n between order n and order $e^{O(n)}$. Then $\underline{REG}(m_*^{\mathcal{F}})$ is of the same order as $REG(m_*^{\mathcal{F}})$ unless t_n^2 is of a smaller order than $\log M_n / n$. Thus, except this situation of small t_n, the lower bounds above match the orders of the upper bounds in the previous theorem.*

For satisfaction of $\sup_{f_\theta \in \mathcal{F}_q(t_n)} \|f_\theta\|_\infty \le L$, consider uniformly bounded functions f_j, then for $0 < q \le 1$,

$$\|\sum_{j=1}^{M_n} \theta_j f_j\|_\infty \le \sum_{j=1}^{M_n} |\theta_j| \|f_j\|_\infty \le \left(\sup_{1\le j\le M_n} \|f_j\|_\infty\right) \|\theta\|_1 \le \left(\sup_{1\le j\le M_n} \|f_j\|_\infty\right) \|\theta\|_q.$$

Thus, under the condition that $\left(\sup_{1\le j\le M_n} \|f_j\|_\infty\right) t_n$ is upper bounded, $\sup_{f_\theta \in \mathcal{F}_q(t_n)} \|f_\theta\|_\infty \le L$ is met.

The lower bounds given in part (iii) of the theorem for the three cases of ℓ_q-aggregation of estimates are of the same order of the upper bounds in the previous theorem, respectively, unless t_n is too small. Hence, under the given conditions, the minimax rates for ℓ_q-aggregation are identified according to the definition of the minimax rate of aggregation in Tsybakov (2003). When no restriction is imposed on the norm of f_0, the lower bounds can certainly approach infinity (e.g., when t_n is really large). That is why $\overline{REG}(\widetilde{m}_*^{\mathcal{F}})$ is introduced. The same can be said for later lower bounds.

For the new case $0 < q < 1$, the ℓ_q-constraint imposes a type of soft-sparsity more stringent than $q = 1$: even more coefficients in the linear expression are pretty much negligible. For the discussion below, assume $m^* < n$. When the radius t_n increases or $q \to 1$, m^* increases given that the ℓ_q-ball enlarges. When $m_* = m^* = M_n < n$, the ℓ_q-constraint is not tight enough to impose sparsity: ℓ_q-aggregation is then simply equivalent to linear aggregation and the risk regret term corresponds to the estimation price of the full model, $M_n\sigma^2/n$. In contrast, when $1 < m_* < M_n \wedge n$, the rate for ℓ_q-aggregation is

$$\sigma^{2-q} t_n^q \left(\frac{\log\left(1 + \frac{M_n}{(n\tau t_n^2)^{q/2}}\right)}{n}\right)^{1-q/2}.$$

When $m^* \le (1 + \log(M_n/m^*))^{q/2}$ or equivalently $m_* = 1$, the ℓ_q-constraint restricts the search space of the optimization problem so much that it suffices to consider at most one f_j and the null model may provide a better risk.

Now let us explain that our ℓ_q-aggregation includes the commonly studied aggregation problems in the literature. First, when $q = 1$, we have the well-known convex or ℓ_1-aggregation (but now with the ℓ_1-norm bound allowed to be general). Second, when $q = 0$, with $k_n = M_n \le n$, we have the linear aggregation. For other $k_n < M_n \wedge n$, we have the aggregation to achieve the best linear performance of only k_n initial estimates. The case $q = 0$ and $k_n = 1$ has a special implication. Observe that from Theorem 5, we deduce that for both the **T-** strategies and **AC-** strategies, under the assumption $\sup_j \|f_j\|_\infty \le L$, our estimator satisfies

$$R(\hat{f}_{F_n}; f_0; n) \le C_0 \inf_{1\le j\le M_n} \|f_j - f_0\|^2 + C_1\sigma^2 \left(1 \wedge \frac{1 + \log M_n}{n}\right),$$

where $C_0 = 1$ for the **AC-C** strategy. Together with the lower bound of the order $\sigma^2\left(1 \wedge \frac{1+\log M_n}{n}\right)$ on the risk regret of aggregation for adaptation given in Tsybakov (2003), we conclude that $\ell_0(1)$-aggregation directly implies the aggregation for adaptation (model selection aggregation). As mentioned earlier, $\ell_0(k_n) \cap \ell_q(t_n)$-aggregation pursues the best

performance of the linear combination of at most k_n initial estimates with coefficients satisfying the ℓ_q-constraint, which includes the D-convex aggregation as a special case (with $q = 1$).

Some additional interesting results on combining procedures are in Audibert (2007, 2009); Birgé (2006); Bunea and Nobel (2008); Catoni (2012); Dalalyan and Tsybakov (2007, 2012a); Giraud (2008); Goldenshluger (2009); Györfi et al. (2002); Györfi and Ottucsák (2007); Wegkamp (2003); Yang (2001).

4. Linear Regression with ℓ_q-Constrained Coefficients under Random Design

Let's consider the linear regression model with M_n predictors $X_1, \ldots, X_{M_n}$. Suppose the data are drawn i.i.d. from the following model

$$Y = f_0(\mathbf{X}) + \varepsilon = \sum_{j=1}^{M_n} \theta_j X_j + \varepsilon. \tag{2}$$

As previously defined, for a function $f(x_1, \ldots, x_{M_n}) : \mathcal{X} \to \mathbb{R}$, the L_2-norm $\|f\|$ is the square root of $Ef^2(X_1, \ldots, X_{M_n})$, where the expectation is taken with respect to P_X, the distribution of $\mathbf{X}$. Denote the $\ell_{q,t_n}^{M_n}$-hull in this context by

$$\mathcal{F}_q(t_n; M_n) = \left\{ f_\theta = \sum_{j=1}^{M_n} \theta_j x_j : \|\theta\|_q \le t_n \right\}, \ 0 \le q \le 1, \ t_n > 0.$$

For linear regression, we assume coefficients of the true regression function f_0 have a sparse ℓ_q-representation ($0 < q \le 1$) or ℓ_0-representation or both, i.e., $f_0 \in \mathcal{F}$ where $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$, $\mathcal{F}_0(k_n; M_n)$ or $\mathcal{F}_q(t_n; M_n) \bigcap \mathcal{F}_0(k_n; M_n)$.

Assumptions BD and $A_{\mathbf{E-G}}$ are still relevant in this section. As in the previous section, for AC-estimators, we consider ℓ_1- and sup-norm constraints.

For each $1 \le m \le M_n \wedge n$ and each subset J_m of size m, let $\mathcal{G}_{J_m} = \{\sum_{j \in J_m} \theta_j x_j : \theta \in \mathbb{R}^m\}$ and $\mathcal{G}_{J_m,s}^L = \{\sum_{j \in J_m} \theta_j x_j : \|\theta\|_1 \le s, \|f_\theta\|_\infty \le L\}$. We introduce now the adaptive estimator $\hat{f}_A$, built with the same strategy used to construct $\hat{f}_{F_n}$ except that we now consider $\mathcal{G}_{J_m}$ and $\mathcal{G}_{J_m,s}^L$ instead of $\mathcal{F}_{J_m}$ and $\mathcal{F}_{J_m,s}^L$.

4.1 Upper Bounds

We give upper bounds for the risk of our estimator assuming $f_0 \in \mathcal{F}_q^L(t_n; M_n)$, $\mathcal{F}_0^L(k_n; M_n)$, or $\mathcal{F}_q^L(t_n; M_n) \cap \mathcal{F}_0^L(k_n; M_n)$, where $\mathcal{F}^L = \{f : f \in \mathcal{F}, \|f\|_\infty \le L\}$ for a positive constant L. Let $\alpha_n = \sup_{f \in \mathcal{F}_0^L(k_n; M_n)} \inf\{\|\theta\|_1 : f_\theta = f\}$ be the maximum smallest ℓ_1-norm needed to represent the functions in $\mathcal{F}_0^L(k_n; M_n)$.

Recall $m^* = \left\lceil 2\left(nt_n^2/\sigma^2\right)^{q/2} \right\rceil \wedge M_n \wedge n$ and m_* equals m^* when $m^* = M_n \wedge n$ and $\left\lceil \frac{m^*}{\left(1+\log \frac{M_n}{m^*}\right)^{q/2}} \right\rceil$ otherwise. For ease of presentation, define $\Psi^{\mathcal{F}}$ as follows:

$$\Psi^{\mathcal{F}_q^L(t_n;M_n)} = \begin{cases} \sigma^2 & \text{if } m_* = n, \\ \frac{\sigma^2 M_n}{n} & \text{if } m_* = M_n < n, \\ \sigma^{2-q} t_n^q \left(\frac{1+\log \frac{M_n}{(nt_n^2\tau)^{q/2}}}{n} \right)^{1-q/2} \wedge \sigma^2 & \text{if } 1 < m_* < M_n \wedge n, \\ \left(t_n^2 \vee \frac{\sigma^2}{n}\right) \wedge \sigma^2 & \text{if } m_* = 1, \end{cases}$$

$$\Psi^{\mathcal{F}_0^L(k_n;M_n)} = \sigma^2 \left(1 \wedge \frac{k_n\left(1+\log \frac{M_n}{k_n}\right)}{n} \right),$$

$$\Psi^{\mathcal{F}_q^L(t_n;M_n)\cap\mathcal{F}_0^L(k_n;M_n)} = \Psi^{\mathcal{F}_q^L(t_n;M_n)} \wedge \Psi^{\mathcal{F}_0^L(k_n;M_n)}.$$

In addition, for lower bound results, let $\underline{\Psi}^{\mathcal{F}_q^L(t_n;M_n)}$ $(0 \le q \le 1)$ and $\underline{\Psi}^{\mathcal{F}_q^L(t_n;M_n)\cap\mathcal{F}_0^L(k_n;M_n)}$ $(0 < q \le 1)$ be the same as $\Psi^{\mathcal{F}_q^L(t_n;M_n)}$ and $\Psi^{\mathcal{F}_q^L(t_n;M_n)\cap\mathcal{F}_0^L(k_n;M_n)}$, respectively, except that when $0 < q \le 1$ and $m_* = 1$, $\underline{\Psi}^{\mathcal{F}_q^L(t_n;M_n)}$ takes the value $\sigma^2 \wedge t_n^2$ instead of $\sigma^2 \wedge \left(t_n^2 \vee \frac{\sigma^2}{n}\right)$ and $\underline{\Psi}^{\mathcal{F}_q^L(t_n;M_n)\cap\mathcal{F}_0^L(k_n;M_n)}$ is modified the same way.

Corollary 10 *Suppose $A_{\mathbf{E-G}}$ holds for the **E-G** strategy respectively, and $\sup_{1\le j\le M_n} \|X_j\|_\infty \le 1$. The estimator $\hat{f}_A$ simultaneously has the following properties.*

*(i) For **T-** strategies, for $\mathcal{F} = \mathcal{F}_q^L(t_n; M_n)$ with $0 < q \le 1$, or $\mathcal{F} = \mathcal{F}_0^L(k_n; M_n)$, or $\mathcal{F} = \mathcal{F}_q^L(t_n; M_n) \cap \mathcal{F}_0^L(k_n; M_n)$ with $0 < q \le 1$, we have*

$$\sup_{f_0\in\mathcal{F}} R(\hat{f}_A; f_0; n) \le C_1 \Psi^{\mathcal{F}},$$

where the constant C_1 does not depend on n.

*(ii) For **AC-** strategies, for $\mathcal{F} = \mathcal{F}_q^L(t_n; M_n)$ with $0 < q \le 1$, or $\mathcal{F} = \mathcal{F}_0^L(k_n; M_n)$, or $\mathcal{F} = \mathcal{F}_q^L(t_n; M_n) \cap \mathcal{F}_0^L(k_n; M_n)$ with $0 < q \le 1$, we have*

$$\sup_{f_0\in\mathcal{F}} R(\hat{f}_A; f_0; n) \le C_1 \Psi^{\mathcal{F}} + C_2 \begin{cases} \frac{\sigma^2 \log(1+\alpha_n)}{n} & \text{for } \mathcal{F} = \mathcal{F}_0^L(k_n; M_n), \\ \frac{\sigma^2 \log(1+t_n)}{n} & \text{otherwise,} \end{cases}$$

where the constants C_1 and C_2 do not depend on n.

We need to point out that closely related work has been done under a fixed design setting. While determining the minimax rate of convergence, Raskutti et al. (2012) derive optimal estimators of a function (only at the design points) in the ℓ_q-hulls for $0 \le q \le 1$, but the estimators require knowledge of q and t_n. Negahban et al. (2012), when applying their general M-estimation methodology to the same fixed design regression setting, give

an optimal estimator of the true coefficient vector (as opposed to the regression function) assumed in any ℓ_q-ball for $0 \le q \le 1$ under the squared error loss. It requires that the predictors satisfy a restricted eigenvalue (RE) condition. Raskutti et al. (2010) show that under broad Gaussian random designs, the RE condition holds with exponentially small exception probability. Therefore, the fixed design performance bounds in Negahban et al. (2012), as well as those in Raskutti et al. (2012) (which do not need the RE condition), can be used to draw some conclusions for Gaussian random designs or more general random design in the latter case. Our risk upper bounds (directly under the global squared L_2 loss) do not require the Gaussian assumption on the covariates nor the RE condition. In addition, differently from the aforementioned two papers, our performance bounds hold without any restriction on the relationship between M and n.

4.2 Lower Bounds

The lower bounds used in the previous section for deriving the minimax rate of aggregation are not suitable for obtaining the minimax rate of convergence for the current regression estimation problem. We make the following near orthogonality assumption on sparse subcollections of the predictors. Such an assumption, similar to the sparse Riesz condition (SRC) (Zhang 2010) under fixed design, is used only for lower bounds but not for upper bounds.

Assumption SRC: For some $\gamma > 0$, there exist two positive constants $\underline{a}$ and $\overline{a}$ that do not depend on n such that for every θ with $\|\theta\|_0 \le \min(2\gamma, M_n)$ we have

$$\underline{a}\|\theta\|_2 \le \|f_\theta\| \le \overline{a}\|\theta\|_2.$$

Theorem 11 *Suppose the noise ε follows a normal distribution with mean 0 and variance $0 < \sigma^2 < \infty$.*

(i) For $0 < q \le 1$, under Assumption SRC with $\gamma = m_*$, we have

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_q(t_n; M_n)} E\|\hat{f} - f_0\|^2 \ge c\underline{\Psi}^{\mathcal{F}_q^L(t_n; M_n)}.$$

(ii) Under Assumption SRC with $\gamma = k_n$, we have

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_0(k_n; M_n) \cap \{f_\theta : \|\theta\|_2 \le a_n\}} E\|\hat{f} - f_0\|^2 \ge c' \begin{cases} \underline{\Psi}^{\mathcal{F}_0^L(k_n; M_n)} & \text{if } a_n \ge \tilde{c}\sigma\sqrt{\frac{k_n\left(1+\log\frac{M_n}{k_n}\right)}{n}}, \\ a_n^2 & \text{if } a_n < \tilde{c}\sigma\sqrt{\frac{k_n\left(1+\log\frac{M_n}{k_n}\right)}{n}}. \end{cases}$$

where $\tilde{c}$ is a pure constant.

(iii) For any $0 < q \le 1$, under Assumption SRC with $\gamma = k_n \wedge m_*$, we have

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_0(k_n; M_n) \cap \mathcal{F}_q(t_n; M_n)} E\|\hat{f} - f_0\|^2 \ge c'' \underline{\Psi}^{\mathcal{F}_q^L(t_n; M_n) \cap \mathcal{F}_0^L(k_n; M_n)}.$$

For all cases, $\inf_{\hat{f}}$ *is over all estimators and the constants* c, $c^{'}$ *and* $c^{''}$ *may depend on* $\underline{a}$, $\overline{a}$, q *and* σ^2.

For the second case (ii), the lower bound is stated in a more informative way because the effect of the bound on $\|\theta\|_2$ is clearly seen.

4.3 The Minimax Rates of Convergence

Combining the upper and lower bounds, we give a representative minimax rate result with the roles of the key quantities n, M_n, q, and k_n explicitly seen in the rate expressions. Below "$\asymp$" means of the same order when L, L_0, q, $t_n = t$, and $\overline{\sigma}^2$ ($\overline{\sigma}^2$ is defined in Corollary 12 below) are held constant in the relevant expressions.

Corollary 12 *Suppose the noise* ε *follows a normal distribution with mean* 0 *and variance* σ^2, *and there exists a known constant* $\overline{\sigma}$ *such that* $0 < \sigma \le \overline{\sigma} < \infty$. *Also assume there exists a known constant* $L_0 > 0$ *such that* $\sup_{1\le j\le M_n} \|X_j\|_\infty \le L_0 < \infty$.

(i) For $0 < q \le 1$, *under Assumption SRC with* $\gamma = m_*$,

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_q^L(t;M_n)} E\|\hat{f} - f_0\|^2 \asymp 1 \wedge \begin{cases} 1 & \text{if } m_* = n, \\ \frac{M_n}{n} & \text{if } m_* = M_n < n, \\ \left(\frac{1+\log \frac{M_n}{(nt^2\tau)^{q/2}}}{n} \right)^{1-q/2} & \text{if } 1 \le m_* < M_n \wedge n. \end{cases}$$

(ii) If there exists a constant $K_0 > 0$ *such that* $\frac{k_n^2\left(1+\log \frac{M_n}{k_n}\right)}{n} \le K_0$, *then under Assumption SRC with* $\gamma = k_n$,

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_0^L(k_n;M_n) \cap \{f_\theta : \|\theta\|_\infty \le L_0\}} E\|\hat{f} - f_0\|^2 \asymp 1 \wedge \frac{k_n\left(1+\log \frac{M_n}{k_n}\right)}{n}.$$

(iii) If $\sigma > 0$ *is actually known, then under the condition* $\frac{k_n^2\left(1+\log \frac{M_n}{k_n}\right)}{n} \le K_0$ *and Assumption SRC with* $\gamma = k_n$, *we have*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_0^L(k_n;M_n)} E\|\hat{f} - f_0\|^2 \asymp 1 \wedge \frac{k_n\left(1+\log \frac{M_n}{k_n}\right)}{n},$$

and for any $0 < q \le 1$, *under Assumption SRC with* $\gamma = k_n \wedge m_*$, *we have*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_0^L(k_n;M_n) \cap \mathcal{F}_q^L(t;M_n)} E\|\hat{f} - f_0\|^2 \asymp 1 \wedge \begin{cases} \frac{k_n\left(1+\log \frac{M_n}{k_n}\right)}{n} & \text{if } m_* > k_n, \\ \left(\frac{1+\log \frac{M_n}{(nt^2\tau)^{q/2}}}{n} \right)^{1-q/2} & \text{if } 1 \le m_* \le k_n. \end{cases}$$

5. Conclusion

Sparse modeling by imposing an ℓ_q-constraint on the coefficients of a linear representation of a target function to be learned has found consensus among academics and practitioners in many application fields, among which, just to mention a few, compressed sensing, signal and image compression, gene-expression, cryptography and recovery of loss data. The ℓ_q-constraints promote sparsity essential for high-dimensional learning and they also are often approximately satisfied on natural classes of signal and images, such as the bounded variation model for images and the bump algebra model for spectra (see Donoho 2006).

In the direction of using the ℓ_1-constraints in constructing estimators, algorithmic and theoretical results have been well developed. Both the Lasso and the Dantzig selector have been shown to achieve the rate $k_n \log(M_n)/n$ under different conditions on correlations of predictors and the hard sparsity constraint on the linear coefficients (see van de Geer and Bühlmann 2009 for a discussion about the sufficient conditions for deriving oracle inequalities for the Lasso). Our upper bound results do not require any of those conditions, but we do assume the sparse Riesz condition for deriving the lower bounds. Computational issues aside, we have seen that the approach of model selection/combination with descriptive complexity penalty has provided the most general adaptive estimators that automatically exploit the sparsity characteristics of the target function in terms of linear approximations subject to ℓ_q-constraints.

In our results, the effective model size m_* (as defined in Section 3.2 and further explained in Appendix B) plays a key role in determining the minimax rate of ℓ_q-aggregation for $0 < q \leq 1$. With the extended definition of the effective model size m_* to be simply the number of nonzero components k_n when $q = 0$ and re-defining m_* to be $m_* \wedge k_n$ under both ℓ_q- $(0 < q \leq 1)$ and ℓ_0-constraints, the minimax rate of aggregation is unified to be the simple form $1 \wedge \frac{m_*\left(1+\log\left(\frac{M_n}{m_*}\right)\right)}{n}$.

The ℓ_q-aggregation includes as special cases the state-of-art aggregation problems, namely aggregation for adaptation, convex and D-convex aggregations, linear aggregation, and subset selection aggregation, and all of them can be defined (or essentially so) by considering linear combinations under ℓ_0- and/or ℓ_1-constraints. Our investigation provides optimal rates of aggregation, which not only agrees with (and, in some cases, improves over) previous findings for the mostly studied aggregation problems, but also holds for a much larger set of linear combination classes. Indeed, we have seen that ℓ_0-aggregation includes aggregation for adaptation over the initial estimates (or model selection aggregation) ($\ell_0(1)$-aggregation), linear aggregation when $M_n \leq n$ ($\ell_0(M_n)$-aggregation), and aggregation to achieve the best performance of linear combination of k_n estimates in the dictionary for $1 < k_n < M_n$ (sometimes called subset selection aggregation) ($\ell_0(k_n)$-aggregation). When M_n is large, aggregating a subset of the dictionary under an ℓ_q-constraint for $0 < q \leq 1$ can be advantageous, which is just $\ell_0(k_n) \cap \ell_q(t_n)$-aggregation. Since the optimal rates of aggregation as defined in Tsybakov (2003), can differ substantially in different directions of aggregation and typically one does not know which direction works the best for the unknown regression function, multi-directional or universal aggregation is important so that the final estimator is automatically conservative and aggressive, whichever is better (see Yang 2004). Our aggregation strategy is indeed multi-directional , achieving the optimal rates over all ℓ_q-aggregation for $0 \leq q \leq 1$ and $\ell_0 \cap \ell_q$-aggregation for all $0 < q \leq 1$.

Our focus in this work is of a theoretical nature to provide an understanding of the fundamental theoretical issues about ℓ_q-aggregation or linear regression under ℓ_q-constraints. Computational aspects will be studied in the future.

Acknowledgments

We thank Yannick Baraud for helpful discussions and for pointing out the work of Audibert and Catoni (2011) that helped us to remove a logarithmic factor in some of our previous aggregation risk bounds. Three anonymous referees are sincerely thanked for their valuable comments on improving our work.

Sandra Paterlini conducted part of this research while visiting the School of Mathematics, University of Minnesota. Her research was partially supported by Fondazione Cassa di Risparmio di Modena for REFIGLO. Yuhong Yang's research was partially supported by NSF grant DMS-1106576.

Appendix A. Metric Entropy and Sparse Approximation Error of $\ell_{q,t_n}^{M_n}$-Hulls

It is well-known that the metric entropy plays a fundamental role in determining minimax-rates of convergence, as shown, e.g., in Birgé (1986); Yang and Barron (1999).

For each $1 \le m \le M_n$ and each subset $J_m \subset \{1, 2, \ldots, M_n\}$ of size m, recall $\mathcal{F}_{J_m} = \{\sum_{j\in J_m} \theta_j f_j : \theta_j \in \mathbb{R}, j \in J_m\}$. Recall also

$$d^2(f_0; \mathcal{F}) = \inf_{f_\theta \in \mathcal{F}} \|f_\theta - f_0\|^2$$

is the smallest approximation error to f_0 over a function class $\mathcal{F}$.

Theorem 13 *(Metric entropy and sparse approximation bound for $\ell_{q,t_n}^{M_n}$-hulls) Suppose $F_n = \{f_1, f_2, ..., f_{M_n}\}$ with $\|f_j\|_{L^2(\nu)} \le 1$, $1 \le j \le M_n$, where ν is a σ-finite measure.*

(i) For $0 < q \le 1$, there exists a positive constant c_q depending only on q, such that for any $0 < \epsilon < t_n$, $\mathcal{F}_q(t_n)$ contains an ϵ-net $\{e_j\}_{j=1}^{N_\epsilon}$ in the $L_2(\nu)$ distance with $\|e_j\|_0 \le 5(t_n\epsilon^{-1})^{2q/(2-q)} + 1$ for $j = 1, 2, ..., N_\epsilon$, where N_ϵ satisfies

$$\log N_\epsilon \le \begin{cases} c_q \left(t_n\epsilon^{-1}\right)^{\frac{2q}{2-q}} \log(1 + M_n^{\frac{1}{q}-\frac{1}{2}} t_n^{-1}\epsilon) & \text{if } \epsilon > t_n M_n^{\frac{1}{2}-\frac{1}{q}}, \\ c_q M_n \log(1 + M_n^{\frac{1}{2}-\frac{1}{q}} t_n\epsilon^{-1}) & \text{if } \epsilon \le t_n M_n^{\frac{1}{2}-\frac{1}{q}}. \end{cases} \tag{3}$$

(ii) For any $1 \le m \le M_n$, $0 < q \le 1$, $t_n > 0$, there exists a subset J_m and $f_{\theta^m} \in \mathcal{F}_{J_m}$ with $\|\theta^m\|_1 \le t_n$ such that the sparse approximation error is upper bounded as follows

$$\|f_{\theta^m} - f_0\|^2 - d^2(f_0; \mathcal{F}_q(t_n)) \le 2^{2/q-1} t_n^2 m^{1-2/q}. \tag{4}$$

Remark 14 *The metric entropy estimate (3) is the best possible. Indeed, if f_j, $1 \le j \le M_n$, are orthonormal functions, then (3) is sharp in order for any ϵ satisfying that ϵ/t_n is bounded away from 1 (see Kühn 2001). Part (i) of Theorem 13 implies Lemma 3 of Raskutti et al. (2012), with an improvement of a $\log(M_n)$ factor when $\epsilon \approx t_n M_n^{\frac{1}{2}-\frac{1}{q}}$, and an improvement from $(t_n\epsilon^{-1})^{\frac{2q}{2-q}}\log(M_n)$ to $M_n \log(1 + M_n^{\frac{1}{q}-\frac{1}{2}} t_n\epsilon^{-1})$ when $\epsilon < t_n M_n^{\frac{1}{2}-\frac{1}{q}}$. These improvements are needed to derive the exact minimax rates for some of the possible situations in terms of M_n, q, and t_n.*

A.1 Proof of Theorem 13

(i) Because $\{e_j\}_{j=1}^{N_\epsilon}$ is an ϵ-net of $\mathcal{F}_q(t_n)$ if and only if $\{t_n^{-1}e_j\}_{j=1}^{N_\epsilon}$ is an ϵ/t_n-net of $\mathcal{F}_q(1)$, we only need to prove the theorem for the case $t_n = 1$. Recall that for any positive integer k, the unit ball of $\ell_q^{M_n}$ can be covered by 2^{k-1} balls of radius ϵ_k in ℓ_1 distance, where

$$\epsilon_k \le c \begin{cases} 1 & 1 \le k \le \log_2(2M_n) \\ \left(\frac{\log_2(1+\frac{2M_n}{k})}{k}\right)^{\frac{1}{q}-1} & \log_2(2M_n) \le k \le 2M_n \\ 2^{-\frac{k}{2M_n}}(2M_n)^{1-\frac{1}{q}} & k \ge 2M_n \end{cases}$$

(c.f., Edmunds and Triebel 1998, page 98). Thus, there are 2^{k-1} functions g_j, $1 \le j \le 2^{k-1}$, such that

$$\mathcal{F}_q(1) \subset \bigcup_{j=1}^{2^{k-1}} (g_j + \mathcal{F}_1(\epsilon_k)).$$

Note that without loss of generality, g_j can be assumed to belong to $\mathcal{F}_q(1)$ (because if not we can replace it by a member in $\mathcal{F}_q(1)$ closest to it in ℓ_1 distance on the coefficient vectors (which is a real distance), the effect of which is merely a change of the constant c above). For any $g \in \mathcal{F}_1(\epsilon_k)$, g can be expressed as $g = \sum_{i=1}^{M_n} c_i f_i$ with $\sum_{i=1}^{M_n} |c_i| \le \epsilon_k$. Following the idea of Maurey's empirical method (see, e.g., Pisier 1981), we define a random function U, such that

$$\mathbb{P}(U = \mathrm{sign}(c_i)\epsilon_k f_i) = |c_i|/\epsilon_k, \quad \mathbb{P}(U = 0) = 1 - \sum_{i=1}^{M_n} |c_i|/\epsilon_k.$$

Then, we have $\|U\|_2 \le \epsilon_k$ a.s. and $\mathbb{E}U = g$ under the randomness just introduced. Let $U_1, U_2, ..., U_m$ be i.i.d. copies of U, and let $V = \frac{1}{m}\sum_{i=1}^m U_i$. We have

$$\mathbb{E}\|V - g\|_2 \le \sqrt{\frac{1}{m}\|\mathbb{V}ar(U)\|_2} \le \sqrt{\frac{1}{m}\mathbb{E}\|U\|_2^2} \le \frac{\epsilon_k}{\sqrt{m}}.$$

In particular, there exists a realization of V, such that $\|V - g\|_2 \le \epsilon_k/\sqrt{m}$. Note that V can be expressed as $\epsilon_k m^{-1}(k_1 f_1 + k_2 f_2 + \cdots + k_{M_n} f_{M_n})$, where k_1, k_2, ..., k_{M_n} are integers, and $|k_1| + |k_2| + \cdots + |k_{M_n}| \le m$. Thus, the total number of different realizations of V is upper bounded by $\binom{2M_n+m}{m}$. Furthermore, $\|V\|_0 \le m$.

If $\log_2(2M_n) \le k \le 2M_n$, we choose m to be the largest integer such that $\binom{2M_n+m}{m} \le 2^k$. Then we have

$$\frac{1}{m} \le \frac{c'}{k}\log_2\left(1 + \frac{2M_n}{k}\right)$$

for some positive constant c'. Hence, $\mathcal{F}_q(1)$ can be covered by 2^{2k-1} balls of radius

$$\epsilon_k \sqrt{c' k^{-1} \log_2 \left(1 + \frac{2M_n}{k}\right)}$$

in L^2 distance.

If $k \geq 2M_n$, we choose $m = M_n$. Then $\mathcal{F}_q(1)$ can be covered by $2^{k-1}\binom{2M_n+m}{m}$ balls of radius $\epsilon_k M_n^{-1/2}$ in L^2 distance. Consequently, there exists a positive constant c'' such that $\mathcal{F}_q(1)$ can be covered by 2^{l-1} balls of radius r_l, where

$$r_l \leq c'' \begin{cases} 1 & 1 \leq l \leq \log_2(2M_n), \\ l^{\frac{1}{2}-\frac{1}{q}} [\log_2(1 + \frac{2M_n}{l})]^{\frac{1}{q}-\frac{1}{2}} & \log_2(2M_n) \leq l \leq 2M_n, \\ 2^{-\frac{l}{2M_n}} (2M_n)^{\frac{1}{2}-\frac{1}{q}} & l \geq 2M_n. \end{cases}$$

For any given $0 < \epsilon < 1$, by choosing the smallest l such that $r_l < \epsilon/2$, we find an $\epsilon/2$-net $\{u_i\}_{i=1}^N$ of $\mathcal{F}_q(1)$ in L^2 distance, where

$$N = 2^{l-1} \leq \begin{cases} \exp\left(c''' \epsilon^{-\frac{2q}{2-q}} \log(1 + M_n^{\frac{1}{q}-\frac{1}{2}} \epsilon)\right) & \epsilon > M_n^{\frac{1}{2}-\frac{1}{q}}, \\ \exp\left(c''' M_n \log(1 + M_n^{\frac{1}{2}-\frac{1}{q}} \epsilon^{-1})\right) & \epsilon < M_n^{\frac{1}{2}-\frac{1}{q}}, \end{cases}$$

and c''' is some positive constant.

It remains to show that for each $1 \leq i \leq N$, we can find a function e_i so that $\|e_i\|_0 \leq 5\epsilon^{2q/(q-2)} + 1$ and $\|e_i - u_i\|_2 \leq \epsilon/2$.

Suppose $u_i = \sum_{j=1}^{M_n} c_{ij} f_j$, $1 \leq i \leq N$, with $\sum_{j=1}^{M_n} |c_{ij}|^q \leq 1$. Let $L_i = \{j : |c_{ij}| > \epsilon^{2/(2-q)}\}$. Then, $|L_i|\epsilon^{2q/(2-q)} \leq \sum |c_{ij}|^q \leq 1$, which implies $|L_i| \leq \epsilon^{2q/(q-2)}$ and also

$$\sum_{j \notin L_i} |c_{ij}| \leq \sum_{j \notin L_i} |c_{ij}|^q [\epsilon^{2/(2-q)}]^{1-q} \leq \epsilon^{\frac{2-2q}{2-q}}.$$

Define $v_i = \sum_{j \in L_i} c_{ij} f_j$ and $w_i = \sum_{j \notin L_i} c_{ij} f_j$. We have $w_i \in \mathcal{F}_1(\epsilon^{\frac{2-2q}{2-q}})$. By the probability argument above, we can find a function w_i' such that $\|w_i'\|_0 \leq m$ and $\|w_i - w_i'\|_2 \leq \epsilon^{\frac{2-2q}{2-q}}/\sqrt{m}$. In particular, if we choose m to be the smallest integer such that $m \geq 4\epsilon^{2q/(q-2)}$. Then, $\|w_i - w_i'\|_2 \leq \epsilon/2$.

We define $e_i = v_i + w_i'$, we have $\|u_i - e_i\|_2 \leq \epsilon/2$, and then we can show that

$$\|e_i\|_0 = \|v_i\|_0 + \|w_i'\|_0 \leq |L_i| + m \leq 5\epsilon^{2q/(q-2)} + 1.$$

(ii) Let $f_\theta^* = \sum_{j=1}^{M_n} c_j f_j = \arg\inf_{f_\theta \in \mathcal{F}_q(t_n)} \|f_\theta - f_0\|^2$ be a best approximation of f_0 over the class $\mathcal{F}_q(t_n)$. For any $1 \leq m \leq M_n$, let $L^* = \{j : |c_j| > t_n m^{-1/q}\}$. Because $\sum_{j=1}^{M_n} |c_j|^q \leq t_n^q$, we have $|L^*| t_n^q / m < \sum |c_j|^q \leq t_n^q$. So, $|L^*| < m$. Also,

$$D := \sum_{j \notin L^*} |c_j| \leq \sum_{j \notin L^*} |c_j|^q [t_n (1/m)^{1/q}]^{1-q} = \sum_{j \notin L^*} |c_j|^q t_n^{1-q} (1/m)^{(1-q)/q} \leq t_n m^{1-1/q}.$$

Define $v^* = \sum_{j\in L^*} c_j f_j$ and $w^* = \sum_{j\notin L^*} c_j f_j$. We have $w^* \in \mathcal{F}_1(D)$. Define a random function U so that $\mathbb{P}(U = D\text{sign}(c_j)f_j) = |c_j|/D$, $j \notin L^*$. Thus, $\mathbb{E}U = w^*$, where $\mathbb{E}$ denotes expectation with respect to the randomness $\mathbb{P}$ (just introduced). Also, $\|U\| \leq D \sup_{1\leq j\leq M_n} \|f_j\| \leq D$. Let $U_1, U_2, ..., U_m$ be i.i.d. copies of U, then $\forall \mathbf{x} \in \mathcal{X}$,

$$\mathbb{E}\left(f_0(\mathbf{x}) - v^*(\mathbf{x}) - \frac{1}{m}\sum_{i=1}^m U_i(\mathbf{x})\right)^2 = \left(f_\theta^*(\mathbf{x}) - f_0(\mathbf{x})\right)^2 + \frac{1}{m}\mathbb{V}ar\left(U(\mathbf{x})\right).$$

Together with Fubini,

$$\mathbb{E}\left\|f_0 - v^* - \frac{1}{m}\sum_{i=1}^m U_i\right\|^2 \leq \|f_\theta^* - f_0\|^2 + \frac{1}{m}E\|U\|^2 \leq \|f_\theta^* - f_0\|^2 + t_n^2 m^{1-2/q}.$$

In particular, there exists a realization of $v^* + \frac{1}{m}\sum_{i=1}^m U_i$, denoted by f_{θ^m}, such that $\|f_{\theta^m} - f_0\|^2 \leq \|f_\theta^* - f_0\|^2 + t_n^2 m^{1-2/q}$. Note that $\|\theta^m\|_1 \leq t_n$ and $\|\theta^m\|_0 \leq 2m-1$. If we consider $\widetilde{m} = \lfloor (m+1)/2 \rfloor$ instead, we have $2\widetilde{m} - 1 \leq m$ and $\widetilde{m} \geq m/2$. The conclusion then follows. This completes the proof of the theorem.

Appendix B. An Insight from the Sparse Approximation Bound Based on Classical Model Selection Theory

Consider general M_n, t_n and $0 < q \leq 1$. With the approximation error bound in Theorem 13, classical model selection theories can provide key insight on what to expect regarding the minimax rate of convergence for estimating a function in the $\ell_{q,t_n}^{M_n}$-hull.

Suppose J_m is the best subset model of size m in terms of having the smallest L_2 approximation error to f_0. Then, the estimator based on J_m is expected to have the risk (under some squared error loss) of order

$$2^{2/q} t_n^2 m^{1-2/q} + \frac{\sigma^2 m}{n}.$$

Minimizing this bound over m, we get the best choice (in order) in the range $1 \leq m \leq M_n \wedge n$:

$$m^* = m^*(q, t_n) = \left\lceil 2\left(nt_n^2\tau\right)^{q/2}\right\rceil \wedge M_n \wedge n,$$

where $\tau = \sigma^{-2}$ is the precision parameter. When $q = 0$ with $t_n = k_n$, m^* should be taken to be $k_n \wedge n$. It is the *ideal model size* (in order) under the ℓ_q-constraint because it provides the best possible trade-off between the approximation error and estimation error when $1 \leq m \leq M_n \wedge n$. The calculation of balancing the approximation error and the estimation error is well-known to lead to the minimax rate of convergence for general full approximation sets of functions with pre-determined order of the terms in an approximation system (see section 4 of Yang and Barron 1999). However, when the terms are not pre-ordered, there are many models of the same size m^*, and one must pay a price for dealing with exponentially many or more models (see, e.g., section 5 of Yang and Barron 1999). The classical model selection theory that deals with searching over a large number of models tells us that the price of searching over $\binom{M_n}{m^*}$ many models is the addition of the term

$\log \binom{M_n}{m^*}/n$ (e.g., Barron and Cover 1991; Yang and Barron 1998; Barron et al. 1999; Yang 1999; Baraud 2000; Birgé and Massart 2001; Baraud 2002; Massart 2007). That is, the risk (under squared error type of loss) of the estimator based on subset selection with a model descriptive complexity term of order $\log \binom{M_n}{m}$ added to the AIC-type of criteria is typically upper bounded in order by the smallest value of

$$\text{(squared) approximation error}_m + \frac{\sigma^2 m}{n} + \frac{\sigma^2 \log \binom{M_n}{m}}{n}$$

over all the subset models, which is called the index of the resolvability of the function to be estimated. Note that $\frac{m}{n} + \frac{\log \binom{M_n}{m}}{n}$ is uniformly of order $m\left(1 + \log\left(\frac{M_n}{m}\right)\right)/n$ over $0 \le m \le M_n$. Evaluating the above bound at m^* in our context yields a quite sensible rate of convergence. Note also that $\log \binom{M_n}{m^*}/n$ (price of searching) is of a higher order than $\frac{m^*}{n}$ (price of estimation) when $m^* \le M_n/2$. Define

$$SER(m) = 1 + \log\left(\frac{M_n}{m}\right) \asymp \frac{m + \log \binom{M_n}{m}}{m}, \qquad 1 \le m \le M_n,$$

to be the ratio of the price with searching to that without searching (i.e., only the price of estimation of the parameters in the model). Here "$\asymp$" means of the same order as $n \to \infty$. Observe that reducing m^* slightly will reduce the order of searching price $\frac{m^* SER(m^*)}{n}$ (since $x(1 + \log(M_n/x))$ is an increasing function for $0 < x < M_n$) and increase the order of the squared bias plus variance (i.e., $2^{2/q} t_n^2 m^{1-2/q} + \frac{\sigma^2 m}{n}$). The best choice will typically make the approximation error $2^{2/q} t_n^2 m^{1-2/q}$ of the same order as $\frac{m(1+\log \frac{M_n}{m})}{n}$ (as also pointed out in Raskutti et al. 2012 from a different analysis). Define

$$m_* = m_*(q, t_n) = \begin{cases} m^* & \text{if } m^* = M_n \wedge n, \\ \left\lceil \frac{m^*}{\left(1+\log \frac{M_n}{m^*}\right)^{q/2}} \right\rceil = \left\lceil \frac{m^*}{SER(m^*)^{q/2}} \right\rceil & \text{otherwise.} \end{cases}$$

We call this the *effective model size* (in order) under the ℓ_q-constraint because evaluating the index of resolvability expression from our general oracle inequality (see Proposition 15 in the Appendix) at the best model of this size gives the minimax rate of convergence, as shown in this work. When $m^* = n$, the minimax risk is of order 1 (or higher sometimes) and thus does not converge. Note that the down-sizing factor $SER(m^*)^{q/2}$ from m^* to m_* depends on q: it becomes more severe as q increases; when $q = 1$, the down-sizing factor reaches the order $\left(1 + \log\left(\frac{M_n}{m^*}\right)\right)^{1/2}$. Since the risk of the ideal model and that by a good model selection rule differ only by a factor of $\log(M_n/m^*)$, as long as M_n is not too large, the price of searching over many models of the same size is small, which is a fact well known in the model selection literature (see, e.g., Yang and Barron 1998, section III.D).

For $q = 0$, under the assumption of at most $k_n \le M_n \wedge n$ nonzero terms in the linear representation of the true regression function, the risk bound immediately yields the rate $\left(1 + \log \binom{M_n}{k_n}\right)/n \asymp \frac{k_n\left(1+\log \frac{M_n}{k_n}\right)}{n}$. Thus, from all above, we expect that $\frac{m_* SER(m_*)}{n} \wedge 1$ is the unifying optimal rate of convergence for regression under the ℓ_q-constraint for $0 \le q \le 1$.

Appendix C.

In this appendix, the theorems in Sections 3-4 are proved, with additional results given as preparations.

C.1 Some General Oracle Inequalities

The proofs of the upper bound results rely on some oracle inequalities, which may be of interest in other applications. Consider the setting in Section 3.2 of the main paper.

Proposition 15 *Suppose $A_{\mathbf{E-G}}$ holds for the **E-G** strategy, respectively. Then, the following oracle inequalities hold for the estimator $\hat{f}_{F_n}$.*

*(i) For **T-C** and **T-Y** strategies,*

$$\begin{aligned} & R(\hat{f}_{F_n}; f_0; n) \\ \leq\ & c_0 \inf_{1\leq m\leq M_n\wedge n}\left(c_1 \inf_{J_m} d^2(f_0;\mathcal{F}_{J_m}) + c_2\frac{m}{n_1} + c_3\frac{1+\log\binom{M_n}{m}+\log(M_n\wedge n)-\log(1-p_0)}{n-n_1}\right) \\ & \wedge c_0\left(\|f_0\|^2 + c_3\frac{1-\log p_0}{n-n_1}\right), \end{aligned}$$

*where $c_0=1$, $c_1=c_2=C_{L,\sigma}$, $c_3=\frac{2}{\lambda_C}$ for the **T-C** strategy; $c_0=C_Y$, $c_1=c_2=C_{L,\sigma}$, $c_3=\sigma^2$ for the **T-Y** strategy.*

*(ii) For **AC-C** and **AC-Y** strategies,*

$$\begin{aligned} & R(\hat{f}_{F_n}; f_0; n) \\ \leq\ & c_0 \inf_{1\leq m\leq M_n\wedge n}\left(R(f_0,m,n) + c_2\frac{m}{n_1} + c_3\frac{1+\log\binom{M_n}{m}+\log(M_n\wedge n)-\log(1-p_0)}{n-n_1}\right) \\ & \wedge c_0\left(\|f_0\|^2 + c_3\frac{1-\log p_0}{n-n_1}\right), \end{aligned}$$

where

$$R(f_0,m,n) = c_1 \inf_{J_m}\inf_{s\geq 1}\left(d^2(f_0;\mathcal{F}^L_{J_m,s}) + 2c_3\frac{\log(1+s)}{n-n_1}\right),$$

*and $c_0=c_1=1$, $c_2=8c(\sigma^2+5L^2)$, $c_3=\frac{2}{\lambda_C}$ for the **AC-C** strategy; $c_0=C_Y$, $c_1=1$, $c_2=8c(\sigma^2+5L^2)$, $c_3=\sigma^2$ for the **AC-Y** strategy.*

From the proposition above, the risk $R(\hat{f}_{F_n}; f_0; n)$ is upper bounded by a multiple of the best trade-off of the different sources of errors (approximation error, estimation error due to estimating the linear coefficients, and error associated with searching over many models of the same dimension). For a model J, let $IR(f_0; J)$ generically denote the sum of these three sources of errors. Then, the best trade-off is $IR(f_0) = \inf_J IR(f_0; J)$, where the infimum is over all the candidate models. Following the terminology in Barron and Cover (1991), $IR(f_0)$ is the so-called index of resolvability of the true function f_0 by the estimation method over the candidate models. We call $IR(f_0; J)$ the index of resolvability at model J. The utility of the index of resolvability is that for f_0 with a given characteristic,

an evaluation of the index of resolvability at the best J immediately tells us how well the unknown function is "resolved" by the estimation method at the current sample size. Thus, accurate index of resolvability bounds often readily show minimax optimal performance of the model selection based estimator.

Proof of Proposition 15.

(i) For the **T-C** strategy,

$$R(\hat{f}_{F_n}; f_0; n) \le \inf_{1\le m\le M_n\wedge n}\left\{C_{L,\sigma}\left(\inf_{J_m} d^2(f_0;\mathcal{F}_{J_m})+\frac{m}{n_1}\right)+ \frac{2}{\lambda_C}\left(\frac{\log(M_n\wedge n)+\log\binom{M_n}{m}-\log(1-p_0)}{n-n_1}\right)\right\}\wedge\left\{\|f_0\|^2-\frac{2}{\lambda_C}\frac{\log p_0}{n-n_1}\right\}.$$

For the **T-Y** strategy,

$$R(\hat{f}_{F_n}; f_0; n) \le C_Y \inf_{1\le m\le M_n\wedge n}\left\{C_{L,\sigma}\inf_{J_m} d^2(f_0;\mathcal{F}_{J_m})+C_{L,\sigma}\frac{m}{n_1}+ \sigma^2\left(\frac{1+\log(M_n\wedge n)+\log\binom{M_n}{m}-\log(1-p_0)}{n-n_1}\right)\right\}\wedge C_Y\left\{\|f_0\|^2+\sigma^2\frac{1-\log p_0}{n-n_1}\right\}.$$

(ii) For the **AC-C** strategy,

$$\begin{aligned} & R(\hat{f}_{F_n}; f_0; n)\\ \le\ & \inf_{1\le m\le M_n\wedge n}\left\{\inf_{J_m}\inf_{s\ge1}\left(d^2(f_0;\mathcal{F}^L_{J_m,s})+c(2\sigma'+H)^2\frac{m}{n_1}+\frac{2}{\lambda_C}\times \left(\frac{\log(M_n\wedge n)+\log\binom{M_n}{m}-\log(1-p_0)}{n-n_1}+\frac{2\log(1+s)}{n-n_1}\right)\right)\right\}\wedge\left\{\|f_0\|^2-\frac{2}{\lambda_C}\frac{\log p_0}{n-n_1}\right\}\\ \le\ & \inf_{1\le m\le M_n\wedge n}\left\{\inf_{J_m}\inf_{s\ge1}\left(d^2(f_0;\mathcal{F}^L_{J_m,s})+8c(\sigma^2+5L^2)\frac{m}{n_1}+\frac{2}{\lambda_C}\times \left(\frac{\log(M_n\wedge n)+\log\binom{M_n}{m}-\log(1-p_0)}{n-n_1}+\frac{2\log(1+s)}{n-n_1}\right)\right)\right\}\wedge\left\{\|f_0\|^2-\frac{2}{\lambda_C}\frac{\log p_0}{n-n_1}\right\}. \end{aligned}$$

For the **AC-Y** strategy,

$$\begin{aligned} & R(\hat{f}_{F_n}; f_0; n)\\ \le\ & C_Y\inf_{1\le m\le M_n\wedge n}\left\{\inf_{J_m}\inf_{s\ge1}\left(d^2(f_0;\mathcal{F}^L_{J_m,s})+c(2\sigma'+H)^2\frac{m}{n_1}+\sigma^2\left(\frac{1+\log(M_n\wedge n)}{n-n_1}+ +\frac{\log\binom{M_n}{m}}{n-n_1}+\frac{-\log(1-p_0)+2\log(1+s)}{n-n_1}\right)\right)\right\}\wedge C_Y\left\{\|f_0\|^2+\sigma^2\frac{1-\log p_0}{n-n_1}\right\}\\ \le\ & C_Y\inf_{1\le m\le M_n\wedge n}\left\{\inf_{J_m}\inf_{s\ge1}\left(d^2(f_0;\mathcal{F}^L_{J_m,s})+8c(\sigma^2+5L^2)\frac{m}{n_1}+\sigma^2\left(\frac{1+\log(M_n\wedge n)}{n-n_1}+ +\frac{\log\binom{M_n}{m}}{n-n_1}+\frac{-\log(1-p_0)+2\log(1+s)}{n-n_1}\right)\right)\right\}\wedge C_Y\left\{\|f_0\|^2+\sigma^2\frac{1-\log p_0}{n-n_1}\right\}. \end{aligned}$$

This completes the proof of Proposition 15.

C.2 Proof of Theorem 5

To derive the upper bounds, we only need to examine the index of resolvability for each strategy. The nature of the constants in Theorem 5 follows from Proposition 15.

(i) For **T-** strategies, according to Theorem 13 and the general oracle inequalities in Proposition 15, for each $1 \le m \le M_n \wedge n$, there exists a subset J_m and the best $f_{\theta^m} \in \mathcal{F}_{J_m}$ such that

$$\begin{aligned} R(\hat{f}_{F_n}; f_0; n) &\le c_0 \left(c_1 \|f_{\theta^m} - f_0\|^2 + 2c_2 \frac{m}{n} + 2c_3 \frac{1 + \log \binom{M_n}{m} + \log(M_n \wedge n) - \log(1-p_0)}{n} \right) \\ &\quad \wedge c_0 \left(\|f_0\|^2 + 2c_3 \frac{1 - \log p_0}{n} \right). \end{aligned}$$

Under the assumption that f_0 has sup-norm bounded, the index of resolvability evaluated at the null model $f_\theta \equiv 0$ leads to the fact that the risk is always bounded above by $C_0 \left(\|f_0\|^2 + \frac{C_2 \sigma^2}{n} \right)$ for some constant $C_0, C_2 > 0$.

For $\mathcal{F} = \mathcal{F}_q(t_n)$, and when $m_* = m^* = M_n < n$, evaluating the index of resolvability at the full model J_{M_n}, we get

$$R(\hat{f}_{F_n}; f_0; n) \le c_0 c_1 d^2(f_0; \mathcal{F}_q(t_n)) + \frac{CM_n}{n} \quad \text{with} \quad \frac{CM_n}{n} = \frac{Cm_* \left(1 + \log\left(\frac{M_n}{m_*}\right)\right)}{n}.$$

Thus, the upper bound is proved when $m_* = m^* = M_n$.

For $\mathcal{F} = \mathcal{F}_q(t_n)$, and when $m_* = m^* = n < M_n$, then clearly $m_* \left(1 + \log\left(\frac{M_n}{m_*}\right)\right)/n$ is larger than 1, and then the risk bound given in the theorem in this case holds.

For $\mathcal{F} = \mathcal{F}_q(t_n)$, and when $1 \le m_* \le m^* < M_n \wedge n$, for $1 \le m < M_n$, and from Theorem 13, we have

$$\begin{aligned} R(\hat{f}_{F_n}; f_0; n) &\le c_0 \Big(c_1 d^2(f_0; \mathcal{F}_q(t_n)) + c_1 2^{2/q-1} t_n^2 m^{1-2/q} + 2c_2 \frac{m}{n} \\ &\quad + 2c_3 \frac{1 + \log \binom{M_n}{m} + \log(M_n \wedge n)}{n} - 2c_3 \frac{\log(1-p_0)}{n} \Big). \end{aligned}$$

Since $\log \binom{M_n}{m} \le m \log\left(\frac{eM_n}{m}\right) = m\left(1 + \log \frac{M_n}{m}\right)$, then

$$\begin{aligned} R(\hat{f}_{F_n}; f_0; n) &\le c_0 c_1 d^2(f_0; \mathcal{F}_q(t_n)) + C \left(2^{2/q} t_n^2 m^{1-2/q} + \frac{m\left(1 + \log \frac{M_n}{m}\right)}{n} + \frac{\log(M_n \wedge n)}{n} \right) \\ &\le c_0 c_1 d^2(f_0; \mathcal{F}_q(t_n)) + C' \left(2^{2/q} t_n^2 m^{1-2/q} + \frac{m\left(1 + \log \frac{M_n}{m}\right)}{n} \right), \end{aligned}$$

where C and C' are constants that do not depend on n, q, t_n, and M_n (but may depend on σ^2, p_0 and L). Choosing $m = m_*$, we have

$$2^{2/q} t_n^2 m^{1-2/q} + \frac{m\left(1+\log\frac{M_n}{m}\right)}{n} \leq C'' \frac{m_*\left(1+\log\left(\frac{M_n}{m_*}\right)\right)}{n},$$

where C'' is an absolute constant. The upper bound for this case then follows.

For $\mathcal{F} = \mathcal{F}_0(k_n)$, by evaluating the index of resolvability from Proposition 15 at $m = k_n$, the upper bound immediately follows.

For $\mathcal{F} = \mathcal{F}_q(t_n) \cap \mathcal{F}_0(k_n)$, both ℓ_q- and ℓ_0-constraints are imposed on the coefficients, the upper bound will go with the faster rate from the tighter constraint. The result follows.

(ii) For **AC-** strategies, three constraints $\|\theta\|_1 \leq s$ ($s > 0$), $\|\theta\|_q \leq t_n$ ($0 \leq q \leq 1$, $t_n > 0$) and $\|f_\theta\|_\infty \leq L$ are imposed on the coefficients. Notice that $\|\theta\|_1 \leq \|\theta\|_q$ when $0 < q \leq 1$, then the ℓ_1-constraint is satisfied by default as long as $s \geq t_n$ and $\|\theta\|_q \leq t_n$ with $0 < q \leq 1$. Using similar arguments as used for **T-**strategies, the desired upper bounds can be easily derived. This completes the proof of Theorem 5.

C.3 Global Metric Entropy and Local Metric Entropy

The derivations of the lower bounds in the main paper require some preparations.

Consider estimating a regression function f_0 in a general function class $\mathcal{F}$ based on i.i.d. observations $(\mathbf{X}_i, Y_i)_{i=1}^n$ from the model

$$Y = f_0(\mathbf{X}) + \sigma \cdot \varepsilon, \tag{5}$$

where $\sigma > 0$ and ε follows a standard normal distribution and is independent of $\mathbf{X}$.

Given $\mathcal{F}$, we say $G \subset \mathcal{F}$ is an *ϵ-packing set* in $\mathcal{F}$ ($\epsilon > 0$) if any two functions in G are more than ϵ apart in the L_2 distance. Let $0 < \alpha < 1$ be a constant.

Definition 1: (*Global metric entropy*) The packing ϵ-entropy of $\mathcal{F}$ is the logarithm of the largest ϵ-packing set in $\mathcal{F}$. The packing ϵ-entropy of $\mathcal{F}$ is denoted by $M(\epsilon)$.

Definition 2: (*Local metric entropy*) The α-local ϵ-entropy at $f \in \mathcal{F}$ is the logarithm of the largest $(\alpha\epsilon)$-packing set in $\mathcal{B}(f, \epsilon) = \{f' \in \mathcal{F} : \| f' - f \| \leq \epsilon\}$. The α-local ϵ-entropy at f is denoted by $M_\alpha(\epsilon \mid f)$. The α-local ϵ-entropy of $\mathcal{F}$ is defined as $M_\alpha^{\text{loc}}(\epsilon) = \max_{f \in \mathcal{F}} M_\alpha(\epsilon \mid f)$.

Suppose that $M_\alpha^{\text{loc}}(\epsilon)$ is lower bounded by $\underline{M}_\alpha^{\text{loc}}(\epsilon)$ (a continuous function), and assume that $M(\epsilon)$ is upper bounded by $\overline{M}(\epsilon)$ and lower bounded by $\underline{M}(\epsilon)$ (with $\overline{M}(\epsilon)$ and $\underline{M}(\epsilon)$ both being continuous).

Suppose there exist ϵ_n, $\bar{\epsilon}_n$, and $\underline{\epsilon}_n$ such that

$$\underline{M}_\alpha^{\text{loc}}(\sigma\epsilon_n) \geq n\epsilon_n^2 + 2\log 2, \tag{6}$$

$$\overline{M}(\sqrt{2}\sigma\bar{\epsilon}_n) = n\bar{\epsilon}_n^2, \tag{7}$$

$$\underline{M}(\sigma\underline{\epsilon}_n) = 4n\bar{\epsilon}_n^2 + 2\log 2. \tag{8}$$

Proposition 16 *(Yang and Barron 1999) The minimax risk for estimating f_0 from model (5) in the function class $\mathcal{F}$ is lower-bounded as the following*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}} E\|\hat{f} - f_0\|^2 \geq \frac{\alpha^2\sigma^2\epsilon_n^2}{8},$$

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}} E\|\hat{f} - f_0\|^2 \geq \frac{\sigma^2 \underline{\epsilon}_n^2}{8}.$$

Let $\underline{\mathcal{F}}$ be a subset of $\mathcal{F}$. If a packing set in $\mathcal{F}$ of size at least $\exp(\underline{M}_\alpha^{loc}(\sigma\epsilon_n))$ or $\exp(\underline{M}(\sigma\underline{\epsilon}_n))$ is actually contained in $\underline{\mathcal{F}}$, then $\inf_{\hat{f}} \sup_{f_0 \in \underline{\mathcal{F}}} E\|\hat{f} - f_0\|^2$ is lower bounded by $\frac{\alpha^2\sigma^2\epsilon_n^2}{8}$ or $\frac{\sigma^2\underline{\epsilon}_n^2}{8}$, respectively.

Proof. The result is essentially given in Yang and Barron (1999), but not in the concrete forms. The second lower bound is given in Yang (2004). We briefly derive the first one.

Let N be an $(\alpha\epsilon_n)$-packing set in $\mathcal{B}(f, \sigma\epsilon_n) = \{f' \in \mathcal{F} : \| f' - f \| \leq \sigma\epsilon_n\}$. Let Θ denote a uniform distribution on N. Then, by applying the upper bound on mutual information displayed in the middle of page 1571 of Yang and Barron (1999), together with the specific form of the K-L divergence between the Gaussian regression densities (see the first paragraph of the proof of Theorem 6 of Yang and Barron 1999 on page 1583), the mutual information between Θ and the observations $(\mathbf{X}_i, Y_i)_{i=1}^n$ is upper bounded by $\frac{n}{2}\epsilon_n^2$, and an application of Fano's inequality (see the proof of Theorem 1 in Yang and Barron 1999, particularly Equation 1 on page 1571) to the regression problem gives the minimax lower bound

$$\frac{\alpha^2\sigma^2\epsilon_n^2}{4}\left(1 - \frac{I\left(\Theta; (\mathbf{X}_i, Y_i)_{i=1}^n\right) + \log 2}{\log|N|}\right),$$

where $|N|$ denotes the size of N. By our way of defining ϵ_n, the conclusion of the first lower bound follows.

For the last statement, we prove for the global entropy case and the argument for the local entropy case similarly follows. Observe that the upper bound on $I\left(\Theta; (\mathbf{X}_i, Y_i)_{i=1}^n\right)$ by $\log(|G|) + n\bar{\epsilon}_n^2$, where G is an $\bar{\epsilon}_n$-net of $\mathcal{F}$ under the square root of the Kullback-Leibler divergence (see Yang and Barron 1999, page 1571), continues to be an upper bound on $I\left(\underline{\Theta}; (\mathbf{X}_i, Y_i)_{i=1}^n\right)$, where $\underline{\Theta}$ is the uniform distribution on a packing set in $\underline{\mathcal{F}}$. Therefore, by the derivation of Theorem 1 in Yang and Barron (1999), the same lower bound holds for $\underline{\mathcal{F}}$ as well. This completes the proof.

C.4 Proof of Theorem 8

Assume $f_0 \in \mathcal{F}$ in each case of $\mathcal{F}$ so that $d^2(f_0; \mathcal{F}) = 0$. Without loss of generality, assume $\sigma = 1$.

(i) We first derive the lower bounds without L_2 or L_∞ upper bound assumption on f_0. To prove case 1 (i.e., $\mathcal{F} = \mathcal{F}_q(t_n)$), it is enough to show that

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_q(t_n)} E\|\hat{f} - f_0\|^2 \geq C_q \begin{cases} \frac{M_n}{n} & \text{if } \widetilde{m}^* = M_n, \\ t_n^q \left(\frac{1+\log\frac{M_n}{(nt_n^2)^{q/2}}}{n}\right)^{1-q/2} & \text{if } 1 < \widetilde{m}_* \leq \widetilde{m}^* < M_n, \\ t_n^2 & \text{if } \widetilde{m}_* = 1, \end{cases}$$

in light of the fact that, by definition, when $\widetilde{m}^* = M_n$, $\widetilde{m}_* = M_n$ and when $1 < \widetilde{m}_* \leq \widetilde{m}^* < M_n$, we have $\frac{\widetilde{m}_*(1+\log\frac{M_n}{\widetilde{m}_*})}{n}$ is upper and lower bounded by multiples (depending only on q)

of $t_n^q \left(\frac{1+\log \frac{M_n}{(nt_n^2)^{q/2}}}{n} \right)^{1-q/2}$. Note that $\widetilde{m}^*$ and $\widetilde{m}_*$ are defined as m^* and m_* except that no ceiling of n is imposed there.

Given that the basis functions are orthonormal, the L_2 distance on $\mathcal{F}_q(t_n)$ is the same as the ℓ_2 distance on the coefficients in $B_q(t_n; M_n) = \{\theta : \|\theta\|_q \le t_n\}$. Thus, the entropy of $\mathcal{F}_q(t_n)$ under the L_2 distance is the same as that of $B_q(t_n; M_n)$ under the ℓ_2 distance.

When $\widetilde{m}^* = M_n$, we use the lower bound tool in terms of local metric entropy. Given the ℓ_q-ℓ_2-relationship $\|\theta\|_q \le {M_n}^{1/q-1/2}\|\theta\|_2$ for $0 < q \le 2$, for $\epsilon \le \sqrt{M_n/n}$, taking $f_0^* \equiv 0$, we have

$$\mathcal{B}(f_0^*; \epsilon) = \{f_\theta : \|f_\theta - f_0^*\| \le \epsilon, \|\theta\|_q \le t_n\} = \{f_\theta : \|\theta\|_2 \le \epsilon, \|\theta\|_q \le t_n\} = \{f_\theta : \|\theta\|_2 \le \epsilon\},$$

where the last equality holds because when $\epsilon \le \sqrt{M_n/n}$, for $\|\theta\|_2 \le \epsilon$, $\|\theta\|_q \le t_n$ is always satisfied. Consequently, for $\epsilon \le \sqrt{M_n/n}$, the $(\epsilon/2)$-packing of $\mathcal{B}(f_0^*; \epsilon)$ under the L_2 distance is equivalent to the $(\epsilon/2)$-packing of $B_\epsilon = \{\theta : \|\theta\|_2 \le \epsilon\}$ under the ℓ_2 distance. Note that the size of the maximum packing set is at least the ratio of volumes of the balls B_ϵ and $B_{\epsilon/2}$, which is 2^{M_n}. Thus, the local entropy $M_{1/2}^{\mathrm{loc}}(\epsilon)$ of $\mathcal{F}_q(t)$ under the L_2 distance is at least $\underline{M}_{1/2}^{\mathrm{loc}}(\epsilon) = M_n \log 2$ for $\epsilon \le \sqrt{M_n/n}$. The minimax lower bound for the case of $\widetilde{m}^* = M_n$ then directly follows from Proposition 16.

When $1 < \widetilde{m}_* \le \widetilde{m}^* < M_n$, the use of global entropy is handy. Applying the minimax lower bound in terms of global entropy in Proposition 16, with the metric entropy order for larger ϵ (which is tight in our case of orthonormal functions in the dictionary) from Theorem 13 the minimax lower rate is readily obtained. Indeed, for the class $\mathcal{F}_q(t_n)$, with $\epsilon > t_n M_n^{\frac{1}{2}-\frac{1}{q}}$, there are constants c' and $\underline{c}'$ (depending only on q) such that

$$\underline{c}' \left(t_n \epsilon^{-1}\right)^{\frac{2q}{2-q}} \log(1 + M_n^{\frac{1}{q}-\frac{1}{2}} t_n^{-1}\epsilon) \le \underline{M}(\epsilon) \le \overline{M}(\epsilon) \le c' \left(t_n \epsilon^{-1}\right)^{\frac{2q}{2-q}} \log(1 + M_n^{\frac{1}{q}-\frac{1}{2}} t_n^{-1}\epsilon).$$

Thus, we see that $\underline{\epsilon}_n$ determined by (8.4) is lower bounded by $c''' t_n^{\frac{q}{2}} \left((1 + \log \frac{M_n}{(nt_n^2)^{q/2}})/n \right)^{\frac{1}{2}-\frac{q}{4}}$, where c''' is a constant depending only on q.

When $\widetilde{m}_* = 1$, note that with $f_0^* = 0$ and $\epsilon \le t_n$,

$$\mathcal{B}(f_0^*; \epsilon) = \{f_\theta : \|\theta\|_2 \le \epsilon, \|\theta\|_q \le t_n\} \supset \{f_\theta : \|\theta\|_q \le \epsilon\}.$$

Observe that the $(\epsilon/2)$-packing of $\{f_\theta : \|\theta\|_q \le \epsilon\}$ under the L_2 distance is equivalent to the $(1/2)$-packing of $\{f_\theta : \|\theta\|_q \le 1\}$ under the same distance. Thus, by applying Theorem 13 with $t_n = 1$ and $\epsilon = 1/2$, we know that the $(\epsilon/2)$-packing entropy of $\mathcal{B}(f_0^*; \epsilon)$ is lower bounded by $\underline{c}'' \log(1 + \frac{1}{2} M_n^{1/q-1/2})$ for some constant $\underline{c}''$ depending only on q, which is at least a multiple of nt_n^2 when $\widetilde{m}^* \le \left(1 + \log \frac{M_n}{\widetilde{m}^*}\right)^{q/2}$. Therefore we can choose $0 < \delta < 1$ small enough (depending only on q) such that

$$\underline{c}'' \log(1 + \frac{1}{2} M_n^{1/q-1/2}) \ge n\delta^2 t_n^2 + 2 \log 2.$$

The conclusion then follows from applying the first lower bound of Proposition 16.

To prove case 2 (i.e., $\mathcal{F} = \mathcal{F}_0(k_n)$), noticing that for $M_n/2 \le k_n \le M_n$, we have $(1+\log 2)/2M_n \le k_n\left(1+\log\frac{M_n}{k_n}\right) \le M_n$, together with the monotonicity of the minimax risk in the function class, it suffices to show the lower bound for $k_n \le M_n/2$. Let $B_{k_n}(\epsilon) = \{\theta : \|\theta\|_2 \le \epsilon, \|\theta\|_0 \le k_n\}$. As in case 1, we only need to understand the local entropy of the set $B_{k_n}(\epsilon)$ for the critical ϵ that gives the claimed lower rate. Let $\eta = \epsilon/\sqrt{k_n}$. Then $B_{k_n}(\epsilon)$ contains the set $D_{k_n}(\eta)$, where

$$D_k(\eta) = \{\theta = \eta I : I \in \{1,0,-1\}^{M_n}, \|I\|_0 \le k\}.$$

Clearly $\|\eta I_1 - \eta I_2\|_2 \ge \eta\left(d_{HM}(I_1,I_2)\right)^{1/2}$, where $d_{HM}(I_1,I_2)$ is the Hamming distance between $I_1, I_2 \in \{1,0,-1\}^{M_n}$. From Lemma 4 of Raskutti et al. (2012) (the result there actually also holds when requiring the pairwise Hamming distance to be strictly larger than $k/2$; see also Lemma 4 of Birgé and Massart 2001 or the derivation of a metric entropy lower bound in Kühn 2001), there exists a subset of $\{I : I \in \{1,0,-1\}^{M_n}, \|I\|_0 \le k\}$ with more than $\exp\left(\frac{k}{2}\log\frac{2(M_n-k)}{k}\right)$ points that have pairwise Hamming distance larger than $k/2$. Consequently, we know the local entropy $M^{loc}_{1/\sqrt{2}}(\epsilon)$ of $\mathcal{F}_0(k_n)$ is lower bounded by $\frac{k_n}{2}\log\frac{2(M_n-k_n)}{k_n}$. The result follows.

To prove case 3 (i.e., $\mathcal{F}_q(t_n) \cap \mathcal{F}_0(k_n)$), for the larger k_n case, from the proof of case 1, we have used fewer than k_n nonzero components to derive the minimax lower bound there. Thus, the extra ℓ_0-constraint does not change the problem in terms of lower bound. For the smaller k_n case, note that for θ with $\|\theta\|_0 \le k_n$, $\|\theta\|_q \le k_n^{1/q-1/2}\|\theta\|_2 \le k_n^{1/q-1/2}\cdot\sqrt{Ck_n\left(1+\log\frac{M_n}{k_n}\right)/n}$ for θ with $\|\theta\|_2 \le \sqrt{Ck_n\left(1+\log\frac{M_n}{k_n}\right)/n}$ for some constant $C > 0$. Therefore the ℓ_q-constraint is automatically satisfied when $\|\theta\|_2$ is no larger than the critical order $\sqrt{k_n\left(1+\log\frac{M_n}{k_n}\right)/n}$, which is sufficient for the lower bound via local entropy techniques. The conclusion follows.

(ii) Now, we turn to the lower bounds under the L_2-norm condition. When the regression function f_0 satisfies the boundedness condition in L_2-norm, the estimation risk is obviously upper bounded by L^2 by taking the trivial estimator $\hat{f} = 0$. In all of the lower boundings in (i) through local entropy argument, if the critical radius ϵ is of order 1 or lower, the extra condition $\|f_0\| \le L$ does not affect the validity of the lower bound. Otherwise, we take ϵ to be L. Then, since the local entropy stays the same, it directly follows from the first lower bound in Proposition 16 that L^2 is a lower order of the minimax risk. The only case remained is that of $\left(1+\log\frac{M_n}{m^*}\right)^{q/2} \le m^* < M_n$. If $t_n^q\left((1+\log\frac{M_n}{(nt^2)^{q/2}})/n\right)^{1-q/2}$ is upper bounded by a constant, from the proof of the lower bound of the metric entropy of the ℓ_q-ball in Kühn (2001), we know that the functions in the special packing set satisfy the L_2 bound. Indeed, consider $\{f_\theta : \theta \in D_{m_n}(\eta)\}$ with m_n being a multiple of $\left(nt_n^2/\left(1+\log\frac{M_n}{(nt_n^2)^{q/2}}\right)\right)^{q/2}$ and η being a (small enough) multiple of $\sqrt{(1+\log\frac{M_n}{(nt_n^2)^{q/2}})/n}$. Then these f_θ have $\|f_\theta\|$ upper bounded by a multiple of $t_n^q\left((1+\log\frac{M_n}{(nt_n^2)^{q/2}})/n\right)^{1-q/2}$ and the minimax lower bound follows from the last statement of Proposition 16. If $t_n^q\left((1+\log\frac{M_n}{(nt_n^2)^{q/2}})/n\right)^{1-q/2}$ is not upper bounded,

we reduce the packing radius to L (i.e., choose η so that $\eta\sqrt{m_n}$ is bounded by a multiple of L). Then the functions in the packing set satisfy the L_2 bound and furthermore, the number of points in the packing set is of a larger order than $nt_n^q\left((1+\log\frac{M_n}{(nt_n^2)^{q/2}})/n\right)^{1-q/2}$. Again, adding the L_2 condition on $f_0 \in \mathcal{F}_q(t)$ does not increase the mutual information bound in our application of Fano's inequality. We conclude that the minimax risk is lower bounded by a constant.

(iii) Finally, we prove the lower bounds under the sup-norm bound condition. For 1), under the direct sup-norm assumption, the lower bound is obvious. For the general M_n case 2), note that the functions f_θ's in the critical packing set satisfies that $\|\theta\|_2 \le \epsilon$ with ϵ being a multiple of $\sqrt{\frac{k_n\left(1+\log\frac{M_n}{k_n}\right)}{n}}$. Then together with $\|\theta\|_0 \le k_n$, we have $\|\theta\|_1 \le \sqrt{k_n}\|\theta\|_2$, which is bounded by assumption. The lower bound conclusion then follows from the last part of Proposition 16. To prove the results for the case $M_n/\left(1+\log\frac{M_n}{k_n}\right) \le bn$, as in Tsybakov (2003), we consider the special dictionary $F_n = \{f_i : 1 \le i \le M_n\}$ on $[0,1]$, where

$$f_i(\mathbf{x}) = \sqrt{M_n} I_{[\frac{i-1}{M_n}, \frac{i}{M_n})}(\mathbf{x}), \qquad i = 1, ..., M_n.$$

Clearly, these functions are orthonormal. By the last statement of Proposition 16, we only need to verify that the functions in the critical packing set in each case do have the sup-norm bound condition satisfied. Note that for any f_θ with $\theta \in D_{k_n}(\eta)$ (as defined earlier), we have $\|f_\theta\| \le \eta\sqrt{k_n}$ and $\|f_\theta\|_\infty \le \eta\sqrt{M_n}$. Thus, it suffices to show that the critical packing sets for the previous lower bounds without the sup-norm bound can be chosen with θ in $D_{k_n}(\eta)$ for some $\eta = O\left(M_n^{-1/2}\right)$. Consider η to be a (small enough) multiple of $\sqrt{\left(1+\log\frac{M_n}{k_n}\right)/n} = O\left(M_n^{-1/2}\right)$ (which holds under the assumption $\frac{M_n}{1+\log\frac{M_n}{k_n}} \le bn$). From the proof of part (ii) without constraint, we know that there is a subset of $D_{k_n}(\eta)$ that with more than $\exp(\frac{k_n}{2}\log\frac{2(M_n-k_n)}{k_n})$ points that are separated in ℓ_2 distance by at least $\sqrt{k_n\left(1+\log\frac{M_n}{k_n}\right)/n}$. This completes the proof.

C.5 Proof of Corollary 10

Since f_0 belongs to $\mathcal{F}_q^L(t_n; M_n)$, or $\mathcal{F}_0^L(k_n; M_n)$, or both, thus $d^2(f_0, \mathcal{F})$ is equal to zero for all cases (except for **AC-** strategies when $\mathcal{F} = \mathcal{F}_0^L(k_n; M_n)$, which we discuss later).

(i) For **T-** strategies and $\mathcal{F} = \mathcal{F}_q^L(t_n; M_n)$. For each $1 \le m \le M_n \wedge n$, according to the general oracle inequalities in the proof of Theorem 5, the adaptive estimator $\hat{f}_A$ has

$$\begin{aligned}\sup_{f_0\in\mathcal{F}} R(\hat{f}_A; f_0; n) \quad &\le \quad c_0\left(2c_2\frac{m}{n} + 2c_3\frac{1+\log\binom{M_n}{m} + \log(M_n\wedge n) - \log(1-p_0)}{n}\right)\\ &\quad \wedge c_0\left(\|f_0\|^2 - 2c_3\frac{\log p_0}{n}\right).\end{aligned}$$

When $m_* = m^* = M_n < n$, the full model J_{M_n} results in an upper bound of order M_n/n.

When $m_* = m^* = n < M_n$, we choose the null model and the upper bound is simply of order one.

When $1 < m_* \le m^* < M_n \wedge n$, the similar argument of Theorem 5 leads to an upper bound of order $1 \wedge \frac{m_*}{n}\left(1+\log\frac{M_n}{m_*}\right)$. Since $(nt_n^2)^{q/2}\left(1+\log\frac{M_n}{(nt_n^2)^{q/2}}\right)^{-q/2} \le m_* \le 4(nt_n^2)^{q/2}\left(1+\log\frac{M_n}{2(nt_n^2)^{q/2}}\right)^{-q/2}$, then the upper bound is further upper bounded by $c_q t_n^q \cdot \left(\frac{1+\log\frac{M_n}{(nt_n^2)^{q/2}}}{n}\right)^{1-q/2}$ for some constant c_q only depending on q.

When $m_* = 1$, the null model leads to an upper bound of order $\|f_0\|^2 + \frac{1}{n} \le t_n^2 + \frac{1}{n} \le 2(t_n^2 \vee \frac{1}{n})$ if $f_0 \in \mathcal{F}_q^L(t_n; M_n)$.

For $\mathcal{F} = \mathcal{F}_0^L(k_n; M_n)$ or $\mathcal{F} = \mathcal{F}_q^L(t_n; M_n) \cap \mathcal{F}_0^L(k_n; M_n)$, one can use the same argument as in Theorem 5.

(ii) For **AC-** strategies, for $\mathcal{F} = \mathcal{F}_q^L(t_n; M_n)$ or $\mathcal{F} = \mathcal{F}_q^L(t_n; M_n) \cap \mathcal{F}_0^L(k_n; M_n)$, again one can use the same argument as in the proof of Theorem 5. For $\mathcal{F} = \mathcal{F}_0^L(k_n; M_n)$, the approximation error is $\inf_{s\ge 1}\left(\inf_{\{\theta:\|\theta\|_1\le s, \|\theta\|_0\le k_n, \|f_\theta\|_\infty \le L\}} \|f_\theta - f_0\|^2 + 2c_3\frac{\log(1+s)}{n}\right) \le \inf_{\{\theta:\|\theta\|_1\le \alpha_n, \|\theta\|_0\le k_n, \|f_\theta\|_\infty \le L\}} \|f_\theta - f_0\|^2 + 2c_3\frac{\log(1+\alpha_n)}{n} = 2c_3\frac{\log(1+\alpha_n)}{n}$ if $f_0 \in \mathcal{F}_0^L(k_n; M_n)$. The upper bound then follows. This completes the proof.

C.6 Proof of Theorem 11

Without loss of generality, we assume $\sigma^2 = 1$ for the error variance. First, we give a simple fact. Let $B_k(\eta) = \{\theta : \|\theta\|_2 \le \eta, \|\theta\|_0 \le k\}$ and $\mathcal{B}_k(f_0; \epsilon) = \{f_\theta : \|f_\theta\| \le \epsilon, \|\theta\|_0 \le k\}$ (take $f_0 = 0$). Then, under Assumption SRC with $\gamma = k$, the $\frac{a}{2\overline{a}}$-local ϵ-packing entropy of $\mathcal{B}_k(f_0; \epsilon)$ is lower bounded by the $\frac{1}{2}$-local η-packing entropy of $B_k(\eta)$ with $\eta = \epsilon/\overline{a}$.

(i) The proof is essentially the same as that of Theorem 8. When $m^* = M_n$, the previous lower bounding method works with a slight modification. When $\left(1+\log\frac{M_n}{m^*}\right)^{q/2} < m^* < M_n$, we again use the global entropy to derive the lower bound based on Proposition 16. The key is to realize that in the derivation of the metric entropy lower bound for $\{\theta : \|\theta\|_q \le t_n\}$ in Kühn (2001), an optimal size packing set is constructed in which every member has at most m_* non-zero coefficients. Assumption SRC with $\gamma = m_*$ ensures that the L_2 distance on this packing set is equivalent to the ℓ_2 distance on the coefficients and then we know the metric entropy of $\mathcal{F}_q(t_n; M_n)$ under the L_2 distance is at the order given. The result follows as before. When $m^* \le \left(1+\log\frac{M_n}{m^*}\right)^{q/2}$, observe that $\mathcal{F}_q(t_n; M_n) \supset \{\beta x_j : |\beta| \le t_n\}$ for any $1 \le j \le M_n$. The use of the local entropy result in Proposition 16 readily gives the desired result.

(ii) As in the proof of Theorem 8, without loss of generality, we can assume $k_n \le M_n/2$. Together with the simple fact given at the beginning of the proof, for $B_{k_n}(\epsilon/\overline{a}) = \{\theta : \|\theta\|_2 \le \epsilon/\overline{a}, \|\theta\|_0 \le k_n\}$, with $\eta' = \epsilon/(\overline{a}\sqrt{k_n})$, we know $B_{k_n}(\epsilon/\overline{a})$ contains the set

$$\{\theta = \eta' I : I \in \{1, 0, -1\}^{M_n}, \|I\|_0 \le k_n\}.$$

For $\theta_1 = \eta' I_1, \theta_2 = \eta' I_2$ both in the above set, by Assumption SRC, $\|f_{\theta_1} - f_{\theta_2}\|^2 \ge \underline{a}^2 \eta'^2 d_{HM}(I_1, I_2) \ge \underline{a}^2\epsilon^2/(2\overline{a}^2)$ when the Hamming distance $d_{HM}(I_1, I_2)$ is larger than $k_n/2$.

With the derivation in the proof of part (i) of Theorem 8 (case 2), we know the local entropy $M^{loc}_{\underline{a}/(\sqrt{2}\bar{a})}(\epsilon)$ of $\mathcal{F}_0(k_n; M_n) \cap \{f_\theta : \|\theta\|_2 \le a_n\}$ with $a_n \ge \epsilon$ is lower bounded by $\frac{k_n}{2} \log \frac{2(M_n - k_n)}{k_n}$. Then, under the condition $a_n \ge C\sqrt{k_n \left(1 + \log \frac{M_n}{k_n}\right)/n}$ for some constant C, the minimax lower rate $k_n \left(1 + \log \frac{M_n}{k_n}\right)/n$ follows from a slight modification of the proof of Theorem 8 with $\epsilon = C'\sqrt{k_n \left(1 + \log \frac{M_n}{k_n}\right)/n}$ for some constant $C' > 0$. When $0 < a_n < C\sqrt{k_n \left(1 + \log \frac{M_n}{k_n}\right)/n}$, with ϵ of order a_n, the lower bound follows.

(iii) For the larger k_n case, from the proof of part (i) of the theorem, we have used fewer than k_n nonzero components to derive the minimax lower bound there. Thus, the extra ℓ_0-constraint does not change the problem in terms of lower bound. For the smaller k_n case, note that for θ with $\|\theta\|_0 \le k_n$, $\|\theta\|_q \le k_n^{1/q-1/2}\|\theta\|_2 \le k_n^{1/q-1/2}\sqrt{Ck_n \left(1 + \log \frac{M_n}{k_n}\right)/n}$ for θ with $\|\theta\|_2 \le \sqrt{Ck_n \left(1 + \log \frac{M_n}{k_n}\right)/n}$. Therefore the ℓ_q-constraint is automatically satisfied when $\|\theta\|_2$ is no larger than the critical order $\sqrt{k_n \left(1 + \log \frac{M_n}{k_n}\right)/n}$, which is sufficient for the lower bound via local entropy techniques. The conclusion follows. This completes the proof.

C.7 Proof of Corollary 12

(i) We only need to derive the lower bound part. Under the assumptions that $\sup_j \|X_j\|_\infty \le L_0 < \infty$ for some constant $L_0 > 0$, for a fixed $t_n = t > 0$, we have $\forall f_\theta \in \mathcal{F}_q(t_n; M_n)$, $\|f_\theta\|_\infty \le \sup_j \|X_j\|_\infty \cdot \sum_{j=1}^{M_n} |\theta_j| \le L_0\|\theta\|_1 \le L_0\|\theta\|_q \le L_0 t$. Then the conclusion follows directly from Theorem 11 (Part (i)). Note that when t_n is fixed, the case $m_* = 1$ does not need to be separately considered.

(ii) For the upper rate part, we use the **AC-C** upper bound. For f_θ with $\|\theta\|_\infty \le L_0$, clearly, we have $\|\theta\|_1 \le M_n L_0$, and consequently, since $\log(1 + M_n L_0)$ is upper bounded by a multiple of $k_n \left(1 + \log \frac{M_n}{k_n}\right)$, the upper rate $\frac{k_n}{n}\left(1 + \log \frac{M_n}{k_n}\right) \wedge 1$ is obtained from Corollary 10. Under the assumptions that $\sup_j \|X_j\|_\infty \le L_0 < \infty$ and $k_n\sqrt{\left(1 + \log \frac{M_n}{k_n}\right)/n} \le \sqrt{K_0}$, we know that $\forall f_\theta \in \mathcal{F}_0(k_n; M_n) \bigcap \{f_\theta : \|\theta\|_2 \le a_n\}$ with $a_n = C\sqrt{k_n \left(1 + \log \frac{M_n}{k_n}\right)/n}$ for some constant $C > 0$, the sup-norm of f_θ is upper bounded by

$$\|\sum_{j=1}^{M_n} \theta_j x_j\|_\infty \le L_0\|\theta\|_1 \le L_0\sqrt{k_n} a_n = CL_0 k_n \sqrt{\frac{1 + \log \frac{M_n}{k_n}}{n}} \le C\sqrt{K_0} L_0.$$

Then the functions in $\mathcal{F}_0(k_n; M_n) \bigcap \{f : \|\theta\|_2 \le a_n\}$ have sup-norm uniformly bounded. Note that for bounded a_n, $\|\theta\|_2 \le a_n$ implies that $\|\theta\|_\infty \le a_n$. Thus, the extra restriction $\|\theta\|_\infty \le L_0$ does not affect the minimax lower rate established in part (ii) of Theorem 11.

(iii) The upper and lower rates follow similarly from Corollary 10 and Theorem 11. The details are thus skipped. This completes the proof.

References

J.-Y. Audibert. No fast exponential deviation inequalities for the progressive mixture rule. Available at `http://arxiv.org/abs/math.ST/0703848v1`, 2007.

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Annals of Statistics*, 37:1591–1646, 2009.

J.-Y. Audibert and O. Catoni. Risk bounds in linear regression through PAC-Bayesian truncation. Preprint at arXiv:1010.0072, 2010.

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

Y. Baraud. Model selection for regression on a random design. *ESAIM Probab.Statist.*, 6: 127–146, 2002.

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

A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields*, 97:113–150, 1986.

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

L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. I. H. Poincaré*, 42:273–325, 2006.

L. Birgé. Model selection for density estimation with L_2-loss. *Probab. Theory Related Fields*, 158:533–574, 2014.

L. Birgé and P. Massart. Gaussian model selection. *Journal of European Math. Society*, 3: 203–268, 2001.

F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Transactions on Information Theory*, 54:1725–1735, 2008.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007.

O. Catoni. The mixture approach to universal model selection. Preprint, LMENS-97-30, Ecole Normale Supèrieure, Paris, France, 1997.

O. Catoni. Universal aggregation rules with exact bias bounds. Preprint 510, Laboratoire de Probabilitès et Modèles Aléatoires, Univ. Paris 6 and Paris 7, France. Available at `http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999`, 1999.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*, volume 1851. Springer, New York, 2004.

O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Ann. I. H. Poincaré*, 48:909–1244, 2012.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Annals of Statistics*, 40:2327–2355, 2012.

A. S. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. *Lecture Notes in Computer Science*, 4539:97–111, 2007.

A. S. Dalalyan and A.B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte Carlo. *Journal of Computer and System Science*, 78:1423–1443, 2012a.

A. S. Dalalyan and A.B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18: 914–944, 2012b.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

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

D. E. Edmunds and H. Triebel. Function spaces, entropy numbers, and differential operators. *Cambridge Tracts in Mathematics*, 120, 1998.

C. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14: 1089–1107, 2008.

A. Goldenshluger. A universal procedure for aggregating estimators. *Annals of Statistics*, 37:542–568, 2009.

L. Györfi and Gy. Ottucsák. Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory*, 53:1866–1872, 2007.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric estimation. *Annals of Statistics*, 28:681–719, 2000.

A. Juditsky, P. Rigollet, and A.B. Tsybakov. Learning by mirror averaging. *Annals of Statistics*, 36:2183–2206, 2008.

T. Kühn. A lower estimate for entropy numbers. *Journal of Approximation Theory*, 110: 120–124, 2001.

G. Leung and A.R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52:3396–3410, 2006.

K. Lounici. Generalized mirror averaging and d-convex aggregation. *Mathematical methods of statistic*, 16:246–259, 2007.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896. Lecture Notes in Mathematics, Springer, Berlin/Heidelberg, 2007.

S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27:538–557, 2012.

A. Nemirovski. Topics in non-parametric statistics. Lecture Notes in Mathematics, Springer, New York, ecole d'eté de probabilités de saint-flour 1998 edition, 2000.

G. Pisier. Remarques sur un résultat non publié de b. maure. *Sém. d'Analyse Fonctionnelle 1980/1981, Exp. V.*, 1981.

G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57:6976–699, 2012.

P. Rigollet and A.B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Annals of Statistics*, 39:731–771, 2010.

A.B. Tsybakov. Optimal rates of aggregation. *Annals of Statistics*, 39:731–771, 2003.

S.A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Statist*, 3:1360–1392, 2009.

Z. Wang, S. Paterlini, F. Gao, and Y. Yang. Adaptive minimax estimation over sparse ℓ_q-hulls. Arxiv preprint arXiv:1108.1961, 2011.

M. Wegkamp. Model selection in nonparametric regression. *Annals of Statistics*, 31:252–273, 2003.

Y. Yang. Minimax optimal density estimation. Ph.D. Dissertation, Department of Statistics, Yale University, 1996.

Y. Yang. Model selection for nonparametric regression. *Statistica Sinica*, 9:475–499, 1999.

Y. Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74:135–161, 2000a.

Y. Yang. Mixing strategies for density estimation. *Annals of Statistics*, 28:75–87, 2000b.

Y. Yang. Adaptive regression by mixing. *Journal of American Statistical Association*, 96: 574–588, 2001.

Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10:25–47, 2004. An older version of the paper is Preprint #1999-17 of Department of Statistics at Iowa State University.

Y. Yang and A.R. Barron. An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory*, 44:95–116, 1998.

Y. Yang and A.R. Barron. Information theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27:1564–1599, 1999.

C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942, 2010.